Text-to-Image Editing and Synthesis

RW of DreamBoth

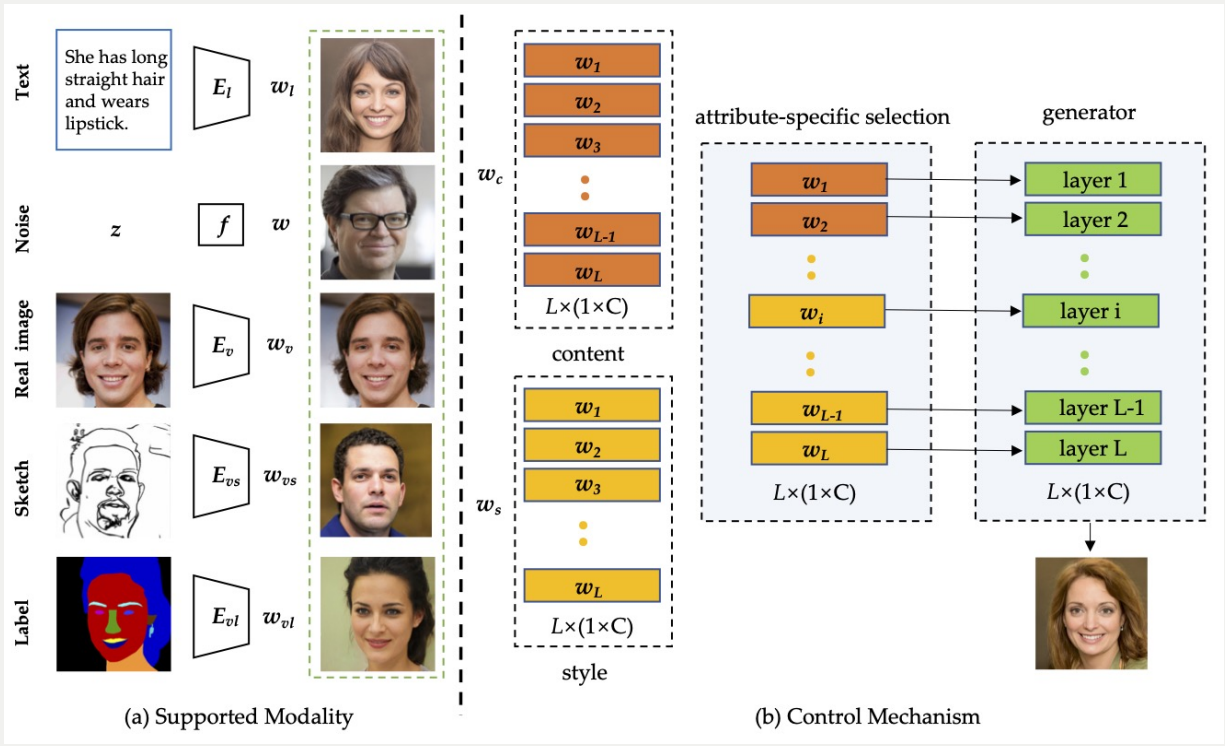
Writer: Chaeeun Ryu

> **Text-driven image manipulation has recently achieved significant progress using GANs combined with image-text representations such as CLIP yielding realistic manipulations using text.**

▼ Example)

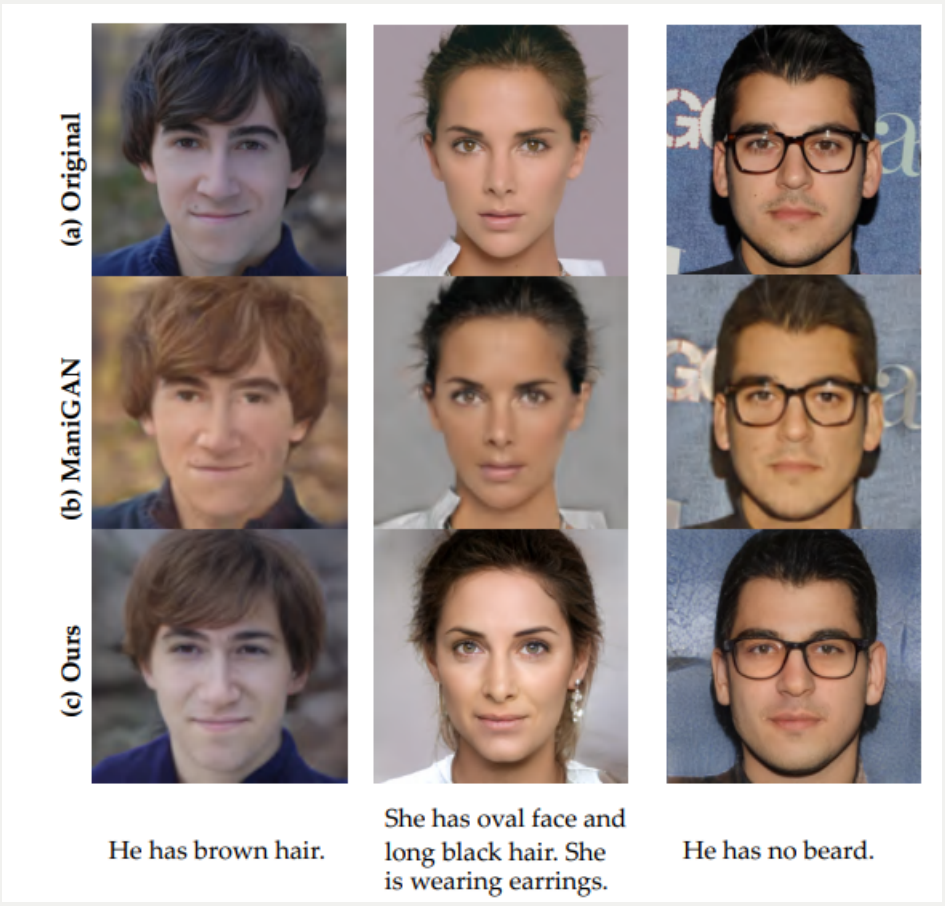
TediGAN: Text-Guided Diverse Face Image Generation and Manipulation

(demo page: https://replicate.com/iigroup/tedigan)

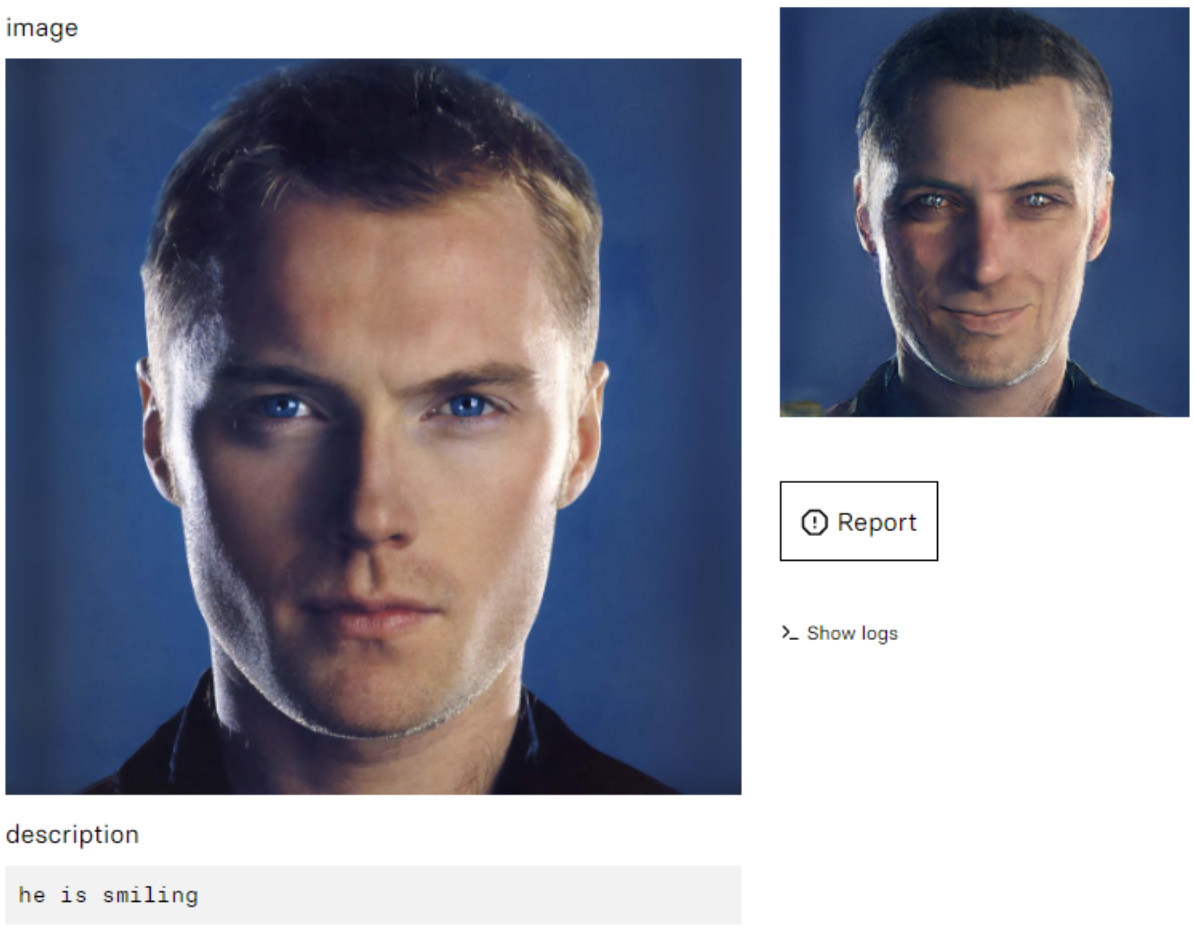


Framework (StyleGAN inversion module, visual-linguistic similarity learning, and instance-level optimization)

- GAN Inversion Module: an image encoder to map the real images to the latent space such that all codes produced by the encoder can be recovered at both the pixel-level and the semantic-level
- Visual-Linguistic Similarity Learning: We then use the hierarchical characteristic of W space to learn the text-image matching by mapping the image and text into the same joint embedding space.
- Instance-Level Optimization: To preserve identity in manipulation, we propose an instance-level optimization, involving the trained encoder as a regularization to better reconstruct the pixel values without affecting the semantic property of the inverted code.



(실제 example 본 개인적 소감: identity를 잘 preserve하는 것 같진 않다. 밑 예시 첨부)



These methods work well on structured scenarios (e.g. human face editing) and can struggle over diverse datasets where subjects are varied.

> **To alleviate this concern, Crowon et al. use VQ-GAN and train over more diverse data.**

VQGAN-CLIP: Open Domain Image Generation and Editing with Natural Language Guidance

: A novel methodology for both tasks which is capable of producing images of high visual quality from text prompts of significant semantic complexity without any training by using a multimodal encoder to guide image generations.

Methodology: Utilize pre-trained models, VQGAN and CLIP.

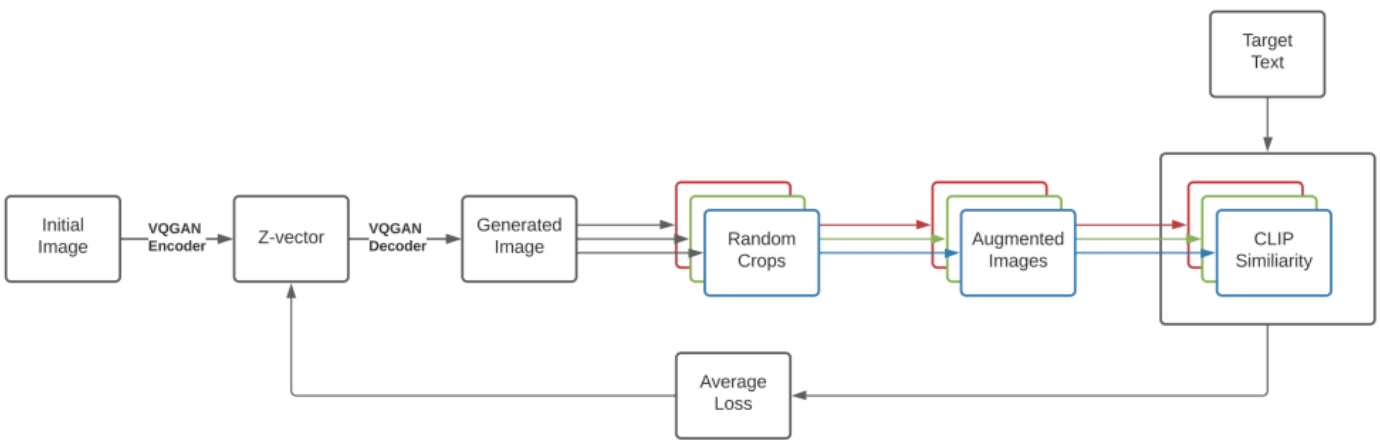


Diagram showing how augmentations are added to stabilize and improve the optimization. Multiple crops, each with different random augmentations, are applied to produce an average loss over a single source generation. This improves the results with respect to a single latent Z-vector.

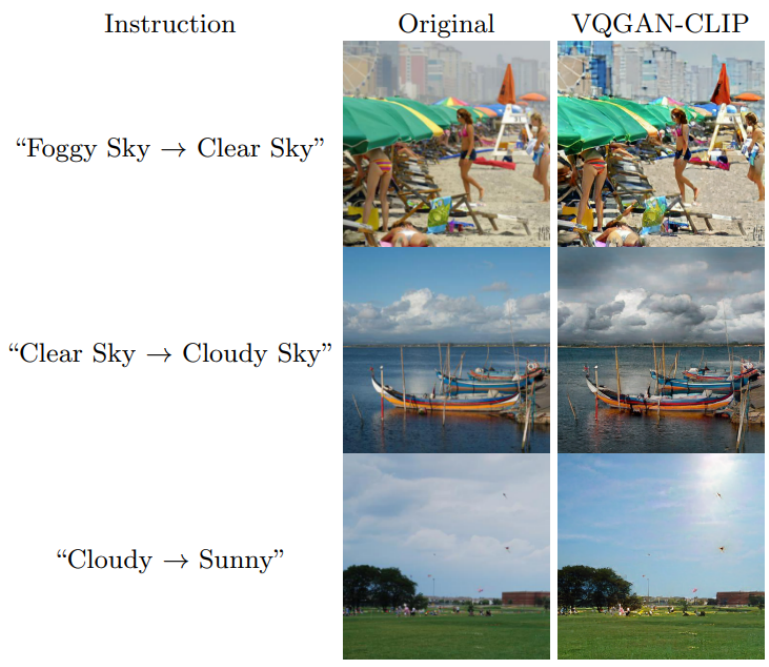
(a) the universal library trending on artstation

(b) a charcoal drawing of a cathedral

(c) a child's drawing of a baseball game

(d) a forest rendered in low poly

Text-based Generation of Images



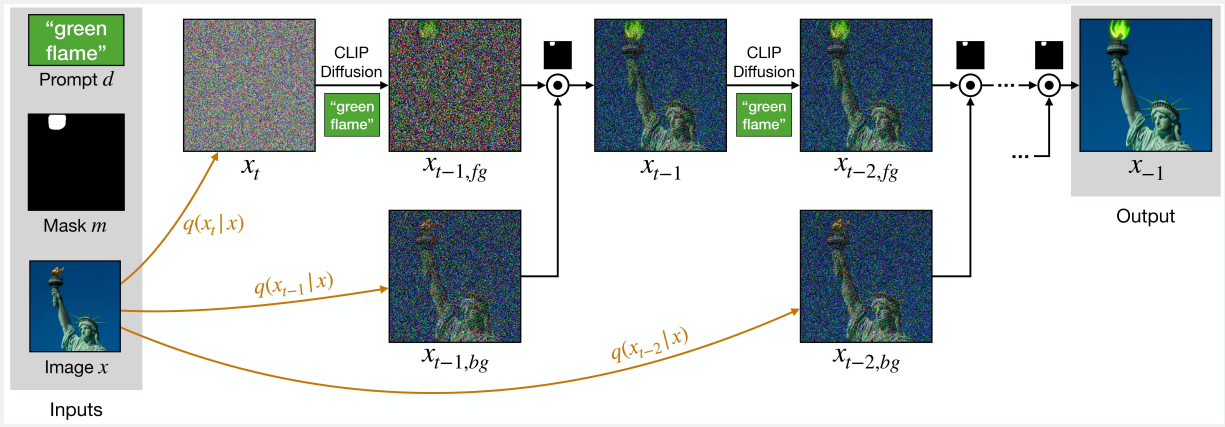
Weather Modification

Color Modification

> **Other works exploit the recent diffusion models, which achieve state-of-the-art generation quality over highly diverse datasets, often surpassing GANs.**

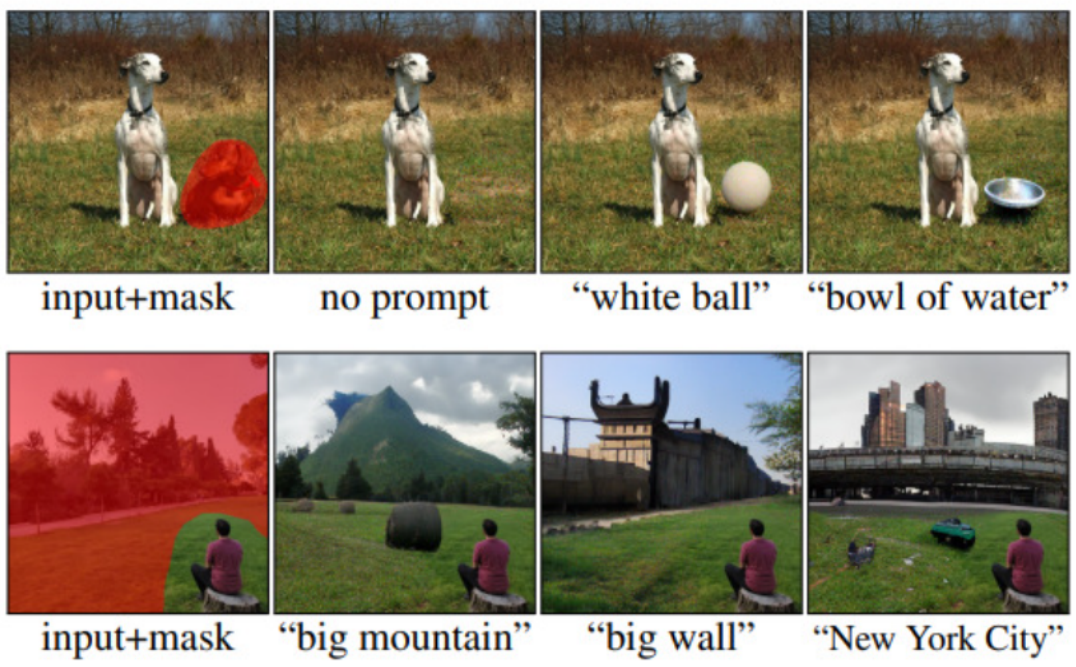
▼ Example)

Blended Diffusion: Text-driven Editing of Natural Images



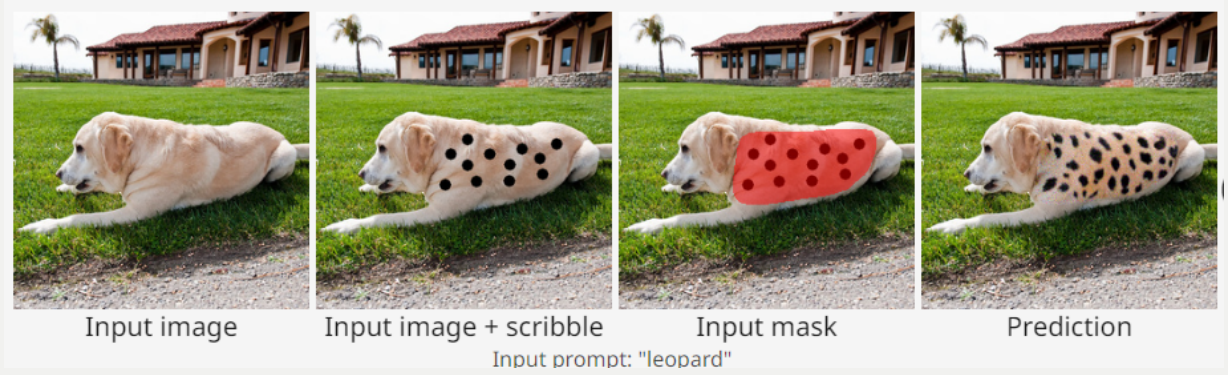
Methodology

Methodology: Given an image x, a guiding text prompt d and a binary mask m that marks the region of interest in the image, our goal is to produce a modified image $\hat{x}$, s.t. the content $\hat{x} \odot m$ is consistent with the text description d, while the complementary area remains as close as possible to the source image, i.e., $x \odot (1-m) \approx \hat{x} \odot (1-m)$, where $\odot$ is element-wise multiplication.



Given an input image and a mask, BLENDED DIFFUSION modifies the masked area according to a guiding text prompt, without affecting the unmasked regions

mask 해당 영역 text에 따라 edit



scribble도 반영 가능!

> **While most works that require only text are limited to global editing, Bar-Tal et al. proposed a text-based localized editing technique without using masks, showing impressive results.**

Text2LIVE: Text-Driven Layered Image and Video Editing

: Rather than directly generating the edited output, our key idea is to generate an *edit layer* (color+opacity) that is composited over the original input. This allows us to constrain the generation process and maintain high fidelity to the original input via novel text-driven losses that are applied directly to the edit layer.

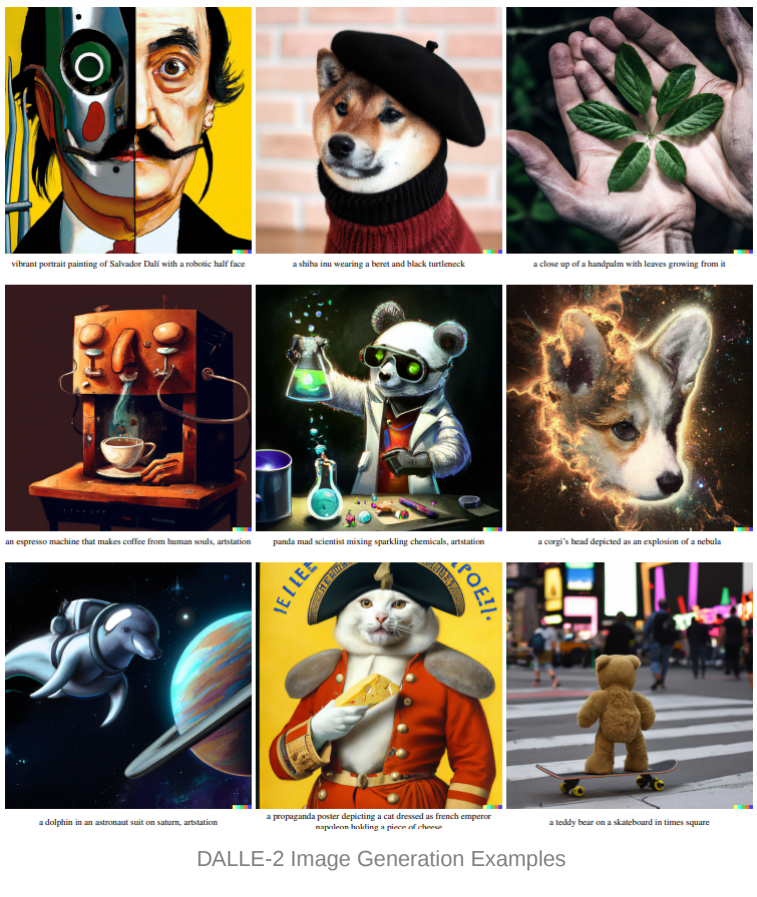
(a) Image text-guided layered editing

(b) Image editing results

https://text2live.github.io/videos/giraffe.mp4

While most of these editing approaches allow modification of global properties or local editing of a given image, none enables generating novel renditions of a given subject in new contexts

> **There also exists work on text-to-image synthesis.**



DALLE-2 Image Generation Examples

These models do not provide fine-grained control over a generated image and use text guidance only. Specifically, it is challenging or impossible to preserve the identity of a subject consistently across synthesized images