

MATHÉMATIQUES  
VISION  
APPRENTISSAGE

---

## **Projet d'Imagerie Numérique : Inpainting par réseau profond avec termes globaux et locaux**

---

Victoria Bami et Clarine Vongpaseut

*encadrées par*  
Nicolas Cherel & Yann Gousseau

[https://github.com/Victoria-brami/Inpainting\\_project\\_mva](https://github.com/Victoria-brami/Inpainting_project_mva)

Décembre - Février 2022

# Table des matières

<b>1</b>	<b>Introduction</b>	<b>2</b>
1.1	Principe d'inpainting . . . . .	2
1.2	Les différentes approches pour l'inpainting (Etat de l'art) . . . . .	2
1.3	Nos objectifs . . . . .	2
<b>2</b>	<b>Inpainting par réseaux profonds : Modèle d'Iizuka et al.</b>	<b>3</b>
2.1	Architecture . . . . .	3
2.2	Entraînement du modèle . . . . .	3
<b>3</b>	<b>Expérimentations sur le modèle</b>	<b>4</b>
3.1	Données et Critères d'Evaluation . . . . .	4
3.1.1	Choix du dataset . . . . .	4
3.1.2	Métriques d'Evaluation Utilisées . . . . .	4
3.2	Etude d'ablation du réseau . . . . .	5
3.2.1	Principe . . . . .	5
3.2.2	Résultats visuels (qualitatifs) . . . . .	6
3.2.3	Résultats quantitatifs . . . . .	6
3.3	Quantification de l'importance des couches du réseau dans la restauration des images . . . . .	7
3.4	Autres Expériences réalisées . . . . .	10
<b>4</b>	<b>Comparaison avec une autre méthode d'inpainting par patches</b>	<b>10</b>
4.1	Principe . . . . .	10
4.2	Analyse comparative avec le modèle d'Iizuka et al. . . . .	10
4.3	Comparaison quantitative avec le modèle d'Iizuka et al. . . . .	13
<b>5</b>	<b>Conclusion et perspectives</b>	<b>13</b>

# 1 Introduction

## 1.1 Principe d'inpainting

L'inpainting est un processus consistant à remplir des zones blanches ("zones manquantes") dans une image afin de reconstruire une image complète. L'inpainting peut permettre de répondre à une multitude de problèmes en imagerie numérique. Cette technique est par exemple fréquemment utilisée pour la restauration d'anciens tableaux, dont certains pans se sont désagrégés. L'inpainting est aussi une solution pour restaurer des vieilles photographies qui ont pu être rayées, tachées ou altérées avec le temps.



FIGURE 1 – Exemple d'inpainting appliqué à la restauration d'oeuvres anciennes

L'inpainting est aussi couramment utilisé aujourd'hui pour créer des trucages photo (par exemple des retraits de meubles pour donner l'illusion qu'un objet vole, etc.), des effets spéciaux ou encore des photomontages.

## 1.2 Les différentes approches pour l'inpainting (Etat de l'art)

Il existe de nombreuses approches pour le problème d'inpainting. Une des approches traditionnelles est celle basé sur la diffusion comme [1], ces méthodes sont généralement efficaces pour reconstruire des petites zones. Les méthodes basées sur des patches, quant à elles, permettent de compléter des régions plus larges comme [2]. Cependant, en raison d'un manque de degré de compréhension élevé de l'image ces deux approches ont généralement des difficultés à remplir des zones hautement complexes. Les approches par apprentissage avec des réseaux convolutionnels [3, 4, 5] cherchent à remédier à ce problème.

## 1.3 Nos objectifs

L'approche du problème d'inpainting par des techniques de Deep Learning a permis d'améliorer considérablement les contenus visuels des images reconstruites. A la différence des approches par patches, qui se basent sur les propriétés des régions voisines de l'images, les réseaux de neurones apprennent à inférer sur une très large banque de données une zone relativement consistante avec le reste de l'image. L'intérêt de cette approche repose sur le fait que ces zones inférées ne proviennent pas de d'autres régions adjacentes de l'image à compléter.

Dans le cadre de notre projet, nous nous sommes en particulier intéressées à une approche d'inpainting par réseaux convolutionnels génératifs [4]. Ce modèle prend en compte à la fois la consistance locale de l'image rendue par inpainting, ainsi que la consistance globale. Nous avons réalisé des études d'ablation à différentes échelles afin de quantifier l'importance des différentes parties du modèle puis réalisé une évaluation qualitative et quantitative. Par la suite nous avons comparé les performances (analyse qualitative et quantitative) du réseau convolutionnel avec un modèle basé par une approche par patches [6] publié la même année.

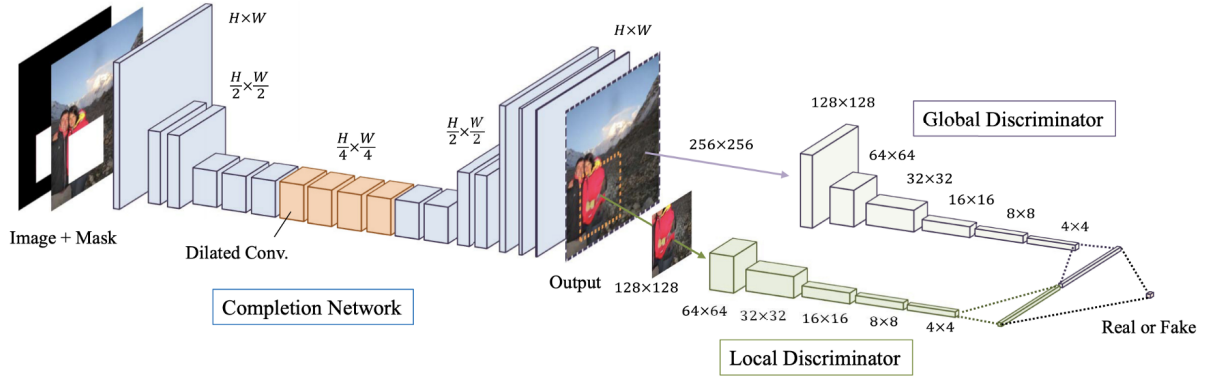


FIGURE 2 – Modèle complet d'Iizuka [4]

## 2 Inpainting par réseaux profonds : Modèle d'Iizuka et al.

### 2.1 Architecture

Iizuka et al. proposent une approche basée sur des modèles génératifs pour implémenter le processus d'inpainting. Le modèle se structure en deux parties. Un premier réseau, qu'il appelle « réseau de complétion » infère la zone manquante sur l'image. Il s'agit d'un réseau type Auto-Encoder (structure en accordéon) de **6 millions de paramètres**, constitué de multiples couches convolutionnelles qui réduisent la taille de l'image à mesure que l'on va en profondeur dans le réseau et augmente le nombre de canaux. Des couches de déconvolution sont ensuite appliquées pour restituer une image complétée avec la résolution initiale.

Il prend en entrée une image RGB de dimension  $H \times W \times 3$  ainsi qu'un masque de taille  $H \times W$  concaténé à l'image d'entrée. Le masque est binaire, avec des 1 aux indexes des zones à compléter et des 0 ailleurs. On concatène le masque à l'image d'entrée pour spécifier au réseau les zones qu'il doit compléter.

### 2.2 Entraînement du modèle

La plus-value du modèle d'Iizuka repose sur l'architecture du discriminateur pour entraîner le réseau de complétion. Celui-ci est subdivisé en deux sous réseaux :

- Un réseau discriminant la consistance globale de l'image. Ce réseau prend en entrée l'image entière complétée par le premier modèle.
- Un réseau discriminant la consistance locale de l'image. Au lieu de prendre l'image entière en entrée, il ne prend en considération qu'une partie de l'image centrée sur la zone complétée, avec une certaine marge sur les bords. Cette marge permet de restituer une certaine continuité des valeurs des pixels et mieux intégrer la zone complétée.

Les deux discriminateurs ont une structure de réseau convolutionnel classique et leur deux sorties (vecteurs de taille 1024) sont concaténés en un seul vecteur.

L'entraînement se décompose en trois temps :

1. Dans un premier temps, on entraîne uniquement le réseau de complétion. On mesure la qualité de reconstruction de l'image en minimisant une fonction de perte  $L2$  sur la zone à compléter pendant l'entraînement.

$$loss(I_{GT}, I_{recons}, M_c) = ||I_{GT}|_{M_c}, I_{recons}|_{M_c}||_2^2$$

Avec  $M_c$  correspondant au masque appliqué à l'image et  $I_{GT}|_{M_c}$  correspondant aux images restreintes aux zones à compléter.

2. Dans un deuxième temps on entraîne uniquement le réseau discriminateur à distinguer les images réelles de celles complétées par le réseau de complétion pré-entraîné en phase 1.
3. Pour finir, les deux réseaux sont entraînés conjointement.

### 3 Expérimentations sur le modèle

Nous présentons dans cette section les différentes études d’ablation que nous avons menées sur le réseau convolutionnel d’Iizuka. Nous présentons les données et métriques que nous avons considérées puis les différentes échelles d’ablations de réseau réalisées.

#### 3.1 Données et Critères d’Evaluation

##### 3.1.1 Choix du dataset

Pour toutes nos expériences (ré-entraînement, ablations, études comparatives), nous avons choisi de travailler sur la base de données CelebA [7] : elle contient plusieurs milliers de portraits de célébrités. Nous avons choisi ces données pour leur facilité d’accès et aussi car des préentraînements ont déjà été effectués avec CelebA. Comparativement à l’autre dataset utilisé dans l’article, Places2, la taille des images de CelebA est plus petite : 218 x 178 x 3 (contre 256 x 256 x 3 pour les données Places2) : la réduction de la taille des images en entrée du réseau permet un gain de temps non négligeable dans le processus d’entraînement.

Type de données	Nombre d’images	Taille d’image
Train	162 079	218 x 178 x 3
Test	40 520	218 x 178 x 3

TABLE 1 – Données utilisées pour nos expérimentations sur le réseau : CelebA dataset

L’utilisation de portraits nous permet aussi de faire différents types d’expériences d’inpainting :

- Reconstruction de régions structurales (par exemple, lorsque le masque cache des régions telles que le nez ou la bouche).
- Reconstruction de régions texturées (par exemple, pour des occlusions partielles des cheveux).

##### 3.1.2 Métriques d’Evaluation Utilisées

L’article d’Iizuka présente une liste exhaustive d’illustrations des sorties de leur réseau de complétion : elles attestent de la qualité des performances du réseau. Néanmoins pour comparer les performances du réseau de complétion avec d’autres, l’évaluation purement visuelle des images complétées est un peu restrictive.

En plus de la qualité visuelle des images produites, nous avons implémenté différentes métriques [8] pour réaliser des comparaisons quantitatives :

- **Mean Squared Error (MSE)** mesure l’erreur quadratique moyenne pixel par pixel entre l’image originale et celle reconstruite :

$$MSE(I_{orig}, I_{recons}) = \frac{1}{H} \frac{1}{W} \sum_{i=1}^H \sum_{j=1}^W (I_{orig}(i, j) - I_{recons}(i, j))^2$$

Nous avons aussi utilisé d’autres métriques : en effet deux images présentant des distorsions différentes peuvent avoir la même valeur de MSE. Ici la valeur du MSE est utilisée à titre indicatif.

- **Peak Signal to Noise Ratio (PSNR)** mesure la distorsion de l’image générée par rapport à celle initiale.

$$PSNR(I_{orig}, I_{recons}) = 10 \log_{10} \left( \frac{255^2}{MSE(I_{orig}, I_{recons})} \right)$$

L’image  $I_{recons}$  est donc supposée d’autant meilleure que  $PSNR(I_{orig}, I_{recons})$  est grande, puisque dans ce cas  $MSE(I_{orig}, I_{recons})$  est de valeur plus faible.

- **Structural Similarity Index Measure (SSIM)** mesure la similarité entre les structures des deux images et prend en compte la luminance  $l$ , le contraste  $c$  et la structure  $s$ .

$$SSIM(I_{orig}, I_{recons}) = l(I_{orig}, I_{recons}) \dot{c}(I_{orig}, I_{recons}) \dot{s}(I_{orig}, I_{recons})$$

Les valeurs du  $SSIM$  sont comprises entre 0 et 1 : la corrélation entre deux images est maximale lorsque  $SSIM(I_{orig}, I_{recons}) = 1$ .

- **Fréchet Inception Distance (FID)** est une autre mesure de similarité et de consistance entre les images reconstruites et les images initiales, généralement utilisée pour évaluer des modèles génératifs. Nous avons recodé entièrement cette métrique. Pour ce faire :
  1. Nous appliquons des masques aléatoirement sur l'ensemble des images d'origine, puis nous calculons les images reconstruites avec le modèle de complétion.
  2. L'ensemble des images reconstruites est mélangé pour éviter tout biais dans le score final. On fait de même avec les images d'entrée.
  3. On calcule la sortie de l'avant dernière couche d'un ResNet50 pré-entraîné de chaque image. Cela nous permet d'obtenir un encodage pour chaque image, à savoir un vecteur de taille (1,2048).
  4. A partir de ces vecteurs, pour chaque banque de données on calcule les paramètres de la distribution gaussienne associée  $(\mu, \Sigma^2)$ .

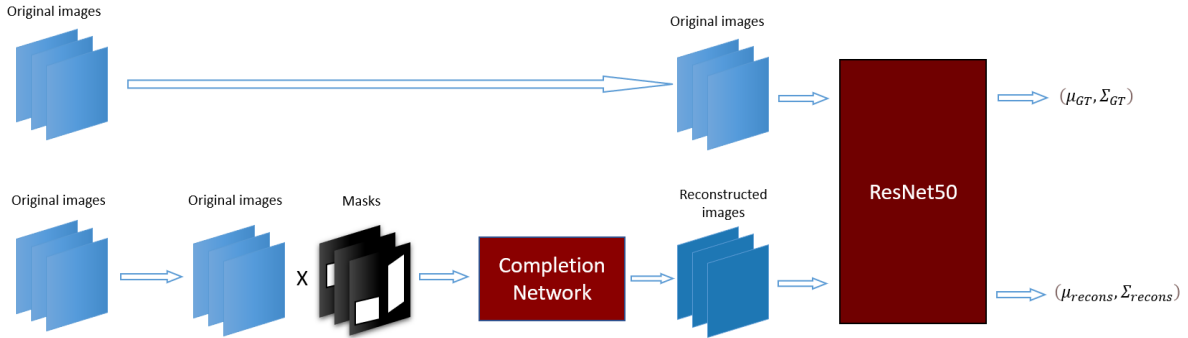


FIGURE 3 – Calcul du score FID lié aux images reconstruites par le réseau

5. On obtient finalement le score FID en calculant la distance de Wassertein entre les distributions  $\mathcal{N}(\mu_{GT}, \Sigma_{GT})$  et  $\mathcal{N}(\mu_{recons}, \Sigma_{recons})$  :

$$FID(data_{GT}, data_{recons}) = \|\mu_{GT} - \mu_{recons}\|_2^2 + Tr(\Sigma_{GT} + \Sigma_{recons} - 2(\Sigma_{GT}^{\frac{1}{2}} \Sigma_{recons} \Sigma_{GT}^{\frac{1}{2}})^{\frac{1}{2}})$$

Notons ici que la valeur du score FID seule n'est pas réellement interprétable. En revanche, en comparant les scores FID pour différents modèles de reconstruction : plus le score FID faible, meilleur le modèle est.

Nous utilisons par la suite ces quatre métriques pour obtenir une évaluation quantitative consistante de nos expériences.

## 3.2 Etude d'ablation du réseau

### 3.2.1 Principe

Dans un premier temps, nous avons voulu évaluer l'importance des composantes du discriminateur dans le modèle. Comme explicité dans la Section 2.2, le discriminateur est composé de deux réseaux convolutionnels : un premier qui donne une consistance globale à la zone à compléter et un second dédié à la consistance locale.

En reprenant le modèle de complétion pré-entraîné en Phase 1, nous avons réentraîné durant les phase 2 (uniquement le discriminateur) et phase 3 (générateur et discriminateur combinés) durant respectivement 10000 et 400000 epochs, afin d'avoir des modèles comparables avec celui initial. Pour chaque phase d'entraînement, les images d'entrée sont coupées aléatoirement à une taille de 160x160. On applique des masques sur les images de manière aléatoire.

Nous nous limitons à un seul masque par image, avec pour dimensions  $(H, W) \in [48, 96]^2$ .

En résumé nous avons réentraîné le réseau :

- Sans discriminateur local.
- Sans discriminateur global.

### 3.2.2 Résultats visuels (qualitatifs)

Nous nous avons cherché à analyser les images des différents modèles de manière qualitative. Afin de pouvoir comparer les modèles entre eux, pour une image donnée le même masque est utilisée pour les prédictions. On distingue différents cas (Table 2) :

- Pour certaines images, il n’y a pas de différences notables entre les différents modèles (1ère ligne).
- De manière générale, on remarque que les zones reconstruites par les modèles entraînés uniquement avec un discriminateur global sont plus floues (2ème colonne).
- On observe que le modèle avec discriminateur local et global est plus performant pour reconstruire des régions structurales du visage. En effet lorsque le nez est masqué (3ème ligne), le nez reconstruit est plus cohérent que celui issu du modèle se basant uniquement sur un discriminateur local. Pour l’image où une oreille est cachée (2ème ligne), on constate qu’une portion de l’oreille est plus nette, tandis que pour les images reconstruites par les modèles avec un discriminateur uniquement local ou global l’oreille est totalement floue.
- Il semble aussi que le modèle proposé par Izuka et al. reconstruit des zones cohérentes pour les zones texturées, comme les cheveux : les régions des images où le masque recouvre des cheveux (4ème et 5ème lignes) sont plus harmonieuses. En effet, même si la texture des cheveux est reproduite par le modèle avec discriminateur local, les cheveux reconstruits par le modèle complet sont plus nets et plus cohérents avec les cheveux non masqués.
- Enfin on remarque que le modèle entraîné avec discriminateur local et global réussit à mieux reconstruire les régions en bordure de la tête lorsque l’éclairage est en contre-jour (dernière ligne) : on voit que la luminosité au niveau du menton reconstruit est en accord avec celle de la bordure du visage.

### 3.2.3 Résultats quantitatifs

Pour chaque modèle, nous avons calculé le score FID sur l’ensemble des 40 520 images test de CelebA. : les scores sont rassemblés dans la table 3. Nous répétons l’évaluation 2 à 3 fois pour chaque modèle pour éviter des biais possibles dans les valeurs, étant donné que les masques sont générés de manière aléatoire sur chaque image.

Modèle	MSE	PSNR	SSIM	FID Score
<b>Global uniq.</b>	0.020	17.221	0.683	$100.75^{\pm 0.2}$
<b>Local uniq.</b>	0.053	12.942	0.595	$69.57^{\pm 0.08}$
<b>Local et Global</b>	<b>0.015</b>	<b>18.447</b>	<b>0.708</b>	<b><math>37.51^{\pm 0.02}</math></b>

TABLE 3 – Scores comparatifs après une étude d’ablation du discriminateur. Le score FID est nettement meilleur lorsque les deux parties du discriminateur sont utilisées dans l’entraînement du réseau de complétion

Les FID scores que l’on obtient sont cohérents avec l’analyse de l’article [4] et celle que nous avons réalisée plus haut. La valeur du FID est plus élevée lorsqu’on retire une partie du discriminateur dans l’entraînement : le réseau reconstruit des images moins naturelles. Il est aussi intéressant de noter que le FID est beaucoup plus élevé lorsque le réseau est entraîné avec un discriminateur purement global comparativement au modèle purement local. L’écart des scores FID est identique entre les 3 modèles. Le discriminateur local joue donc un rôle prépondérant sur le discriminateur global dans la consistance de l’image complétée.

Nous observons des résultats différents pour les autres métriques d’évaluation. Les scores MSE, SSIM, PSNR sont nettement meilleurs lorsque les deux réseaux discriminants sont utilisés en entraînement. Toutefois, l’indice de similarité de structure est plus élevé pour le modèle purement global que pour le modèle uniquement local, alors

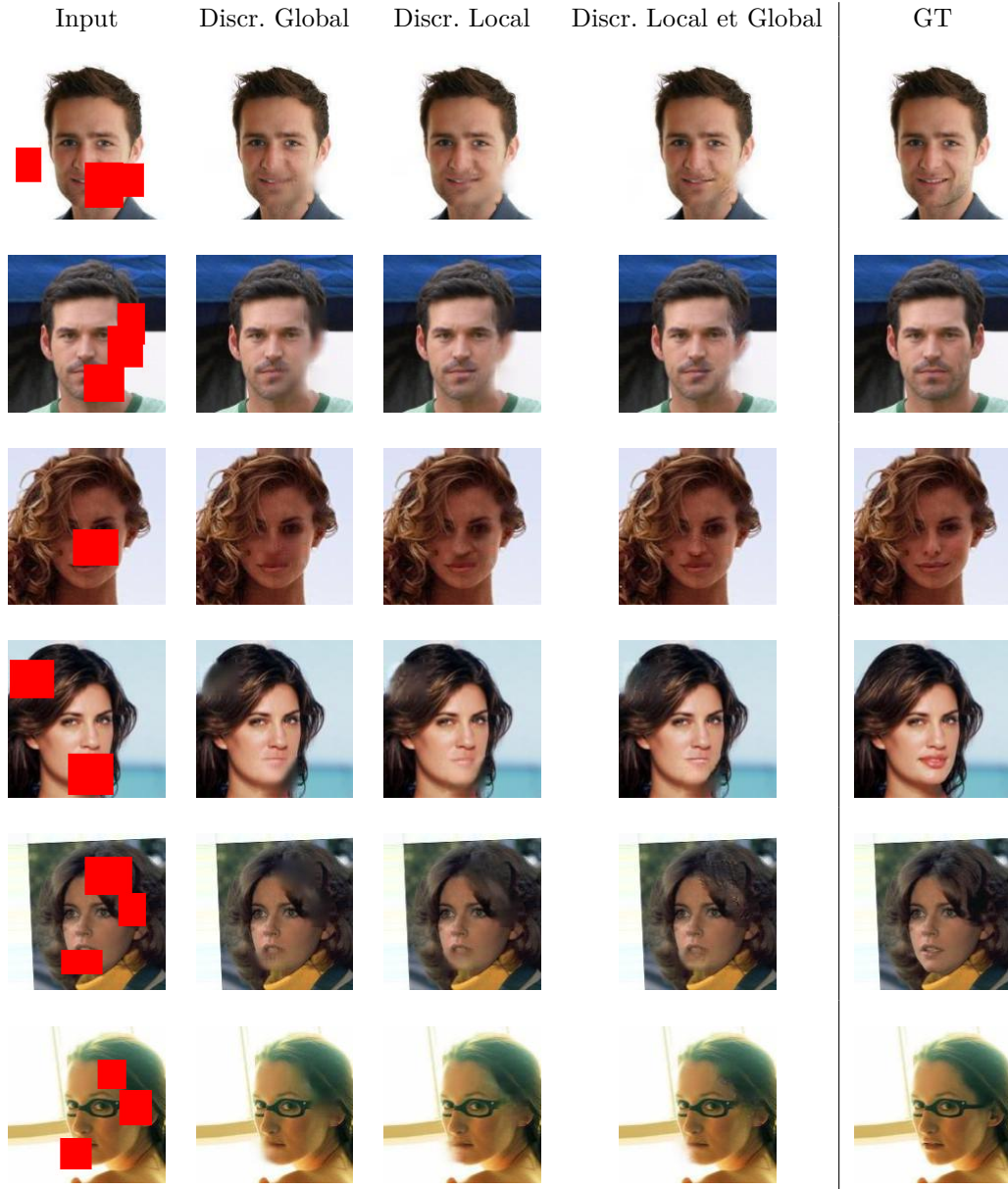


TABLE 2 – Résultats comparatifs d’inpainting

que les images nous apparaissent plus floues. De même, le score MSE est plus faible que pour le modèle purement global.

### 3.3 Quantification de l’importance des couches du réseau dans la restauration des images

Une seconde idée pour évaluer l’importance des paramètres du réseau consiste à étudier les informations apprises dans chaque couche du réseau une fois celui-ci entraîné. Pour une couche donnée du réseau, nous avons retiré un canal de façon aléatoire puis recalculé le score FID du réseau sans celui-ci.

Le réseau de complétion comportant 17 couches convolutionnelles, nous avons étudié uniquement les canaux situés sur les couches 1, 2, 3, 4, 10, 14, 15, 16. Nous avons volontairement pris les premières et dernières couches du réseau. Nous avons pensé qu’elles pourraient être les plus informatives. Nous avons aussi considéré une couche aléatoire en milieu de réseau (ici la dixième couche) pour visualiser aussi les performances des couches intermédiaires.



Nous avons réalisé l'inférence du réseau de multiples fois, à chaque fois en effaçant un canal aléatoirement dans une des 8 couches données.



TABLE 4 – Scores FID obtenus après en fonction du retrait des sorties d'un canal d'une couche de convolution (1, 2, 3, 4, 10, 14, 15 et 16). La ligne horizontale correspond au score FID du modèle de base (de valeur 37.51). Certains retraits canaux, en particulier dans la première couche, influencent fortement le score final.

Nous avons remarqué que le retrait de certains canaux particuliers permettait de diminuer le score FID du modèle de façon significative : par exemple, le retrait du canal 0 de la première couche fait chuter le score de 37.51 à 32.52. A l'inverse, le retrait de d'autres canaux perturbent le modèle. Cet effet est particulièrement visible sur les 3 premières couches notamment, où le score FID est plus élevé pour une dizaine de canaux.

Ces différences sont légèrement visibles lorsque nous visualisons certaines images (Table 5).










Modèle	Image d'entrée	Sortie	Image réelle
Sans changement			
Canal 44 de Conv1			
Canal 0, 102, 44			

TABLE 5 – Exemple d'images où certains canaux ont été retirés. En troisième ligne, le retrait de certains canaux bien choisis permet de mieux rendre la texture des cheveux

Pour cet exemple, la zone complétée en haut du front semble meilleure lorsqu'on a retiré les trois canaux bien choisis : la texture des cheveux est mieux restaurée que celle rendue par le modèle original. Par ailleurs, la suppression du canal 44 de la première couche affecte la sortie du réseau : les trois zones complétées sont plus floues et les détails moins bien rendus. Ce canal semble donc garder en mémoire des informations de l'image essentielles pour sa complétion.

Quantitativement, on observe bien l'impact du retrait du canal 44 de la première couche dans le tableau 6, en particulier pour la valeur du MSE.

Modèle	MSE	PSNR	SSIM
Sans changement	<b>0.00191</b>	27.181	<b>0.9249</b>
Retrait canal 44 Conv1	0.00200	26.988	0.9174
Retrait canal 0, 102 et 44 de Conv 1,2,14	0.00189	<b>27.217</b>	0.9244

TABLE 6 – Comparaison des scores de reconstruction pour l’image de la figure 5

### 3.4 Autres Expériences réalisées

Dans le but d’étudier l’importance des différentes couches du réseau de complétion, nous avons aussi essayé de remplacer de manière aléatoire certains filtres par le filtre identité. Les images obtenues étaient difficilement interprétables et nous avons donc abandonné cette piste.

## 4 Comparaison avec une autre méthode d’inpainting par patches

### 4.1 Principe

Nous avons cherché à comparer les performances du modèle d’Iizuka et al. avec une méthode par optimisation globale de patches [6], qui repose sur la résolution du problème d’optimisation suivant.

$$\min_{\phi} \sum_{p \in \mathcal{H}} d^2(W_p, W_{p+\phi(p)}) \quad (1)$$

où  $\mathcal{H}$  est le masque appliqué à l’image,  $W_p$  est un patch autour du pixel  $p$ ,  $\phi$  est un champs de vecteur indiquant la position du *nearest neighbor patch* NN et  $d(.,.)$  définit une distance entre patches. La spécificité de la méthode présentée par Newson et al. est que les textures sont prises en compte dans le calcul de la distance entre patches  $d$ , mais aussi dans l’étape de reconstruction de l’image.



FIGURE 4 – Exemple où la texture est bien recréée lors de l’inpainting [6]. A gauche, l’image d’entrée avec en vert le contour de la zone masquée et à droite l’image complétée. On constate que la zone complétée n’est pas homogène et que des plis ont été créés.

### 4.2 Analyse comparative avec le modèle d’Iizuka et al.

Nous avons voulu comparer les performances de ces deux approches (inférence par patch avec des patches de dimensions 7 x 7 et par un réseau). Nous avons pris aléatoirement 280 images de l’ensemble des images test de CelebA pour lesquelles nous avons générés des masques aléatoires. Pour la génération des masques binaires, nous nous sommes restreints à un nombre de rectangles allant de 1 à 3, et des dimensions de telle sorte que le masquage n’occupe pas plus de 25% de l’image originelle. Afin de pouvoir comparer les modèles entre eux, pour chaque image donnée nous avons gardé le même masque. Nous rassemblons quelques exemples visuels dans la table 7.

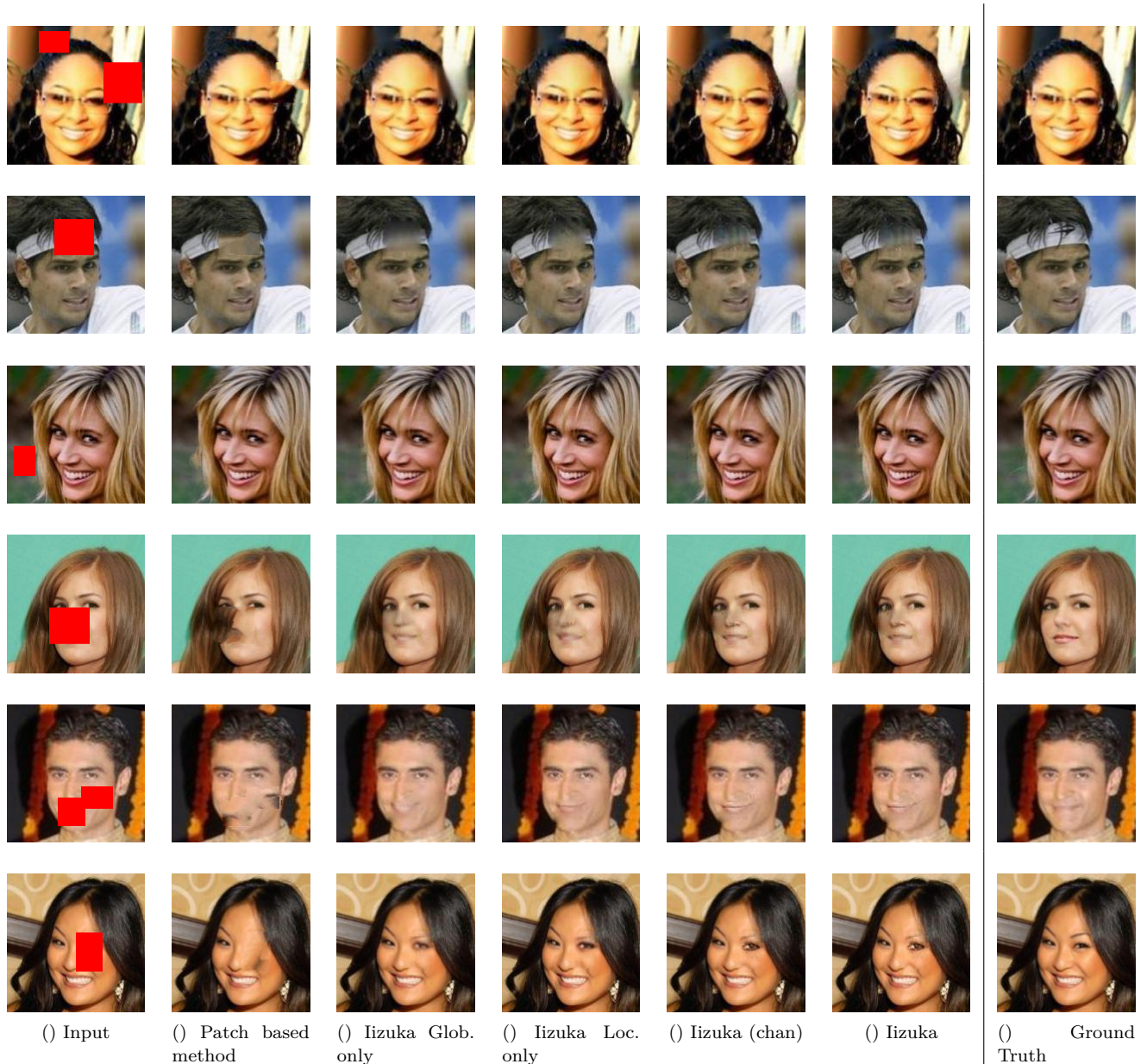


TABLE 7 – Visualisation de la complétion des différents modèles. Nous avons rassemblés tous les différents modèles que nous avons pu tester (sauf le retrait du channel 44 de la première convolution qui renvoie systématiquement des résultats brouillés (voir section précédente). Les prédictions du modèle de propagation de patches manquent de consistance et ne sont pas plausibles. On perd souvent la continuité des traits de visages.

On remarque que la méthode par patches de Newson et al. performe bien lorsque que le masque recouvre l’arrière-plan. Incluant une notion de texture pour la distance entre des patches ainsi que pour l’inpainting en lui même, cette méthode complète de manière satisfaisante les régions texturées telles que les cheveux (table 8). Toutefois, la méthode par patches ne peut reconstruire des zones structurales totalement obstruées telles qu’un nez ou une bouche (table 7). Dans le cas où un seul oeil serait caché par le masque, cette méthode ne peut utiliser le deuxième oeil car l’algorithme de recherche du NN approximatif d’un patch explore uniquement son voisinage. Pour ces cas particuliers, on constate que les zones reconstruites ne sont pas uniformes. Ayant des images de dimension 160 x 160, nous aurions sûrement dû utiliser des patches de dimensions inférieure à 7 x 7 conformément à l’article [6], option malheureusement non disponible sur la démonstration IPOL.

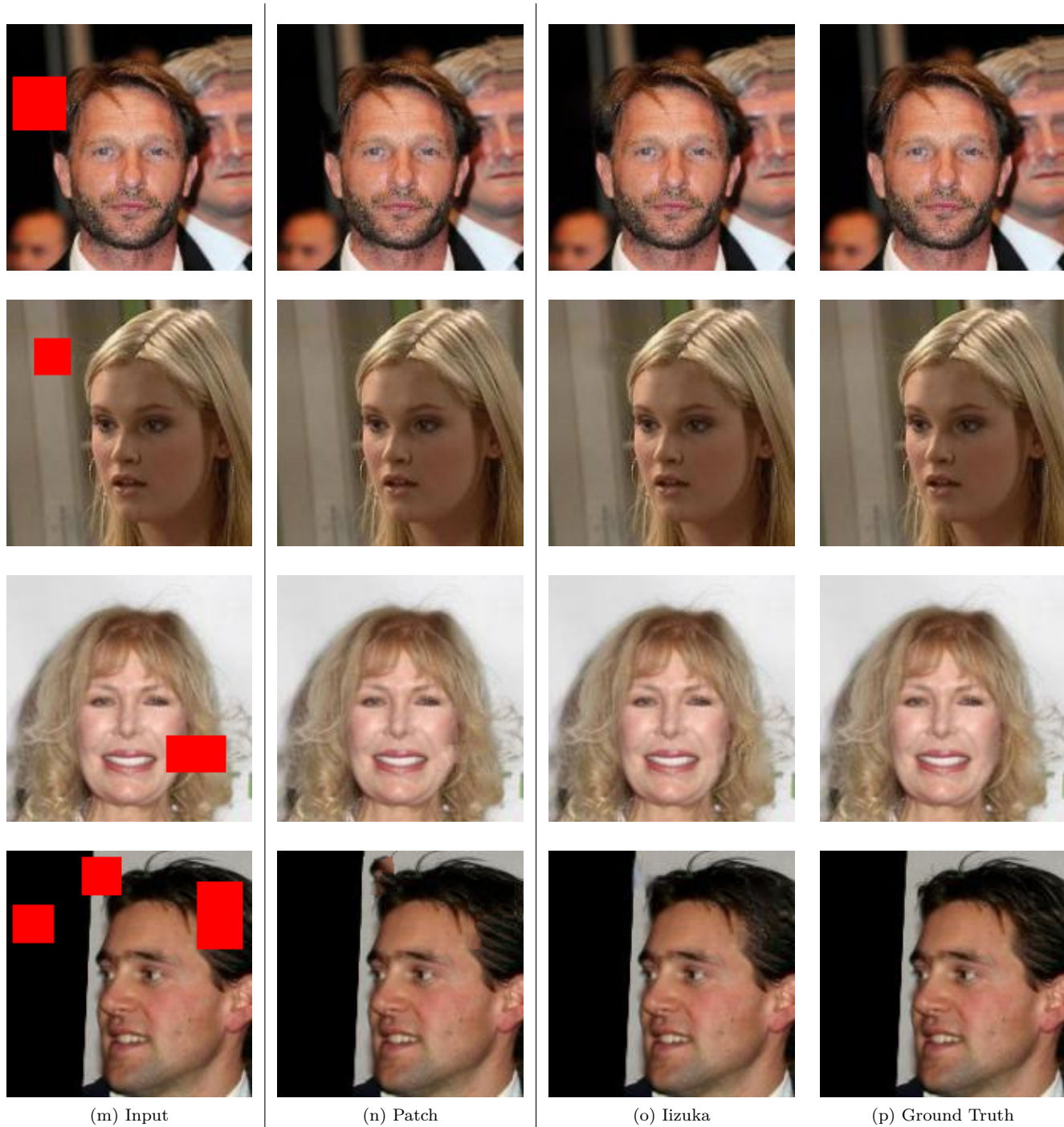


TABLE 8 – Comparaison de la complétion du modèle d’Iizuka et al. (CNN) et celui de Newson et al. (patch based) pour des images où le masque recouvre une partie de l’arrière plan et/ou les cheveux. Pour ces images, la méthode par patches reconstruit des images cohérentes

### 4.3 Comparaison quantitative avec le modèle d’Iizuka et al.

Modèle	MSE↓	PSNR↑	SSIM↑	FID↓
Iizuka (Global)	0.0011	31.938	0.972	8.83
Iizuka (Local)	0.0011	31.907	0.971	8.643
Iizuka (retrait Canal 44)	0.0016	29.706	0.964	10.891
Par Patch [6]	0.0036	26.373	0.943	33.296
Iizuka	<b>0.0010</b>	<b>31.983</b>	<b>0.973</b>	<b>7.743</b>

TABLE 9 – Étude comparative sur un batch de 280 images de celebA. Quantitativement, Iizuka obtient des meilleurs scores : son modèle synthétise et infère des nouvelles zones non vues dans l’image à la différence de la méthode par patches.

Quantitativement, on observe bien que les scores du modèle d’Iizuka et al. sont meilleurs pour la reconstruction de photographies de visages. Le score FID du modèle par réseau convolutionnel est nettement supérieur (Table 9). En effet les images rendues par le modèle de réseau sont beaucoup plus réalistes et vraisemblables, d’où cette différence nette. La méthode par patches ne pouvant pas créer de nouvelles images à partir de zones non vues, il échoue presque tout le temps dans l’inférence d’un membre du visage (oeil, nez, bouche, etc.).

## 5 Conclusion et perspectives

Notre étude du modèle d’Iizuka et al. confirme l’importance des discriminateurs global et local pour le modèle de complétion proposé. Utiliser uniquement un discriminateur global lors de l’entraînement du modèle conduit à des zones reconstruites floues, tandis qu’utiliser uniquement un discriminateur local permet d’avoir des résultats localement satisfaisant en terme de textures mais qui manque de cohérence à une échelle plus globale. Ces résultats sont en accord avec les différentes métriques d’évaluation que nous avons calculées. En effet, le MSE est plus faible lorsqu’on inclut la totalité du réseau discriminateur et le score FID est beaucoup plus élevé pour ce modèle.

Nous nous sommes aussi intéressées à l’importance des couches du réseau de complétion. L’étude d’ablation sur les canaux réalisée montre en particulier l’importance de la première couche de convolution du modèle. Le retrait de certains canaux bien choisis permettent dans certains cas d’obtenir de meilleurs résultats mais quantitativement il est difficile de conclure.

Nous avons enfin comparé ce modèle avec une méthode par patches avec notion de textures. Du fait de cette spécialisation sur les textures, la méthode par patches fonctionne bien pour l’inpainting de zones texturées. Toutefois comme cette méthode repose sur la minimisation d’une distance entre un patch et un NN-patch, cet algorithme ne peut inférer une zone qui n’existe dans l’image/inférer à partir d’une région loin du masque.

Par ailleurs, de nouveaux modèles de deep learning semblent réaliser la tâche d’inpainting de manière plus robuste. On aurait notamment pu étendre notre étude sur le modèle Palette, publié en 2021 [9]. Palette est un réseau inspiré de l’architecture U-Net et comportant des couches de Self-Attention, se différenciant par le fait que les masques pris en entrée sont bruités par un bruit Gaussien.



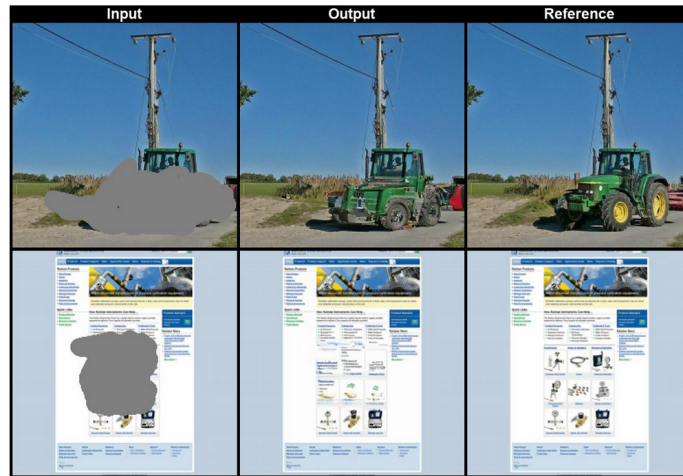


FIGURE 5 – Inpainting par le modèle Palette. Palette semble compléter de façon beaucoup plus consistante les images. Il prédit parfois plus de la moitié des objets à compléter comme sur cette image.

## Références

- [1] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, “Image inpainting,” in *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pp. 417–424, 2000.
- [2] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman, “Patchmatch : A randomized correspondence algorithm for structural image editing,” *ACM Trans. Graph.*, vol. 28, no. 3, p. 24, 2009.
- [3] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, “Context encoders : Feature learning by inpainting,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- [4] S. Iizuka, E. Simo-Serra, and H. Ishikawa, “Globally and Locally Consistent Image Completion,” *ACM Transactions on Graphics (Proc. of SIGGRAPH 2017)*, vol. 36, no. 4, p. 107, 2017.
- [5] Z. Yi, Q. Tang, S. Azizi, D. Jang, and Z. Xu, “Contextual residual aggregation for ultra high-resolution image inpainting,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [6] A. Newson, A. Almansa, Y. Gousseau, and P. Pérez, “Non-Local Patch-Based Image Inpainting,” *Image Processing On Line*, vol. 7, pp. 373–385, 2017. <https://doi.org/10.5201/ipol.2017.189>.
- [7] Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep learning face attributes in the wild,” in *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.
- [8] D. M. K.Silpa, “Comparison of Image Quality Metrics,” *International Journal of Engineering Research Technology (IJERT)*, vol. 1, no. 4, pp. 373–385, June - 2012. <https://www.ijert.org>.
- [9] C. Saharia, W. Chan, H. Chang, C. A. Lee, J. Ho, T. Salimans, D. J. Fleet, and M. Norouzi, “Palette : Image-to-image diffusion models,” *arXiv preprint arXiv :2111.05826*, 2021.

## Acronymes

**FID** Fréchet Inception Distance. 5, 13

**MSE** Mean Squared Error. 4, 6, 7, 9, 13

**PSNR** Peak Signal to Noise Ratio. 4, 6, 13

**SSIM** Structural Similarity Index Measure. 4, 6, 13