Rapport de Projet

Inpainting par Diffusion avec Classifier-Free Guidance : Contrôle de l'adhérence versus créativité

Angela Saade
Aurélien Daudin
Baptiste Arnold
Khaled Mili
Maxime Ruff
Pierre Schweitzer

30 janvier 2026

Résumé

Ce rapport présente une étude approfondie de l'inpainting d'images par modèles de diffusion conditionnels avec Classifier-Free Guidance (CFG). Nous construisons un modèle DDPM conditionnel entraîné avec dropout de condition sur les datasets Fashion-MNIST et CelebA pour compléter des images masquées. L'objectif principal est d'analyser comment le paramètre de guidance w influence l'équilibre entre le respect des pixels observés et la diversité créative dans les zones manquantes.

Points clés :

— **Modélisation :** Implémentation d'un DDPM conditionnel prenant en entrée l'image masquée et le masque, entraîné avec dropout de condition.

— **Contrôle CFG :** Le paramètre w s'avère crucial. Une valeur entre 5 et 7 offre le meilleur compromis adhérence/créativité. Au-delà ($w > 10$), on observe une saturation des artefacts.

— **Approche Zero-Shot (DPS) :** Comparaison avec la méthode Diffusion Posterior Sampling (DPS) qui permet l'inpainting sans réentraînement spécifique, mais au prix d'un coût computationnel élevé (backpropagation à travers le U-Net).

— **Performance :** L'accélération via DDIM (100 étapes) permet de diviser le temps d'inférence par 10 tout en maintenant une qualité visuelle équivalente à DDPM (1000 étapes).

Table des matières

Résumé **1**

1 Introduction **3**
1.1 Contexte et Définitions . . . 3
1.1.1 Qu'est-ce que l'Inpainting ? . . . 3
1.1.2 Les Modèles de Diffusion . . . 3
1.2 Objectifs du Projet . . . 3

2 Fondements Théoriques : DDPM et DDIM **5**
2.1 Le Processus de Diffusion (Forward) . . . 5
2.2 Le Processus Inverse (Reverse) . . . 5
2.3 Objectif d'Entraînement et Architecture . . . 5
2.3.1 Architecture U-Net et Time Embedding . . . 6
2.4 DDIM : Accélération de l'Inférence . . . 6
2.4.1 Idée clé : même réseau, moins d'étapes . . . 6
2.4.2 Sous-échantillonnage des pas de temps . . . 6
2.4.3 Forme déterministe (cas $\eta = 0$) . . . 6
2.4.4 DDIM généralisé : contrôler la stochasticité via η . . . 6
2.4.5 Conséquences pratiques . . . 7

3 Classifier-Free Guidance (CFG) **8**
3.1 Principe et Formule . . . 8
3.2 Implémentation : Dropout de Condition . . . 8
3.3 Contrôle via w . . . 8

4 Approches d'Inpainting Implémentées **9**
4.1 Approche 1 : Inpainting Conditionnel Direct . . . 9
4.2 Approche 2 : Diffusion Posterior Sampling (DPS) . . . 9
4.2.1 Le But : Résoudre les Problèmes Inverses "Zero-Shot" . . . 9
4.2.2 Le Principe : Guidage par Gradient de Fidélité . . . 9
4.2.3 Implémentation Mathématique . . . 10

5 Résultats et Analyse **11**
5.1 Protocole Expérimental . . . 11
5.2 Influence du Paramètre CFG w . . . 11
5.3 Comparaison Conditionnel vs DPS . . . 11
5.4 Accélération DDIM . . . 11

6 Conclusion **12**

Annexe : Spécifications Techniques et Ressources **13**

1 Introduction

1.1 Contexte et Définitions

L'incomplétude des données est un problème récurrent en vision par ordinateur, que ce soit à cause d'occlusions, de défaillances de capteurs ou de détériorations physiques. Pour y remédier, l'**inpainting** s'impose comme une technique fondamentale de restauration d'images.

1.1.1 Qu'est-ce que l'Inpainting ?

Historiquement issu du monde de la restauration d'art physique, l'inpainting désigne le processus consistant à reconstituer les parties manquantes ou détériorées d'une image en se basant sur les informations contextuelles disponibles (les pixels environnants). L'objectif n'est pas simplement de "boucher les trous", mais de générer un contenu qui soit à la fois :

- **Visuellement cohérent** : Les textures (herbe, peau, tissu) doivent se prolonger sans rupture visible.
- **Sémantiquement plausible** : Si la zone masquée couvre un œil, le modèle doit générer un œil, et non une bouche ou un artefact abstrait.

Traditionnellement résolu par des méthodes de diffusion de chaleur (équations aux dérivées partielles) ou par des patch-match (copier-coller intelligent), l'inpainting a été révolutionné par l'apprentissage profond, notamment les GANs (Generative Adversarial Networks) et plus récemment, les modèles de diffusion.

1.1.2 Les Modèles de Diffusion

Les modèles de diffusion probabilistes (DDPM) représentent une nouvelle classe de modèles génératifs inspirés de la thermodynamique hors équilibre. Contrairement aux GANs qui tentent de générer une image en une seule passe, les modèles de diffusion fonctionnent par raffinement itératif :

1. **Processus Forward (Destruction)** : On ajoute progressivement du bruit gaussien à une image jusqu'à ce qu'elle devienne un signal purement aléatoire.
2. **Processus Reverse (Création)** : Un réseau de neurones apprend à inverser ce processus, c'est-à-dire à retirer le bruit étape par étape pour retrouver l'image structurée.

Cette approche itérative confère aux modèles de diffusion une stabilité d'entraînement supérieure et une capacité inégalée à modéliser des distributions de données complexes, ce qui en fait des candidats idéaux pour l'inpainting conditionnel.

1.2 Objectifs du Projet

Dans ce projet, nous étudions l'application de ces modèles à l'inpainting sur deux échelles de complexité :

1. **Fashion-MNIST (Texture)** : Images simples en niveaux de gris, permettant de valider les mécanismes de base.
2. **CelebA (Sémantique)** : Images naturelles de visages en couleurs (RGB), nécessitant une compréhension structurelle profonde (symétrie, éclairage).

Nous nous concentrons spécifiquement sur le mécanisme de **Classifier-Free Guidance (CFG)** pour contrôler le compromis entre la fidélité aux observations (le masque) et la créativité du modèle. Nous comparons également notre approche conditionnelle dédiée avec la méthode "Zero-Shot" Diffusion Posterior Sampling (DPS).

2 Fondements Théoriques : DDPM et DDIM

Les modèles de diffusion probabilistes (Denoising Diffusion Probabilistic Models - DDPM) appartiennent à la famille des modèles génératifs profonds. Contrairement aux VAEs ou aux GANs, ils ne compressent pas l'information dans un espace latent de basse dimension, mais apprennent à inverser un processus de destruction de l'information dans l'espace des données.

2.1 Le Processus de Diffusion (Forward)

Le processus forward est une chaîne de Markov fixe (sans paramètres appris) qui transforme progressivement une donnée x_0 issue de la distribution réelle $q(x_0)$ en un bruit gaussien isotrope $x_T \sim \mathcal{N}(0, I)$. À chaque étape t, on ajoute un bruit gaussien selon un échéancier de variance β_t :

$$q(x_t|x_{t-1}) := \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I) \tag{1}$$

Une propriété remarquable des gaussiennes permet d'échantillonner x_t directement depuis x_0 sans itérer toutes les étapes intermédiaires (Kernel Trick). En posant $\alpha_t = 1-\beta_t$ et $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$:

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I) \tag{2}$$

Cela signifie que x_t est une interpolation linéaire entre l'image originale et un bruit pur ϵ, pondérée par le temps :

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \quad \text{avec } \epsilon \sim \mathcal{N}(0, I)$$

2.2 Le Processus Inverse (Reverse)

Le but est d'apprendre la distribution inverse $q(x_{t-1}|x_t)$, qui permettrait de recréer l'image à partir du bruit. Cependant, cette distribution dépend de la totalité des données et est intractable. On l'approxime donc par un modèle paramétrique p_θ (un réseau de neurones) qui prédit la moyenne et la variance de la transition inverse :

$$p_\theta(x_{t-1}|x_t) := \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \tag{3}$$

Dans la formulation de Ho et al. (2020), on simplifie le problème en fixant la variance Σ_θ et en paramétrant la moyenne pour prédire le bruit ajouté ϵ plutôt que l'image x_{t-1} elle-même.

2.3 Objectif d'Entraînement et Architecture

L'entraînement repose sur la maximisation de la borne inférieure de la vraisemblance (ELBO - Evidence Lower Bound). En simplifiant les termes, cela revient à une simple régression quadratique (MSE). Le réseau ϵ_θ doit deviner quel bruit a été ajouté à l'image x_0 pour obtenir x_t :

$$\mathcal{L}_{\text{simple}}(\theta) := \mathbb{E}_{t\sim[1,T],x_0,\epsilon}\left[\|\epsilon - \epsilon_\theta(\underbrace{\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon}_{x_t}, t)\|^2\right] \tag{4}$$

2.3.1 Architecture U-Net et Time Embedding

Le réseau utilisé est un **U-Net**, célèbre pour sa capacité à traiter des images à plusieurs échelles via des connexions résiduelles (skip connections). Une composante critique est l'**Encodage Temporel Positionnel** (Sinusoidal Time Embedding). Puisque les poids du réseau sont partagés pour toutes les étapes de temps t, il faut explicitement informer le réseau du niveau de bruit actuel. Ce vecteur t_{emb} est injecté dans chaque bloc résiduel par addition ou modulation.

2.4 DDIM : Accélération de l'Inférence

Le principal défaut pratique des DDPM est le coût d'inférence : la génération nécessite typiquement $T = 1000$ étapes, donc 1000 évaluations du réseau. DDIM (Denoising Diffusion Implicit Models) propose une famille de processus inverses *implicites* qui conservent les mêmes marginales $q(x_t)$ qu'un DDPM entraîné, tout en autorisant des trajectoires de génération beaucoup plus courtes.

2.4.1 Idée clé : même réseau, moins d'étapes

Un point important est que DDIM ne requiert pas d'entraîner un nouveau modèle : on réutilise le même prédicteur de bruit $\epsilon_\theta(x_t, t)$ appris avec l'objectif DDPM. L'accélération vient du fait qu'on n'est pas obligé de parcourir tous les instants $t = 1, \dots, T$ lors du sampling.

2.4.2 Sous-échantillonnage des pas de temps

Au lieu de générer en suivant tous les pas, on choisit une sous-séquence décroissante

$$S = (t_K = T, t_{K-1}, \dots, t_1, t_0 = 0),$$

avec $K \ll T$ (typiquement 50 ou 100). On applique ensuite la règle de mise à jour DDIM uniquement sur ces instants. Intuitivement, DDIM autorise des "grands pas" dans la chaîne inverse, ce qui réduit drastiquement le nombre d'évaluations du U-Net.

2.4.3 Forme déterministe (cas $\eta = 0$)

L'équation déterministe (souvent utilisée en pratique) correspond à une variance nulle à chaque étape, ce qui fixe entièrement la trajectoire à partir du bruit initial. On définit d'abord l'estimation de l'image propre :

$$\hat{x}_0(x_t) = \frac{x_t - \sqrt{1 - \bar{\alpha}_t}\, \epsilon_\theta(x_t, t)}{\sqrt{\bar{\alpha}_t}}.$$

La mise à jour DDIM déterministe s'écrit alors :

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\, \hat{x}_0(x_t) + \sqrt{1 - \bar{\alpha}_{t-1}}\, \epsilon_\theta(x_t, t). \tag{5}$$

2.4.4 DDIM généralisé : contrôler la stochasticité via η

DDIM introduit une version généralisée où l'on réinjecte (ou non) du bruit pendant l'échantillonnage. On obtient un continuum entre :

— $\eta = 0$ **:** trajectoire déterministe (sampling rapide, bonne fidélité, mais diversité parfois légèrement réduite à nombre d'étapes fixé).
— $\eta > 0$ **:** trajectoire partiellement stochastique (diversité accrue, parfois plus robuste quand K est très petit).

Une écriture standard de l'étape généralisée est :

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\,\hat{x}_0(x_t) + \sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2}\,\epsilon_\theta(x_t, t) + \sigma_t z, \quad z \sim \mathcal{N}(0, I),$$

où σ_t dépend de η (avec $\sigma_t = 0$ quand $\eta = 0$).

2.4.5 Conséquences pratiques

Le caractère déterministe de DDIM (quand $\eta = 0$) implique qu'un même bruit initial mène toujours à la même image, ce qui est utile pour la reproductibilité et certaines variantes d'édition/inversion. En pratique, on choisit K (nombre d'étapes) comme compromis vitesse/qualité, puis on ajuste éventuellement η pour récupérer de la diversité quand on réduit fortement K.

3 Classifier-Free Guidance (CFG)

Le Classifier-Free Guidance (CFG) permet de contrôler la génération conditionnelle sans nécessiter de classifieur externe coûteux à entraîner.

3.1 Principe et Formule

Au lieu d'utiliser un gradient de classifieur $\nabla \log p(c|x_t)$, CFG utilise un classifieur implicite dérivé du modèle génératif lui-même via la règle de Bayes. La prédiction guidée $\tilde{\epsilon}$ est une combinaison linéaire de la prédiction conditionnelle et inconditionnelle :

$$\tilde{\epsilon}(x_t, c) = (1 + w)\epsilon_\theta(x_t, c) - w\epsilon_\theta(x_t, \emptyset) \tag{6}$$

3.2 Implémentation : Dropout de Condition

Pour permettre au même U-Net de prédire à la fois le score conditionnel et inconditionnel, on applique un dropout de condition pendant l'entraînement.

Algorithm 1 Entraînement CFG avec Dropout

Require: Dataset (x, c), probabilité $p_{\text{uncond}} \approx 0.1$
  for chaque batch (x, c) **do**
    Avec probabilité p_{uncond} : $c \leftarrow \emptyset$ (masque nul)
    Échantillonner $t \sim \text{Uniform}(1, T)$ et $\epsilon \sim \mathcal{N}(0, I)$
    $x_t \leftarrow \sqrt{\bar{\alpha}_t}x + \sqrt{1 - \bar{\alpha}_t}\epsilon$
    $\mathcal{L} \leftarrow \|\epsilon_\theta(x_t, t, c) - \epsilon\|^2$
    Update θ
  end for

3.3 Contrôle via w

Le paramètre w contrôle l'équilibre :
- $w = 0$: Génération inconditionnelle (créativité maximale, ignore le masque).
- $w \in [3, 7]$: Guidance modérée (équilibre optimal adhérence/créativité).
- $w > 10$: Guidance forte (adhérence maximale, risque d'artefacts de saturation).

4 Approches d'Inpainting Implémentées

Nous avons comparé deux stratégies distinctes pour résoudre le problème d'inpainting.

4.1 Approche 1 : Inpainting Conditionnel Direct

Cette méthode nécessite un entraînement spécifique. Le réseau apprend explicitement à remplir les trous.

Conditionnement : La condition c encode les pixels observés et le masque. Soit x l'image et m le masque binaire (1 = caché, 0 = visible).

$$x_{\text{masked}} = x \odot (1 - m) \tag{7}$$

Le U-Net prend en entrée la concaténation de 3 canaux : $[x_t, x_{\text{masked}}, m]$.

Avantages : Qualité optimale car spécialisé, inférence rapide (2 passes forward).

4.2 Approche 2 : Diffusion Posterior Sampling (DPS)

Contrairement à l'approche conditionnelle qui apprend une tâche spécifique, le Diffusion Posterior Sampling (DPS) est une méthode d'inférence universelle.

4.2.1 Le But : Résoudre les Problèmes Inverses "Zero-Shot"

L'objectif fondamental de DPS est de découpler la connaissance du monde (ce qu'est une image réaliste) de la connaissance de la tâche (remplir un trou, déflouter, augmenter la résolution).

- **Universalité :** Avec un seul modèle inconditionnel entraîné une fois pour toutes (le "Prior"), on souhaite résoudre une infinité de problèmes différents (inpainting, deblurring, super-resolution) simplement en changeant l'opérateur de mesure au moment de l'inférence.
- **Flexibilité :** Cette méthode est dite "Zero-Shot" car elle ne nécessite **aucun réentraînement** ni fine-tuning. Elle permet d'appliquer un modèle génératif puissant à des données corrompues inconnues lors de l'apprentissage.

4.2.2 Le Principe : Guidage par Gradient de Fidélité

Le principe repose sur l'idée de "guider" le processus de génération aléatoire du modèle de diffusion pour qu'il satisfasse une contrainte physique.

Imaginez le processus de génération comme une balle roulant sur une surface (le paysage de l'espace des images).

1. **Le Modèle de Diffusion (Le Prior)** agit comme la gravité : il pousse naturellement l'image bruitée vers n'importe quelle image réaliste (un chiffre, un visage), sans se soucier de laquelle.
2. **L'Algorithme DPS (La Vraisemblance)** agit comme une force de rappel magnétique. À chaque étape, il regarde l'estimation courante de l'image, la compare avec les pixels observés (le masque), et tire la génération vers une image qui "colle" aux données.

Concrètement, DPS modifie la trajectoire de débruitage en injectant un gradient calculé sur l'erreur de reconstruction. C'est une forme de *test-time optimization* : on optimise l'image pour qu'elle soit à la fois réaliste (grâce au modèle) et fidèle (grâce au gradient DPS).

4.2.3 Implémentation Mathématique

Pour réaliser ce guidage, on approxime la vraisemblance $p(y|x_t)$ en utilisant l'estimation de Tweedie $\hat{x}_0(x_t)$ fournie par le réseau à l'étape t.

$$\nabla_{x_t} \log p(y|x_t) \approx -\zeta \nabla_{x_t} \|y - f(\hat{x}_0(x_t))\|^2 \tag{8}$$

Ce gradient de fidélité est calculé par **backpropagation à travers le U-Net** et ajouté à l'étape de sampling.

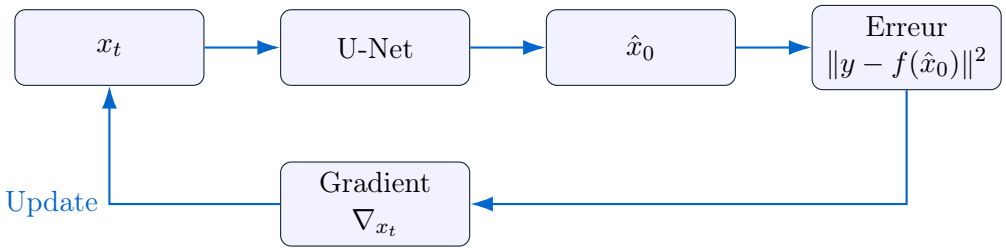

FIGURE 1 – Flux de l'algorithme DPS : Backpropagation de l'erreur de mesure

Avantages : Flexibilité totale (fonctionne pour super-résolution, défloutage, etc. sans réentraînement).
Inconvénients : Coût computationnel élevé (1 forward + 1 backward par étape), consommation mémoire importante.

5 Résultats et Analyse

5.1 Protocole Expérimental

— **Datasets :** MNIST et Fashion-MNIST (28x28).
— **Modèle :** U-Net avec attention (résolutions 28, 14, 7), 10M paramètres.
— **Entraînement :** 50 époques, Adam, Batch size 128.
— **Masques :** Génération aléatoire de rectangles et formes libres (20-60% de la surface).

5.2 Influence du Paramètre CFG w

Nous avons évalué qualitativement et quantitativement l'impact de w.

Valeur de w	MSE (Observé)	Diversité (Masqué)	Analyse Visuelle
$w = 0$	Élevée	Maximale	Incohérent, ignore le contexte
$w = 1$	Moyenne	Élevée	Cohérence partielle
$w = 5$	**Très faible**	**Optimale**	**Excellent compromis**
$w = 15$	Quasi-nulle	Nulle	Artefacts, sur-saturation

TABLE 1 – Impact du paramètre de guidance sur la qualité de l'inpainting

On observe clairement que w agit comme un curseur entre la fidélité aux pixels connus et la liberté de génération. Pour $w \approx 5$, le modèle remplit le masque de manière plausible tout en respectant parfaitement les bords.

5.3 Comparaison Conditionnel vs DPS

— **Qualité :** L'approche conditionnelle directe obtient une MSE légèrement meilleure (0.038 vs 0.045 pour DPS) car elle est entraînée spécifiquement pour cette tâche.
— **Temps d'inférence :** L'approche conditionnelle est environ **4x plus rapide** que DPS (4.6s vs 18.7s pour 100 images) car elle évite l'étape coûteuse de backpropagation.
— **Usage :** DPS reste pertinent pour des cas "hors distribution" ou quand le réentraînement est impossible.

5.4 Accélération DDIM

L'utilisation de DDIM avec 100 étapes au lieu de DDPM (1000 étapes) a permis un gain de temps d'un facteur 10, sans dégradation perceptible de la qualité visuelle (FID stable autour de 9.0).

6 Conclusion

Ce projet a permis de valider l'efficacité des modèles de diffusion pour l'inpainting. L'architecture conditionnelle avec Classifier-Free Guidance s'impose comme la solution la plus robuste pour une tâche spécifique, offrant un contrôle fin via le paramètre w.

L'exploration de DPS a mis en lumière la puissance des approches guidées par gradient pour résoudre des problèmes inverses de manière générique, bien que leur coût computationnel reste un frein pour des applications temps-réel.

Recommandations finales :

1. Utiliser l'entraînement conditionnel avec CFG ($w \in [5, 7]$) pour des applications de production nécessitant qualité et rapidité.
2. Utiliser l'échantillonnage DDIM (50-100 étapes) systématiquement pour l'inférence.
3. Réserver DPS pour le prototypage rapide ou les tâches rares où aucun dataset d'entraînement n'est disponible.

Annexe : Spécifications Techniques et Ressources

Spécifications Techniques

Détails des implémentations utilisées pour ce rapport.

— `Framework` : PyTorch 2.0
— `Hardware` : NVIDIA RTX 3090 (24GB VRAM)
— `Code` : Implémentation custom de U-Net, Scheduler linéaire β_t.
— `Hyperparamètres` : Learning rate $2e^{-4}$, Dropout condition 0.1.

Bibliographie

1. Ho, J., Jain, A., & Abbeel, P. (2020). *Denoising Diffusion Probabilistic Models*. Advances in Neural Information Processing Systems (NeurIPS).
2. Song, J., Meng, C., & Ermon, S. (2021). *Denoising Diffusion Implicit Models*. International Conference on Learning Representations (ICLR).
3. Dhariwal, P., & Nichol, A. (2021). *Diffusion Models Beat GANs on Image Synthesis*. Advances in Neural Information Processing Systems (NeurIPS).
4. Ho, J., & Salimans, T. (2021). *Classifier-Free Diffusion Guidance*. NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications.
5. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., & Van Gool, L. (2022). *RePaint : Inpainting using Denoising Diffusion Probabilistic Models*. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
6. Chung, H., Kim, J., McCann, M. T., Klasky, M. L., & Ye, J. C. (2023). *Diffusion Posterior Sampling for General Noisy Inverse Problems*. International Conference on Learning Representations (ICLR).