October 31, 2025

# Résumé général de l'article

**Titre :** Fader Networks : Manipuler des images en faisant glisser des attributs **Auteurs :** Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic Denoyer et Marc'Aurelio Ranzato (Facebook AI Research et Sorbonne Université). **Conférence :** NIPS 2017, Long Beach, Californie, États-Unis.

## Contexte et motivation

Cet article s'inscrit dans le domaine de la vision par ordinateur et plus précisément dans les modèles génératifs conditionnels. Le problème étudié est la **manipulation contrôlée d'images naturelles** : il s'agit de modifier certains attributs visuels (comme le genre, l'âge, le sourire, la présence de lunettes, etc.) tout en conservant l'identité, la structure et le réalisme de l'image.

Ce type de transformation est particulièrement difficile car le processus d'apprentissage est **non supervisé** : il n'existe pas de paires d'images montrant le même individu avec des attributs différents (par exemple la même personne avec et sans lunettes). La plupart des approches existantes reposent sur des réseaux adversariaux complexes opérant dans l'espace des pixels, ce qui rend l'entraînement instable et difficile à généraliser à plusieurs attributs simultanément.

## Objectif et contribution principale

Les auteurs proposent un nouveau modèle appelé **Fader Networks**. Son objectif est de :

- apprendre une **représentation latente désentrelacée** de l'image, où les attributs choisis peuvent être modifiés indépendamment du contenu général ;
- permettre un contrôle **continu et explicite** sur chaque attribut — à la manière d'un curseur ou *fader* dans une console audio ;

- simplifier la procédure d'entraînement, tout en conservant la qualité visuelle des images générées.

L'idée clé est d'imposer une **invariance de l'espace latent** vis-à-vis des attributs : le code latent extrait par l'encodeur ne doit contenir aucune information directe sur les attributs, afin que le décodeur soit obligé d'utiliser les valeurs d'attributs fournies en entrée pour reconstruire correctement l'image.

## Architecture proposée

Le modèle repose sur une architecture **encodeur–décodeur** accompagnée d'un **discriminateur adversarial** dans l'espace latent.

- L'encodeur $E(x)$ extrait une représentation latente $z$ d'une image $x$.
- Le décodeur $D(z, y)$ reconstruit l'image à partir du code latent et d'un vecteur d'attributs $y$.
- Le discriminateur $P(y|z)$ tente de prédire les attributs à partir du code latent.

L'encodeur est entraîné à **tromper le discriminateur** pour rendre $z$ invariant aux attributs, tout en minimisant une erreur de reconstruction :

$$L(\theta_{enc}, \theta_{dec}|\theta_{dis}) = \|D(E(x), y) - x\|_2^2 - \lambda_E \log P(1 - y|E(x)).$$

Le paramètre $\lambda_E$ contrôle le compromis entre fidélité et invariance.

Ainsi, à l'inférence, l'utilisateur peut faire varier librement les valeurs des attributs $y$ pour générer différentes versions d'une même image.

## Implémentation

Les Fader Networks sont implémentés avec des couches convolutionnelles $4 \times 4$ (stride 2, padding 1) et des *leaky-ReLU*. L'encodeur comporte 7 couches de convolutions et le décodeur est symétrique, utilisant des convolutions transposées pour le sur-échantillonnage. Les attributs sont injectés sous forme de canaux additionnels à chaque niveau du décodeur. L'entraînement est effectué avec Adam ($\eta = 0.002$, $\beta_1 = 0.5$) et un schéma de montée progressive du coefficient $\lambda_E$ jusqu'à 0.0001 pour stabiliser la convergence.

## Expérimentations et résultats

Les auteurs évaluent leur modèle sur deux jeux de données :

- **CelebA** : plus de 200 000 images de visages annotées par 40 attributs ;

- **Oxford-102 Flowers** : 9 000 images de fleurs réparties en 102 classes.

**Résultats sur CelebA :**

- Les Fader Networks parviennent à **modifier de manière réaliste** des attributs tels que le genre, le sourire, la présence de lunettes, etc.
- L'évaluation par des humains (Mechanical Turk) montre que les images générées obtiennent des scores de naturalité de plus de 80 % pour certains attributs, surpassant largement les modèles IcGAN.
- Les échanges (*swaps*) d'attributs produisent des effets clairs, tout en maintenant l'identité du visage.
- Le modèle permet également la **modification simultanée de plusieurs attributs** (par exemple : genre + sourire + ouverture des yeux).

**Résultats sur Oxford-102 Flowers :**

- Le modèle réussit à modifier la couleur des fleurs selon l'attribut choisi (ex. : degré de  rose ) tout en conservant le fond et la forme de la fleur.
- Ces expériences démontrent la **généralisation du principe d'invariance latente** à d'autres types d'images que les visages.

## Analyse et conclusion

Les Fader Networks montrent qu'il est possible d'obtenir une **manipulation fine et contrôlable d'images** à l'aide d'un entraînement adversarial dans l'espace latent, sans recourir à un GAN en sortie. Cette approche :

- simplifie le processus d'apprentissage ;
- permet de contrôler continûment les attributs générés ;
- produit des images réalistes, naturelles et cohérentes ;
- se généralise facilement à plusieurs attributs ou domaines.

Les auteurs concluent que ce type de modèle constitue une **alternative prometteuse aux GANs classiques**, notamment pour les applications d'édition automatique d'images et de génération contrôlée. Ils suggèrent que cette approche pourrait être étendue à d'autres domaines comme la parole ou le texte, où la génération est difficilement différentiable.

**En résumé :** Les Fader Networks représentent une avancée majeure vers des modèles génératifs explicables et contrôlables, offrant un compromis élégant entre simplicité d'entraînement, qualité visuelle et contrôle des attributs.