Apprentissage d'observateurs idéaux pour l'évaluation de la qualité d'images en Tomographie par Rayons X en vues éparses

Romain VO[1,2] Antoine OLISLAEGERS[2]

[1]CNRS, ENS de Lyon, Laboratoire de Physique, Lyon, France

[2]Université Paris-Saclay, CEA, List, F-91190, Palaiseau, France

Résumé – Dans cet article, nous tentons de répondre à une question primordiale en imagerie : *Comment évaluer la qualité d'une image ?* En nous concentrant sur des applications de tomographie par rayons X, nous présentons un cas limite d'évaluation à l'aide du PSNR, qui ne parvient pas à capturer la qualité visuelle d'une image (voir Fig. 1). Nous nous appuyons sur le concept d'*observateurs idéaux* pour proposer une nouvelle méthode d'évaluation de la qualité d'une image, et surtout **une métrique** capable de classer différents algorithmes de restauration en termes de *qualité* des images produites. En effet, en formulant le problème comme un *test d'hypothèse*, l'approche par *observateur idéal* offre une manière intuitive de quantifier la qualité d'une image.

Abstract – In this work, we try to find an answer to a longstanding question in computational imaging: *How to assess the quality of an image ?* Focusing on X-ray computed tomography (CT), we present a failure case of PSNR, which fails to capture the visual quality of a reconstruction (see Fig. 1). We build on the concept of *ideal observers* to propose a new way to assess the quality of an image, and most importantly **a metric** that is able to rank different image restoration algorithms in terms of the *quality* of the images they produce. The *ideal observer* framework provides an intuitive way to quantify the quality of an image.

1 Introduction

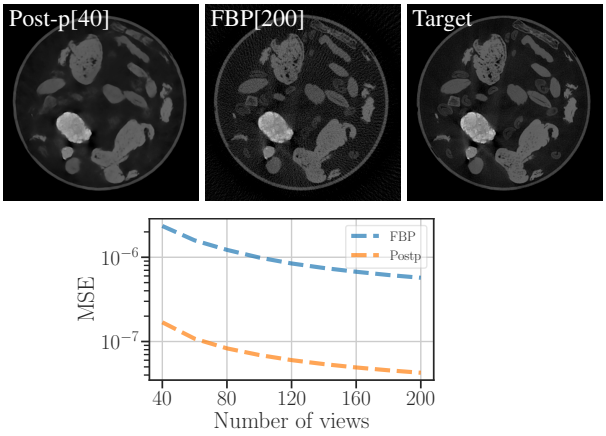

Figure 1 – Sur le jeu de données **2DeteCT** [7]: évolution de la MSE des reconstructions FBP (*courbe bleue*) et des reconstructions post-traitées par un réseau de neurones profond (*courbe orange*) en fonction du nombre de vues. De *gauche* à *droite*: une reconstruction FBP à 40 vues post-traitée par un réseau de neurones profond, une reconstruction FBP à 200 vues, et l'image cible.

L'évaluation de la qualité d'image (*Image Quality Assessment – IQA*) est une étape cruciale dans le développement de systèmes d'imagerie tels que la tomographie par rayons X (CT). Un processus d'évaluation de la qualité bien défini aide les utilisateurs à établir des normes, par la suite garantissant que les images produites sont de qualité suffisante pour la tâche à accomplir, qu'il s'agisse de diagnostic, de surveillance ou d'inspection industrielle. Cela permet également de comprendre les limites d'un système, d'identifier ses modes de défaillance et ses biais – autant d'étapes nécessaires à son amélioration.

Dans ce contexte, les métriques standardes d'évaluation à partir d'images de référence (*full reference* IQA – FR-IQA) telles que le maximum du rapport signal sur bruit (PSNR) et l'indice de similarité structurelle (SSIM) sont largement utilisées pour évaluer la qualité des images [14]. Cependant, ces métriques se retrouvent dans de nombreux cas inadéquates. Le PSNR a par exemple tendance à privilégier les images lisses au détriment des images riches en détail [2].

La Fig. 1 illlustre ce problème: on y observe que l'erreur quadratique (*Mean Squared Error – MSE*) d'une reconstruction à 40 vues post-traitée par un réseau de neurones profond (*post-processing*) est inférieure à celle d'une reconstruction par rétro-projection filtrée (*Filtered Back Projection – FBP*) à 200 vues, alors que cette dernière est visuellement de meilleure qualité. Ce constat est critique car il peut mener à la sélection d'algorithmes sous-optimaux. En effet, une grande partie de la recherche sur le dévelopemment de nouveaux algorithmes de reconstruction se concentre sur l'amélioration de ce type de métrique, qui ne sont pas toujours corrélées avec la réussite d'une tâche en aval sur ces mêmes images [2].

Dans cet article, nous explorons l'utilisation du critère de Bayes et plus spécifiquement des observateurs ideaux (*Ideal Observers – IO*) [15, 1, 11] comme cadre général pour évaluer la qualité d'une image. L'évaluation par observateur part de l'idée largement admise selon laquelle un système d'imagerie doit être optimisé pour une tâche spécifique [1, 2, 6], *e.g.* en imagerie médicale, cela signifie optimiser la capacité du médecin à détecter une lésion.

Pour une tâche de détection binaire, ce processus peut être formulé comme un test d'hypothèse: un observateur est présenté avec l'hypothèse nulle (H_0), *i.e.* l'image ne contient pas de signal, et l'hypothèse alternative (H_1), *i.e.* l'image contient un signal. Dans ce contexte, l'observateur idéal [15] fournit une formule pour calculer la statistique de test optimale et fixe une limite supérieure sur la performance de tout

observateur, humain ou algorithme, effectuant la tâche. Il est optimal car l'observateur idéal minimise la probabilité d'erreur moyenne, *i.e.* la probabilité de faire une mauvaise décision, étant donné les données [15]. Malheureusement, l'observateur idéal est analytiquement intractable et n'est utilisé que dans des modèles relativement simples où l'image et le signal à détecter suivent des distributions statistiques simples [9]. Une alternative coûteuse consiste à calculer une approximation linéaire tel que l'observeur d'*Hotelling* [3].

Des méthodes récentes explorent la possibilité d'apprendre l'observateur idéal à partir de données [8, 17, 16]. Les résultats obtenus sur des ensembles simples, avec un observateur idéal analytique, suggèrent que l'apprentissage de l'observateur idéal à partir de données est prometteur et peut être utilisé pour évaluer la qualité des images. Cependant, peu de travaux ont tenté d'utiliser des observateurs appris pour classer des algorithmes d'imagerie [16, 10].

Contributions Ce travail vise à fournir un aperçu clair de l'évaluation par observateur. Nous nous concentrons sur la tomographie par rayons X en vues éparses et comparons deux algorithmes de reconstruction: la rétro-projection filtrée (FBP) et une méthode de post-traitement [4] par apprentissage. Nous faisons varier le nombre de vues pour simuler différentes conditions d'imagerie et évaluons comment le post-traitement affecte la performance mesurée par un observateur appris.

2 Observateurs

Une tâche de détection binaire peut être formulée comme un *test d'hypothèse* : étant donné une image $\boldsymbol{x} \in \mathbb{R}^n$, un *observateur* doit choisir entre deux hypothèses: l'*hypothèse nulle* (H_0) , l'image $\boldsymbol{x}$ ne contient aucun signal d'intérêt, et l'*hypothèse alternative* (H_1), l'image contient un signal aléatoire $\boldsymbol{s} \in \mathbb{R}^n$:

$$\begin{cases} H_0 : \boldsymbol{x} = \boldsymbol{x}_b, \\ H_1 : \boldsymbol{x} = \boldsymbol{x}_b + \boldsymbol{s}, \end{cases} \tag{1}$$

où $\boldsymbol{x}_b \in \mathbb{R}^n$ désigne une image de "fond", soulignant que, quel que soit le signal, l'image $\boldsymbol{x}$ peut toujours être décomposée en un fond et un signal résiduel $\boldsymbol{s} \in \mathbb{R}^n$.

2.1 Observateur Idéal

Face à ces deux hypothèses (1), l'*observateur idéal* (IO) correspond au test statistique $\Lambda_{io}(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R}$ qui minimise la probabilité moyenne d'erreur. Il est défini comme le rapport de vraisemblance entre les deux hypothèses [15] :

$$\Lambda_{io}(\boldsymbol{x}) = \frac{p(\boldsymbol{x}|H_1)}{p(\boldsymbol{x}|H_0)}. \tag{2}$$

Il est *optimal* car il établit une borne supérieure sur la performance de détection de tout observateur sur cette tâche.

2.2 Observateur Idéal Appris

L'*observateur idéal* étant analytiquement intractable, l'approche par apprentissage consiste à observer que toute transformation monotone du rapport de vraisemblance $\Lambda_{io}(\boldsymbol{x})$ est également un test statistique optimal. Comme le note Zhou *et al.* [17], la probabilité postérieure $\mathbb{P}(H_1|\boldsymbol{x})$ est une telle transformation :

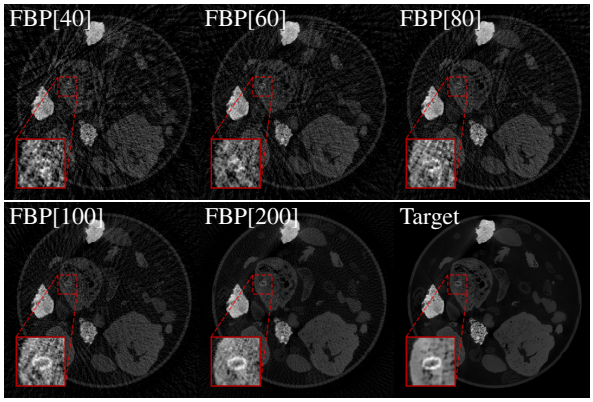

Figure 2 – **2DeteCT** : Disparition progressive du signal binaire synthétique en fonction du nombre de vues.

$$\mathbb{P}(H_1|\boldsymbol{x}) = \frac{(\mathbb{P}(H_1)/\mathbb{P}(H_0))\Lambda(\boldsymbol{x})}{1 + (\mathbb{P}(H_1)/\mathbb{P}(H_0))\Lambda(\boldsymbol{x})}. \tag{3}$$

Cela signifique qu'en utilisant (3), l'observateur ideal peut être appris en approximant la probabilité postérieure $\mathbb{P}(H_1|\boldsymbol{x})$ sur une distribution donnée. Soit $\mathrm{y} \in \{0, 1\}$ la variable aléatoire associée à chaque hypothèse, et $\mathbf{x}$ la variable aléatoire associée aux données, on peut montrer que minimiser la divergence de Kullback-Leibler entre la véritable probabilité postérieure et la postérieure approximée est équivalent à minimiser l'entropie croisée binaire (BCE) :

$$\begin{aligned} &\min_{\psi} \mathrm{KL}(\mathbb{P}(\mathrm{y}|\mathbf{x}) \,||\, \mathbb{P}(\mathrm{y}|\mathbf{x}, \psi)) \\ &= \min_{\psi} \mathbb{E}_{(\mathbf{x},\mathrm{y})} \left[-\mathrm{y} \log \Lambda_\psi(\mathbf{x}) - (1 - \mathrm{y}) \log \left(1 - \Lambda_\psi(\mathbf{x})\right) \right], \end{aligned} \tag{4}$$

ici $\Lambda_\psi(\boldsymbol{x})$ avec des paramètres ψ est la fonction approximant la probabilité postérieure de H_1 sachant $\boldsymbol{x}$, *i.e.* $\mathbb{P}(\mathrm{y} = H_1|\boldsymbol{x}, \psi)$. Naturellement, Λ_ψ peut être paramétré par un réseau de neurones profond, afin de pouvoir réutiliser des architectures efficaces [17] développées pour des tâches de classification.

3 Méthodologie

3.1 Problème Inverse

Dans ce travail, nous considérons le problème de tomographie par rayons X en vues éparses. Étant donné un opérateur de projection discret $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ et des mesures de projection $\boldsymbol{b} \in \mathbb{R}^m$, l'objectif est de trouver la véritable image $\boldsymbol{x}^\dagger \in \mathbb{R}^n$ telle que

$$\boldsymbol{b} = \boldsymbol{A}\boldsymbol{x}^\dagger + \boldsymbol{\varepsilon},$$

avec $\boldsymbol{\varepsilon} \in \mathbb{R}^m$ le bruit de mesure. L'opérateur de rétro-projection filtrée (FBP) $\boldsymbol{A}^+$ est utilisé pour générer des reconstructions bruitées $\tilde{\boldsymbol{x}} = \boldsymbol{A}^+\boldsymbol{b}$, ensuite post-traitées par un réseau de neurones profond D_ϕ pour obtenir une estimation de la vérité terrain $\hat{\boldsymbol{x}} = D_\phi(\tilde{\boldsymbol{x}})$. Comme illustré dans les Figures 1 and 2, nous simulons des conditions d'imagerie de qualité différente en faisant varier le nombre de vues $k > 0$ utilisées pour reconstruire les images. On suppose que les images cibles $\mathbf{x}^\dagger$ sont échantillonnées à partir d'une distribution $\pi_{\mathbf{x}^\dagger}$ et que les images d'entrée $\tilde{\mathbf{x}}_k$ sont échantillonnées à partir d'une distribution $\pi_{\tilde{\mathbf{x}}_k}$, où l'indice k désigne le nombre de vues utilisées pour reconstruire l'image $\tilde{\mathbf{x}}_k$. nous supposons également que les images post-traitées $\hat{\mathbf{x}}_k$ sont échantillonnées à partir d'une autre distribution $\pi_{\hat{\mathbf{x}}_k}$.

3.2 Détection de Signal Binaire

Pour tester cette méthode d'évaluation par observateur, nous utilisons le jeu de données **2DeteCT** [7]. Les mesures brutes sont agrégées afin de simuler un détecteur de 478 cellules, et des reconstructions de 512^2 pixels. Nous créons une tâche synthétique de détection de signal binaire. On fait l'hypothèse qu'une image $\boldsymbol{x}$ contient un signal $\boldsymbol{s}$, qui est la réalisation d'une variable aléatoire $\mathbf{s}$, et on le modélise comme une forme géométrique simple, *i.e.* une ellipse, avec une taille, une orientation et une position aléatoires dans l'image (Fig. 2).

3.3 Configurations d'Apprentissage

Post-traitement Pour le réseau de post-traitement, nous considérons une configuration simple, avec un réseau de type UNet [12] entraîné à prédire la vérité terrain $\boldsymbol{x}^\dagger$ à partir de la reconstruction FBP en vues éparses $\tilde{\boldsymbol{x}}$. Nous considérons un opérateur général, qui prend également k, le nombre de vues, comme entrée, *i.e.* $D_\phi(\cdot, k)$, qui est entraîné à prédire l'image propre à partir de n'importe quelle configuration de vues éparses :

$$\min_\phi \mathbb{E}_k \mathbb{E}_{\mathbf{x}_k, \mathbf{x}} \left[\|D_\phi(\tilde{\mathbf{x}}_k, k) - \mathbf{x}^\dagger\|_2^2 \right]. \tag{5}$$

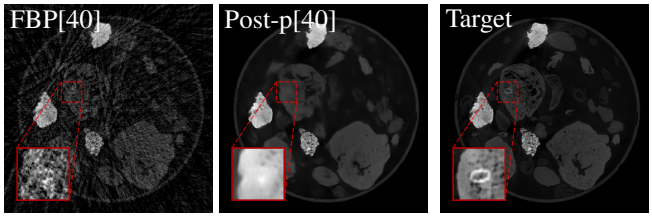

Figure 3 – **2DeteCT**: illustrations de reconstructions à 40 vues avec signal. Sur la reconstruction FBP, le signal est à peine visible, tandis que l'opération de post-traitement lisse complètement la région qui l'entoure.

Apprentissage de l'observateur Nous insistons sur le fait que pour approximer correctement l'observateur idéal (4), l'opérateur appris doit minimiser l'entropie croisée binaire pour la distribution d'images sur laquelle il est évalué. Cela signifie qu'utiliser un classifieur entraîné sur des images propres donne des résultats sous-optimaux lorsqu'il est évalué sur des échantillons FBP en vues éparses ou post-traités (voir Fig. 4–*gauche*). Par conséquent, pour maximiser la performance de l'*observateur* appris sur chaque cas, nous décidons d'entraîner un opérateur différent pour chaque configuration d'imagerie, c'est-à-dire que nous entraînons un classifieur spécialisé pour chaque nombre de vues k et pour chaque algorithme de reconstruction (voir Fig. 4–*droite*). Une approche moins coûteuse consisterait à *ajuster* un classifieur entraîné sur images propres pour chaque nouvelle configuration d'imagerie, mais nous ne l'explorons pas dans cet article.

Optimisation relative à la tâche Nous considérons l'objectif de minimisation dans (5) avec une légère modification, l'entropie croisée binaire est ajoutée comme terme de régularisation avec un paramètre $\lambda > 0$

$$\min_\phi \mathbb{E}_k \mathbb{E}_{(\tilde{\mathbf{x}}_k, \mathbf{x}^\dagger, \mathbf{y})} \Big[\|D_\phi(\tilde{\mathbf{x}}_k, k) - \mathbf{x}^\dagger\|_2^2 - \lambda \log \mathbb{P}(\mathbf{y} | D_\phi(\tilde{\mathbf{x}}_k, k), \hat{\psi}) \Big] \tag{6}$$

où $\hat{\psi}$ sont les paramètres de l'*observateur* entraîné sur des images propres.

3.4 Configuration d'évaluation

Pour évaluer la performance d'une configuration d'imagerie spécifique, nous calculons l'erreur quadratique moyenne (MSE) ainsi que l'AUC[ROC] de son *observateur idéal* appris (voir Fig. 4).

4 Résultats

4.1 Approximer l'observateur idéal

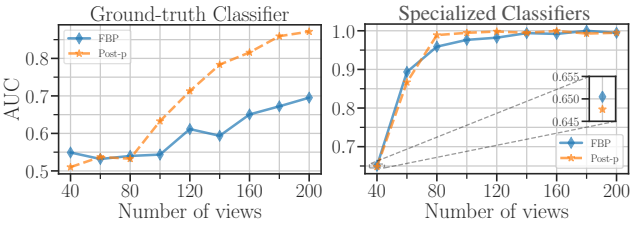

Figure 4 – **2DeteCT**: *(gauche)* évolution de l'AUC calculée par un classifieur entraîné sur des images propres, évalué sur des images bruitées $\tilde{\mathbf{x}}_k$ et des images post-traitées $\hat{\mathbf{x}}_k$. *(droite)* évolution de l'AUC calculée par des classifieurs spécialisés.

Dans la section suivante, nous nous réfèrerons interchangeablement à l'AUC comme la métrique *basée sur la tâche*. Relativement à notre problème initial Figure 1, on note que sur la Fig. 4, l'AUC classe correctement les reconstructions FBP avec 200 vues comme meilleures qu'une reconstruction post-traitée avec 40 vues.

Nous analysons maintenant si, pour un nombre donné de vues, l'opération de post-traitement améliore réellement les performances en classification. De manière remarquable, malgré l'amélioration de la MSE d'un ordre de grandeur (Fig. 1), l'AUC à 40 et 60 vues diminue après l'opération de post-traitement. Il apparait alors que le signal soit si faible que l'opération de post-traitement supprime le signal de l'image, ou du moins diminue son intensité. Cela est confirmé par les résultats qualitatifs dans Fig. 3 où nous voyons que le post-traitement avec 40 vues lisse complètement l'image autour du signal. Pour un nombre plus élevé de vues, on remarque qu'avec un observateur correctement entraîné, le gain de performance apporté par le post-traitement de la reconstruction FBP reste limité Fig. 4. Nous faisons cette première observation et émettons l'hypothèse que maximiser le PSNR d'une reconstruction peut avoir un impact limité sur la réussite de détection en aval. Une approche plus spécifique consiste à optimiser le système d'imagerie, y compris l'algorithme de reconstruction, en fonction de la tâche en aval, ici la classification. C'est l'idée derrière l'optimisation *basée sur la tâche* (6).

4.2 Optimisation *basée sur la tâche*

Nous complètons notre analyse en comparant les performances de l'algorithme de reconstruction *basé sur la tâche* avec la méthode de post-traitement standarde. Comme attendu, la régularisation de l'algorithme de reconstruction améliore les performances en classification (Fig. 5).

Nous fournissons également des résultats qualitatifs pour illustrer l'impact de l'optimisation *basée sur la tâche* sur la visibilité du signal pour de faibles nombres de vues.(Fig. 6).

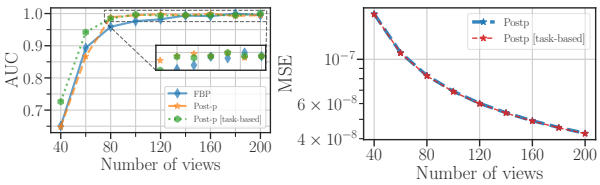

Figure 5 – **2DeteCT**: nous complètons les résultats Figure 4 en comparant les performances de l'algorithme de reconstruction *basé sur la tâche* (courbe *verte*). Nous montrons l'AUC de l'observateur appris, et la MSE des reconstructions.

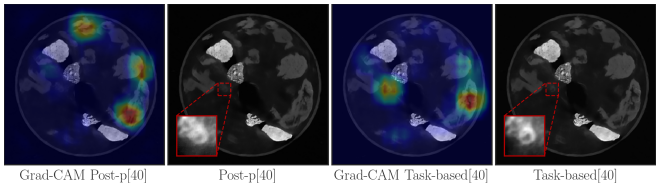

Figure 6 – **2DeteCT**: nous illustrons les améliorations apportées par l'optimisation *basée sur la tâche* de l'algorithme de reconstruction. Nous affichons les résultats à 40 vues et complétons l'illustration avec des cartes de chaleur Grad-CAM [13] pour interpréter le changement de performance en classification. Ici, l'*observateur* "remarque" la présence d'un signal - nous interprétons l'amélioration de l'AUC comme une diminution de l'entropie du classifieur.

5 Conclusion

Nous avons présenté une méthodologie d'évaluation par observateur pour quantifier la qualité d'une image relativement à une tâche en aval. Nous avons vu que la performance de la tâche reflète correctement la qualité globale des images produites. Nous avons montré comment le développement (et l'optimisation) d'un algorithme de reconstruction devait s'effectuer relativement à la tâche pour laquelle il est deployé. En ce sens, nous avons observé que les algorithmes *basé-tâche* donnaient les meilleurs résultats, validant effectivement l'approche. L'étape suivante serait d'évaluer l'approche par observateur sur des algorithmes plus avancés, *e.g.* méthodes *déroulées*, *plug-and-play* ou *diffusion*, afin d'observer si l'optimisation *basée sur la tâche* est bien déterminante pour la qualité de l'image ou si des algorithmes plus complexes peuvent compenser l'absence de régularisation par la tâche.

References

[1] Harrison H. Barrett, Jie Yao, Jannick P. Rolland et Kyle J. Myers : Model observers for assessment of image quality. *In National Academy of Sciences*, 1993.

[2] Anna Breger, Ander Biguri, Malena Sabaté Landman, Ian Selby, Nicole Amberg, Elisabeth Brunner, Janek Gröhl, Sepideh Hatamikia, Clemens Karner, Lipeng Ning, Sören Dittmer, Michael Roberts, AIX-COVNET Collaboration et Carola-Bibiane Schönlieb : A study of why we need to reassess full reference image quality assessment with medical images. *Journal of Imaging Informatics in Medicine*, 2025.

[3] Sebastian Eckel, Peter Huthwaite, Michael Lowe, Andreas Schumm et Pierre Guérin : Establishment and validation of the Channelized Hotelling Model Observer for image assessment in industrial radiography. *NDT & E International*, 2018.

[4] Yo Seob Han, Jaejun Yoo et Jong Chul Ye : Deep Residual Learning for Compressed Sensing CT Reconstruction via Persistent Homology Analysis. 2016.

[5] Zohaib Amjad Khan, Giuseppe Valenzise, Aladine Chetouani et Frédéric Dufaux : Towards an image utility assessment framework for machine perception. *In European Signal Processing Conference*, 2022.

[6] Maximilian B. Kiss, Sophia B. Coban, K. Joost Batenburg, Tristan van Leeuwen et Felix Lucka : 2DeteCT - A large 2D expandable, trainable, experimental Computed Tomography dataset for machine learning. *Scientific Data*, 2023.

[7] Matthew A. Kupinski, Darrin C. Edwards, Maryellen L. Giger et Charles E. Metz : Ideal observer approximation using Bayesian classification neural networks. *IEEE Transactions on Medical Imaging*, 2001.

[8] Matthew A. Kupinski, John W. Hoppin, Eric Clarkson et Harrison H. Barrett : Ideal-observer computation in medical imaging with use of Markov-chain Monte Carlo techniques. *Journal of the Optical Society of America*, 2003.

[9] Kaiyan Li, Weimin Zhou, Hua Li et Mark A. Anastasio : Assessing the Impact of Deep Neural Network-Based Image Denoising on Binary Signal Detection Tasks. *IEEE Transactions on Medical Imaging*, 2021.

[10] Robert N. McDonough, Anthony D. Whalen et Anthony D. Whalen : *Detection of Signals in Noise*. Academic Press, 2nd ed édition, 1995.

[11] Olaf Ronneberger, Philipp Fischer et Thomas Brox : U-Net: Convolutional Networks for Biomedical Image Segmentation. *In Medical Image Computing and Computer-Assisted Intervention*, 2015.

[12] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh et Dhruv Batra : Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. 2020.

[13] Zhou Wang, Alan Conrad Bovik, Hamid Rahim Sheikh et Eero P. Simoncelli : Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 2004.

[14] Anthony Whalen : Statistical theory of signal detection and parameter estimation. *IEEE Communications Magazine*, 1984.

[15] Xiaohui Zhang, Varun A. Kelkar, Jason Granstedt, Hua Li et Mark A. Anastasio : Impact of deep learning-based image super-resolution on binary signal detection. *Journal of Medical Imaging*, 2021.

[16] Weimin Zhou, Hua Li et Mark A. Anastasio : Approximating the Ideal Observer and Hotelling Observer for binary signal detection tasks by use of supervised learning methods. *IEEE Transactions on Medical Imaging*, 2019-10.