Projet IMAGE
Edition du genre d'un portrait
Etat de l'art

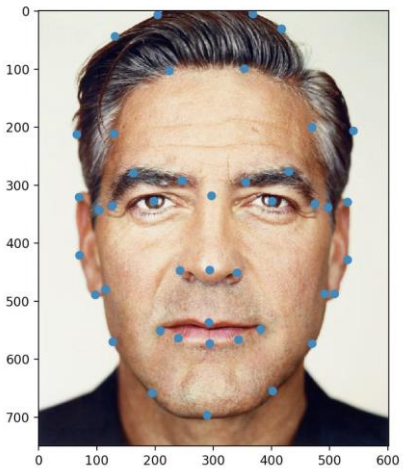
Au travers de ce document, nous essaierons de décrire l'ensemble des méthodes pouvant être utiles à la réalisation de notre projet.

Morphing facial

Cette méthode repose sur quatre grandes étapes :

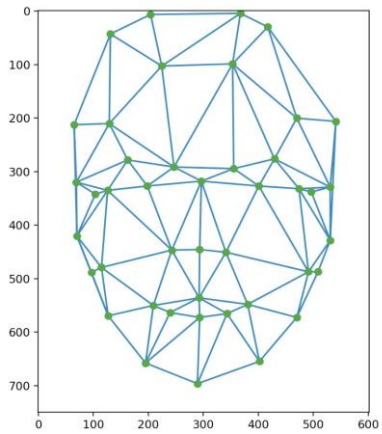
- **Définition des points de correspondance** :

Cette étape consiste à identifier et marquer précisément des points clés sur les visages masculins et féminin qui serviront de repères pour les transformations ultérieures. Ils sont sélectionnés pour capturer la structure globale du visage par exemple les coins des yeux, du nez, de la bouche ainsi que les contours du visage et les lignes de la mâchoire.



- **Triangulation de Delaunay**

La triangulation de Delaunay est une méthode largement utilisée en géométrie informatique et en traitement d'images pour diviser un ensemble de points en un maillage de triangles non superposés, de sorte que les triangles produits respectent certaines propriétés spécifiques. Dans notre cas, la méthode est utilisée pour diviser les points de correspondance définis sur les visages en un ensemble de triangle que nous utiliserons par la suite comme base au morphing.
(Explication de l'algorithme : Méthode de Delaunay)



- **Calcul du visage moyen de la population :**

Le but de ce calcul est de créer une représentation composite qui capture les caractéristiques moyennes des visages d'un groupe de personnes données. On peut utiliser des bases de données de visages existantes qui fournissent des visages avec des points de correspondance annotés.
En faisant ce calcul on obtient une forme moyenne pour les visages masculins et féminins, en prenant en compte les caractéristiques spécifiques à chaque sexe. Cette approche permet d'obtenir un aperçu global des traits faciaux moyens caractéristiques des hommes et des femmes.

- **Changer genre d'une image :**

Pour changer un visage masculin en un visage féminin, on se base sur le calcul du visage moyen déjà fait pour déformer sa forme vers la forme féminine moyenne et ensuite extrapoler sa couleur en conséquence pour ajuster les tonalités et les nuances afin de mieux correspondre aux caractéristiques colorimétriques communes des visages féminins.

Réseaux génératifs antagonistes (GAN)

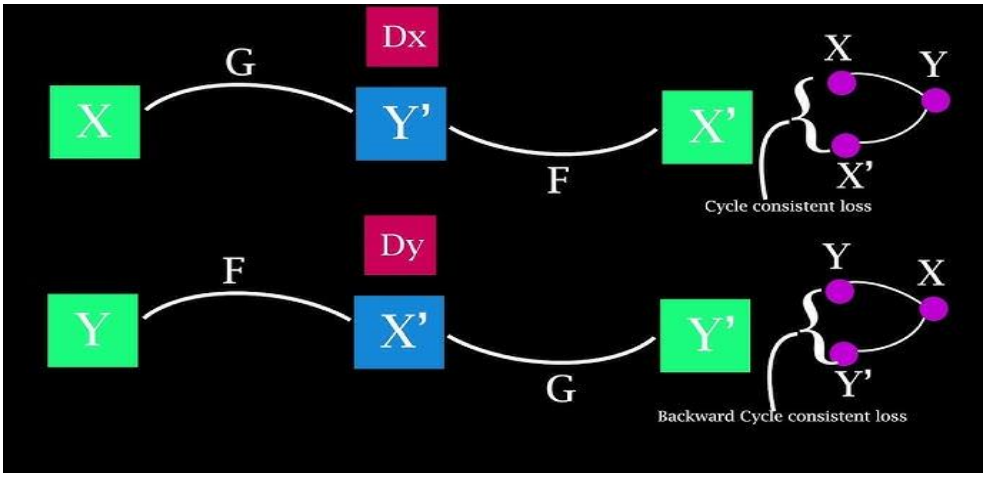
Il existe deux approches de la famille des GAN qui sont : une approche d'entraînement non supervisée avec CycleGAN et une approche d'entraînement supervisée avec pix2pixHD.

Méthode non supervisée avec CycleGAN :

Cette méthode utilise une variation du CycleGAN appelée UNIT (Translation Unpaired Image-to-Image Network). CycleGAN est un type de réseau de transfert de style non supervisé. Il s'agit d'une implémentation modifiée des GAN qui a la capacité de transférer le style d'une image à une autre.

Le CycleGAN se compose de deux générateurs et de deux discriminateurs. Le premier générateur de ce réseau transfère le style des images étiquetées A vers les images étiquetées B. Le deuxième générateur tente de les transférer de B vers A. Chacun des discriminateurs décide si les styles étiquetés A et B correspondent réellement aux images générées. Le nom "CycleGAN" provient de la fonction de perte supplémentaire qui contrôle l'application

conséquente des deux générateurs B2A(A2B(x)) et A2B(B2A(x)), ainsi que de la comparaison des images générées avec les images initialement données au réseau.



Méthode supervisée avec pix2pixHD :

Cette méthode repose sur l'utilisation d'une architecture de réseau de neurones appelée pix2pixHD pour la transformation d'images homme-femme. Cette approche nécessite un ensemble de données d'entraînement composé de paires de photos, chaque paire contenant un exemple d'une personne représentée en tant qu'homme et un exemple d'une personne représentée en tant que femme, les deux photos de la paire étant une représentation de la même personne.

Pix2pix est un cadre GAN conditionnel pour la traduction d'images. Il se compose d'un générateur G et d'un discriminateur D. L'objectif du générateur G est de traduire des cartes d'étiquettes sémantiques en images réalistes, tandis que le discriminateur D vise à distinguer les images réelles des images traduites. Le framework fonctionne dans un contexte supervisé, où l'ensemble de données d'entraînement est fourni sous la forme d'un ensemble de paires d'images correspondantes.

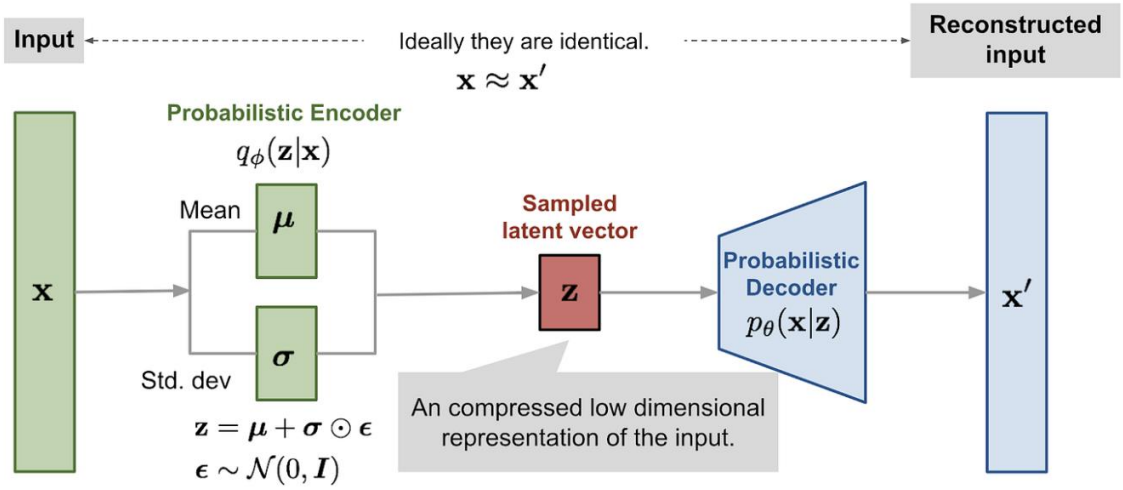
La principale différence entre pix2pix et pix2pixHD réside dans le fait qu'avec pix2pixHD, le générateur est divisé en deux parties qui permettent la génération d'images haute résolution. Cette architecture améliorée vise à surmonter les limitations de l'entraînement instable du framework pix2pix, ce qui permet de générer des images de meilleure qualité et résolution.

Variational Auto Encoder (VAE)

Un auto-encodeur variationnel (VAE) est un algorithme d'IA génératif qui utilise l'apprentissage profond pour générer de nouveaux contenus. L'auto-encodeur est une structure de réseau neuronal très simple qui se compose de deux parties, l'encodeur et le décodeur. Il apprend à compresser l'entrée originale dans une forme compressée dans un espace vectoriel

beaucoup plus petit, et le décodeur apprend à reconstruire les données compressées dans l'entrée originale avec une certaine perte.

Les auto-encodeurs traditionnels apprennent à compresser et à reconstruire des données, mais n'aident pas vraiment à générer de nouvelles données. C'est là que l'encodeur variationnel (VAE) s'avère utile. Le VAE apprend la distribution des données au lieu d'une simple image compressée, et en utilisant la distribution, nous pouvons décoder et générer de nouvelles données. L'encodeur tente d'apprendre les paramètres φ pour compresser l'entrée de données x en un vecteur latent z, et la sortie de codage z est tirée d'une densité gaussienne avec des paramètres φ. En ce qui concerne le décodeur, son entrée est le codage z, la sortie de l'encodeur. Il paramétrise la reconstruction x' sur les paramètres θ, et la sortie x' est tirée de la distribution des données.



L'utilisation des auto-encodeurs variationnels (VAE) pour changer le genre d'une image de portrait nécessite l'entraînement du modèle VAE sur des images prétraitées et normalisées de personnes de sexe masculin et féminin : l'encodeur apprend à comprimer l'image d'entrée en un vecteur latent et le décodeur apprend à reconstruire l'image d'origine à partir de ce vecteur latent. Une fois que le VAE est entraîné avec les données d'images de portrait, il doit être ajusté pour effectuer le changement de genre (une réorganisation des caractéristiques apprises par le modèle pour représenter les attributs de genre spécifiques). Ensuite, il faut ajuster ou modifier le vecteur latent correspondant dans l'encodeur pour refléter les attributs du genre souhaité. Après ces étapes, le décodeur intervient pour générer une nouvelle image qui représente l'individu avec le genre modifié.

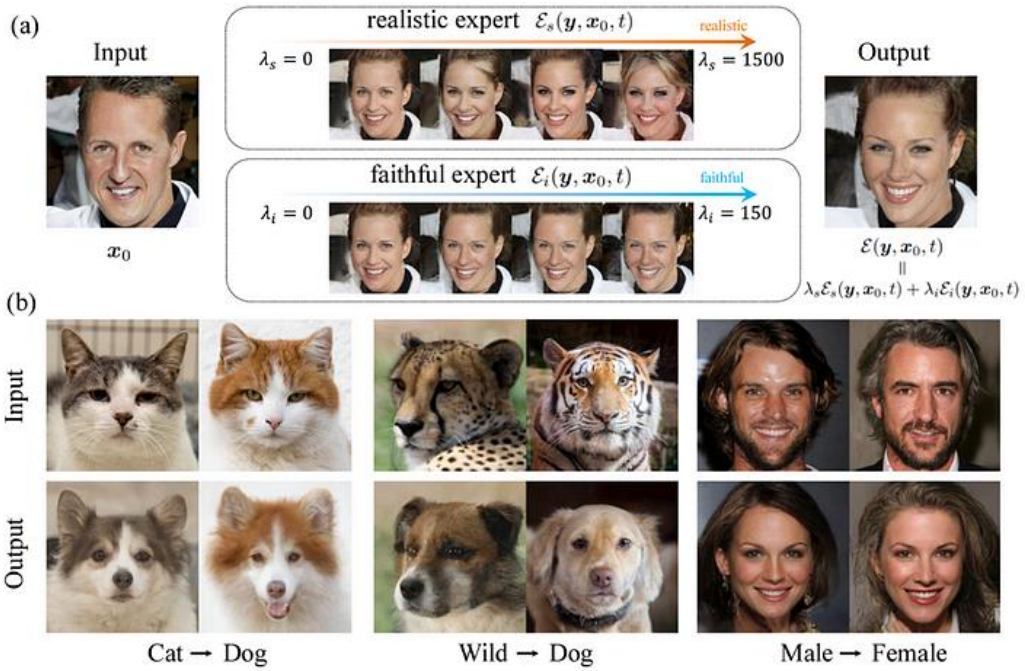
Diffusion Models (VAE)

Les modèles de diffusion sont des modèles probabilistes utilisés en traitement d'images pour la génération et la transformation d'images. Ils sont basés sur la modélisation itérative des distributions de probabilité des pixels dans une image. Ces modèles représentent chaque pixel comme une distribution de probabilité conditionnelle des valeurs de pixel précédents dans

l'image. En utilisant des itérations successives, ces modèles diffusent de l'information à travers les pixels pour générer une image pixel par pixel, en prenant en compte les pixels précédents.

Il existe des modèles de traduction d'image à image, comme Palette, basé sur des modèles de diffusion conditionnelle, et utilisé notamment dans le cadre de 4 tâches spécifiques : l'inpainting, l'uncropping, la colorisation, ainsi que la restauration JPEG.
(Papier associé : Palette: Image-to-Image Diffusion Models)

Toutefois, cette méthode ignore complètement les données d'entraînement du domaine de l'image d'entrée, rendant des solutions suboptimales. C'est pourquoi un autre modèle a proposé d'utiliser des équations différentielles stochastiques guidées par l'énergie (energy-guided stochastic differential equations ou EGSDE), utilisant une fonction d'énergie pré-entraînée sur les domaines source et cible (dans notre cas par exemple, homme-femme ou inversement) pour finalement obtenir une transformation réaliste et relativement fidèle d'image à image sans paire.
(Papier associé : EGSDE: Unpaired Image-to-Image Translation via Energy-Guided Stochastic Differential Equations)


(a)
Input
x_0
realistic expert $\mathcal{E}_s(y, x_0, t)$
realistic
$\lambda_s = 0$
$\lambda_s = 1500$
faithful expert $\mathcal{E}_i(y, x_0, t)$
faithful
$\lambda_i = 0$
$\lambda_i = 150$
Output
$\mathcal{E}(y, x_0, t)$
||
$\lambda_s\mathcal{E}_s(y, x_0, t) + \lambda_i\mathcal{E}_i(y, x_0, t)$
(b)
Input
Output
Cat → Dog
Wild → Dog
Male → Female