SAPIENZA
UNIVERSITÀ DI ROMA

StyleGAN encoder for image-to-image translation

Vision and Perception project

September, 2021

(Amila Sikalo - 1938032) & (Olga Sorokoletova - 1937430)

Project Goal and Overview

- Given: **Variety of image-to-image translation tasks**

- Goal: **Generic framework - pixel2style2pixel (pSp)**
  - **Novelty**: ENCODER that directly generates style vectors in $\mathcal{W}^+$
  - New methodology for utilizing pre-trained StyleGAN generator.

Advantages over previous algorithms

SAPIENZA UNIVERSITÀ DI ROMA

- No additional optimization over latent space;
- Domain-independent;
- Supports multi-modal synthesis;
- Support for tasks without p2p correspondence;
- No adversary required.

StyleGAN

- STYLE-based generator architecture (operates **globally** instead of locally);
- State-of-the-art visual quality on the high resolution images;
- Disentangled LATENT SPACE $\mathcal{W}$.

01	$\mathcal{W}$	• Meaningful manipulations can be performed • "Invert first, edit later" approach • 512-dimensional vector w
02	$\mathcal{W}^+$	• More accurate reconstruction • Per-image optimization • 18 different 512-dimensional vectors w
03	Encoder	• Accelerates optimization process • Infer an approximate vector • Additional optimization is required

Trade-off: FAST vs ACCURATE inversion (?)

New encoder

- Idea: **Feature Pyramid Network**

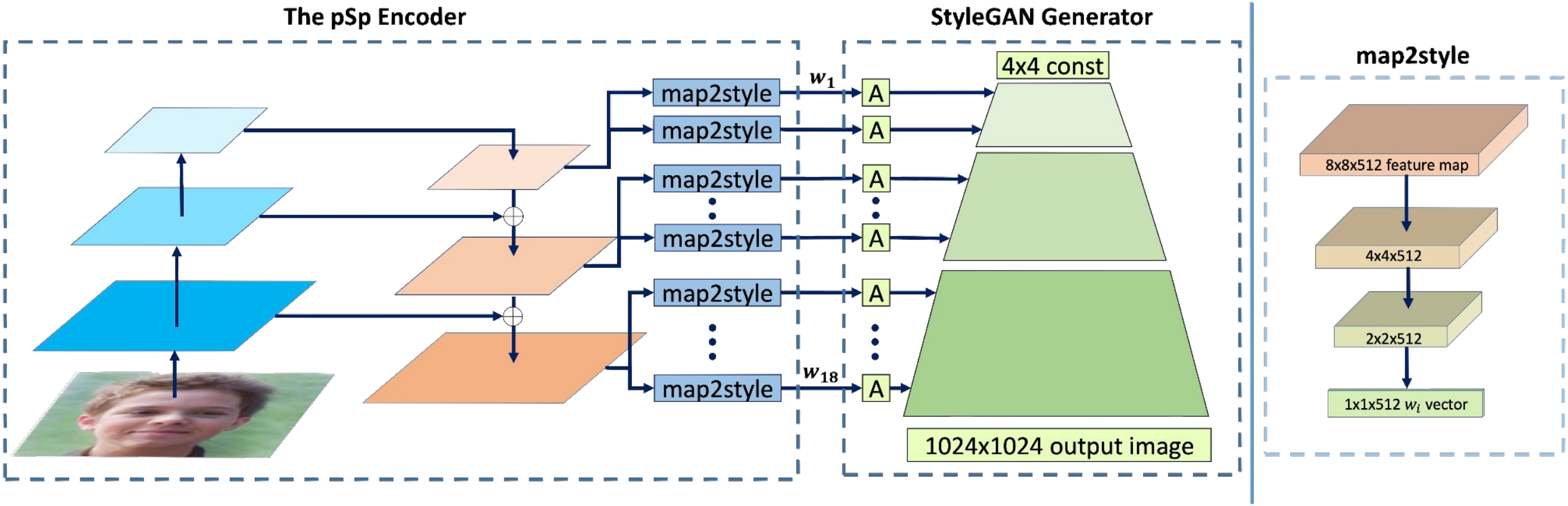
  Style vectors are extracted from different pyramid scales and inserted in correspondence to their spatial scales.

- **Motivation**: Different style inputs correspond to different levels of detail, which are roughly divided into three groups — coarse, medium and fine.
- **Result:** Latent space manipulations without requiring time-consuming optimization.

- Math part: $\mathrm{pSp}(x) \coloneqq G(E(x)+\mathrm{avg}(w))$

  Encoder aims to learn the latent code with respect to the average style vector of the pre-trained generator.

Architecture

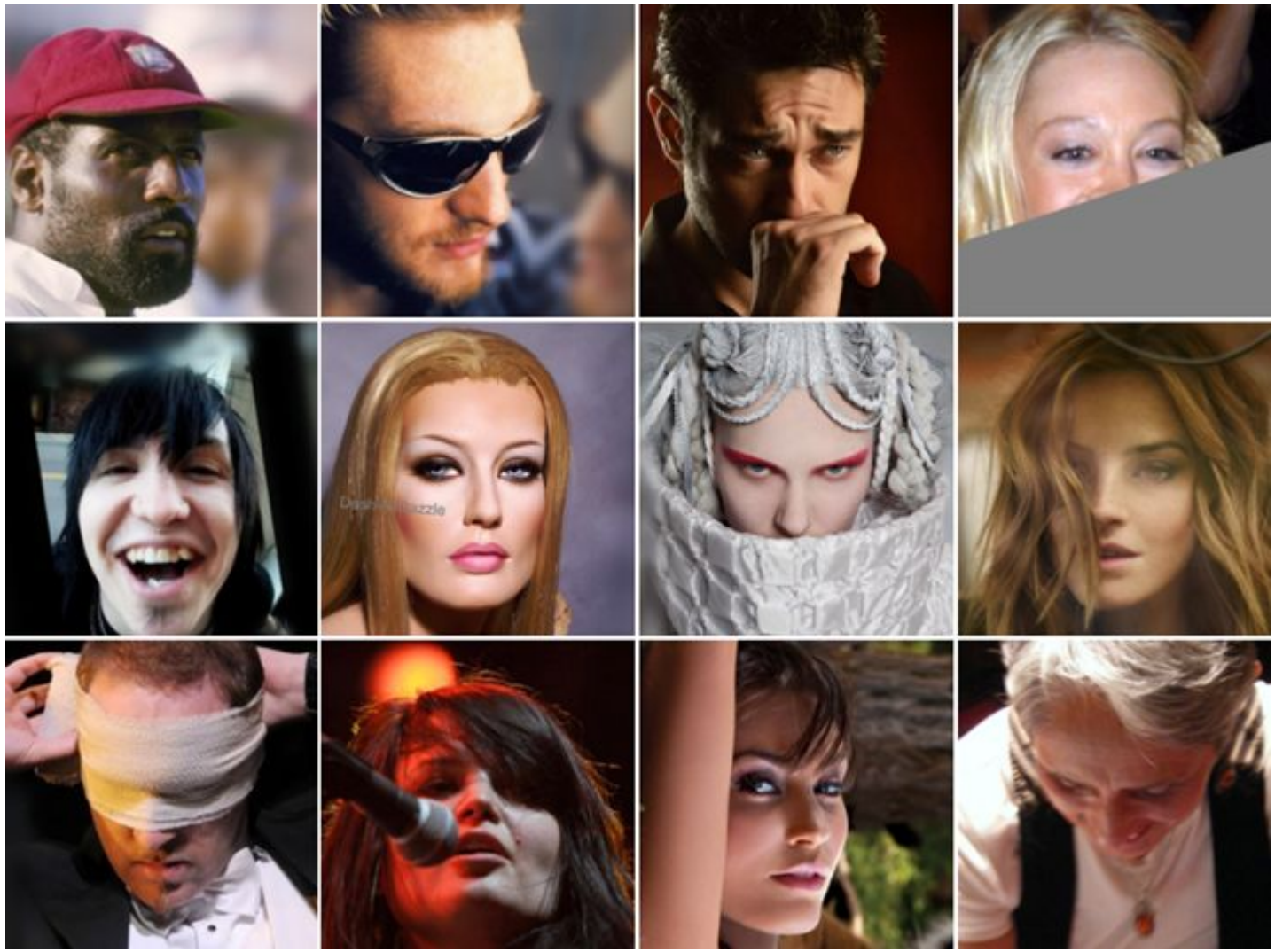


GAN Inversion

We use pSp to find the latent code of real images in the latent space of a pre-trained StyleGAN generator.

CelebA Dataset

- 202 599 celebrity images, each with **40** attribute annotations.
- The images cover large pose variations and background clutter.
- 10 177 number of identities
- 5 landmark locations
- Resizing
- Rescaling



Results

StyleGAN results

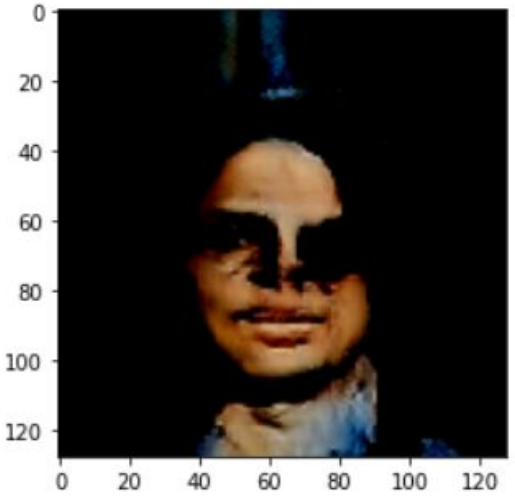
Model resolution: 4 x 4

Model resolution: 8 x 8

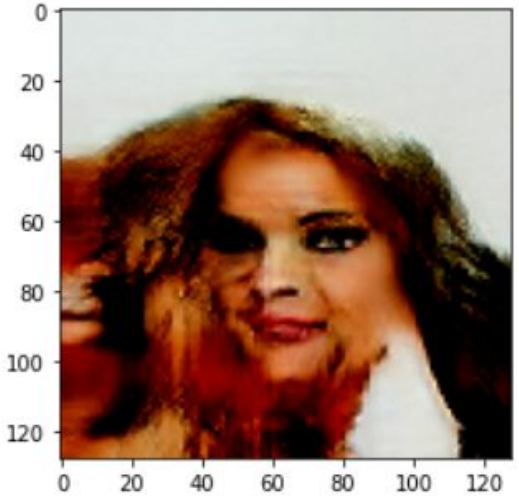
Model resolution: 16 x 16

Model resolution: 128 x 128

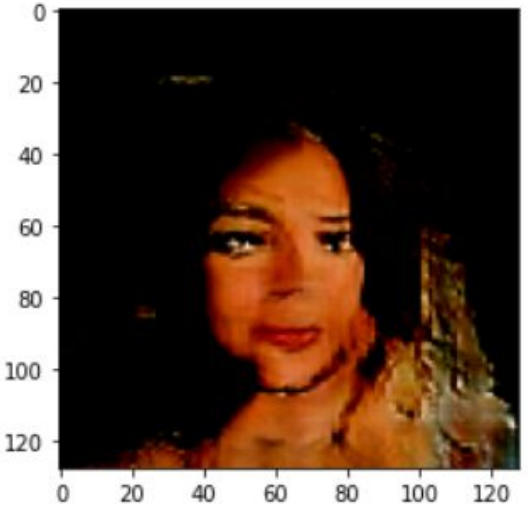
pSp with Encoder



Epochs: 10

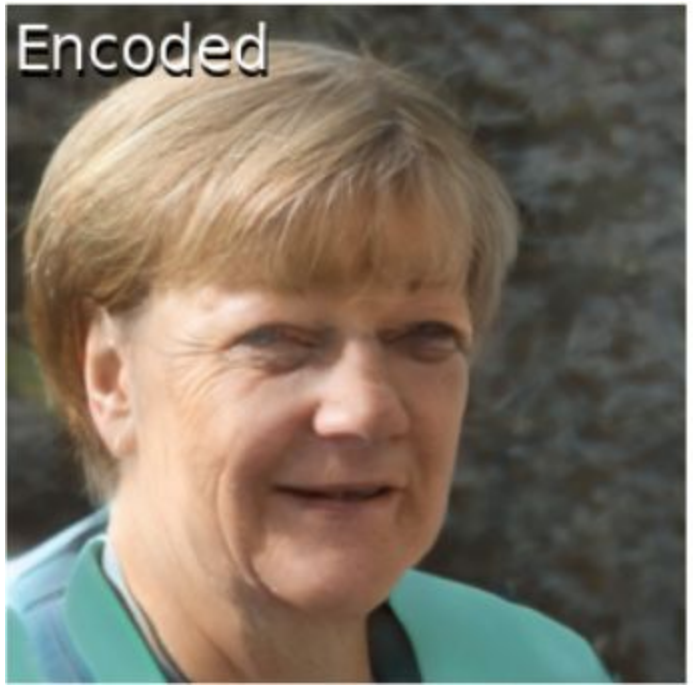


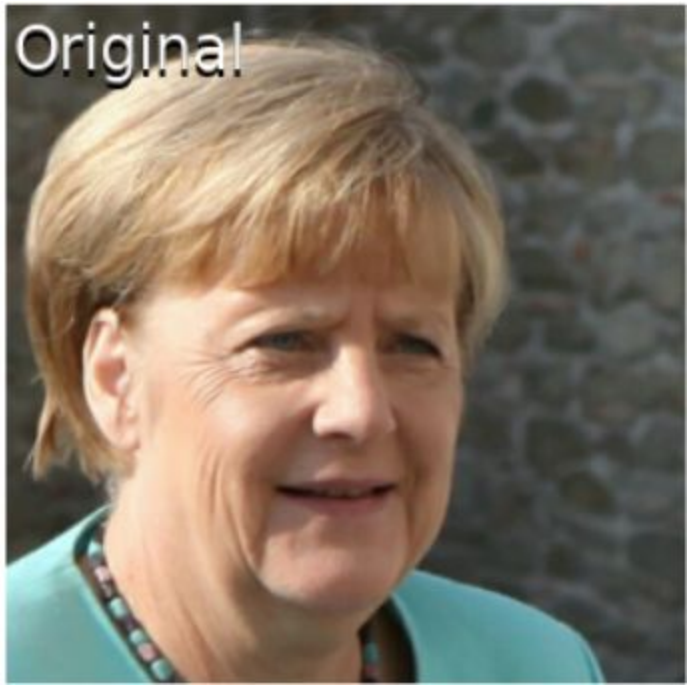
Epochs: 20



Epochs: 40

Encoder





Wrap-Up and Conclusions

- APPLICABILITY
  - PsP can be used to **directly encode translation tasks** into StyleGAN, thereby supporting input images that do not reside in the StyleGAN domain (**out-of-domain support**);
  - Capable of solving a **wide variety of image-to-image translation tasks** (e.g face frontalization, conditional image synthesis, ect), **requiring only minimal changes** (training losses and methodology);
  - Generates the **high-quality images**;
- EXTENSIONS
  - Going **beyond the facial domain**;
- LIMITATIONS
  - Method is limited to images that can be generated by **StyleGAN**;
  - Globality of approach introduces a challenge in **preserving finer details** of the input image e.g. earrings, background details);