MAP, déconvolution aveugle et a priori appris : un subtil mélange

Minh Hai NGUYEN[1,2] Edouard PAUWELS[3] Pierre WEISS[1,2]

[1]Institut de Recherche en Informatique de Toulouse (IRIT), 31400 Toulouse, France

[2]Centre de Biologie Intégrative (CBI), 31400 Toulouse, France

[3]Toulouse School of Economics, 31000 Toulouse, France

Résumé – Nous étudions les propriétés de l'estimateur du Maximum A Posteriori (MAP) pour la déconvolution aveugle, lorsqu'on utilise des a priori reposant sur les modèles de diffusion. Nous montrons empiriquement que pour ces a priori, les images floues sont plus probables. En conséquence, l'estimateur MAP produit des filtres proches de la masse de Dirac et des images floues. Nous montrons toutefois que la loi a posteriori possède un minimum local favorable. Nous proposons aussi une stratégie d'optimisation qui permet d'éviter le minimum global. Elle repose sur une initialisation adéquate avec une méthode d'optimisation alternée spécifique.

Abstract – We study the properties of the Maximum A Posteriori (MAP) estimator for blind deconvolution when using priors based on diffusion models. We empirically show that with these priors, blurry images are more likely. As a result, the MAP estimator produces filters close to the Dirac mass and blurry images. However, we demonstrate that the posterior distribution has a favorable local minimum. This leads us to propose an optimization strategy that avoids the global minimum. This strategy relies on proper initialization with a specific alternating optimization method.

1 Introduction

La déconvolution aveugle consiste à retrouver une image $\bar{x}$ et un noyau de convolution $\bar{h}$ à partir d'une observation floue et bruitée y :

$$y = \bar{x} \star \bar{h} + b, \tag{1}$$

où $\star$ est l'opérateur de convolution et b est un bruit gaussien additif d'écart-type σ. D'un point de vue bayésien, nous supposons que $\bar{x}$ et $\bar{h}$ sont des réalisations de deux vecteurs aléatoires indépendants $\boldsymbol{x}$ et $\boldsymbol{h}$ avec des lois respectives $p_{\boldsymbol{x}}$ et $p_{\boldsymbol{h}}$. Pour simplifier, on suppose aussi que la loi $p_{\boldsymbol{h}}$ est uniforme sur un certain domaine $\mathcal{H}$. La loi a posteriori satisfait donc :

$$-\log p(x, h|y) \propto \frac{\|h \star x - y\|_2^2}{2\sigma^2} - \log p_{\boldsymbol{x}}(x). \tag{2}$$

L'estimateur le plus utilisé dans la littérature est sans doute le *Maximum A Posteriori* (MAP), qui cherche à maximiser la loi a posteriori $p(x, h|y)$ par rapport au couple (x, h), ou de manière équivalente :

$$(\hat{x}, \hat{h}) = \underset{x,h}{\arg\min} \frac{1}{2} \|h \star x - y\|_2^2 + \sigma^2 q(x), \tag{3}$$

avec $q(x) = -\log p_{\boldsymbol{x}}(x)$, le potentiel de l'image. Plus cette valeur est petite, plus l'image est probable. On note aussi $L_y(x, h)$ la fonction à minimiser dans (3).

Le choix du potentiel a priori $q(x)$ est crucial. Les approches classiques reposent sur des notions de parcimonie, qui peut être exprimée dans le domaine spatial, des ondelettes ou du gradient (variation totale). Cependant, ces a priori « artisanales » souffrent d'un problème majeur [3, 5, 4, 1] : le minimiseur global de (3) correspond à la solution « triviale », c'est-à-dire un noyau de Dirac pour $\hat{h}$ et une image floue proche de y pour $\hat{x}$. Cette défaillance s'explique par le fait que les a priori parcimonieux favorisent les images floues par rapport aux images nettes. Autrement dit, le potentiel $q(x)$ est plus faible lorsque l'image x est floue. Cette propriété peut être démontrée formellement grâce à des inégalités de type Hölder. Certains travaux [1, 5] ont toutefois montré que l'image et le noyau réels peuvent être trouvés à un maximum local de la densité a posteriori. Ainsi, ils cherchent des *bons* maxima locaux de la loi a posteriori.

Les modèles génératifs, tels que les modèles de diffusion [6, 2], permettent d'apprendre un a priori plus expressif sur les images naturelles. Cela soulève la question abordée ici : *Les modèles de diffusion peuvent-ils résoudre les problèmes du MAP pour la déconvolution aveugle ?*

Nous répondons à cette question par une analyse théorique et empirique ; et mettons en évidence trois éléments clés : 1/ Même avec des a priori avancés reposant sur des modèles de diffusion, l'estimateur MAP favorise toujours la solution triviale, révélant une limite fondamentale de cette approche. 2/ Les points critiques du second ordre du potentiel q, correspondent à des minimiseurs stables de (3). 3/ Nous proposons une stratégie d'initialisation simple pour éviter la convergence vers la solution triviale.

2 Les images floues sont plus probables

Nous présentons brièvement les modèles de diffusion et montrons comment évaluer le potentiel a priori sous-jacent, q. Nous montrons également que les modèles de diffusion pré-entrainés favorisent les images floues.

2.1 Préliminaires sur les modèles de diffusion

Les modèles de diffusion reposent sur *l'Equation Différentielle Stochastique* (EDS) suivante [6] :

$$d\boldsymbol{x}_t = f(\boldsymbol{x}_t, t)\, dt + g(t)\, d\boldsymbol{w}_t, \tag{4}$$

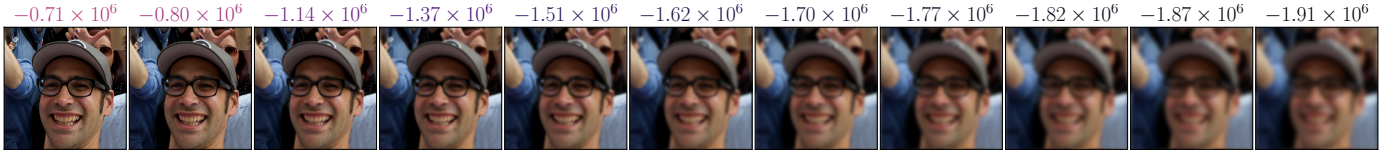

FIGURE 1 : Variation du potentiel $q(h_\theta \star x)$ avec un flou gaussien. Une valeur plus faible indique une image plus probable.

où f est une fonction de drift, g est une fonction de diffusion, et $\boldsymbol{w}_t$ est un processus brownien. Dans ce modèle, $\boldsymbol{x}_0$ suit la loi $p_{\boldsymbol{x}}$ et, pour un temps T assez grand et certaines f, g, la loi de $\boldsymbol{x}_T$ est proche d'une loi gaussienne, avec des paramètres dépendant de f et g. Ce processus peut être inversé en temps avec l'EDS suivante, qui est ensuite discrétisée en pratique :

$$d\boldsymbol{x}_t = \left[f(\boldsymbol{x}_t, t) - g(t)^2 \nabla \log p_t(\boldsymbol{x}_t)\right] dt + g(t) d\boldsymbol{w}_t. \quad (5)$$

Cela permet de simuler des échantillons selon la loi $p_{\boldsymbol{x}}$, en $t = 0$, à partir de réalisations gaussiennes pour $\boldsymbol{x}_T$. L'entraînement d'un modèle de diffusion revient à estimer la fonction de score $\nabla \log p_t$, par « score matching ». L'estimateur du score est généralement donné par un réseau de neurones. Plus de détails techniques peuvent être retrouvés dans l'article original [6].

Calculer le potentiel d'une image Il existe aussi un processus *déterministe* (équation de Fokker-Planck) dont les trajectoires partagent la même distribution marginale, défini par *l'Equation Différentielle Ordinaire* (EDO) :

$$d\boldsymbol{x}_t = v(\boldsymbol{x}_t, t) dt,$$

où $v(\boldsymbol{x}_t, t) \overset{\text{def}}{=} f(\boldsymbol{x}_t, t) + \frac{1}{2} g(t)^2 \nabla \log p_t(\boldsymbol{x}_t)$. Cette EDO nous permet aussi de calculer le potentiel d'une image x_0, à l'aide de la formule suivante [6] :

$$q(x_0) = -\log p_{\boldsymbol{x}}(x_0) = -\log p_T(x_T) - \int_0^T \nabla \cdot v(x_t, t)\, dt.$$

Dans cette expression, p_T est la densité d'une gaussienne de moyenne nulle et de covariance connue, la trajectoire x_t est obtenue en partant de x_0 et en résolvant l'EDO par un schéma numérique, où le terme de divergence dans l'intégrale est estimé à l'aide de l'estimateur de Hutchinson-Skilling et la différentiation automatique. Ce processus est coûteux en calcul, nécessitant des centaines d'évaluations de $\nabla \log p_t$.

Modèle pré-entrainés Nous utilisons les modèles (UNet) entrainés sur FFHQ-256 de [6] ; FFHQ-64, AFHQ-64 et ImageNet-64 de [2].

2.2 Observation empirique

L'échec du MAP avec des a priori reposant sur la parcimonie peut être expliqué par le fait que *les images floues sont plus vraisemblables que les images nettes*. Ce phénomène a été démontré pour la variation totale et la parcimonie [3, 4, 1]. Nous examinons empiriquement si cette propriété s'applique également aux a priori issus des modèles de diffusion.

Expérience Nous utilisons un ensemble d'images nettes que nous convoluons avec quatre types de flous (Figure 2). En partant d'une image x de la base de données, nous mesurons le potentiel $q(h_\theta \star x)$ pour différentes valeurs de θ.

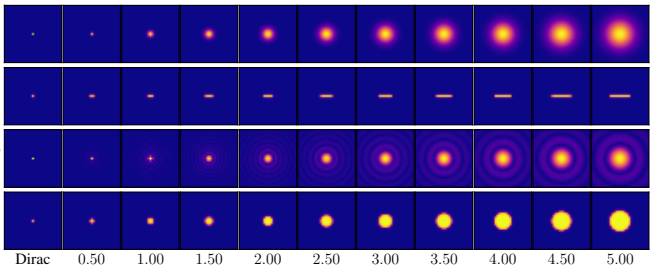

FIGURE 2 : Une paramétrisation univariée avec un paramètre θ. On a $h_0 = \delta$, et plus θ est élevé, plus le flou est important. De haut en bas : gaussien, mouvement, Airy, défocalisation.

Les résultats pour 100 images de la base d'entraînement du modèle, pour différents modèles, sont présentés en Figure 3. On observe que le potentiel $q(h_\theta \star x)$ diminue strictement avec θ, indiquant que les images floues sont plus vraisemblables. Une illustration qualitative sur une image est présentée en Figure 1. Cette observation montre que les potentiels des modèles de diffusion, ont tendance à favoriser les images floues, comme les a priori de parcimonie.

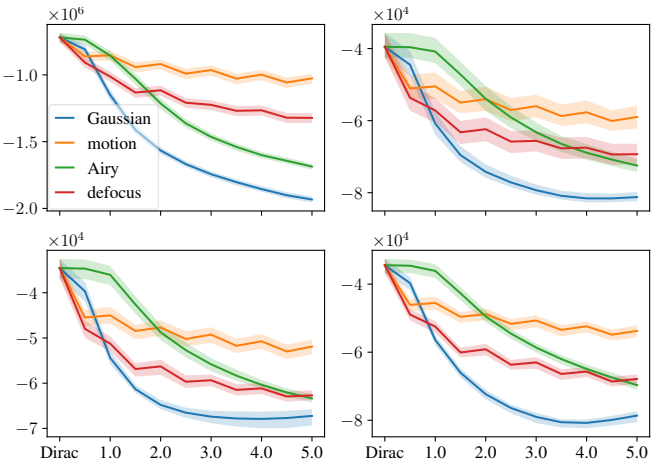

FIGURE 3 : Evolution du potentiel $q(h_\theta \star x)$ pour différents modèles. En haut : FFHQ-256, ImageNet-64. En bas : FFHQ-64 et AFHQ-64.

3 Propriétés de la loi a posteriori

3.1 La solution globale du MAP est triviale

Nous proposons ci-dessous une condition suffisante simple sous laquelle le MAP possède un minimum global indésirable :

Enoncé 1. *Soit $\mathcal{H} = \{h_\theta, \theta \in \mathbb{R}^K\}$ un ensemble des noyaux de flou. Si $\delta \in \mathcal{H}$, où δ est le dirac, et que, pour toute image x*

et tout noyau $h_\theta \in \mathcal{H}$,

$$q(h_\theta \star x) \leq q(x), \tag{6}$$

alors l'ensemble des minimiseurs globaux dans (3)*, s'il n'est pas vide, contient le noyau de Dirac et l'image débruitée obtenue avec le MAP.*

La condition (6) a été vérifiée empiriquement dans la partie précédente, pour différents modèles de diffusion et différentes familles de flou. L'énoncé 1 fournit ainsi une explication (les images floues sont favorisées par le potential a priori) à l'observation empirique selon laquelle le déconvolution aveugle avec l'estimateur MAP aboutit à une solution triviale.

Ce résultat est similaire aux résultats de Levin *et al.* [3] à propos de la variation totale. Cependant, leur preuve repose sur une mise à l'échelle qui ne s'applique pas lorsque le noyau est contraint dans un compact. Les arguments ont été affinés dans [4] avec une contrainte de simplexe. Notre approche fournit une condition générale qui englobe les précédentes.

3.2 Minima locaux du potentiel a posteriori

Le résultat le plus technique de ce papier stipule que, pour la déconvolution aveugle, les points critiques du second ordre du potentiel q (les points $\bar{x}$ tels que $\nabla q(\bar{x}) = 0$ et $\nabla^2 q(\bar{x}) \succeq 0$) sont des minima locaux stables de la loi a posteriori.

Théoreme 1. *Soit $\bar{x}$ un point critique du second ordre de q et $J_{\bar{x}}(\theta) = \frac{\partial}{\partial \theta}(h_\theta \star \bar{x})$ la jacobienne de $\theta \mapsto h_\theta \star \bar{x}$. Si :*

$$\ker \nabla^2 q(\bar{x}) \cap \ker h_{\bar{\theta}} = \{0\}, \tag{7}$$

$$\ker J_{\bar{x}}(\bar{\theta}) = \{0\}, \tag{8}$$

$$(h_{\bar{\theta}} \star \ker \nabla^2 q(\bar{x})) \cap \operatorname{Im} J_{\bar{x}}(\bar{\theta}) = \{0\}, \tag{9}$$

alors pour tout $\sigma > 0$, il existe $r, \epsilon > 0$ tels que pour tout bruit b avec $\|b\| \leq \epsilon$, il existe un minimiseur local unique $(x_b^\star, \theta_b^\star)$ de (3) *dans la boule de rayon r autour de $(\bar{x}, \bar{\theta})$ avec :*

$$\|x_b^\star - \bar{x}\| = \mathcal{O}(\|b\|) \quad \textit{et} \quad \left\|\theta_b^\star - \bar{\theta}\right\| = \mathcal{O}(\|b\|).$$

Le théorème 1 est positif : sous certaines conditions géométriques, la loi a posteriori possède un minimum stable autour du couple $(\bar{h}, \bar{x})$ qu'on souhaite retrouver. En particulier, il caractérise précisément les images $\bar{x}$ qui peuvent être retrouvées par la procédure MAP lorsqu'on utilise des a priori appris : ce sont les points critique du second ordre du potentiel q. On peut mettre ces résultats en perspective de l'échantillonnage compressif qui stipule que les points d'intérêts sont les points parcimonieux. Ce résultat ne permet toutefois pas de caractériser la géométrie du bassin d'attraction que nous explorons de manière empirique dans la section suivante.

4 Stratégies d'optimisation

4.1 Présentation générale de la méthode

Nous proposons ci-dessous l'algorithme 1 qui semble permettre de converger vers des maxima locaux favorables de la loi a posteriori. Il repose sur 3 idées :

- **Initialisation :** $x_0 = y$ et un noyau $h(\theta_0)$ plus « grand » que le vrai noyau $h(\bar{\theta})$.
- **Optimisation :** minimisation exacte sur l'image x (N étapes de gradient) et minimisation partielle (1 étape de gradient) sur le paramètre du noyau θ.
- **Réinitialisation :** on repart à intervalle régulier de l'image y lors de la minimisation sur x pour éviter de tomber dans de mauvais minima locaux de $L_y(\cdot, h)$.

L'intuition derrière cet algorithme est la suivante : en partant d'un noyau trop « gros » et en effectuant une minimisation exacte sur la variable x, on va obtenir une image « trop nette ». Ainsi, l'étape de minimisation sur le noyau h favorise un noyau « trop gros » et évite ainsi de tomber dans le bassin d'attraction de la masse de Dirac, qui est le minimiseur global du négatif log-posterieur (3).

Algorithme 1 : Algorithme évitant le *piège du Dirac*

Pour $n = 0 \dots N_{\text{ext}} - 1$ **faire**
  $\hat{x}_{n+1,0} = \begin{cases} y & \text{si } \operatorname{mod}(n, 100) = 0 \\ x_{n,N-1} & \text{sinon} \end{cases}$
  Pour $i = 0 \dots N_{\text{in}} - 1$ **faire**
    $\hat{x}_{n+1,i+1} = \hat{x}_{n+1,i} - \gamma_x \nabla_x L_y(\hat{x}_{n+1,i}, h(\hat{\theta}_n))$
  fin
  $\hat{\theta}_{n+1} \leftarrow \hat{\theta}_{n+1} - \gamma_h \nabla_\theta L_y(\hat{x}_{n+1}, h(\hat{\theta}_n))$
fin

4.2 Résultats numériques

Nous considérons les noyaux de convolution $h(\theta) = P(\theta)$ avec $\theta \in \mathbb{R}^{K \times K}$, $K = 21$ et P est la projection orthogonale sur le simplexe. On considère 2 initialisations différentes :

- *Noyau étalé (grand)* : $h(\hat{\theta}_0) = \frac{1}{K^2}\mathbf{1}_{K \times K}$ (image constante).
- *Petit noyau* : $h(\hat{\theta}_0) =$ gaussienne d'écart-type 1.5.

Les résultats obtenus avec l'algorithme 1 sont présentés sur la Figure 4 et la Figure 5. On utilise 4 noyaux de convolution différents et le potentiel de [6], entrainé sur FFHQ-256. On remarque qu'on retrouve une bonne estimation du couple $(\bar{x}, h(\bar{\theta}))$ avec une initialisation étalée $h(\hat{\theta}_0)$. L'initialisation avec un petit noyau mène à la solution triviale.

5 Conclusion

Le MAP pour la déconvolution aveugle possède une limitation fondamentale : il coincide avec la sulution triviale, même avec des a priori de diffusion. Nous avons cependant démontré que si l'image à retrouver est proche d'un minimum local de la loi a priori, la loi a posteriori possède un minimum local favorable. Celui-ci peut-être obtenu numériquement, grâce à un algorithme de minimisation alternée spécifique couplé avec une initialisation « étalée » ou « grande » du noyau.

Remerciement Ce travail a bénéficié d'une aide de l'ANR Micro-Blind ANR-21-CE48-0008 et des ressources HPC de GENCI-IDRIS (Grant 2021-AD011012210R3).

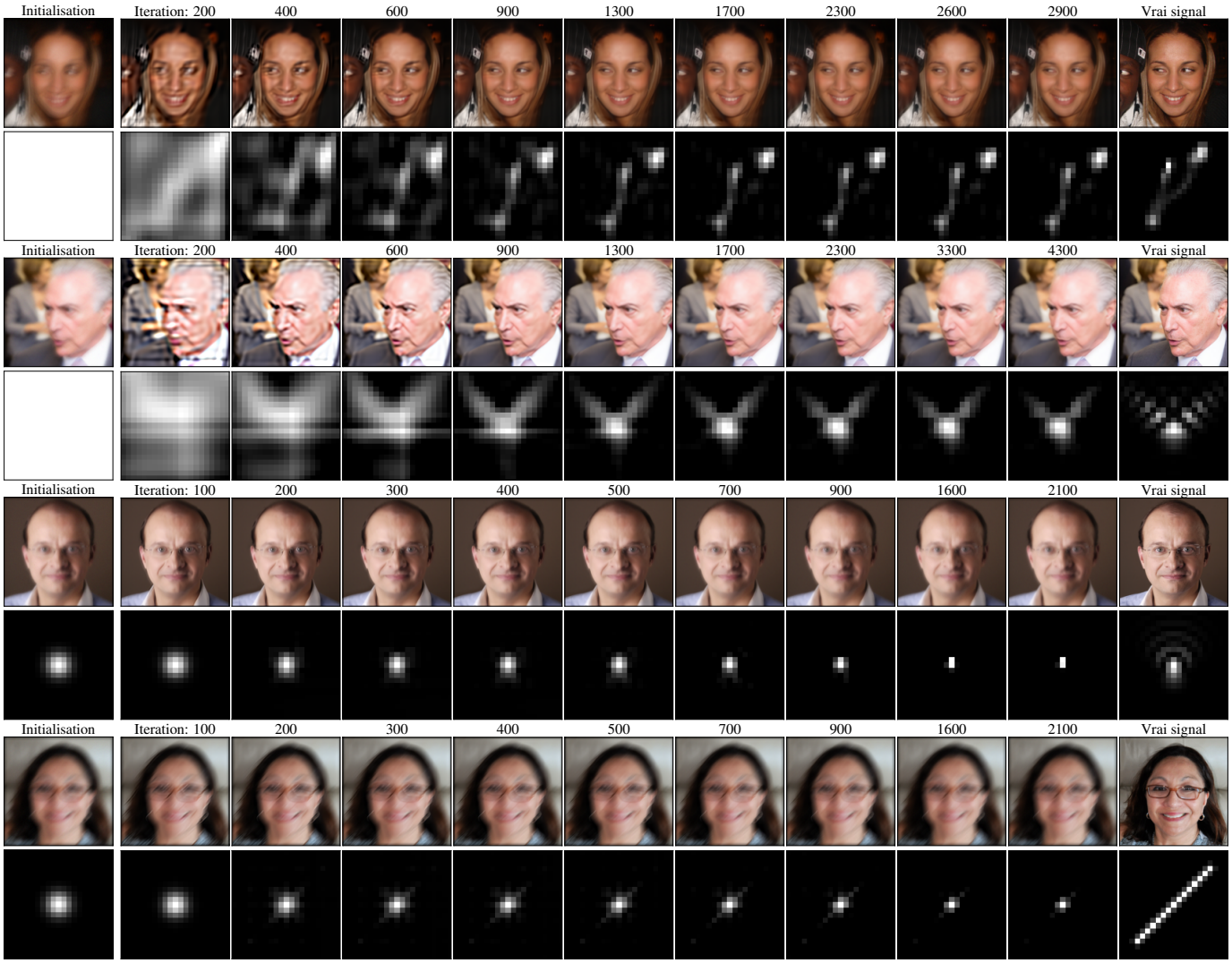

FIGURE 4 : Illustration du résultat des itérations de l'Algorithme 1 et de l'influence de l'initialisation sur le noyau $h(\hat{\theta}_0)$. Une initialisation « petite » de noyau donne une convergence vers la solution triviale. On utilise le modèle FFHQ-256 de [6].

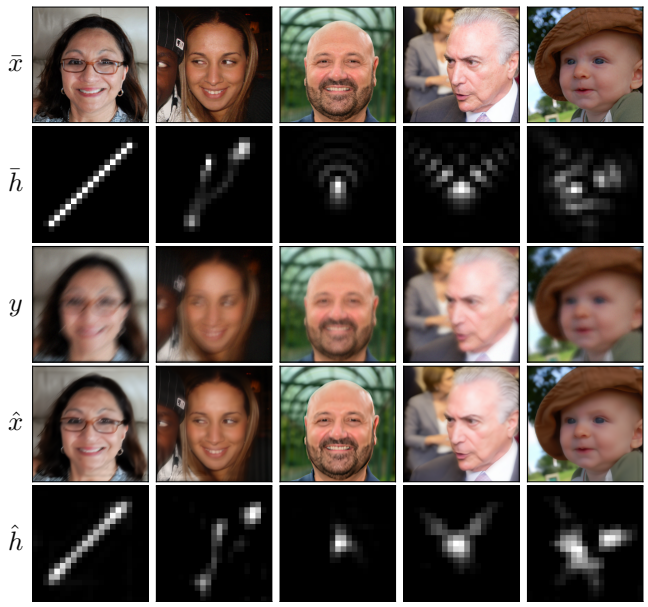

FIGURE 5 : Résultats numériques sur différentes couples $(\bar{x}, \bar{h})$. L'algorithme 1 est initialisé avec $h(\hat{\theta}_0) = \frac{1}{K^2}\mathbf{1}_{K\times K}$. Le niveau de bruit est de $\sigma = 0.01$.

Références

[1] Alexis BENICHOUX, Emmanuel VINCENT et Rémi GRIBONVAL : A fundamental pitfall in blind deconvolution with sparse and shift-invariant priors. *In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6108–6112. IEEE, 2013.

[2] Tero KARRAS, Miika AITTALA, Timo AILA et Samuli LAINE : Elucidating the design space of diffusion-based generative models. *In Proc. NeurIPS*, 2022.

[3] Anat LEVIN, Yair WEISS, Fredo DURAND et William T. FREEMAN : Understanding and evaluating blind deconvolution algorithms. *In 2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1964–1971, 2009.

[4] Daniele PERRONE et Paolo FAVARO : Total variation blind deconvolution : The devil is in the details. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2909–2916, 2014.

[5] Renu M RAMESHAN, Subhasis CHAUDHURI et Rajbabu VELMURUGAN : Joint map estimation for blind deconvolution : When does it work ? *In Proceedings of the Eighth Indian Conference on Computer Vision, Graphics and Image Processing*, pages 1–7, 2012.

[6] Yang SONG, Jascha SOHL-DICKSTEIN, Diederik P KINGMA, Abhishek KUMAR, Stefano ERMON et Ben POOLE : Score-based generative modeling through stochastic differential equations. *In International Conference on Learning Representations*, 2020.