StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery

(Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., & Lischinski, D. (2021))

RAMI LAB.

Sang-Hyun KIM

Contents

- **Background**
  - **StyleGAN2**
  - **CLIP**
- **Introduction**
- **StyleCLIP Text-Driven Manipulation**
  - **Latent Optimization**
  - **Latent Mapper**
  - **Global Directions**
- **Comparisons and Evaluation**
- **Conclusions**
- **Q & A**

Background

- **StyleGAN2**
  - **StyleGAN을 기반**

Configuration	FFHQ, 1024×1024				LSUN Car, 512×384			
	FID ↓	Path length ↓	Precision ↑	Recall ↑	FID ↓	Path length ↓	Precision ↑	Recall ↑
A Baseline StyleGAN [24]	4.40	212.1	**0.721**	0.399	3.27	1484.5	**0.701**	0.435
B + Weight demodulation	4.39	175.4	0.702	0.425	3.04	862.4	0.685	0.488
C + Lazy regularization	4.38	158.0	0.719	0.427	2.83	981.6	0.688	0.493
D + Path length regularization	4.34	**122.5**	0.715	0.418	3.43	651.2	0.697	0.452
E + No growing, new G & D arch.	3.31	124.5	0.705	0.449	3.19	471.2	0.690	0.454
F + Large networks (StyleGAN2)	**2.84**	145.0	0.689	**0.492**	**2.32**	**415.5**	0.678	**0.514**
Config A with large networks	3.98	199.2	0.716	0.422	–	–	–	–

Table 1. Main results. For each training run, we selected the training snapshot with the lowest FID. We computed each metric 10 times with different random seeds and report their average. *Path length* corresponds to the PPL metric, computed based on path endpoints in $\mathcal{W}$ [24], without the central crop used by Karras et al. [24]. The FFHQ dataset contains 70k images, and the discriminator saw 25M images during training. For LSUN CAR the numbers were 893k and 57M. ↑ indicates that higher is better, and ↓ that lower is better.

Background

- **StyleGAN2 (Droplet Artifact)**

  - **최종 결과물에 물방울 무늬 같은 noise가 반복해서 나타나는 현상.**

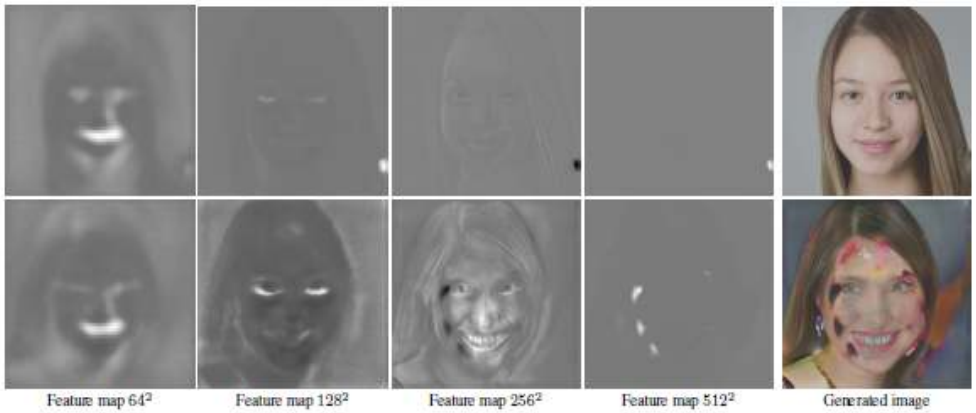
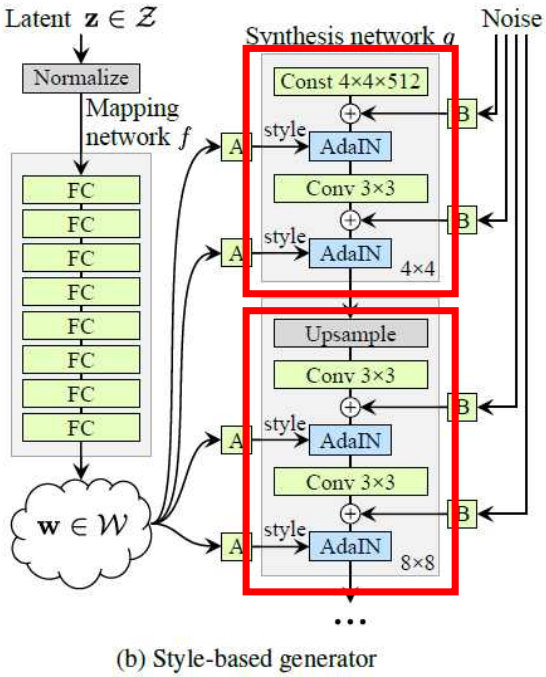
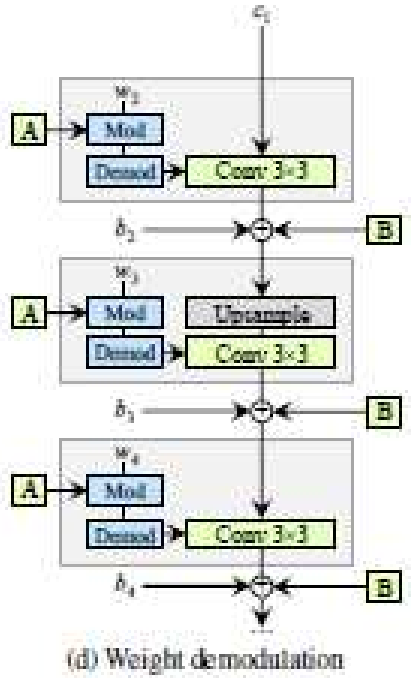


Figure 15. An example of the importance of the droplet artifact in StyleGAN generator. We compare two generated images, one successful and one severely corrupted. The corresponding feature maps were normalized to the viewable dynamic range using instance normalization. For the top image, the droplet artifact starts forming in 64^2 resolution, is clearly visible in 128^2, and increasingly dominates the feature maps in higher resolutions. For the bottom image, 64^2 is qualitatively similar to the top row, but the droplet does not materialize in 128^2. Consequently, the facial features are stronger in the normalized feature map. This leads to an overshoot in 256^2, followed by multiple spurious droplets forming in subsequent resolutions. Based on our experience, it is rare that the droplet is missing from StyleGAN images, and indeed the generator fully relies on its existence.



(b) Style-based generator



(d) Weight demodulation

Background

- **StyleGAN2 (Phase artifact)**

  - **이빨이나 눈동자 같이 detail 한 부분에서 disentanglment 가 자연스럽지 않은 현상.**

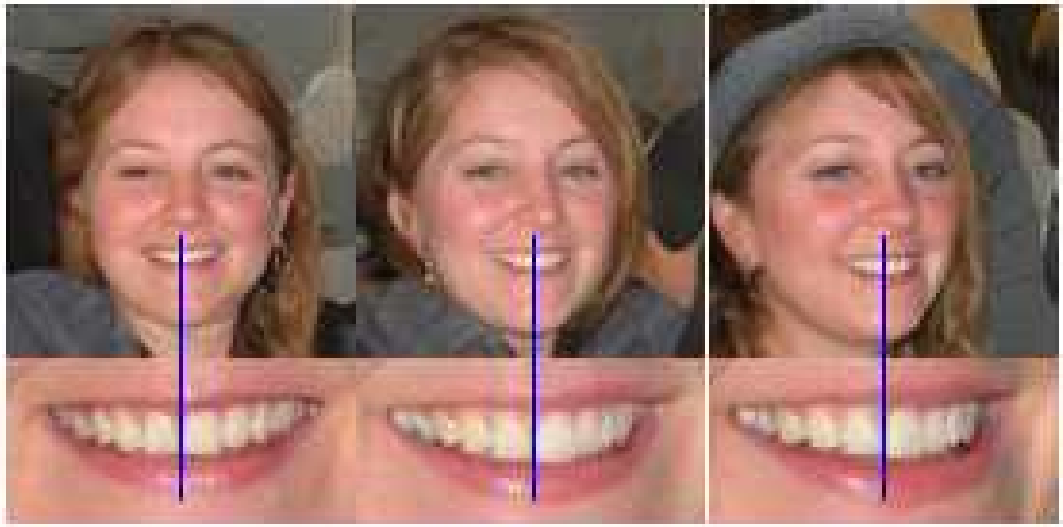
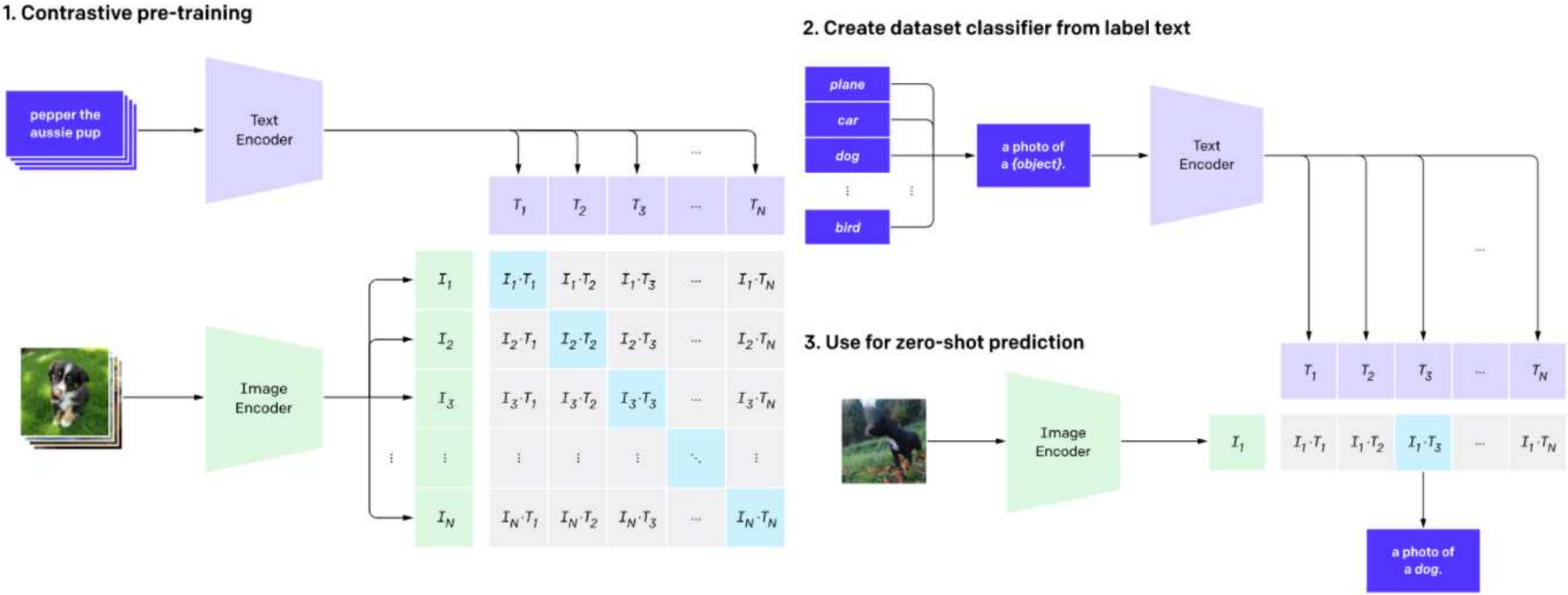


Figure 6. Progressive growing leads to "phase" artifacts. In this example the teeth do not follow the pose but stay aligned to the camera, as indicated by the blue line.

Background

- **CLIP (Contrastive Language-Image Pre-training)**



Introduction

- **StyleGAN의 disentangl한 latent space를 사용해서 image manipulation.**

- **Problem**
  - **Latent vector를 찾기 위한 수동적인 방법.**

- **Solution: StyleCLIP**

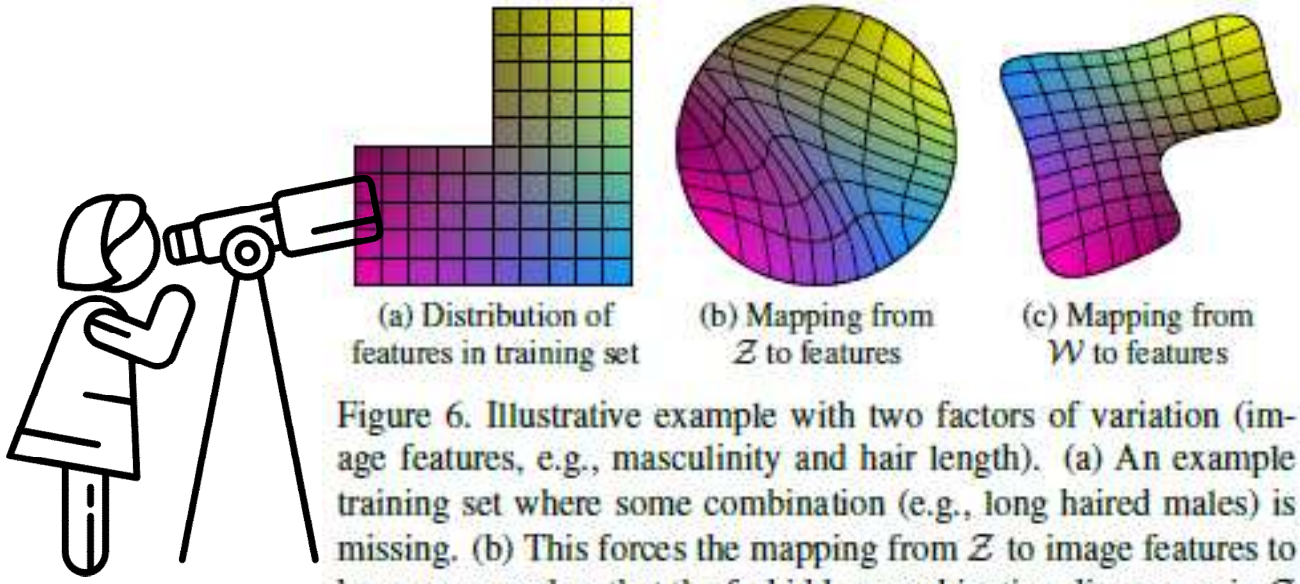


Figure 6. Illustrative example with two factors of variation (image features, e.g., masculinity and hair length). (a) An example training set where some combination (e.g., long haired males) is missing. (b) This forces the mapping from $\mathcal{Z}$ to image features to become curved so that the forbidden combination disappears in $\mathcal{Z}$ to prevent the sampling of invalid combinations. (c) The learned mapping from $\mathcal{Z}$ to $\mathcal{W}$ is able to "undo" much of the warping.

그림 출처: https://www.pngwing.com/ko/free-png-ptovd

Introduction

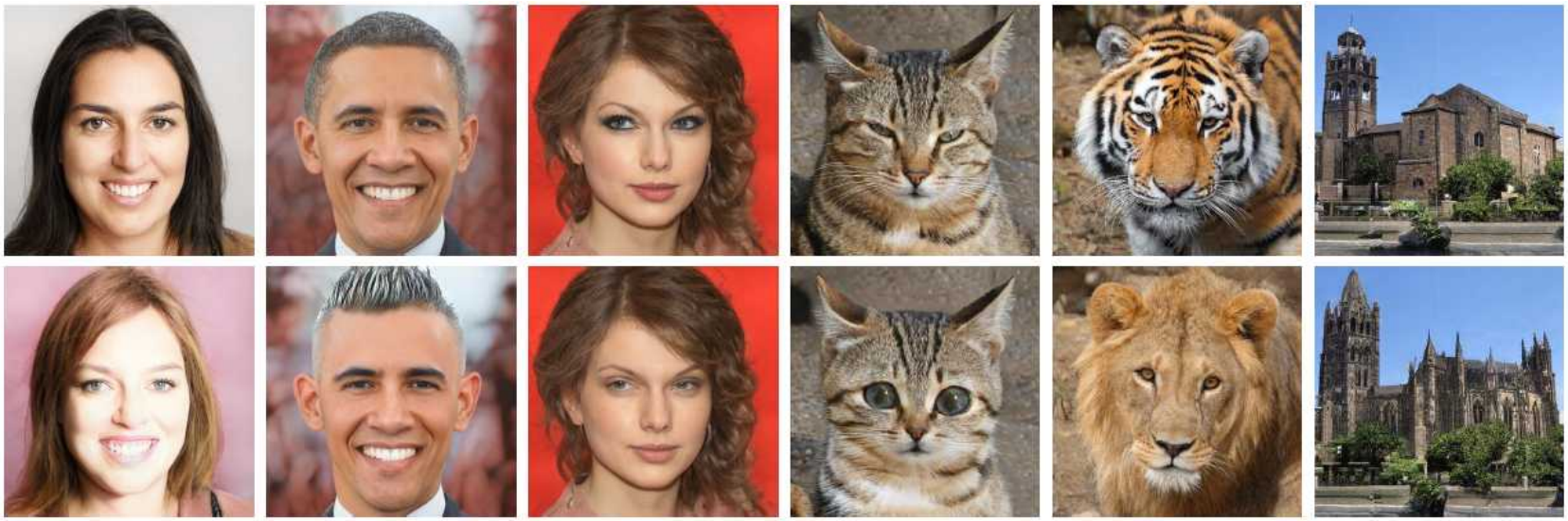


Figure 1. Examples of text-driven manipulations using StyleCLIP. Top row: input images; Bottom row: our manipulated results. The text prompt used to drive each manipulation appears under each column.

StyleCLIP Text-Driven Manipulation

- **In this paper, they investigate three techniques that combine CLIP with StyleGAN.**
  - **Text-guided latent optimization, where a CLIP model is used as a loss network. This is the most versatile approach, but it requires a few minutes of optimization to apply a manipulation to an image.**
  - **A latent residual mapper, trained for a specific text prompt. Given a starting point in latent space (the input image to be manipulated), the mapper yields a local step in latent space.**
  - **A method for mapping a text prompt into an input-agnostic (global) direction in StyleGAN's style space, providing control over the manipulation strength as well as the degree of disentanglement.**

	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.

Method

- **Latent Optimization**

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

- $G(w)$: StyleGAN2 Generator
- D_{CLIP} : The cosine distance between the CLIP embeddings of its two arguments

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

- R: Pretrained ArcFace network for face recognition
- <•, •>: Computes the cosine similarity between it's arguments.

	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	$\mathcal{S}$

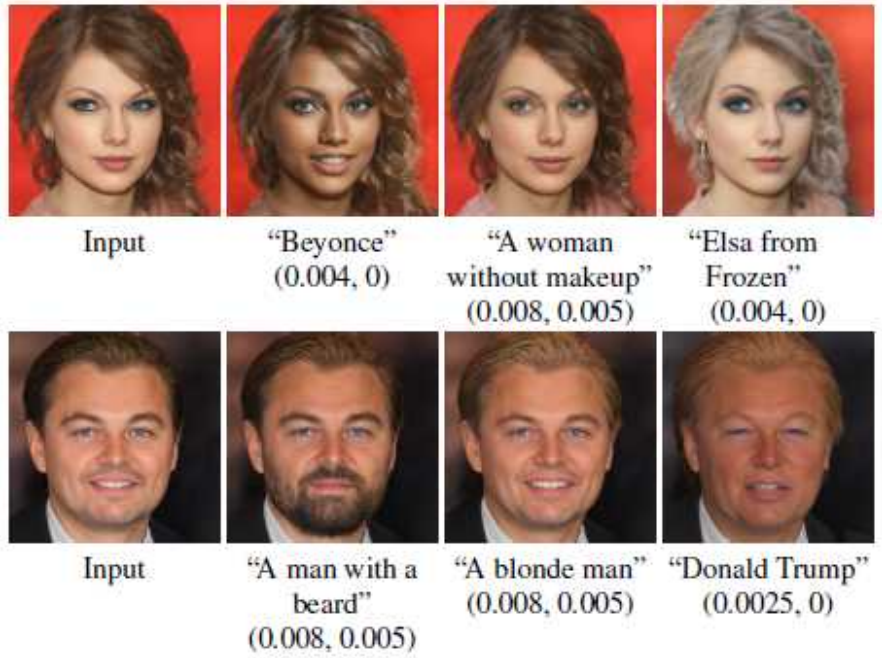


Figure 3. Edits of real celebrity portraits obtained by latent optimization. The driving text prompt and the $(\lambda_{\text{L2}}, \lambda_{\text{ID}})$ parameters for each edit are indicated under the corresponding result.

Method

- **Latent Mapper**

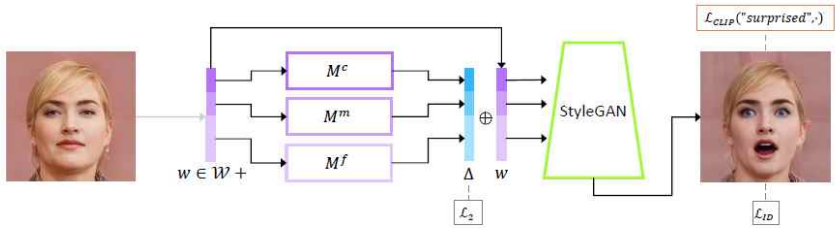


Figure 2. The architecture of our text-guided mapper (using the text prompt "surprised", in this example). The source image (left) is inverted into a latent code w. Three separate mapping functions are trained to generate residuals (in blue) that are added to w to yield the target code, from which a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

- **Latent code of the input image.**

$$w = (w_c, w_m, w_f)$$

- **One can choose to train only a subset of the three mappers.**

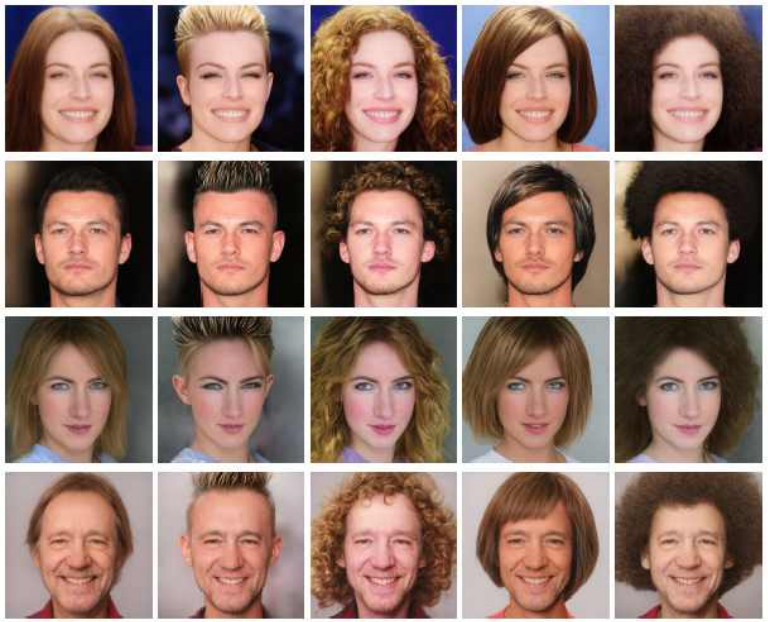
$$M_t(w) = (M_t^c(w_c), M_t^m(w_m), M_t^f(w_f))$$

- **Loss**

$$\mathcal{L}_{\text{CLIP}}(w) = D_{\text{CLIP}}(G(w + M_t(w)), t)$$

$$\mathcal{L}(w) = \mathcal{L}_{\text{CLIP}}(w) + \lambda_{L2} \left\| M_t(w) \right\|_2 + \lambda_{\text{ID}} \mathcal{L}_{\text{ID}}(w)$$

	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	$\mathcal{S}$



Input | "Mohawk hairstyle" | "Curly hair" | "Bob-cut hairstyle" | "Afro hairstyle"

	Mohawk	**Afro**	**Bob-cut**	**Curly**	**Beyonce**	**Taylor Swift**	**Surprised**	**Purple hair**
Mean	0.82	0.84	0.82	0.84	0.83	0.77	0.79	0.73
Std	0.096	0.085	0.095	0.088	0.081	0.107	0.893	0.145

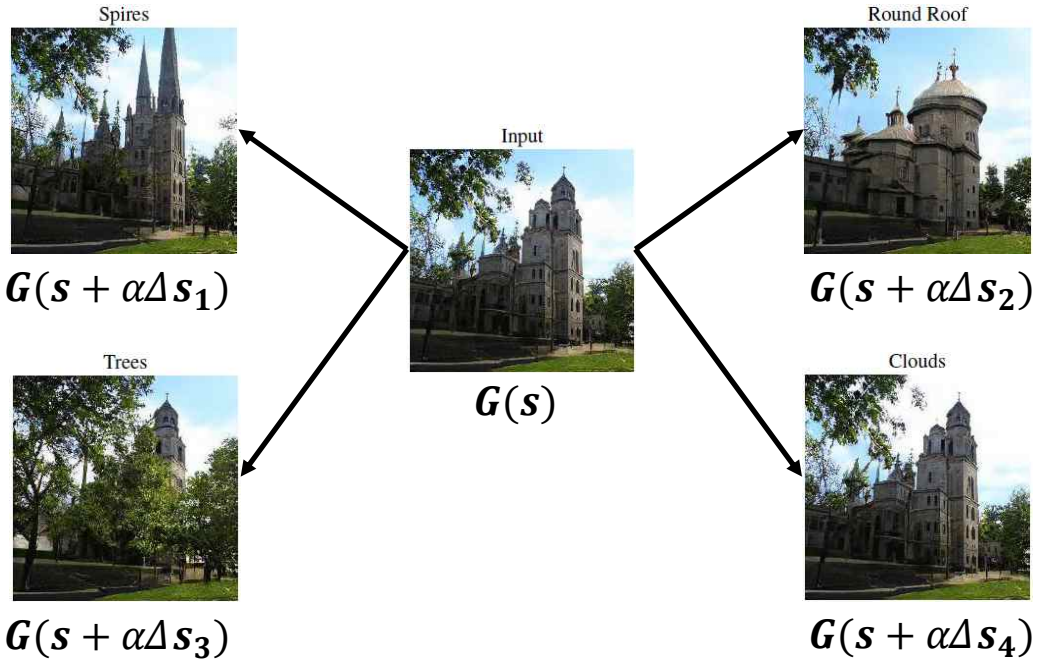
Table 2. Average cosine similarity between manipulation directions obtained from mappers trained using differnt text prompts.

Method

	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	$\mathcal{S}$

- **Global Direction**
  - **The manipulation strength is controlled by α.**

$$G(s + \alpha\Delta s)$$



Method

- **Global Direction**

  - **Parameter β may be used to control the degree of disentanglement in the manipulation.**

  - **Using higher threshold values results in more disentangled manipulations.**

  - **But at the same time the visual effect of the manipulation is reduced.**

	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	$\mathcal{S}$

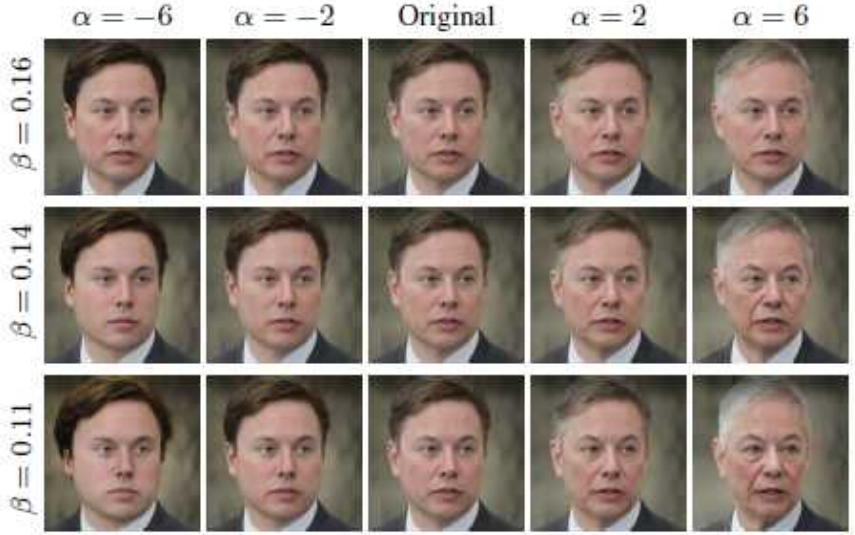


Figure 6. Image manipulation driven by the prompt "grey hair" for different manipulation strengths and disentanglement thresholds. Moving along the Δs direction, causes the hair color to become more grey, while steps in the $-\Delta s$ direction yields darker hair. The effect becomes stronger as the strength α increases. When the disentanglement threshold β is high, only the hair color is affected, and as β is lowered, additional correlated attributes, such as wrinkles and the shape of the face are affected as well.

Comparisons and Evaluation

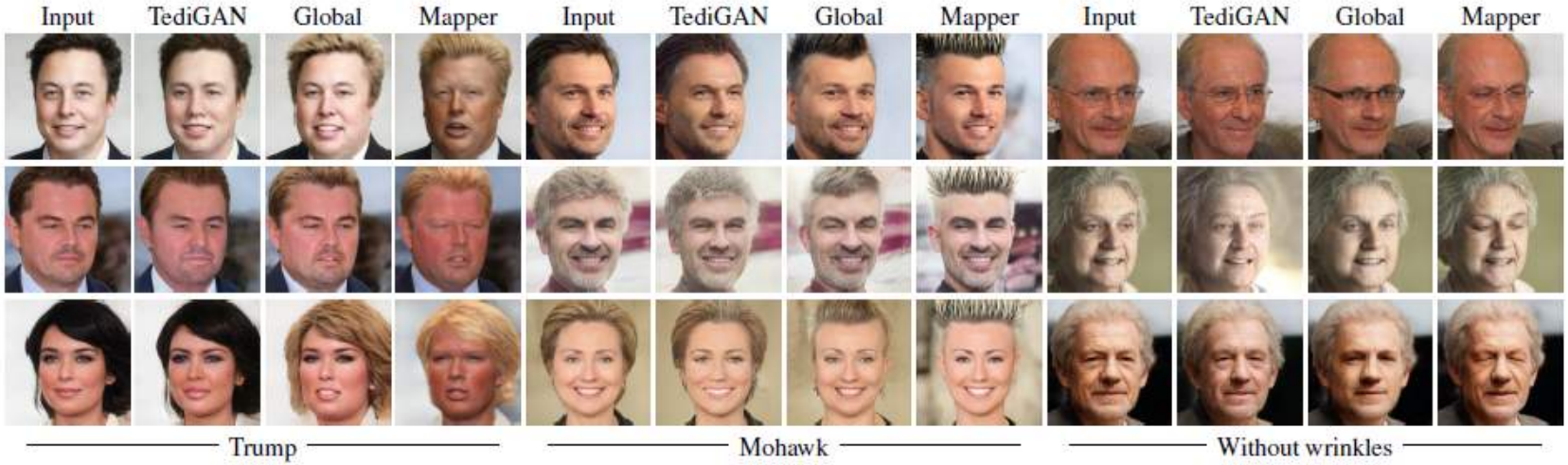


Figure 9. We compare three methods that utilize StyleGAN and CLIP using three different kinds of attributes.

Limitations

- **The methods rely on a pretrained Style-GAN generator and CLIP model for a joint language-vision embedding.**

- **Drastic manipulations in visually diverse datasets are difficult to achieve.**

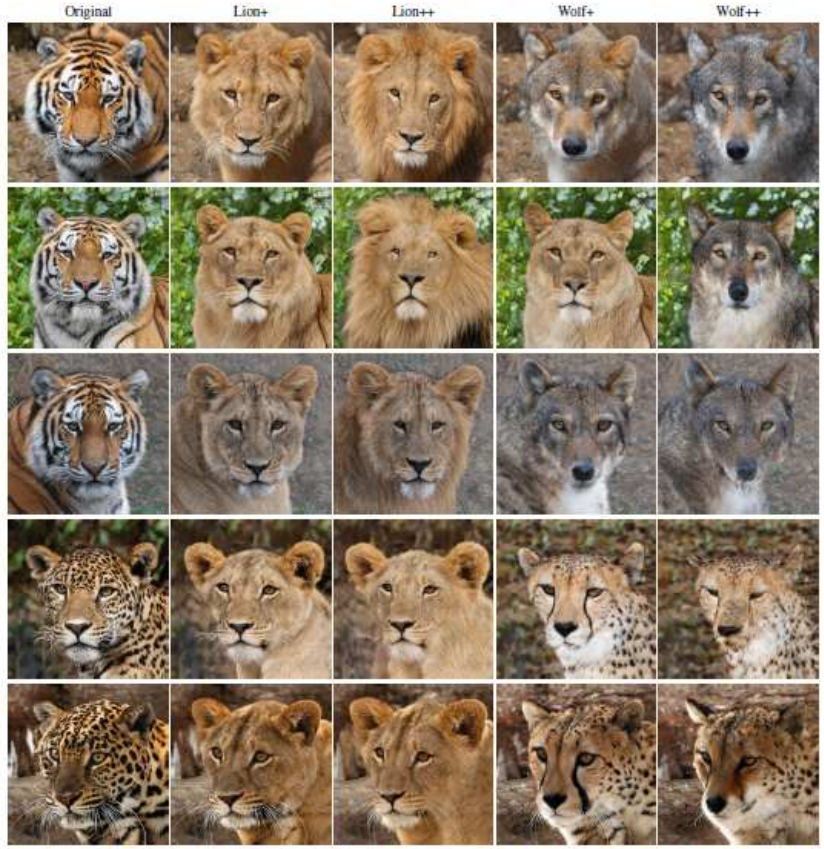


Figure 24. Drastic manipulations in visually diverse datasets are sometimes difficult to achieve using our global directions. Here we use StyleGAN-ada [17] pretrained on AFHQ wild [5], which contains wolves, lions, tigers and foxes. There is a smaller domain gap between tigers and lions, which mainly involves color and texture transformations. However, there is a larger domain gap between tigers and wolves, which, in addition to color and texture transformations, also involves more drastic shape deformations. This figure demonstrates that our global directions method is more successful in transforming tigers into lions, while failing in some cases to transform tigers to wolves. The "+" and "++" indicate medium and strong manipulation strength, respectively.

Conclusion

- **Propose three methods that bridge the gap between the latent space of StyleGAN and CLIP to realize text-driven image manipulation.**
  - **Introduce an optimization method that utilizes a CLIP-based loss.**
  - **Introduce a latent mapper that infers a text-guided latent manipulation.**
  - **Present a method for mapping text prompts to input-agnostic global directions.**

Reference

- **StyleGAN:** https://arxiv.org/abs/1812.04948

- **CLIP:** https://openai.com/blog/clip

Q & A