Impact de la compression des normales sur l'évaluation de qualité subjective de nuages de points éclairés

A. Tious, T. Vigier, V. Ricordel
Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000
Nantes, France
amar.tious@univ-nantes.fr

Résumé

L'avènement de nouveaux cas d'usage intégrant des représentations 3D de contenus naturels a mené au développement de nouvelles méthodes pour la compression, le rendu et l'évaluation de qualité de nuage de points (point clouds). Cependant, ces méthodes ne traitent que les coordonnées et la couleur des points, et ne prennent pas en compte d'autres attributs pouvant jouer un rôle dans l'optimisation du rendu de point clouds. Dans cet article, nous nous intéressons aux normales des points qui permettent un rendu avec ombrage des point clouds. Nous ajoutons le traitement des normales au codec V-PCC en considérant 2 approches : encoder les normales comme attributs, ou bien les recalculer après décodage à partir du point cloud décompressé. Nous comparons ces 2 approches en terme de qualité visuelle dans une expérience d'évaluation subjective de point clouds avec ombrage. Nous démontrons alors les avantages que présente la compression des normales pour différents types de contenus et cas d'usages. Nous concluons en identifiant les prochaines étapes vers un rendu physique réaliste des point clouds.

Mots clefs

Nuage de Points, Point cloud, Compression vidéo, Évaluation de qualité subjective, Rendu physique réaliste, V-PCC

1 Introduction

L'imagerie volumétrique permet de représenter des scènes et objets réels en 3D. Ces données peuvent être capturées avec des scanners LiDAR pour de grandes scènes ou des studios et rigs multi-caméras pour les objets et personnes. Un nuage de points (point cloud) décrit ce type de capture comme un ensemble de points caractérisés par leurs coordonnées 3D et divers attributs décrivant la texture et la surface du contenu. C'est une bonne alternative aux maillages 3D (mesh), car plus facile à stocker et moins coûteux en calcul [1]. Suite aux avancées technologiques des réalités étendues, l'utilisation de point clouds est devenue intéressante dans des domaines comme la vidéo 3D immersive, l'architecture, les musées virtuels et la télécommunication [2]. Ces applications nécessitent de nouvelles méthodes de compression des données, de rendu et d'évaluation de qualité. A ce sujet, les groupes de travail MPEG-I [1] et JPEG Pleno [3] ont défini des normes pour la compression et pour les conditions de tests subjectifs d'évaluation de qualité de point clouds. Les tests subjectifs consistent à faire évaluer la qualité visuelle de point clouds par des observateurs humains afin d'évaluer l'effet des déformations causées par des méthodes de compression, de rendu ou de reconstruction sur le contenu original.
Cependant, la plupart des études d'évaluation de qualité considèrent uniquement la compression des coordonnées et des couleurs des points sans prendre en compte d'autres attributs ou des méthodes de rendu avancées qui peuvent pourtant avoir un impact important sur la qualité perçue du contenu 3D [4]. Rossoni et al. [5] proposent un pipeline théorique tenant compte des propriétés de surface et de matériau pour un rendu physique réaliste (PBR). Cependant, aucune implémentation n'a été réalisée jusqu'à présent, et donc aucune étude d'évaluation de qualité n'a pu valider son intérêt.

FIGURE 1 – *Les sources sélectionnées, rendues sans ombrage.*

Pour cet article, nous avons mené une étude d'évaluation de qualité subjective visant à comparer deux approches permettant de considérer les normales (coordonnées des vecteurs normaux aux plans de surface des points) en plus des coordonnées et des couleurs dans la compression et le rendu de point clouds. La première approche est celle que nous proposons dans [6] qui est d'encoder les normales en tant qu'attribut comme les couleurs. La seconde approche est de recalculer les normales après compression à partir de la géométrie déformée du point cloud décodé, comme fait précédemment dans [7]. La seconde méthode est à ce jour le cas de figure utilisé dans les rares cas où

TABLEAU 1 – *Débits des niveaux de qualité en bits/point (géométrie+normales+couleurs). nE n'a pas de débit car les normales sont calculées et non transmises.*

			Normales			
			n1	n3	n5	nE
			0.07	0.15	0.36	0
Géométrie	g1	0.03	7.20	7.28	7.49	7.13
	g3	0.05	7.22	7.30	7.51	7.15
	g5	0.12	7.29	7.37	7.58	7.22

les normales sont utilisées après transmission [7]. Le codec V-PCC (video-based point cloud compression) étant le standard de MPEG-I offrant le meilleur rapport qualité visuelle/coût de transmission [8, 9], nous concentrons notre étude sur l'utilisation de ce codec. Notre objectif est d'évaluer l'intérêt d'encoder et transmettre les normales plutôt que de les recalculer dans le cas où les objets seront ensuite rendus en 3D avec un éclairage virtuel.

2 Méthodologie

2.1 Génération de Stimuli

Pour prendre en compte différents cas de contenu et d'applications, nous avons choisi des point clouds sources de différentes catégories sémantiques, parmi ceux de BASICS [10], l'un des datasets d'évaluation de qualité de point cloud le plus riche à ce jour. Six sources ont été sélectionnées parmi celles du dataset selon 3 catégories (2 "Humans", 2 "Objects", 2 "Buildings"). Tous ces modèles sont des numérisations de contenus réels possédant des normales de surfaces en attributs, et ayant été voxelisés dans une grille de 1024x2024x2024 points. La figure 1 présente les six modèles choisis.

Pour chaque source, nous avons généré 12 versions déformées en nous inspirant des niveaux de qualité recommandés par [3]. Pour observer séparément l'impact de la compression des différents paramètres, nous croisons 3 niveaux de qualité des normales ($n1, n3, n5$) avec 3 niveaux de qualité de géométrie ($g1, g3, g5$). Pour chaque niveau de géométrie, nous avons également généré un niveau de qualité nE pour lequel les normales ont été recalculées plutôt que compressées, puisque l'estimation de celles-ci est basée sur la géométrie. Étant donné que la distorsion de la couleur des points impacterait trop fortement la qualité perçue [10], nous maintenons un niveau de compression sans perte pour cet attribut (débit moyen=7.1bpp). Ainsi, les distorsions sont causées seulement par la compression des normales et des coordonnées des points. Le Tableau 1 résume tous les niveaux de qualité utilisés en indiquant, pour chacun, la somme des débits de la géométrie, des normales et des couleurs. Il est important de noter que la couleur, même si ajustée à un niveau de compression juste suffisant, serait toujours bien plus coûteuse en débit que la géométrie. La méthode que nous utilisons pour compresser et estimer les normales est décrite dans notre précédent article [6].

FIGURE 2 – *Exemple de point cloud déformé avec des normales compressées, et recalculées.*

2.2 Condition de tests

Nous avons suivi les conditions de tests recommandées par ISO/IEC N91058 [3] pour l'évaluation de qualité de point cloud. Les participants étaient placés face à un écran IPS Dell 24" (P2422H) à une distance de 3 fois la hauteur de l'écran (soit 90cm). La luminosité et le gamut de l'écran ainsi que l'éclairage de la salle de test ont été calibrés selon la norme ITU-R BT.500 [11]. La tâche des participants était de noter les déformations des stimuli sur l'échelle DSIS (Double Stimulus Impairment Scale)[11] à 5 niveaux (*Très gênant, Gênant, Légèrement gênant, Perceptible mais pas gênant, Imperceptible*). Les point clouds ont été présentés aux observateurs dans des vidéos de 10 secondes capturées par une caméra virtuelle tournant autour de l'objet. La trajectoire de la caméra est la même que dans l'expérience de [10] : 420° autour de l'axe y (-30° à 390°) et 60° autour de l'axe x (0° à 60°). Pour rendre visible la distorsion des normales, nous avons généré ces vidéos via Unity3D, en utilisant un shader avec une taille de point ajustable et un ombrage de surface [5, 12] afin que les objets soient éclairés et pleins. La scène enregistrée est éclairée par une lumière ambiante à 25% d'intensité et une lumière directionnelle à 100% d'intensité située 30° au-dessus et face à l'objet, comme conseillé par [13] pour simuler un éclairage naturel. L'interface de test est illustrée en Figure 3.

FIGURE 3 – *Capture d'écran de l'interface de test.*

Pour familiariser les participants avec l'interface et la tâche, la session de test était précédée d'une session d'entraînement pendant laquelle 6 versions déformées d'une source non séléctionnée pour le test étaient montrées. Pendant la session principale, chaque participant devait noter les 72 (6x12) versions déformées de point clouds, résultant en une durée moyenne de test de 15 minutes.

Nous avons collecté les scores de 20 utilisateurs au total (14 hommes et 6 femmes) âgés entre 21 et 49 ans, ayant tous une vision correcte ou corrigée, dont 45% sont non-experts, et 55% sont experts mais non familiers au type de stimulus utilisé.

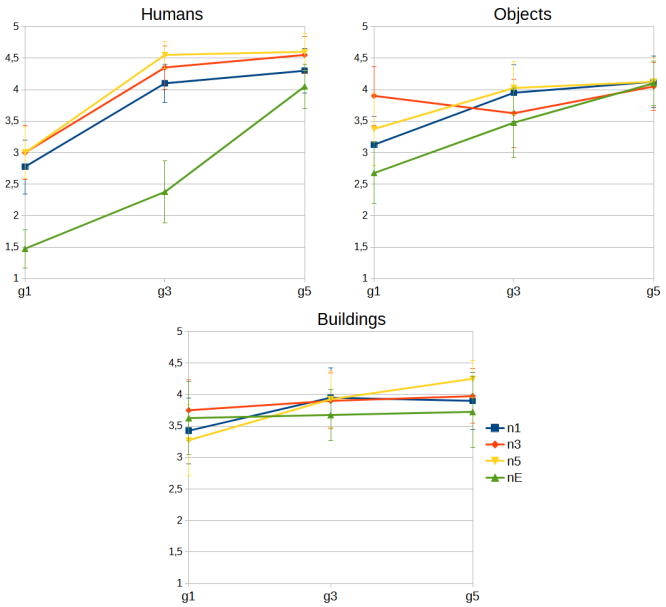


FIGURE 4 – *MOS pour tous les niveaux de qualité et chaque catégorie de contenu (avec intervalle de confiance de 95%.*

3 Résultats & Discussion

3.1 Scores de qualité et Discriminabilité

Les scores d'opinion moyen (MOS) pour chaque niveau de qualité et pour chaque catégorie de contenu sont représentés en Figure 4, avec les intervalles de confiance ($\gamma = 95\%$). Comme on pouvait s'y attendre, en l'absence de déformation de la couleur (qui est le principal facteur d'influence sur la notation dans les précédentes études), la tâche d'évaluation est plus difficile car l'ambiguïté des déformations est plus élevée. La difficulté est telle que pour les catégories "Objects" et "Buildings", les MOS ne permettent pas de discerner de classement des déformations. Cela semble indiquer que pour ces contenus, le niveau de compression de la géométrie et des normales avec V-PCC n'a pas d'impact sur la qualité perçue. Cela implique qu'utiliser le niveau de qualité le moins coûteux ($g1 + nE$) peut être satisfaisant. Cependant, ces résultats ne sont pas assez significativement fiables pour conclure. Sur la Figure 5 est représentée la variation de la discriminabilité des MOS en fonction du nombre de participants. Ici, la discriminabilité est le taux de paires de déformations avec des scores de qualité significativement différents [14] (test de Student avec $p - value > 0.05$). Nous pouvons voir que pour l'intégralité de nos données, même avec 20 participants, la discriminabilité ne converge pas encore vers un maximum, ce qui peut suggérer que plus de participants serait nécessaire pour obtenir des résultats significatifs.

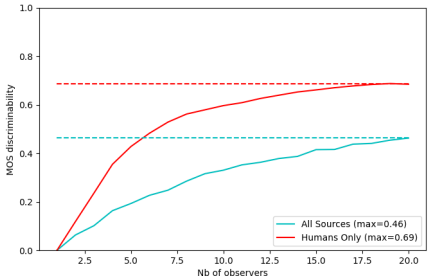


FIGURE 5 – *Évolution de la discriminabilité des MOS en fonction du nombre de participants, pour l'ensemble des stimuli et pour seulement la catégorie "Humans".*

Cependant, la même conclusion ne s'applique pas aux point clouds de la catégorie "Humans". Le taux de paires de déformations significativement différentes est plus élevé si on ne considère que ces 2 sources là, et il converge à partir de 18 participants, indiquant que les observations que l'on peut faire sur les MOS de ces contenus sont significatives. Pour ces modèles, on voit en Fig.4 que recalculer les normales après compression (nE) rend visible les distorsions de la géométrie pour les niveaux $g1\&g3$. Cet effet révélateur est mis en évidence par un rendu anormalement ombré des objets et résulte en une mauvaise qualité perçue par les observateurs. En comparaison, on voit que des normales encodées (quel que soit le niveau de qualité) ont un effet de masquage sur les éventuels défauts de géométrie et permettent donc une meilleure qualité visuelle du rendu. De plus, la différence moindre entre les niveaux $n1, n3, n5$ de qualité des normales montre que le coût de transmission des normales peut être minimisé. Aussi les combinaisons $g3 + n1$ et $g1 + n1$ ont des débits équivalents et des qualités de rendu supérieures à respectivement $g5 + nE$ et $g3+nE$, ce qui veut dire que l'encodage des normales peut même permettre d'économiser des bits sur la transmission de la géométrie.
Pour des cas d'usage où les point clouds sont utilisés pour représenter des humains dans des scènes avec éclairage virtuel (télécommunication et vidéo volumétriques), il y a donc un réel avantage à encoder les normales plutôt que de les recalculer. D'autant plus, il faut noter que la version publique du codec V-PCC utilisée dans cette expérience date de 2020. Celui-ci est aujourd'hui bien plus performant à bas débit.

3.2 Limites et Améliorations

Bien que significatives, nos observations sur la catégorie "Humans" ne sont basées que sur deux point clouds sources et ne couvrent pas la diversité de ce type de contenu. Une nouvelle expérience centrée sur cette catégorie sémantique avec plus de sources permettrait de robustifier ces résultats. L'effet de masquage de déformations géométriques qu'ont

les normales transmises sur le rendu n'est également pas la seule piste pour améliorer la compression et le rendu de point clouds. La compression des attributs de couleurs était hors du cadre de cet article, mais de futurs tests s'y intéressant pourraient révéler quel niveau de masquage existe entre l'ombrage et les niveaux de déformations de couleurs. Mis à part les normales, il y a aussi d'autres propriétés de matériaux, de texture ou de surface qui pourraient être ou représentés par d'autres attributs [5] (réflectance, transparence, aspect métallique, ...) et leur traitement en transmission mériterait d'être étudié.
Vu que nos résultats semblent concerner principalement des cas d'usage impliquant du contenu dynamique et des applications immersives, Il sera important d'étudier comment d'autres facteurs d'influence, propre au cas de vidéos volumétriques et de réalité étendue [15, 16] (degrés de liberté, interaction, attention visuelle, ...), affecte la perception de ces déformations de compression et rendus.
Enfin, l'impact de ces mêmes déformations sur des métriques de qualité objective pour le point clouds et sur leur performance en termes de corrélation avec des scores d'évaluation subjectives est également à considérer. Le développement des nouvelles métriques, ou l'adaptation de métriques existantes, adaptés à l'évaluation de qualité de rendu est donc nécessaire, comme nous suggérons dans un précédent article [6]. Concernant les autres types de contenus que les humains, des normales compressées, même si elle ne semble pas faire de différence en qualité à priori, pourrait malgré tout s'avérer utile pour la vision par ordinateur. Elles sont plus fidèles à la surface du point cloud avant compression et pourraient servir à corriger des déformations géométriques.

4 Conclusion

Nous avons montré que compresser les normales peut être avantageux en termes de qualité et/ou de coût de transmission pour des cas d'usages tels que la télécommunication et le divertissement immersifs où les point clouds représentent des humains. Nous encourageons donc l'utilisation de cette méthode. Des tests supplémentaires avec plus de stimuli de cette catégorie viendrait robustifier ces résultats. Plus d'études et expériences seraient également intéressantes afin d'évaluer la compression de nouveaux attributs et le rendu physique réaliste de point clouds dans des contextes d'usages, en réalité étendue.

Références

[1] Sebastian Schwarz et al. Emerging mpeg standards for point cloud compression. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 9(1) :133–148, 2018.

[2] Stuart Perry. JPEG Pleno Point Cloud Use Cases and Requirements, v1.6. *ISO/IEC JTC 1/SC 29/WG1*, 100096, 2022.

[3] Stuart Perry. Jpeg pleno point cloud common test conditions v3.6. *ISO/IEC JTC1/SC29/WG1 N*, 86044, Avril 2021.

[4] Kenneth Vanhoey et al. Visual Quality Assessment of 3D Models : On the Influence of Light-Material Interaction. *ACM Transactions on Applied Perception*, 15(1) :1–18, Octobre 2017.

[5] Marco Rossoni et al. Physically-based Rendering of Animated Point Clouds for eXtended Reality. *Journal of Computing and Information Science in Engineering*, pages 1–9, Septembre 2023.

[6] Amar Tious et al. Impact of point cloud normals compression on objective quality assessment. Dans *EUSIPCO*, Lyon, France, 2024. To be published.

[7] Alireza Javaheri et al. Point cloud rendering after coding : Impacts on subjective and objective quality. *IEEE Transactions on Multimedia*, 23 :4049–4064, 2021.

[8] Emin Zerman, Cagri Ozcinar, Pan Gao, et Aljosa Smolic. Textured Mesh vs Coloured Point Cloud : A Subjective Study for Volumetric Video Compression. Dans *QoMEX 2020*, pages 1–6, Mai 2020.

[9] Cheng-Hao Wu et al. Quantitative Comparison of Point Cloud Compression Algorithms With PCC Arena. *IEEE Transactions on Multimedia*, 25 :3073–3088, 2023.

[10] Ali Ak et al. BASICS : Broad Quality Assessment of Static Point Clouds in a Compression Scenario. *IEEE Transactions on Multimedia*, pages 1–13, 2024.

[11] ITU-R Recommendation BT.500-14. Methodology for the subjective assessment of the quality of television pictures, Octobre 2019.

[12] Amar Tious et al. Physically-based Lighting of 3D Point Clouds for Quality Assessment. Dans *ACM IMX '23*, pages 423–426, New York, NY, USA, 2023.

[13] James P. O'Shea et al. The assumed light direction for perceiving shape from shading. Dans *APGV '08*, pages 135–142, New York, NY, USA, 2008.

[14] Andreas Pastor et al. Comparison of conditions for omnidirectional video with spatial audio in terms of subjective quality and impacts on objective metrics resolving power. Dans *ICASSP 2024*, pages 8210–8214, 2024.

[15] Jesús Gutiérrez et al. Subjective Evaluation of Dynamic Point Clouds : Impact of Compression and Exploration Behavior. Dans *EUSIPCO*, pages 675–679, Septembre 2023. ISSN : 2076-1465.

[16] Silvia Rossi et al. Behavioural Analysis in a 6-DoF VR System : Influence of Content, Quality and User Disposition. Dans *ACM IXR '22*, pages 3–10, New York, NY, USA, Octobre 2022.

ALE : active learning extension for object detection

T. Oriol[1,3] J. Pasquet[1,2] J. Cortet[3]

[1] AMIS, Université de Montpellier 3, Montpellier, France

[2]TETIS - Inrae, AgroParisTech, Cirad, CNRS, Univ. Montpellier, Montpellier, France

[3]UMR 5175 Centre d'Ecologie Fonctionnelle et Evolutive, Université Paul Valéry Montpellier 3, Université de Montpellier, EPHE, CNRS, IRD, CEFE UMR, Montpellier, France

{theo.oriol,jerome.pasquet,jerome.cortet}@univ-montp3.fr

Abstract

Monitoring human activities impact on soil biodiversity over time is a costly and resource-intensive challenge. Modern technologies like deep learning offer a promising solution because they can analyze large datasets much faster than humans. However, deep learning relies on extensively annotated datasets, and annotating these samples is both time-consuming and expensive, complicating its application. This paper introduces a novel active learning approach called Active Learning Extension (ALE), which aims at improving model performance in object detection tasks while minimizing the need for extensive data annotation. Traditional active learning methods typically rely solely on prediction uncertainty to select images for annotation, which can be suboptimal when introducing new classes. ALE addresses this limitation by considering both the uncertainty and the number of predictions. This dual consideration leads to significant improvements, particularly in scenarios like Collembola detection, where creating and updating datasets is highly time-intensive. Our evaluation demonstrates that ALE significantly enhances model performance compared to state-of-the-art methods. The results underscore the importance of selecting challenging examples and accounting for the number of predictions to optimize active learning in object detection.

keywords

Deep learning, Active learning, Object detection.

1 Introduction

With the rise of environmental concerns, the need for tools to monitor the impact of human activities on soil over time has become urgent. Various metrics have been developed to assess soil quality [1, 2], one of which is biodiversity [3]. Collembola, commonly known as springtails, are a class of arthropods that, like other soil organisms, are sensitive to changes in soil properties such as pH, temperature, soil moisture, and nutrient availability. Consequently, they are currently used as a biodiversity indicator, particularly in agricultural and forest practices [4, 5]. Collembola play a crucial role in nutrient cycling and soil aggregation within their ecosystems, and soils can contain thousands of these individuals per square meter [6]. However, using Collembola as an indicator generates a substantial amount of data [7], which can take months to process due to the specialized expertise and the identification having to be done using a microscope [8]. This makes the process very time-consuming. Over the last few years, deep learning models have emerged as promising tools in ecology. The identification of Collembola using deep learning has already been demonstrated [9, 10]. However, to enhance the performance of these tools and enable them to identify a larger pool of species, there is a need to add more data, which conflicts with the time-consuming nature of manual Collembola identification. Given that expert identification is so time-consuming, optimizing the annotation process using active learning is a solution [11, 12, 13, 14, 15]. The challenge is that state-of-the-art active learning for object detection has been designed to improve models on already existing classes, not on new ones. In this paper, we introduce a new active learning technique to add new species to the Collembola datasets. The premise of this technique is that when adding new species to the datasets, it is more efficient to add more annotations with less uncertainty than fewer annotations with more uncertainty. Since the new species, have at first no annotations in the datasets, adding even a small amount of annotations can have a significant impact on the model results.

2 Related

2.1 Active learning

The use of deep learning requires extensive training data, but annotating new samples can often be time-consuming. Active learning aims to maximize model performance while minimizing the number of samples that need to be annotated. This is especially relevant for Collembola detection, where creating and updating datasets is very time-consuming. State-of-the-art active learning for object de-

tection typically follows this process : first, a model is trained on a base dataset. Then, the model makes predictions on a pool of unannotated images. Each prediction is evaluated based on its uncertainty, as it is more beneficial to provide the model with challenging examples that bring new information rather than easy predictions that do not significantly enhance model performance. After each prediction is evaluated, the images receive a score by aggregating the evaluated prediction scores. The top-scoring images are then annotated by an expert.

2.2 Metrics

Least confidence. The least confidence is one of the two main metrics used to evaluate the uncertainty in a model's predictions. It is based on the premise that the smaller the difference between the highest probability and the second highest probability, the greater the uncertainty. The formula is as follows :

$$LC = 1 - (p_1 - p_2) \tag{1}$$

Here, p_1 is the highest probability and p_2 is the second highest probability. The higher the least confidence value, the greater the uncertainty.

Entropy. Entropy is the second main metric used to evaluate the uncertainty. It considers that the flatter the distribution of probabilities is, the more unsure the model is, to do that the entropy is calculated using the following formula :

$$entropy = -\sum_{i=0}^{N} p_i log(p_i) \tag{2}$$

where N represents the length of the probability distribution, and p_i the probability p at the index i.

Aggregation of Detection Metric

The scoring for each image is determined by aggregating the scores S of their predictions. The state-of-the-art aggregation methods include the sum, mean, and maximum of the scores. Images without predictions receive a score of 0.

$$A_{\text{sum}} = \sum_{i=1}^{N} S_i \tag{3}$$

The sum aggregation method tends to prefer images with more annotations but does not consider the number of annotations.

$$A_{\text{max}} = \max_{i \in 1,2,\ldots,N} S_i \tag{4}$$

The maximum aggregation method focuses on identifying images with the most challenging predictions and ignores the number of predictions.

$$A_{\text{mean}} = \frac{1}{N} \sum_{i=1}^{N} S_i \tag{5}$$

The mean aggregation method, similar to the maximum method, does not consider the number of predictions and favors images with only difficult predictions.

2.3 Model

We used Yolov5x6, yolov5 biggest version. Yolov5 is an advanced object detection model developed by Ultralytics as an extension of Yolov3 [16]. It is a one-step detector, meaning it simultaneously detects and classifies objects. It comprises a backbone (CSPDarknet), a neck, and a prediction head (Figure 1). The backbone extracts features from the image, which are then mixed and combined by the neck for prediction. The detection head uses these features to propose bounding boxes and classes. To generate these proposals, the image is divided into multiple grids of various scales, with each cell proposing N objects. Yolov5 achieves precise detection by using anchors to predict box coordinates and different scale aspect ratios. Anchors facilitate coordinate prediction by using different ratios and sizes tailored to fit the data, in this case, Collembola. Instead of directly predicting Collembola coordinates, Yolov5 identifies which cell contains the center of the Collembola and predicts the height and width ratio of the anchor used for the prediction. This approach simplifies the task, greatly improving the accuracy of the model's coordinate predictions.

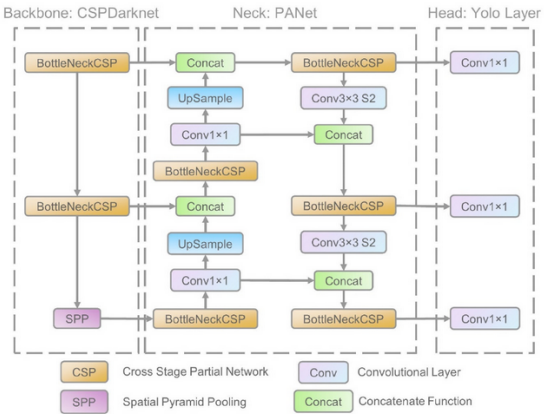


FIGURE 1 – *Yolov5 architecture [17].*

3 Proposed Method

3.1 Datasets

The data used in this paper was initially developed to benchmark object detection for Collembola on microscope images using deep-learning techniques [9]. It includes seven taxa of interest : *Ceratophysella denticulata* and *Ceratophysella gibbosa* (CER), *Hemisotoma ther-*

mophila (HEM-THE), *Hypogastrura manubrialis* (HYP-MAN), *Lepidocyrtus cyaneus* and *Lepidocyrtus lanuginosus* (LEP), *Metaphorura affinis* (MET-AFF), *Isotomiella minor* (ISO-MIN), and *Parisotoma notabilis* (PAR-NOT), along with a category "Other" for Collembola from unannotated species. All Collembola specimens were sourced from 12 different projects, more details on the projects are available in the original paper [9]. Eleven of these projects were used for training, while BISES, the 12th and largest project, was reserved for evaluation to ensure the model did not train on previously seen projects. It is important to note that no instances of *Hypogastrura manubrialis* HYP-MAN were available in the BISES project so they won't be part of the experiments. Within the "Other" category, two species were extracted and annotated to create the new classes, *Pseudosinella alba* (PSE-ALB) and *Sphaeridia pumilis* (SPH-PUM), with 37 and 54 annotations on BISES , respectively. The Collembola of these species were hidden with a white square (Figure 2) in the training set to prevent the model from training on these new taxa while preserving other annotations due to limited data. (Table 1) refers to the number of annotations per species (including the new one) in the base dataset.

FIGURE 2 – *Example of obscured collembola using a white square.*

3.2 Active learning

This paper presents a novel active learning approach called "ALE" (Active Learning Extension), which enhances the current object detection active learning paradigm that typically focuses solely on prediction uncertainty. ALE introduces the consideration of the number of predictions, leading to substantial improvements, especially when new classes are introduced through active learning. Increasing the number of annotations, even with just a few additional ones, on new classes with very few annotations, can significantly improve the model's performance.

ALE. ALE core idea is that when incorporating a new class during active learning, solely relying on prediction uncertainty for image selection is not optimal. It is more effective to consider both uncertainty and the number of predictions. In state-of-the-art object detection active learning, prediction confidence is typically measured using metrics like entropy or least confidence. Each image is then assigned a score using aggregation techniques such as sum, maximum, or mean, and images with the highest score are selected. However, when adding new classes, prioritizing annotations with lower uncertainty might be more efficient than fewer annotations with higher uncertainty. To incorporate the number of predictions into image selection, an extension to the aggregation of uncertainty metrics is employed, as shown in Equation 6.

$$ALE = A(S) \times \sqrt{N} \tag{6}$$

Where $A(S)$ represents the aggregation function of uncertainty scores, and N is the total number of predictions on the image.

This paper introduces the square root as an extension. It was chosen empirically because it rapidly prioritizes more annotations. We tested various combinations of evaluation metrics, and aggregation techniques to determine the most suitable approach for active learning.

4 Experimental protocol and evaluation

4.1 Experiment

The experiment aimed to test our models as follows : First, we trained a base model using the eleven projects of Collembola as training data and the twelfth (BISES) as validation data. Then, we built multiple datasets using ALE with different aggregation and metric combinations, as well as state-of-the-art active learning techniques for comparison. Additionally, we created 3 random selections and used the one with the best overall performance to compare our technique against random sampling. All of these methods selected images from the validation set, which were added to the training set. The models were fine-tuned on the updated training datasets and evaluated on the new validation sets. The experiment was conducted three times, each with a different selection of images to be added to the training set : 20, 50, and 100 images.

4.2 Evaluation

Since the models were evaluated on different validation datasets, comparing them using their own validation results was not feasible. To accurately compare all models, we would need to use only the common images across all validation datasets, ensuring no model is evaluated on images it was trained on. However, this approach is impractical due to the number of models and the limited number of images. To address this, we evaluated the models in pairs by creating new datasets that include only the common evaluation images for each pair. This method allowed us to directly compare each model against another, determining which

TABLE 1 – *Distribution of Annotations in Training and Validation on the Base Dataset.*

	OTHER	CER	CRY THE	HYP MAN	ISO MIN	LEP	MET AFF	PAR NOT	PSE ALB	SPH PUM
train	410	228	126	109	87	144	106	111	0	0
valid	163	92	96	0	62	121	23	151	37	54

one performs better on their shared evaluation dataset. The evaluation metric used for these comparisons was the mean average precision (mAP).

4.3 Training protocols

Every model was trained until convergence with these parameters : an initial learning rate of 0.01 with a weight decay of 0.005, and the Adam optimizer with a beta1 of 0.937. Data augmentation techniques were applied during training to artificially increase the dataset size, enhancing the model's ability to generalize and improve accuracy on the test dataset. Various transformations were employed, including random crop, mosaic, and color distortions such as brightness, contrast, saturation, hue, Gaussian blur, random scaling, random rotation, and random horizontal flipping. These transformations introduce variations that enhance the model's performance and adaptability to different input scenarios.

4.4 Results

The results of the experiment, presented in Tables 2, 3, and 4, illustrate the following comparisons : The left column lists the baseline model trained on the original dataset, the best-performing model from random sampling, and various state-of-the-art models combining multiple metrics and aggregation techniques. the top row represent every ALE versions. Each model is labeled with an evaluation metric followed by an aggregation technique. For example, a model may be labeled as "Entr Sum" or "ALE LC Max," indicating, respectively, the use of the entropy metric with the "Sum" aggregation and the ALE approach with LC (Least Confidence) as the metric, using a maximum-based aggregation and square root extension.

The columns display the average results and variances obtained after ensemble training for each version of the active learning model, ensuring a more robust evaluation. The percentages presented indicate the average improvement in mean accuracy (mAP) compared to other techniques and are accompanied by the standard deviation of the results. This methodology helps capture the stability and robustness of each approach.

TABLE 2 – *Comparison of methods with 20 images active learning.*

	ALE Entr Max	ALE Entr Sum	ALE LC Max	ALE LC Sum
Baseline Model	**12.80%**	**14.80%**	**16.10%**	**11.50%**
Rand	**4.42%±0.48**	**4.44%±2.05**	**6.78%±1.27**	**2.96%±0.38**
Entr Max	**0.43%±0.26**	**0.73%±2.06**	**5.00%±1.68**	-0.73%±0.53
Entr Sum	-1.77%±0.30	-1.97%±2.17	**0.57%±1.40**	-3.53%±0.50
LC Max	**0.17%±0.31**	**0.10%±2.30**	**2.87%±1.56**	-1.03%±0.81
LC Sum	-0.43%±0.30	-0.60%±2.12	**2.00%±1.42**	-2.27%±0.50

TABLE 3 – *Comparison of methods with 50 images active learning.*

	ALE Entr Max	ALE Entr Sum	ALE LC Max	ALE LC Sum
Baseline Model	**22.10%**	**24.20%**	**25.10%**	**25.20%**
Rand	**5.16%±5.97**	**11.23%±4.66**	**7.44%±6.06**	**10.35%±5.05**
Entr Max	-0.65%±2.34	**4.72%±1.42**	**2.08%±1.07**	**4.02%±1.07**
Entr Sum	-3.45%±2.21	**1.75%±1.58**	-1.20%±1.07	**1.10%±1.19**
LC Max	-3.57%±2.22	**2.10%±1.47**	-0.72%±1.03	**1.45%±1.11**
LC Sum	-2.85%±2.22	**2.73%±1.51**	-0.40%±1.08	**1.90%±1.21**

TABLE 4 – *Comparison of methods with 100 images active learning.*

	ALE Entr Max	ALE Entr Sum	ALE LC Max	ALE LC Sum
Baseline Model	**32.60%**	**35.90%**	**33.20%**	**31.30%**
Rand	**5.22%±1.14**	**8.80%±0.54**	**6.32%±0.66**	**6.24%±1.42**
Entr Max	**2.33%±0.66**	**4.83%±0.12**	**3.77%±1.07**	**2.90%±2.19**
Entr Sum	-2.60%±0.75	**0.43%±0.13**	-1.43%±1.18	-2.20%±1.70
LC Max	-1.87%±0.84	**1.90%±0.15**	**0.20%±1.30**	**0.93%±1.20**
LC Sum	-1.60%±0.93	**1.77%±0.17**	0.00%±1.42	**0.50%±0.70**

When 20 images are selected (Table 2), "ALE LC Max" shows the best performance with a score of 16.10%, making it more effective than other ALE methods compared to the Baseline model. "ALE Entr Sum" (14.80%) and "ALE Entr Max" (12.80%) also demonstrate strong results. "ALE LC Sum," with 11.50%, is the least effective among the ALE methods tested. "ALE LC Max" is the only model that outperforms all state-of-the-art models and random sampling. The results from the 20-image selection indicate that even with a small image set, model performance increases significantly.

For the selection of 50 images (Table 3), "ALE LC Sum" achieves a score of 25.20%. "ALE LC Max" (25.10%) and "ALE Entr Sum" (24.20%) closely follow, suggesting that increasing the number of images significantly enhances performance. Two models outperform both state-of-the-art methods and random sampling : "ALE LC Sum" and "ALE Entr Sum."

With the selection of 100 images (Table 4), "ALE Entr Sum" achieves the best performance with a score of 35.90%, surpassing "ALE LC Max" (33.20%) and "ALE LC Sum" (31.30%). Once again, it outperforms all other state-of-the-art models and random sampling.

The results indicate that, in most cases, the ALE versions outperform their state-of-the-art counterparts. However, many of them still fall short of other state-of-the-art models, particularly "Entr Sum." Notably, "ALE Entr Sum" stands out, outperforming every other model in nearly every comparison, except in the 20-image sample, where it is surpassed by "Entr Sum" and "LC Sum."

We would expect the improvement of "ALE Entr Sum"

over "Entr Sum" to follow a linear trend as sample size increases. However, while there is a noticeable improvement from 20 to 50 samples, at 100 samples, even though "ALE Entr Sum" continues to lead, the difference in results between the two models becomes less pronounced. This can be explained by the fact that, in the BISES project, the number of images containing three or more annotations is 62 (Table 5). Consequently, "ALE Entr Sum" maximizes its effectiveness with a selection of 50 images (Table 3), capturing diversity without introducing redundancy. However, when the selection increases to 100 images, this distinction fades, as a larger number of images become common across "ALE Entr Sum" and "Entr Sum" (Table 6), reducing the advantage of the ALE.

TABLE 5 – *Number of annotations per image in the BISES project.*

Number of annotations per image	1	2	>=3
Number of image	414	69	62

TABLE 6 – *Progression of result comparison between the "ALE Entr Sum" and "Entr Sum" models, based on sample count and non-common images in the new class dataset.*

Number of samples	Number of non-common images	Performance comparison
20	3	-1,97%
50	6	1,75%
100	4	0,43%

In Figures 3 and 4, we can observe that while "ALE Entr Sum" does not consistently outperform "Entr Sum" across every class, it demonstrates a notably higher performance on both of the new classes, *Pseudosinella alba* (PSE-ALB) and *Sphaeridia pumilis* (SPH-PUM). This suggests that "ALE Entr Sum" effectively adapts to novel data, showing a significant advantage over "Entr Sum" in handling unfamiliar categories by adding more annotations, even if its overall improvement is not uniform across all classes.

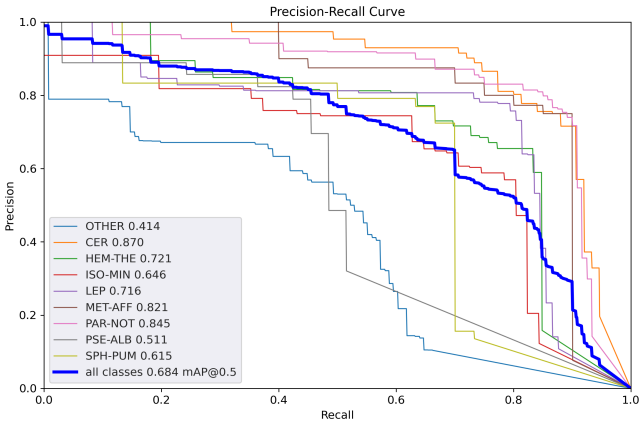


FIGURE 3 – *The AP and mAP of the top-performing "Entr Sum" version from the ensemble, evaluated on its shared dataset with the best "ALE Entr Sum" version from the ensemble (using the 50-image sample version).*

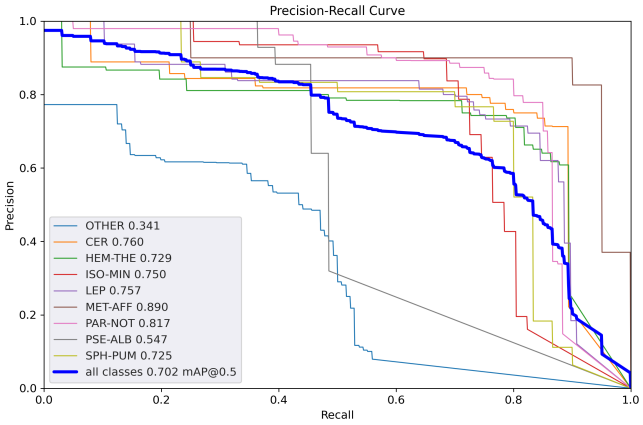


FIGURE 4 – *The AP and mAP of the top-performing "ALE Entr Sum" version from the ensemble, evaluated on its shared dataset with the best "Entr Sum" version from the ensemble (using the 50-image sample version).*

5 Conclusion

In conclusion, this study demonstrates the potential of ALE (Active Learning Extension) to enhance and expand model performance in ecological applications, particularly when adding new classes to an already trained model. ALE's approach, which combines uncertainty with prediction quantity, is especially effective on datasets containing images with varying numbers of predictions, enabling it to outperform traditional active learning methods. This capability allows ALE not only to improve model accuracy but also to extend its applicability by efficiently integrating new classes without necessitating a complete retraining on the entire dataset. In ecological monitoring, where precise species identification is essential, ALE enables the seamless addition of new taxa, allowing researchers to expand bio-

diversity assessments over time. This progressive approach supports adaptive model growth and offers an efficient, scalable solution to the costly, time-intensive process of annotating ecological data, making ALE a valuable tool for advancing biodiversity research and environmental monitoring.

6 Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

7 Acknowledgment

Théo Oriol is thankful for the financial support provided by the University of Paul Valery and the Occitanie Region. This work was granted access to the HPC resources of IDRIS under the allocation 20XX-AD011014028 made by GENCI

Références

[1] Dorothy Stone, Karl Ritz, BG Griffiths, Alberto Orgiazzi, et RE Creamer. Selection of biological indicators appropriate for european soil monitoring. *Applied Soil Ecology*, 97 :12–22, 2016.

[2] Anne Turbé, Arianna De Toni, Patricia Benito, Patrick Lavelle, Perrine Lavelle, Nuria Ruiz Camacho, Wim H van Der Putten, Eric Labouze, et Shaleindra Mudgal. Soil biodiversity : functions, threats and tools for policy makers. 2010.

[3] Jack H Faber, Rachel E Creamer, Christian Mulder, Jörg Römbke, Michiel Rutgers, J Paulo Sousa, Dorothy Stone, et Bryan S Griffiths. The practicalities and pitfalls of establishing a policy-relevant and cost-effective soil biological monitoring scheme. *Integrated environmental assessment and management*, 9(2) :276–284, 2013.

[4] Jean-François Ponge, Servane Gillet, Florence Dubs, E Fedoroff, Lucienne Haese, José Paulo Sousa, et Patrick Lavelle. Collembolan communities as bioindicators of land use intensification. *Soil biology and biochemistry*, 35(6) :813–826, 2003.

[5] José Paulo Sousa, Thomas Bolger, Maria Manuela Da Gama, Tuomas Lukkari, Jean-François Ponge, Carlos Simón, Georgy Traser, Adam J Vanbergen, Aoife Brennan, Florence Dubs, et al. Changes in collembola richness and diversity along a gradient of land-use intensity : a pan european study. *Pedobiologia*, 50(2) :147–156, 2006.

[6] T Larsen, Per Schjønning, et J Axelsen. The impact of soil compaction on euedaphic collembola. *Applied Soil Ecology*, 26(3) :273–281, 2004.

[7] Pascal Querner et Alexander Bruckner. Combining pitfall traps and soil samples to collect collembola for site scale biodiversity assessments. *Applied Soil Ecology*, 45(3) :293–297, 2010.

[8] Philippe JANSSEN, Marc FUHR, et Jean-Jacques BRUN. Effets de l'ancienneté du couvert forestier et de la maturité des peuplements sur la biodiversité des forêts de chartreuse, 2015.

[9] Théo Oriol, Jerome Pasquet, et Jérôme Cortet. Automatic identification of collembola with deep learning techniques. *Ecological Informatics*, 81 :102606, 2024.

[10] Stanislav Sys, Stephan Weißbach, Lea Jakob, Susanne Gerber, et Clément Schneider. Collembolai, a macrophotography and computer vision workflow to digitize and characterize samples of soil invertebrate communities preserved in fluid. *Methods in Ecology and Evolution*, 13(12) :2729–2742, 2022.

[11] Clemens-Alexander Brust, Christoph Käding, et Joachim Denzler. Active learning for deep object detection. *arXiv preprint arXiv :1809.09875*, 2018.

[12] Weiping Yu, Sijie Zhu, Taojiannan Yang, et Chen Chen. Consistency-based active learning for object detection. Dans *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3951–3960, 2022.

[13] JR van Bommel. Active learning during federated learning for object detection. *University of Twente Enschede : Enschede, The Netherlands*, 2021.

[14] Jiwoong Choi, Ismail Elezi, Hyuk-Jae Lee, Clement Farabet, et Jose M Alvarez. Active learning for deep object detection via probabilistic modeling. Dans *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10264–10273, 2021.

[15] Ying Li, Binbin Fan, Weiping Zhang, Weiping Ding, et Jianwei Yin. Deep active learning for object detection. *Information Sciences*, 579 :418–433, 2021.

[16] Joseph Redmon et Ali Farhadi. Yolov3 : An incremental improvement. *arXiv preprint arXiv :1804.02767*, 2018.

[17] Renjie Xu, Haifeng Lin, Kangjie Lu, Lin Cao, et Yunfei Liu. A forest fire detection system based on ensemble learning. *Forests*, 12(2) :217, 2021.

Vers une nouvelle génération de schémas de lifting basés sur les réseaux de neurones et application à la compression d'images

T. Dardouri[1], M. Kaaniche[2,3], A. Benazza-Benyahia[4], J.-C. Pesquet[3], G. Dauphin[2]
[1] Novelis, R&D laboratory, 75012, Paris, France
[2] Université Sorbonne Paris Nord, L2TI, UR 3043, F-93430, Villetaneuse, France.
[3] Centre de Vision Numérique, Univ. Paris-Saclay, CentraleSupélec, INRIA, 91190 Gif-sur-Yvette, France
[4] Université de Carthage, SUP'COM, LR11TIC01, COSIM Lab., 2083, El Ghazala, Tunisie
tdardouri@novelis.io, {mounir.kaaniche, gabriel.dauphin}@univ-paris13.fr
benazza.amel@supcom.rnu.tn, jean-christophe.pesquet@centralesupelec.fr

Résumé

Les schémas de lifting ont connu un grand succès en analyse et traitement d'images, et plus particulièrement, en compression d'images. Dans ce contexte, l'optimisation des opérateurs de lifting (i.e., les opérateurs de prédiction et de mise à jour) joue un rôle crucial dans la conception de nouveaux schémas de codage efficaces et adaptés aux images d'entrée. À cet égard, nous proposons, dans cet article, d'explorer le potentiel des réseaux neuronaux dans le contexte de structures de lifting 2D non séparables. Contrairement aux travaux précédents où différents modèles de réseaux neuronaux sont utilisés pour toutes les étapes de prédiction et de mise à jour, notre conception repose sur un nouveau modèle de réseau de neurones convolutif multi-tâches qui prend en compte les similitudes entre deux étapes de prédiction. Les simulations effectuées sur deux bases d'images usuelles montrent l'intérêt de l'architecture proposée pour la compression d'images.

Mots clefs

Transformées en ondelettes, schémas de lifting adaptatifs, optimisation, réseaux de neurones, compression d'images.

1 Introduction

Les ondelettes ont suscité beaucoup d'intérêt dans la comunauté de traitement du signal et d'images grâce à leurs bonnes propriétés de scalabilité en qualité et en résolution ainsi que leur capacité d'offrir une analyse multi-échelle des données. Par exemple, elles ont été largement adoptées dans diverses tâches de traitement [1, 2, 3] de différents types de contenus multimédias tels que les images 2D et 3D, la vidéo, l'audio, etc [4, 5].

Pour produire les coefficients d'ondelettes, le schéma de lifting (*Lifting Scheme* (LS)) s'est avéré être un outil efficace permettant une mise en oeuvre rapide et une reconstruction parfaite du signal d'entrée. Ainsi, un schéma de lifting conventionnel repose sur une étape de prédiction suivie d'une étape de mise à jour permettant de générer respectivement les coefficients de détails et d'approximation. Alors que la norme de codage d'images JPEG2000 utilise certains filtres prédéfinis avec des poids fixes, de nombreux efforts ont été déployés pour mieux adapter ces filtres au contenu des données d'entrée et améliorer l'efficacité des codeurs basés sur le schéma de lifting. Pour cela, différentes techniques d'optimisation ont été développées pour la conception des opérateurs (ou filtres) de prédiction et de mise à jour. La plupart de ces techniques ont été consacrées au filtre de prédiction, qui est souvent optimisé en minimisant un certain critère défini sur les coefficients de détail. Les critères utilisés comprennent les normes ℓ_2 [6] et ℓ_1[7] ainsi que l'entropie [8, 9]. Cependant, l'optimisation de l'opérateur de mise à jour est plus difficile, et a été peu explorée dans la littérature [6, 10, 11].

En plus de ces approches d'optimisation traditionnelles, certaines méthodes mettant en jeu les réseaux neuronaux ont été récemment proposées. En effet, les tâches de prédiction et de mise à jour ont été réalisées à l'aide de réseaux de neurones convolutifs (CNN) [12, 13] et de réseaux de neurones entièrement connectés (FCNN) [14, 15]. De tels schémas peuvent être considérés comme une première catégorie de méthodes de compression d'images basées sur l'apprentissage profond, une autre catégorie de méthodes, inspirées par les auto-encodeurs, a également été présentée dans la littérature. L'architecture commune à la plupart de ces méthodes comprend trois modules : (1) transformée d'analyse non linéaire, (2) quantification et codage entropique, et (3) transformée de synthèse non linéaire [16, 17, 18]. Notons que d'autres méthodes de codage prédictif (intra) ont également été proposées [19, 20].

Motivé par les nombreux avantages des représentations issues des schémas de lifting et les résultats prometteurs obtenus par les modèles FCNN [14, 15], l'objectif de cet article est d'exploiter davantage les réseaux neuronaux dans les systèmes de codage d'images basés sur les schémas de lifting.

Le reste de l'article est organisé comme suit. La section 2 rappelle le concept des schémas de lifting basés sur les réseaux de neurones. Ensuite, la section 3 décrit l'architecture proposée. Enfin, les résultats expérimentaux sont pré-

sentés dans la section 4 et des conclusions sont tirées dans la section 5.

2 Travaux connexes

Récemment, nous avons proposé de nouvelles structures de lifting reposant sur les réseaux de neurones [14, 15]. Pour cela, une structure de lifting 2D non séparable, qui présente l'avantage de réduire le nombre d'étapes de lifting par rapport à la décomposition séparable, a été adoptée. Cette structure de lifting est composée de trois étapes de prédiction suivi d'une étape de mise à jour [11]. Plus précisément, la structure 2D, illustrée dans la Fig. 1, consiste à décomposer une image d'entrée $X_j(m,n)$ en quatre composantes polyphases : $\mathbf{X}_{0,j}(m,n) = \mathbf{X}_j(2m,2n)$, $\mathbf{X}_{1,j}(m,n) = \mathbf{X}_j(2m,2n+1)$, $\mathbf{X}_{2,j}(m,n) = \mathbf{X}_j(2m+1,2n)$ et $\mathbf{X}_{3,j}(m,n) = \mathbf{X}_j(2m+1,2n+1)$. Ensuite, trois étapes de prédiction sont appliquées sur $\mathbf{X}_{3,j}(m,n)$, $\mathbf{X}_{2,j}(m,n)$ et $\mathbf{X}_{1,j}(m,n)$ afin de générer respectivement les coefficients de détails diagonaux $\mathbf{X}_{j+1}^{(HH)}(m,n)$, verticaux $\mathbf{X}_{j+1}^{(LH)}(m,n)$ et horizontaux $\mathbf{X}_{j+1}^{(HL)}(m,n)$. Enfin, une étape de mise à jour est appliquée à $\mathbf{X}_{0,j}(m,n)$ afin de produire les coefficients d'approximation $\mathbf{X}_{j+1}(m,n)$. Les étapes de prédiction et de mise à jour sont effectuées en utilisant quatre modèles FCNN désignés par $f_j^{(o)}$, avec $o \in \{HH, LH, HL, LL\}$.

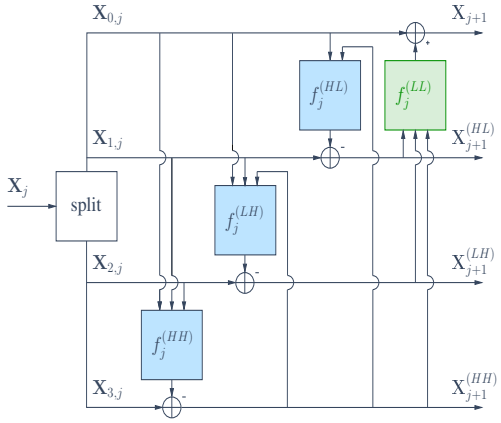


FIGURE 1 – *Structure d'analyse du FCNN-LS [14].*

Notons ici que les différents modèles FCNN (de prédiction et de mise à jour) peuvent être appris séparément à chaque niveau de résolution j [14]. Il est également possible d'optimiser conjointement les différents réseaux FCNN [15].

3 Nouvelle architecture basée sur un réseau CNN multi-tâches

3.1 Motivation

Une limitation majeure du modèle FCNN utilisé est qu'il ne tient pas compte des fortes corrélations locales dans l'image d'entrée. Pour surmonter ce problème et améliorer davantage les performances de prédiction, nous proposons d'abord de recourir à un réseau de neurones convolutionnels (CNN). De plus, dans les approches précédemment proposées, quatre modèles $f_j^{(o)}$ sont utilisés pour générer la sous-bande d'approximation ainsi que les trois sous-bandes de détails. Cependant, la Fig. 1 montre qu'une fois les coefficients de détails diagonaux sont générés, les deux étapes de prédiction suivantes peuvent être effectuées simultanément pour produire les coefficients de détails verticaux et horizontaux. Il est important de souligner que ces deux étapes de prédiction sont assez similaires et partagent certains entrées. Par conséquent, il devient plus intéressant de concevoir un nouveau modèle CNN multi-tâches (désigné dans la suite par MT-CNN) pour réaliser conjointement ces deux étapes de prédiction.

3.2 Modèles et méthodes d'apprentissage

La structure d'analyse de l'architecture de lifting, illustrée dans la Fig. 2, est composée des trois modèles CNN suivants. Le premier modèle, noté $C_j^{(HH)}$, correspond à la première étape de prédiction qui vise à générer les coefficients de détails diagonaux $\mathbf{X}_{j+1}^{(HH)}$. Le deuxième, désigné par $C_j^{(HL,LH)}$, effectue simultanément les deux tâches de prédiction permettant de générer les coefficients de détails verticaux $\mathbf{X}_{j+1}^{(LH)}$ et horizontaux $\mathbf{X}_{j+1}^{(HL)}$. Enfin, le dernier modèle, désigné par $C_j^{(LL)}$, concerne l'étape de mise à jour pour produire les coefficients d'approximation $\mathbf{X}_{j+1}$.

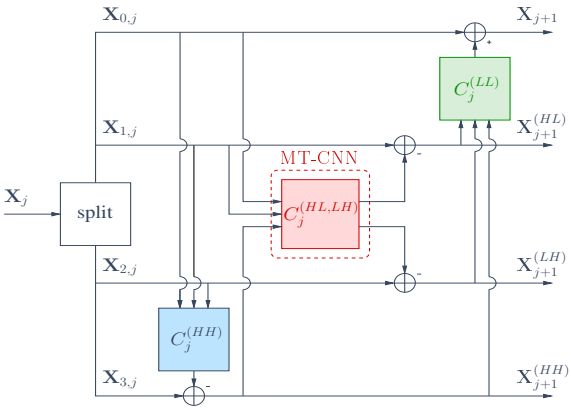


FIGURE 2 – *Structure d'analyse de l'architecture de lifting à base d'un réseau CNN multi-tâches.*

Les modèles CNN impliqués et leurs stratégies d'apprentissage sont décrits par la suite.

Étape de prédiction diagonale basée sur un CNN : De manière similaire à l'étape de prédiction basée sur le FCNN, la première étape de prédiction reposant sur le CNN vise à obtenir les coefficients de détails diagonaux en calculant la différence entre les composantes polyphases originales et celles prédites. La structure de l'architecture CNN retenue comprend cinq couches de convolution utilisant respectivement les nombres de canaux suivants : 32, 16, 16, 32 et 1. La taille des noyaux de la première couche est de 7×7, tandis que celles associées aux couches suivantes sont de 3×3. Nous utilisons également la fonction d'activation Gaussian Error Linear Unit (GELU). Le

modèle CNN retenu dépend d'un vecteur de paramètres $\mathbf{\Theta}_j^{(HH)}$ qui est appris en minimisant un critère d'erreur quadratique moyenne.

Étapes de prédiction horizontale et verticale basées sur un CNN multi-tâches : Une fois les coefficients de détails diagonaux générés, on peut procéder aux deuxième et troisième étapes de prédiction pour produire simultanément les coefficients de détails verticaux et horizontaux. En raison de la similitude entre ces deux étapes, un nouveau modèle CNN multi-tâches est proposé pour ces étapes de prédiction. Le modèle MT-CNN proposé, noté par $C_j^{(HL,LH)}$ dans la Fig. 2, est construit en utilisant le schéma de partage des paramètres [21] comme indiqué dans la Fig. 3.

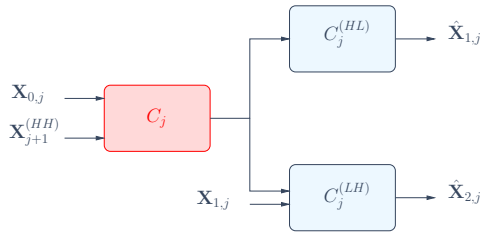


FIGURE 3 – *Schéma de partage des paramètres pour le modèle MT-CNN proposé* $C_j^{(HL,LH)}$.

Plus précisément, le modèle MT-CNN se compose d'un modèle CNN partagé suivi de deux modèles spécifiques à chaque tâche. En fait, dans une structure NSLS typique (comme illustré dans la Fig. 1), le calcul des coefficients de détail horizontaux et verticaux nécessite deux signaux de référence communs $\mathbf{X}_{0,j}$ et $\mathbf{X}_{j+1}^{(HH)}$. Ces deux canaux constitueront d'abord les entrées du modèle partagé noté C_j. Ensuite, la sortie de C_j sert de référence dans les deux modèles CNN spécifiques à chaque tâche illustrés dans les branches supérieure et inférieure du réseau, et désignés respectivement par $C_j^{(HL)}$ et $C_j^{(LH)}$.

Sur la Fig. 1, en plus des deux canaux d'entrée communs utilisés par le modèle CNN partagé, la génération des coefficients de détail verticaux $\mathbf{X}_{j+1}^{(LH)}$ utilise un troisième signal de référence correspondant à $\mathbf{X}_{1,j}$. Pour cette raison, un canal supplémentaire $\mathbf{X}_{1,j}$ a été inclus en tant qu'entrée du modèle $C_j^{(LH)}$. Enfin, les couches de sortie des deux modèles spécifiques à chaque tâche $C_j^{(HL)}$ et $C_j^{(LH)}$ permettent de générer les sous-bandes de détails horizontaux $\mathbf{X}_{j+1}^{(HL)}$ et verticaux $\mathbf{X}_{j+1}^{(LH)}$. Pour apprendre le modèle conjoint $C_j^{(HL,LH)}$, une approche d'apprentissage multi-tâches est adoptée. En effet, le modèle est appris en optimisant la somme des fonctions de coût spécifiques à chaque tâche, reposant sur un critère d'erreur quadratique moyenne.

Étape de mise à jour basée sur un CNN : Après les étapes de prédiction, une étape de mise à jour basée sur un réseau CNN est finalement effectuée pour calculer les coefficients d'approximation $\mathbf{X}_{j+1}$. En effet, les sous-bandes de détails générées constitueront les trois canaux d'entrée du modèle CNN de mise à jour $C_j^{(LL)}$. Son canal de sortie sera utilisé pour produire les coefficients d'approximation $\mathbf{X}_{j+1}$. Il convient de noter ici que la structure utilisée pour $C_j^{(LL)}$ est similaire à celle de $C_j^{(HH)}$. Par conséquent, elle est entraînée de la même manière.

4 Résultats expérimentaux

L'architecture proposée a été entraînée en utilisant la base de données Flickr composée de 8,000 images de différentes tailles [1]. Deux bases de données de test ont été considérées : Kodak [2] (composée de 24 images de taille 768×512) et Tecnick [3] [22] (où 30 images, de taille 1200×1200, ont été sélectionnées). Notre approche a été comparée à JPEG2000 ainsi que d'autres méthodes d'état de l'art basées sur les réseaux de neurones. Plus précisément, les tableaux 1 et 2 montrent les gains de notre méthode par rapport à CNN-LS[12] et FCNN-LS [14], en terme de métrique de Bjøntegaard. Notons ici que la qualité de reconstruction est évaluée en utilisant la métrique perceptuelle PieAPP [23], qui s'est avérée plus pertinente que les métriques traditionnelles telles que le PSNR et le SSIM. Les résultats obtenus à bas et moyens débits (aux points $\{0.07, 0.1, 0.15, 0.2\}$ et $\{0.2, 0.25, 0.3, 0.4\}$ bpp, respectivement) montrent des gains significatifs de l'architecture proposée, en terme de réduction de débit, par rapport à FCNN-LS [14] and CNN-LS [12].

	gain de débit (in %)		**gain de PieAPP**	
Bases	**bas**	**moyens**	**bas**	**moyens**
Kodak	-13.80	-8.74	-0.11	-0.06
Tecnick	-19.87	-14.41	-0.14	-0.07

TABLEAU 1 – *Gain du MT-CNN-LS par rapport à FCNN-LS [14] en terme de métrique de Bjøntegaard.*

	gain de débit (in %)		**gain de PieAPP**	
Bases	**bas**	**moyens**	**bas**	**moyens**
Kodak	-57.01	-26.57	-0.55	-0.23
Tecnick	-51.82	-19.40	-0.44	-0.12

TABLEAU 2 – *Gain du MT-CNN-LS par rapport à CNN-LS [12] en terme de métrique de Bjøntegaard.*

5 Conclusion

Dans cet article, une nouvelle architecture de schéma de lifting basée sur un réseau CNN multi-tâches a été proposée. Le potentiel de cette architecture a été démontré dans le contexte de la compression d'images, et mériterait d'être exploré pour d'autres tâches d'analyse d'images.

1. https ://www.kaggle.com/datasets/adityajn105/flickr8k
2. https ://www.r0k.us/graphics/kodak/
3. https ://testimages.org/

Références

[1] J.-H. Jacobsen, A. W. M. Smeulders, et E. Oyallon. i-RevNet : Deep invertible networks. Dans *International Conference on Learning Representations*, pages 1–11, Vancouver, Canada, May 2018.

[2] J. J. Huang et P. L. Dragotti. LINN : Lifting inspired invertible neural network for image denoising. Dans *European Signal and Image Processing Conference*, pages 1–5, Dublin, Ireland, September 2021.

[3] T.-S. Nguyen, M. Luong, M. Kaaniche, L. H. Ngo, et A. Beghdadi. A novel multi-branch wavelet neural network for sparse representation based object classification. *Pattern Recognition*, 135 :109155, 2023.

[4] Y. Xing, M. Kaaniche, B. Pesquet-Popescu, et F. Dufaux. Adaptive non separable vector lifting scheme for digital holographic data compression. *Applied Optics*, 54(1) :A98–A109, January 2015.

[5] E. Martinez-Enriquez, J. Cid-Sueiro, F. Diaz de Maria, et A. Ortega. Directional transforms for video coding based on lifting on graphs. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(4) :933–946, November 2016.

[6] A. Gouze, M. Antonini, M. Barlaud, et B. Macq. Design of signal-adapted multidimensional lifting schemes for lossy coding. *IEEE Transactions on Image Processing*, 13(12) :1589–1603, December 2004.

[7] M. Kaaniche, B. Pesquet-Popescu, A. Benazza-Benyahia, et J.-C. Pesquet. Adaptive lifting scheme with sparse criteria for image coding. *EURASIP Journal on Advances in Signal Processing : Special Issue on New Image and Video Representations Based on Sparsity*, 2012(1) :1–22, January 2012.

[8] J. Solé et P. Salembier. Generalized lifting prediction optimization applied to lossless image compression. *IEEE Signal Processing Letters*, 14(10) :695–698, October 2007.

[9] A. Benazza-Benyahia, J.-C. Pesquet, J. Hattay, et H. Masmoudi. Block-based adaptive vector lifting schemes for multichannel image coding. *EURASIP International Journal of Image and Video Processing*, 2007(1) :10 pages, January 2007.

[10] B. Pesquet-Popescu. *Two-stage adaptive filter bank*. First filling date 1999/07/27, official filling number 99401919.8, European patent number EP1119911, 1999.

[11] M. Kaaniche, A. Benazza-Benyahia, B. Pesquet-Popescu, et J.-C. Pesquet. Non separable lifting scheme with adaptive update step for still and stereo image coding. *Elsevier Signal Processing : Special issue on Advances in Multirate Filter Bank Structures and Multiscale Representations*, 91(12) :2767–2782, January 2011.

[12] H. Ma, D. Liu, R. Xiong, et F. Wu. iWave : CNN-based wavelet-like transform for image compression. *IEEE Transactions on Multimedia*, 22(7) :1667–1697, July 2020.

[13] H. Ma, D. Liu, N. Yan, H. Li, et F. Wu. End-to-end optimized versatile image compression with wavelet-like transform. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, September 2020.

[14] T. Dardouri, M. Kaaniche, A. Benazza-Benyahia, et J.-C. Pesquet. Dynamic neural network for lossy-to-lossless image coding. *IEEE Transactions on Image Processing*, 31 :569–584, December 2021.

[15] T. Dardouri, M. Kaaniche, A. Benazza-Benyahia, G. Dauphin, et J.-C. Pesquet. Joint learning of fully connected network models in lifting based image coders. *IEEE Transactions on Image Processing*, 33 :134–148, March 2023.

[16] J. Ballé, V. Laparra, et E. P. Simoncelli. End-to-end optimized image compression. Dans *International Conference on Learning Representations*, pages 1–27, Toulon, France, April 2017.

[17] E. Agustsson, M. Tschannen, F. Mentzer, R. Timofte, et V. G. Luc. Generative adversarial networks for extreme learned image compression. Dans *International Conference on Learning Representations*, pages 1–31, New Orleans, LA, USA, May 2019.

[18] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, et N. Johnston. Variational image compression with a scale hyperprior. Dans *International Conference on Learning Representations*, pages 1–47, Vancouver, Canada, May 2018.

[19] J. Li, B. Li, J. Xu, R. Xiong, et W. Gao. Fully connected network-based intra prediction for image coding. *IEEE Transactions on Image Processing*, 27(7) :3236–3247, July 2018.

[20] T. Dumas, A. Roumy, et C. Guillemot. Context-adaptive neural network-based prediction for image compression. *IEEE Transactions on Image Processing*, 29(1) :679–693, August 2019.

[21] S. Vandenhende, S. Georgoulis, W. V. Gansbeke, M. Proesmans, D. Dai, et L. V. Gool. Multi-task learning for dense prediction tasks : A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7) :3614–3633, 2022.

[22] N. Asuni et A. Giachetti. Test images : A large data archive for display and algorithm testing. *Journal of Graphics Tools*, 17(4) :113–125, February 2015.

[23] E. Prashnani, H. Cai, Y. Mostofi, et P. Sen. PieAPP : Perceptual image-error assessment through pairwise preference. Dans *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–10, Salt Lake City, UT, USA, June 2018.

Compression efficace d'images basée sur un modèle de représentation d'état

Bouzid Arezki, Anissa Mokraoui, Fangchen Feng
L2TI, Université Sorbonne Paris Nord
99, avenue Jean-Baptiste Clément, 93430 Villetaneuse, France
{bouzid.arezki, anissa.mokraoui, fangchen.feng}@univ-paris13.fr

Résumé

Les réseaux neuronaux profonds ont révolutionné le domaine de la compression d'images en surpassant les approches traditionnelles. Cependant, leur complexité élevée limite souvent leur utilisation. Pour pallier ce problème, diverses stratégies telles que la distillation des connaissances et les architectures légères ont été explorées pour réduire la complexité tout en maintenant de bonnes performances. Cet article présente une nouvelle architecture de compression d'images basée sur un modèle de représentation d'état. Cette architecture trouve un équilibre entre performance et complexité de calcul, la rendant adaptée aux applications pratiques. Les évaluations expérimentales confirment que notre architecture obtient de bonnes performances en termes de BD-rate, tout en réduisant considérablement la complexité de calcul et la latence par rapport aux autres méthodes compétitives de l'état de l'art.

Mots clefs

Compression d'images, Modèle d'espace d'état, Complexité de calcul, Débit-distorsion

1 Introduction

Les méthodes de compression basées sur les réseaux neuronaux profonds [1, 2] offrent des performances en constante amélioration par rapport aux approches traditionnelles. Elles incluent généralement un autoencodeur variationnel hiérarchique à deux niveaux avec un *hyper-prior* comme modèle d'entropie, composé de deux ensembles d'encodeurs et de décodeurs. Ces méthodes sont souvent trop complexes pour une utilisation en temps réel. Des modèles réduits sont proposés pour diminuer cette complexité, souvent en sacrifiant la performance débit-distorsion [2]. Pour pallier cela, des techniques comme la distillation des connaissances, les architectures légères et les mécanismes d'attention légers [3] ont été développées. Récemment, les modèles à espace d'état (SSM) et leur variante, le modèle Mamba, ont suscité un grand intérêt en vision par ordinateur. Cependant, leur adoption a été limitée par des exigences élevées en calcul et en mémoire. Mamba [4] surmonte ces limitations en intégrant un mécanisme de sélection, améliorant ainsi le raisonnement contextuel.

Dans cet article, nous introduisons une architecture de compression d'images basée sur un modèle à espace d'état, qui met l'accent sur la performance en termes de débit-distorsion, la complexité de calcul et la latence. Des expériences approfondies montrent que notre méthode parvient à atteindre de bonnes performances en termes de compression tout en réduisant de manière significative la complexité de calcul et la latence par rapport aux méthodes compétitives de l'état de l'art. Cette affirmation est étayée par la Fig. 1 (BD-rate versus complexité de calcul), qui situe clairement notre méthode de compression parmi les approches les plus compétitives de l'état de l'art.

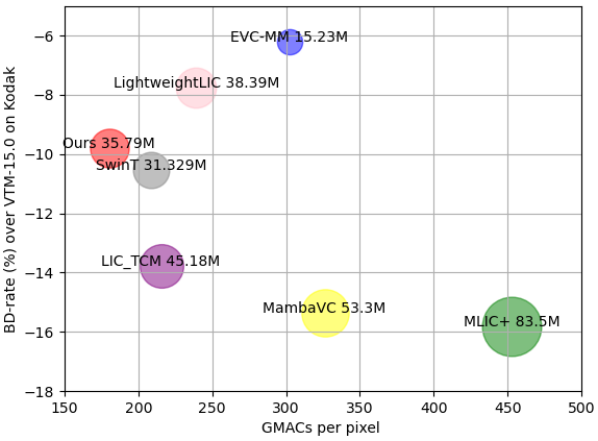


FIGURE 1 – *BD-rate (VTM-15.0 [5] comme référence) vs complexité de calcul (GMAC) sur Kodak [6]. Le rayon des cercles indique le nombre de paramètres du modèle.*

2 Préliminaires

Introduisons ci-dessous quelques notions utiles sur les modèles de représentation d'état *space state model* (SSM) sur lesquels notre méthode de compression s'appuie.

La transformation du SSM dans S4 [4] est dérivée du modèle classique d'espace d'état, qui associe un signal d'entrée unidimensionnel $x(t) \in \mathbb{R}$ à un signal de sortie unidimensionnel $y(t) \in \mathbb{R}$ via un état latent $h(t) \in \mathbb{R}^N$ de dimension N, nous pouvons formuler le processus à l'aide d'équations différentielles ordinaires linéaires (EDO) :

$$\begin{aligned} h'(t) &= \mathbf{A}h(t) + \mathbf{B}x(t), \\ y(t) &= \mathbf{C}h(t), \end{aligned} \tag{1}$$

où $\mathbf{A} \in \mathbb{R}^{N\times N}$, $\mathbf{B} \in \mathbb{R}^{N\times 1}$, $\mathbf{C} \in \mathbb{R}^{1\times N}$ sont des paramètres du réseau de neurones.

Pour traiter la séquence d'entrée discrète $x = [x_0, x_1, \ldots, x_{L-1}] \in \mathbb{R}^L$, les paramètres de l'équation (1) sont discrétisés en utilisant une taille de pas Δ, représentant la résolution de l'entrée continue $x(t)$ [4]. En particulier, les paramètres continus $\mathbf{A}$ et $\mathbf{B}$ sont convertis en paramètres discrets $\overline{\mathbf{A}}$ et $\overline{\mathbf{B}}$ par la technique *zero-order hold (ZOH)*, définie comme suit :

$$\begin{aligned} \overline{\mathbf{A}} &= \exp(\Delta\mathbf{A}), \\ \overline{\mathbf{B}} &= (\Delta\mathbf{A})^{-1}(\exp(\Delta\mathbf{A}) - \mathbf{I}).\Delta\mathbf{B}. \end{aligned} \tag{2}$$

Aprés la discrétisation de $\mathbf{A}$ et $\mathbf{B}$, l'equation (1) est reformulée comme suit :

$$\begin{aligned} h_t &= \overline{\mathbf{A}}h_{t-1} + \overline{\mathbf{B}}x_t, \\ y_t &= \mathbf{C}h_t. \end{aligned} \tag{3}$$

Les SSM peuvent alors être calculés efficacement à l'aide de RNNs. Le processus récursif peut être reformulé et calculé comme une convolution :

$$\begin{aligned} \overline{\mathbf{K}} &= (\mathbf{C}\overline{\mathbf{B}}, \mathbf{C}\overline{\mathbf{A}\mathbf{B}}, \ldots, \mathbf{C}\overline{\mathbf{A}}^{L-1}\overline{\mathbf{B}}), \\ y &= x * \overline{\mathbf{K}}, \end{aligned} \tag{4}$$

où L représente la longueur de la séquence d'entrée x ; et $\overline{\mathbf{K}} \in \mathbb{R}^L$ le noyau de convolution SSM.

Mamba [7] intègre la dépendance des données pour capturer l'information contextuelle dans l'équation (1) en introduisant une nouvelle méthode de paramétrage pour les SSM incluant un mécanisme de sélection conditionnelle basé sur l'entrée, appelé S6. Bien que la nature récurrente des SSM limite la possibilité d'une parallélisation complète, Mamba propose une optimisation grâce à des techniques de paramétrage structuré et à un algorithme de balayage parallèle efficace sur le plan matériel. En conséquence, de nombreux travaux ont adapté Mamba du traitement du langage naturel (NLP) au domaine de la vision [8].

3 Méthode de compression proposée

Notre méthode repose sur une architecture à deux niveaux. Dans un premier temps, l'image d'entrée x est encodée par l'encodeur génératif pour obtenir $y = g_a(x)$. Ensuite, le hyper-latent $z = h_a(y)$ est extrait via l'encodeur du réseau hyper-prior. Le hyper-latent quantifié $\hat{z}$ est alors modélisé et codé à l'aide d'un codage entropique basé sur un modèle factorisé *Factorized model* avant d'être transmis à travers $h_s(\hat{z})$. Les sorties de h_s et du modèle contextuel *context model* sont ensuite utilisées par le réseau de paramètres d'entropie *Entropy Parametres* [1], qui génère les paramètres μ et σ d'un modèle d'entropie gaussien conditionnel $P(y|\hat{z}) = \mathcal{N}(\mu, \sigma^2)$ pour modéliser y. Le vecteur latent quantifié $\hat{y} = Q(y)$ est codé au final via un codage entropique (codage/décodage arithmétique AE/AD) et transmis à $\hat{x} = g_s(\hat{y})$ pour reconstruire l'image $\hat{x}$. L'architecture détaillée de notre méthode est schématisée dans la Fig. 2.

Nous proposons d'utiliser le *context model* décrit dans [1]. Contrairement à l'architecture présentée dans [12], notre choix vise à développer une architecture optimisée pour l'efficacité computationnelle.

Les encodeurs génératifs et hyper-prior, g_a et h_a, sont construits avec le bloc de fusion *patch merge* et le bloc *Visual State Space (VSS)* illustré dans la Fig 3. Le bloc de *patch split* utilise l'opération *Depth-to-Space* [2] pour le sous-échantillonnage, une couche de normalisation et une couche linéaire pour projeter l'entrée à une certaine profondeur C_i. Dans g_a, la profondeur C_i de vecteur latent augmente à mesure que le réseau devient plus profond, ce qui permet une représentation de l'image de plus en plus abstraite. À chaque étape, nous réduisons la résolution par un facteur 2. Comparé à la couche convolutionnelle utilisée dans MambaVC [12], le bloc de fusion *patch merge* que nous avons choisi est plus simple à implémenter et offre également une complexité de calcul moindre.

Le bloc VSS, proposé initialement dans [8], consiste en une seule branche de réseau avec deux modules résiduels, suivant l'architecture du bloc Transformer classique. Plus précisément, chaque niveau de profondeur de notre méthode se compose d'une séquence de blocs VSS ou le nombre de blocs à chaque niveau i est noté d_i (voir Fig. 2). A partir d'une carte de caractéristiques d'entrée $\mathbf{f} \in \mathbb{R}^{H\times W\times C}$, nous obtenons $\mathbf{f}''$ à partir d'un premier module résiduel :

$$\mathbf{f}'' = \mathbf{f} + \mathbf{f}', \tag{5}$$

où $\mathbf{f}'$ est obtenu comme suit :

$$\begin{aligned} \mathbf{f}' = \mathrm{MLP}_2(\mathrm{LN}_2(\mathrm{SS2D}(\\ \sigma(\mathrm{DWConv}(\mathrm{MLP}_1(\mathrm{LN}_1((f)))))))). \end{aligned} \tag{6}$$

Comme illustré par la Fig. 3 (voir (a) and (b)), la sortie du VSS bloc est donnée par :

$$\mathbf{f}_{out} = \mathrm{MLP}_3(\mathrm{LN}_3(\mathbf{f}'')) + \mathbf{f}'', \tag{7}$$

où LN représente la couche de normalisation ; SS2D le module de balayage sélectif 2D ; σ la fonction d'activation SiLU ; DWConv la convolution depthhwise ; et MLP la projection linéaire. Contrairement au bloc VSS dans [8], nous utilisons RMSNorm [16] à la place de LayerNorm comme couche de normalisation. Nous avons observé empiriquement que RMSNorm améliore significativement la vitesse de convergence lors de l'entraînement. Nous adoptons l'approche de balayage sélectif proposée dans [8], qui adapte un mécanisme de sélection dépendant de l'entrée [7] aux données 2D sans compromettre ses avantages. SS2D comprend trois étapes illustrées dans la Fig. 3 (c) : Balayage croisé *Cross-scan* : déplie les caractéristiques d'entrée en séquences le long de quatre chemins de traversée distincts ; Balayage sélectif *Selective scan* : traite chaque chemin avec un S6 distinct en parallèle ; Fusion croisée *Cross-merge* : remodèle et fusionne ensuite les séquences résultantes pour former les caractéristiques de sortie, intégrant efficacement l'information provenant d'autres pixels dans différentes directions par rapport à chaque pixel courant. Cela facilite l'établissement de champs réceptifs globaux dans l'espace 2D.

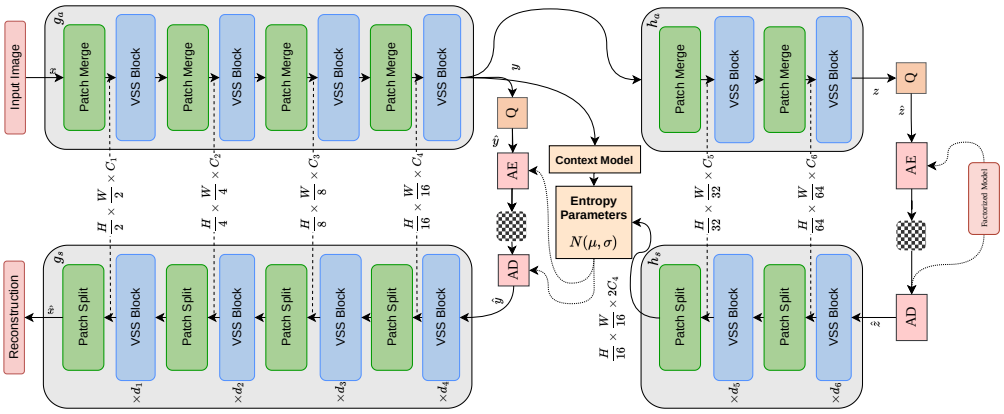


FIGURE 2 – *Architecture de notre méthode proposée de compression d'images.*

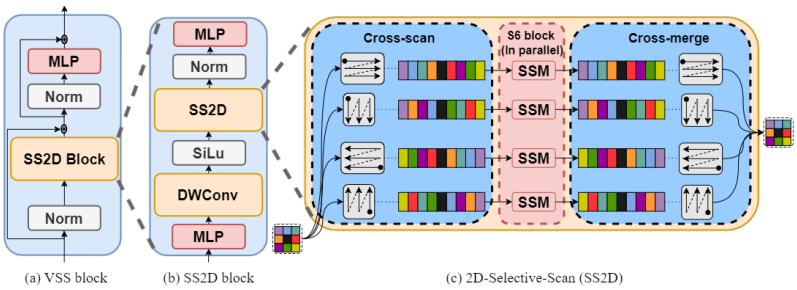


FIGURE 3 – *Le bloc VSS [8] se compose d'un bloc SS2D qui effectue des balayages sélectifs selon 4 parcours parallèles.*

4 Analyse des performances

Notre modèle a été entraîné sur le jeu de données CLIC20 [9] en utilisant une fonction de perte $L = D + \lambda R$, où R représente le débit binaire et D la distorsion. Nous utilisons l'erreur quadratique moyenne (MSE) comme mesure de distorsion dans l'espace de couleur RGB. Le multiplicateur de Lagrange λ ajuste le compromis débit-distorsion (RD), avec des valeurs choisies $\lambda \in \{100, 50, 30, 10\}$. Chaque lot d'entraînement comporte 8 découpes aléatoires des images de taille 256×256. Nous avons effectué 1 million (1M) d'itérations avec l'optimiseur ADAM et un taux d'apprentissage fixé à 10^{-4}. notre modèle a été implémentée avec Pytorch en utilisant la bibliothèque CompressAI [17].

Nous proposons d'évaluer les performances de notre modèle sur trois jeux de données : Kodak [6], JPEG-AI [10], et CLIC20 [9] incluant les catégories mobile et professionnelle. Toutes les images ont été complétées par des zéros dans cas où leur taille n'est pas un multiple de 256. Nous comparons notre modèle aux modèles compétitifs de l'état de l'art [2,3,12–15] choisis pour leurs performances (débit-distorsion, complexité de calcul). Les expériences ont été réalisées sur un GPU A100 80 Go et un CPU Intel Xeon Gold 6330 3,10 GHz.

Le tableau 1 présente le BD-rate des modèles compétitifs sur les 3 datasets, avec un débit couvrant la plage de 0.4 à 1.2 bits par pixel (bpp), et le VTM-15.0 [5] comme référence. Notre modèle atteint un BD-rate de -21.75% et une augmentation relative de 4.17% par rapport au modèle LIC_TCM [15]. Ce dernier reconnu pour ses performances supérieures parmi les méthodes comparées. Cependant, ce modèle présente une complexité de calcul et un nombre de paramètres significativement plus élevés, comme illustré par Fig. 1. Celle-ci compare les performances en termes de BD-rate et de GMACs [1] des différentes méthodes sur le jeu de données Kodak. L'analyse du graphique confirme bien que notre modèle offre un excellent compromis entre le BD-rate, la complexité de calcul et le nombre de paramètres.

Le tableaux 2 compare la complexité de calcul évaluée en termes de MACs[1] pour trois résolutions différentes, soulignant la réduction significative apportée par notre approche en termes de complexité de calcul.

Le tableau 3 donne la latence moyenne sur 2000 images à une résolution de 256×256 sur le GPU utilisé. Notre modèle affiche des temps de décodage compétitifs par rapport à lightweightLIC [3], tout en maintenant des temps de codage similaires.

1. Les MAC et les FLOP sont calculés à l'aide de la bibliothèque calflops. `https://github.com/MrYxJ/calculate-flops.pytorch`

Methodes	Kodak [6]	CLIC2020 [9]	JPEG-AI [10]	La moyenne
BPG444 [11]	29.86%	32.77%	43.77%	35.46%
SwinT* [2]	-10.52%	-6.47%	-2.78%	-6.03%
MambaVC* [12]	-15.37%	-16.69%	-12.47%	-14.69%
SwinNPE [13]	-5.85%	-17.50%	-23.56%	-15.63%
LightweightLIC [3]	-7.76%	-23.60%	-29.86%	-20.40%
MLIC+* [14]	**-15.86%**	-	-15.89%	-
LIC_TCM [15]	-13.76%	**-30.65%**	**-33.37%**	**-25.92%**
Méthode proposée	-9.81%	-29.91%	-25.55%	-21.75%

TABLEAU 1 – *BD-rate performance (VTM-15.0 [5] comme référence). $*$ Chiffres extraits des articles (car les modèles pré-entrainés ne sont pas disponibles). L'évaluation de MLIC+ sur le jeu de données CLIC2020 n'est pas fournie dans [14].*

Résolution	Méthode proposée	SwinT [2]	LIC_TCM [15]	LightweightLIC [3]	MambaVC [12]	MLIC+ [14]
768×512	**180.053G**	208.789G	215.316G	239.213G	326.112G	452.622G
1024×768	**360.106G**	417.570G	430.632G	478.425G	652.224G	905.243G
1280×1280	**750.222G**	869.956G	897.150G	996.719G	OM	1.8859T
#paramètres (M)	35.79	32.34	45.18	38.39	53.30	83.50

TABLEAU 2 – *Multiply-Add Cumulation (MACs), mesuré sur un GPU A100 80 Go et un CPU Intel Xeon Gold 6330 3,10 GHz, a différentes résolutions d'images. OM signifie Out of Memory.*

Méthodes	Latence GPU (ms)	
	Encodage	Décodage
MambaVC [12]	**14.01**	73.36
LIC_TCM [15]	15.23	52.46
SwinT [2]	14.25	20.86
LightweightLIC [3]	15.62	**15.56**
Méthode proposée	20.24	19.93

TABLEAU 3 – *Latence moyenne sur 2000 images (résolution 256×256) sur un GPU A100 80 Go et un processeur Intel Xeon Gold 6330 3,10 GHz.*

5 Conclusion

Nous avons proposé une approche de compression efficace d'images basée sur le modèle de représentation d'état. Celle-ci présente des performances compétitives en termes de débit-distorsion tout en réduisant de manière significative la complexité de calcul et la latence.

Références

[1] David Minnen, Johannes Ballé, et George D Toderici. Joint autoregressive and hierarchical priors for learned image compression. *NeurIPS*, 31, 2018.

[2] Yinhao Zhu, Yang Yang, et Taco Cohen. Transformer-based transform coding. Dans *ICLR*, 2021.

[3] Ziyang He, Minfeng Huang, Lei Luo, Xu Yang, et Ce Zhu. Towards real-time practical image compression with lightweight attention. *Expert Systems with Applications*, 252 :124142, 2024.

[4] Albert Gu, Karan Goel, et Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv*, 2021.

[5] Benjamin Bross, Ye-Kui Wang, Yan Ye, Shan Liu, Jianle Chen, Gary Sullivan, et Jens-Rainer Ohm. Overview of the versatile video coding (vvc) standard and its applications. *TCSVT*, 31 :3736–3764, 10 2021.

[6] Rich Franzen. Kodak lossless true color image suite. `http://r0k.us/graphics/kodak`, 1999.

[7] Albert Gu et Tri Dao. Mamba : Linear-time sequence modeling with selective state spaces. *arXiv*, 2023.

[8] Yue Liu et Tian et al. Vmamba : Visual state space model. *arXiv*, 2024.

[9] Lucas Theis George Toderici et al. Clic 2020 : Challenge on learned image compression. 2020.

[10] JPEG-AI. Jpeg-ai test images. `https://jpegai.github.io/test_images/`, 2020.

[11] Fabrice Bellard. Bpg image format. `http://bellard.org/bpg/`, 2018.

[12] Shiyu Qin, Jinpeng Wang, Yimin Zhou, Bin Chen, Tianci Luo, Baoyi An, Tao Dai, Shutao Xia, et Yaowei Wang. Mambavc : Learned visual compression with selective state spaces. *arXiv*, 2024.

[13] Bouzid Arezki, Fangchen Feng, et Anissa Mokraoui. Convolutional transformer-based image compression. Dans *SPA*, pages 154–159. IEEE, 2023.

[14] Wei Jiang, Jiayu Yang, Yongqi Zhai, Peirong Ning, Feng Gao, et Ronggang Wang. Mlic : Multi-reference entropy model for learned image compression. ACM, Octobre 2023.

[15] Jinming Liu, Heming Sun, et Jiro Katto. Learned image compression with mixed transformer-cnn architectures. Dans *CVPR*, pages 14388–14397, 2023.

[16] Biao Zhang et Rico Sennrich. Root mean square layer normalization. *NeurIPS*, 32, 2019.

[17] Jean Bégaint, Fabien Racapé, Simon Feltman, et Akshay Pushparaja. Compressai : a pytorch library and evaluation platform for end-to-end compression research, 2020.

Gestion du déséquilibre de classe au sein du classifieur de séries d'images astronomiques *ConvEntion*

Anass BAIROUK[1], Marc CHAUMONT[1,2], Dominique FOUCHEZ[3],
Jérome PASQUET[4], Frédéric COMBY[1]
[1] *LIRMM, Équipe ICAR, Université de Montpellier, CNRS, Montpellier, France*
[2] *Université de Nîmes, Nîmes, France*
[3] *CPPM, Université d'Aix-Marseille, Marseille, France*
[4] *TETIS, Université Paul Valéry Montpellier 3, Montpellier, France*
{anass.bairouk, marc.chaumont, frederic.comby}@lirmm.fr, dominique.fouchez@cppm.in2p3.fr, jerome.pasquet@univ-montp3.fr

Résumé

En classification de séquences d'images astronomiques, l'approche état-de-l'art (ConvEntion) repose sur l'utilisation d'une architecture basé sur une convolution 3D et un transformer. Cette architecture ConvEntion ne gère pas parfaitement le déséquilibre de classes. Dans cet article, nous proposons d'améliorer cela en s'appuyant sur les méthodologies auto-supervisées. Nous réduisons la variance intra-classe en passant par une architecture à deux branches. Chacune des deux branches traite une version augmentée de la donnée d'entrée. Dans le même temps, nous conservons la contrainte de classification ce qui nous permet de faire l'apprentissage sur un petit ensemble de données labellisées. Les résultats de notre modèle ICT-ConvEntion nous ont permis d'obtenir une amélioration de l'exactitude (accuracy) de 2.3% et du score F1 de 4.7% sur la base SDSS Supernova Survey.

Mots clefs

Astrophysique, Séquence d'images, Deep Learning, Classification, Transformer, DINO, BYOL, ConvEntion.

1 Introduction

Afin de déterminer les constantes astrophysiques et mieux comprendre l'univers, les astrophysiciens ont notamment besoin de détecter les supernovae Ia [1]. La détection consiste à pointer un télescope vers une zone du ciel, à suivre l'événement durant un certain nombre de nuits [2], puis à classer l'évenement. Historiquement, la séquence d'imagettes centrée sur l'événement était transformée en plusieurs séries de scalaires (une par bande) et appelée courbe de lumière. Ces courbes de lumière étaient alors utilisées pour classer le phénomène dans une des classes d'intérêt comme par exemple une supernova Ia.

1. Les supernovae de type Ia sont particulièrement intéressantes en raison de leur mécanisme d'explosion standard produisant une énergie de flux presque identique pour chaque supernova, ce qui permet de déduire la distance lumineuse de la supernova qui a explosé.
2. Une supernova est observable durant 100 à 200 nuits.

Nous avons récemment proposé l'approche « ConvEntion » qui permet de traiter directement la série d'imagettes issue du télescope. L'approche repose entre autre sur l'utilisation d'un "Transformer", le formatage des données afin de gérer l'absence d'imagettes et de bandes de fréquence, la réduction de la complexité via l'utilisation d'un 3DCNN, etc. Nous avons publié une version courte dans la conférence française GRETSI [1] et une version journal dans Astronomy & Astrophysics [2].
ConvEntion dépasse largement l'état de l'art sur la base Sloan Digital Sky Survey (SDSS) [3] sur une tache de classification à 4 classes difficiles. Nous améliorons l'exactitude ("accuracy") de 13% par rapport à l'approaches préliminaire de [4] et 14% par rapport à la meilleur approche basée courbe de lumière [5].
Dans cet article, nous proposons de mieux gérer le déséquilibre des classes en forçant les représentations latentes de différentes versions d'une même séquence d'image à être proches. Pour cela, nous avons repris le principe d'apprentissage de type « teacher-student » présents dans les approches auto-supervisé comme BYOL [6] et DINO [7]. Ainsi, en plus d'avoir un terme de perte (i.e. loss) pour la classification supervisée, nous ajoutons un terme de perte non-supervisé en conjonction d'une architecture « à la manière » de BYOL/DINO.
Dans la section 2 nous présentons notre proposition appelée ICT-ConvEntion (pour *Inbalanced Classes Treatment* ConvEntion) qui est une amélioration de ConvEntion. Dans la section 3 nous présentons brièvement les résultats [3].

2 Architecture utilisée pour l'apprentissage de ICT-ConvEntion

ICT-ConvEntion reprend la structure d'apprentissage des approches auto-supervisées récentes comme [6, 7] etc. Cependant, ICT-ConvEntion est une architecture qui peut apprendre avec des petits ensembles de données. Par ailleurs comme les méthodes contrastives [9, 10], notre approche

3. Ces résultats sont plus amplement détaillés dans la thèse d'Anass Bairouk [8].

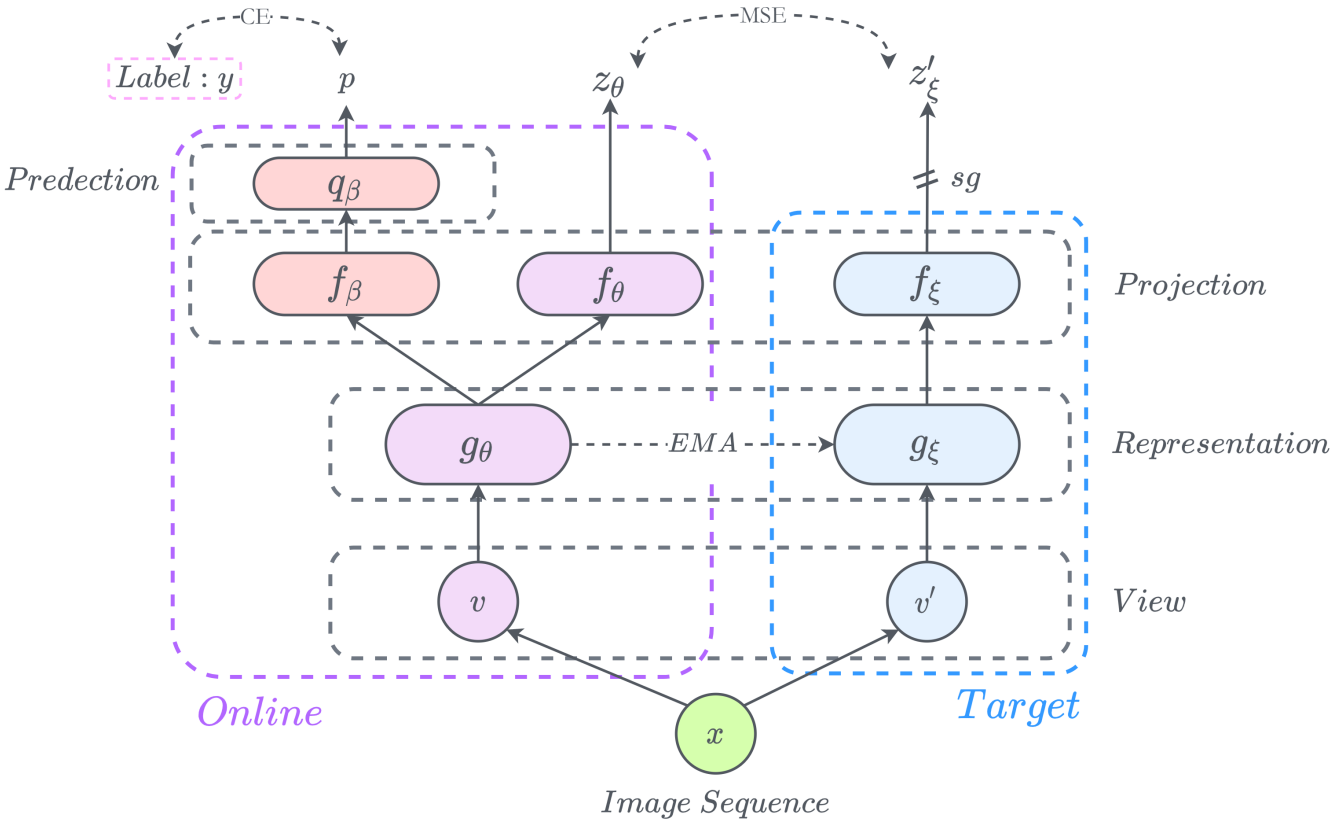


FIGURE 1 – *L'architecture générale d'ICT-ConvEntion. L'apprentissage d'ICT-Convention consiste à minimiser la différence entre z_θ et z'_ξ en utilisant l'erreur quadratique moyenne, ainsi qu'à minimiser l'entropie croisée entre y et p , où θ représente les poids, ξ représente les poids mis à jour par moyenne mobile exponentielle (EMA) de θ, et sg signifie "stop-gradient" (le sg empêche la rétropropagation sur le réseau cible).*

repose sur l'augmentation des données pour l'apprentissage de la représentation. C'est en partie grâce cette augmentation que nous gérons mieux le déséquilibre des classes.

La figure 1 illustre la structure d'apprentissage d'ICT-ConvEntion. Dans un premier temps, nous créons deux séquences d'images augmentées (des vues) v et v', à partir de la séquence originale. Plusieurs techniques d'augmentation ont été utilisées, notamment la suppression aléatoire d'images de la séquence pour en créer une nouvelle, la rotation des images de la séquence, le retournement horizontal et vertical, et le décalage temporel pour créer une séquence plus courte que l'original.

De la même manière que pour BYOL/DINO, nous utilisons deux réseaux neuronaux dont l'architecture est identique (nous utilisons notre architecture ConvEntion [2]), appelés réseau *en ligne* et réseau *cible*. Le réseau en ligne (resp. cible) est défini par un ensemble de poids θ, β (resp. ξ), et est composé de trois (resp. deux) étapes. A la première étape, se trouve un encodeur g_θ (resp. g_ξ) qui reprend l'architecture de ConvEntion. L'objectif de l'encodeur est l'apprentissage de la représentation. La deuxième étape contient deux parties pour le réseau en ligne, f_β et f_θ afin de projeter la représentation issue de g_θ. Le réseau cible ne contient qu'une seul brique de projection f_ξ. La projection f_β est utilisée par la troisième étape (pour la classification supervisée). Les projections f_θ et f_ξ sont utilisées pour minimiser la distance entre les représentations des réseaux en ligne et cible. La dernière étape est spécifique au réseau en ligne et à la partie classification. Cette étape contient une couche de prédiction q_β qui est utilisée pour minimiser l'entropie croisée avec le label vérité terrain. A noter que g_θ et f_θ partagent la même architecture respectivement avec g_ξ et f_ξ.

Après avoir obtenu les deux vues augmentées, v et v', de la séquence d'images, nous transmettons la première vue v', au réseau cible pour qu'elle passe par l'encodeur et le projecteur afin d'obtenir la sortie z'_ξ. Pendant ce temps, nous faisons passer v par le réseau en ligne pour obtenir la projection z_θ et la prédiction de classe p. Nous normalisons ensuite les projections des deux réseaux $\tilde{z}_\theta = z_\theta / ||z_\theta||_2$ et $\tilde{z'}_\xi = z'_\xi / ||z'_\xi||_2$.

La fonction de perte inspirée de l'apprentissage auto-supervisé, est appliquée sur les deux projections normalisées comme suit :

$$\mathcal{L}_{MSE}=||\tilde{z}_\theta - \tilde{z}'_\xi||_2^2 = 2 - 2 \cdot \frac{\langle z_\theta, z'_\xi \rangle}{||z_\theta||_2 \cdot ||z'_\xi||_2}. \quad (1)$$

La fonction de perte de classification est calculée entre le label, y, et la prédiction du réseau en ligne, p, comme suit :

$$\mathcal{L}_{CE} = -\sum_{c=1}^{M} y_{x,c} \log(p_{v,c}), \quad (2)$$

où M est le nombre de classes, c la classe, tandis que x et v sont respectivement la séquence d'entrée et la nouvelle vue générée à partir de la séquence d'entrée. Nous symétrisons les deux pertes $\mathcal{L}_{MSE}$ et $\mathcal{L}_{CE}$, en envoyant v' au réseau en ligne, et v au réseau cible, afin d'obtenir deux autres fonction de perte $\tilde{\mathcal{L}}_{MSE}$ et $\tilde{\mathcal{L}}_{CE}$. Nous calculons ensuite la perte totale comme suit :

$$\mathcal{L} = \mathcal{L}_{MSE} + \mathcal{L}_{CE} + \tilde{\mathcal{L}}_{MSE} + \tilde{\mathcal{L}}_{CE}. \quad (3)$$

A noter que les paramètres du réseau cible sont mis à jour à l'aide d'une moyenne mobile exponentielle (EMA) des paramètres du réseau en ligne. Les poids ξ sont mis à jour en suivant la relation $\xi \leftarrow \tau\xi + (1-\tau)\theta$ où $\tau \in [0, 1]$ est le taux de déclin (decay). Les projections f_β, f_θ, f_ξ consistent en deux couches entièrement connectées avec une activation ReLU et un "dropout" entre les deux. Pour le réseau en ligne, la couche de prédiction q_β consiste en une couche entièrement connectée qui prend en entrée la sortie de f_β et produit une sortie avec une dimension qui correspond au nombre de classes dans l'ensemble de données.

3 Détails expérimentaux et résultats/discussions

3.1 La base de données issue du SDSS

Le Sloan Digital Sky Survey (SDSS) [3] est un programme d'étude qui recueille des images, des spectres et des informations descriptives de millions d'objets célestes à l'aide d'un télescope dédié équipé d'instruments photométriques et spectroscopiques. Nous utilisons dans cet article le SDSS Supernova Survey, une composante de l'extension du SDSS-II portant sur la période 2005 à 2008. L'étude des supernovae consistait à imager la même région du ciel tous les deux soirs, en utilisant cinq filtres à large bande pour construire une vaste base de données d'images afin de découvrir de nouveaux objets célestes.

La base de données issue de cette étude comprend des séquences d'images d'étoiles variables galactiques, de noyaux actifs de galaxie (AGN), de supernovae (SNe) et d'autres transitoires astronomiques. Certaines de ces séquences sont également complétées du résultat spectroscopique, pour notamment identifier avec certitude les SNe et mesurer leur décalage vers le rouge. Cette base de données comporte un fort déséquilibre des classes, qui peut constituer un obstacle important pour l'apprentissage profond ainsi que la présence de classes dont les caractéristiques sont très similaires.

Nous avons séparé le jeu de données en deux ensembles de données : l'un contenant des données de typage photométrique (dont la vérité terrain peut être erronée) et l'autre des données confirmées par spectroscopie (la vérité terrain est certaine). Nous avons d'abord entraîné le modèle sur les données photométriques, puis nous avons utilisé l'apprentissage par transfert pour affiner le modèle sur les données confirmées par spectroscopie. Le tableau 1 résume la partition et les classes des données que nous avons utilisées dans ce travail.

Class	Pre-Train	FineTune	Test
AGN	362	362	182
SNIa	1448	400	99
Variable	1290	1290	645
SNOther	2041	72	17

TABLEAU 1 – *Nombre d'objets dans chaque ensemble de données. Le Pre-Train ne contient que des données photométriques, le FineTune et le Test ne contiennent que des données confirmées par spectroscopie.*

3.2 Hypers-paramètres et détails d'implémentation

Le modèle a été entraîné en utilisant l'optimiseur adamw, une taille de lot de 128 répartie sur 4 GPU. Le taux d'apprentissage a été fixé à 2×10^{-3} et un « dropout » de 0,3. Nous diminuons le taux d'apprentissage avec un cosinus « schedule ». Le taux de décroissance pour l'EMA est fixé à $\tau = 0,99$. Pour la brique « ConvEntion » nous avons utilisé les mêmes paramètres que ceux utilisés dans l'article original [2] avec $K = 3$ et $M = 99$. Le nombre de couches de ConvEntion est fixé à $L = 2$ et le nombre de têtes d'attention à $T = 4$. Les modèles sont entraînés à l'aide d'une validation croisée de cinq plis. Toutes les architectures présentées dans cet article suivent ce même processus et sont implémentées en utilisant PyTorch.

3.3 Résultats et discussions

Le tableau 2 contient les évaluations sur les quelques approches état-de-l'art utilisant des séries d'images astronimiques. Notre solution obtient une exactitude (accuracy) de 82,18% et un score F1 [4] de 75,33%. Notre proposition ICT-ConvEntion permet d'obtenir un gain en score F1 d'environ 5% par rapport à l'approche ConvEntion. Par ailleurs, les exactitudes (accuracy) par classe sont bien plus équilibrées pour ICT-ConvEntion (cela va de 77% à 83%) que pour ConvEntion (cela va de 67% à 81%) [5]. Ces résultats confirment l'intérêt à introduire des fonctions de perte

4. $F1 = 2\frac{presicion.rappel}{precision+rappel}$.
5. Voir les matrices de confusion dans la thèse d'Anass Bairouk[8].

Model	Accuracy (%)	F1 (%)
ICT-ConvEntion	**82,18**	**75,33**
ConvEntion [2]	79,83	70,62
CNN+GRU [4]	66,39	63,22
CNN+LSTM [11]	64,08	60,65

TABLEAU 2 – *Comparaison des performances en termes de score F1 moyen et de précision moyenne sur 5 plis de validation croisée.*

contraignant l'espace latent. Cela permet, entre autres, de mieux prendre en compte le déséquilibre des classes, mais également d'augmenter l'exactitude (accuracy) de la classification. La fonction de perte MSE (Eq. 1) réduit efficacement la variance intra-classe, ce qui permet à la fonction de perte de classification (Eq. 2) d'augmenter la séparabilité des classes. Notons tout de même que le temps d'apprentissage de la méthode ICT-ConvEntion est assez coûteux, puisqu'il prend trois fois plus de temps que l'apprentissage du modèle ConvEntion. La réduction du temps d'apprentissage est donc une des possibles pistes de recherches futures.

4 Conclusion

En conclusion, l'approche ICT-ConvEntion proposée pour traiter la classification des séries temporelles d'images astronomiques s'est avérée efficace pour augmenter l'exactitude (accuracy) de la classification. Les résultats montrent une augmentation de l'exactitude de 2,3% et du score F1 de 4,7% sur la base le SDSS Supernova Survey par rapport à l'état de l'art ConvEntion. Cette amélioration est grandement due à une meilleure gestion du déséquilibre des classes grâce à une fonction de perte réduisant les distances inter-classes, qui est obtenue via une architecture « à la BYOL/DINO », tout en gardant la fonction de perte pour la classification.

Remerciement

Ce travail a été réalisé grâce au soutien du projet ANR DEEPDIP (ANR-19-CE31-0023). Ce travail utilise les données du Sloan Digital Sky Survey (SDSS).

Références

[1] Anass Bairouk, Marc Chaumont, Dominique Fouchez, Jérome Pasquet, et Frédéric Comby. ConvEntion : Classification des séries chronologiques d'images astronomiques à l'aide d'attention convolutive. Dans *GRETSI 2022 - 28e Colloque Francophone de Traitement du Signal et des Images*, numéro 001-034 dans 28, pages 137–140, Nancy, France, Septembre 2022.

[2] Anass Bairouk, Marc Chaumont, Dominique Fouchez, Frédéric Comby, Jerome Pasquet, et Julian Bautista. Astronomical image time series classification using CONVolutional attENTION (ConvEntion). *Astronomy and Astrophysics - A&A*, 673 :A141, 2023.

[3] Jon A. Holtzman et al.. The sloan digital sky survey-ii : Photometry and supernova ia light curves from the 2005 data. *The Astronomical Journal*, 136(6) :2306, nov 2008.

[4] Catalina Gómez, Mauricio Neira, Marcela Hernández Hoyos, Pablo Arbeláez, et Jaime E Forero-Romero. Classifying image sequences of astronomical transients with deep neural networks. *Monthly Notices of the Royal Astronomical Society*, 499(3) :3130–3138, 10 2020.

[5] A Möller et T de Boissière. SuperNNova : an open-source framework for Bayesian, neural network-based supernova classification. *Monthly Notices of the Royal Astronomical Society*, 491(3) :4277–4293, 12 2019.

[6] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, et Michal Valko. Bootstrap your own latent a new approach to self-supervised learning. Dans *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, et Armand Joulin. Emerging properties in self-supervised vision transformers. Dans *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9630–9640. IEEE, 2021.

[8] Anass Bairouk. *Astronomical Image Time-series Classification Using Deep Learning*. Theses, Université de Montpellier, Octobre 2023.

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, et Geoffrey Hinton. A simple framework for contrastive learning of visual representations. Dans *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

[10] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, et Ross Girshick. Momentum contrast for unsupervised visual representation learning. Dans *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735, 2020.

[11] Rodrigo Carrasco-Davis, Guillermo Cabrera-Vives, Francisco Förster, Pablo A. Estévez, Pablo Huijse, Pavlos Protopapas, Ignacio Reyes, Jorge Martínez-Palomera, et Cristóbal Donoso. Deep learning for image sequence classification of astronomical events. *Publications of the Astronomical Society of the Pacific*, 131(1004) :108006, sep 2019.

Évaluation de méthodes d'édition des attributs faciaux basées sur des modèles génératifs profonds*

L. Bour S. Bougleux C. Charrier O. Lézoray

Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, 14000 Caen, France

lilian.bour,bougleux,christophe.charrier,olivier.lezoray}@unicaen.fr

Résumé

Nous proposons une évaluation approfondie de modèles génératifs permettant l'édition de visages. Plusieurs critères sont évalués : la qualités des images générées, la modification effective des attributs faciaux visés et conservation de l'identité. Les trois modèles évalués sont StarGAN, VecGAN et DiffAE. Les résultats montrent que StarGAN produit des résultats de faible qualités. Tandis que VecGAN et DiffAE ont une bonne conservation de l'identité et qualité visuelle, mais ne sont pas encore assez efficaces dans la modification de l'attribut visé de manière non entrelacée.

Mots clefs

Edition d'attributs faciaux, Modèles génératifs, Conservation de l'identité, Entrelacement des attributs.

1 Introduction

Les modèles génératifs peuvent permettre d'éditer les attributs faciaux d'images de visages en ajoutant de la `barbe` ou changeant la couleur des `cheveux` par exemple. Ces éditions peuvent être faites en ciblant directement l'ajout ou la suppression d'un attribut facial (`barbe`, `genre`, ...). Une difficulté majeure que ces modèles rencontrent est un problème d'entrelacement d'attributs. En effet, changer la `barbe` sur un visage ne devrait pas changer le `genre` de ce dernier. L'objectif de cet article est de proposer un cadre pour l'évaluation approfondie de ces méthodes selon trois axes : 1) La qualité des images générées, 2) la conservation de l'identité après édition, 3) la capacité à correctement modifier les attributs ciblés sans entrelacement. Trois modèles sont sélectionnés pour cette évaluation : StarGAN [1], VecGAN [2], et DiffAE [3]. Ils sont basés sur des GAN ou bien un modèle de diffusion. Nous les présentons dans la suite.

2 Modèles d'édition d'attributs faciaux

StarGAN [1] est un GAN dont le générateur est basé sur styleGAN et le discriminateur sur patchGAN. Il est entrainé afin de minimiser une fonction cout composée de trois termes : une loss adversariale afin de générer des images réalistes, une loss de classification pour assurer la modification de l'attribut visé, et une loss garantissant la bonne reconstruction de l'image après édition.

VecGAN [2] pour lequel les directions d'édition des attributs faciaux sont apprises dans un espace latent et régularisées afin d'être orthogonales entre elles pour un désentrelacement. Le générateur est un auto-encoder avec un méchanisme d'attention. VecGAN est entrainé à partir de cinq termes : une loss adversariale pour le réalisme de l'image, une loss de reconstruction, une loss de consistence entre image originale et éditée, une loss d'orthogonalité entre les directions d'éditions et une loss de désentrelacement.

DiffAE [3], est un modèle de diffusion basé sur le modèle Denoising Diffusion Implicit Models (DDIMs) [4] qui permet de convertir une image en une variable latente en effectuant le processus de diffusion inverse. Une des contributions phare de ce modèle est l'utilisation d'une information sémantique extraite de l'image source afin de guider l'édition du visage.

3 Méthode d'évaluation

Nous considérons un ensemble $I = \{I_i\}_i$ composé de n_I visages réprésentant $n_C \leq n_I$ personnes. $I(c) \subset I$ est l'ensemble des visages pour une identité c, parmi les C identités de I. Notons I_i^k l'image éditée selon l'attribut a_k et $I^k = \{I_i^k\}_i$ l'ensemble des images générées à partir de I. Trois types d'opérations sont possibles : l'*ajout*, la *suppression* d'un attribut, ou bien *reconstruction* de l'image source (sans modification). Nous détaillons dans la suite les mesures d'évaluation considérées.

3.1 Mesures de qualité

La mesure de similarité structurelle (SSIM) [5], quantifie la similarité perceptuelle entre une image source I_i et une éditée I_i^k. Notée SSIM(I_i, I_i^k), elle permet d'évaluer les changements globaux. La distance de Fréchet (*Frechet Inception Distance*–FID) [6], permet de quantifier la qualité visuelle d'un ensemble d'images éditées I^k et la diversité des images éditées par rapport aux images originales I.

3.2 Mesures de préservation d'identité

Les changements d'identités causés par l'édition sont évalués à partir d'un vecteur de carcatéristiques, pouvant

*Ce travail a bénéficié d'un financement de Saint-Lô Agglomération et de la Région Normandie.

être extrait grâce à des modèles de reconnaissance faciale, comme Swinface [7], ou encore avec des landmarks extraits d'un visage, tel que la librairie DLIB. Une image I_i^k génerée à partir de I_i, est comparée à toutes les images sources dans I, en utilisant la distance L_2 : $d(I_i^k, I_j) = \|fi(I_j) - fi(I_i^k)\|_2$, où fi est le vecteur caractéristique d'une image. Une identité c pouvant correspondre à plusieurs images sources $I(c)$, la proximité à c est mesurée par la distance moyenne aux images dans $I(c)$: $\bar{d}(I_i^k, c) = \frac{1}{|I(c)|} \sum_{I_j \in I(c)} d(I_i^k, I_j)$.
Dans un premier temps, la conservation de l'identitée est évaluée en utilisant la distance entre l'identité de I_i et les autres identités, en moyenne sur toutes les images sources I_i avec :

$$\mathrm{IDD}^k = \frac{1}{n_I} \sum_i \frac{(n_\mathrm{C} - 1)\bar{d}(I_i^k, c_i)}{\sum_{c \neq c_i} \bar{d}(I_i^k, c)}. \quad (1)$$

Dans un deuxième temps, elle l'est cette fois ci de manière plus globale, en mesurant l'aire sous la courbe ROC (AUC ROC) dans un but d'authentification : on essaye de retrouver l'image source I_i d'une image éditée I_i^k.

3.3 Mesures des changements d'attributs

Nous évaluons les changements d'attributs ciblés (a_k) ou non, en utilisant un classifieur d'attribut facial fa, fournissant un vecteur binaire $fa(I_i) \in \{0,1\}^{n_A}$. Chaque composante $fa(I_i)_l$ indiquant la présence ou non de l'attribut a_l parmis n_A attributs.

$$\mathrm{FAC}_{=}^k = \frac{1}{n_I} \sum_i |fa(I_i)_k - fa(I_i^k)_k| \quad (2)$$

$$\mathrm{maxFAC}_{\neq}^k = \frac{1}{n_A - 1} \max_{i, l \neq k} |fa(I_i)_l - fa(I_i^k)_l| \quad (3)$$

$$\mathrm{FAC}_{\neq}^k = \frac{1}{n_I(n_A - 1)} \sum_i \sum_{l \neq k} |fa(I_i)_l - fa(I_i^k)_l|. \quad (4)$$

$\mathrm{FAC}_{=}^k$ mesure en moyenne les changements pour a_k, il devrait être proche de 1. $\mathrm{maxFAC}_{\neq}^k$ et $\mathrm{FAC}_{\neq}^k$ mesurent l'entrelacement maximum et moyen pour les autres attributs, les valeurs devraient être proches de 0.

4 Expériences

Trois datasets sont utilisés : Le premier est constitué d'un échantillon de 1000 images tirées aléatoirement de la partie test de CelebA[8], le second est la base Face Research Lab London Set (FRL) [9] et le dernier créé par nos soins (Custom), composé de 28 images. Les attributs faciaux utilisés, en fonction des disponibilités des modèles sont : `Frange`, `Age`, `Sourire`, `Genre`, `Lunettes`, `Cheveux bruns`, `Barbe`, `Cheveux blonds`, `Maquillage` et `Cheveux noirs`.

4.1 Qualité Visuelle

Dans le tableau 1, FID et SSIM présentent la qualité des images résultantes (figure 1). Ils permettent de mettre de côté le modèle StarGAN car pas assez performant. En outre il ne permet la modification que dans un sens selon la présence ou non de l'attribut visé. Pour FID et SSIM, DiffAE obtient de meilleurs résultats et donc produit des images de meilleures qualités.

TABLEAU 1 – *FID moyenne (best ↓) et SSIM (best ↑) pour l'ajout (AJO), la suppression (SUP) et reconstruction (REC).*

DiffAE						
Dataset	SUP		REC		AJO	
	FID↓	SSIM↑	FID↓	SSIM↑	FID↓	SSIM↑
Custom	68.50	**0.88**	**17.47**	**0.98**	**58.34**	**0.88**
FRL	**43.60**	**0.89**	**17.32**	**0.98**	**41.48**	**0.89**
CelebA	32.93	**0.88**	**12.88**	**0.98**	31.80	**0.88**
VecGAN						
Dataset	SUP		REC		AJO	
	FID↓	SSIM↑	FID↓	SSIM↑	FID↓	SSIM↑
Custom	**58.73**	0.83	45.28	0.85	68.32	0.76
FRL	44.32	0.86	34.94	0.88	61.99	0.81
CelebA	27.83	0.86	21.04	0.88	34.20	0.81
StarGAN						
Dataset	Attributs modifiés en fonction de leur présence					
	FID↓		SSIM↑			
Custom	116.69		0.75			
FRL	65.84		0.80			
CelebA	**25.34**		0.79			

4.2 Conservation de l'identité

La mesure IDD permet de mettre en avant une bonne conservation de l'identité de DiffAE dans les cas de reconstruction et d'ajout, comme montré dans le tableau 2. Tandis que VecGAN est meilleur dans les cas de suppression. L'attribut `Genre` est celui modifiant le plus l'identité, cette opération peut facilement amener des changements structurels au niveau du visage. De la même manière l'AUC ROC montre une légère supériorité de DiffAE, qui conserve globalement mieux l'identité.

4.3 Entrelacement des attributs

Les résultats du tableau 3 montrent que pour la reconstruction, DiffAE est plus efficace. Pour la suppression, ce dernier modifie mieux l'attribut ciblé a_k, mais à un entrelacement ($\mathrm{FAC}_{\neq}^k$, $\mathrm{maxFAC}_{\neq}^k$) plus élevé. Au contraire, pour l'ajout, VecGAN modifie mieux l'attribut ciblé et a un entrelacement plus élevé. Les résultats mettent en évidence que l'entrelacement des attributs n'est pas général mais spécifique, par exemple entre `Genre` et `Barbe` comme montré dans la figure 1.

5 Conclusion

L'évaluation réalisée a permis de mettre en avant DiffAE [3], un modèle de diffusion permettant de modifier un grand nombre d'attributs faciaux, tout en conservant bien l'identité des visages et en produisant des images de bonnes qualités. VecGAN [2] est un peu moins bon, mais est plus efficace pour ajouter un attribut. StarGAN [1] quant à lui, produit des résultats avec de fortes distorsions. Pour finir, DiffAE ne fournit pas toujours des résultats satisfaisants lorsqu'il s'agit de correctement modifier les attributs faciaux sans entrelacement, et les résultats présentent des artefacts

TABLEAU 2 – *Changement d'identité mesurés par l'IDD (↓) avec des Landmarks et Swinface. (Suppression : Sup, Reconstruction : Rec, Ajout : Ajo, DiffAE : D, VecGAN : V) La moyenne ne prend pas en compte* `Maquillage` *et* `Pas de Barbe` *pour DiffAE.*

Dataset	**Custom**						**FRL**						**CelebA**					
Operation	**Sup**		**Rec**		**Ajo**		**Sup**		**Rec**		**Ajo**		**Sup**		**Rec**		**Ajo**	
Modeles	**D**	**V**	**D**	**V**	**D**	**V**	**D**	**V**	**D**	**V**	**D**	**V**	**D**	**V**	**D**	**V**	**D**	**V**
Landmarks																		
`Frange`	0.25	0.17	0.16	0.19	0.28	0.28	0.37	0.29	0.21	0.24	0.36	0.45	0.86	0.86	0.86	0.86	0.87	0.87
`Age`	0.32	0.29	0.16	0.21	0.34	0.40	0.48	0.49	0.21	0.34	0.45	0.33	0.88	0.87	0.86	0.86	0.88	0.90
`Sourire`	0.57	0.36	0.16	0.24	0.66	0.52	0.69	0.22	0.21	0.23	0.78	0.72	0.90	0.87	0.86	0.87	0.96	0.94
`Genre`	0.43	0.48	0.16	0.35	0.42	0.31	0.62	0.77	0.21	0.52	0.58	0.46	0.91	0.91	0.86	0.87	0.90	0.88
`Lunettes`	0.47	0.33	0.16	0.26	0.40	0.47	N/A	N/A	0.21	0.36	0.54	0.66	0.90	0.91	0.86	0.86	0.88	0.90
`Cheveux bruns`	0.25	0.19	0.16	0.20	0.22	0.20	0.27	0.22	0.21	0.27	0.28	0.25	0.83	0.82	0.86	0.86	0.87	0.87
`Barbe`	0.26		0.16		0.41		0.36		0.21		0.44		0.87		0.86		0.87	
`Cheveux blonds`	0.19	0.16	0.16	0.23	0.25	0.39	0.26	0.25	0.21	0.30	0.35	0.63	0.88	0.87	0.86	0.88	0.86	0.86
`Maquillage`	0.39		0.16		0.39		0.61		0.21		0.53		0.87		0.86		0.91	
`Cheveux noirs`	0.23	0.18	0.16	0.21	0.25	0.27	0.34	0.27	0.21	0.28	0.31	0.34	0.87	0.86	0.86	0.86	0.87	0.87
Moyennes	0.34	**0.27**	**0.16**	0.24	**0.35**	0.36	0.43	**0.36**	**0.21**	0.32	**0.46**	0.48	0.88	**0.87**	**0.86**	0.87	0.89	0.89
SwinFace																		
`Frange`	0.18	0.10	0.01	0.04	0.23	0.30	0.23	0.12	0.01	0.04	0.25	0.33	0.63	0.56	0.43	0.45	0.59	0.56
`Age`	0.37	0.26	0.01	0.10	0.42	0.43	0.40	0.39	0.01	0.19	0.44	0.22	0.69	0.56	0.43	0.48	0.66	0.63
`Sourire`	0.39	0.15	0.01	0.07	0.36	0.32	0.51	0.02	0.01	0.03	0.36	0.35	0.68	0.54	0.43	0.47	0.66	0.64
`Genre`	0.55	0.48	0.01	0.31	0.53	0.37	0.52	0.59	0.01	0.33	0.51	0.50	0.75	0.67	0.43	0.52	0.80	0.73
`Lunettes`	0.43	0.44	0.01	0.13	0.25	0.46	N/A	N/A	0.01	0.20	0.27	0.57	0.60	0.70	0.43	0.48	0.61	0.70
`Cheveux bruns`	0.09	0.04	0.01	0.06	0.10	0.09	0.09	0.05	0.01	0.13	0.11	0.09	0.50	0.44	0.43	0.48	0.48	0.47
`Barbe`	0.28		0.01		0.26		0.28		0.01		0.21		0.61		0.43		0.58	
`Cheveux blonds`	0.14	0.06	0.01	0.14	0.15	0.47	0.19	0.08	0.01	0.16	0.18	0.56	0.58	0.50	0.43	0.63	0.54	0.54
`Maquillage`	0.45		0.01		0.48		0.55		0.01		0.44		0.77		0.43		0.72	
`Cheveux noirs`	0.25	0.04	0.01	0.10	0.22	0.28	0.22	0.07	0.01	0.16	0.24	0.41	0.55	0.42	0.43	0.48	0.60	0.60
Moyennes	0.30	**0.20**	**0.01**	0.12	**0.28**	0.34	0.31	**0.19**	**0.01**	0.16	**0.29**	0.38	0.62	**0.55**	**0.43**	0.50	0.62	**0.61**

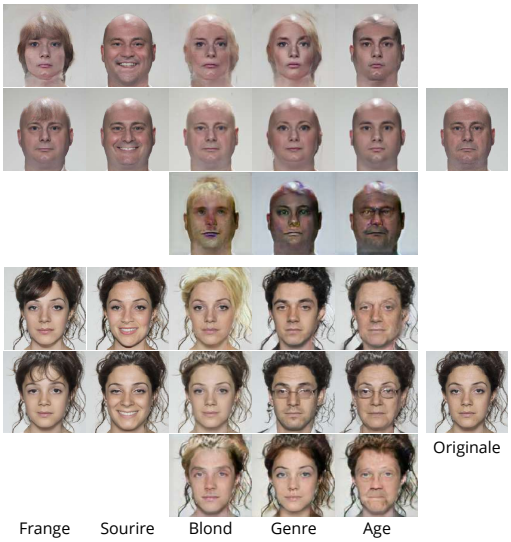


FIGURE 1 – *Exemples de résultats pour StarGAN (bas), DiffAE (Milieu) et VecGAN (Haut).*

visuels. Ces défauts sont des pistes d'améliorations à explorer pour les modèles génératifs d'éditions d'attributs faciaux.

Références

[1] Yunjey Choi, Min-Je Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, et Jaegul Choo. Stargan : Unified generative adversarial networks for multi-domain image-to-image translation. Dans *CVPR*, pages 8789–8797, 2018.

[2] Yusuf Dalva, Hamza Pehlivan, Oyku Irmak Hatipoglu, Cansu Moran, et Aysegul Dundar. Image-to-image translation with disentangled latent vectors for face editing. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(12) :14777–14788, 2023.

[3] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, et Supasorn Suwajanakorn. Diffusion autoencoders : Toward a meaningful and decodable representation. Dans *CVPR*, pages 10609–10619, 2022.

[4] Jiaming Song, Chenlin Meng, et Stefano Ermon. Denoising diffusion implicit models. Dans *ICLR*, 2021.

[5] Zhou Wang, A.C. Bovik, H.R. Sheikh, et E.P. Simoncelli. Image quality assessment : from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4) :600–612, 2004.

[6] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, et Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. Dans *NeurIPS*, pages 6626–6637, 2017.

[7] Lixiong Qin, Mei Wang, Chao Deng, Ke Wang, Xi Chen, Jiani Hu, et Weihong Deng. Swinface : A multi-task transformer for face recognition, expression recognition, age estimation and attribute estimation. *IEEE Trans. Circuits Syst. Video Technol.*, pages 1–1, 2024.

[8] Ziwei Liu, Ping Luo, Xiaogang Wang, et Xiaoou Tang. Deep learning face attributes in the wild. Dans *ICCV*, pages 3730–3738, 2015.

[9] DeBruine Lisa et Jones Benedict. Face research lab london set, 2017.

TABLEAU 3 – *Changement d'attributs faciaux mesurés par $FAC^k_=$, $FAC^k_{\neq}$, $maxFAC^k_{\neq}$, pour chaque type d'édition. (DiffAE : D, VecGAN : V, Modification correcte de la cible : (>0.5), Entrelacement : (>0.5) . La moyenne ne prend pas en compte* `Maquillage` *et* `Pas de Barbe` *pour DiffAE.*

Reconstruction																		
Dataset	**Custom**						**FRL**						**CelebA**					
Mesure	$FAC^k_=\downarrow$		$FAC^k_{\neq}\downarrow$		$maxFAC^k_{\neq}\downarrow$		$FAC^k_=\downarrow$		$FAC^k_{\neq}\downarrow$		$maxFAC^k_{\neq}\downarrow$		$FAC^k_=\downarrow$		$FAC^k_{\neq}\downarrow$		$maxFAC^k_{\neq}\downarrow$	
Modeles	**D**	**V**	**D**	**V**	**D**	**V**	**D**	**V**	**D**	**V**	**D**	**V**	**D**	**V**	**D**	**V**	**D**	**V**
`Frange`	0.04	0.07	0.05	0.06	0.29	0.25	0.00	0.06	0.06	0.05	0.23	0.17	0.01	0.06	0.03	0.03	0.21	0.07
`Pas de Barbe`	0.04		0.05		0.29		0.07		0.06		0.23		0.04		0.03		0.21	
`Cheveux noirs`	0.00	0.39	0.05	0.09	0.29	0.43	0.04	0.45	0.06	0.10	0.23	0.39	0.02	0.19	0.03	0.05	0.21	0.14
`Cheveux blonds`	0.04	0.57	0.05	0.13	0.29	0.39	0.00	0.63	0.06	0.12	0.23	0.48	0.01	0.60	0.03	0.09	0.21	0.34
`Cheveux bruns`	0.00	0.21	0.05	0.09	0.29	0.29	0.03	0.23	0.06	0.11	0.23	0.47	0.01	0.16	0.03	0.05	0.21	0.17
`Lunettes`	0.04	0.39	0.05	0.07	0.29	0.29	0.00	0.57	0.06	0.09	0.23	0.26	0.00	0.24	0.03	0.04	0.21	0.12
`Maquillage`	0.14		0.05		0.29		0.20		0.05		0.23		0.12		0.03		0.21	
`Genre`	0.07	0.43	0.05	0.18	0.29	0.54	0.07	0.54	0.06	0.21	0.23	0.67	0.00	0.24	0.03	0.06	0.21	0.25
`Sourire`	0.04	0.14	0.05	0.06	0.29	0.29	0.06	0.07	0.06	0.04	0.23	0.18	0.02	0.22	0.03	0.04	0.21	0.19
`Age`	0.00	0.00	0.05	0.10	0.29	0.21	0.00	0.00	0.06	0.10	0.23	0.35	0.01	0.05	0.03	0.05	0.21	0.15
Moyenne	**0.03**	0.28	**0.05**	0.10	**0.29**	0.33	**0.02**	0.32	**0.06**	0.10	**0.23**	0.37	**0.01**	0.22	**0.03**	0.05	0.21	**0.18**
Suppression																		
Dataset	**Custom**						**FRL**						**CelebA**					
Mesure	$FAC^k_=\uparrow$		$FAC^k_{\neq}\downarrow$		$maxFAC^k_{\neq}\downarrow$		$FAC^k_=\uparrow$		$FAC^k_{\neq}\downarrow$		$maxFAC^k_{\neq}\downarrow$		$FAC^k_=\uparrow$		$FAC^k_{\neq}\downarrow$		$maxFAC^k_{\neq}\downarrow$	
Modeles	**D**	**V**	**D**	**V**	**D**	**V**	**D**	**V**	**D**	**V**	**D**	**V**	**D**	**V**	**D**	**V**	**D**	**V**
`Frange`	0.17	0.17	0.14	0.10	0.67	0.50	0.36	0.27	0.12	0.06	0.45	0.27	0.54	0.51	0.09	0.06	0.24	0.18
`Pas de Barbe`	0.11		0.14		0.42		0.10		0.15		0.65		0.06		0.08		0.39	
`Cheveux noirs`	0.78	0.44	0.13	0.06	0.44	0.33	0.77	0.18	0.12	0.06	0.53	0.16	0.56	0.22	0.07	0.03	0.29	0.12
`Cheveux blonds`	0.11	0.33	0.10	0.09	0.33	0.33	0.59	0.24	0.12	0.09	0.71	0.41	0.39	0.16	0.06	0.03	0.23	0.15
`Cheveux bruns`	0.25	0.00	0.09	0.07	0.63	0.38	0.48	0.22	0.08	0.05	0.44	0.19	0.33	0.17	0.05	0.03	0.28	0.12
`Lunettes`	1.00	1.00	0.15	0.13	1.00	1.00	N/A	N/A	N/A	N/A	N/A	N/A	0.85	0.85	0.08	0.11	0.30	0.27
`Maquillage`	0.78		0.22		0.78		0.29		0.23		1.00		0.65		0.19		0.95	
`Genre`	0.65	0.47	0.28	0.27	0.94	0.88	0.68	0.92	0.28	0.34	1.00	0.92	0.95	0.79	0.18	0.15	0.88	0.65
`Sourire`	0.80	0.10	0.14	0.08	0.60	0.50	1.00	0.00	0.18	0.05	1.00	0.50	0.98	0.63	0.10	0.07	0.55	0.45
`Age`	0.00	0.00	0.17	0.15	0.48	0.33	0.00	0.00	0.18	0.18	0.67	0.47	0.36	0.10	0.11	0.09	0.40	0.33
Moyenne	**0.47**	0.31	0.15	**0.12**	0.64	**0.53**	**0.55**	0.26	0.15	**0.12**	0.69	**0.42**	**0.62**	0.43	0.09	**0.07**	0.40	**0.28**
Addition																		
Dataset	**Custom**						**FRL**						**CelebA**					
Mesure	$FAC^k_=\uparrow$		$FAC^k_{\neq}\downarrow$		$maxFAC^k_{\neq}\downarrow$		$FAC^k_=\uparrow$		$FAC^k_{\neq}\downarrow$		$maxFAC^k_{\neq}\downarrow$		$FAC^k_=\uparrow$		$FAC^k_{\neq}\downarrow$		$maxFAC^k_{\neq}\downarrow$	
Modeles	**D**	**V**	**D**	**V**	**D**	**V**	**D**	**V**	**D**	**V**	**D**	**V**	**D**	**V**	**D**	**V**	**D**	**V**
`Frange`	0.86	0.86	0.16	0.20	0.41	0.41	0.90	0.99	0.16	0.18	0.48	0.48	0.74	0.69	0.09	0.08	0.25	0.18
`Pas de Barbe`	0.22		0.13		0.56		0.00		0.09		0.37		0.10		0.10		0.50	
`Cheveux noirs`	0.53	0.89	0.13	0.14	0.47	0.47	0.71	0.89	0.13	0.16	0.73	0.73	0.37	0.55	0.07	0.09	0.21	0.27
`Cheveux blonds`	0.37	0.95	0.11	0.28	0.32	0.74	0.31	0.99	0.13	0.29	0.48	0.73	0.17	0.41	0.08	0.07	0.29	0.20
`Cheveux bruns`	0.30	0.40	0.09	0.09	0.25	0.35	0.39	0.39	0.08	0.08	0.28	0.44	0.16	0.19	0.05	0.04	0.22	0.16
`Lunettes`	0.89	0.89	0.11	0.16	0.30	0.37	1.00	1.00	0.10	0.20	0.31	0.71	0.92	0.90	0.06	0.12	0.19	0.33
`Maquillage`	1.00		0.25		0.95		0.85		0.21		0.66		0.88		0.16		0.80	
`Genre`	0.64	0.36	0.28	0.19	0.82	0.45	0.49	0.47	0.24	0.19	0.96	0.69	1.00	0.84	0.20	0.15	0.89	0.59
`Sourire`	0.11	0.11	0.13	0.13	0.83	0.67	0.25	0.25	0.12	0.13	0.58	0.58	0.72	0.66	0.12	0.09	0.77	0.48
`Age`	0.00	0.00	0.16	0.27	0.71	0.71	0.00	0.00	0.20	0.06	1.00	0.67	0.15	0.14	0.11	0.12	0.52	0.41
Moyenne	0.46	**0.56**	**0.14**	0.18	**0.51**	0.52	0.51	**0.62**	**0.14**	0.16	**0.60**	0.63	0.53	**0.55**	0.10	0.10	0.42	**0.33**

Attaque d'une méthode d'obscuration d'images par bit-flipping

N. Hutte[1] W. Puech[1]

[1] LIRMM, Univ. Montpellier, CNRS, Montpellier, France

Résumé

De nos jours, il est de plus en plus nécessaire de sécuriser les données multimédia. Le chiffrement intégral est efficace et sûr, mais il ne garantit pas un bon compromis entre la sécurité et d'autres exigences. L'obscuration d'images est une solution possible, préservant l'intégrité des données multimédia tout en obscurcissant leur contenu. En 2021, Aprilpyone et Kiya [1] ont proposé une méthode d'obscuration d'image par bit-flipping, utilisant une clé secrète permettant d'obscurcir et de garder secrète l'image originale, tout en étant réversible. Toutefois, cette méthode présente une faille exploitable sujette à une attaque. Dans cet article, nous présentons une méthode pour attaquer les images obscurcies par bit-flipping, afin de reconstruire l'image originale sans aucune connaissance de la clé secrète.

Mots clefs

Sécurisation des images, obscuration d'images, attaque.

1 Introduction

Depuis plus de dix ans, la transmission et l'archivage sur le cloud d'images et de vidéos ont augmenté de façon spectaculaire. Afin de garantir leur confidentialité, il est nécessaire de les sécuriser visuellement. De nombreuses études sur les communications sécurisées, efficaces et flexibles ont été rapportées [2]. Pour les données multimédia, le chiffrement intégral avec une sécurité éprouvée (tel que les chiffrements RSA ou AES) est l'option la plus sûre. Toutefois, de nombreuses applications multimédia nécessitent un compromis entre sécurité et d'autres exigences, telles qu'un traitement peu coûteux et la préservation du format. C'est pourquoi des méthodes de chiffrement sélectif ont été étudiées [3]. Parmi les approches visant à protéger la vie privée tout en maintenant la qualité et l'intégrité des données, l'obscuration d'images (comme le flou [4] ou le mélange des pixels [5]) est une solution possible. Ces techniques sont pertinentes pour des applications en anonymisation et en confidentialité des images, dans des domaines tels que la vidéo-surveillance, la télé-médecine ou les réseaux sociaux, par exemple. Les méthodes réversibles permettent de reconstruire l'image originale à l'aide d'une clé secrète. En 2021, Aprilpyone et Kiya ont proposé une méthode réversible d'obscuration d'image bloc par bloc par bit-flipping, basée sur l'inversion d'un sous-ensemble de pixels pour chaque bloc d'une image [1].

Dans cet article, nous proposons une attaque de la méthode d'obscuration par bit-flipping proposée par Aprilpyone et Kiya [1]. Sur la base d'une analyse des bits de poids fort de tous les pixels de l'image obscurcie, nous proposons de reconstruire le motif binaire qui a été généré à partir d'une clé secrète au cours de l'étape d'obscuration. Nous pouvons alors reconstruire deux versions possibles pour chaque composante de couleur de l'image et, en les combinant, obtenir la reconstruction de huit images couleur. Sur la base d'un classifieur binaire, nous pouvons alors déduire laquelle de ces huit images peut correspondre à l'image originale, le tout sans aucune connaissance de la clé secrète.

2 Attaque proposée

De nombreuses attaques contre les méthodes d'obscuration d'images ont été proposées, en particulier pour les systèmes de chiffrement d'images basés sur les mélanges de pixels [6]. À notre connaissance, aucune attaque n'a été proposée pour les méthodes d'obscuration par bit-flipping. Dans cette section, nous présentons une attaque de la méthode d'obscuration par bit-flipping [1].

2.1 Obscuration par bit-flipping [1]

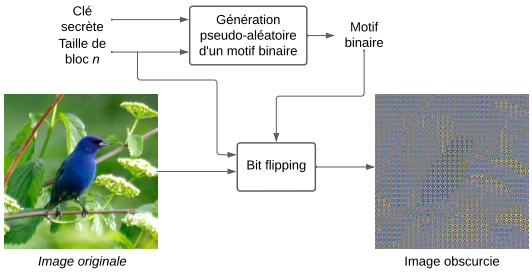


FIGURE 1 – *Obscuration d'images par bit-flipping [1].*

Dans la méthode d'obscuration d'images bloc par bloc proposée par Aprilpyone et Kiya [1], tous les bits d'un sous-ensemble de pixels de chaque bloc sont inversés. Cela est effectué séparément pour chaque composante couleur RGB, sur la base d'un motif binaire généré pseudo-aléatoirement par une clé secrète. Tous les blocs de $n \times n$ pixels de l'image subissent le même traitement, sur la base du même motif binaire. Comme l'illustre la figure 1, les principales étapes de cette méthode sont les suivantes :

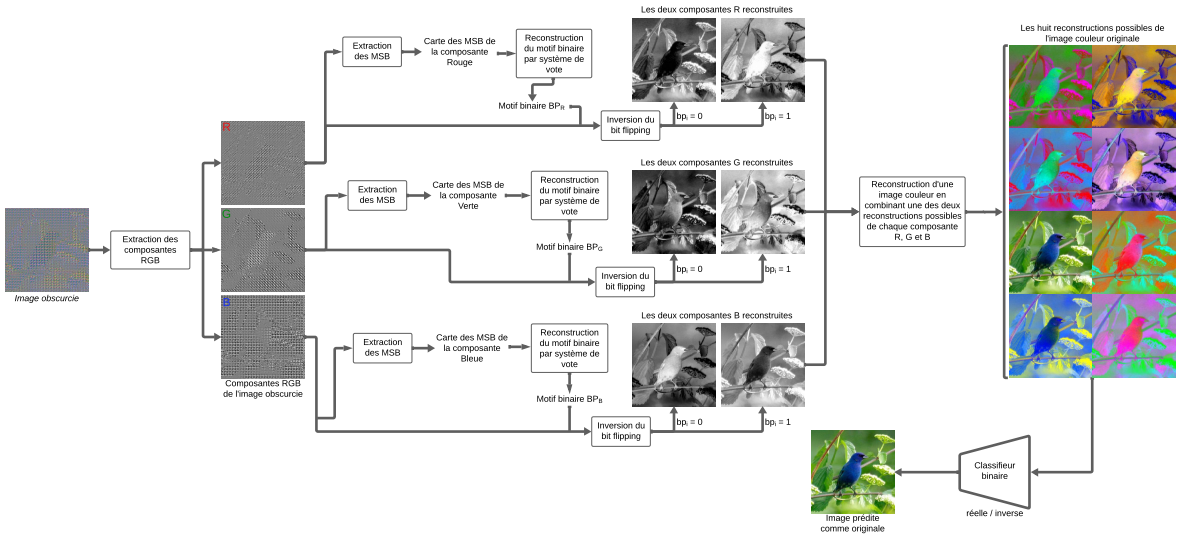


FIGURE 2 – *Attaque proposée pour la méthode d'obscuration par bit-flipping [1].*

1. Générer un motif binaire $\mathbf{BP}$ pseudo-aléatoire par composante couleur, de taille $n \times n$ bits, avec n le côté des blocs de pixels : $\mathbf{BP} = \{bp_0, \ldots, bp_i, \ldots, bp_{n^2-1}\}$.
2. Diviser chaque composante couleur de l'image en blocs de taille $n \times n$ pixels.
3. Lire les n^2 pixels p_i de chaque bloc de pixels et appliquer à chacun leur nouvelle valeur, telle que :

$$p'_i = \begin{cases} p_i, & \text{si } bp_i = 0, \\ p_i \oplus (2^L - 1), & \text{si } bp_i = 1, \end{cases} \quad (1)$$

avec L le nombre de bits par pixel pour composante couleur ($L = 8$ bits dans cet article).

Un point important dans cette méthode est qu'un seul motif binaire par composante de couleur est généré de manière pseudo-aléatoire. Il est ensuite utilisé de la même manière pour tous les blocs de pixels de l'image.

2.2 L'attaque proposée

La faille que nous exploitons dans la méthode par bit-flipping est qu'un motif binaire unique est utilisé pour l'ensemble des blocs de chaque composante de couleur d'une image. Comme ce motif binaire $\mathbf{BP}$ est appliqué de la même manière à tous les blocs, nous pouvons les analyser afin de le reconstruire, puis l'utiliser pour attaquer la méthode et tenter de reconstruire l'image originale sans la clé secrète. Cependant, pour cette attaque, nous supposons connaître le taille des blocs de pixels de l'image obscurcie. Pour chaque composante d'une image obscurcie, comme illustré en figure 2, nous appliquons d'abord les trois étapes suivantes de manière indépendante :

1. Diviser chaque composante couleur en blocs de taille $n \times n$ pixels et extraire le MSB (bit de poids fort) de chaque pixel.
2. Reconstruire le motif binaire de la composante couleur en votant sur les MSB extraits de chaque bloc et décider s'il s'agit d'un 0 ou d'un 1.
3. A partir du motif binaire reconstruit, pour chaque composante, appliquer l'opération inverse du bit-flipping à la composante couleur de l'image obscurcie. Toutefois, comme nous ne savons pas quel sous-ensemble de pixels a été inversé, nous pouvons reconstruire deux composantes différentes, inversées l'une par rapport à l'autre.

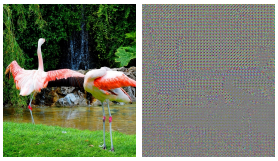
Ces trois premières étapes nous permettent d'obtenir 6 composantes différentes, avec deux versions reconstruites pour chaque composante couleur. À partir de ces 6 composantes, comme illustré en figure 2, nous pouvons reconstruire 8 images couleur différentes, dont une seule d'entre elles devrait correspondre à l'image originale. Pour déterminer laquelle de ces 8 images est prédite comme étant une image originale, comme illustré en figure 2, nous utilisons un modèle de classification binaire avec deux classes : "réelle" et "inverse".

3 Résultats expérimentaux

3.1 Exemple détaillé de l'attaque proposée

Nous appliquons d'abord notre attaque proposée à l'image Flamingo de 500×437 pixels de la base de données ILSVRC2012 (ImageNet Large Scale Visual Recognition Challenge 2012) [7], illustrée dans la figure 3.a. La figure 3.b montre les résultats obtenus avec la méthode d'obscuration par bit-flipping [1] avec des blocs de 10×10 pixels. A l'issue de cette obscuration, nous obtenons un PSNR [8] de 6,05 dB, un SSIM [9, 8] de 0,02, un UACI [10] de 26,95 %, un NPCR [10] de 10,98 % et un EDR [11] de 0,45.

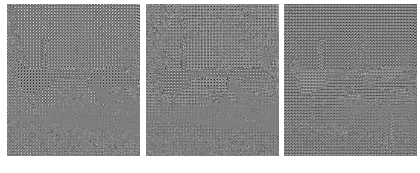
Pour l'attaque, la première étape consiste à décomposer l'image obscurcie, figure 3.b., en trois composantes cou-



(a) (b)

FIGURE 3 – *Obscuration par bit-flipping : a) Image originale issue de la base ILSVRC2012 [7], b) L'image obscurcie par bit-flipping [1] correspondante.*

leur, comme illustré dans la première ligne de figure 4. Sur la base du système de vote, nous pouvons alors reconstruire un motif binaire pour chaque composante couleur, comme illustré en seconde ligne de la figure 4.



(a) (b) (c)

(d) (e) (f)

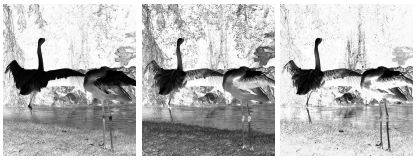
FIGURE 4 – *Les premières étapes de la méthode d'attaque proposée appliquée à l'image obscurcie Flamingo, figure 3.b : a), b) et c) Composantes RGB de l'image obscurcie, d), e) et f) Les motifs binaires déduits.*

À partir des trois composantes RGB de l'image obscurcie et de leurs motifs binaires reconstruits correspondants, illustrés en figure 4, nous pouvons alors reconstruire, pour chaque composante, deux images possibles en fonction du sous-ensemble de pixels que nous décidons d'inverser. Ainsi, à partir de l'image présentée dans la figure 3.b., nous reconstruisons six composantes couleur, deux par composante (figure 5.a à figure 5.f), et finalement nous avons huit combinaisons possibles pour reconstruire l'image originale comme illustré de la figure 5.g à la figure 5.n.
Pour déterminer laquelle des huit images reconstruites est la bonne, nous utilisons un classifieur binaire [1] qui a été entraîné pour deux classes : les images "réelles" (les images originales que nous recherchons) et les images "inversées" (les images dont au moins un des plans RGB est inversé, ce qui implique des couleurs erronées). Pour entraîner ce classifieur binaire, nous avons utliser les 50 000 images de validation de la base ILSVRC2012 [7] et appliqué une inversion aléatoire des composantes (inversion d'au moins une des trois composantes RGB), de manière équiprobable.

1. https://github.com/A-Jatin/CNN-implementation-for-binary-image-classification

Pour chaque image, les motifs binaires sont différents. 80 % de la base de données générée est utilisée pour l'entraînement, et 20 % pour tester l'attaque proposée.

(a) (b) (c)



(d) (e) (f)

(g) (h) (i) (j)

(k) (l) (m) (n)

FIGURE 5 – *Étapes suivantes de l'attaque proposée figure 4 : a) à f) Les possibles composantes RGB reconstruites, deux par composante, g) à n) Les huit images reconstruites possibles.*

Nous obtenons ainsi 80 000 images (40 000 "réelles" et 40 000 "inversées") dédiées à l'entraînement. Le classifieur binaire, conçu pour des images de taille (3, 128, 128) et composé de 4 couches (2 pour la convolution et 2 pour la partie fully connected) est entraîné pendant 50 epochs, avec une batch size de 32, en utilisant l'optimiseur Adam, et ReLU et Sigmoïde comme fonctions d'activation. Nous obtenons une précision de validation de 0,9714 %, sur les 20 000 images restantes (10 000 "réelles" et 10 000 "inverses").
En utilisant ce modèle entraîné, nous pouvons alors prédire si une image reconstruite par notre attaque correspond à une image "réelle" ou "inverse". Des huit images reconstruites présentées en figure 5, comme indiqué dans le tableau 1, nous constatons que seule la figure 5.g est prédite comme une image "réelle", tandis que les sept autres images reconstruites sont prédites comme des images "inverses". Et en effet, l'image reconstruite illustrée figure 5.g correspond bien à l'image originale. Nous avons donc réussi à reconstruire l'image originale sans la clé secrète, uniquement avec l'image obscurcie.

Image	Classe prédite	Probabilité
Originale	réelle	0.999
figure 5.g	réelle	0.999
figure 5.h	inverse	1.0
figure 5.i	inverse	1.0
figure 5.j	inverse	0.999
figure 5.k	inverse	1.0
figure 5.l	inverse	1.0
figure 5.m	inverse	1.0
figure 5.n	inverse	0.999

TABLEAU 1 – *Prédictions obtenues par le classifieur binaire pour les huit images reconstruites figure 5*

.

3.2 Analyse sur une plus large base d'images

Nous testons l'attaque proposée sur les 10 000 images restantes de la base de données présentée. Ces images ont toutes été obscurcies avec une clé différente. Après notre attaque, nous obtenons alors 80 000 images reconstruites.

Résultats		Nombre d'images	Total
Correctement prédites	Parfaitement reconstruites	8 063	9 201
	Partiellement reconstruites	1 138	
Incorrectement prédites		799	

TABLEAU 2 – *Prédiction pour les 10 000 images obscurcies attaquées par la méthode proposée.*

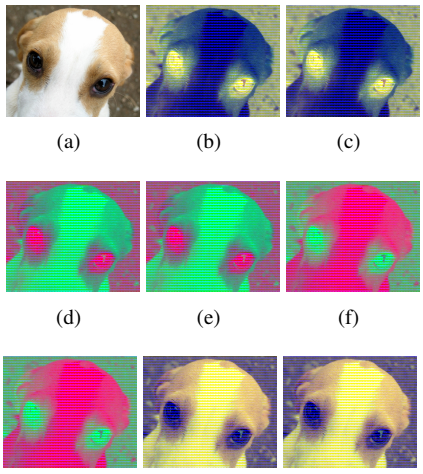
Le tableau 2 montre les résultats obtenus pour ces 10 000 images obscurcies. Parmi les 10 000 images obscurcies, le classifieur binaire a réussi à prédire 9 201 images comme étant la bonne image "réelle". Parmi ces images prédites comme "réelle", 8 063 sont parfaitement reconstruites (PSNR$= \infty$), les motifs binaires ont donc été parfaitement reconstruits. 1 138 images sont correctement prédites mais partiellement reconstruites, et conduisent à un PSNR moyen de 30 dB, comme illustré en figure 6. Cela est dû à un motif binaire mal reconstruit pendant la phase de vote, entraînant une inversion incomplète, et ainsi des différences imperceptibles ou des artefacts réguliers.

(a) (b)

FIGURE 6 – *Image partiellement reconstruite classifiée comme "réelle" : a) Image originale, b) Image reconstruite.*

Concernant les 799 images obscurcies que nous n'avons pas réussi à reconstruire comme images "réelles" avec notre attaque, plusieurs scénarios sont possibles :

— Les images sans aucune reconstruction prédite comme "réelle" (comme illustré sur la figure 7).
— Les images comportant au moins deux reconstructions prédites comme "réelles", dont la correcte, comme le montre l'illustration figure 8).
— Images avec au moins une reconstruction prédite comme "réelle", mais sans la correcte.

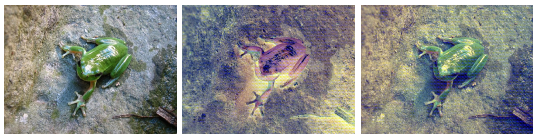


(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

FIGURE 7 – *Exemple avec aucune des huit images prédite comme "réelle" : a) Image originale, b) à i) Les huit images reconstruites possibles, toutes prédites comme "inverses".*



(a) (b) (c)

FIGURE 8 – *Exemple avec deux images prédites comme "réelles" : a) Image originale, b) Image reconstruite incorrecte prédite comme "réelle" (PSNR = 9,17 dB), c) Image reconstruite correcte prédite comme "réelle" (PSNR = 21,07dB).*

4 Conclusion

Dans cet article, nous avons proposé une attaque de la méthode d'obscuration par bit-flipping qui ne nécessite que la taille des blocs comme paramètre pour reconstruire l'image originale. Cette attaque consiste en une analyse bloc par bloc de chaque composante couleur, afin de reconstruire le motif binaire original. Les résultats montrent qu'il est possible de reconstruire parfaitement l'image originale, même si parfois nous obtenons une image avec des artefacts. En perspective, nous envisageons d'appliquer notre système de vote que sur les blocs homogènes, puis, de développer une détection automatique de la taille des blocs.

Remerciements

Ce travail a bénéficié d'une aide de l'état gérée par l'Agence National de la Recherche au titre de France 2030 portant la référence ANR-22-PECY-0011.

Références

[1] Maungmaung Aprilpyone et Hitoshi Kiya. Block-Wise Image Transformation With Secret Key for Adversarially Robust Defense. *IEEE Transactions on Information Forensics and Security*, 16 :2709–2723, 2021.

[2] R.L. Lagendijk, Zekeriya Erkin, et Mauro Barni. Encrypted signal processing for privacy protection : Conveying the utility of homomorphic encryption and multiparty computation. *IEEE Signal Processing Magazine*, 30(1) :82–105, 2013.

[3] W. Puech, Z. Erkin, M. Barni, S. Rane, et R. L. Lagendijk. Emerging cryptographic challenges in image and video processing. Dans *2012 19th IEEE International Conference on Image Processing*, pages 2629–2632, 2012.

[4] Steven Hill, Zhimin Zhou, Lawrence K. Saul, et Hovav Shacham. On the (In)effectiveness of Mosaicing and Blurring as Tools for Document Redaction. *Proceedings on Privacy Enhancing Technologies*, 2016 :403 – 417, 2016.

[5] Kenta Kurihara, Sayaka Shiota, et Hitoshi Kiya. An encryption-then-compression system for JPEG standard. Dans *2015 Picture Coding Symposium (PCS)*, pages 119–123, 2015.

[6] Tatsuya Chuman, Kenta Kurihara, et Hitoshi Kiya. On the security of block scrambling-based ETC systems against jigsaw puzzle solver attacks. Dans *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2157–2161, 2017.

[7] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, et Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3) :211–252, 2015.

[8] Alain Horé et Djemel Ziou. Image Quality Metrics : PSNR vs. SSIM. Dans *2010 20th International Conference on Pattern Recognition*, pages 2366–2369, 2010.

[9] Zhou Wang, A.C. Bovik, H.R. Sheikh, et E.P. Simoncelli. Image quality assessment : from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4) :600–612, 2004.

[10] Guanrong Chen, Yaobin Mao, et Charles K Chui. A symmetric image encryption scheme based on 3D chaotic cat maps. *Chaos, Solitons & Fractals*, 21(3) :749–761, 2004.

[11] Deok-Han Kim et Young-Gab Kim. A Method for De-Identification Analysis of Encrypted Video. Dans *2024 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT & NCON)*, pages 233–236, 2024.

Protection suffisante d'objet 3D par déformations géométriques

Khélian Larvet[1,2] William Puech[1] Jean-Pierre Pedeboy[2]

[1] LIRMM, Univ. Montpellier, CNRS, Montpellier, France
[2] STRATEGIES, Rungis, France

Résumé

Les récents progrès des applications multimédia ont considérablement contribué à la popularité des objets 3D dans divers domaines, avec une fréquence accrue de leur stockage et transmission en ligne. Il devient vital de les sécuriser afin de protéger leur intégrité, leur confidentialité et leur propriété intellectuelle. Dans cet article, nous présentons une nouvelle méthode de protection qui se distingue des méthodes de l'état de l'art basées sur de l'ajout de bruit. Notre approche consiste en l'application de déformations globales réversibles, contrôlées par une clé secrète afin de rendre un objet 3D inintelligible. Les résultats expérimentaux montrent que notre méthode peut s'avérer intéressante pour une protection suffisante bloquant toutes copies numériques ainsi que toutes impressions 3D des objets 3D protégés. Nous présentons des résultats expérimentaux et évaluons la robustesse de notre approche face à des attaques par des utilisateurs.

Mots clefs

Sécurité multimédia, déformations géométriques 3D, sécurité 3D, sécurité visuelle 3D.

1 Introduction

Le développement des objets 3D a révolutionné notre société, offrant une représentation plus intuitive de l'espace et des caractéristiques des objets qui nous entourent. Leur importance ne cesse de croître et leurs applications sont diverses, couvrant des domaines tels que les jeux, les films, l'architecture, la médecine, ainsi que divers secteurs économiques, sociaux et commerciaux [1, 2, 3]. Cette popularité entraîne une augmentation du stockage et de la transmission en ligne des objets 3D, qui contiennent souvent des informations sensibles. Ces informations, telles que des données personnelles ou des secrets commerciaux dans le secteur industriel, deviennent vulnérables aux attaques, et leur divulgation non autorisée entraîne des pertes économiques considérables [4, 5, 6]. Il devient donc crucial de préserver la confidentialité des objets 3D et d'en assurer la sécurité pendant leur stockage et leur transmission.
Il existe principalement deux grandes catégories de méthodes pour protéger les objets 3D, à savoir l'insertion de données cachées et le chiffrement. L'insertion de données cachées consiste à intégrer de manière invisible un message secret dans un objet 3D [7, 8, 9]. Ces approches peuvent être robustes, notamment pour la protection des droits d'auteur, ou fragiles, afin de garantir l'intégrité des objets 3D.

Les méthodes de chiffrement incluent les méthodes de chiffrement complet et sélectif. Les méthodes de chiffrement complet sécurisent un objet 3D en convertissant intégralement son contenu en un fichier binaire chiffré et inintelligible, garantissant ainsi une confidentialité totale. En revanche, les méthodes de chiffrement sélectif chiffrent uniquement des informations spécifiques des objets 3D [10]. Ces approches permettent de conserver le format original des objets et peuvent être compatibles avec une étape de compression, tout en offrant différents niveaux de sécurité visuelle, ajustés aux besoins spécifiques de l'utilisateur.

De nombreux travaux ont porté sur le chiffrement sélectif des objets 3D. Gschwantner *et al.* proposent un système de protection des objets 3D utilisant des maillages 3D progressifs pour appliquer un chiffrement différent à chaque partie de l'objet 3D [11]. Eluard *et al.* ont introduit le principe de chiffrement préservant la géométrie, avec une méthode de chiffrement basée sur des permutations pseudo-aléatoires des sommets [12]. Dans les travaux de Beugnon *et al.*, les coordonnées (X, Y, Z) des sommets des objets 3D sont converties en format binaire, puis certains bits sont sélectionnés pour être chiffrés, ce qui permet de présenter des niveaux variés de sécurité visuelle [13]. À partir de ce principe, en 2022, Jansen van Rensburg *et al.* ont proposé une méthode de chiffrement hiérarchique permettant de générer plusieurs versions déchiffrées d'un même objet 3D [14]. En 2024, Li *et al.* ont abordé l'adaptabilité limitée de [14] en proposant une méthode de chiffrement prenant en charge un déchiffrement hiérarchique adaptatif [15]. Enfin, Zhao *et al.* ont développé un chiffrement XOR avec des permutations basées sur un système chaotique [16].

Dans cet article, contrairement à l'état de l'art en chiffrement sélectif consistant à protéger les objets 3D principalement par ajout de bruit sur les sommets, nous proposons une nouvelle méthode de protection réversible des objets 3D, basée sur des déformations géométriques. Pour cela, pour chacun des axes X, Y et Z, nous appliquons un redimensionnement pseudo-aléatoire basé sur une clé secrète, sachant que chacun des trois facteurs de redimensionnement appliqués est contraint par les deux autres, cela afin de garantir une protection visuelle suffisante.

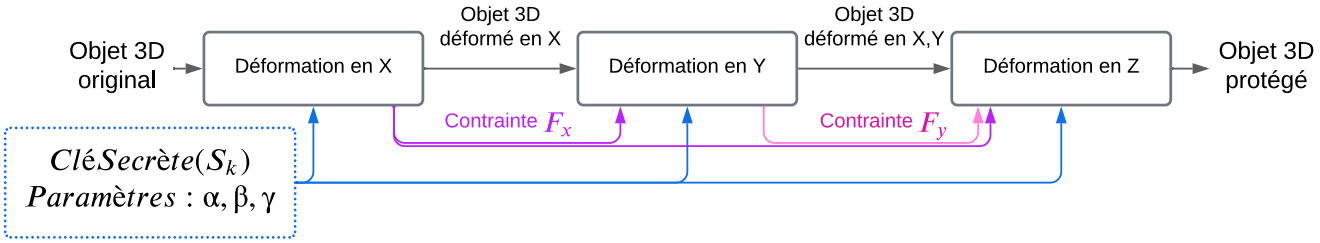


FIGURE 1 – *Aperçu général du processus de protection des objets 3D par déformations pour chacun des axes X, Y et Z.*

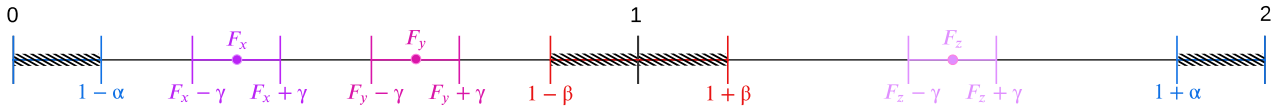


FIGURE 2 – *Illustration des contraintes sur la sélection des facteurs de redimensionnement.*

2 Méthode de protection d'objet 3D

La méthode que nous proposons repose sur des déformations géométriques obtenues par redimensionnement des axes X, Y et Z. Les facteurs de redimensionnement F_X, F_Y et F_Z sont générés à l'aide d'un générateur pseudo-aléatoire dont la graine est une clé secrète, garantissant ainsi la réversibilité du processus. Bien que la forme générale de l'objet reste accessible, cette méthode offre une protection suffisante de l'objet 3D en déformant suffisamment ses détails pour les protéger. La figure 1 illustre une vue d'ensemble de la méthode proposée. Afin de ne pas se retrouver dans la situation d'une transformation globale qui ne serait qu'un simple agrandissement ou au contraire qu'une simple réduction de la taille de l'objet ($F_X \approx F_Y \approx F_Z$), une fois que le premier redimensionnement suivant un axe est appliqué, des contraintes supplémentaires sont appliquées aux autres facteurs de redimensionnement.

Par conséquent, avec notre approche, les déformations par redimensionnement présentent des cas particuliers qui nécessitent une attention particulière, tels qu'un redimensionnement uniforme ou proche de l'identité. Pour éviter ces déformations qui conservent la plupart des détails de l'objet 3D original, nous proposons un algorithme qui contraint la génération pseudo-aléatoire des facteurs de redimensionnement afin de respecter quatre contraintes spécifiques. Comme illustré en figure 2, nous notons α, le paramètre d'amplitude de la zone aléatoire autour de l'identité, β le paramètre d'amplitude de la zone interdite autour de l'identité, et γ le paramètre d'amplitude de la zone interdite autour d'un facteur de redimensionnement F choisi.

Les contraintes illustrées en figure 2, sont les suivantes :

1. La portée maximale aléatoire autour de l'identité est : $[1-\alpha; 1+\alpha]$.
2. La zone interdite autour de l'identité est définie par : $[1-\beta; 1+\beta]$, avec $\beta < \alpha$.
3. La zone interdite autour de chaque facteur de redimensionnement F_i, avec $i \in \{X, Y, Z\}$, afin d'éviter des facteurs de redimensionnement trop similaires est : $[F_i-\gamma; F_i+\gamma]$, avec $\gamma < (\alpha-\beta)/2$.
4. Les facteurs de redimensionnement ne peuvent pas tous les trois, soit agrandir, soit réduire la forme de l'objet 3D. Cela implique qu'au moins un facteur de redimensionnement soit supérieur à l'identité et au moins un inférieur à l'identité.

En combinant toutes ces contraintes, les zones possibles pour la sélection des facteurs de redimensionnement sont :

$$[1-\alpha; 1-\beta] \cup [1+\beta; 1+\alpha] \setminus \bigcup_{F_i \in X,Y,Z} [F_i-\gamma; F_i+\gamma], \tag{1}$$

où F_i est l'ensemble des facteurs de redimensionnement déjà générés. Ces contraintes permettent d'assurer une diversité suffisante dans les redimensionnements tout en évitant les déformations uniformes ou proches de l'identité.

Algorithme 1 : Redimensionnement 3D

Input : Objet 3D original avec n sommets $v_i : O$;
Facteurs de redimensionnement : F_x, F_y et F_z ;
Output : Objet 3D protégé avec n sommets $v'_i : O'$
for chaque sommet v_i de O **do**
  $v'_{i,x} \leftarrow v_{i,x} \times F_x$
  $v'_{i,y} \leftarrow v_{i,y} \times F_y$
  $v'_{i,z} \leftarrow v_{i,z} \times F_z$
end for
return O'

Afin d'appliquer ces redimensionnements suivant chaque axe, nous appliquons l'algorithme de redimensionnement décrit dans l'algorithme 1. A partir d'un objet 3D original O et des facteurs de redimensionnement générés F_X, F_Y et F_Z pseudo-aléatoirement à partir d'une clé secrète, et respectant les contraintes résumés en équation 1, tous les sommets de l'objet original O sont déformés afin de générer un objet 3D protégé O'.

L'objet 3D original peut alors être reconstruit de manière totalement réversible à partir de l'objet 3D protégé O' en appliquant les déformations géométriques inverse. En effet, pour la phase de décodage, à partir de la clé secrète il est possible de générer à nouveau les trois facteurs de redimensionnement F_X, F_Y et F_Z et d'appliquer sur l'objet 3D protégé O' les transformations inverses $\frac{1}{F_X}$, $\frac{1}{F_Y}$ et $\frac{1}{F_Z}$ respectivement sur chacun des trois axes.

3 Résultats

Dans cette section nous présentons et analysons des résultats expérimentaux obtenus avec la méthode de protection d'objets 3D que nous proposons. En section 3.1, nous présentons des résultats sur un ensemble d'objets 3D et mesurons leurs déformations sur la base d'un calcul RMSE. En section 3.2, nous présentons une expérience proposée à des utilisateurs afin d'attaquer par déformation des objets 3D protégés par notre méthode.

3.1 Protection par déformation

Nous utilisons le format 3D PLY avec une précision flottante de 32 bits selon la norme IEEE 754 pour définir les sommets des objets 3D. Les objets 3D peuvent être déformés avec des facteurs de déformation allant jusqu'à 1×10^{-38}. Nous avons mené une expérience en sélectionnant 10 objets 3D[1,2] et en leur appliquant 5 déformations aléatoires, avec les paramètres : $\alpha = 0.9$, $\beta = 0.25$ et $\gamma = 0.25$. La première ligne du tableau 1 illustre 4 de ces 10 objets.

Objet 3D original O			
			
Objet 3D protégé O'			
			
{1.52, 0.41, 0.13}	{1.26, 0.32, 1.76}	{0.18, 0.74, 1.88}	{1.57, 1.85, 0.17}
RMSE entre l'objet 3D original O et l'objet 3D déformé O'			
1.274	1.003	1.003	0.883
RMSE entre l'objet 3D original O et l'objet 3D reconstruit O''			
7.45×10^{-8}	7.73×10^{-8}	6.11×10^{-8}	4.26×10^{-8}

TABLEAU 1 – *Protection d'objets 3D par déformation avec $\alpha = 0.9$, $\beta = 0.25$, $\gamma = 0.25$, sur 4 objets originaux[1,2] avec différents facteurs de redimensionnement {F_X, F_Y, F_Z}.*

La deuxième ligne du tableau 1 présente ces 4 objets après protection par déformation. Nous observons que, bien que les objets soient déformés, il est encore possible de reconnaître la forme de ces objets, la protection est dite "suffisante". La valeur de la racine de l'erreur quadratique moyenne (RMSE) entre un objet 3D original et ce même objet 3D protégé est en moyenne égale à 1. Après reconstruction, en moyenne le RMSE entre l'objet reconstruit et l'objet original, est de l'ordre de 10^{-8}, ce qui peut être considéré comme réversible.

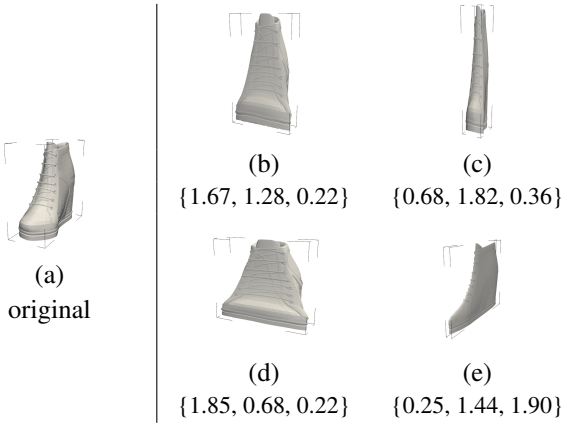


FIGURE 3 – *4 déformations {F_X,F_Y,F_Z} de l'objet 3D "Chaussure".*

La notion de protection "suffisante" s'appuie sur les travaux de Pommer *et al.* [1] et de Beugnon *et al.* [2, 3] qui définissent la sécurité visuelle d'un objet 3D comme l'accessibilité de cet objet par le système visuel humain (SVH). Trois niveaux de sécurité visuelle ont ainsi été établis pour les objets 3D. Au niveau confidentiel, ni la forme ni le contenu de l'objet 3D ne sont accessibles au SVH. Au niveau suffisant, seule la forme de l'objet 3D est visible, tandis que le contenu reste caché. Enfin, au niveau transparent, la forme et le contenu de l'objet 3D sont accessibles, mais les détails de haute qualité restent protégés et ne peuvent pas être utilisés.
Dans notre cas, notre méthode atteint un niveau de sécurité visuelle suffisant où la forme de l'objet 3D reste accessible, tout en étant suffisamment déformés pour empêcher une utilisation précise, comme par exemple une impression 3D.

3.2 Attaques des objets 3D protégés

Nous avons sollicité 22 participants avec le protocole suivant : chaque utilisateur se voit présenter une déformation choisie aléatoirement pour chacun des 10 objets 3D[1,2]. Leur objectif est de manipuler l'objet 3D protégé afin d'essayer de retrouver l'objet 3D original à un factuer d'échelle près.
Pour ce faire, nous avons mis en place une interface web permettant de charger un objet 3D et de le manipuler dans l'espace avec des fonctions de déplacement et de rotation. Des curseurs ont été ajoutés pour modifier les facteurs de déformation en chaque axe X, Y et Z de l'objet 3D, comme illustré figure 4.

1. https://free3d.io
2. https://www.artec3d.com/3d-models

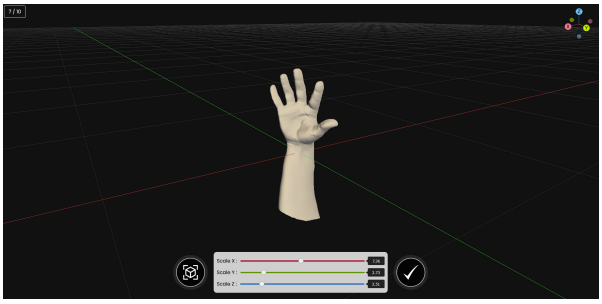


FIGURE 4 – *Interface de l'application.*

Pour mesurer la correspondance d'un objet 3D reconstruit par un utilisateur par rapport à sa version originale et protégée, nous utilisons les ratios des dimensions. Pour l'objet original O, l'objet protégé O' et l'objet reconstruit par un utilisateur O'', nous notons les ratios R de la manière suivante :

$$R_{xy} = \frac{x}{y}, \quad R_{xz} = \frac{x}{z}, \quad R_{yz} = \frac{y}{z}. \tag{2}$$

Pour chaque objet, les ratios sont définis comme suit :

$$R^{O}_{xy} = \frac{O_x}{O_y}, \quad R^{O}_{xz} = \frac{O_x}{O_z}, \quad R^{O}_{yz} = \frac{O_y}{O_z}, \tag{3}$$

$$R^{O'}_{xy} = \frac{O'_x}{O'_y}, \quad R^{O'}_{xz} = \frac{O'_x}{O'_z}, \quad R^{O'}_{yz} = \frac{O'_y}{O'_z}, \tag{4}$$

$$R^{O''}_{xy} = \frac{O''_x}{O''_y}, \quad R^{O''}_{xz} = \frac{O''_x}{O''_z}, \quad R^{O''}_{yz} = \frac{O''_y}{O''_z}. \tag{5}$$

Prenons l'exemple du ratio R_{xy}. Lorsque $R^{O''}_{xy} = R^{O}_{xy}$, nous souhaitons obtenir un score de 1.0, indiquant une correspondance parfaite avec l'objet original, et lorsque $R^{O''}_{xy} = R^{O'}_{xy}$, un score de 0.0, indiquant une correspondance avec l'objet protégé. En utilisant ces conditions, nous définissons la fonction linéaire suivante :

$$S_{xy} = \alpha_{xy} \times R^{O''}_{xy} + \beta_{xy}, \tag{6}$$

avec $R^{O''}_{xy} \in [R^{O}_{xy}; R^{O'}_{xy}]$. En résolvant pour α_{xy} et β_{xy}, nous obtenons :

$$\alpha_{xy} = \frac{1.0}{R^{O}_{xy} - R^{O'}_{xy}}, \quad \beta_{xy} = -\frac{R^{O'}_{xy}}{R^{O}_{xy} - R^{O'}_{xy}}. \tag{7}$$

En substituant ces valeurs, nous obtenons le score :

$$S_{xy} = \frac{R^{O''}_{xy} - R^{O'}_{xy}}{R^{O}_{xy} - R^{O'}_{xy}}. \tag{8}$$

Les scores pour les autres ratios R_{xz} et R_{yz} sont calculés de la même manière :

$$S_{xz} = \frac{R^{O''}_{xz} - R^{O'}_{xz}}{R^{O}_{xz} - R^{O'}_{xz}}, \quad S_{yz} = \frac{R^{O''}_{yz} - R^{O'}_{yz}}{R^{O}_{yz} - R^{O'}_{yz}}. \tag{9}$$

La correspondance globale est la moyenne des trois scores :

$$S = \frac{1}{3}\left(S_{xy} + S_{xz} + S_{yz}\right). \tag{10}$$

Objet 3D	**Score μ (%)**	**Score σ (%)**
Main	84.04 %	14.27 %
Voiture	92.92 %	4.13 %
Sculpture	91.18 %	6.54 %
Chaussure	67.31 %	30.85 %
Chaise	75.19 %	14.12 %
Chandelier	85.06 %	9.63 %
Dague	80.06 %	23.06 %
Clé	75.71 %	20.58 %
Vis	80.03 %	15.61 %
Fleur	86.52 %	12.80 %
Total	81.81 %	18.49 %

TABLEAU 2 – *Score moyen de reconstruction par 22 utilisateurs et écart type des objets 3D.*

Lors de la reconstruction d'objets 3D, nous avons observé que les utilisateurs tendent à identifier et à reconstituer les formes probablement circulaires ou sphériques de l'objet original. Nos résultats, illustrés dans le tableau 2, montrent qu'en moyenne sur 22 utilisateurs, les objets 3D ont été reconstruits avec un score moyen de 81,8 % et un écart type de 18,4 %. Notons que, malgré ces scores relativement élevés, les objets attaqués par les utilisateur ne permettent pas d'effectuer un copie numérique ou une impression 3D de qualités. Les objets 3D restent suffisamment protégés par notre méthode.

4 Conclusion

Dans cet article, nous avons introduit une méthode innovante de protection des objets 3D basée sur des déformations géométriques. Contrairement aux techniques traditionnelles d'ajout de bruit, notre approche utilise des redimensionnements contrôlés par un générateur pseudo-aléatoire, garantissant la réversibilité. Les résultats montrent qu'un attaquant peut reconstruire l'objet original avec une précision moyenne de 81,81%, ce qui montre que notre méthode est suffisante pour protéger les objets 3D contre une copie ou une impression 3D illégale par exemple. Cette approche ouvre donc de nouvelles perspectives pour la protection 3D via des déformations géométriques.

À l'avenir, nous avons pour objectif d'explorer de nouvelles déformations géométriques afin d'évaluer leur potentiel de confidentialité. Nous envisageons également d'appliquer des déformations géométriques locales, car nos déformations ont jusqu'à présent été globales. Cette approche pourrait créer de nouveaux effets et élargir les possibilités de protection des objets 3D.

Références

[1] Andreas Pommer et Andreas Uhl. Application scenarios for selective encryption of visual data. Dans *Multimedia and Security Workshop, ACM Multimedia*, 2002.

[2] Sebastien Beugnon, William Puech, et Jean-Pierre Pedeboy. From visual confidentiality to transparent format-compliant selective encryption of 3D objects. Dans *2018 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6. IEEE, 2018.

[3] Sébastien Beugnon, Bianca Jansen van Rensburg, Naima Amalou, William Puech, et Jean-Pierre Pedeboy. A 3D Visual Security (3DVS) score to measure the visual security level of selectively encrypted 3D objects. *Signal Processing : Image Communication*, 108 :116832, 2022.

[4] Marianna Lezzi, Mariangela Lazoi, et Angelo Corallo. Cybersecurity for Industry 4.0 in the current literature : A reference framework. *Computers in Industry*, 103 :97–110, 2018.

[5] Siva Chaitanya Chaduvula, Adam Dachowicz, Mikhail J Atallah, et Jitesh H Panchal. Security in cyber-enabled design and manufacturing : A survey. *Journal of Computing and Information Science in Engineering*, 18(4) :040802, 2018.

[6] Dazhong Wu, David W Rosen, Lihui Wang, et Dirk Schaefer. Cloud-based design and manufacturing : A new paradigm in digital manufacturing and design innovation. *Computer-aided design*, 59 :1–14, 2015.

[7] Kai Wang, Guillaume Lavoué, Florence Denis, et Atilla Baskurt. A comprehensive survey on three-dimensional mesh watermarking. *IEEE Transactions on Multimedia*, 10(8) :1513–1527, 2008.

[8] N Medimegh, S Belaid, et N Werghi. A survey of the 3D triangular mesh watermarking techniques. *International Journal of Multimedia*, 1(1) :33–39, 2015.

[9] Chang-Min Chou et Din-Chang Tseng. Technologies for 3d model watermarking : A survey. *International Journal of Computer Science and Network Security*, 7 :328–334, 2007.

[10] Ayoub Massoudi, Frédéric Lefebvre, Christophe De Vleeschouwer, Benoit Macq, et J-J Quisquater. Overview on selective encryption of image and video : challenges and perspectives. *Eurasip Journal on information security*, 2008(1) :179290, 2008.

[11] Michael Gschwandtner et Andreas Uhl. Protected progressive meshes. Dans *International Symposium on Visual Computing*, pages 35–48. Springer, 2009.

[12] Marc Éluard, Yves Maetz, Gwenaël Doërr, R Technicolor, et D France. Geometry-preserving encryption for 3D meshes. *Actes de COmpression et REprésentation des Signaux Audiovisuels*, pages 7–12, 2013.

[13] Sebastien Beugnon, William Puech, et Jean-Pierre Pedeboy. From visual confidentiality to transparent format-compliant selective encryption of 3D objects. Dans *2018 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6. IEEE, 2018.

[14] Bianca Jansen van Rensburg, William Puech, et Jean-Pierre Pedeboy. A format compliant encryption method for 3D objects allowing hierarchical decryption. *IEEE Transactions on Multimedia*, 25 :7196–7207, 2022.

[15] Shimin Li, Ruoyu Zhao, Qingxiao Guan, Junxin Chen, et Yushu Zhang. A 3D model encryption method supporting adaptive visual effects after decryption. *Advanced Engineering Informatics*, 59 :102319, 2024.

[16] Ruoyu Zhao, Yushu Zhang, Shimin Li, Wenying Wen, Shuang Yi, et Rushi Lan. 3D mesh encryption with differentiated visual effect and high efficiency based on chaotic system. *Expert Systems with Applications*, 238 :122140, 2024.

3DOF+Quantization: 3DGS quantization for large scenes with limited Degrees of Freedom

Matthieu GENDRIN
Stéphane PATEUX
Théo LADUNE
Orange Innovation

July 12, 2024

Abstract

3DGS [Kerbl et al., 2023] is a major breakthrough in 3D scene reconstruction. With a number of views of a given object or scene, the algorithm trains a model composed of 3D gaussians, which enables the production of novel views from arbitrary points of view. This freedom of movement is referred to as 6DoF for 6 degrees of freedom: a view is produced for any position (3 degrees), orientation of camera (3 other degrees). On large scenes, though, the input views are acquired from a limited zone in space, and the reconstruction is valuable for novel views from the same zone, even if the scene itself is almost unlimited in size. We refer to this particular case as 3DoF+, meaning that the 3 degrees of freedom of camera position are limited to small offsets around the central position. Considering the problem of coordinate quantization, the impact of position error on the projection error in pixels is studied. It is shown that the projection error is proportional to the squared inverse distance of the point being projected. Consequently, a new quantization scheme based on spherical coordinates is proposed. Rate-distortion performance of the proposed method are illustrated on the well-known Garden scene.

1 Introduction

3DGS [Kerbl et al., 2023] has opened new possibilities in terms of novel view synthesis of 3D scenes. The quality and training performance are such that a major part of the 3D research community has switched to this model. With this success comes the need to compress such models, which is achieved in several ways in the literature. Papantonakis *et al.* [Papantonakis et al., 2024] proposes to optimize the number of gaussians and the color coefficients, and to use a codebook-based quantization method. Scaffold-GS ([Lu et al., 2024]) introduces a structured description of the model to obtain additional compression performance. HAC ([Chen et al., 2024]) builds upon Scaffold-GS adding

entropy minimization for further rate savings. These previous work provides compelling rate-distortion, without any hypothesis on the degrees of freedom of the camera. As a complement, this paper analyses how the 3DoF+ hypothesis can be leveraged to perform a more accurate bit allocation to the spatial coordinates, giving more precision to the gaussians near the cameras, to the expense of gaussians located further away. Note that this work is complementary to the existing methods.

2 Preliminaries

3D Gaussian Splatting (3DGS) [Kerbl et al., 2023] models a 3D scene with 3D gaussians, and renders viewpoints through a differentiable splatting and tile-based rasterization. Each Gaussian is defined by a 3D covariance matrix $\mathbf{\Sigma} \in \mathbb{R}^{3\times3}$ and location (mean) $\boldsymbol{\mu} \in \mathbb{R}^3$, where $\mathbf{x} \in \mathbb{R}^3$ is a random 3D point, and $\mathbf{\Sigma}$ is defined by a diagonal matrix $\mathbf{S} \in \mathbb{R}^{3\times3}$ representing scaling and rotation matrix $\boldsymbol{R_g} \in \mathbb{R}^{3\times3}$ to guarantee its positive semi-definite characteristics, such that $\mathbf{\Sigma} = \boldsymbol{R_g S S}^\top \boldsymbol{R_g}^\top$.

$$G(\mathbf{x}) = exp(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})) \tag{1}$$

To render an image from a random viewpoint, 3D Gaussians are first splatted to 2D, and render the pixel value $\mathbf{C} \in \mathbb{R}^3$ using α-composed blending, where $\alpha \in \mathbb{R}$ measures the contribution to this pixel of each Gaussian after 2D projection, $\mathbf{c} \in \mathbb{R}^3$ is view-dependent color modeled by Spherical Harmonic (SH) coefficients, and N is the number of sorted Gaussians contributing to the rendering. The 3DGS rendering is illustrated in figure 1.

$$\mathbf{C} = \sum_{n\in N} c_n \alpha_n \prod_{j=1}^{n-1}(1-\alpha_j) \tag{2}$$

Note that the gaussian calculation exposed in eq. (1) is actually done in 2D after projection of the gaussians center and covariance matrix on the screen image. The projection of the center is done with the classical formulation given in eq. (3).

3 Position dependent quantization methodology

Perspective projection For the sake of simplicity, let us first consider a toy system with a single camera. Any point with coordinates (x, y, z) in the camera referential is projected to the image plan as (u, v):

$$\begin{aligned} u &= f\frac{x}{z} \\ v &= f\frac{y}{z} \end{aligned} \tag{3}$$

Figure 1: Gaussian render example. The rendering has been modified on the left part to highlight the gaussians, while the right part is the standard 3DGS rendering.

Under a high-rate hypothesis, quantizing the point coordinates (x, y, z) with a scalar, unitary quantizer can be modeled as adding an independent noise δ to the coordinates.

1. the impact on (u, v) of the noise δ on x, y is proportional to $\frac{1}{z}$
2. the impact on (u, v) of the noise δ on z is proportional to $\frac{1}{z^2}$

Of course, we don't want to encode the model for one precise camera, but This gives the intuition of how the 3DoF+ assumption can be leveraged: the z coordinate of the local referential is to be quantized differently than the (x, y). For example, we could use larger quantization steps for x, y, z when z is large. And even larger step for z than for x, y. [1] To generalize from this toy example, we note that z is approximately the distance between the camera and the gaussian, and that the cameras are all in the same area in 3DoF+. As such, the reasoning holds for all cameras.
Figure 3 illustrates a typical 3DoF+ scene, with the camera positions in a small spatial zone compared to the distance to the closest points.

[1] Note that this encourages to use spherical coordinates here for performing positional quantization. Simple scalar quantization on spherical coordinates would then lead to well adapted vector quantization of positions.

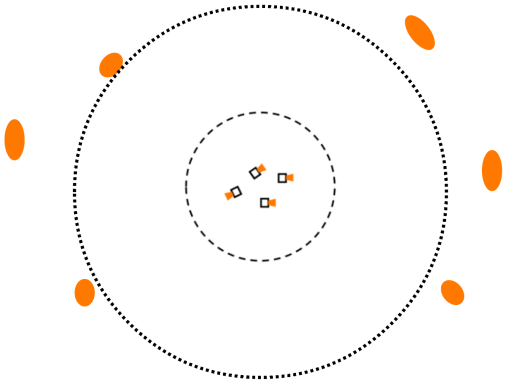


Figure 2: 3DoF+ scene.
Inner circle, radius R_i, is the limit of the possible camera poses.
Outer circle, radius R, defines the minimum gaussian distance to the center.

Spherical coordinates To make the distance between a gaussian and a camera more explicit, the gaussian coordinates are now expressed in spherical coordinates (ρ, θ, ϕ). The origin is set in the area of the cameras centers. With this referential and the 3DoF+ assumption, the ρ coordinate of any point approximates the distance between this point and any camera.
Now on, we'll consider a point of spherical coordinates $\rho\mathbf{d}$ is the unit direction vector defined by:

$$\mathbf{d} = (\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta)^T \tag{4}$$

Considering that a camera can point to any direction, we will work on the 360° projection. This projection associates each point with its projection on the sphere of radius f. We will now work in a referential with the camera as center, keeping the orientation of the world referential. This referential will be referred to as the local referential.

$$\mathbf{P} = \mathbf{P}_0 + \rho\mathbf{d} \tag{5}$$

Where $\mathbf{P}_0$ is the world origin in the local referential.
The projection $\mathbf{p}$ of $\mathbf{P}$ on the sphere is:

$$\mathbf{p} = \frac{\mathbf{P}}{||\mathbf{P}||} \tag{6}$$

And the derivation of this projection gives (cf details in appendix):

$$\begin{aligned}
\frac{\partial \mathbf{p}}{\partial \theta} =& f\frac{\rho}{||\mathbf{P}||}(\cos\theta\cos\phi, \cos\theta\sin\phi, -\sin\theta)^T + O(\epsilon) \\
\frac{\partial \mathbf{p}}{\partial \phi} =& f\frac{\rho\sin\theta}{||\mathbf{P}||}(-\sin\phi, \cos\phi, 0)^T + O(\epsilon) \\
\frac{\partial \mathbf{p}}{\partial \rho} =& -f\frac{1}{\rho^2}((\mathbf{P}_0^T\mathbf{d})\mathbf{d} - \mathbf{P}_0) + o(\epsilon^2) \\
\epsilon =& \frac{||\mathbf{P}_0||}{||\mathbf{P}||}
\end{aligned} \tag{7}$$

This means that uniform quantization is relevant for θ and ϕ, since their impact is bounded by finite values close to 1. The impact of ρ quantization on the other hand depends on the value of ρ itself. Thus quantizing the coordinate uniformly would be suboptimal.
It is proposed to parameterize ρ as: $\rho = \frac{1}{t}$, yielding:

$$\frac{\partial \mathbf{p}}{\partial t} = f((\mathbf{P}_0^T\mathbf{d})\mathbf{d} - \mathbf{P}_0) + o(\epsilon^2) \tag{8}$$

Which is a good candidate for uniform quantization, since it does not depend on the position of $\mathbf{P}$.

Center vs periphery Since the calculations assume the gaussians are far away, compared to the distance between the cameras. In most cases, this hypothesis is not verified for all gaussians, and quantization scheme can not be used for the whole scene. A center zone is thus defined, and uses a uniform quantization instead of the proposed model. Outside of the center zone lies the peripheral zone where the proposed schema is used.
In short:

1. In the center, x, y, z are quantized uniformly
2. In the periphery, $\theta, \phi, 1/\rho$ are quantized uniformly

The center is defined as the points matching: $\rho < R$ with R being roughly twice the distance of the cameras from the center of the scene.
[2]

4 Experiments

Test conditions We tested the proposed quantization a reconstruction of the Garden scene, from [Barron et al., 2022], after 30k iterations on the original

[2]The simple fact to divide the scene in center vs periphery is already a way to quantize the coordinates of the central part with a finer step, without changing the quantization step of the rest of the scene. This optimization can be used on x, y, z coordinate, it is tested in the ablation study below.

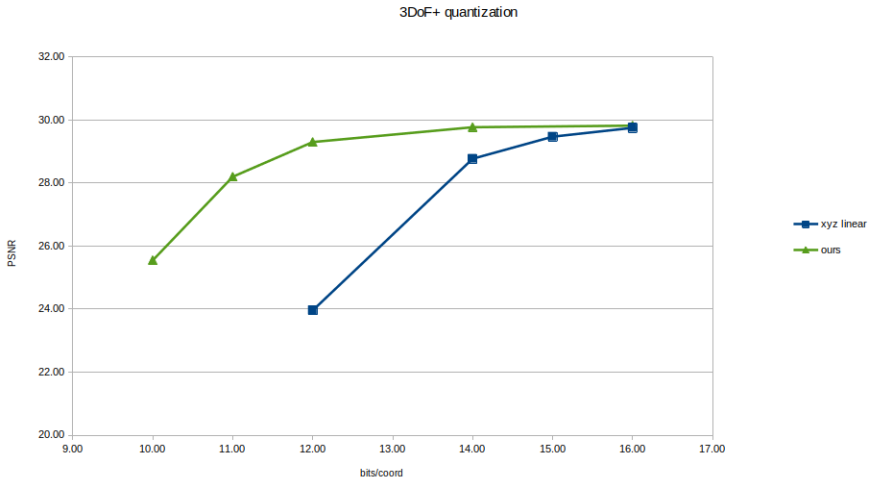


Figure 3: PSNR versus bits/coord

code of 3DGS. The PSNR is evaluated on the training views, with the configurations:

1. uniform: each x,y,z coordinate is quantized independently, the step depending on the extent of the scene
2. ours: we use $R = 1.5$ times the radius of the training cameras positions (bigger values of R did not improve the quality of the novel views)

Table 1: Results on garden scene, mip-nerf dataset

bits / coord	**uniform**	**ours**
16	29.76	29.82
14	28.77	29.77
12	23.96	29.30

The measures listed in table 1 show a clear improvement in terms of PSNR when lowering the number of bits per coordinate.

Discussions This document proposes an analysis of the impact of quantization noise in terms of projection on the screen plan. Another impact of quantization noise comes from the use of the position to define in which order the gaussians are drawn. The analysis of this aspect is left for future work. The split of the model points in center vs periphery is a new information, which should be added to the information to be coded. One may argue that the order of the points in the file is an easy way to encode this information. If the center

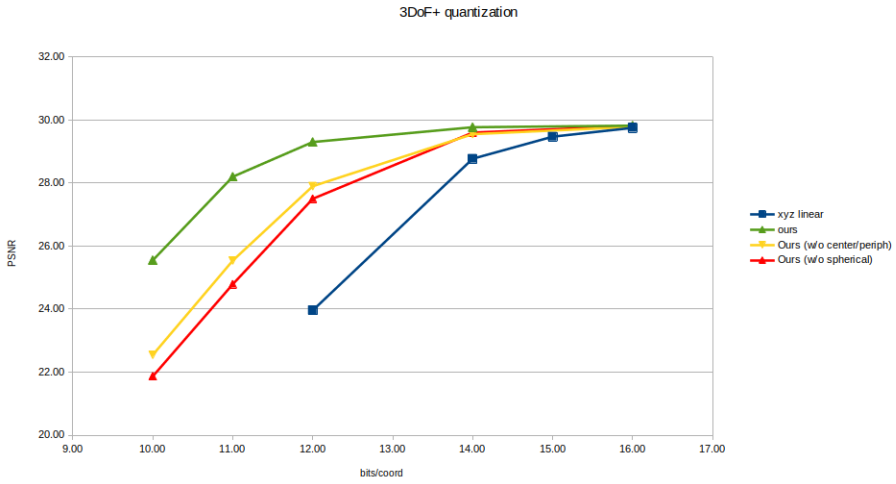


Figure 4: Ablations.

points are transfered first, only the index of the first periphery point has to be provided to differentiate the two populations. A more basic way to transfer this information would be to add one bit per gaussian, which costs 0.33 bit per coordinate. With this extra cost, the proposed solution keeps better than uniform quantization.

5 Ablation study

The proposed quantization scheme includes the use of spherical coordinates and the split of the scene in two parts: center vs periphery. Spherical coordinates, including the inversion of ρ, bring some value by giving more precision to points close to the center. The added value of this part is illustrated by *"w/o center/periphery"*, where $1/\rho, \theta, \phi$ are quantize independently of the gaussian position. The split in center vs periphery is another way to enable finer precision on the center, without sacrifying the precision of the periphery. This part is illustrated by *"w/o spherical"*, where x, y, z are quantized, with two sets of bounds: $[-R, R]$ for the center gaussians, and the extent of the whole scene for the periphery. The figure 4 shows the PSNR reached at different bits/coord values.

6 Conclusions

This article proposes a simple parameterization of spatial coordinates, which minimizes the projection error due to the positions quantization, compared to a standard uniform quantization. This straightforward technique does not in-

terfere with 3DGS training and is compatible with many other compression algorithms. Though exposed in a simple 3DoF+ context, this technique can be adapted to many 3D scenes to avoid storing and transmitting more information than needed for zones that will be viewed far away at render time.

Appendix A Partial derivations

This section details the partial derivatives exposed in the document.
Unit vector of the spherical coordinates:

$$\begin{aligned}
\mathbf{d} =& (\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta)^T \\
\frac{\partial \mathbf{d}}{\partial \theta} =& (\cos\theta\cos\phi, \cos\theta\sin\phi, -\sin\theta)^T \\
\mathbf{d}^T \frac{\partial \mathbf{d}}{\partial \theta} =& 0 \\
\frac{\partial \mathbf{d}}{\partial \phi} =& (-\sin\theta\sin\phi, \sin\theta\cos\phi, 0)^T \\
\mathbf{d}^T \frac{\partial \mathbf{d}}{\partial \phi} =& 0
\end{aligned} \tag{9}$$

Point in the camera referential:

$$\begin{aligned}
\mathbf{P} =& \mathbf{P}_0 + \rho\mathbf{d} \\
\frac{\partial \mathbf{P}}{\partial \rho} =& \mathbf{d} \\
\frac{\partial \mathbf{P}}{\partial \mathbf{d}} =& \rho\mathbf{I}
\end{aligned} \tag{10}$$

With $\mathbf{I}$ the 3×3 identity matrix.
Projection of the point on the unit sphere:

$$\begin{aligned}
||\mathbf{P}|| =& \sqrt{p_x^2 + p_y^2 + p_z^2} \\
\frac{\partial ||\mathbf{P}||}{\partial \mathbf{P}} =& (p_x/\sqrt{p_x^2 + p_y^2 + p_z^2}, p_y/\sqrt{p_x^2 + p_y^2 + p_z^2}, p_z/\sqrt{p_x^2 + p_y^2 + p_z^2}) \\
=& \frac{1}{||\mathbf{P}||}\mathbf{P}^T \\
\mathbf{p} =& \frac{1}{||\mathbf{P}||}\mathbf{P} \\
\frac{\partial \mathbf{p}}{\partial \mathbf{P}} =& \frac{1}{||\mathbf{P}||}\mathbf{I} + \mathbf{P}(\frac{-1}{||\mathbf{P}||^2}\frac{\partial ||\mathbf{P}||}{\partial \mathbf{P}}) \\
=& \frac{1}{||\mathbf{P}||}\mathbf{I} + \mathbf{P}(\frac{-1}{||\mathbf{P}||^2}\frac{1}{||\mathbf{P}||}\mathbf{P}^T) \\
=& \frac{1}{||\mathbf{P}||}\mathbf{I} - \frac{1}{||\mathbf{P}||^3}\mathbf{P}\mathbf{P}^T
\end{aligned} \tag{11}$$

Derivation of the projection by angles:

$$
\begin{aligned}
\frac{\partial \mathbf{p}}{\partial \theta} =& (\frac{1}{||\mathbf{P}||}\mathbf{I} - \frac{1}{||\mathbf{P}||^3}\mathbf{P}\mathbf{P}^T)\frac{\partial \mathbf{P}}{\partial \mathbf{d}}\frac{\partial \mathbf{d}}{\partial \theta} \\
=& (\frac{1}{||\mathbf{P}||}\mathbf{I} - \frac{1}{||\mathbf{P}||^3}\mathbf{P}\mathbf{P}^T)\rho\frac{\partial \mathbf{d}}{\partial \theta} \\
=& \frac{1}{||\mathbf{P}||}\rho\frac{\partial \mathbf{d}}{\partial \theta} - \frac{1}{||\mathbf{P}||^3}\mathbf{P}\mathbf{P}^T(\rho\frac{\partial \mathbf{d}}{\partial \theta}) \\
=& \frac{\rho}{||\mathbf{P}||}(\frac{\partial \mathbf{d}}{\partial \theta} - \frac{1}{||\mathbf{P}||}(\mathbf{P}_0^T\frac{\partial \mathbf{d}}{\partial \theta})\mathbf{p}) \\
=& \frac{\rho}{||\mathbf{P}||}\frac{\partial \mathbf{d}}{\partial \theta} + O(\epsilon) \\
=& \frac{\rho}{||\mathbf{P}||}(\cos\theta\cos\phi, \cos\theta\sin\phi, -\sin\theta) + O(\epsilon) \\
\frac{\partial \mathbf{p}}{\partial \phi} =& (\frac{1}{||\mathbf{P}||}\mathbf{I} - \frac{1}{||\mathbf{P}||^3}\mathbf{P}\mathbf{P}^T)\frac{\partial \mathbf{P}}{\partial \mathbf{d}}\frac{\partial \mathbf{d}}{\partial \phi} \\
=& \frac{1}{||\mathbf{P}||}\rho\frac{\partial \mathbf{d}}{\partial \phi} - \frac{1}{||\mathbf{P}||^3}\mathbf{P}\mathbf{P}^T(\rho\frac{\partial \mathbf{d}}{\partial \phi}) \\
=& \frac{\rho}{||\mathbf{P}||}(\frac{\partial \mathbf{d}}{\partial \phi} - \frac{1}{||\mathbf{P}||}(\mathbf{P}_0^T\frac{\partial \mathbf{d}}{\partial \phi})\mathbf{p}) \\
=& \frac{\rho}{||\mathbf{P}||}\frac{\partial \mathbf{d}}{\partial \phi} + O(\epsilon) \\
=& \frac{\rho\sin\theta}{||\mathbf{P}||}(-\sin\phi, \cos\phi, 0) + O(\epsilon)
\end{aligned} \tag{12}
$$

Derivation of the projection by ρ:

$$
\begin{aligned}
\frac{\partial \mathbf{p}}{\partial \rho} =& (\frac{1}{||\mathbf{P}||}\mathbf{I} - \frac{1}{||\mathbf{P}||^3}\mathbf{P}\mathbf{P}^T)\frac{\partial \mathbf{P}}{\partial \rho} \\
=& \frac{1}{||\mathbf{P}||}\mathbf{d} - \frac{1}{||\mathbf{P}||^3}\mathbf{P}\mathbf{P}^T\mathbf{d} \\
=& \frac{1}{||\mathbf{P}||}\mathbf{d} - \frac{1}{||\mathbf{P}||^3}(\mathbf{P}_0 + \rho\mathbf{d})(\rho + \mathbf{P}_0^T\mathbf{d}) \\
=& \frac{1}{||\mathbf{P}||^3}(||\mathbf{P}||^2\mathbf{d} - \rho\mathbf{P}_0 - (\mathbf{P}_0^T\mathbf{d})\mathbf{P}_0 - \rho^2\mathbf{d} - \rho(\mathbf{P}_0^T\mathbf{d})\mathbf{d}) \\
=& \frac{1}{||\mathbf{P}||^3}((\rho^2 + 2\rho\mathbf{P}_0^T\mathbf{d} + ||\mathbf{P}_0||^2)\mathbf{d} - \rho\mathbf{P}_0 - (\mathbf{P}_0^T\mathbf{d})\mathbf{P}_0 - \rho^2\mathbf{d} - \rho(\mathbf{P}_0^T\mathbf{d})\mathbf{d}) \\
=& \frac{1}{||\mathbf{P}||^3}((\rho\mathbf{P}_0^T\mathbf{d} + ||\mathbf{P}_0||^2)\mathbf{d} - (\rho + \mathbf{P}_0^T\mathbf{d})\mathbf{P}_0) \\
=& \frac{\rho}{||\mathbf{P}||}\frac{1}{||\mathbf{P}||^2}((\mathbf{P}_0^T\mathbf{d})\mathbf{d} - \mathbf{P}_0) + \frac{1}{||\mathbf{P}||}(\frac{||\mathbf{P}_0||^2}{||\mathbf{P}||^2}\mathbf{d} - \frac{\mathbf{P}_0^T\mathbf{d}}{||\mathbf{P}||}\frac{\mathbf{P}_0}{||\mathbf{P}||}) \\
=& \frac{\rho}{||\mathbf{P}||}\frac{1}{||\mathbf{P}||^2}((\mathbf{P}_0^T\mathbf{d})\mathbf{d} - \mathbf{P}_0) + O(\epsilon^2) \\
=& \frac{1}{||\mathbf{P}||^2}((\mathbf{P}_0^T\mathbf{d})\mathbf{d} - \mathbf{P}_0) + O(\epsilon^2)
\end{aligned} \tag{13}
$$

Please note that in these equations, $\mathbf{p}$ is the projection on the unit sphere, as in the paper it's the projection on the sphere of radius f.

References

[Barron et al., 2022] Barron, J. T., Mildenhall, B., Verbin, D., Srinivasan, P. P., and Hedman, P. (2022). Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5470–5479.

[Chen et al., 2024] Chen, Y., Wu, Q., Cai, J., Harandi, M., and Lin, W. (2024). Hac: Hash-grid assisted context for 3d gaussian splatting compression. *arXiv preprint arXiv:2403.14530*.

[Kerbl et al., 2023] Kerbl, B., Kopanas, G., Leimkühler, T., and Drettakis, G. (2023). 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1.

[Lu et al., 2024] Lu, T., Yu, M., Xu, L., Xiangli, Y., Wang, L., Lin, D., and Dai, B. (2024). Scaffold-gs: Structured 3d gaussians for view-adaptive rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20654–20664.

[Papantonakis et al., 2024] Papantonakis, P., Kopanas, G., Kerbl, B., Lanvin, A., and Drettakis, G. (2024). Reducing the memory footprint of 3d gaussian

splatting. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 7(1):1–17.

Amélioration de codecs de parole par modèle de diffusion

Romain Buguet* Stéphane Ragot Thomas Muller
Orange Innovation, Lannion
{prenom.nom}@orange.com

Résumé

Dans cet article on étend une méthode de post-traitement appelée SPF (Score-based Post-Filter), utilisant un modèle de diffusion et visant à réduire le bruit de codage audio. On s'intéresse d'abord au post-traitement du codec Opus (à 24 kbit/s) en modifiant le signal d'entrée de référence lors de l'entraînement. Le post-traitement est ensuite étendu au codec AMR-WB dans le cas multi-débits (6,6, 8,85 et 12,65 kbit/s), avec une amélioration de qualité significative.

Mots clefs

Codage audio, amélioration de la parole, modèles de diffusion, évaluation de la qualité audio.

1 Introduction

Les approches traditionnelles de codage audio mono – dont l'état de l'art est représenté par des codecs comme EVS [1] ou Opus [2] pour les applications conversationnelles – sont basées sur une représentation du signal combinant des blocs de traitement de manière experte. Depuis 2017, on observe une émergence de méthodes de bout en bout de codage audio par réseaux de neurones, comme le vocodage basé WaveNet [3], et plus récemment des codecs de type autoencodeurs basés GAN (Generative Adversarial Networks), comme SoundStream [4], AudioDec [5] ou DAC [6], dont certains offrent un compromis débit/qualité jusque-là inatteignable. Les modèles de diffusion commencent à donner des résultats très prometteurs en codage audio [7, 8, 9] ; dans ce type de modèles, la génération consiste à inverser un processus de diffusion stochastique à partir d'un réseau de neurones.

L'objet de cet article est d'étudier et d'étendre la méthode ScoreDec récemment proposée dans [9] pour améliorer la qualité de codecs de parole. Celle-ci consiste à appliquer à la sortie d'un codec de parole existant (AudioDec ou Opus dans [9]) un post-traitement appelé SPF (Score-based Post-Filter) utilisant un modèle de diffusion. Ce post-traitement réutilise en fait la méthode SGMSE (*Score-based Generative Model for Speech Enhancement*) de [10].

Le post-traitement est une méthode classique d'amélioration de qualité après codage avec des approches "traditionnelles" par traitement du signal [11, 12, 13, 14] ou plus récemment par réseaux de neurones [15, 16, 17, 18, 19]. La spécificité de l'approche ScoreDec est d'appliquer un débruitage par modèle de diffusion.

Cet article est organisé comme suit. La méthode ScoreDec est revue à la section 2. L'amélioration de ScoreDec pour Opus et l'extension à AMR-WB sont détaillées à la section 3, avant de présenter les résultats expérimentaux à la section 4 et conclure à la section 5.

*Romain Buguet était en stage de fin d'études quand ce travail a été réalisé.

2 Revue de la méthode ScoreDec

2.1 Modèles de diffusion

Modèles de diffusion Les modèles de diffusion sont une classe de modèles génératifs inspirés de la physique statistique [20] et développés récemment dans le contexte de la génération d'images [20, 21]. Le mécanisme des modèles de diffusion repose sur deux processus stochastiques, l'un vers l'avant, dit *forward*, et l'autre rétrograde, appelé processus *backward*. Au cours du premier, la structure des données, de distribution p inconnue, est progressivement détruite et évolue lentement vers une distribution cible connue et simple à échantillonner, le plus souvent gaussienne. Cette transformation est modélisée par une chaîne de Markov $\{\mathbf{x}_t\}_{t=0}^{T}$ issue de $\mathbf{x}_0 \sim p$, indexée par la variable de temps t, et dont les états correspondent à des niveaux de bruits de plus en plus élevés, si bien que la distribution de l'état final $\mathbf{x}_T$ est proche d'une loi normale centrée réduite. À l'inverse, le processus *backward* génère un échantillon $\mathbf{x}_0 \sim p_\theta$ (où p_θ est proche de p) par débruitages successifs, à partir d'un bruit gaussien $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$. L'apprentissage consiste à estimer le bruit ajouté à chaque étape du processus *forward* afin de le soustraire graduellement lors de l'inférence.

Modèles génératifs basés sur le score Plutôt que d'estimer directement la distribution p des données au moyen d'un modèle p_θ, les modèles génératifs basés sur le score estiment la *fonction de score (de Stein)* $\nabla_{\mathbf{x}} \log p(\mathbf{x})$ des données à l'aide d'un *modèle de score* s_θ [22]. Dans [23], Song et al. proposent le formalisme suivant pour le processus *forward*, basé sur l'équation différentielle stochastique (EDS)

$$\mathrm{d}\mathbf{x}_t = \mathbf{f}(t, \mathbf{x}_t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}_t, \tag{1}$$

où le champ de vecteurs $\mathbf{f}$ est un terme de dérive (*drift*), gouvernant le comportement moyen de l'équation, g est un coefficient contrôlant la quantité de bruit injectée à chaque instant, et $\mathbf{w}_t$ est un mouvement brownien (processus de Wiener). La variable temporelle t évolue ici continûment entre l'instant initial $t = 0$ et l'instant final $t = T$. D'après [24], le processus *backward* correspond à la solution de l'EDS rétrograde associée à (1)

$$\mathrm{d}\mathbf{x}_t = \left[-\mathbf{f}(t, \mathbf{x}_t) + g(t)^2 \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)\right] \mathrm{d}t + g(t)\mathrm{d}\bar{\mathbf{w}}_t, \tag{2}$$

qui fait intervenir la fonction de score de la distribution p_t des données au temps t, ainsi qu'un processus de Wiener rétrograde $\bar{\mathbf{w}}_t$. Le score est approximé par $s_\theta(\mathbf{x}_t, t)$, modélisé par un réseau de neurones. Une fois le modèle entraîné, de nouveaux échantillons sont produits par simulation du processus *backward* (2) à l'aide de méthodes de résolution numérique d'EDS, après avoir substitué le modèle s_θ à la fonction de score.

2.2 Modèle SGMSE

SGMSE [10] est un modèle génératif basé sur le score pour l'amélioration de la parole. Le signal de parole observé est noté $\mathbf{y}$ et correspond à une version bruitée du signal de parole pure $\mathbf{x}$. Le modèle SGMSE inclut un processus *forward* dont le but est de

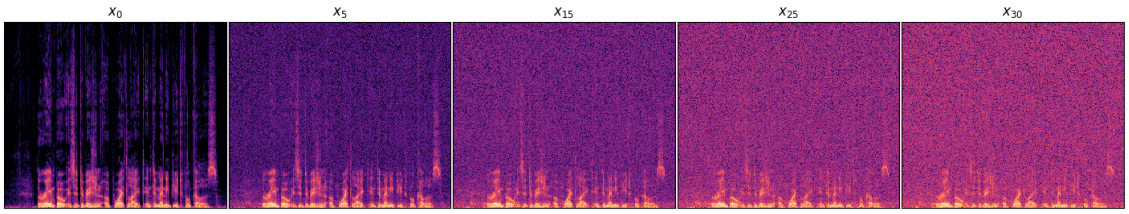


FIGURE 1 – *Spectrogramme de $\mathbf{x}_t$ (ici montré uniquement en amplitude) à différentes itérations du processus* forward.

transférer la distribution - inconnue - de la parole pure $\mathbf{x}$ vers une distribution gaussienne centrée sur les données bruitées (avec une dérive progressive de $\mathbf{x}$ vers $\mathbf{y}$ lors du processus de diffusion), et un processus *backward* réalisant la transformation inverse. Dans ce modèle, les données sont traitées sous forme de spectrogrammes complexes. Avec les notations du paragraphe précédent, le terme de *drift* $\mathbf{f}(t, \mathbf{x}_t)$ de l'équation 1 est remplacé par

$$\mathbf{f}(\mathbf{x}_t, \mathbf{y}) := \gamma(\mathbf{y} - \mathbf{x}_t), \tag{3}$$

et le coefficient de diffusion est donné par

$$g(t) := \sigma_{\min} \left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^t \sqrt{2 \log \left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)}, \tag{4}$$

où $\gamma > 0$ est un coefficient de raideur et les hyperparamètres $\sigma_{\min}$ et $\sigma_{\max}$ contrôlent la diffusion. Notons que la fonction $\mathbf{f}$ à l'équation 3 dépend de $\mathbf{y}$ et ne dépend pas explicitement du temps t – pour simplifier les notations, on garde cependant le même symbole $\mathbf{f}(.)$.

Inférence SGMSE. À l'inférence, les audio dégradés $y(n)$ sont normalisés par $M = \max_n |y(n)|$ puis transformés en spectrogrammes complexes $\mathbf{y} \in \mathbb{C}^{K \times F}$ par transformée de Fourier discrète à court-terme (STFT) et compression d'amplitude, où K et F sont les nombres de trames temporelles et de raies fréquentielles. Le modèle de diffusion SGMSE, basé sur l'architecture de réseau de neurones NCSN++ (pour *Noise Conditional Score Network*), synthétise une version améliorée $\hat{\mathbf{x}}$ de $\mathbf{y}$, qui est ensuite convertie dans le domaine temporel et renormalisée. Pour cela, une version fortement corrompue $\mathbf{x}_T$ de $\mathbf{y}$ est échantillonnée selon une distribution gaussienne complexe centrée en $\mathbf{y}$. L'intervalle $[0, T]$ est discrétisé en N sous-intervalles de longueur ΔT, et l'équation *backward* est résolue entre $t = T$ et $t = t_\epsilon$, où $t_\epsilon \simeq 0$, à l'aide de méthodes numériques telles que la méthode d'Euler-Maruyama. La limitation à t_ϵ permet d'éviter des instabilités numériques pouvant se produire pour t proche de 0.

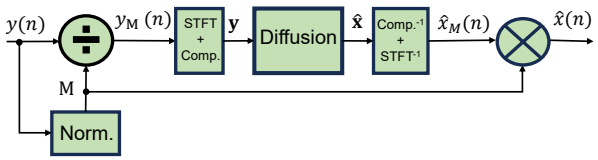


FIGURE 2 – *Modèle SGMSE (inférence).*

Entraînement SGMSE. À chaque étape de l'entraînement, un temps t est d'abord échantillonné selon une distribution uniforme sur l'intervalle $[t_\epsilon, T]$, puis un couple $(\mathbf{x}_0, \mathbf{y})$ de spectrogrammes purs/dégradés est choisi aléatoirement dans la base de données. Connaissant $\mathbf{x}_0$ et $\mathbf{y}$, la distribution de $\mathbf{x}_t$ peut être déterminée explicitement : on peut donc échantillonner directement $\mathbf{x}_t$ et calculer la fonction de score correspondante. Une distance l_2 entre le modèle de score et le score est enfin calculée, puis les paramètres du réseau sont actualisés.

2.3 Post-traitement SPF dans ScoreDec

La méthode ScoreDec [9] est résumée à la figure 3, où un codeur existant comme Opus (incluant un préfiltre passe-haut noté H) prend en entrée le signal $x(n)$, puis le décodeur reconstruit un signal $y(n)$. La méthode ScoreDec revient à post-traiter $y(n)$ avec le modèle de débruitage SGMSE décrit à la figure 2 ; le bruit de codage est ainsi réduit pour améliorer la qualité audio.

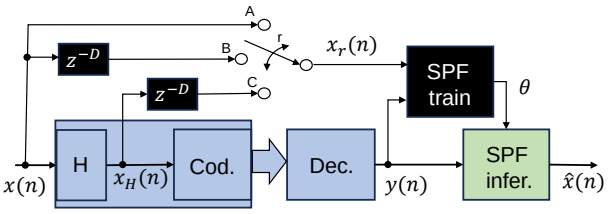


FIGURE 3 – *Modèle ScoreDec avec 3 points de référence ($r = A, B, C$) – $r = A$ correspond au cas traité dans [9].*

Dans [9], le codec Opus est traité comme une "boîte noire" : l'entraînement du post-traitement SPF – qui donne les paramètres θ du modèle de diffusion – est ainsi réalisé en prenant comme point de référence $r = \text{A}$, soit $x_r(n) = x(n)$.

3 Méthode proposée

3.1 Amélioration de SPF pour Opus

Nous proposons d'améliorer ScoreDec, en considérant Opus non pas comme une "boîte noire" mais en prenant en compte des caractéristiques connues du codec. En particulier, Opus induit un retard algorithmique de D échantillons entre l'entrée $x(n)$ et la sortie $y(n)$ – par défaut, à 48 kHz, $D = 312$ (6,5 ms). De plus, Opus inclut un préfiltre passe-haut H (de fréquence de coupure adaptative autour de 60–70 Hz).
On propose donc de modifier la procédure d'entraînement de ScoreDec, comme indiqué à la figure 3 : le signal de référence est remplacé soit par le signal retardé $x_r(n) = x(n - D)$ (cas $r =$ B), soit par le signal préfiltré et retardé $x_r(n) = x_H(n - D)$ (cas $r =$ C) – le signal $y(n)$ en sortie du codec Opus reste le même peu importe le signal d'entrée de référence. Ainsi, le post-filtre SPF ne compense pas le retard algorithmique d'Opus (cas $r =$ B ou C), et ne modélise pas le préfiltre H (cas $r =$ C). La capacité du modèle de diffusion est ainsi concentrée sur la réduction du bruit de codage induit par Opus.

3.2 Extension de ScoreDec à AMR-WB

Nous proposons aussi d'étendre la méthode ScoreDec au codec AMR-WB [25]. Le post-traitement SPF pour AMR-WB suit le

principe de la figure 3, mis à part que les blocs de codage et décodage correspondent respectivement au codeur et décodeur AMR-WB. À la différence d'Opus, on traite le cas multi-débits en considérant les trois modes d'AMR-WB à plus bas débit ($6,6$, $8,85$ et $12,65$ kbits/s) qui sont utilisés en téléphonie mobile. Il est alors possible d'étudier si un post-traitement SPF entrainé à un débit donné reste optimal pour un autre débit d'AMR-WB.
À noter que le retard algorithmique d'AMR-WB est de $D = 95$ échantillons à 16 kHz. AMR-WB inclut aussi un préfiltre passe-haut (avec une fréquence de coupure de 31 Hz). Cependant, ce préfiltre opère sur une bande basse 0–6,4 kHz après décimation à 12,8 kHz ; pour simplifier, on se restreint au cas $r = \text{B}$ de la figure 3 pour AMR-WB.

4 Expériences

4.1 Protocole expérimental

Bases de données : Pour Opus, le modèle ScoreDec est entraîné et testé sur la même base de données Valentini [26] à 48 kHz (parole pure uniquement) que dans [9], pour des questions de reproductibilité et de comparaison directe avec [9]. Pour AMR-WB, cette base Valentini est ré-échantillonnée à 16 kHz.

Paramètres des modèles testés : La méthode ScoreDec pour Opus est configurée par défaut comme dans [9] (avec des trames de 320 échantillons). Pour AMR-WB, on reprend les paramètres SGMSE de [10] (avec des trames de 128 échantillons).

Evaluation de qualité : Les intervalles de confiance à 95% sont donnés avec chaque note moyenne de qualité. La qualité est évaluée par les métriques objectives PESQ [27] et SI-SDR [28] comme dans [9]. Le SI-SDR s'interprète comme un rapport signal à bruit, mais il est difficile à corréler avec une évaluation subjective. Le score PESQ prédit la note MOS (Mean Opinion score) sur l'échelle de qualité d'écoute : $1\rightarrow$ mauvaise, $2\rightarrow$ médiocre, $3\rightarrow$ passable, $4\rightarrow$ bonne, $5\rightarrow$ excellente.

Configuration matérielle : Les entraînements et inférences de SPF sont réalisés sur un GPU A100 avec 40 Go de RAM. Un entraînement SPF dure en moyenne 96 h pour Opus et 30 h pour AMR-WB (sur la base Valentini). L'inférence sur la base de test Valentini prend 3 h pour Opus et 45 min pour AMR-WB (pour 45 min de parole).

4.2 Résultats pour Opus

Les résultats de qualité objective pour Opus (à 24 kbit/s) sont résumés dans le tableau 1. Le cas $r = \text{A}$ du post-traitement SPF (noté Opus_SPF) correspond à [9] – les poids pré-entrainés du modèle de [9] n'étant pas disponibles, ce modèle a été ré-entrainé mais nos résultats sont très proches de ceux de [9] dans les mêmes conditions. Les résultats de [9] sont rappelés au tableau 1 à des fins de comparaison. Notre version ré-entrainée d'Opus_SPF donne des résultats objectifs légèrement meilleurs que ceux présentés dans [9]. La seule différence significative concerne le score SI-SDR pour Opus. Contrairement à [9], la qualité d'Opus est ici mesurée en prenant le signal $x_H(n - D)$ comme signal d'entrée d'Opus avant calcul de la métrique SI-SDR ; le SI-SDR passe ainsi de $-20,62$ dB dans [9] à $12,48$ dB pour la même base de test ; cela élimine le biais dû au retard algorithmique et au préfiltre, on observe un gain de seulement $4,01$ dB en SI-SDR entre Opus et Opus_SPF avec $r = \text{A}$; on confirme une amélioration du score PESQ d'environ 0.09 pour Opus_SPF avec $r = \text{A}$.
Les scores PESQ d'Opus_SPF modifié selon nos propositions à la section 3.1 sont légèrement meilleurs en changeant l'entraînement du modèle (en passant à $r = \text{B}$ puis C) – le SI-SDR montre un gain plus net.

Codec	r	SI-SDR (dB)	PESQ (MOS-LQO$_{wb}$)
Opus [9]	–	-20,62	4,21
Opus (nos tests)	–	$12,48 \pm 0,09$	$4,24 \pm 0,01$
Opus_SPF [9]	A	16.20	4.29
Opus_SPF (nos tests)	A	$16,49 \pm 0,12$	$4,33 \pm 0,02$
	B	$16,66 \pm 0,12$	$4,34 \pm 0,01$
	C	$\mathbf{17,08 \pm 0,13}$	$\mathbf{4,35 \pm 0,02}$

TABLEAU 1 – *Qualité objective sur la base de test Valentini pour Opus à 24 kbit/s avec ou sans post-traitement SPF (avec un point de référence r = A, B ou C à l'entraînement tel que défini à la figure 3).*

Un test d'écoute de type "RefAB" a été réalisé par 5 sujets experts, avec 30 double phrases de parole (8s) en français (normalisées en niveau d'écoute). Ce test subjectif a confirmé les scores PESQ : il n'y pas de différence statistiquement significative entre Opus_SPF avec $r = \text{A}$ et $r = \text{C}$.
Cette étude sur Opus a repris les conditions de [9] où Opus opère à 24 kbit/s, alors qu'Opus a déjà une bonne qualité sur la parole pure à ce débit. Par la suite, il serait intéressant de tester Opus à un débit inférieur à 24 kbit/s.

4.3 Résultats pour AMR-WB

Les résultats de qualité objective pour AMR-WB (à 6,6, 8,85 et 12,65 kbit/s) sont résumés au tableau 2 et illustrés à la figure 4. Le post-traitement SPF pour AMR-WB (abrégé en SPF_x, où x est le débit d'entraînement) est entraîné séparément pour chaque mode (x = 6,6, 8,85 et 12,65), puis sur une base de données comportant un mélange de ces trois débits – ce dernier cas est noté "SPF_MR" (MR pour Multi-Rate).
La comparaison inclut AMR-WB et sa version post-traitée, SPF_x ou SPF_MR. Le post-traitement SPF_x entraîné pour AMR-WB à x kbit/s améliore significativement le score PESQ au même débit (avec un écart de 0,80, 0,52 et 0,37 à 6,6, 8,85 et 12,65 kbit/s, respectivement) ; cette nette amélioration est confirmée par des écoutes informelles sur la base de test Valentini.
Comme on pourrait l'attendre, le post-traitement SPF_x est surtout optimisé pour le débit pour lequel il a été entraîné. En termes de score PESQ, le modèle donnant le meilleur résultat à un débit donné est celui qui a été entrainé pour ce même débit ; c'est également le cas en termes de SI-SDR, sauf au débit de 6,6 kbit/s où le modèle à 8,85 kbit/s obtient un score SI-SDR numériquement supérieur – on note cependant que les modèles entrainés à 6,6, 8,85 et 12,65 kbit/s sont en fait équivalents (en tenant compte de la marge d'erreur). La métrique SI-SDR est connue pour être faiblement corrélée à une évaluation subjective (perceptive), on considère ici PESQ comme donnant une meilleure indication (prédiction) de la qualité perçue du signal post-traité.
Parmi les méthodes précédemment proposées de post-traitement neuronal, on retient ici la méthode par réseaux convolutionnels de [15] qui s'applique aussi à AMR-WB ; dans le tableau IV de [15] (avec une base en anglais et allemand) les tests limités au débit de 12,65 kbit/s donnent un score PESQ de 3,60 pour AMR-WB et 3,85 pour la meilleure variante de post-traitement, soit un gain de 0,25 en termes de score PESQ. Ici, pour le même débit de 12,65 kbit/s le tableau 2 montre un score PESQ de 3,58 pour AMR-WB et 3,95 pour SPF_12,65, soit un gain de 0,37 en termes de score PESQ. Même si ces résultats ne sont pas directement comparables car ils ne sont pas obtenus dans les mêmes conditions

de tests, on peut s'attendre à ce que le débruitage par modèle de diffusion donne de meilleurs résultats étant donné que l'approche considérée de type SGMSE relâche certaines contraintes (étant plus complexe et non causale).

Le post-traitement SPF_MR entraîné en multi-débits a une performance plus homogène que SPF_x mais sous-optimale. Par la suite, il sera intéressant de donner le débit de décodage comme contexte (conditionnement) au modèle de diffusion pour viser une performance optimale en multi-débits. Par ailleurs, on pourra aussi ajouter à la comparaison du tableau 2 plus de débits d'AMR-WB (par exemple le débit maximal de 23,85 kbit/s aussi utilisé en téléphonie) et comparer la qualité avec le décodage AMR-WB amélioré dans EVS (EVS-IO pour InterOperable) [1] qui utilise des techniques classiques de post-traitement [14].

Codec	Mode	SI-SDR (dB)	PESQ (MOS-LQO$_{wb}$)
AMR-WB		$6,00 \pm 0,10$	$2,60 \pm 0,02$
SPF_6,6		$11,30 \pm 0,09$	$\mathbf{3,40 \pm 0,02}$
SPF_8,85	6,6	$\mathbf{11,38 \pm 0,10}$	$3,35 \pm 0,02$
SPF_12,65		$11,34 \pm 0,09$	$3,29 \pm 0,02$
SPF_MR		$10,95 \pm 0,10$	$3,31 \pm 0,02$
AMR-WB		$7,20 \pm 0,11$	$3,08 \pm 0,03$
SPF_6,6		$12,17 \pm 0,02$	$3,54 \pm 0,02$
SPF_8,85	8,85	$13,64 \pm 0,10$	$\mathbf{3,71 \pm 0,03}$
SPF_12,65		$\mathbf{13,71 \pm 0,10}$	$3,66 \pm 0,02$
SPF_MR		$12,87 \pm 0,10$	$3,61 \pm 0,02$
AMR-WB		$7,99 \pm 0,11$	$3,58 \pm 0,03$
SPF_6,6		$12,50 \pm 0,10$	$3,58 \pm 0,03$
SPF_8,85	12,65	$13,65 \pm 0,11$	$3,73 \pm 0,02$
SPF_12,65		$\mathbf{15,38 \pm 0,11}$	$\mathbf{3,95 \pm 0,02}$
SPF_MR		$14,82 \pm 0,11$	$3,90 \pm 0,02$

TABLEAU 2 – *Qualité objective sur la base de test Valentini pour AMR-WB, SPF_x (où x est le débit d'entraînement) ou SPF_MR (pour Multi-Rate) – le post-traitement SPF est entraîné avec le point de référence $r = B$ tel que défini à la figure 3.*

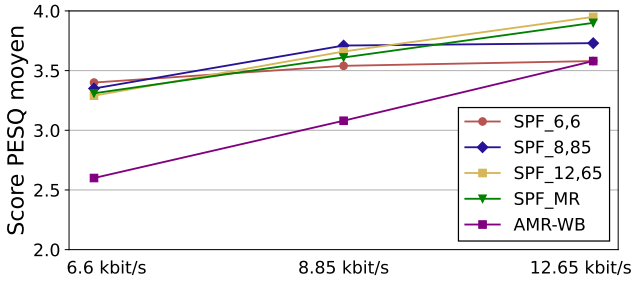


FIGURE 4 – *Représentation graphique des résultats pour AMR-WB donnés au tableau 2.*

5 Conclusion

Dans cet article, la méthode de post-traitement SPF de [9] a été améliorée pour Opus et étendue à AMR-WB. Le gain en qualité apporté par le post-traitement SPF est prometteur et significatif pour AMR-WB, cependant cette méthode traite des séquences audio complètes (et non par trame), avec une complexité très élevée et sans opérer en temps réel par trame courte (≤ 20 ms) ; il sera intéressant de corriger ces aspects en modifiant le modèle SGMSE sous-jacent. L'approche supervisée de [29] pourra aussi être testée dans le contexte du codage. L'extension de cette étude à d'autres contenus audio que la parole (pure) sera aussi nécessaire.

Références

[1] M. Dietz et al. Overview of the EVS codec architecture. Dans *Proc. ICASSP*, 2015.

[2] IETF RFC 6716. Definition of the Opus Audio Codec, 2012.

[3] W.B. Kleijn et al. Wavenet based low rate speech coding. arXiv :1712.01120, 2017.

[4] N. Zeghidour et al. SoundStream : An End-to-End Neural Audio Codec. *IEEE/ACM TASLP*, 30, 2021.

[5] Y.-C. Wu et al. Audiodec : An Open-Source Streaming High-Fidelity Neural Audio Codec. Dans *Proc. ICASSP*, 2023.

[6] R. Kumar et al. High-Fidelity Audio Compression with Improved RVQGAN. Dans *Proc. NIPS*, 2023.

[7] R. San Roman et al. From discrete tokens to high-fidelity audio using multi-band diffusion. arXiv :2308.02560, 2023.

[8] Y. Haici et al. Generative de-quantization for neural speech codec via latent diffusion. Dans *Proc. ICASSP*, 2024.

[9] Y.-C. Wu et al. ScoreDec : A Phase-Preserving High-Fidelity Audio Codec with a Generalized Score-Based Diffusion Post-Filter. Dans *Proc. ICASSP*, 2024.

[10] J. Richter et al. Speech enhancement and dereverberation with diffusion-based generative models. *IEEE/ACM TASLP*, 31 :2351–2364, 2023.

[11] V. Ramamoorthy et N. S. Jayant. Enhancement of ADPCM speech by adaptive postfiltering. *AT&T Bell Laboratories Technical Journal*, 63(8) :1465–1475, 1984.

[12] J.H. Chen et A. Gersho. Adaptive postfiltering for quality enhancement of coded speech. *IEEE Transactions on Speech and Audio Processing*, 3(1) :59–71, 1995.

[13] J.-L. Garcia, C. Marro, et B. Kövesi. A PCM coding noise reduction for ITU-T G.711.1. Dans *Proc. Interspeech*, 2008.

[14] T. Vaillancourt, R. Salami, et M. Jelínek. New post-processing techniques for low bit rate CELP codecs. Dans *Proc. ICASSP*, 2015.

[15] Z. Zhao, H. Liu, et T. Fingscheidt. Convolutional Neural Networks to Enhance Coded Speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(4) :663–678, 2019.

[16] Srikanth Korse, Kishan Gupta, et Guillaume Fuchs. Enhancement of Coded Speech Using a Mask-Based Post-Filter. Dans *Proc. ICASSP*, 2020.

[17] Kishan Gupta, Srikanth Korse, Bernd Edler, et Guillaume Fuchs. A DNN Based Post-Filter to Enhance the Quality of Coded Speech in MDCT Domain. Dans *Proc. ICASSP*, 2022.

[18] S. Korse, N. Pia, K. Gupta, et G. Fuchs. PostGAN : A GAN-Based Post-Processor to Enhance the Quality of Coded Speech. Dans *Proc. ICASSP*, 2022.

[19] J. Büthe, J.-M. Valin, et A. Mustafa. Lace : A Light-Weight, Causal Model for Enhancing Coded Speech Through Adaptive Convolutions. Dans *Proc. WASPAA*, 2023.

[20] J. Sohl-Dickstein et al. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. Dans *Proc. ICML*, 2015.

[21] J. Ho et al. Denoising diffusion probabilistic models. Dans *Proc. NeurIPS*, 2020.

[22] Y. Song et al. Generative modeling by estimating gradients of the data distribution. Dans *Proc. NeurIPS*, 2019.

[23] Y. Song et al. Score-based generative modeling through stochastic differential equations. Dans *Proc. ICLR*, 2021.

[24] B. D. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their App.*, 12(3) :313–326, 1982.

[25] B. Bessette et al. The adaptive multirate wideband speech codec (AMR-WB). *IEEE TSAP*, 10(8) :620–636, 2002.

[26] Valentini-Botinhao C. et al. Noisy speech database for training speech enhancement algorithms and TTS models. https ://datashare.ed.ac.uk/handle/10283/2791, 2017.

[27] A.W. Rix et al. Perceptual evaluation of speech quality (PESQ) – a new method for speech quality assessment of telephone networks and codecs. Dans *Proc. ICASSP*, 2001.

[28] J. Le Roux et al. SDR – Half-baked or Well Done ? Dans *Proc. ICASSP*, 2019.

[29] J.E. Ayilo et al. Diffusion-based speech enhancement with a weighted generative-supervised learning loss. Dans *Proc. ICASSP*, 2024.

Primer design for DNA storage random access

Jérémy Mateos[1,3], Dominique Lavenier[2], Melpomeni Dimopoulou[3], Anthony Genot[4], Marc Antonini[1]
[1] I3S Laboratory, Côte d'Azur University, CNRS, UMR 7271, Sophia Antipolis, France
[2] University of Rennes, Inria, CNRS, IRISA, Campus de Beaulieu, Rennes, France
[3] Pearcode, Sophia-Antipolis, France
[4] LIMMS (IRL2820)/CNRS-IIS, University of Tokyo, Tokyo, Japan
Email: mateos@i3s.unice.fr, dominique.lavenier@irisa.fr, melpomeni@pearcode.io, genot@iis.u-tokyo.ac.jp, am@i3s.unice.fr

Abstract

DNA is a promising candidate for data storage due to its high density and long-term stability. However, accessing specific data, known as random access, is challenging. This process uses primers, short DNA segments that act as identifiers. Efficient random access depends on high-quality primers, constrained by DNA structure. This paper introduces a method to generate primers that meet strict biochemical criteria, avoiding sequences that form problematic shapes. The proposed tool uses computational models to predict primer binding affinity and specificity, allowing users to adjust parameters for lab protocols, enhancing data retrieval efficiency and optimization.

Keywords

DNA data storage, PCR-based random access

1 Introduction

Unlike traditional storage media which face problems of longevity, integrity, ecology and energy consumption, DNA offers an incredible data density and stability over long periods, lasting thousands of years if stored under optimal conditions.

The DNA data storage workflow consists of six steps:

1. **Encoding:** Digital data is converted into sequences of A, C, G, T, cut into short chunks called oligonucleotides, and formatted with addressing fields,
2. **DNA synthesis:** Synthetic oligonucleotides are created based on the encoded information
3. **Storage:**DNA molecules are stored in controlled environments to ensure stability and longevity,
4. **Data retrieval:** Specific segments of stored DNA data are accessed randomly, mixing multiple DNA molecules in the same container,
5. **Sequencing:** DNA sequencers read and translate DNA molecules into A, C, G, T sequences,
6. **Decoding:** Retrieved sequences are reorganized and corrected to recover the original data.

This paper examines the efficiency of the data recovery phase in DNA-based storage according to the quality of the addressing primers (small adressing DNA sequence). Millions of files are stored in DNA molecules within the same space. Random access involves selecting DNA molecules corresponding to a file through PCR (Polymerase Chain Reaction) [1] [2]. PCR allows selective duplication of these molecules using primers that identify the sequences associated to one file. This technique, adapted for DNA data storage, ensures specific file retrieval, similar to accessing digital archives.

PCR-based random access [3] is an innovative modification of the classical PCR, commonly used in molecular biology, but requires precise primer design. Primers must attach only to short segments at the extremities of the oligonucleotides and should not form undesired shapes or loops to avoid non specific retrieval or data loss. To our knowledge, no software exists specifically for designing primers for DNA data storage, as existing tools are tailored for genomics and unsuitable for this purpose [4]. High-quality primer design is crucial for accurate file extraction, given the complexity of managing millions of sequences in the same location.

2 Molecular random access

As introduced previously, DNA data storage involves mixing millions of DNA oligonucleotides in a single container. Files are represented by DNA oligonucleotides with specific primer pairs at their extremities that barcode the files. Those files are present with multiple copies of each oligonucleotides. Accessing files uses a "PCR-based random access" method to amplify and retrieve specific DNA oligos among the others.

Unlike traditional PCR, DNA data storage primers are designed specifically for files and not derived from existing sequences. Each file has a unique molecular address provided by a primer, which must bind, called hybridization, efficiently to prevent incorrect amplification leading to data loss.

The PCR is a mix containing DNA, primers, DNA polymerase, cations (Na+ and Mg2+), and nucleotides. It involves three main steps, forming a cycle, that are repeated multiple times to replicate the DNA:

1. **Denaturation:** The reaction mix is heated to around 94-98°C to initiate replication.
2. **Annealing:** The temperature is lowered to allow the primers to bind to their complementary sequence on the oligo. This temperature is based on the **melting temperature (Tm)** depending on the primers being used, which will be explained in the following paragraph.
3. **Extension:** The temperature is raised to 72°C, the temperature to activate the DNA polymerase. DNA

polymerase synthesizes a new DNA strand by adding nucleotides to the primers, creating a complementary strand to each of the original strands.

Effective primer design is crucial for robust random access in large DNA data storage systems, ensuring reliable file amplification without cross-hybridization. Key factors include the melting temperature (Tm), the point at which 50% of double-stranded DNA separates into single strands. Tm depends on parameters like Na+ and Mg2+ concentrations in the PCR mixture.
The challenge in DNA storage is designing large sets of primers to address many files while ensuring specificity and compatibility under the same PCR conditions, especially maintaining consistent Tm.

3 Primer Design

As higlighted in the previous section, it is mandatory to design good primers to ensure an effective and specific selection of oligonucleotides encoding a specific document during the PCR process. The primer design plays a major role in the success of the PCR-based random access. In DNA data storage, unlike genomic studies, numerous primers must coexist with minimal interference. Thus, the process is to firstly design primers with specific characterisctics to optimize the PCR. Secondly, all the primers are checked to ensure their compatibility with each other and with the dataset oligos. The design of the primers is implemented in a set of tools called DSPT (Dna Storage Primer Tools). The following programs, written in C, are currently available :

- **DSPgen** generate a set of primers according to a rigorous list of criteria to meet the PCR requirements. Those criterias are designed by biochemical and biotechnology constraints. This is an adaptation of the IThOS software previously developped for designing primers for genomic purpose [5]. More recent methods for calculating the melting temperature [6] [1] , which increase precision, have been added. Additionally, supplementary filters have been implemented to better meet DNA storage requirements. One of these filters addresses secondary structure, avoiding sequences that can form undesired shapes or loops that hinder DNA amplification and data retrieval. At last, it allows the generation of thousands of primers in less than one second.
- **DSPham** checks the Hamming distance between primers and eliminates the minimum set necessary to maximize the number of primers with a Hamming distance above a user-defined threshold.
- **DSPhyb** detects potential hybridization, from a thermodynamical point of view, between primers and DNA sequences by computing the primers stability called ΔG (Gibbs free energy G) between small similar regions of both primers and DNA sequences.

[1] https://eu.idtdna.com/calc/Analyzer/Home/definitions

The DSPT package is available on the following gitlab: https://gitlab.inria.fr/molecularxiv-pc-2/dna-storage-primer-tools

4 Experiments results

For the experiments, the tools have been tested on the JPEG-AIC-03 [7] dataset encoded with JPEGDNA-SFC4-S-R [8]. It includes 10 images and represents a total of 99156 oligonucleotides of length equal to 300. Different primer sets have been generated (using DSPgen) and checked (using DSPhyb) for potential hybridization with the oligonucleotides provided in the JPEG-AIC-03 dataset. Figure 1 illustrates the number of detected hybridization sites. These results suggest that the encoding process can be improved. Introducing thermodynamic verification for each oligonucleotide during encoding could enhance the quality of the oligonucleotides. This improvement would prevent the oligonucleotides from binding to each other and forming secondary structures. Additionally, this tool could be used to verify headers, indexes, and other metadata to ensure accurate data retrieval.

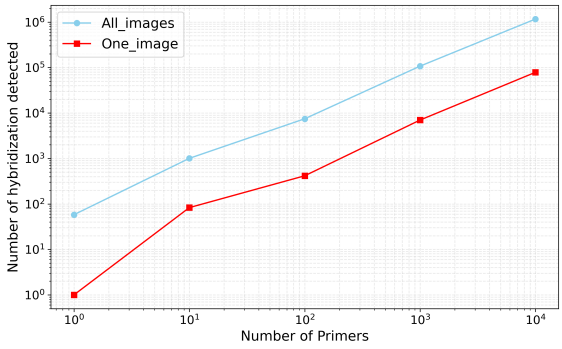


Figure 1: ***DSPhyb - Hybridization detection :*** *Number of hybridization detected while checking if the primers generated can hybridize to the payload of the JPEG-AIC-03 dataset encoded with JPEGDNA-SFC4-S-R.*

5 Conclusion

The DSP software demonstrated excellent performance, particularly in generating efficient primers, which are crucial for PCR-based random access commonly used in molecular biology. With DSP tools, researchers can achieve greater precision and efficiency in their experiments. Additionally, the tools have short execution times, allowing for efficient testing of multiple PCR configurations. Although the software is still under development, the current versions are very promising. The next tool to be developed will check primer compatibility, and a parallel version is planned to further reduce execution times. However, to fully validate these tools further tests are needed, as well as wetlab experiments to ensure the primers meet all specifications.

References

[1] IR Lehman. Discovery of dna polymerase. Journal of Biological Chemistry, 278(37):34733–34738, 2003.

[2] Lilit Garibyan et Nidhi Avashia. Research techniques made simple: polymerase chain reaction (pcr). The Journal of investigative dermatology, 133(3):e6, 2013.

[3] Lee Organick, Siena Dumas Ang, Yuan-Jyue Chen, Randolph Lopez, Sergey Yekhanin, Konstantin Makarychev, Miklos Z Racz, Govinda Kamath, Parikshit Gopalan, Bichlien Nguyen, et al. Random access in large-scale dna data storage. Nature biotechnology, 36(3):242–248, 2018.

[4] Jian Ye, George Coulouris, Irena Zaretskaya, Ioana Cutcutache, Steve Rozen, et Thomas L Madden. Primer-blast: a tool to design target-specific primers for polymerase chain reaction. BMC bioinformatics, 13:1–11, 2012.

[5] Nouri Ben Zakour, Michel Gautier, Rumen Andonov, Dominique Lavenier, Marie-Françoise Cochet, Philippe Veber, Alexei Sorokin, et Yves Le Loir. GenoFrag: software to design primers optimized for whole genome scanning by long-range PCR amplification. Nucleic Acids Research, 32(1):17–24, 01 2004.

[6] Richard Owczarzy, Bernardo G Moreira, Yong You, Mark A Behlke, et Joseph A Walder. Predicting stability of dna duplexes in solutions containing magnesium and monovalent cations. Biochemistry, 47(19):5336–5353, 2008.

[7] Michela Testolina, Vlad Hosu, Mohsen Jenadeleh, Davi Lazzarotto, Dietmar Saupe, et Touradj Ebrahimi. Jpeg aic-3 dataset: Towards defining the high quality to nearly visually lossless quality range. pages 55–60, 2023.

[8] Xavier Pic, Eva Gil San Antonio, Melpomeni Dimopoulou, et Marc Antonini. Rotating labeling of entropy coders for synthetic dna data storage. Dans 2023 24th International Conference on Digital Signal Processing (DSP), pages 1–5, 2023.

Édition visuelle pilotée par l'audio à l'aide d'outils de synthèse vocale

Rémi Decelle[1] Serge Miguet[1] Thibault Jaillon[2]

[1] Univ Lyon, Univ Lyon 2, CNRS, INSA Lyon, UCBL, LIRIS, UMR5205

[2] Mon Petit Placement

{remi.decelle, serge.miguet}@univ-lyon2.fr thibault@monpetitplacement.fr

Résumé

L'édition visuelle ou Facial Reenactment est une tâche complexe qui nécessite une compréhension approfondie de divers outils pour obtenir de bons résultats. Malgré les progrès récents, plusieurs défis subsistent, tels que la synchronisation des lèvres, l'absence de mouvements labiaux pendant les silences et la préservation de l'identité. L'utilisation des spectrogrammes de Mel pour représenter l'audio est limitée pour capturer les nuances et les expressions faciales. Dans notre approche, nous utilisons des outils de synthèse vocale tels que le réseau EnCodec pour fournir des caractéristiques audio et textuelles, extraites à l'aide du modèle CLIP. Ces caractéristiques devraient permettre une meilleure qualité visuelle et une compréhension plus nuancée des mots prononcés et des expressions faciales. Nous avons également construit un jeu de données francophones. Les expériences nous encouragent à approfondir cette approche, qui donne des résultats équivalents à l'état de l'art.

Mots clefs

Deepfake, Animation Faciale, Synthèse Vidéo, Génération conditionnée.

1 INTRODUCTION

L'apprentissage profond permet désormais de faire du doublage visuel automatique (ou édition visuelle ou bien encore *facial reenactment*) en synchronisant précisément les mouvements de lèvres avec l'audio. Cette technologie a de nombreuses applications, notamment dans le montage vidéo, la post-production et la traduction en temps réel, utiles dans divers secteurs tels que le cinéma, les jeux vidéo ou l'éducation.

Bien que des progrès aient été réalisés dans le domaine du doublage visuel automatique, il reste des défis à relever pour obtenir un rendu naturel. Les méthodes précédentes basées sur les repères faciaux et les réseaux neuronaux en 2D génèrent des visages déformés car elles ne parviennent pas à séparer correctement le mouvement de la tête et l'expression faciale. Les méthodes utilisant une représentation de l'animation du visage basée sur des données ont également des difficultés à produire des vidéos de haute qualité. Cependant, l'utilisation d'un modèle facial 3D comme extracteur de caractéristiques dans les méthodes d'apprentissage profond semble donner des résultats plus précis et naturels.

Sur la base des observations précédentes, une nouvelle architecture est proposée pour le doublage visuel automatique, utilisant des méthodes d'édition visuelle basées sur l'*inpainting*. Le fait d'utiliser uniquement des données audio comme données d'entraînement pour la génération peut conduire à des résultats désynchronisés. Pour limiter ce problème, nous proposons d'utiliser des couches telles que la normalisation adaptative des instances (AdaIn) et l'utilisation d'outils Speech-To-Text [1], qui permet d'extraire le texte à partir de l'audio, pour obtenir des mouvements labiaux plus précis et plus naturels. Le schéma de l'architecture de notre méthode s'inspire de [2] et est présenté dans la figure 1. Nous considérons tout d'abord les coefficients de mouvement de la 3DMM (3D Morphable Model) comme une représentation latente du mouvement et de l'expression du visage.

Notre architecture pour le doublage visuel automatique intègre les caractéristiques audio et textuelles extraites à partir des réseaux EnCodec [3] et CLIP [4] respectivement. Elle se compose de trois modules : extraction des caractéristiques 3D de la personne, prédiction du mouvement labial, et augmentation de la résolution de l'image. Les résultats préliminaires montrent que notre méthode est au moins aussi performante que l'état de l'art. Nous proposons également un nouveau jeu de données francophone pour une utilisation concrète et un déploiement en entreprise.

Les principales contributions sont :

- l'utilisation du texte et de l'audio ;
- un nouveau jeu de données francophone ;
- des résultats encourageant à approfondir.

2 État de l'art

Génération vidéo conditionné par l'audio Dans les méthodes qui n'utilisent que l'entrée audio pour la génération, la collecte de données audio et vidéo pour l'entraînement et le ré-entraînement sont généralement nécessaires. RAD-NeRF [5] décompose la représentation du visage dans un espace de grande dimension en trois grilles de caractéristiques à faible dimension, ce qui permet de générer le visage en temps réel. En raison du manque d'in-

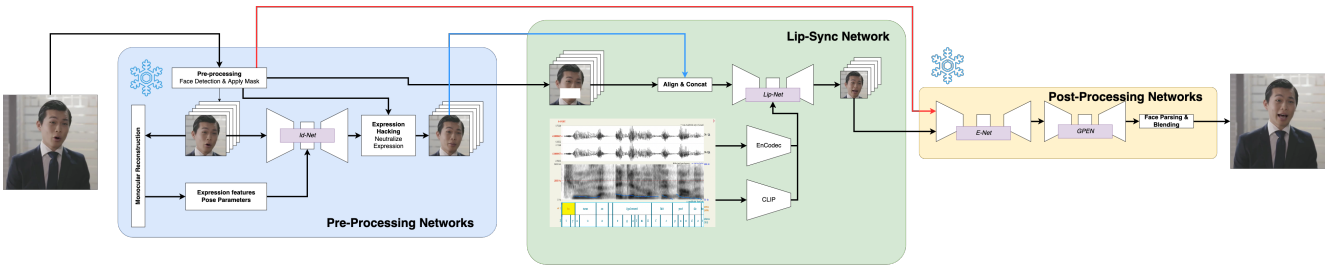


FIGURE 1 – *Architecture générale de la solution proposée.*

formations préalables, ces tâches peinent encore à rendre des expressions réalistes et des mouvements naturels. Plus récemment, SadTalker [6] propose un nouveau système conditionné par l'audio. À partir d'une seule image, en utilisant les coefficients 3D de la personne issus d'une réconstruction monoculaire 3D, il améliore la synchronisation des différents mouvements et la qualité vidéo. VideoReTalking [2] propose d'insérer une modèle 3D neutre pour préserver au mieux l'identité de la personne. Les coefficients 3D obtenus à partir d'une image sont modifiés pour faire paraître la personne sans expressions faciales marquées. Enfin, Hyperlips [7] utilise des hypergraphes pour mettre à jour les poids dans un réseau GAN. Cette méthode est proche des modèles de diffusion conditionnée par de l'audio. La méthode consistant à piloter l'audio avec la source vidéo est la plus poussée, car elle permet d'obtenir des expressions faciales suffisamment réalistes. Cependant, des défauts restent perceptibles visuellement.

Génération vidéo conditionnée par le texte Dans ce domaine, il y a eu plusieurs études sur la synthèse de vidéo à partir d'un texte. *Text-based Mouth Editing* [8] est une technique d'édition d'une vidéo existante avec une nouvelle entrée de texte. Cette méthode effectue une recherche de visèmes pour localiser les segments vidéo dont les mouvements de bouche correspondent au texte édité. Zhang et al. [9] ont proposé *Text2Video* pour animer des vidéos. Cette méthode est basée sur un dictionnaire de positions de phonèmes et ils ont entraîné un GAN pour générer des vidéos à partir de positions de phonèmes interpolées.

3 Notre approche

Notre approche multimodale utilise l'audio et le texte extrait pour prédire la partie inférieure du visage. Pour éviter les résultats indésirables causés par la corrélation entre le mouvement des lèvres et le discours dans les émissions TV, nous avons créé un réseau *lip-sync* qui prend en entrée une image masquée, l'image de la personne neutre, le nouvel audio souhaité et le texte extrait.

3.1 Caractéristiques faciales : 3DMM

Nous effectuons une détection du visage et des *landmarks* pour extraire uniquement le visage à modifier et mieux réintégrer la prédiction dans la vidéo d'origine. Nous extrayons les coefficients 3D de la personne à partir d'une image, que nous modifions pour obtenir un modèle 3D neutre reconverti en une image passée dans le réseau.

3.2 Caractéristiques audio

Plusieurs méthodes utilisent le spectogramme de Mel comme entrée audio. À partir de là, un réseau entraîné extrait des caractéristiques. Cependant, ces caractéristiques audio sont corrélées à la vidéo, limitant l'expressivité.
Pour réduire l'impact de cet entraînement, nous proposons d'utiliser un réseau pré-entraîné pour extraire ces caractéristiques. Pour cela, nous utilisons le réseau pré-entraîné EnCodec [3].
Le réseau encodeur prend le signal audio d'une durée d qui peut-être décrit par une séquence $x \in [-1, 1]^{C \times T}$ avec C le nombre de canaux et $T = d \times f_{sr}$ le nombre d'échantillons pour un taux d'échantillonnage f_{sr}.
Nous prenons 200 ms d'audio pour une frame. L'encodeur produit alors un vecteur caractéristique de taille 15 pour un signal à 24 kHz. Si l'audio d'entrée n'est pas à 24 kHz, il est ré-échantillonné. La configuration choisie correspond à l'encodeur à un bitrate de 24kbps pour un audio reconstruit à 24 kHz. Nous avons choisi un *bitrate* grand afin d'avoir le plus de caractéristiques audio. Dans cette configuration, en donnant un signal de 200 ms, les caractéristiques extraites sont une matrice $M_a \in \mathbb{R}^{15 \times 32}$. En effet, $15 = \frac{24000 * 0.2}{320}$ est le pas de temps sous-échantillonné et 32 est le nombre de quantificateurs du réseau. La matrice de sorti est convertie en un vecteur audio $f_a \in \mathbb{R}^{420 \times 1}$.

3.3 Caractéristiques textuelles

L'un des points clé de notre méthode est d'incorporer du texte, pouvant être extrait à partir de réseau *Speech-To-Text* [1]. Nous utilisons *Montreal-Forcer-Aligner* pour aligner le texte et l'audio [10]. Cela permet d'être sûr que chaque mot correspond bien au segment audio sélectionné. Nous extrayons les caractéristiques textuelles à l'aide du modèle CLIP [4]. Les mots prononcés pendant la séquence sélectionnée sont convertis en un vecteur puis le modèle CLIP génère un vecteur de caractéristiques textuelles $f_t \in \mathbb{R}^{512 \times 1}$.

3.4 Lip-Sync Network

Notre réseau est basé sur un cadre conditionnel reposant sur l'*inpainting*. Nous utilisons les images originales masquées et les caractéristiques audio et textuelles comme condition. La figure 1, montre l'architecture générale de

notre solution. La partie pré-traitement, qui consiste à l'extraction des caractéristiques 3D et leur neutralisation, est gelée. De même, pour les deux réseaux de post-traitement pour augmenter la résolution de l'image générée. Le réseau de génération de mouvement labiale est un réseau auto-encoder proche de celui de [2].

3.5 Fonctions de pertes

Nous avons entraîné le réseau sur une combinaison linéaire des fonctions suivantes : *perceptual loss*, *lip-sync discriminator* pour la qualité visuelle, *audio-visual synchronization* [11], la *L2-loss* sur le logarithme du spectre de Fourier réduit [12] et *L1-loss* dans le domaine spatial.
Soit $\mathcal{I}^{gt}$ l'images de référence, $\mathcal{I}^{lr}$ l'image générée. La *L1-loss* est définie comme :

$$\mathcal{L}_1 = ||\mathcal{I}^{gt} - \mathcal{I}^{lr}||_1$$

La *perceptual loss* est définie par

$$\mathcal{L}_{perc} = \sum_{l \in layers} ||f^l_{vgg}(\mathcal{I}^{gt}) - f^l_{vgg}(\mathcal{I}^{lr})||_2^2$$

où indique f^l_{vgg} le vecteur caractéristiques obtenus à la l-ième couche du réseau VGG19. La *audio-visual synchronization loss* est quant à elle définie par :

$$\mathcal{L}_{sync} = \frac{1}{N} \sum_{i=1}^{N} -\log P_{sync}$$

avec :

$$P_{sync} = \frac{v \cdot a}{\max(||v||_2, ||a||_2)}$$

où v et a sont respectivement les caractéristiques latentes du réseau SyncNet [11] de la vidéo et de l'audio.
Le discriminateur pour la qualité visuelle est donnée par la formule usuelle des GANs.
Enfin, la *L2-loss* du logarithme de la transformée de Fourier réduite est :

$$\mathcal{L}_{spec} = \frac{1}{\hat{H}} \sum_{k=0}^{\hat{H}-1} || \log\left(\tilde{S}(\mathcal{I}^{lr})\right)[k] - \log\left(\tilde{S}(\mathcal{I}^{gt})\right)[k]||_2^2$$

avec $\hat{H} = \frac{H}{\sqrt{2}}$, S est le carré de la magnitudes des composantes de Fourier et

$$\tilde{S}(r) = \frac{1}{2\pi} \int_0^{2\pi} S(r, \theta) \mathrm{d}\theta$$

La fonction de perte est une combinaison linéaire des fonctions introduites :

$$\mathcal{L} = \lambda_1 \mathcal{L}_1 + \lambda_2 \mathcal{L}_{perc} + \lambda_2 \mathcal{L}_{sync} + \lambda_3 \mathcal{L}_{spec} + \lambda_4 \mathcal{L}_{gan}$$

3.6 Post-traitement pour une haute résolution

La sortie du réseau est de taille 96×96 pixels. Afin d'augmenter la qualité et pour l'intégrer au mieux dans la vidéo d'origine, nous procédons à deux augmentations de résolution. Une première qui permet de préserver au mieux l'identité de la personne et une deuxième pour passer à la taille 512×512 pixels.

FIGURE 2 – *Exemples de visage du jeu de données French HDTF que nous avons constitué.*

TABLEAU 1 – *Détail du jeu de données francophone*

# Vidéos	# Personnes	Durée Totale
3529	382+	16H 23min 27s

4 Nouveau jeu de données

Le manque de jeux de données francophones limite la mise en production de ce type d'outils dans le monde francophone. C'est pourquoi nous avons créé un jeu de données francophones, nommé *French HDTF dataset*, en collectant des vidéos publiques de l'administration française des dernières années. Quelques exemples de ce jeu de données sont montrés dans la figure 2. Le jeu de données est résumé dans le tableau 1. Dans le cadre de l'édition visuelle, la détection de *landmarks* et du visage produisent une fenêtre d'au minimum 256×256 pixels. Avec la grande résolution des vidéos d'origine, il est possible d'aller jusqu'à des vidéos de visage de taille 512×512 pixels.

5 Expériences

Nous présentons le jeu de données d'entraînement, fournissons des détails d'implémentation et comparons les résultats quantitatifs avec des méthodes de l'état de l'art.

5.1 Jeu de données

Nous avons utilisé le jeu de données LRS2 [13] avec des vidéos de résolution 160p de différents programmes de la BBC. L'ensemble est traité en utilisant la détection des visages et en redimensionnant l'image d'entrée à 96×96 pixels. Les caractéristiques audio et textuelles sont déjà extraites.

5.2 Critères d'évaluation

Pour évaluer la méthode nous avons retenu les critères largement utilisés dans le domaine, à savoir : *Frechet Inception Distance* (FID), PNSR, SSIM, CPBD et *landmarks metric distance* (LMD). Pour la synchronisation labiale, nous l'évaluons avec le score LSE-D et LSE-C [11].

5.3 Résultats

Le tableau tableau 2 montre les résultats quantitatifs sur le jeu LRS2. Les résultats indiquent que notre approche est similaire à l'état de l'art, notamment pour le FID, CPBD

TABLEAU 2 – *Comparaison avec des méthodes de l'état de l'art sur le jeu LRS2 [13].*

Méthode	FID ↓	CPBD↑	PNSR ↑	SSIM ↑	LMD ↓	LSE-C ↑	LSE-D ↓
Wav2Lip	21.911	0.271	31.794	0.894	1.471	9.641	7.202
MakeItTalk	26.829	0.206	-	-	-	4.937	10.231
ATVGNet	-	-	32.812	0.871	1.984	4.610	8.445
PC-AVS	25.602	0.208	-	-	-	8.959	6.435
VideoReTalking	5.193	0.283	-	-	-	6.519	7.089
IP LAP	-	-	33.281	0.891	1.494	3.435	9.398
SadTalker	22.057	0.335	-	-	-	7.290	7.772
Our	22.432	0.290	32.914	0.820	1.203	5.939	8.193

et PSNR. Nous obtenons le meilleur résultat pour le LMD. Une explication serait la fonction de perte lié au domaine fréquentielle qui n'avait jamais été utilisée jusqu'alors. On peut noter que la méthode SadTalker [6] n'a pas fourni de score pour ce critère, mais il est probable qu'en le calculant les résultats soient meilleurs car prédiction des *landmarks* est incluse dans leur méthode. Le LSE-C est le meilleur pour Wav2Lip [11] car le réseau est entraîné pour ça. Les autres méthodes utilisent ce réseau dans leur fonction de perte de synchronisation audio-vidéo sans ré-entraînement. De même pour LSE-D, Wav2Lip est la meilleure méthode, à l'exception de deux méthodes VideoReTalking et PC-AVS. Cela s'explique par un ré-entraînement du réseau discriminateur permettant de baisser davantage ce score.

6 Conclusions

Ces résultats préliminaires nous encouragent à effectuer des comparaisons avec d'autres jeux de données tels que celui que nous avons constitué et HDTF [14]. Des études approfondies doivent également être menées, notamment une étude qualitative impliquant des personnes et des études dans lesquelles le module audio ou textuel est retiré pour évaluer l'impact respectif sur chaque module.

Références

[1] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, et Ilya Sutskever. Robust speech recognition via large-scale weak supervision. Dans *International conference on machine learning*, pages 28492–28518. PMLR, 2023.

[2] Kun Cheng, Xiaodong Cun, Yong Zhang, Menghan Xia, Fei Yin, Mingrui Zhu, Xuan Wang, Jue Wang, et Nannan Wang. Videoretalking : Audio-based lip synchronization for talking head video editing in the wild. Dans *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022.

[3] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, et Yossi Adi. High fidelity neural audio compression.

[4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. Dans *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[5] Jiahe Li, Jiawei Zhang, Xiao Bai, Jun Zhou, et Lin Gu. Efficient region-aware neural radiance fields for high-fidelity talking portrait synthesis. Dans *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7568–7578, 2023.

[6] Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, et Fei Wang. SadTalker : Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. type : article.

[7] Yaosen Chen, Yu Yao, Zhiqiang Li, Wei Wang, Yanru Zhang, Han Yang, et Xuming Wen. Hyperlips : Hyper control lips with high resolution decoder for talking face generation. *arXiv preprint arXiv :2310.05720*, 2023.

[8] Ohad Fried, Ayush Tewari, Michael Zollhöfer, Adam Finkelstein, Eli Shechtman, Dan B Goldman, Kyle Genova, Zeyu Jin, Christian Theobalt, et Maneesh Agrawala. Text-based editing of talking-head video. 38(4) :1–14.

[9] Sibo Zhang, Jiahong Yuan, Miao Liao, et Liangjun Zhang. Text2video : Text-driven talking-head video synthesis with personalized phoneme-pose dictionary. Dans *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2659–2663. IEEE, 2022.

[10] Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, et Morgan Sonderegger. Montreal forced aligner : Trainable text-speech alignment using kaldi. Dans *Interspeech*, volume 2017, pages 498–502, 2017.

[11] K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, et C.V. Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. Dans *Proceedings of the 28th ACM International Conference on Multimedia*, pages 484–492. ACM.

[12] Katja Schwarz, Yiyi Liao, et Andreas Geiger. On the frequency bias of generative models. *Advances in Neural Information Processing Systems*, 34 :18126–18136, 2021.

[13] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, et Andrew Zisserman. Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence*, 44(12) :8717–8727, 2018.

[14] Zhimeng Zhang, Lincheng Li, Yu Ding, et Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. Dans *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3661–3670, 2021.

Vers des mesures subtiles de performance en Réalité Virtuelle : l'exemple d'une tâche de mobilité

Yujie Huang[1] Alexandre Bruckert[1] Patrick Le Callet[1,2]

[1] Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes

[2] Institut universitaire de France (IUF)

{Yujie.Huang}@etu.univ-nantes.fr

{Alexandre.Bruckert, Patrick.Lecallet}@univ-nantes.fr

Résumé

De nombreuses pathologies de la vision peuvent significativement affecter la qualité de vie (Quality of Life, QoL) de patients, de part leurs impacts sur la façon d'effectuer diverses tâches visuelles. Il est important alors de distinguer les notions de fonction visuelle, c'est-à-dire les capacités ophtalmiques du système visuel (acuité visuelle, vision des couleurs, sensibilité au contraste), et de vision fonctionnelle, faisant référence à la capacité d'accomplir différentes tâches visuelles (lecture, navigation, reconnaissance d'objets, etc). Pour la première, il existe de nombreux outils d'évaluation quantitatifs standardisés en milieu clinique. Cependant, les outils d'évaluation actuels de la vision fonctionnelle sont soit trop subjectifs, soit non standardisés. Plusieurs protocoles d'évaluation de la vision fonctionnelle fondés sur une évaluation de l'orientation et de la mobilité (O&M) en environnement virtuel ont cependant récemment été proposés. Dans cette étude, nous proposons une étude de données issues d'un tel test, afin de mettre en évidence les avantages et inconvénients des protocoles existants. Nous constatons notamment des niveaux de difficulté variables dans différentes configurations environnementales, menant à un biais potentiel dans l'interprétation des résultats de ces tests. Nous avons également identifié des causes potentielles d'erreurs. En combinant ces résultats à une analyse de sujets sains et de patients utilisant des métriques précédentes, nous montrons que se fier uniquement aux indicateurs statistiques est insuffisant, soulignant le besoin de nouvelles méthodologies de traitement de ces données.

Mots clefs

Vision fonctionnelle, Réalité virtuelle, Qualité de vie, Orientation et mobilité, Modélisation.

1 Introduction

Les déficiences visuelles représentent un problème de santé publique majeur qui affecte la qualité de vie d'individus de tous âges [1]. En France, on estime que le nombre de personnes souffrant de déficiences visuelles dépasse 2% de la population totale [2], ce qui pose de nombreux défis en matière de soins médicaux et sociaux. Face à cette réalité, des investissements importants sont réalisés pour développer des dispositifs de remédiation, des prothèses, ou encore de nouvelles thérapies géniques prometteuses [3].

Cependant, un frein à l'adoption de telles thérapies est le manque d'outils pour mesurer objectivement la vision fonctionnelle. En effet, il est important de séparer les notions de fonction visuelle et de vision fonctionnelle. La fonction visuelle est liée à "la performance des yeux" [4], telle que l'acuité visuelle ou la sensibilité au contraste, et de nombreux outils de mesure standardisés existent. La vision fonctionnelle quant à elle est liée à "la performance de la personne dans les tâches visuelles" [4]. Les méthodes de mesure pour évaluer cette vision fonctionnelle présentent cependant certaines lacunes importantes : les questionnaires manquent d'objectivité [5] ; le test physique d'orientation et de mobilité (O&M) manque de reproductibilité [1], est coûteux et chronophage ; le test d'O&M basé sur la réalité virtuelle actuel nécessite un déplacement physique du patient [6, 7]. Nous avons donc proposé un test O&M basé sur la réalité virtuelle [8], qui présente trois avantages principaux par rapport aux tests précédents : la facilité de configuration de l'environnement, une sécurité accrue pour les participants et l'accès à plus de données comportementales pendant le test.

Dans cette étude, nous utilisons les données de sujets sains et de patients malvoyants recueillies lors de ce test. Sur la base des données des sujets sains, nous avons tout d'abord analysé la similarité des différents labyrinthes de test. Nous montrons que des différences de difficultés peuvent exister entre les différentes configurations de parcours. Nous avons ensuite étudié les caractéristiques des erreurs commises par les participants et avons constaté que la plupart des objets manqués avaient les valeurs RGB les plus basses. En examinant le nombre de points de fixation oculaires sur chaque objet manqué pendant le test, nous avons déduit l'existence de trois types d'erreurs différents. En ajoutant les données des patients, nous avons démontré que l'utilisation des seules métriques de temps et de nombre d'erreurs sur le parcours n'est pas suffisante pour évaluer la

vision fonctionnelle de manière précise. Il est nécessaire de combiner ces indicateurs avec des données issues du comportement des participants pour fournir une évaluation robuste.

2 Travaux connexes

La vision fonctionnelle est traditionnellement mesurée à l'aide de méthodes subjectives et objectives. Les méthodes subjectives reposent largement sur des auto-évaluations par les patients à travers des questionnaires ou des échelles de notation [5]. Ces méthodes sont faciles à mettre en œuvre mais ne produisent que des scores subjectifs, qui ne sont pas toujours fiables, et peuvent manquer de détails. Les méthodes objectives sont principalement basées sur le test d'Orientation et de Mobilité (O&M) [6]. Le test O&M le plus couramment utilisé aujourd'hui est le Multi-Luminance Mobility Test (MLMT) [1]. Pour représenter la performance du sujet, la durée du test et le nombre de collisions pendant le parcours sont enregistrés. Cependant, ces parcours de mobilité physique présentent plusieurs limitations en raison des conditions environnementales et de la difficulté de reproductibilité [1].

La réalité virtuelle permet quant à elle de concevoir des environnements divers et facilement modifiables à faible coût. Ainsi, sur la base du MLMT, une version basée sur la réalité virtuelle VR-O&M a été proposée et validée en utilisant le même système de notation que le MLMT [6]. Malgré l'avantage de la reproductibilité, cette méthode nécessite encore que les participants se déplacent physiquement. De plus, les systèmes de notation des méthodes actuelles sont basés sur des métriques simples, principalement la durée du test et le nombre d'erreurs. Par conséquent, il reste nécessaire d'évaluer ces outils ainsi que leurs limites, afin de pouvoir proposer un ensemble de méthodes d'analyse de données adapté.

Pour surmonter les lacunes des méthodes existantes, nous avons conçu un nouveau test O&M basé sur la réalité virtuelle [8]. Le test comprend huit labyrinthes différents, chacun ayant la même longueur de parcours et le même nombre d'objets (deux pour l'entraînement et six pour les tests). Un exemple est illustré dans la Figure 1. Lors de ce test, les participants restent assis pour garantir leur sécurité. Il leur est donné la consigne d'atteindre la sortie du labyrinthe, tout en détruisant les objets qu'ils détectent, en les touchant pendant 2 secondes. Dans notre test, le fait de ne pas détruire un objet est considéré comme une erreur. En utilisant cet environnement, nous avons collecté des données auprès de sujets sains et de patients (42 sujets sains, 9 patients). Sur la base de cette étude, nous avons mené plusieurs analyses préliminaires.

Tous les tests précédents d'orientation et de mobilité (O&M) ont utilisé des configurations de parcours différentes pour chaque test, ce qui est également le cas dans notre étude. Pour le test MLMT, ainsi que ses équivalents en VR actuels, le système de notation est défini comme :

$$TimeScore = t_{duration} + t_{penalties} \quad (1)$$

FIGURE 1 – *Structure d'un labyrinthe*

$$AccuracyScore = \frac{N_{penalties}}{N_{obstacles}} \quad (2)$$

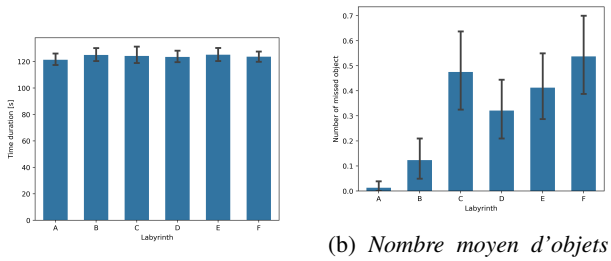
où $t_{penalties}$ était fixé arbitrairement à 15 secondes par erreur simple, c'est-à-dire par collision, déviation du parcours et erreurs de contournement ; et à 30 secondes pour les erreurs de redirection. Les scores dans chaque configuration ont ensuite été comparés empiriquement pour s'assurer qu'aucune n'était particulièrement facile ou difficile. Cependant, dans le test VR-O&M, la difficulté de chaque configuration n'a pas été vérifiée rigoureusement, ce qui pourrait entraîner des résultats d'analyse biaisés.

3 Analyse et résultat

Afin de garantir une comparabilité entre les différents participants et les sessions de tests, ainsi que pour éliminer l'effet de mémoire, il est important que les labyrinthes conçus soient de difficulté équivalente. Nous avons donc analysé et vérifié cette équivalence à partir des données préliminaires, en utilisant les données des sujets sains, plus nombreuses que celles des patients. Sur 479 sessions valides, chaque labyrinthe comptait en moyenne 80 $\pm$ 1.5 essais.

Nous utilisons deux métriques différentes pour cette analyse. En effet, le protocole de collecte étant différent par rapport à l'approche MLMT [6], nous ne pouvons pas directement appliquer le système de score utilisé lors de cette étude 1, 2. De plus, ces scores reposent sur une attribution de pénalité de temps arbitraire, peu indiquée dans l'objectif d'une méthode répliquable d'analyse. Nous nous concentrerons donc sur les variables brutes $t_{duration}$ et n_{error}. Comme illustré dans sur la Figure 2a, aucune différence significative de durée de parcours n'a été trouvée (Kruskal–Wallis test : P=0.771), indiquant une similarité entre toutes les configurations de labyrinthes. Ces résultats étaient attendus, étant donné la contrainte de similarité de longueur lors de la création des différents parcours.

Une autre variable que nous utilisons est le nombre d'objets manqués. Comme le montre la Figure 2b, une différence significative a été observée en termes de nombre d'objets manqués (Kruskal–Wallis test : P$\ll$0.001). Nous avons ensuite réalisé un test par paires de Dunn, dont les résultats indiquent que les participants ont commis significativement moins d'erreurs dans les Labyrinthes A et B par rapport aux Labyrinthes C, D, E et F. Cela démontre que malgré le maintien du même nombre d'objets, certaines configurations affectent significativement la difficulté de réussir à

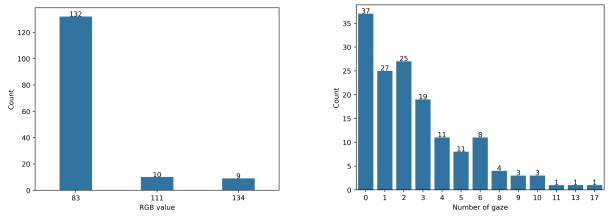


(a) *Durée moyenne du test pour chaque labyrinthe*

(b) *Nombre moyen d'objets manqués pour chaque labyrinthe*

FIGURE 2 – *Scores des participants sains sur les différents labyrinthes*

détecter puis détruire chaque objet.
Dans les tests précédents d'orientation et de mobilité (O&M), n_{error} est une métrique utilisée pour évaluer la performance. Cependant, nous avons constaté que parmi les sujets sains, il existe une différence significative dans le nombre d'erreurs à travers différents labyrinthes, qui n'est donc a priori pas due à une baisse de la vision fonctionnelle. Il est donc important de comprendre la cause de ces erreurs. Nous avons extrait certaines caractéristiques de tous les objets manqués, y compris les valeurs RGB, les niveaux de luminance et le nombre de regards portés sur ces objets pendant le test. Comme le montre la Figure 3a, la plupart (87%) des objets manqués ont la plus basse valeur RGB. De plus, comme indiqué dans la Figure 3b, la plupart (91%) des objets manqués ont été regardés moins de 8 fois. Il semble alors que, pour les sujets sains, les objets plus sombres sont plus susceptibles d'être manqués. Il est à noter que certains objets ont tout de même été manqués malgré un nombre élevé de points de fixations. Nous supposons que ces cas représentent un autre type d'erreur, probablement causé par une mauvaise utilisation de l'environnement de test. De plus, les objets avec des fixations autour de 10 fois pourraient indiquer ceux qui sont intrinsèquement difficiles à distinguer du fond.



(a) *RGB*

(b) *Nombre de fixations*

FIGURE 3 – *Nombre d'erreurs selon différentes caractéristiques*

Il semble alors nécessaire de distinguer les types d'erreurs. Les erreurs dues à l'inattention et à une mauvaise utilisation de l'environnement de test ne sont pas liées à la vision fonctionnelle, et doivent être exclues dans son l'évaluation. En revanche, les erreurs dues à la difficulté peuvent être utilisées pour évaluer la gravité de l'état du patient. Comme mentionné précédemment, les différentes configurations de nos six labyrinthes de test entraînent une différence significative dans le nombre d'erreurs. Par conséquent, identifier les configurations qui entraînent des types spécifiques d'erreurs sera notre direction de recherche future. Une fois que nous aurons compris quelles configurations provoquent quelles erreurs, nous pourrons modifier la difficulté de notre environnement de test en manipulant ces configurations. Cela aidera à réduire l'occurrence d'erreurs non pertinentes et à améliorer le pouvoir discriminant de l'environnement de test en ajustant la difficulté du labyrinthe.

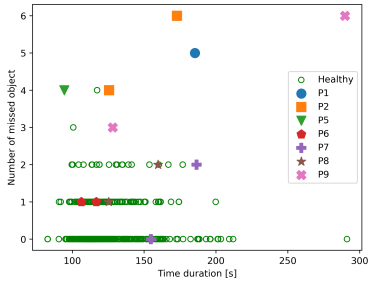


FIGURE 4 – *Durée et nombre d'objets manqués par les sujets sains et les patients*

Bien que nous disposions seulement de 12 résultats de tests disponibles dans le groupe de patients malvoyants (7 sur 9 patients ayant réussi à compléter au moins un test après l'entraînement), nous visons toujours à évaluer le pouvoir discriminatoire du test. Comme le montre la Figure 4, en utilisant uniquement la durée et le nombre d'objets manqués, certains patients comme P1 et P2 peuvent être séparés du reste de la cohorte. D'autres patients, comme P5, P7 et P8, ont réalisé des performances similaires aux sujets sains en termes de durée et de nombre d'objets manqués. Ce résultat montre que nous ne pouvons pas déterminer si une personne a une vision fonctionnelle dégradée basée uniquement sur ces deux métriques courantes. Nous émettons l'hypothèse que ces résultats peuvent être biaisés par des erreurs dues à l'inattention et à une mauvaise utilisation de l'environnement de test.

4 Conclusion et perspective

Dans cette étude, nous avons exploité les données collectées à partir de notre test O&M basé sur la réalité virtuelle pour à la fois des sujets sains et des patients malvoyants. Nos analyses initiales ont porté sur la comparabilité des différents labyrinthes de test en utilisant les données des sujets sains. Les résultats ont indiqué que les labyrinthes sont comparables en termes de temps nécessaire pour compléter le test, mais non comparables en termes de nombre d'erreurs. Les différentes configurations du parcours ont causé cette différence de difficulté entre les labyrinthes.
Des analyses supplémentaires des caractéristiques des objets manqués ont révélé que la plupart avaient les va-

leurs RGB les plus basses. En examinant le nombre de regards portés sur chaque objet manqué pendant le test, nous avons identifié trois types différents d'erreurs : manquement d'objet par négligence, par mauvaise utilisation de l'environnement de test, et en raison de la difficulté d'observation. Nous avons illustré l'importance de distinguer les types d'erreurs afin d'éliminer le biais des erreurs non liées à la vision.

En intégrant les données des patients malvoyants, nous avons démontré que l'utilisation uniquement des métriques de temps et de précision, comme dans les tests précédents, est insuffisante pour une évaluation nuancée de la vision fonctionnelle. Des métriques d'évaluation plus robustes devraient être proposées.

Étant donné la richesse de nos données comportementales, les analyses futures se concentreront sur l'explication des différents comportements et interactions dans l'environnement virtuel. Deux approches peuvent être envisagées. Les modèles stochastiques ont le potentiel d'extraire des caractéristiques ou de découvrir des motifs sous-jacents. Par exemple, Shaily et al. [9] ont proposé l'utilisation de Modèles de Markov Cachés (HMMs) pour les données séquentielles dans la reconnaissance des activités humaines, tandis que Ben-Gal et al. [10] ont employé des chaînes de Markov pour la modélisation de la mobilité. Les méthodes basées sur les graphes sont également efficaces pour la représentation des données. Rossi et al. [11] ont utilisé une approche basée sur les graphes pour regrouper les sujets ayant des patterns de navigation similaires en réalité virtuelle. Avec ces diverses stratégies de modélisation, des méthodes d'apprentissage associées peuvent être explorées plus en profondeur.

À l'avenir, les tests d'orientation et de mobilité basés sur la réalité virtuelle pourraient ne pas se limiter aux tâches traditionnelles de "navigation". Les chercheurs pourraient explorer le potentiel d'utiliser des tâches plus immersives et liées au mode de vie, telles que la cuisine ou la conduite, pour tirer pleinement parti des capacités de la réalité virtuelle.

Références

[1] Daniel C Chung, Sarah McCague, Zi-Fan Yu, Satha Thill, Julie DiStefano-Pappas, Jean Bennett, Dominique Cross, Kathleen Marshall, Jennifer Wellman, et Katherine A High. Novel mobility test to assess functional vision in patients with inherited retinal dystrophies. *Clinical & experimental ophthalmology*, 46(3) :247–259, 2018.

[2] Seth R Flaxman, Rupert RA Bourne, Serge Resnikoff, Peter Ackland, Tasanee Braithwaite, Maria V Cicinelli, Aditi Das, Jost B Jonas, Jill Keeffe, John H Kempen, et al. Global causes of blindness and distance vision impairment 1990–2020 : a systematic review and meta-analysis. *The Lancet Global Health*, 5(12) :e1221–e1234, 2017.

[3] Albert M Maguire, Katherine A High, Alberto Auricchio, J Fraser Wright, Eric A Pierce, Francesco Testa, Federico Mingozzi, Jeannette L Bennicelli, Guishuang Ying, Settimio Rossi, et al. Age-dependent effects of rpe65 gene therapy for leber's congenital amaurosis : a phase 1 dose-escalation trial. *The Lancet*, 374(9701) :1597–1605, 2009.

[4] August Colenbrander. Assessment of functional vision and its rehabilitation. *Acta ophthalmologica*, 88(2) :163–173, 2010.

[5] Vijaya K Gothwal, Jan E Lovie-Kitchin, et Rishita Nutheti. The development of the lv prasad-functional vision questionnaire : a measure of functional vision performance of visually impaired children. *Investigative ophthalmology & visual science*, 44(9) :4131–4139, 2003.

[6] Tomas S Aleman, Alexander J Miller, Katherine H Maguire, Elena M Aleman, Leona W Serrano, Keli B O'Connor, Emma C Bedoukian, Bart P Leroy, Albert M Maguire, et Jean Bennett. A virtual reality orientation and mobility test for inherited retinal degenerations : testing a proof-of-concept after gene therapy. *Clinical Ophthalmology*, pages 939–952, 2021.

[7] Jean Bennett, Elena M Aleman, Katherine H Maguire, Jennifer Nadelmann, Mariejel L Weber, William M Maguire, Ayodele Maja, Erin C O'Neil, Albert M Maguire, Alexander J Miller, et al. Optimization and validation of a virtual reality orientation and mobility test for inherited retinal degenerations. *Translational Vision Science & Technology*, 12(1) :28–28, 2023.

[8] Audrey Crozet, Lucas Communier, Toinon Vigier, Pierre Lebranchu, et Patrick Le Callet. A virtual mobility test to evaluate functional vision of visual impaired patients. Dans *IMXw'23 : ACM International Conference on Interactive Media Experiences Workshops*. ACM, 2023.

[9] Shagun Shaily et Veenu Mangat. The hidden markov model and its application to human activity recognition. Dans *2015 2nd International Conference on Recent Advances in Engineering & Computational Sciences (RAECS)*, pages 1–4. IEEE, 2015.

[10] Irad Ben-Gal, Shahar Weinstock, Gonen Singer, et Nicholas Bambos. Clustering users by their mobility behavioral patterns. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 13(4) :1–28, 2019.

[11] Silvia Rossi, Francesca De Simone, Pascal Frossard, et Laura Toni. Spherical clustering of users navigating 360 content. Dans *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4020–4024. IEEE, 2019.

Limitations de la crypto-compression de vidéos HD

Erwan Reinders[1,2] Pauline Puteaux[3] Samuel Brau[2] William Puech[1]

[1] LIRMM, Univ. Montpellier, CNRS, Montpellier, France
[2] DroneGeofencing, Nîmes, France
[3] Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France

Résumé

Dans les scénarios de vidéo surveillance, il est essentiel de sécuriser le contenu visuel pendant la transmission et le stockage. Comme ces séquences vidéo peuvent représenter de gros volumes de données, il est également nécessaire de les compresser. De plus, les capteurs vidéo évoluent constamment, offrant des vidéos à des résolutions de plus en plus élevées. Le codec H.264 est une norme de compression vidéo largement utilisée dans la vidéo surveillance, même aujourd'hui. Dans cet article, nous proposons une analyse des différentes techniques de crypto-compression basées sur le codec H.264, compte tenu de cette évolution de la résolution vidéo. Nous montrons qu'il est quantitativement difficile de trouver des différences entre les vidéos crypto-compressées à basse résolution et à haute résolution, alors que visuellement, une plus grande partie du contenu original est reconnaissable à haute résolution.

Mots clefs

Sécurité multimédia, analyse de la crypto-compression vidéo, évolution de la résolution, sécurité visuelle.

1 Introduction

Les capteurs vidéo des drones modernes fournissent des vidéos en haute définition. Pour en optimiser la transmission, une étape de compression est nécessaire. Dans le but d'être conforme aux normes internationales de compression, la norme d'encodage vidéo H.264 reste le format le plus utilisé pour ce type d'application [1, 2]. Selon la nature de ce que le drone survole, il est essentiel de sécuriser cette transmission vidéo par du chiffrement. Le but du chiffrement vidéo est de s'éloigner du contenu vidéo original, de sorte qu'aucune information visuelle ne puisse en être déduite, tandis que le but de la compression vidéo est de réduire la quantité de données utilisées pour décrire le contenu visuel. Lorsqu'un contenu visuel doit être compressé et chiffré, le chiffrement peut être effectué avant, pendant ou après la compression. En 2018, Chuman *et al.* ont proposé une méthode ETC (Encryption Then Compression) pour chiffrer une image avant de la transmettre par un canal de communication qui effectue une compression JPEG (comme les réseaux sociaux) [3]. Pour réaliser le chiffrement, l'image est découpée en bloc de 8×8 pixels. Ces blocs sont ensuite mélangés entre eux et transformés (rotation, inversion).
Lorsque le chiffrement est effectué après la compression, le contenu compressé est considéré comme un flux binaire. Les méthodes de chiffrement standard, telles que AES par exemple, peuvent être appliquées à l'ensemble de ce flux binaire issu de l'étape de compression [4]. Dans ce cas, le chiffrement modifie la sémantique et la syntaxe des données compressées, qui ne peuvent plus être décompressées.

Enfin, la compression et le chiffrement peuvent être réalisés conjointement, dans ce que l'on appelle les méthodes de crypto-compression. En 2011, Shahid *et al.* ont proposé deux nouvelles méthodes de crypto-compression basées sur AES (en mode CFB) et sur les deux différents codeurs entropiques de la norme H.264, appelées SE-CAVLC et SE-CABAC [5]. En 2013, Dubois *et al.* ont présenté une nouvelle méthode de crypto-compression, basée sur la méthode SE-CAVLC. Cette méthode permet d'ajuster le nombre de coefficients à chiffrer dans une frame, pour la même sécurité visuelle [6]. Pour ce faire, une nouvelle mesure, appelée TSSIM, est proposée. Cette métrique mesure le SSIM [7] de la différence absolue entre une frame originale et la frame précédente, avec la différence absolue entre ces deux mêmes frames mais après chiffrement. Plus la valeur du TSSIM est petite, plus la sécurité visuelle est grande.

Parallèlement à cela, les capteurs vidéo ont considérablement évolué, permettant de produire des vidéos à plus haute résolution. La résolution vidéo est passée de formats QCIF (176×144 pixels) ou CIF (352×288 pixels) à des formats HD (720p : 1280×720 pixels, 1080p : 1920×1080 pixels), voire UHD comme la 4K (2160×3840 pixels).

Dans cet article, nous proposons d'analyser l'évolution de la sécurité visuelle des méthodes de crypto-compression vidéo H.264 appliquées à des vidéos de résolutions de plus en plus élevées. Dans la section 2, nous analysons et dessinons les limites du chiffrement vidéo par rapport à l'évolution de la qualité vidéo. Nous présentons ensuite dans la section 3 une analyse expérimentale de la méthode de crypto-compression de Shahid *et al.* [5]. Enfin, nous concluons en section 4.

2 Analyse théorique et limitation des méthodes de chiffrement vidéo

Dans la section 2.1 nous présentons l'évolution de la résolution vidéo, tandis que dans la section 2.2 nous donnons un aperçu des éléments syntaxiques produits par le codeur H.264. Enfin, dans la section 2.3, nous détaillons la méthode de l'état de l'art sur laquelle nous nous concentrons dans nos expériences.

2.1 Évolution de la résolution des vidéos

Comme illustré figure 1, la résolution des données vidéos a considérablement augmentée depuis la norme CIF (Common Intermediate Format), qui offrait une résolution de 352×288 pixels, ou même QCIF, avec une résolution de 176×144 pixels, vers des normes beaucoup plus élevées, telles que la Full HD (1920×1080 pixels), la 4K (3840×2160 pixels), et même la 8K (7680×4320 pixels). Le format CIF était largement utilisé dans les premiers systèmes de vidéo surveillance et vidéo conférence.

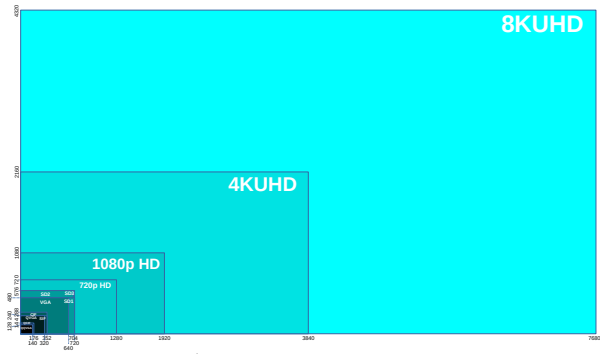


FIGURE 1 – Évolution de la résolution des vidéos.

Dans le cas des vidéos de drones, plus la qualité de la vidéo est bonne et plus celle-ci peut être utilisée ultérieurement (reconnaissance de personnes, suivi des cibles). Cependant, quelle que soit la résolution de la vidéo d'entrée, l'encodeur vidéo H.264 ne traite que les blocs de 16×16 pixels, appelés des macroblocks (MB). Cela signifie que pour une même vidéo enregistrée à différentes résolutions, un MB, en termes de pixels, contient plus d'informations visuelles globales et moins de détails dans une vidéo à basse résolution que dans une vidéo à haute résolution.

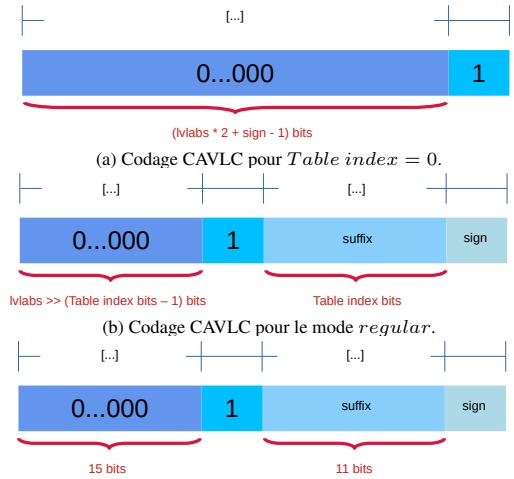
(a) Frame QCIF. (b) Frame HD 1080p.

FIGURE 2 – Contenu d'un MB de 16×16 pixels selon la résolution vidéo. La même frame en résolution : a) QCIF, le MB contient une grande partie de la tête, b) HD 1080p, le MB contient seulement un oeil de la même tête.

La figure 2 illustre un exemple du contenu d'un MB en fonction de la résolution vidéo. Nous constatons que pour la même vidéo, un MB QCIF (figure 2a) couvre plus de contenu qu'un MB 1080p (figure 2b), même si les deux MB contiennent le même nombre de pixels (16×16).

2.2 Éléments syntaxiques dans H.264

L'encodeur vidéo H.264 est composé de différentes étapes, telles que la prédiction, la transformation, la quantification et le codage. Pour obtenir un flux binaire compressé, trois redondances sont exploitées : les redondances spatiales (au sein d'une même frame), temporelles (entre deux frames différentes) et statistiques (codage entropique avec CAVLC ou CABAC).



(a) Codage CAVLC pour $Table\ index = 0$.

(b) Codage CAVLC pour le mode $regular$.

(c) Codage CAVLC pour le mode $escape$.

FIGURE 3 – Schémas du codage des NZ dans CAVLC.

Le codage CAVLC est une méthode de codage adaptative, variable en taille et à tables fixes. Ces tables VLC sont définies arbitrairement dans le codeur et leur choix dépend de seuils fixes. Pour chaque sous-MB de 4×4 pixels d'un MB à coder, les coefficients prédits/transformés/quantifiés sont d'abords ré-ordonnées, de sorte à regrouper les coefficients de même fréquence entre eux. On peut alors dissocier deux grands types de coefficients à coder : les coefficients nuls (codés par RLE) et les coefficients non nuls (ou NZ). Tous les NZ consécutifs égaux à ± 1 en fin d'ordre, au maximum des trois derniers, sont considérés comme des $trailing - ones$, et seul leur signe et leur nombre sont codés. La figure 3 illustre le codage des coefficients NZ restant, avec $lvlabs$ la valeur absolue du coefficient à coder par CAVLC, $sign$ pour le signe de ce coefficient (0 pour positif, 1 pour négatif), et $Table\ index$ faisant référence à la table de longueur fixe utilisée. $Table\ index$ est réévalué pour chaque nouveau NZ à coder dans le sous-MB, et qui n'est pas un $trailing\text{-}one$. La figure 3a montre le processus de codage du coefficient pour $Table\ index = 0$ (le premier coefficient du sous-MB à coder, sous certaines conditions). La figure 3b montre le processus de codage pour toutes les autres valeurs de $Table\ index$, et la figure 3c montre le processus de codage lorsque la valeur du coefficient à coder est supérieure à $15 \ll (Table\ index - 1)$, à l'exception $Table\ index = 0$, où le seuil est de 8.

2.3 Analyse de la résolution sur une méthode spécifique de crypto-compression

En 2011, Shahid *et al.* ont proposé deux méthodes de crypto-compression pour H.264 AVC, appelées SE-CAVLC et SE-CABAC, pour les deux codeurs présents dans H.264 : CAVLC et CABAC [5]. Dans la méthode SE-CAVLC, les coefficients codés dans la sortie du codeur CAVLC sont chiffrés. Pour ce faire, comme l'illustre la figure 4, l'encodeur vidéo prédit d'abord les pixels d'un

MB, par rapport à la même image ou aux voisins temporels (images précédentes ou suivantes), puis transforme uniquement le résultat de cette prédiction à l'aide d'une DCT 4×4 entière, quantifie ces pixels prédits transformés, et les réordonne afin de regrouper les coefficients à coder par fréquences.

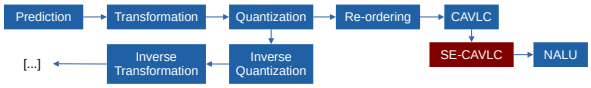


FIGURE 4 – Schéma général de la méthode SE-CAVLC [5].

Dans SE-CAVLC, l'espace de chiffrement est défini comme les valeurs qui ont la même longueur de code par rapport à la table VLC utilisée ($Table\ index$) pour le codage de ce coefficient. Pour ce faire, les coefficients dont $Table\ index = 0$ ne sont pas chiffrés (figure 3a), car seul le suffixe des coefficients codés est chiffré (figure 3b et 3c). Cela permet de maintenir la conformité du format, car sinon la table VLC à sélectionner pour encoder les coefficients suivants pourrait être différente, ce qui introduirait des erreurs de lecture potentielles lors du décodage. Comme c'est la valeur absolue du coefficient qui est prise en compte lors du choix de la table VLC, le signe du coefficient codé est également chiffré. Enfin, le signe des $trailling\text{-}ones$ est chiffré, ce qui n'interfère pas avec le décodage du flux binaire.

3 Analyse expérimentale

FIGURE 5 – Frame $\#45$ de la vidéo $Office$.

La frame brute $\#45$ (figure 5) compressée avec $QP = 18$ produit, pour les résolutions QCIF (176×144), CIF (352×288), HD 720p (1280×720) et HD 1080p (1280×1080) les images illustrées figure 6a, figure 6c, figure 6e et figure 6g respectivement. Dans la figure 6b, la frame $\#45$ a été crypto-compressée en résolution QCIF, tandis que dans la figure 6d la frame $\#45$ a été crypto-compressée en résolution CIF. En augmentant la résolution, dans la figure 6f et figure 6h, nous pouvons voir la même image crypto-compressée pareillement pour des résolution HD 720p et HD 1080p, respectivement. Nous observons, à partir de la même méthode de crypto-compression, que les résultats sont visuellement différents en fonction de la résolution.

La figure 7 illustre le même processus avec $QP = 36$. Nous pouvons observer que plus la quantification est importante et plus l'impact de la crypto-compression sur la sécurité visuelle est faible, et plus la résolution de la vidéo crypto-compressée est élevée et plus il est difficile de maintenir un niveau suffisant de sécurité visuelle. En réalisant une analyse sur les 100 premières frames de la vidéo $Office$ (tableau 1), Nous obtenons un PSNR moyen

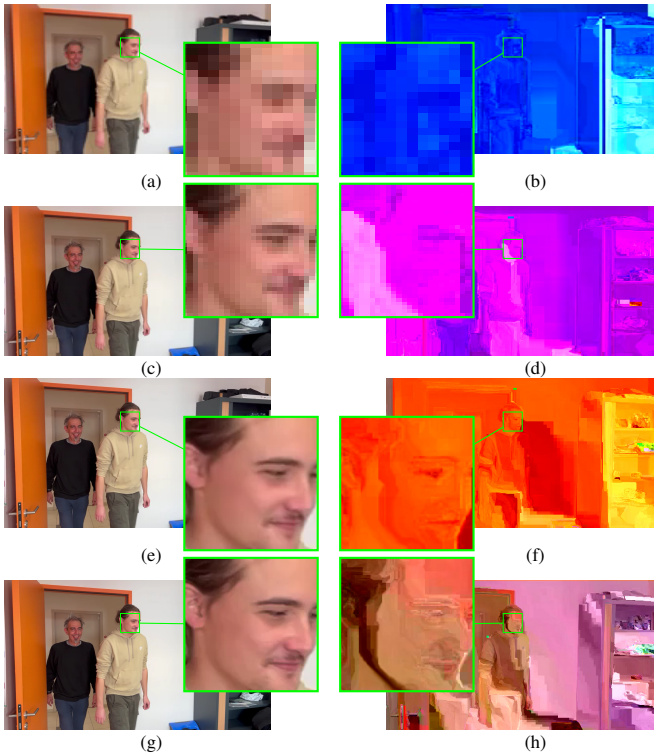


FIGURE 6 – Résultats de la compression et crypto-compression (à différentes résolutions) sur la frame $\#45$ de la vidéo Office avec $QP = 18$: frames compressées en 1^{ere} colonne, et frames crypto-compressées en 2^{eme}. a) et b) En résolution QCIF (176×144), c) et d) En CIF (352×288), e) et f) En HD 720p (1280×720) et g) et h) En HD 1080p (1920×1080).

entre la vidéo originale et les versions crypto-compressées ($QP = 18$) de $8,985\ dB$ pour la vidéo QCIF, $9,577\ dB$ pour la vidéo CIF, $9,442\ dB$ pour la vidéo HD 720p, et $9,753\ dB$ pour la vidéo HD 1080p. Nous ne constatons donc pas d'augmentation réelle du PSNR à mesure que la résolution augmente, alors qu'entre la vidéo originale et les versions compressées, l'augmentation est plus significative, avec $46,123\ dB$ pour la vidéo QCIF, $47,861\ dB$ pour la vidéo CIF, $49,622\ dB$ pour la vidéo HD 720p, et $50,444\ dB$ pour la vidéo HD 1080p. La même analyse est effectuée pour $QP = 36$. Nous pouvons constater qu'entre la vidéo originale et les versions crypto-compressées, les valeurs de PSNR et de SSIM sont similaires pour toutes les résolutions vidéo.

Dans le tableau 2, nous présentons les résultats des métriques UACI et NPCR, entre la vidéo $Office$ originale et celle crypto-compressée, pour les 100 premières frames. Nous pouvons constater que les valeurs de ces métriques sont similaires entre les résolutions, et entre les QPs.

Dans la figure 8, l'algorithme de détection des contours de Canny est appliqué à la composante Y de la frame $\#45$ de la vidéo $Office$ ($\sigma = 0, s_{min} = 25, s_{max} = 50$). Sur les images compressées de la première colonne ($QP = 18$), nous observons que les principales informations visuelles sont accentuées par le filtre de Canny : les deux protagonistes entrant dans le bureau, le cadre de la porte et l'étagère sur la droite. Ces informations sont accentuées en

FIGURE 7 – Résultats de la compression et crypto-compression (à différentes résolutions) sur la frame #45 de la vidéo Office avec $QP = 36$: frames compressées en 1^{ere} colonne, et frames crypto-compressées en 2^{eme}. a) et b) En résolution QCIF (176×144), c) et d) En CIF (352×288), e) et f) En HD 720p (1280×720) et g) et h) En HD 1080p (1920×1080).

QP	Métriques		QCIF	CIF	HD 720p	HD 1080p
18	PSNR (dB)	$OvsC$	46, 123	47, 861	49, 622	50, 444
		$OvsCC$	8, 985	9, 577	9, 442	9, 753
	SSIM	$OvsC$	1	1	1	0, 999
		$OvsCC$	0, 052	0, 057	0, 099	0, 115
36	PSNR (dB)	$OvsC$	32, 571	35, 521	38, 255	39, 599
		$OvsCC$	11, 376	12, 508	11, 915	11, 646
	SSIM	$OvsC$	0, 995	0, 996	0, 999	0, 999
		$OvsCC$	0, 266	0, 295	0, 338	0, 326

TABLEAU 1 – PSNR et SSIM entre la vidéo $Office$ originale et les versions compressées ($OvsC$) et crypto-compressées ($OvsCC$) à des résolutions différentes (100 premières frames).

QP	Métriques	QCIF	CIF	HD 720p	HD 1080p
18	UACI	0, 303	0, 285	0, 281	0, 271
	NPCR	0, 995	0, 994	0, 994	0, 993
36	UACI	0, 223	0, 199	0, 209	0, 217
	NPCR	0, 994	0, 993	0, 993	0, 993

TABLEAU 2 – UACI et NPCR entre la video $Office$ originale et les versions crypto-compressées à des résolutions différentes (100^{ere} frames).

basse résolution, comme QCIF et CIF (1^{ere} et 2^{eme} lignes respectivement), et en haute résolution, comme HD 720p et HD 1080p (3^{eme} et 4^{eme} lignes respectivement). Lorsqu'une étape de crypto-compression est effectuée (dernière

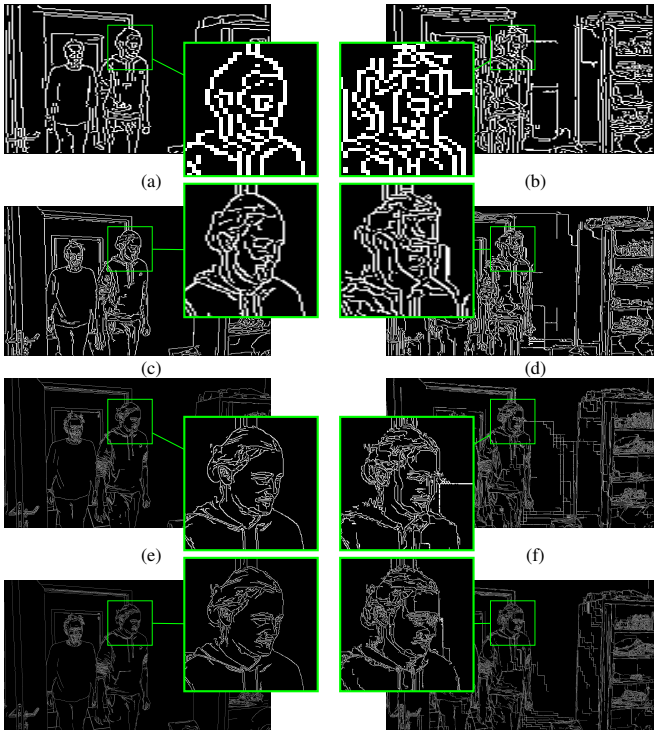


FIGURE 8 – Détection des contours de Canny sur la frame #45 de la vidéo $Office$: frames compressées en 1^{ere} colonne, frames crypto-compressées en 2^{eme}. La 1^{ere} ligne sont des frames en résolution QCIF (176×144), la 2^{eme} en CIF (352×288), la 3^{eme} en HD 720p (1280×720) et la 4^{eme} en HD 1080p (1920×1080).

colonne avec le même $QP = 18$), le résultat est conforme aux analyses précédentes. Sans connaître la vidéo originale, il est difficile de comprendre en basse résolution la construction visuelle de la scène vidéo crypto-compressée. Cependant, plus la résolution est élevée, plus les informations visuelles crypto-compressées sont compréhensibles. Ce phénomène est directement lié à la nature des données chiffrées lors de l'étape de compression.

4 Conclusion

Dans cet article, nous proposons une analyse des méthodes de crypto-compression H.264 sur des vidéos à haute résolution, et démontrons les limites de ces méthodes en termes de sécurité visuelle. Ces limites proviennent principalement de la taille limitée d'un MB dans l'encodeur vidéo. Si, d'un point de vue statistique, une vidéo crypto-compressée semble sûre, quelle que soit sa résolution, ce n'est pas le cas visuellement. Dans cette optique, il serait intéressant de calculer une métrique permettant de mieux mesurer la sécurité visuelle, autrement qu'au moyen d'une métrique appliquée image par image. Cette mesure peut être basée sur une analyse préalable de l'image, telle que la détection des contours. En outre, une seule image vidéo peut parfois suffire au système visuel humain pour comprendre l'ensemble du contenu d'une vidéo crypto-compressée. Il est donc essentiel de prendre en compte l'aspect de l'intégration temporelle afin de sécuriser une vidéo.

Références

[1] ITU Telecom. Advanced video coding for generic audiovisual services. *ITU-T Recommendation H. 264*, 2003.

[2] Thomas Wiegand, Gary J Sullivan, Gisle Bjontegaard, et Ajay Luthra. Overview of the H.264/AVC video coding standard. *IEEE Transactions on circuits and systems for video technology*, 13(7) :560–576, 2003.

[3] Tatsuya Chuman, Warit Sirichotedumrong, et Hitoshi Kiya. Encryption-then-compression systems using grayscale-based image encryption for JPEG images. *IEEE Transactions on Information Forensics and security*, 14(6) :1515–1525, 2018.

[4] Vincent Rijmen et Joan Daemen. Advanced encryption standard. *Proceedings of federal information processing standards publications, national institute of standards and technology*, 19 :22, 2001.

[5] Zafar Shahid, Marc Chaumont, et William Puech. Fast protection of H.264/AVC by selective encryption of CAVLC and CABAC for I and P frames. *IEEE Transactions on Circuits and Systems for Video Technology*, 21(5) :565–576, 2011.

[6] Loïc Dubois, William Puech, et Jacques Blanc-Talon. Confidentiality metrics and smart selective encryption for HD H.264/AVC videos. Dans *21st European Signal Processing Conference (EUSIPCO 2013)*, pages 1–5. IEEE, 2013.

[7] Zhou Wang, A.C. Bovik, H.R. Sheikh, et E.P. Simoncelli. Image quality assessment : from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4) :600–612, 2004.

Towards Light-weight Transformer-based Quality Assessment Metric for Augmented Reality

Aymen Sekhri[1,2] Seyed Ali Amirshahi[2] Mohamed-Chaker Larabi[1]

[1] CNRS, Université de Poitiers, XLIM, Poitiers, France

[2] Norwegian University of Science and Technology, Gjøvik, Norway

{aymen.sekhri, chaker.larabi}@univ-poitiers.fr, s.ali.amirshahi@ntnu.no

Abstract

This Paper introduces transformAR, a lightweight transformer-based model for objective quality assessment in AR applications. This approach utilizes pre-trained vision transformer-based encoders to capture image content information, computes distance vectors for quantifying distortions, and employs cross-attention-based decoders to model perceptual quality features. The model integrates adapted regularization techniques and label smoothing to mitigate overfitting. Experimental results demonstrate the effectiveness of transformAR, surpassing existing state-of-the-art methods.

Keywords

Augmented Reality, Vision Transformer, Image Processing, Image Quality Assessment.

1 Introduction

Augmented Reality (AR) overlays computer-generated information onto the real world via devices like smartphones and head-mounted displays, enhancing experiences in fields such as navigation, education, entertainment, and healthcare [1]. Ensuring high Quality of Experience (QoE) requires objective quality assessment methods that account for various factors impacting visual perception [2, 3]. Image quality is crucial for QoE [4], typically measured through complex and time-consuming psychophysical experiments, resulting in scarce subjective datasets for AR Image Quality Assessment (AR-IQA).

As most previous studies addressed geometric and textural degradation in 3D meshes and point clouds [5, 6]. However, Duan et al. [7] introduced CFIQA (Confusing Image Quality Assessment) and ARIQA datasets to simulate AR scenarios. They demonstrated that classical 2D metrics like PSNR, SSIM [8], and VIF [9] are ineffective for AR, necessitating advanced metrics. They explored LPIPS [10] using CNN-based feature extractors like SqueezeNet [11], AlexNet [12], and VGG [13]. Duan et al. also proposed CFIQA, a CNN-based model using VGG and ResNet [14]. This model processes features from reference and superimposed images at each convolution layer, generating distance feature maps that are refined by channel and spatial attention mechanisms to predict quality scores. The ARIQA model extends this approach with two superimposed images from the same reference but different quality scores. ARIQA+ incorporates edge detection features. However, CNN-based models face limitations due to local connectivity and translation invariance, restricting their ability to capture global patterns [15].

To address these limitations, we propose a transformer-based AR quality assessment metric, therefore, our contributions include :

- Adapting a lightweight encoder-decoder transformer framework to capture global quality features with minimal data.
- Introducing label smoothing for quality scores to reduce model overconfidence.
- We account for perceptual confusion by feeding the model with background and foreground images in addition to the fused content.

This approach mimics human observers by considering both global and localized regions for accurate quality perception [16].

2 Proposed Method

Our approach, transformAR, adapts the Vision Transformer (ViT) architecture [17, 18] for AR quality assessment. It comprises content-aware encoders, quality-aware decoders, and regressors. Below, we provide an overview of each component.

2.1 Content-aware Encoders :

We use three frozen ViT encoders with self-supervised pretrained weights via a method called, DINO [19]. These encoders capture global content information from both reference and distorted images. ViT divides an input image into non-overlapping patches, which are then processed into vectors. Using self-attention, ViT focuses on different parts of the image, enhancing feature extraction [18].

Due to data scarcity, using the full 12 transformer blocks in DINO led to overfitting. We found that using only the first two blocks was sufficient to map the input image into useful representations for quality assessment. The dataset

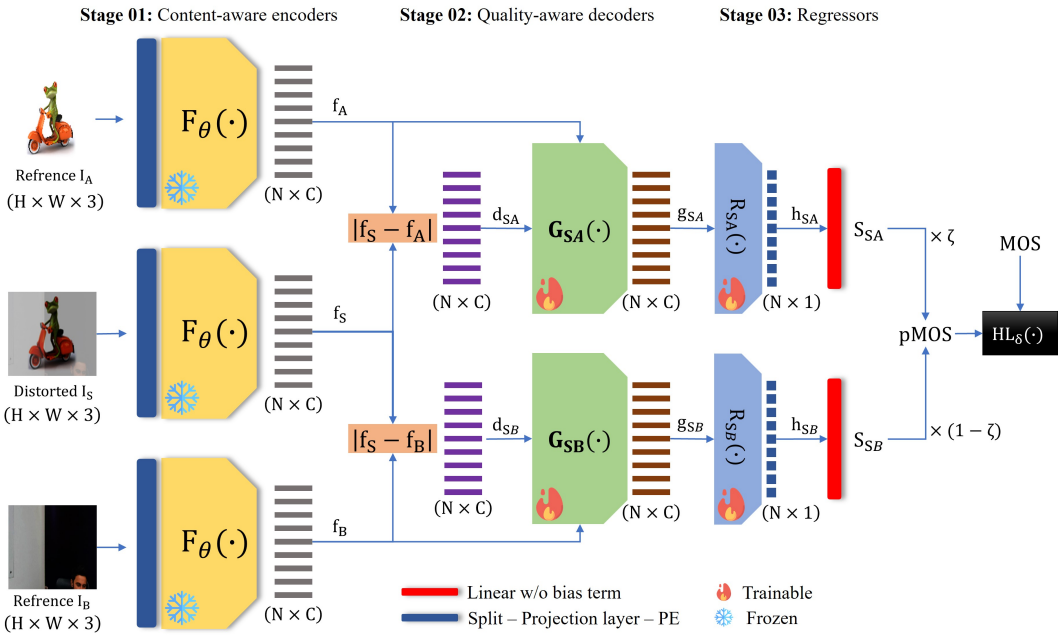


FIGURE 1 – *Illustrates our proposed architecture, showing all components.*

includes three input images : the superimposed image I_S or the distorted image, the background image I_B, and the AR image I_A or the foreground. I_S is calculated by :

$$I_S = \lambda \circ D(I_A) + (1 - \lambda) \circ I_B, \tag{1}$$

where $D(\cdot)$ denotes distortions and λ is the mixing value. Each encoder $F(\cdot)$ generates vectors $f_* = \{f_{*_i}\}_{i=0}^{N}$ for each image, with $* = \{S, A, B\}$. We use the l_1 distance to compute the sequence of distance vectors between superimposed and reference vectors :

$$\begin{cases} d_{SA_i} = |f_{S_i} - f_{A_i}| \\ d_{SB_i} = |f_{S_i} - f_{B_i}| \end{cases} \tag{2}$$

Here, d_{SA} and d_{SB} denote the distances between the representations of I_S and I_A, and I_S and I_B, respectively, which serve as inputs to the decoders described in the next section.

2.2 Quality-aware Decoders

We adapt a transformer decoder [17] without the masked self-attention mechanism. Instead, cross-attention (CA) is used, where queries come from the reference vectors and keys/values from the distance vectors. Decoders $G_{SA}(\cdot)$ and $G_{SB}(\cdot)$ use CA :

$$\begin{cases} \text{CA}(Q_{f_A}, K_{d_{SA}}, V_{d_{SA}}) = \text{softmax}\left(\frac{Q_{f_A} K_{d_{SA}}^T}{\sqrt{d_k}}\right) V_{d_{SA}} \\ \text{CA}(Q_{f_B}, K_{d_{SB}}, V_{d_{SB}}) = \text{softmax}\left(\frac{Q_{f_B} K_{d_{SB}}^T}{\sqrt{d_k}}\right) V_{d_{SB}} \end{cases} \tag{3}$$

where Q_{f_A} and Q_{f_B} are queries from f_A and f_B. Keys/values are from d_{SA} and d_{SB}, and d_k is the key vector dimension. The output embeddings are normalized, followed by a skip connection and projected via a multi-layer perceptron, producing vectors $g_{SA} = G_{SA}(f_A, d_{SA})$ and $g_{SB} = G_{SB}(f_B, d_{SB})$. This captures long-range quality information based on the distortions and visual confusion information that come from the distance vectors.

2.3 Regressors

Two regression modules $R_{SA}(\cdot)$ and $R_{SB}(\cdot)$ produce quality scores for each patch x_i from I_S : $h_{SA} = R_{SA}(g_{SA})$ and $h_{SB} = R_{SB}(g_{SB})$. Scores are aggregated using a linear layer with parameters W_{SA} and W_{SB}, resulting in final scores S_{SA} and S_{SB}. The predicted MOS (*pMOS*) is then computed as :

$$pMOS = \zeta S_{SA} + (1 - \zeta) S_{SB}, \tag{4}$$

where ζ is set to 0.51 based on iterative experimentation.

2.4 Training Procedure

Our training procedure includes key aspects such as loss function choice and addressing overfitting with regularization techniques. We selected the Huber loss [20] for its balance between robustness to outliers and sensitivity to small errors :

$$HL_\delta(y, \hat{y}) = \begin{cases} \frac{1}{2}(y - \hat{y})^2 & \text{if } |y - \hat{y}| \leq \delta, \\ \delta |y - \hat{y}| - \frac{1}{2}\delta^2 & \text{otherwise.} \end{cases} \tag{5}$$

To mitigate overfitting, we use elastic net regularization [21], which combines Lasso and Ridge methods. The overall loss function is :

$$L(y, \hat{y}, \beta) = HL_\delta(y, \hat{y}) + \lambda(\alpha \|\beta\|_1 + (1 - \alpha)\|\beta\|_2^2), \tag{6}$$

where β represents the learnable weights, λ controls the penalty terms, and α balances l_1 and l_2 penalties.

We also apply label smoothing, a technique traditionally used in classification to reduce model overconfidence,

(a) *AR image* I_A (b) *Distorted image* I_S (c) *Background image* I_B

FIGURE 2 – *Illustration of reference and superimposed (distorted) images from the used dataset [7].*

adapted here to regression. Given data scarcity and the overconfidence in predicting MOS, we introduce small random noises to the MOS :

$$y_\epsilon = y + \lambda_n \epsilon, \tag{7}$$

where λ_n is a uniform random value between [-1, 1], and ϵ is a normal distribution random number. This approach reduces overfitting and slightly improves performance.
For the implementation, we trained the model using the AdamW optimizer [22] with a learning rate of $1e^{-4}$, a batch size of 32 images, and for 150 epochs. A learning rate scheduler reduced the rate if no improvement was seen in 10 epochs. The implementation in PyTorch was run on an NVIDIA Tesla V100S-PCIE-32GB GPU.

3 EXPERIMENTAL RESULTS

We evaluated our approach on the ARIQA dataset (Figure 2), containing 560 superimposed images with associated MOS. Following [7], we divided the dataset into 50 folds, splitting each fold into 280 training and 280 testing samples without scene repetition, as detailed in Equations 8 and 9.

$$\mathcal{D} = \{[I_{A_i}, I_{B_i}, I_{S_i}, MOS_i]\}_{i=1}^{560} \tag{8}$$

$$\mathcal{X} = \{(\mathcal{T}_i, \mathcal{S}_i) \mid \mathcal{T}_i \cap \mathcal{S}_i = \emptyset\}_{j=1}^{50} \text{where} \bigcup_{j=1}^{50} \mathcal{S}_j = \mathcal{D} \tag{9}$$

3.1 Dataset

The ARIQA dataset includes 20 background images (10 indoor, 10 outdoor) and 20 AR images (web pages, natural images, and graphical representations). Each AR image has six degraded levels using JPEG compression, scaling, and contrast adjustment. Visual confusion is considered a distortion [7] with mixing thresholds $\lambda \in [0.26, 0.42, 0.58, 0.74]$, resulting in 560 stimuli. 23 participants evaluated the images using HTC VIVE Pro Eye VR headsets.

3.2 Comparison to State-of-the-Art

Table 1 compares our method to state-of-the-art approaches. Averaging metrics across 50 folds, our method outperforms others, including ARIQA+, with fewer parameters (15.32M).

TABLEAU 1 – *Comparison with state-of-the-art performance.*

Model	SRCC↑	KRCC↑	PLCC↑	# params
LPIPS [10]	0.7624	0.5756	0.7591	20.02 M
CFIQA [7]	0.7787	0.5863	0.7695	20.12 M
ARIQA [7]	0.7902	0.5967	0.7824	20.12 M
ARIQA+ [7]	0.8124	0.6184	0.8136	35.02 M
TransformAR (ours)	**0.8267**	**0.6359**	**0.8251**	**15.32 M**

3.3 Ablation Study

An ablation study using five folds from $\mathcal{X}$ evaluated the impact of removing components like the decoder, l_1-distance, label smoothing, elastic net, and Huber loss. Each component significantly impacted the model's performance (Table 2).

TABLEAU 2 – *Ablation study results on 5 folds.*

Model \ Criteria	SRCC↑	KRCC↑	PLCC↑
w/o decoder	0.6161	0.4408	0.6242
w/o l_1-distance	0.4365	0.4443	0.2960
w/o label smoothing	0.8269	0.6361	0.8221
w/o elastic net	0.8427	0.6567	0.8471
w/o Huber loss	0.8374	0.6525	0.8443
all combined	**0.8461**	**0.6582**	**0.8481**

4 CONCLUSION

This paper introduces an efficient and lightweight objective quality assessment metric for AR scenarios based on the transformer architecture. To address data scarcity, we simplified the model using two encoder blocks and one decoder block. Elastic net regularization and label smoothing were employed to enhance model robustness. Our proposed method surpassed widely used metrics like LPIPS and existing ARIQA metrics, achieving superior performance with significantly fewer parameters. The emerging field of AR quality assessment presents opportunities for advancement. Future research will focus on transformer-based approaches in realistic AR scenarios and developing specialized AR-IQA datasets to enhance objective quality metrics.

Références

[1] R. Vertucci, S. D'Onofrio, S. Ricciardi, et M. De Nino. History of augmented reality. Dans *Springer Handbook of Augmented Reality*, pages 35–50. Springer, 2023.

[2] International Telecommunication Union. Itu-t recommendation g.1036. `https://www.itu.int/rec/T-REC-G.1036-202207-I`, 2022. Accessed on October 27, 2023.

[3] A. Perkis, C. Timmerer, S. Baraković, et al. Qualinet white paper on definitions of immersive media experience (imex). *arXiv preprint arXiv :2007.07032*, 2020.

[4] J. Xu, C. Lin, W. Zhou, et Z. Chen. Subjective quality assessment of stereoscopic omnidirectional image. Dans *Advances in Multimedia Information Processing–PCM 2018 : 19th Pacific-Rim Conference on Multimedia, Hefei, China, September 21-22, 2018, Proceedings, Part I 19*, pages 589–599. Springer, 2018.

[5] E. Alexiou, E. Upenik, et T. Ebrahimi. Towards subjective quality assessment of point cloud imaging in augmented reality. Dans *2017 IEEE 19th Int. Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6, 2017.

[6] J. Gutiérrez, T. Vigier, et P. Le Callet. Quality evaluation of 3d objects in mixed reality for different lighting conditions. *Electronic Imaging*, 32 :1–7, 2020.

[7] H. Duan, X. Min, Y. Zhu, et al. Confusing image quality assessment : Toward better augmented reality experience. *IEEE Transactions on Image Processing*, 31 :7206–7221, 2022.

[8] Z. Wang, A. Bovik, H. Sheikh, et E. Simoncelli. Image quality assessment : from error visibility to structural similarity. *IEEE Transactions on Image Processing (TIP)*, 13(4) :600–612, 2004.

[9] H. Sheikh et A. Bovik. Image information and visual quality. *IEEE Transactions on Image Processing (TIP)*, 15(2) :430–444, 2006.

[10] R. Zhang, P. Isola, A. Efros, et al. The unreasonable effectiveness of deep features as a perceptual metric. Dans *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018.

[11] F. N Iandola, S. Han, M. Moskewicz, K. Ashraf, W. Dally, et K. Keutzer. Squeezenet : Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size. Dans *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[12] A. Krizhevsky, I. Sutskever, et G. Hinton. Imagenet classification with deep convolutional neural networks. Dans *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 1097–1105, 2012.

[13] K. Simonyan et A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv :1409.1556*, 2014.

[14] K. He, X. Zhang, S. Ren, et J. Sun. Deep residual learning for image recognition. Dans *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[15] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, et A. Dosovitskiy. Do vision transformers see like convolutional neural networks ? *Advances in Neural Information Processing Systems*, 34 :12116–12128, 2021.

[16] M. Cheon, S. Yoon, B. Kang, et J. Lee. Perceptual image quality assessment with transformers. Dans *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 433–442, 2021.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N Gomez, Ł. Kaiser, et I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[18] A. Dosovitskiy, L. Beyer, A.r Kolesnikov, D. Weissenborn, et al. An image is worth 16x16 words : Transformers for image recognition at scale. *arXiv preprint arXiv :2010.11929*, 2020.

[19] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, et A. Joulin. Emerging properties in self-supervised vision transformers. Dans *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[20] P. Huber. Robust estimation of a location parameter. Dans *Breakthroughs in statistics : Methodology and distribution*, pages 492–518. Springer, 1992.

[21] H. Zou et T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B : Statistical Methodology*, 67(2) :301–320, 2005.

[22] I. Loshchilov et F. Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv :1711.05101*, 2017.

Évaluation des mesures d'équité pour les systèmes biométriques par génération de biais contrôlés

KN. SANON[1 2] J. DI MANNO[2] T.GERNOT[1] C. CHARRIER[1] C. ROSENBERGER[1]

[1] Université Caen Normandie, ENSICAEN, CNRS, Normandie Univ,
GREYC UMR6072, F-14000 Caen, France
[2] FIME EMEA, 14000 Caen, France

{neily.sanon, tanguy.gernot, christophe.charrier}@unicaen.fr
{christophe.rosenberger}@ensicaen.fr, {joel.dimanno}@fime.com

Résumé

Les biais dans les systèmes biométriques, tels que ceux liés au genre et à l'ethnie, sont des enjeux technologiques et sociétaux cruciaux à considérer. Dans ce travail, nous suivons un scénario de boîte grise avec une transparence limitée, permettant l'accès au score de comparaison biométrique et au seuil de décision. Différentes métriques d'évaluation ont été proposées pour quantifier l'équité d'un système biométrique. Nous proposons, tout d'abord, une méthode pour analyser les biais dans un système équitable et, ensuite, une étude comparative des métriques à l'état de l'art, en nous concentrant sur la corrélation entre les biais et les métriques dans les systèmes de reconnaissance faciale. Dans notre protocole expérimental, nous utilisons différentes bases de visages et des extracteurs de visages avec différentes fonctions de perte. Les résultats expérimentaux permettent d'évaluer la capacité des métriques à quantifier correctement ou non les biais dans les systèmes biométriques.

Mots clefs

Système biométrique, Évaluation de l'équité, Performance.

1 Introduction

Les systèmes biométriques authentifient les individus en utilisant des caractéristiques physiques ou comportementales uniques et sont désormais largement utilisés dans diverses industries. Cependant, leur efficacité et leur équité peuvent être compromises par différents types biais(Dans notre cas, nous considérerons les biais liés au genre).
Cet article porte sur l'évaluation des biais dans les systèmes biométriques, en particulier dans les situations de boîte grise où les mécanismes internes ne sont pas tous accessibles. Notre étude vise à améliorer l'équité et la fiabilité de ces technologies et à établir de nouvelles références pour les évaluations futures dans des contraintes similaires. Les contributions clés incluent :

1. Une revue des métriques existantes pour l'estimation des biais dans les systèmes biométriques.
2. Une méthode pour contrôler les biais après l'extraction des caractéristiques comme vérité terrain pour valider les métriques d'équité.
3. Un protocole expérimental utilisant deux bases de données et systèmes biométriques faciaux avec des fonctions de perte.
4. Des résultats expérimentaux montrant les avantages de la méthodologie proposée pour comparer les métriques d'équité.

L'article est structuré comme suit : la section 2 couvre les bases des systèmes biométriques et des biais, la section 3 examine les recherches sur l'évaluation de l'équité, la section 4 décrit notre protocole expérimental, la section 5 présente les résultats, et enfin, la conclusion aborde la discussion et les perspectives futures.

2 Contexte

2.1 Système biométrique

Un système biométrique vérifie ou identifie un utilisateur en utilisant des caractéristiques uniques. Les modalités typiques incluent les empreintes digitales, le visage, la voix, l'iris, etc. Le système fonctionne suivant plusieurs étapes définies ci-après :

Capture : Le système capture des données biométriques brutes, telles qu'une image du visage de l'utilisateur qui sera ensuite détecté par un détecteur de visage permettant de localiser la zone du visage dans l'image.

Extraction : Les caractéristiques sont extraites de l'échantillon de visage utilisant généralement des réseaux neuronaux convolutifs tels que Inception ou ResNet50.

Comparaison : Les caractéristiques extraites sont comparées à une base de données pour de l'identification ou au modèle biométrique de référence pour de l'authentification. Ceci s'opérant grâce à des calculs de distance telles que le cosinus ou Manhattan, générant ainsi un score.

Décision : Si le score est en dessous (ou au-dessus) d'un seuil fixé, la vérification de l'identité est accordée.

2.2 Définition des biais

Dans cette étude, nous interprétons le biais comme une déviation par rapport à une norme. Il peut se manifester sous différentes formes : statistique (déviations numériques par rapport aux valeurs attendues), moral (déviations par rapport aux normes éthiques), ou encore sous d'autres aspects légaux, sociaux et psychologiques. Nous considérons un système biométrique équitable s'il fonctionne de manière cohérente avec des bases de données biaisées et non biaisées, que ce soit en termes de données démographiques, de condition physique, de qualité de l'image, d'environnement ou d'accessoires. En classification, cela signifie une catégorisation précise indépendamment du biais introduit. Pour notre étude sur l'authentification faciale impliquant le genre, l'équité implique une précision de reconnaissance égale entre ces groupes. Dans la section suivante, nous analysons l'état de l'art dans l'évaluation de l'équité des systèmes biométriques.

3 Etat de l'art

Un ensemble de données biométriques se compose d'utilisateurs $U_i,\ i = [1,\ldots,N]$ (où N est le nombre d'utilisateurs), chacun avec des échantillons biométriques $S_{i,j},\ j = [1,\ldots,M]$ (où M est le nombre d'échantillons par utilisateur). Les utilisateurs appartiennent à des catégories démographiques telles que le genre $d_i = \{\text{masculin}, \text{féminin}\},\ i = [1,\ldots,N]$. Un système biométrique est équitable s'il fonctionne de manière cohérente pour toutes les catégories d'utilisateurs. La performance est évaluée en examinant deux principales erreurs : le taux de fausses correspondances (FMR) et le taux de fausses non-correspondances (FNMR). Les fausses correspondances sont souvent dues à des caractéristiques non uniques, tandis que les fausses non-correspondances peuvent résulter du bruit de l'échantillon, de modèles de mauvaise qualité ou de changements dans les données biométriques au fil du temps. Le seuil de décision τ a un impact significatif sur ces erreurs. L'étude des biais dans les systèmes biométriques est un domaine en pleine croissance, mettant en lumière les défis pour garantir l'équité des performances. En 2021, Drozdowski et al. [1] ont exploré l'impact des caractéristiques démographiques sur la performance de reconnaissance, en soulignant les problèmes entre différents groupes ethniques, genres et âges. Howard et al. dans [2] ont montré que des facteurs environnementaux tels que l'éclairage peuvent introduire des biais dans la reconnaissance faciale.

Différentes approches abordent les biais dans les systèmes biométriques. Schuckers et al. [3] offrent une perspective statistique, en considérant les variations pouvant survenir par hasard (erreur de type I). Fang et al. [4] ont proposé la métrique Accuracy Balanced Fairness (ABF).

Nous avons retenu quatre métriques, à savoir trois sur les résultats différentiels et une sur la performance différentielle, telles que définies ci-après :

1. **Taux de Discrépance d'Équité (FDR)** [5] : Mesure les différences de performance entre les groupes démographiques en comparant le FMR et le FNMR à un seuil donné τ.

$$FDR = 1 - (\alpha \times A(\tau) + (1-\alpha) \times B(\tau)) \quad (1)$$

où

$$A(\tau) = \max_{i,j} \left(\left| FMR^{d_i}(\tau) - FMR^{d_j}(\tau) \right| \right)$$

$$B(\tau) = \max_{i,j} \left(\left| FNMR^{d_i}(\tau) - FNMR^{d_j}(\tau) \right| \right)$$

représentant l'écart maximal de fausses correspondances et de fausses non correspondances suivant les démographies considérées.

2. **Taux d'Inégalité (IR)** [6] : Évalue les disparités en calculant le rapport des valeurs maximales et minimales de FMR et FNMR.

$$IR = \left(\frac{\max_{d_i} FMR(\tau)}{\min_{d_i} FMR(\tau)} \right)^{\alpha} \times \left(\frac{\max_{d_i} FNMR(\tau)}{\min_{d_j} FNMR(\tau)} \right)^{1-\alpha} \quad (2)$$

3. **Taux d'Agrégation de Gini pour l'Équité Biométrique (GARBE)** [7] : Utilise le coefficient de Gini pour mesurer l'équité en agrégeant FMR et FNMR.

$$G_x = \left(\frac{n}{n-1} \right) \left(\frac{\sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|}{2n^2 \bar{x}} \right) \quad (3)$$

$$\text{GARBE}(\tau) = \alpha A(\tau) + (1-\alpha) B(\tau) \quad (4)$$

où $A(\tau)$ and $B(\tau)$ sont les coefficients de Gini pour le FMR et le FNMR respectivement.

4. **Indice de Séparation de l'Équité (SFI)** [8] : Quantifie la capacité à distinguer les scores authentiques des imposteurs entre les groupes démographiques d_i (où $i \in$ [1,D]).

$$SFI_N = 1 - \frac{2}{D} \sum_{i=1}^{D} |z_{Si} - z_{S\text{mean}}| \quad (5)$$

où $z_{Si} = |\mu_{Gi} - \mu_{Ii}|, \quad i = 1, 2, \ldots, D$ et $z_{S\text{mean}} = \frac{1}{D} \sum_{i=1}^{K} z_{Si}$

Ces métriques suivent des approches similaires, et évaluer leur fiabilité est un défi. Ce travail propose de comparer ces métriques en utilisant un protocole rigoureux défini.

4 Protocole expérimental

Dans cette section, nous détaillons les expériences suivies pour comparer les métriques d'équité décrites ci dessus.

4.1 Base de données

Dans ce travail, nous avons utilisé deux ensembles de données publiques de visages. Le premier est LFW10, un sous-ensemble du dataset LFW [9], qui contient 158 sujets avec plus de 10 apparitions. Le second est DemogPairs [10], connu pour son équité en matière de genre et de ces trois ethnicités (Asiatique, Noir, Blanc).

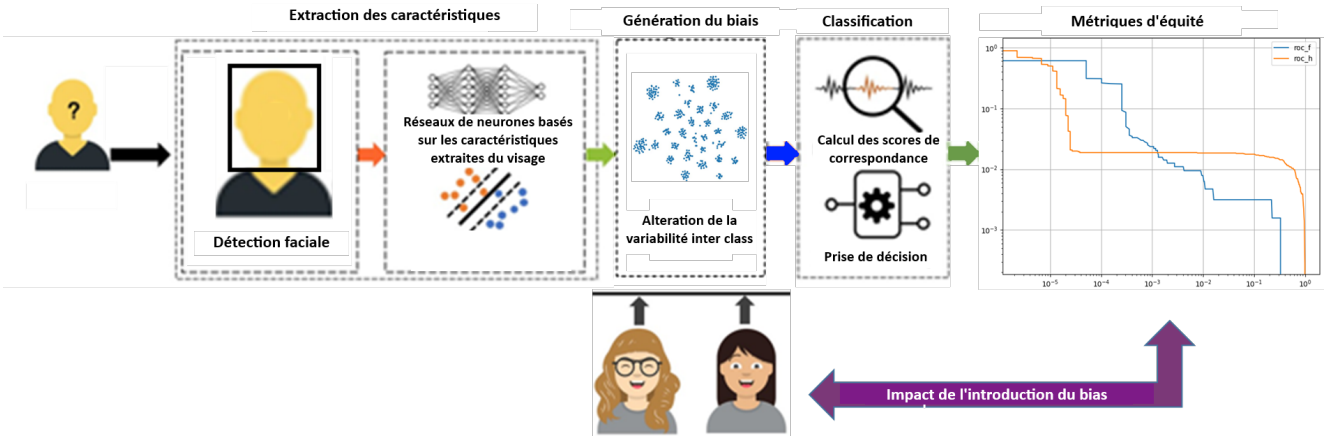


FIGURE 1 – *Méthodologie proposée pour l'évaluation des métriques d'équité.*

4.2 Systèmes biométriques

Nous utilisons MTCNN [11] pour la détection des visages, connu pour sa haute précision. MTCNN est pré-entraîné sur VGGFace, qui se compose de 59,3% de sujets masculins. Pour l'extraction des caractéristiques, nous utilisons les modèles CNN pré-entraînés suivants, chacun étant entraîné avec des fonctions de perte spécifiques pour une extraction précise (512 caractéristiques par modèle) :

Inception-ResNet V1 + Softmax combine l'architecture Inception pour l'extraction multi-échelle de caractéristiques avec ResNet pré-entraîné sur VGGFace2.

InsightFace + ArcFace [12] utilise la perte angulaire additive pour améliorer la précision de la reconnaissance en introduisant une marge angulaire dans la fonction de perte.

4.3 Scénarios de test

Nous avons utilisé deux scénarios : 1 non biaisé et un 2nd biaisé. Initialement, nous avons utilisé des échantillons biométriques sans modification. Ensuite, nous avons ajouté du bruit Gaussien aux caractéristiques pour biaiser le genre femme. Les étapes, résumées dans la Figure 1, sont :

1. **Extraction des caractéristiques et des étiquettes :** Extraire les caractéristiques et les étiquettes des base de données.
2. **Génération de biais :** Ajouter du bruit Gaussien pour biaiser les caractérisques d'une démographie, en occurence les femmes.
3. **Calcul des scores de correspondance :** Déterminer les scores légitimes et imposteurs.
4. **Analyse des taux d'erreurs (FMR et FNMR) :** Calculer les taux de fausses correspondances et de fausses non-correspondances sur différents seuils.
5. **Évaluation des performances :** Appliquer les FNMR et FMR en tenant compte du genre avec des visualisations pour montrer l'impact démographique sur les biais et la précision.
6. **Calcul des métriques d'équité :** Calculer les métriques sur les données non biaisées et biaisées, en contrôlant le biais avec la valeur de bruit σ.
7. **Calcul des corrélations :** Utiliser la corrélation de Pearson pour quantifier comment les métriques d'équité répondent au biais synthétique, en visualisant les relations entre les métriques et le biais.

5 Résultats expérimentaux

Les résultats obtenus avec ce protocole peuvent être divisés en trois parties : la performance initiale des systèmes, le comportement des systèmes biométriques face aux biais, et l'évaluation des métriques corrélant biais et bruit.

5.1 Évaluation des performances

Nous évaluons l'efficacité des deux systèmes biométriques proposées. Nous calculons l'aire sous la courbe (AUC) pour mesurer la capacité du système à minimiser les taux de fausses correspondances et de fausses non-correspondances à travers différents systèmes et bases de données. Une AUC minimale indique une distinction efficace entre différents groupes. Nous avons observé une AUC de 0,02 pour les deux systèmes, suggérant que nos systèmes sont alignés avec la définition de l'équité de la section 2.2. Ensuite, nous examinons les différences de performance entre les catégories démographiques en utilisant la technique de biais synthétique proposée.

5.2 Introduction de biais

L'objectif de cette technique d'altération est de biaiser les systèmes pour évaluer la sensibilité des métriques aux biais. Nous introduisons du bruit avec un σ allant de la moyenne de l'écart-type intra-groupe $\sigma_{\text{std intra group}}$ à quatre fois cette valeur ($\sigma \in [0, 4 \times \sigma_{\text{std intra group}}]$), obtenant ainsi dix valeurs de bruit dans cet intervalle. La moyenne des écarts-type intra-groupe est de 0,039 pour Demogpairs, et de 0,044 pour LFW10. La Figure 2 illustre l'impact de ce bruit sur les femmes, projeté à l'aide de l'algorithme t-Distributed Stochastic Neighbor Embedding (t-SNE).

TABLE 1 – *Corrélation (Valeur absolue en pourcentage) - Genre*

Métrique	FDR					IR					GARBE					SFI
α (si il existe)	0	0.25	0.5	0.75	1	0	0.25	0.5	0.75	1	0	0.25	0.5	0.75	1	
ARCFACE/Demogp	83	83	83	83	80	83	89	**89**	88	80	64	74	76	74	70	83
ARCFACE/LFW10	83	82	82	82	82	83	89	**90**	**90**	83	83	82	81	77	53	81
INCEPTION/Demogp	83	83	83	82	81	83	87	**88**	87	81	78	80	82	83	84	84
INCEPTION/LFW10	84	84	84	82	79	84	**88**	**88**	87	79	80	80	80	80	80	85

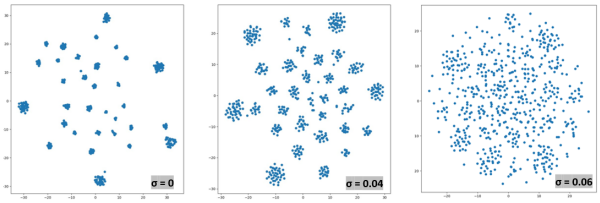


FIGURE 2 – *Introduction de bruit (des caractéristiques initiales aux biais progressifs)*

5.3 Métriques d'équité

Nous évaluons l'équité de ces systèmes biométriques après l'introduction de biais. Les points clés à considérer sont : **1-)** Le biais a été introduit en ajoutant des perturbations aux variations intra-groupe, augmentant le taux de fausses non-correspondances et dégradant l'expérience utilisateur. L'analyse doit se concentrer sur cet aspect. **2-)** La valeur alpha dans FDR, IR et GARBE vise à équilibrer le FMR et le FNMR. Une valeur alpha plus basse est critique lorsque le FNMR est élevé. **3-)** La valeur d'équité maximale pour FDR est 1, tandis que pour IR et GARBE elle est 0. SFI augmente avec l'équité.

Sans biais, Arcface et Inception ResNet semblent équitables pour le genre. Lorsque le biais est introduit, toutes les métriques reflètent un changement dans l'expérience utilisateur, montrant une transition de l'équité à l'iniquité à un certain taux. Les corrélations entre les niveaux de bruit et les valeurs des métriques ont été calculées, en tenant compte de l'aire sous les courbes des métriques en utilisant la règle trapézoïdale.

Le comportement des métriques à travers les datasets et les systèmes est résumé dans le Tableau 1. Demogpairs montre des corrélations stables (80%-89%), indiquant une capture efficace des biais. LFW10, cependant, montre des corrélations variables, surtout pour GARBE avec un α de 0,75 (53%-80%). Cela suggère que les métriques sont moins robustes face à différents types de biais. IR démontre la meilleure corrélation globale. La structure de cette métrique semble intéressante pour une investigation plus approfondie.

6 Conclusion et perspectives

Notre étude des biais dans les systèmes biométriques montre un fort intérêt en raison des exigences strictes de certification en termes de précision et d'équité. Nous comparons les métriques de biais en introduisant des perturbations affectant le taux de fausses non-correspondances. Les métriques semblent stables par rapport aux biais, avec IR en tête. Les recherches futures pourraient ajouter des biais par variation intra-groupe tenant compte à la fois du volet sécurité et expérience utilisateur et développer des mesures qui ne reposent pas sur des paramètres déterministes (comme le paramètre alpha) mais prenant en compte les seuils.Aussi, on pourrait étendre l'étude à d'autres bases de données et extracteurs.

Références

[1] Pawel Drozdowski, Christian Rathgeb, et Christoph Busch. Demographic Fairness in Face Identification : The Watchlist Imbalance Effect, Juin 2021. arXiv :2106.08049 [cs].

[2] John J Howard, Yevgeniy B Sirotin, et Arun R Vemury. The effect of broad and specific demographic homogeneity on the imposter distributions and false match rates in face recognition algorithm performance. 2019.

[3] Michael Schuckers, Sandip Purnapatra, Kaniz Fatima, Daqing Hou, et Stephanie Schuckers. Statistical Methods for Assessing Differences in False Non-Match Rates Across Demographic Groups, Août 2022.

[4] Meiling Fang, Wufei Yang, Arjan Kuijper, Vitomir Struc, et Naser Damer. Fairness in face presentation attack detection. *Pattern Recognition*, 147 :110002, Mars 2024.

[5] Tiago De Freitas Pereira et Sebastien Marcel. Fairness in Biometrics : A Figure of Merit to Assess Biometric Verification Systems. 4(1), Janvier 2020.

[6] P. Grother. Demographic differentials in face recognition algorithms. EAB Virtual Event Series - Demographic Fairness in Biometric Systems, 2021.

[7] John J. Howard, Eli J. Laird, Rebecca E. Rubin, Yevgeniy B. Sirotin, Jerry L. Tipton, et Arun R. Vemury. Evaluating Proposed Fairness Models for Face Recognition Algorithms. Lecture Notes in Computer Science, Cham, 2023.

[8] Ketan Kotwal et Sebastien Marcel. Fairness Index Measures to Evaluate Bias in Biometric Recognition, Juin 2023.

[9] Gary B Huang, Manu Ramesh, Tamara Berg, et Erik L-M. Labeled Faces in the Wild : A Database for Studying Face Recognition in Unconstrained Environments.

[10] Isabelle Hupont et Carles Fernández. Demogpairs : Quantifying the impact of demographic imbalance in deep face recognition. 2019.

[11] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, et Yu Qiao. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. 23(10), Octobre 2016.

[12] Jiankang Deng, Jia Guo, Niannan Xue, et Stefanos Zafeiriou. ArcFace : Additive Angular Margin Loss for Deep Face Recognition. Juin 2019.

Enhancing Medical Image Security : An Adaptable and robust watermarking Scheme

S.S. ZANEIDOU[1] M. OUTTAS[2] M. MOSSI IDRISSA[1] K. KPALMA[2]

[1] Physics department, Abdou Moumouni University of Niamey, Niger

[2] Univ. Rennes, INSA Rennes, CNRS, IETR - UMR 6164, Rennes, France

Abstract

In recent years, digital image watermarking has attracted much attention from researchers because of its great ability to ensure the security and protection of information against unauthorized access or illegal modification. In this paper, a robust watermarking scheme for medical images based on the discrete wavelet transform (DWT), the singular value decomposition (SVD), and the Arnold transform (AT) has been proposed. Several image modalities including ultrasound, CT scan, PET scan, X-ray, and MRI were used to test our proposed method. On the other hand, two metrics PSNR and SSIM were used to assess the imperceptibility while the NC metric was used to evaluate the robustness. Experimental results without attacks showed that our method offers good imperceptibility with a PSNR above 45dB and SSIM close to 1. After applying the attacks, the results showed that the method presents good robustness with NC close to 1. In conclusion, our proposed method offers a good compromise between imperceptibility and robustness which allows it to be a valuable method in the field of e-health.

Keywords

Medical image watermarking, DWT, SVD, Arnold transform

1 Introduction

The development of information and communication systems and the high use of electronic management of medical records have enabled the sharing or exchange of digital medical images around the world for several services such as telemedicine, teleradiology, telediagnosis, and teleconsultation. For that, a system for effectively sharing the medical files of patients between several hospitals or centers must ensure good protection and security of these data. To solve this problem, the Watermarking system appears as one of the most effective methods for ensuring the integrity and authenticity of these data. It aims to secure storage and transmission and could ensure the authentication of information and only legal duplication in the exchange of health information.s [1].

The digital watermarking technique consists of inserting information called the watermark into the original image. In general, this watermark can be a logo, text, or image [2]. Watermarking techniques can be classified based on two domains. Firstly, a spatial domain where the watermark is inserted directly by modifying the values of the pixels of the original image. Secondly, the frequential domain, where the watermark is inserted in the coefficients of transformation of the original [3] [4]. The requirements of a good watermarking system are [5] :

- robustness, which is the effectiveness of a watermarking system against the usual image processing operations such as filtering, compression, geometric transformation, ... commonly called attacks [5] ;
- imperceptibility, which means that after the insertion of the watermark, the watermarked image's quality must remain faithful to the original and the watermarked image must ensure the same diagnostic reliability [5] [3] ;
- capacity, refers to the amount of information to be embedded in the original image [2] [6]

In the context of medical images, these properties are particularly important, because, in addition to the sensitivity of the information they contain, they are designed to perform a diagnostic task.

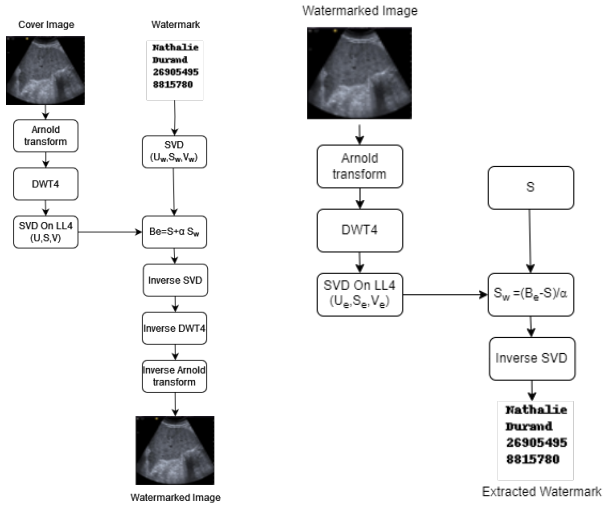
Several studies have been developed in the field of medical image watermarking. Movahed et al. [4] proposed a watermarking method dedicated to Computerized Tomography (CT) medical images. et al. [7] proposed a watermarking method based on “multiple watermarking systems", where they used a Discrete Wavelet Transform (DWT) and Discrete Cosine Transform (DCT), with a quantization step of the frequency coefficients. Mahyudin *et al.* [1] introduced the Arnold Transform in their algorithm to improve the robustness of the watermarking process. Recently, Vaidya *et al.* [8] proposed the embedding of a patient's fingerprint in different modalities using hybrid transform. Some studies have proposed deep-learning methods for watermarking purposes in medical images. Singh *et al.*[9] provides a comprehensive review of watermarking techniques in deep learning environments, emphasizing their accuracy and learning ability. However, the explainability of deep learning methods in medical applications is an ongoing challenge and has not been addressed in medical watermarking, which is a major concern in the medical community.

In this paper, we propose a versatile watermarking of medical images based on the DWT SVD and the Arnold transform. DWT is used for its compatibility with compression and its robustness against several attacks. The Arnold transform is applied for its encryption effect to increase the security of the watermark. The rest of this paper is organized as follows : Section 2 presents the proposed watermarking scheme. Section 3 presents further experiments, results, and discussion on the method. The paper is concluded in section 4.

2 PROPOSED WATERMARKING SCHEME

2.1 Insertion process

To insert the watermark, we first applied image processing techniques on both the original and the watermark. Then Arnold transform, DWT, and SVD are applied, and the entire algorithm is represented in Figure1(a).



(a) Watermark embedding process

(b) Watermark extracting process

FIGURE 1 – *Watermarking process*

The watermarked image is generated. At the end of the process, the result is evaluated using the performance metrics : PSNR and SSIM.

2.2 Extraction process

The watermark extraction process is the reverse of the insertion one as presented in Figure1(b). At the end of this process, the NC is used to assess the robustness of the system.

2.3 Performance metrics

In this study, we used two categories of metrics to assess the performance of the system. The first category aims to evaluate the imperceptibility and image quality of the watermarked image. This is typically measured using metrics such as the peak-signal-to-noise ratio (PSNR) and the structural similarity index measure (SSIM). The second one aims to assess the robustness of the watermarking system against various attacks. This is typically measured using the normalized correlation (NC) coefficient that evaluates the similarity between the original and the attacked watermark.

TABLEAU 1 – *Results without an attack on US Image*

Gain factor(α)	PSNR(dB)	SSIM	NC
0.01	47.7279	0.9930	0.9903
0.02	47.7778	0.9931	0.9976
0.03	47.7778	0.9931	0.9987
0.04	47.3529	0.9918	0.9986
0.05	47.3529	0.9918	0.9984
0.06	46.5723	0.9891	0.9999
0.07	46.5723	0.9891	0.9999
0.08	45.5992	0.9852	0.9996
0.09	45.5992	0.9852	1.0

3 RESULTS AND DISCUSSION

In this part, we present the evaluation results and a comparison with the results proposed in the literature. The size of the original image used is 1024×1024 and 64×64 for the watermark. Images from several medical imaging modalities such as US [10], CT, PET-scan, X-ray [11], and MRI [12] are used.

The numerical results without attack for the different gain factor values are presented in Table 1 where the maximum value of $PSNR = 47.7778dB$ and $SSIM = 0.9931$ is given at $\alpha = 0.02$ and $\alpha = 0.03$. The results for several image modalities without attack, the comparison of the results without attack with the existent results, and the comparison of the results with the attack of our system with the literature's results are presented in table 2, 3 and 4. These tests are carried out for $\alpha = 0.07$ this later value was determined empirically in this context of medical imaging it is valuable at this stage and before an automation process in ongoing work.

Table 1 presents the evaluation of PSNR, SSIM, and NC according to the gain factor. It shows that the PSNR and SSIM values decrease when the gain factor increases. This explains that the greater the gain factor, the more important the watermark is in the watermarked image, which will reduce the value of NC.

It is worth mentioning that our proposed method has a lower PSNR and high SSIM than the method proposed in [1]. This can be explained by the fact that the watermark used in our method contains more characters than the one used by Mahyudin *et al.* [1]. This indicates that the proposed scheme has a greater capacity to embed information. Furthermore, the SSIM is more correlated to the human visual

TABLEAU 2 – *Comparison of the proposed method with the existing watermarking scheme without attack*

Image modality	Watermarking scheme in [1]			Watermarking scheme in [8]			Watermarking scheme in [13]			Prpposed method		
	PSNR	SSIM	NC	PSNR	SSIM	NC	PSNR	SSIM	NC	PSNR	SSIM	NC
Pet-Scan	-	-	-	36.56	0.9821	1.0	24.5103	-	0.9327	46.6764	0.9975	1.0
X-Ray	50.7885	0.9957	1.0	35.43	0.9857	1.0	28.0304	-	0.9488	46.4854	0.9948	1.0
MRI	50.6744	0.9586	1.0	37.07	0.9815	1.0	38.0842	-	0.9869	46.0375	0.9828	0.9784
CT	50.7628	0.9983	1.0	33.85	0.9844	1.0	36.9234	-	0.9154	46.7419	0.9830	0.9868
US	50.7897	0.9539	1.0	36.47	0.9878	1.0	21.4548		0.8865	46.5723	0.9891	0.9999

TABLEAU 3 – *Robustness (NC values) of the proposed VS existing ones under attacks applied on US images*

Attacks/Noise	Noise density	Scheme [13]	Scheme in [1]	Proposed
		NC	NC	NC
Gaussian	0.0001	0.9785	—	0.9996
	0.0005	0.8311	—	0.9996
Rotation	5°	0.8908	—	0.9990
	10°	0.8913	—	0.9596
Salt and pepper	0.0001	0.9975	—	0.9997
	0.001	0.8761	0.9981	0.9999
	0.05	—	0.9290	0.9988
Speckle Noise	0.01	0.8277	0.8814	0.9589
	0.05	—	0.9765	0.9986

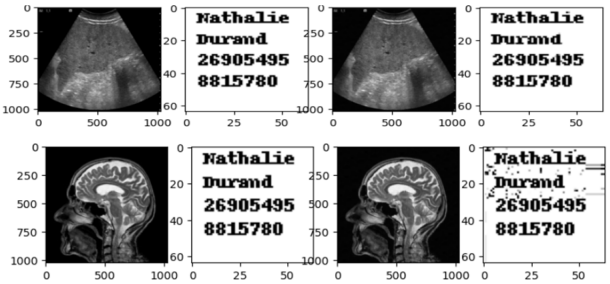


FIGURE 2 – *Watermarked US and MRI images with embedded watermark .From left to right : cover image, watermark, watermarked image, and extracted watermark*

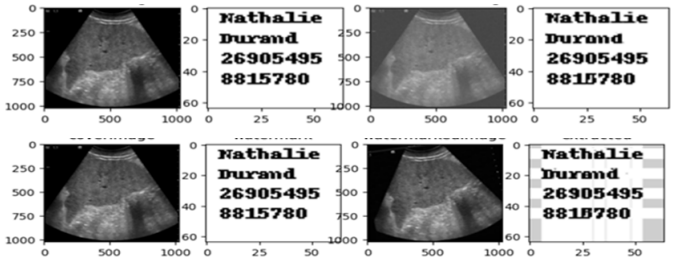


FIGURE 3 – *Visual inspection of US watermarked image with embedded watermark under attack. From left to right : cover image, watermark, watermarked image under attack and extracted watermark*

system than the PSNR in terms of perceptual and diagnostic quality [14] considering the simplest metric to implement. We can conclude that the proposed method offers good imperceptibility, which makes it a valuable method usable in medical applications. To evaluate the robustness of our method, we applied and simulated different types of attacks : a rotation and 3 types of noise, including Gaussian noise, speckle noise, and salt and pepper noise. The results are summarised in Table 2. The values of NC are very close to the value of NC without attacks, indicating that the proposed watermarking scheme has good resistance to these attacks. Table 3 presents a comparison with the method proposed in [1] and [13]. The results indicate that the proposed method has mostly a higher NC for all noise and rotation than that presented by the scheme in [1] and [13]. This can be explained by the presence of the Arnold transform in our algorithm, which increases the security and survivability of the watermark against attacks. Also, one could see that the proposed method has high results in terms of robustness in watermark recovery under noises thanks to the multiplying factor α that allows to management of the reinforcement of the watermark.

Table 4 shows the comparison with the method proposed by VAIDYA *et al.* [8]. It turns out that the proposed method presents an NC higher against all the attacks used and for all image modalities except MRI. The main difference is that authors in [8] used a larger watermark size, which reduces the quality of the watermarked image as shown in table 2. However, the extracted watermark is very similar to the original watermark compared to the state-of-the-art methods. We can therefore conclude that the proposed method offers good robustness against several types of attacks, including adding noise, filtering, and rotation.

For a preview of the rendering quality, Figure 3 shows the resulting images after the embedding of the watermark and the recovered watermark, and Figure 3 offers a visual inspection of images with embedded watermark under attack and the recovered watermark.

4 Conclusion

In this paper, we proposed a robust and adaptive medical image watermarking algorithm based on DWT-SVD and the Arnold transform. DWT is used for its compatibility with compression and its robustness against several attacks. The Arnold transform is applied to increase the security of

TABLEAU 4 – *Performance Comparison on different modalities under attacks*

Attacks	Watermarking Scheme in [8]					Proposed method				
	US	MRI	CT-Scan	Pet-Scan	X-ray	US	MRI	CT-Scan	Pet-Scan	X-ray
Gaussian noise (0,0.002)	0.9745	0.9613	0.9361	0.9708	0.9673	0.9994	0.9737	0.9998	1.0	1.0
Salt and Pepper (0.001)	0.9941	0.9914	0.9792	0.9950	0.9935	0.9999	0.9905	0.9999	1.0	1.0
Median filter (3x3)	0.9890	0.9882	0.9314	0.9833	0.9797	0.9996	0.9792	0.9998	1.0	1.0

the watermark. Several modalities are used for the realization of the tests namely, CT, US, PET-Scan, X-ray, and MRI. The result without attack shows that our method offers good imperceptibility where all the PSNR values are greater than 45dB, those of SSIM are close to 1 and the NC values are very close to 1. Imperceptibility is a key factor in medical imaging. It enables reliable diagnosis and ensures that the quality of the diagnosis is not compromised. Several attacks have been applied, and the experimental results showed that our method has a high robustness. Therefore, the proposed scheme is valuable in enhancing the security of data used in e-health applications. As ongoing work, we are trying to insert the watermark in the region of non-interest of the image in a retrospective dataset with a real attack and to complete the security scheme for the transmission chain far beyond the watermarking process.

Références

[1] Muhammad Fachri Mahyudin, Ledya Novamizanti, et Sofia Sa'idah. Robust watermarking using arnold and hybrid transform in medical images. Dans *2021 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)*, pages 180–185. IEEE, 2021.

[2] David-Octavio Muñoz-Ramirez, Volodymyr Ponomaryov, Rogelio Reyes-Reyes, Volodymyr Kyrychenko, Oleksandr Pechenin, et Alexander Totsky. A robust watermarking scheme to jpeg compression for embedding a color watermark into digital images. Dans *2018 IEEE 9th International Conference on Dependable Systems, Services and Technologies (DESSERT)*, pages 619–624. IEEE, 2018.

[3] Muhammad Arslan Usman et Muhammad Rehan Usman. Using image steganography for providing enhanced medical data security. Dans *2018 15th IEEE Annual Consumer Communications & Networking Conference (CCNC)*, pages 1–4. IEEE, 2018.

[4] Reza Akbari Movahed, Mohammad Reza Rezaeian, et Shirin Ghasemi. An image watermarking algorithm for medical computerized tomography images. Dans *2019 5th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS)*, pages 1–5. IEEE, 2019.

[5] Ruizhen Liu et Tieniu Tan. An svd-based watermarking scheme for protecting rightful ownership. *IEEE transactions on multimedia*, 4(1) :121–128, 2002.

[6] R Eswaraiah et E Sreenivasa Reddy. A fragile roi-based medical image watermarking technique with tamper detection and recovery. Dans *2014 Fourth International Conference on Communication Systems and Network Technologies*, pages 896–899. IEEE, 2014.

[7] Jianfeng Lu, Meng Wang, Junping Dai, Qianru Huang, Li Li, et Chin-Chen Chang. Multiple watermark scheme based on dwt-dct quantization for medical images. *J. Inf. Hiding Multim. Signal Process.*, 6(3) :458–472, 2015.

[8] S Prasanth Vaidya. Fingerprint-based robust medical image watermarking in hybrid transform. *The Visual Computer*, 39(6) :2245–2260, 2023.

[9] Himanshu Kumar Singh et Ashutosh Kumar Singh. Comprehensive review of watermarking techniques in deep-learning environments. *Journal of Electronic Imaging*, 32 :031804 – 031804, 2022.

[10] M Outtas, L Zhang, O Deforges, A Serir, W Hamidouche, et Y Chen. Subjective and objective evaluations of feature selected multi output filter for speckle reduction on ultrasound images. *Physics in Medicine Biology*, 63(18) :185014, sep 2018.

[11] https ://radiopaedia.org/cases.

[12] https ://radiologie-nogent.fr/examen/irm/.

[13] Sonika Thakur, Amit Singh, Basant Kumar, et S. Ghrera. *Improved DWT-SVD-Based Medical Image Watermarking Through Hamming Code and Chaotic Encryption*, pages 897–905. 01 2020.

[14] Allister Mason, James A Rioux, Sharon E. Clarke, Andreu F. Costa, Matthias Helge Schmidt, Valerie Keough, Thien Huynh, et Steven D. Beyea. Comparison of objective image quality metrics to expert radiologists' scoring of diagnostic quality of mr images. *IEEE Transactions on Medical Imaging*, 39 :1064–1072, 2020.