école
normale
supérieure
paris—saclay

Image Inpainting using the Latent space of StyleGAN2

DELIRES

Mathis Wauquiez

M2 Mathématiques Vision Apprentissage

March 26, 2025

école normale supérieure paris-saclay

Objectives

Figure 1: **Example of image inpainting.**

- Use StyleGAN for image inpainting
- Analyze the robustness of the inpainting algorithm
- Add semantic constraints to the inpainting

Quick reminder: StyleGAN/StyleGAN2 Generators

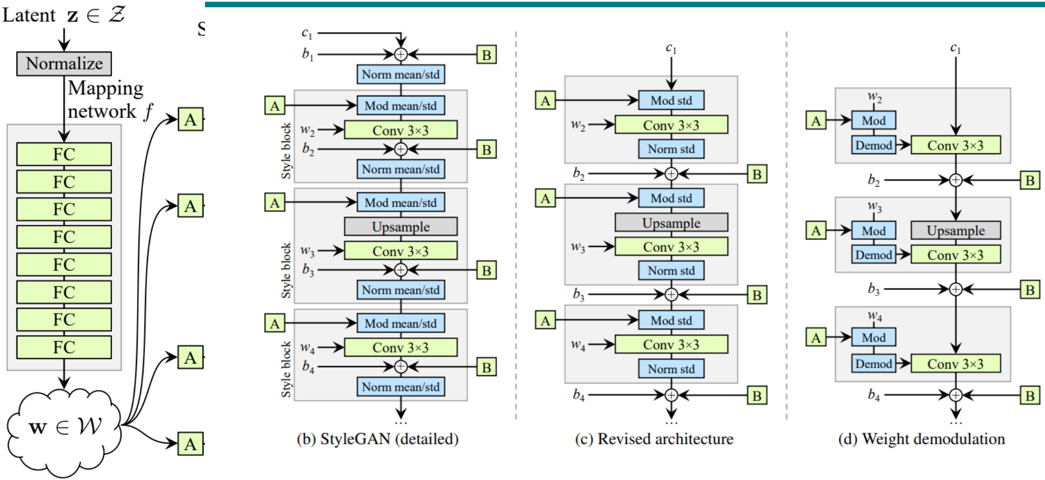

Figure: StyleGAN2 Generator Architecture

Problem Statement - Inversion

Objective: Solve

$$w^* = \arg\min_{w \in \mathcal{W}^+} \|(x - G(w)) \odot M\|_2^2$$

Method:

- `Adam` optimizer
- Initialize $w = w_{\text{avg}}$
- With an `ExponentialLR` scheduler

Naive inversion - Results

$$w^* = \arg\min_{w \in \mathcal{W}^+} \|(x - G(w)) \odot M\|_2^2$$

LPIPS

- LPIPS: Learned Perceptual Image Patch Similarity
- Similarity measure, using deep features
- More aligned with human perception than pixel-wise metrics
- Captures higher-level perceptual differences than MSE

$$\text{LPIPS}(x, y) = \sum_l w_l \| f_l(x) - f_l(y) \|_2^2$$

Problem Statement - Inversion

Objective: Solve

$$
\begin{aligned}
w^* &= \arg\min_{w \in \mathcal{W}^+} \lambda_1 \mathrm{MSE}\big(x \odot M, y \odot M\big) + \lambda_2 \mathrm{LPIPS}\big(x \odot M, y \odot M\big) \\
&= \arg\min_{w \in \mathcal{W}^+} \lambda_1 \mathrm{MSE}_M\big(x, y\big) + \lambda_2 \mathrm{LPIPS}_M\big(x, y\big)
\end{aligned}
$$

école normale supérieure paris-saclay

MSE + LPIPS - Results

$$w^* = \arg\min_{w \in \mathcal{W}^+} \lambda_1 \mathrm{MSE}_M(x, y) + \lambda_2 \mathrm{LPIPS}_M(x, y)$$

Figure: Classifier constrained generations

Why does the generations vary?

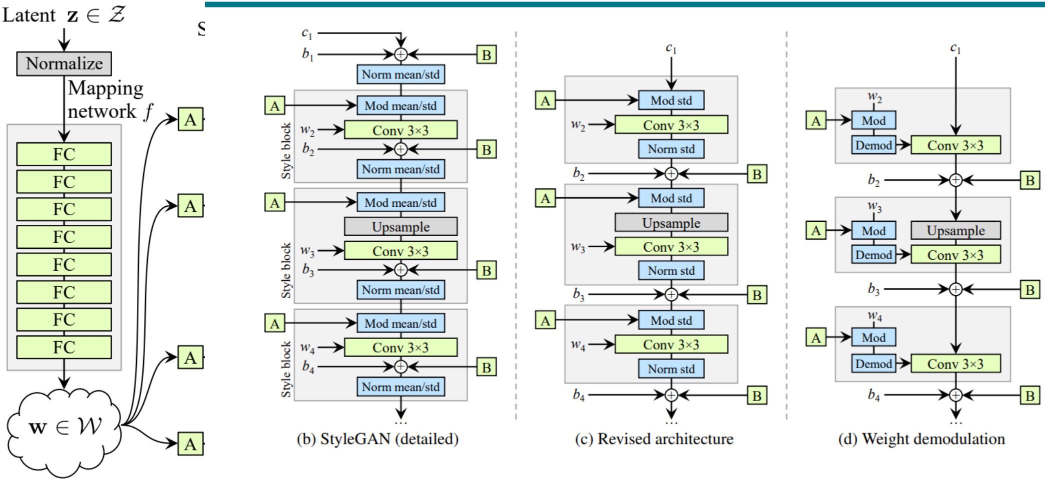

Figure: StyleGAN2 Generator Architecture

Adding semanticity constraints - Classifier

Let f_θ be our classifier, used to control one attribute. We use a Vit-B/16 classifier, pretrained on the CelebA attributes dataset.

$$w^* = \arg\min_{w \in \mathcal{W}^+} \lambda_1 \text{MSE}_M(x, y) + \lambda_2 \text{LPIPS}_M(x, y) + \lambda_3 \text{BCE}(f_\theta(x), y)$$

Small improvementt:

- ϵ-margin formulation: $\mathcal{L}_{\text{class}} = \max(\text{BCE}(f_\theta(x), y), \epsilon)$
- `LogSumExp` relaxation: $\mathcal{L}_{\text{class}} = \log(\exp(f_\theta(x), y) + exp(\epsilon))$

Classifier results

Figure: **Modified attribute:** Big Lips

Figure: **Modified attribute:** No smile

Figure: Classifier constrained generations

Semanticity constraint - CLIP

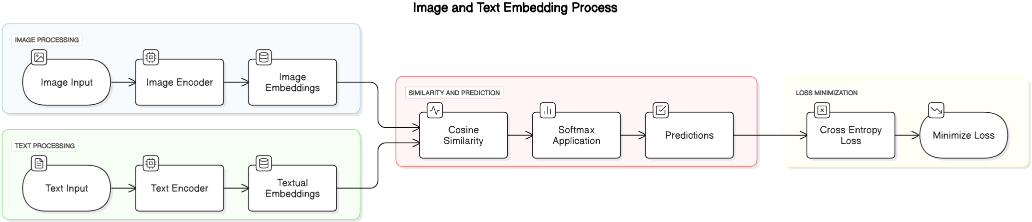

Figure: CLIP methodology

Key point: CLIP enables us to encode both textual descriptions and images into vector embeddings. We can then measure the similarity between these embeddings using cosine similarity.

Adding semanticity constraints

We used the CLIP/Vit-B/16 model as the image encoder. The objective function is defined as follows, where C_ϕ represents our image encoder and c denotes the textual embeddings:

$$w^* = \arg\min_{w \in \mathcal{W}^+} \left(\lambda_1 \mathrm{MSE}_M(x, y) + \lambda_2 \mathrm{LPIPS}_M(x, y) - \lambda_3 \text{cosine similarity}(C_\phi(x), c)\right)$$

CLIP Results

$$w^* = \arg\min_{w \in \mathcal{W}^+} \left(\lambda_1 \mathrm{MSE}_M(x, y) + \lambda_2 \mathrm{LPIPS}_M(x, y) - \lambda_3 \text{cosine similarity}(C_\phi(x), c) \right)$$

Figure: **Prompt:** A woman with purple lipstick

Figure: **Prompt:** A woman with a serious expression, not smiling

Conclusion

- Naive inversion is not sufficient
- Our generations are stochastic
- It is easy to guide the generation if we have a classifier
- Diversity of the generations
- Some masks are more difficult than others
- However, they are truly rare, and CLIP is more generic
- $\approx$ 20min/image on a GTX 1050 Ti
- L2 regularization or discriminator-based losses do not noticibly improve the results