Edition Multi-Pivots : Vers une Edition Multi-Directionnelle et Démêlée

Neil Farmer, Catherine Soladié, Gabriel Cazorla, Renaud Séguier

Edition Multi-Pivots : Vers une Edition Multi-Directionnelle et Démêlée

Neil Farmer[1,2] Catherine Soladié[2] Gabriel Cazorla[1] Renaud Séguier[2]

[1] Chanel Parfums Beauté, Innovation Research and Development, Pantin, France
[2] CentraleSupélec, IETR UMR CNRS 6164, France

neil.farmer@chanel.com

Résumé

Les réseaux antagonistes génératifs (GAN) permettent d'éditer des images en manipulant leurs caractéristiques. Cependant, ces manipulations ne sont pas toujours démêlées. Par exemple, lorsqu'une ride spécifique est modifiée, d'autres caractéristiques liées à l'âge sont souvent modifiées également. Cet article propose une nouvelle méthode d'édition démêlée. L'approche présentée est basée sur des images pivots qui permettent d'apprendre des directions d'édition pour une image d'entrée. Ces pivots sont basés sur une image réelle (l'entrée) et des modifications synthétiques de l'image réelle. Bien que notre principal cas d'applications d'édition soit les rides, notre méthode peut être étendue à d'autres tâches d'édition, telles que l'édition de la couleur des cheveux ou du rouge à lèvres. Les résultats qualitatifs et quantitatifs montrent que notre Edition Multi-Pivots (EMP) fournit un niveau plus élevé de démêlage et une édition plus réaliste que les méthodes de l'état de l'art. Le code est disponible sur GitHub.

Mots Clef

Edition Démêlée, GAN, Ride

Abstract

Generative Adversarial Networks (GANs) enable image editing by manipulating image features. However, these manipulations still lack disentanglement. For example, when a specific wrinkle is edited, other age-related features or facial expressions are often changed as well. This paper proposes a new method for disentangled editing. The presented approach is based on two pivot images that allow learning an editing direction for an input image. These pivots are based on a real image (the input) and a synthetic modification of the real image along the desired editing direction. Although our primary focus is on wrinkle editing applications, our method can be extended to other editing tasks, such as hair color or lipstick editing. Qualitative and quantitative results show that our Multi-Pivotal Tuning Editing provides a higher level of disentanglement and a more realistic editing than state-of-the-art methods.

Keywords

Disentangled Editing, GAN, Wrinkle

1 Introduction

Les GANs [1] ont émergé au cours de la dernière décennie pour générer des images à partir de variables latentes [2]. Les capacités génératives des GANs ont été appliquées à l'édition d'images [3]. Ces techniques d'édition d'images nécessitent d'abord l'inversion de l'image souhaitée, c'est-à-dire la projection de l'image dans l'espace latent du GAN. L'inversion et l'édition conduisent à de nombreuses applications, telles que l'édition d'attributs et la restauration d'images. L'objectif de cet article est de traiter la manipulation de rides faciales spécifiques tout en préservant d'autres caractéristiques liées à l'âge. Ce type d'édition peut être utilisé pour prévisualiser les effets des produits anti-âge [4], pour créer des données synthétiques modifiables afin de comprendre la perception de l'âge [5, 6], ou pour l'augmentation des données dans la formation des réseaux neuronaux [7]. Pour réaliser ce type d'édition, les méthodes de vieillissement du visage [8, 9, 10] ne conviennent pas car ces méthodes visent à modifier toutes les rides en même temps. Ces méthodes ne permettent donc pas de démêler les rides, c'est-à-dire à éditer indépendamment chaque ride. D'autre part, la plupart des approches d'édition se concentrent sur des caractéristiques larges, telles que les cheveux, ou sur toutes les caractéristiques d'une région d'intérêt (ROI). Par exemple, l'édition de la couleur des lèvres implique également l'édition du sourire car ces deux éditions partagent une même ROI. Cependant, l'édition des rides nécessite de se concentrer sur des caractéristiques spécifiques au sein d'une région d'intérêt. Pour résoudre ce problème, nous proposons les contributions suivantes :

— Inspirés par le PTI [11], nous proposons de finetuner un GAN pour éditer des images. Notre méthode repose sur des directions d'édition définies par des pivots : l'image réelle à éditer et des versions modifiées de celle-ci. Cette approche garantit que les parties statiques restent inchangées (démêlées).
— Nous proposons de créer des fausses éditions via des modifications de l'image réelle, appelées pseudo-vérité terrain (PGT), pour obtenir les pivots.
— Pour faciliter le démêlement des caractéristiques,

FIGURE 1 – Exemple d'édition de rides démêlées avec notre Edition Mulit-Pivots (EMP). Les lignes représentent les images éditées pour une direction d'édition de rides donnée avec une force allant de -1 à +1. Sur les trois premières lignes (visage entier), nous pouvons voir que l'édition des rides est démêlée, c'est-à-dire que les autres parties de l'image (en particulier les autres rides) ne changent pas. Les trois dernières lignes (cadrage sur la ride) montrent l'évolution réaliste des rides.

les directions d'édition sont orthogonales entre elles et aux dimensions ayant la plus grande variance expliquée de l'espace latent.

La figure 1 montre des exemples de modifications démêlées que notre méthode peut effectuer.

2 Etat de l'art

StyleGAN [12, 13, 14] est le modèle le plus utilisé et le plus efficace (présentant le plus faible FID [15]) pour la synthèse de visages [16]. Les GAN ont été largement utilisés pour diverses applications [17] et notamment pour l'édition d'images [18, 19]. Ces applications nécessitent d'abord une inversion, c'est-à-dire la projection de l'image dans l'espace latent du GAN. Après avoir présenté les méthodes d'inversion du GAN (2.1), nous présenterons les méthodes d'édition du GAN (2.2).

2.1 Inversion des GANs

L'inversion du GAN est au centre de toute édition de GAN, puisqu'une mauvaise inversion conduira à une mauvaise édition. I2S [18, 19] a introduit des stratégies pour inverser efficacement les images, notamment en utilisant l'espace W+ ou en commençant par la variable latente moyenne. Cependant, le compromis entre distorsion et éditabilité [11] montre que si l'inversion est plus simple dans l'espace W+, l'édition est plus efficace dans l'espace W. Plusieurs méthodes de raffinement [11, 20] ont été proposées pour répondre à ce compromis. Parmi ces méthodes, le PTI (Pivotal Tuning Inversion) [11] fixe la variable latente trouvée après l'inversion (appelée pivot) et finetune le générateur pour faire correspondre l'image générée à l'image cible. Nous avons étendu l'idée du PTI [11] à la tâche d'édition en créant des directions d'édition synthétiques.

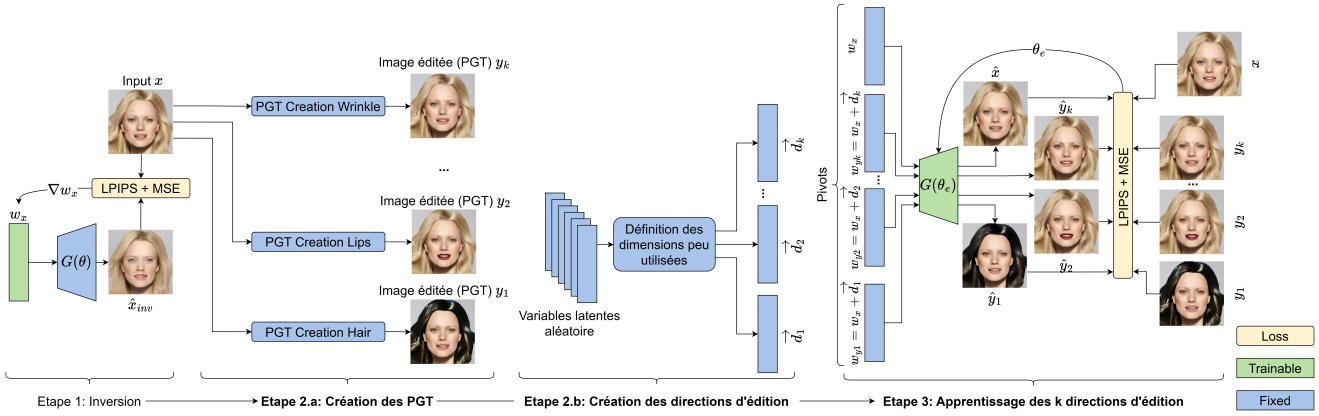

FIGURE 2 – Schéma global d'Edition Multi-Pivots. L'idée principale (étape 3) est de finetuner un GAN $G(\theta)$ à l'aide de variables latentes fixes (pivots). Ces variables définissent les directions d'édition souhaitées. Le premier pivot (étape 1) w_x est obtenu en inversant l'entrée x. Les autres pivots $w_{y1}...w_{yk}$ correspondent, respectivement, au premier pivot plus une direction d_1, ..., d_k. Ces directions sont obtenues en utilisant les dimensions de l'espace latent peu utilisées pour la génération des images (étape 2.b). Une Pseudo Vérité Terrain (PGT) est associée à chaque direction. Ces PGT sont obtenues en modifiant l'entrée x (étape 2.a).

2.2 Edition des GANs

Les avancées pour l'inversion des GANs ont permis une édition plus réaliste. Il existe plusieurs approches pour manipuler les attributs d'une image via les GANs [21].

Liaison de l'espace latent aux pixels de l'image Une première approche consiste à rechercher des modifications maximisant la direction à l'intérieur d'une région d'intérêt (ROI) tout en minimisant les perturbations à l'extérieur de la ROI [22, 23]. Resefa [22] utilise la matrice jacobienne pour relier une région d'intérêt de l'image à l'espace latent. Toutefois, cette approche conduit à des caractéristiques emmêlées à l'intérieur et à l'extérieur de la région d'intérêt pour certaines éditions. Cet emmêlement s'explique par la faible dimensionnalité de l'espace latent par rapport au nombre de pixels de l'image.

Modification du GAN Pour résoudre ce problème d'emmêlement de l'espace latent du GAN, certaines approches proposent de modifier le GAN [24, 10, 25, 26, 27]. StyleMapGAN [24] propose de remodeler l'espace latent pour maintenir l'information spatiale démêlée. VecGAN [25, 26] a modifié un StyleGAN et l'a réentrainé avec des nouvelles loss. Malgré un espace latent spatialement démêlé, ces méthodes emmêlent les caractéristiques à l'intérieur d'une région. Pour résoudre ce problème, nous proposons d'ajouter des objectifs pour démêler une caractéristique donnée au lieu d'une région.

Edition Textuelle Afin de ne pas limiter l'édition à un nombre restreint de tâches définies, des travaux antérieurs ont également exploré l'édition par du texte avec CLIP. CLIP (Contrastive Language-Image Pre-training) [28] est un réseau neuronal qui a formé conjointement des encodeurs d'images et de textes pour coupler ces modalités. StyleClip [29] et FEAT [30] ont utilisé cette capacité pour l'édition de texte. Bien qu'il soit prometteur pour l'édition d'images classiques, CLIP [28] n'a pas de connaissances sur les rides. Il n'est donc pas possible de modifier ces caractéristiques avec ces méthodes.

Composition d'Image Une autre approche consiste à envisager l'édition par interpolation des features map intermédiaires [31, 32]. SOAT [31] a exploré les propriétés de StyleGAN et a montré des résultats prometteurs en utilisant une interpolation des features map entre l'image d'entrée et une image de référence générée. Malgré la haute qualité des images éditées, ces méthodes manquent de personnalisation de l'image. Les résultats de l'édition sont similaires pour toutes les images de la région éditée. Notre Edition Multi-Pivots (EMP) montre des modifications personnalisées pour chaque image en finetunant le GAN avec l'image d'entrée et des Pseudo Vérité Terrain (PGT) correspondant aux éditions souhaitées.

Manipulation des variables latentes La dernière approche consiste à modifier la représentation latente de l'image d'entrée obtenue après inversion [33, 34, 35, 36, 37, 38]. Cependant la recherche d'une direction démêlée dans un espace latent GAN est difficile. Parihar et al. [39] proposent de créer une paire synthétique d'images positives et négatives. Après avoir inversé ces paires, ils ont pu déterminer une direction dans l'espace latent du GAN pour éditer la caractéristique souhaitée. Toutefois, cette approche est limitée aux caractéristiques qui sont déjà démêlées dans l'espace latent du GAN. Notre Edition Multi-Pivots (EMP) pousse cette idée plus loin en forçant le GAN à démêler la caractéristique souhaitée à l'aide de paires synthétiques composées de l'image réelle et de ses PGT.

3 Méthode

Après avoir présenté l'architecture (3.1), nous décrivons chacune des contributions : Pseudo-vérité terrain (PGT)

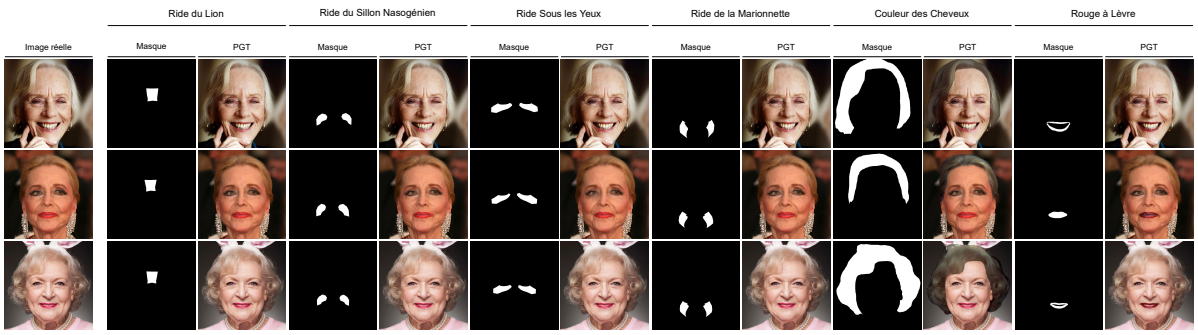

FIGURE 3 – Exemples de PGT et de masque d'édition. Dans le contexte de l'édition des rides, il convient de noter que nous avons choisi de supprimer les rides plutôt que de les augmenter. Pour les deux autres tâches, nous modifions grossièrement la couleur des cheveux ou des lèvres.

pour l'édition (3.2), création des directions orthogonales (3.3), et apprentissage des directions (3.4).

3.1 Formulation et Architecture

Étant donné une image réelle x, l'édition d'images démêlées consiste à générer une image $\hat{x}'$ telle que seules la ou les caractéristiques données de x diffèrent dans $\hat{x}'$. Dans cette étude, nous nous concentrons sur trois caractéristiques faciales : les rides (rides du lion, rides de la marionnette, rides nasolabiale et rides sous les yeux), la couleur des lèvres et la couleur des cheveux. Bien que notre étude se concentre sur l'édition de caractéristiques faciales spécifiques, notre EMP peut être appliquée à d'autres tâches d'édition. L'émergence des GANs a rendu possible l'édition de caractéristiques d'images en modifiant la projection latente d'une image réelle. En d'autres termes, étant donné une image réelle x et un GAN pré-entraîné G, il est possible de modifier la projection latente w_x de l'image réelle avec une direction d'édition $\overrightarrow{d}$, de telle sorte que

$$\hat{x}' = G(w_x + \alpha \times \overrightarrow{d}) \tag{1}$$

où $\hat{x}'$ est l'image éditée, w_x est la projection latente de x, G est un GAN pré-entraîné, α est la force d'édition, et $\overrightarrow{d}$ est la direction d'édition linéaire souhaitée dans l'espace latent. Nous avons choisi de contraindre les directions pour une édition linéaire des attributs. Notre Edition Multi-Pivots (EMP) trouve k directions d'édition $\overrightarrow{d_1}$, ..., $\overrightarrow{d_k}$ et les démêle des autres caractéristiques de l'image. L'ensemble du processus est illustré sur la figure 2 et chacune des 4 étapes est décrite ci-dessous.

Inversion La première étape pour éditer une image x avec un GAN pré-entraîné G consiste à inverser l'image à éditer (étape 1 dans la Fig. 2). L'objectif de l'inversion est de trouver la variable latente w_x telle que $G(w_x) \approx x$, où x est l'image d'entrée à inverser. Comme dans [11], nous trouvons la variable latente w_x en utilisant le processus d'optimisation suivant :

$$w_x = \underset{w_x}{argmin}\, \mathcal{L}_{per}(x, G(w_x, \theta)) + \|x - G(w_x, \theta)\|_2 \tag{2}$$

où $G(w_x, \theta)$ est l'image générée par le GAN G avec la variable latente w_x et les poids θ, x est l'image réelle à inverser, $\mathcal{L}_{per}$ est la fonction de perte perceptuelle [40], et $\|\|_2$ est la norme L2. L'image inversée est notée $\hat{x}_{inv}$.

Création des PGT Pour guider le réseau dans le processus d'apprentissage de l'édition, des fausses cibles sont créées : les pseudo-vérités terrain (PGT). Ces PGT sont créés via des techniques de traitement d'image traditionnelles à partir de l'image d'entrée x. Pour k directions d'édition, les PGT sont désignées par y_1, ..., y_k (étape 2.a de la figure 2). Le nombre de directions k est compris entre 1 et la dimensionnalité de l'espace latent, soit 512 pour StyleGAN. Plus de détails dans 3.2.

Création des directions d'édition Pour réaliser l'édition, une direction d'édition est associée à chacune des k pseudo-vérités terrain (PGT). Ces directions sont notées $\overrightarrow{d_1}$, ..., $\overrightarrow{d_k}$, respectivement pour les PGT y_1, ..., y_k. Ces directions sont obtenues via l'utilisation des dimensions de l'espace latent peu utilisées pour la génération des images (étape 2.b de la figure 2), plus de détails dans 3.3.

Apprentissage des directions d'édition L'image inversée $\hat{x}_{inv}$ est souvent floue par rapport à l'image originale et l'identité est perdue. Pour résoudre ce problème, le PTI [11] finetune le générateur G de sorte que l'image générée $\hat{x}_{inv}$ soit similaire à l'image réelle x [11]. Cette opération de finetuning crée une copie de l'image d'entrée dans l'espace latent du générateur avec les poids θ_{pt}. Nous avons poussé plus loin cette idée de finetuning du générateur pour éditer les attributs d'une image réelle x. Une fois que le générateur G est finetuné, il peut éditer les caractéristiques souhaitées sur l'image x en suivant les directions d'éditions $\overrightarrow{d_1}$, ..., $\overrightarrow{d_k}$. Plus de détails dans 3.4.

3.2 Pseudo-Vérité Terrain

Il n'est pas toujours possible d'obtenir une vérité terrain correspondant à l'édition souhaitée. Pour résoudre ce problème, nous introduisons la pseudo-vérité terrain. Une pseudo-vérité, ou Pseudo Ground Truth (PGT), est une fausse édition créée à l'aide de méthodes de vision par ordinateur appliquées à l'image réelle x. Pour notre Edition

FIGURE 4 – Comparaison globale des méthodes de l'état de l'art Resefa [22], StyleMapGAN [24], et SOAT [31] avec notre EMP pour l'édition de la ride du lion (quatre premières lignes) et de la ride sous les yeux (quatre dernières lignes). La variation (Δ) entre les images éditées le long d'une direction d'édition illustre que les éditions avec l'EMP sont plus démêlées.

Multi-Pivots (EMP), le PGT est utilisé pour guider le fine-tuning en montrant au générateur ce à quoi l'image d'entrée devrait ressembler après l'édition (étape 2.a de la Fig. 2). Le PGT peut représenter soit l'édition souhaitée, soit son opposé. Par exemple, pour l'édition des rides, le PGT peut augmenter ou supprimer la ride souhaitée. La figure 3 montre des exemples avec la zone d'édition correspondante (masque) et le PGT pour différentes tâches d'édition (édition des rides, de la couleur des lèvres et de la couleur des cheveux). Comme le montre la figure 3, le PGT n'est pas destiné à être une cible d'édition réaliste, mais à éditer la caractéristique souhaitée tout en laissant les autres inchangées. Plus précisément, pour la tâche d'édition des rides, elles sont identifiées via un filtre hessien (HHF) [41], puis supprimées à l'aide de l'algorithme de Poisson blending [42]. Pour l'édition de la couleur des lèvres et des cheveux, la couleur est copiée à partir d'une image de référence et appliquée à l'image d'entrée x à l'aide de la méthode de correspondance d'histogramme.

3.3 Création des directions d'édition

Härkönen et al. [43] ont démontré que seules 100 dimensions de l'espace W sont nécessaires pour générer une image sans perte de qualité visible. Ces résultats ont été obtenus en utilisant les 100 premières dimensions d'une PCA entrainée à partir de 10 000 variables latentes issues de W. Les 412 dernières dimensions contrôlent donc des modifications imperceptibles. En se basant sur ces résultats, nous proposons de créer k directions d'éditions dans les dernières dimensions (les dimensions ne représentant que 5% de l'énergie totale) de l'espace latent transformé par une PCA. Cela permet d'ajouter de nouvelles informations sans altérer les capacités génératives du GAN. La formulation de transformation de la PCA est :

$$w' = (w - B) \cdot C^{\intercal} \tag{3}$$

avec w une variable latente, w' est la projection de w dans l'espace de la PCA, B est la moyenne empirique de l'espace W et C la matrice des composantes principales. Les m premières composantes principales sont alors définies :

$$p_1, p_2, ..., p_m = C_{m,*} \tag{4}$$

avec p_1, ..., p_m les m premières principales composantes et $C_{m,*}$ les m première ligne de matrice des composantes de la PCA. Pour ne pas interférer avec la capacité générative du GAN les directions d'édition doivent donc être orthogonales à ces premières composantes p_1, ..., p_m. De plus, nous avons choisi de forcer les directions d'édition à être orthogonales entre elles pour faciliter l'indépendance de chacune des directions. L'orthogonalité des directions peut être garantie via le procédé de Gram-Schmidt. Pour

un ensemble de vecteurs v_1, ..., v_k, le procédé de Gram-Schmidt définit les vecteur orthogonaux u_1, ..., u_k comme :

$$u_k = v_k - \sum_{j=1}^{k-1} proj_{u_j}(v_k) \tag{5}$$

$$proj_u(v) = \frac{\langle v, u \rangle}{\langle u, u \rangle} u \tag{6}$$

avec $\langle , \rangle$ le produit scalaire et v_1, ..., v_k des vecteurs quelconques. Les k dimensions d_1, ..., d_k sont alors définies de manière suivante :

$$\begin{aligned} \overrightarrow{d_1} &= a_1 - \sum_{j=1}^{m} proj_{p_j}(a_1) \\ &\vdots \\ \overrightarrow{d_k} &= a_k - \sum_{i=1}^{k-1} proj_{\overrightarrow{d_i}}(a_k) - \sum_{j=1}^{m} proj_{p_j}(a_k) \end{aligned} \tag{7}$$

avec d_1, ..., d_k des directions d'édition, a_1, ..., a_k des variables aléatoires suivant une loi uniforme $\mathcal{U}(0, 1)$ et p_1, ..., p_m les premières principales dimensions de la PCA.

3.4 Apprentissage des directions

Pour apprendre des directions d'édition appropriées (étape 3 de la figure 2), nous projetons avec le PTI [11] les PGT dans les directions définies en 3.3. Pour réaliser une telle projection, plusieurs cibles sont nécessaires : une image réelle x et k cibles d'édition y_1, ..., y_k (détaillée dans 3.2). Les poids θ du GAN sont libérés pour apprendre la direction d'édition. Pendant cet apprentissage, les pivots restent fixes. L'objectif de l'apprentissage est le suivant :

$$\begin{aligned} \theta_e = \underset{\theta_e}{argmin}\ \mathcal{L}_{per}(x, G(w_x, \theta_e)) + \|x - G(w_x, \theta_e)\|_2 \\ + \sum_{i=0}^{k} \mathcal{L}_{per}(y_i, G(w_x + \overrightarrow{d_i}, \theta_e)) + \left\|y_i - G(w_x + \overrightarrow{d_i}, \theta_e)\right\|_2 \end{aligned} \tag{8}$$

où θ_e sont les poids accordés, $\mathcal{L}_{per}$ est la perte perceptuelle [40], $\|\|_2$ est la norme L2.

4 Résultat

Dans cette section, après avoir présenté les détails d'implémentations (4.1), nous comparons qualitativement notre EMP avec les méthodes d'édition de l'état de l'art (4.2). Ensuite, nous évaluons quantitativement les résultats (4.3). Pour confirmer les résultats quantitatifs, nous avons mené une étude auprès des utilisateurs (4.4). Enfin, nous examinons les limites de la méthode que nous proposons (4.5).

4.1 Détails d'Implémentation

Le GAN utilisé est un StyleGAN-XL [14] pour la Fig. 1 et un StyleGAN-2 [13] dans le cas contraire. Pour les étapes d'inversion et de finetuning, nous utilisons les mêmes hyperparamètres que ceux décrits par Roich et al. [11]. Nous

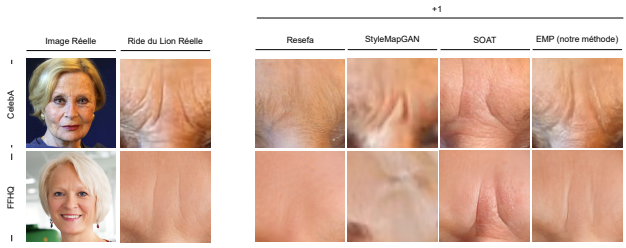

FIGURE 5 – Comparaison locale avec les méthodes de l'état de l'art démêlées Resefa [22], StyleMapGAN [24], et SOAT [31] pour l'édition des rides du lion.

fixons le nombre d'itérations pour l'inversion à 1000, et le learning rate pendant l'inversion est fixé à 5×10^{-3}. Comme le suggère I2S [18], la variable latente inversée est initialisée sur la moyenne de 10 000 variables latentes aléatoires. Nous entrainons ensuite le générateur pour 350 itérations avec un learning rate de 3×10^{-4}. Pour l'inversion et le finetuning, nous utilisons l'optimiseur Adam [44]. Pour l'apprentissage de l'édition, nous utilisons $\lambda_{L_2} = 10$ et $\lambda_{L_{LPIPS}} = 1$. Le nombre de dimensions m de la PCA figée est fixé à 196 dimensions (soit 95% de l'énergie). Notre approche prend moins de 5 minutes sur un seul GPU Nvidia GeForce RTX 3080 Laptop 16 Go pour des images d'une résolution de 1024×1024. L'inversion initiale prend environ 2 minutes et peut être accélérée à l'aide d'encodeurs tels que psp [45] ou e4e [46]. L'apprentissage de l'édition prend environ 3 minutes. Toutes les expériences qualitatives et quantitatives ont été réalisées sur un échantillon de femmes âgées avec des rides pour l'édition des rides, et sur un autre échantillon de femmes pour l'édition des lèvres et des cheveux. Tous les échantillons ont été extraits du FFHQ [12] et de CelebA [47].

4.2 Résultats Qualitatifs

Nous analysons d'abord les résultats qualitativement selon deux critères : le démêlage global de l'édition (Fig. 4) et la qualité locale de l'édition (Fig. 5).

Démêlage Global Nous commençons par analyser qualitativement les résultats du démêlage global. La figure 4 montre quatre visages avec des rides (deux de l'ensemble de données CelebA [47] et deux de l'ensemble de données FFHQ [22]). Pour chaque méthode de l'état de l'art, nous comparons les modifications qui augmentent les rides (+1) et celles qui les suppriment (-1). Une carte de chaleur des différences entre l'augmentation et la suppression des rides est également fournie pour aider à identifier la localisation et l'intensité des changements (Δ). Nous observons que Resefa [22] modifie l'expression du visage pendant l'édition. Nous observons que StyleMapGAN [24] édite une zone plus large que la zone d'édition définie, en particulier pour les images échantillonnées à partir de l'ensemble de données FFHQ. Nous observons que SOAT [31] modifie toutes les caractéristiques de la région d'intérêt, y compris les pores de la peau, le teint de la peau et les rides. Toutes

		Identity preservation			Image quality		
	Metrics	$IC \downarrow$	$LD_o \downarrow$	$MSE_o \downarrow$	$SSIM \uparrow$	$FID \downarrow$	$KID \downarrow$
Lion	Resefa [22]	0.32	154	0.89	0.859	89	1.70
	StyleMapGAN [24]	0.27	**29**	1.43	0.963	119	2.37
	SOAT [31]	**0.08**	51	**0.09**	**0.995**	76	0.87
	EMP	0.11	67	0.21	0.969	**31**	**0.31**
Sous les Yeux	Resefa [22]	0.30	208	0.93	0.840	94	1.68
	StyleMapGAN [24]	0.29	**44**	1.68	0.946	144	4.24
	SOAT [31]	0.14	97	0.23	**0.983**	80	0.41
	EMP	**0.10**	92	**0.20**	0.972	**31**	**0.28**
Marionette	Resefa [22]	0.36	213	0.96	0.841	91	1.16
	StyleMapGAN [24]	0.24	**42**	2.17	0.926	126	2.55
	SOAT [31]	0.14	116	0.34	**0.974**	82	0.94
	EMP	**0.09**	91	**0.22**	0.967	**33**	**0.42**
Nasogénien	Resefa [22]	0.25	150	1.11	0.842	104	1.29
	StyleMapGAN [24]	0.30	**54**	2.02	0.931	149	3.64
	SOAT [31]	0.15	117	0.24	**0.985**	93	0.50
	EMP	**0.10**	85	**0.10**	0.972	**36**	**0.10**
Lèvre	Resefa [22]	0.48	129	0.94	0.842	66	0.51
	StyleMapGAN [24]	0.45	66	2.00	0.940	95	2.48
	StyleClip [29]	0.16	68	**0.29**	**0.975**	69	0.90
	FEAT [30]	0.23	103	1.37	0.724	77	1.71
	EMP	**0.15**	**65**	0.34	0.965	**28**	**0.17**
Cheveux	StyleGANEX [10]	0.54	144	1.97	0.579	70	2.59
	VecGAN++ [26]	0.29	**27**	2.73	**0.781**	73	2.94
	StyleClip [29]	0.30	102	1.04	0.584	77	1.59
	FEAT [30]	0.36	105	1.10	0.674	61	0.37
	CtrlHair [48]	0.28	**27**	3.61	0.548	63	1.21
	EMP	**0.21**	87	**0.61**	0.718	**41**	**0.11**

TABLE 1 – Comparaison de la préservation de l'identité et de la qualité de l'image pour les méthodes de l'état de l'art et notre EMP. Les résultats MSE et KID sont arrondis à $1e^{-2}$. Chaque méthode est affectée à ses tâches disponibles.

ces limitations des méthodes de l'état de l'art sont justifiées par leur manque de précision. En effet, StyleMapGAN [24] et SOAT [31] modifient les caractéristiques à l'intérieur d'une région donnée sans autre indication, ce qui produit des modifications non souhaitées à l'intérieur de la région d'intérêt. En revanche, notre EMP ne montre pas changement visible en dehors de la modification des rides.

Qualité Locale Nous comparons la qualité de l'édition locale (Fig. 5). La première ligne est une image issue de CelebA [47] et la seconde de FFHQ [22]. Dans cette partie, nous concentrons notre analyse sur l'évolution des rides afin d'étudier le réalisme des modifications. Pour ce faire, nous faisons un zoom sur la région d'intérêt pour chaque méthode de l'état de l'art et notre EMP pour l'édition de la ride du lion. Alors qu'elle devrait augmenter la ride, nous observons que Resefa [22] supprime les rides. Cela peut s'expliquer par le manque d'informations sur la ride après l'inversion et le mauvais démêlage des caractéristiques dans la zone d'édition. Inversement, nous constatons que SOAT [31] perd les caractéristiques originales de la ride en modifiant la forme et la localisation de la ride. SOAT [31] étant une technique de composition d'images, la texture et la couleur de la peau sont également modifiées. Nous remarquons que StyleMapGAN raccourcit et agrandit la ride. Nous observons également que les résultats de StyleMapGAN [24] sont flous. Enfin, nous constatons que notre EMP agrandit la ride en prolongeant la ride originale.

4.3 Résultats Quantitatifs

Nous avons effectué une analyse quantitative sur 50 images par tâche d'édition. Cette analyse qualitative vise à évaluer notre EMP par rapport à l'état de l'art selon deux critères (Tableau 1) : la préservation de l'identité et la qualité de l'image. Les métriques sont appliquées entre les directions d'éditions +1 et -1. Les images sont échantillonnées de manière égale à partir des ensembles de données CelebA [47] et FFHQ [12]. Les méthodes de l'état de l'art sont affectées à des tâches d'édition spécifiques pour lesquelles elles ont été conçues. Ainsi, les méthodes de l'état de l'art diffèrent

	Evaluation	Identity	Quality	Realism
Lion	Resefa [22]	15	16	15
	StyleMapGAN [24]	12	11	8
	SOAT [31]	5	7	7
	EMP	**67**	**67**	**71**
Sous les Yeux	Resefa [22]	10	11	9
	StyleMapGAN [24]	10	10	11
	SOAT [31]	9	9	13
	EMP	**71**	**70**	**67**
Marionette	Resefa [22]	13	13	13
	StyleMapGAN [24]	9	9	9
	SOAT [31]	15	15	18
	EMP	**63**	**63**	**60**
Nasogénien	Resefa [22]	14	12	15
	StyleMapGAN [24]	8	7	11
	SOAT [31]	18	19	16
	EMP	**60**	**61**	**57**
Lèvre	Resefa [22]	7	4	–
	StyleMapGAN [24]	10	8	–
	StyleClip [29]	21	24	–
	FEAT [30]	10	11	–
	EMP	**63**	**64**	–
Cheveux	StyleGANEX [10]	5	7	–
	VecGAN++ [26]	23	19	–
	StyleClip [29]	14	12	–
	FEAT [30]	5	12	–
	CtrlHair [48]	12	14	–
	EMP	**40**	**36**	–

TABLE 2 – Résultats de l'étude utilisateurs (en pourcentages). L'Identité, Qualité et Réalisme sont, respectivement, les mesures de la préservation de l'identité, de la qualité de l'image et du réalisme de l'évolution des rides. Chaque méthode est évaluée sur ses tâches d'édition.

selon le type de tâche d'édition.

Préservation de l'Identité La préservation de l'identité est fondamentale pour les méthodes d'édition. Nous l'avons mesurée qualitativement à l'aide de trois mesures :

- Conservation de l'identité (IC) : Elle vise à mesurer la préservation de l'identité au cours de l'édition en calculant la distance entre les caractéristiques d'un réseau de reconnaissance faciale FaceNet [49].
- Distance entre les points morphologiques en dehors de la zone éditée (LD_o) : Ce paramètre évalue la préservation de la forme et de l'expression du visage en mesurant la distance entre les points morphologiques calculés via MediaPipe [50].
- Erreur quadratique moyenne (MSE_o) : Elle mesure les modifications sur l'image globale (distance en pixels) en dehors de la zone d'édition.

Pour toutes les tâches d'édition (tableau 1), l'EMP obtient les meilleurs résultats pour la conservation de l'identité (IC), de bons résultats pour la distance des points morphologiques (LD_o) et pour l'erreur quadratique moyenne en dehors de la zone éditée (MSE_o).

Qualité d'image $SSIM$ [51], FID [15] et KID [52] sont utilisés pour mesurer quantitativement la qualité de l'image après l'édition. Nous constatons que pour FID [15] et KID [52], l'EMP surpasse toutes les méthodes de l'état de l'art. Pour $SSIM$, l'EMP montre de bons résultats par rapport aux méthodes de l'état de l'art. SOAT [31] obtient d'excellents résultats sur $SSIM$. Ces résultats confirment l'observation de la figure 4. Cela peut s'expliquer par le fait que SOAT [31] édite une zone définie des features map, ce qui permet d'obtenir une qualité d'image équivalente à celle générée. Pour chaque tâche et chaque mesure, notre EMP produit des résultats similaires ou meilleurs que les méthodes les plus récentes.

4.4 Etude Utilisateurs

Nous avons mené une étude auprès d'utilisateurs afin d'évaluer notre méthode. Nous avons sélectionné au hasard 10 images d'entrée pour chacune des 6 tâches d'édition étudiées dans cet article. Il a été demandé à 25 participants de choisir la meilleure image éditée parmi ces échantillons en fonction de trois aspects : la conservation de l'identité, la qualité de l'image et le réalisme de l'évolution (pour l'édition des rides). Pour chacune des 6 tâches d'édition et pour chaque utilisateur, 5 images sont choisies au hasard. Chaque image est présentée avec ses éditions aux participants. L'ordre des tâches et des images est aléatoire pour chaque participant. Les participants ont été invités à voter pour le résultat qu'ils préféraient. Enfin, nous avons calculé le pourcentage de votes pour chaque méthode dans chaque tâche d'édition (tableau 2). Notre méthode reçoit le pourcentage de votes le plus élevé pour chaque aspect évalué.

4.5 Discussion

Pour l'édition des rides, le PGT requiert une ride pour pouvoir la retirer. Cela limite l'approche aux personnes ayant une ride. L'approche pourra être étendue avec un meilleur PGT. Notre méthode repose sur l'apprentissage du générateur, elle est plus lente que toutes les autres méthodes de l'état de l'art. Ce problème peut être résolu à l'avenir en utilisant des techniques d'apprentissage.

5 Conclusion

Nous avons proposé une nouvelle méthode d'édition, appelée Edition Multi-Pivots (EMP). Cette approche d'édition repose sur le finetuning d'un GAN avec une image et des PGT associés (i.e., une image éditée via des techniques de traitement d'image) pour générer une édition significative dans l'espace latent du GAN. Les résultats quantitatifs et qualitatifs montrent que nous surpassons les méthodes de l'état de l'art pour le démêlement de l'espace latent, le réalisme et la préservation de l'identité au cours de l'édition.

Références

[1] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *arXiv*, 2014.

[2] A. Aggarwal, M. Mittal, and G. Battineni, "Generative adversarial network : An overview of theory and applications," *International Journal of Information Management Data Insights*, vol. 1, p. 100004, Apr 2021.

[3] D. Bau, J.-Y. Zhu, J. Wulff, W. Peebles, B. Zhou, H. Strobelt, and A. Torralba, *Seeing what a GAN cannot generate*, p. 4501–4510. IEEE, Oct 2019.

[4] P. Vatiwutipong, S. Vachmanus, T. Noraset, and S. Tuarob, "Artificial intelligence in cosmetic dermatology : A systematic literature review," *IEEE access : practical innovations, open solutions*, vol. 11, p. 71407–71425, 2023.

[5] J. Aznar-Casanova, N. Torro-Alves, and S. Fukusima, "How much older do you get when a wrinkle appears on your face ? modifying age estimates by number of wrinkles.," *Neuropsychology, Development, and Cognition. Section B, Aging, Neuropsychology and Cognition*, vol. 17, p. 406–421, Jan 2010.

[6] A. Nkengne, C. Bertin, G. N. Stamatas, A. Giron, A. Rossi, N. Issachar, and B. Fertil, "Influence of facial skin attributes on the perceived age of caucasian women.," *Journal of the European Academy of Dermatology and Venereology*, vol. 22, p. 982–991, Aug 2008.

[7] X. Wang, K. Wang, and S. Lian, "A survey on face data augmentation for the training of deep neural networks," *Neural Computing and Applications*, vol. 32, p. 15503–15531, Oct 2020.

[8] Z. Huang, S. Chen, J. Zhang, and H. Shan, "Pfagan : Progressive face aging with generative adversarial network," *IEEE Transactions on Information Forensics and Security*, vol. 16, p. 2031–2045, 2021.

[9] G. Antipov, M. Baccouche, and J.-L. Dugelay, *Face aging with conditional generative adversarial networks*, p. 2089–2093. IEEE, Sep 2017.

[10] S. Yang, L. Jiang, Z. Liu, and C. C. Loy, *StyleGANEX : StyleGAN-Based Manipulation Beyond Cropped Aligned Faces*, p. 20943–20953. IEEE, Oct 2023.

[11] D. Roich, R. Mokady, A. H. Bermano, and D. Cohen-Or, "Pivotal tuning for latent-based editing of real images," *ACM Transactions on Graphics (TOG)*, vol. 42, p. 1–13, Feb 2023.

[12] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4396–4405, 2019.

[13] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[14] A. Sauer, K. Schwarz, and A. Geiger, "Stylegan-xl : Scaling stylegan to large diverse datasets," in *ACM SIGGRAPH 2022 Conference Proceedings*, SIGGRAPH '22, (New York, NY, USA), Association for Computing Machinery, 2022.

[15] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, (Red Hook, NY, USA), p. 6629–6640, Curran Associates Inc., 2017.

[16] S. Bond-Taylor, A. Leach, Y. Long, and C. G. Willcocks, "Deep generative modelling : A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models.," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, p. 7327–7347, Nov 2022.

[17] W. Xia, Y. Zhang, Y. Yang, J.-H. Xue, B. Zhou, and M.-H. Yang, "Gan inversion : A survey," *IEEE transactions on pattern analysis and machine intelligence*, p. 1–17, 2022.

[18] R. Abdal, Y. Qin, and P. Wonka, *Image2stylegan : how to embed images into the stylegan latent space ?*, p. 4431–4440. IEEE, Oct 2019.

[19] R. Abdal, Y. Qin, and P. Wonka, *Image2stylegan++ : how to edit the embedded images ?*, p. 8293–8302. IEEE, Jun 2020.

[20] A. Bhattad, V. Shah, D. Hoiem, and D. A. Forsyth, "Make it so : Steering stylegan for any image inversion and editing," *arXiv*, 2023.

[21] M. Liu, Y. Wei, X. Wu, W. Zuo, and L. Zhang, "Survey on leveraging pre-trained generative adversarial networks for image editing and restoration," *Science China Information Sciences*, vol. 66, p. 151101, May 2023.

[22] J. Zhu, Y. Shen, Y. Xu, D. Zhao, and Q. Chen, "Region-based semantic factorization in GANs," in *International Conference on Machine Learning (ICML)*, 2022.

[23] B. Li, Q. Wang, J. Pei, Y. Yang, and X. Ji, "Which style makes me attractive ? interpretable control discovery and counterfactual explanation on stylegan," *arXiv*, 2022.

[24] H. Kim, Y. Choi, J. Kim, S. Yoo, and Y. Uh, *Exploiting Spatial Dimensions of Latent in GAN for Real-time Image Editing*, p. 852–861. IEEE, Jun 2021.

[25] Y. Dalva, S. F. Altındiş, and A. Dundar, *VecGAN : Image-to-Image Translation with Interpretable Latent Directions*, vol. 13676 of *Lecture notes in computer science*, p. 153–169. 2022.

[26] Y. Dalva, H. Pehlivan, O. I. Hatipoglu, C. Moran, and A. Dundar, "Image-to-image translation with disentangled latent vectors for face editing.," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, p. 14777–14788, Dec 2023.

[27] A. Suwała, B. Wójcik, M. Proszewska, J. Tabor, P. Spurek, and M. Śmieja, "Face identity-aware disentanglement in stylegan," *arXiv*, 2023.

[28] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, and et al., "Learning transferable visual models from natural language supervision," *arXiv*, 2021.

[29] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, and D. Lischinski, *StyleCLIP : Text-Driven Manipulation of StyleGAN Imagery*, p. 2065–2074. IEEE, Oct 2021.

[30] X. Hou, L. Shen, O. Patashnik, D. Cohen-Or, and H. Huang, "Feat : Face editing with attention," *arXiv*, 2022.

[31] M. J. Chong, H.-Y. Lee, and D. Forsyth, "Stylegan of all trades : Image manipulation with only pretrained stylegan," *arXiv*, 2021.

[32] E. Schubert, A. Lang, and G. Feher, *Accelerating Spherical k-Means*, vol. 13058 of *Lecture notes in computer science*, p. 217–231. Springer International Publishing, 2021.

[33] S. Khodadadeh, S. Ghadar, S. Motiian, W.-A. Lin, L. Boloni, and R. Kalarot, *Latent to Latent : A Learned Mapper for Identity Preserving Editing of Multiple Face Attributes in StyleGAN-generated Images*, p. 3677–3685. IEEE, Jan 2022.

[34] G. Balakrishnan, R. Gadde, A. Martinez, and P. Perona, "Rayleigh eigendirections (reds) : Gan latent space traversals for multidimensional features," *arXiv*, 2022.

[35] Y. Nitzan, K. Aberman, Q. He, O. Liba, M. Yarom, Y. Gandelsman, I. Mosseri, Y. Pritch, and D. Cohen-Or, "Mystyle : A personalized generative prior," *ACM Trans. Graph.*, vol. 41, nov 2022.

[36] C. Naveh and Y. Hel-Or, "Orthogan : Multifaceted semantics for disentangled face editing," *arXiv*, 2022.

[37] Y. Shen, C. Yang, X. Tang, and B. Zhou, "Interfacegan : interpreting the disentangled face representation learned by gans.," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, p. 2004–2018, Apr 2022.

[38] Y. Liu, Q. Li, Q. Deng, and Z. Sun, "Towards spatially disentangled manipulation of face images with pretrained stylegans," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, p. 1725–1739, Apr 2023.

[39] R. Parihar, A. Dhiman, T. Karmali, and V. R, "Everything is there in latent space : Attribute editing and attribute style manipulation by stylegan latent space exploration," in *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, (New York, NY, USA), p. 1828–1836, Association for Computing Machinery, 2022.

[40] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, *The unreasonable effectiveness of deep features as a perceptual metric*, p. 586–595. IEEE, Jun 2018.

[41] C.-C. Ng, M. H. Yap, N. Costen, and B. Li, *Automatic wrinkle detection using hybrid hessian filter*, vol. 9005 of *Lecture notes in computer science*, p. 609–622. Springer International Publishing, 2015.

[42] P. Pérez, M. Gangnet, and A. Blake, *Poisson image editing*, p. 313–318. ACM, Jul 2003.

[43] E. Härkönen, A. Hertzmann, J. Lehtinen, and S. Paris, "Ganspace : Discovering interpretable gan controls," in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds.), vol. 33, pp. 9841–9850, Curran Associates, Inc., 2020.

[44] D. P. Kingma and J. Ba, "Adam : A method for stochastic optimization," *arXiv*, 2014.

[45] E. Richardson, Y. Alaluf, O. Patashnik, Y. Nitzan, Y. Azar, S. Shapiro, and D. Cohen-Or, *Encoding in Style : a StyleGAN Encoder for Image-to-Image Translation*, p. 2287–2296. IEEE, Jun 2021.

[46] O. Tov, Y. Alaluf, Y. Nitzan, O. Patashnik, and D. Cohen-Or, "Designing an encoder for stylegan image manipulation," *arXiv*, 2021.

[47] 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.

[48] X. Guo, M. Kan, T. Chen, and S. Shan, *GAN with Multivariate Disentangling for Controllable Hair Editing*, vol. 13675 of *Lecture notes in computer science*, p. 655–670. Springer Nature Switzerland, 2022.

[49] F. Schroff, D. Kalenichenko, and J. Philbin, *FaceNet : A unified embedding for face recognition and clustering*, p. 815–823. IEEE, Jun 2015.

[50] Google, *MediaPipe*. Google, 2019.

[51] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment : from error visibility to structural similarity.," vol. 13, p. 600–612, Apr 2004.

[52] M. Binkowski, D. J. Sutherland, M. Arbel, and A. Gretton, "Demystifying mmd gans," *arXiv*, 2018.