StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery

21. 04.29 신한이

목차

0 Abstract

- StyleGAN의 다양한 도메인에서 사실적인 이미지를 생성하는 능력에 영감을 받아, 최근 연구는 StyleGAN의 잠재 공간(latent spaces)를 사용하는 방법을 이해하는 데에 초점을 맞추고 있다.

- 그러나, 의미 있는 latent manipulations를 발견하는 것은 인간의 노동이 필요하거나 주석이 달린 이미지를 필요로 한다.

- 본 연구에서는 최근에 소개된 CLIP 모델과 StyleGAN을 더해주어 소모적인 수동 작업이 필요하지 않다.

-> **CLIP**: 사용자가 제공한 텍스트에 대응하여 입력 잠재 벡터(input latent vector)를 수정하는 최적화 체계를 소개.

-> **StyleGAN**: 대화식 텍스트를 기반으로 이미지 조작(manipulation)을 가능하게 함.

0 Abstract

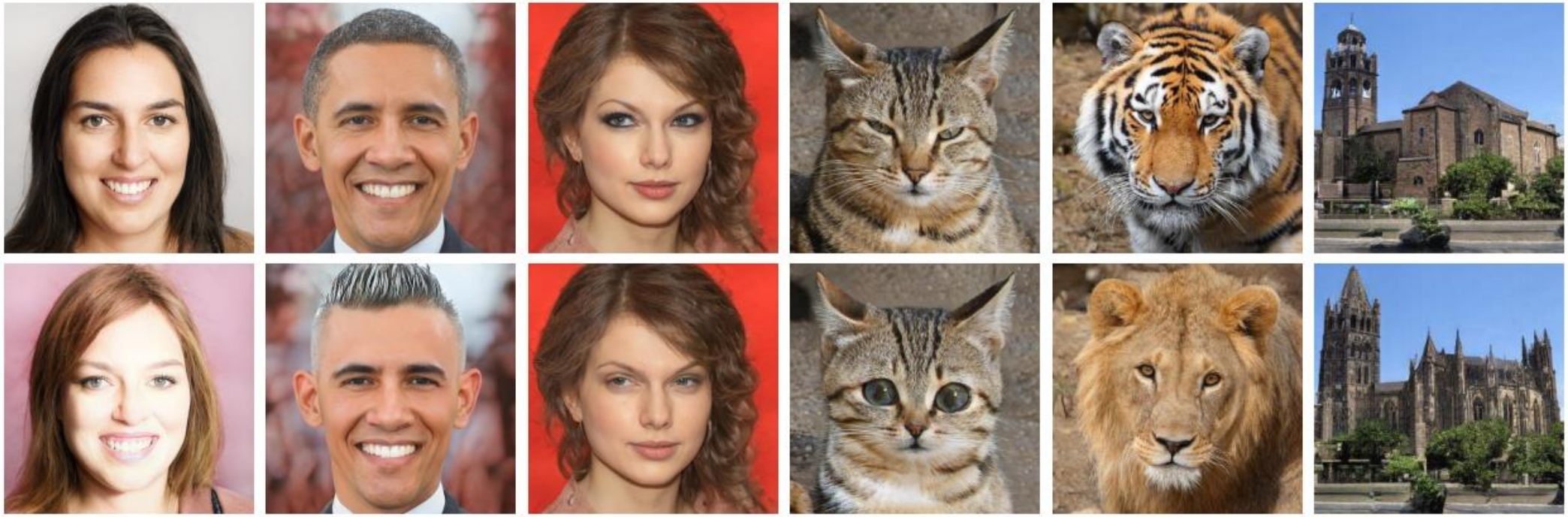


Figure 1. Examples of text-driven manipulations using StyleCLIP. Top row: input images; Bottom row: our manipulated results. The text prompt used to drive each manipulation appears under each column.

1 Introduction

2019 **StyleGAN**
A Stule-Based Generator Architecture for Generative Adversarial Networks

CLIP 2021. Feb
Learning Transferable Visual Models From Natural Language Supervision

2021. Mar **StyleCLIP**
StyleCLIP: Text-Driven Manipulations of StyleGAN Imagery

1 Introduction

StyleGAN

- 고해상도의 이미지 생성에 효과적.
- Disentanglement
  of semantic feactures
- StyleGAN의 Latent Vector
  : Coarse, Middle, Fine styles
- Image Manipulation Step
  : 1. Encoding step (Gradient Descent)
    2. Manipulation step (Learned Direction)

CLIP

- Image encoder 와 Text encode를 jointly 하게 학습시킴.
- 어떠한 image와 text에 대해 유사한 semantic 특징을 찾아냄.

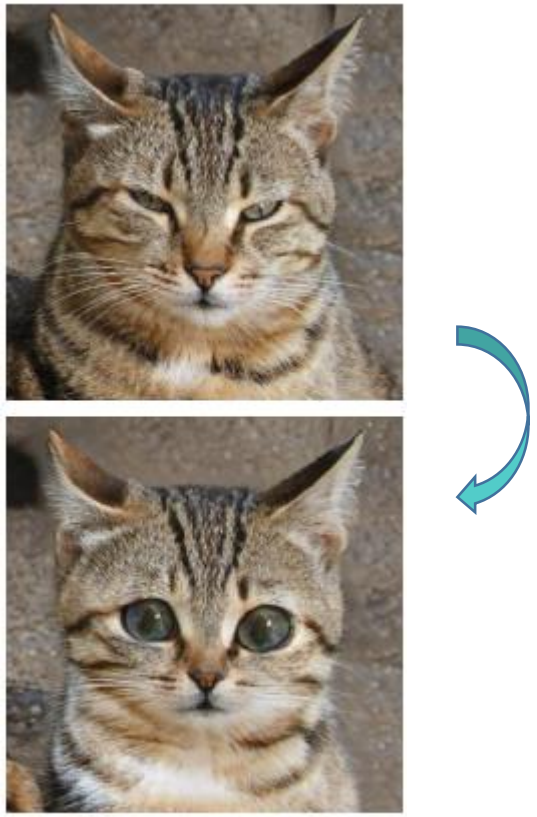
1 Introduction

StyleCLIP = StyleGAN + CLIP (main idea)

- StyleGAN에서 이미지를 manipulate 할 때, 직관적으로 text로 설명

ex)

Input image : "고양이 사진" + text : "Cute"
=> "Cute 고양이 사진" 생성

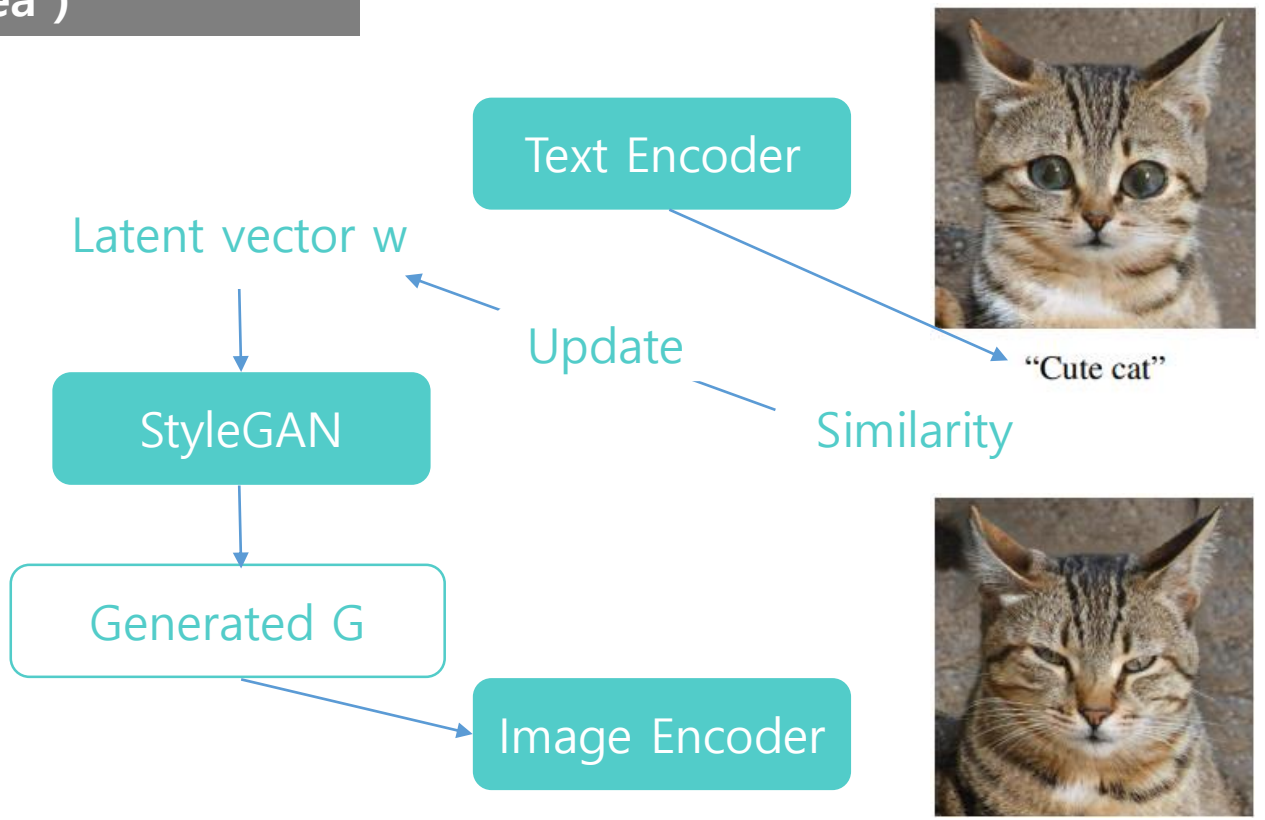


"Cute cat"

1 Introduction

StyleCLIP = StyleGAN + CLIP (main idea)

- StyleGAN과 CLIP을 합치기 위해 세가지 방법(methods)을 제안.
  1. Optimization method
  2. Latent mapper
  3. Global directions



3 StyleCIIP Text-Driven Manipulation

	pre-proc.	train time	infer. time	input image dependent	latent space
optimizer	–	–	98 sec	yes	$\mathcal{W}+$
mapper	–	10 – 12h	75 ms	yes	$\mathcal{W}+$
global dir.	4h	–	72 ms	no	$\mathcal{S}$

Table 1. Our three methods for combining StyleGAN and CLIP. The latent step inferred by the optimizer and the mapper depends on the input image, but the training is only done once per text prompt. The global direction method requires a one-time pre-processing, after which it may be applied to different (image, text prompt) pairs. Times are for a single NVIDIA GTX 1080Ti GPU.

4 Latent Optimization Method 1

Latent Optimization : A simple approach for leveraging CLIP to guide image manipulation

$$\arg\min_{w \in \mathcal{W}+} D_{\text{CLIP}}(G(w), t) + \lambda_{\text{L2}} \|w - w_s\|_2 + \lambda_{\text{ID}} \mathcal{L}_{\text{ID}}(w), \tag{1}$$

$$\mathcal{L}_{\text{ID}}(w) = 1 - \langle R(G(w_s)), R(G(w)) \rangle, \tag{2}$$

(1)
- (Clip loss)+(특정한 이미지를 임베딩)
- Latent vector update

(2)
- G : A pretrained StyleGAN generator
- D_{CLIP} is the cosine distance

=> 일반적으로 200-300 iterations 요구 , several minutes 소요

5 Latent Mapper Method 2

세 가지의 개별 네트워크: **C**oarse, **M**edium, **F**ine

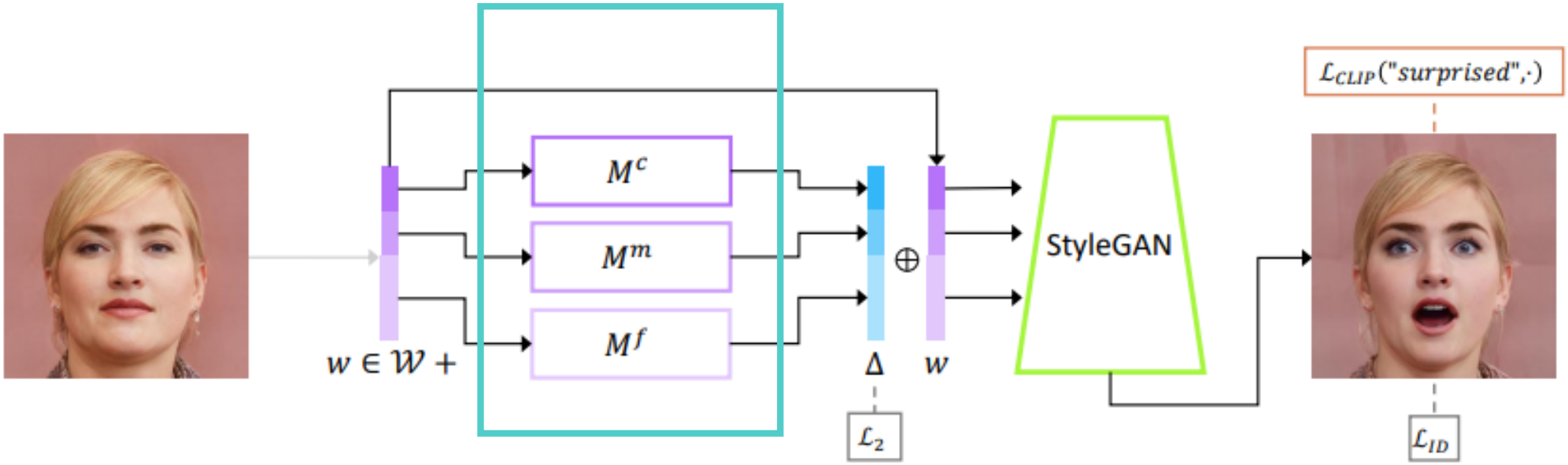
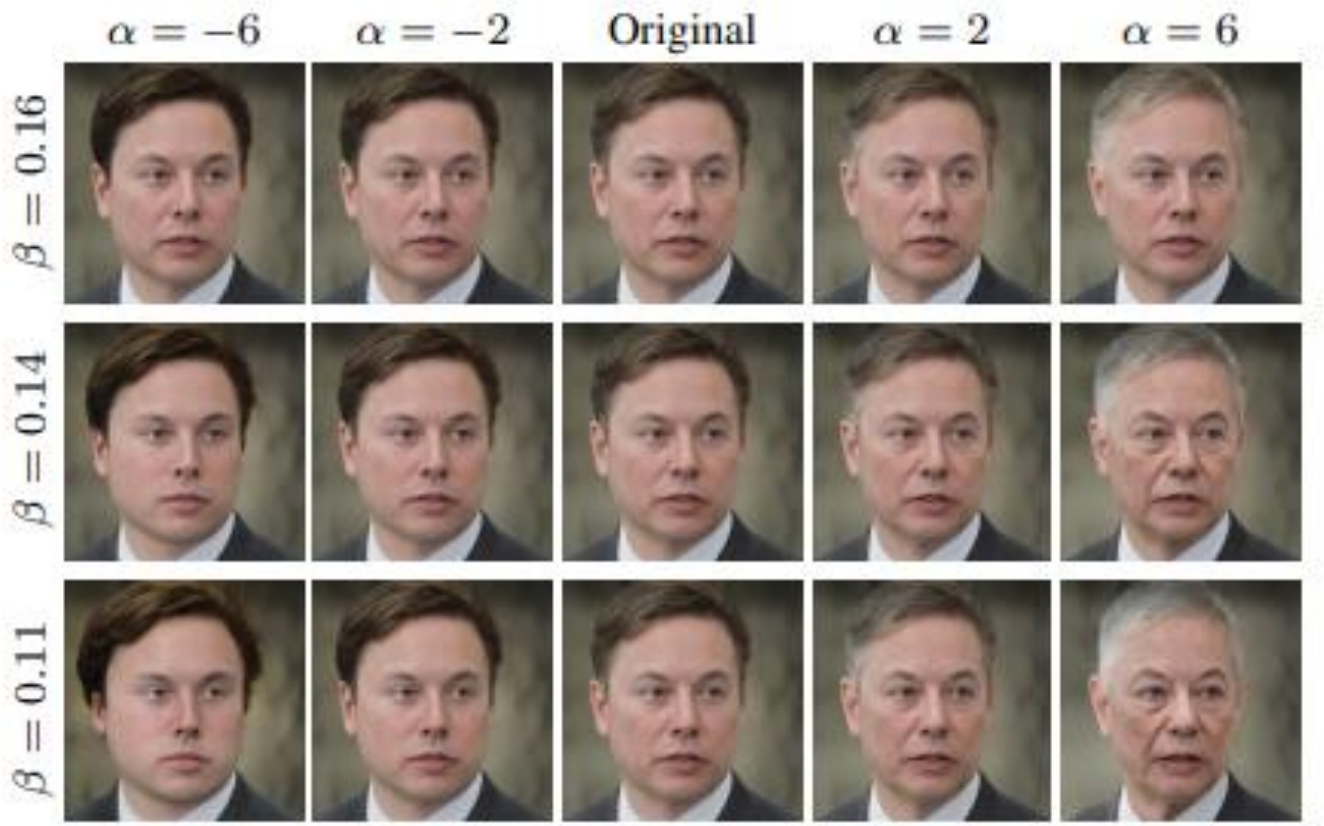


Figure 2. The architecture of our text-guided mapper (using the text prompt "surprised", in this example). The source image (left) is inverted into a latent code w. Three separate mapping functions are trained to generate residuals (in blue) that are added to w to yield the target code, from which a pretrained StyleGAN (in green) generates an image (right), assessed by the CLIP and identity losses.

=> 인코더 네트워크 학습은 several hours 소요

6 Global Directions Method 3

Global Directions : 특정한 이미지 input에 대해, 방향만 적용하면 쉽게 이미지에 적용 가능.



- Text prompt를 global direction 으로 **매핑(mapping)** 하는 방법

- 하나의 텍스트에 대해 **global direction vector**을 찾은 뒤, **어떠한 이미지라도 input**으로 사용할 수 있다.

참조 자료

- Code and video are available on
  https://github.com/orpatashnik/StyleCLIP

- Paper
  https://arxiv.org/pdf/2103.17249.pdf

- Refer
  https://www.youtube.com/watch?v=hFC7DSh9RIw

- Refer
  http://aidev.co.kr/deeplearning/10338