Blended Diffusion for Text-driven Editing of Natural Images [CVPR 2022]

M25037 장정우, GitHub repository url:
https://github.com/JeongWoo7780/Generative_Model_2025-blended-diffusion

※ 이 보고서 및 GitHub repository에 포함된 모든 그림은 직접 생성한 것입니다.

Getting Started

설치

0. Github repository에서 git clone을 통해 로컬에 파일 다운

1. 가상환경 생성

```
conda create --name blended-diffusion python=3.9
conda activate blended-diffusion
pip3 install ftfy regex matplotlib lpips kornia opencv-python torch==1.9.0+cu111 torchvision==0.10.0+cu
```

``` bash
conda create --name blended-diffusion python=3.9
conda activate blended-diffusion
pip3 install ftfy regex matplotlib lpips kornia opencv-python torch==1.9.0+cu111 torchvision==0.10.0+cu111 -f https://download.pytorch.org/whl/torch_stable.html
```

Numpy 오류가 발생하는 경우, 낮은 버전의 재설치가 필요합니다.

```
pip3 install "numpy<2.0"
```

2. “checkpoints” 디렉토리 생성 및 사전 학습 diffusion 모델 폴더로 다운로드

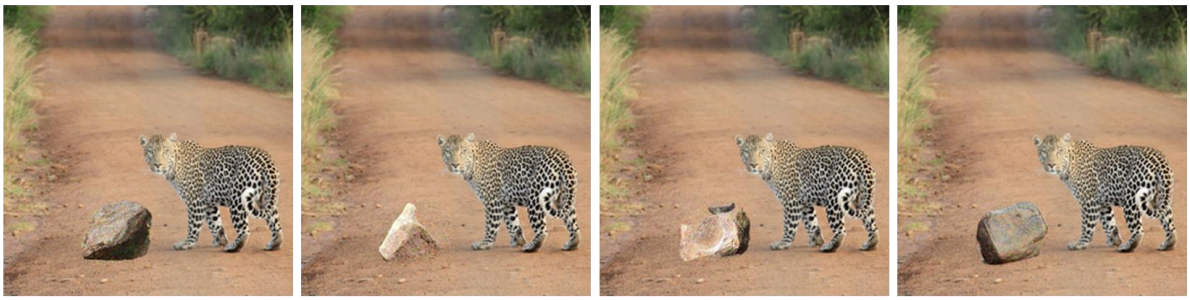
Image Generation

텍스트 기반 다중 합성 결과 생성 예시:

```
python main.py -p "rock" -i "input_example/img.png" --mask "input_example/mask.png" --output_path "outp
```

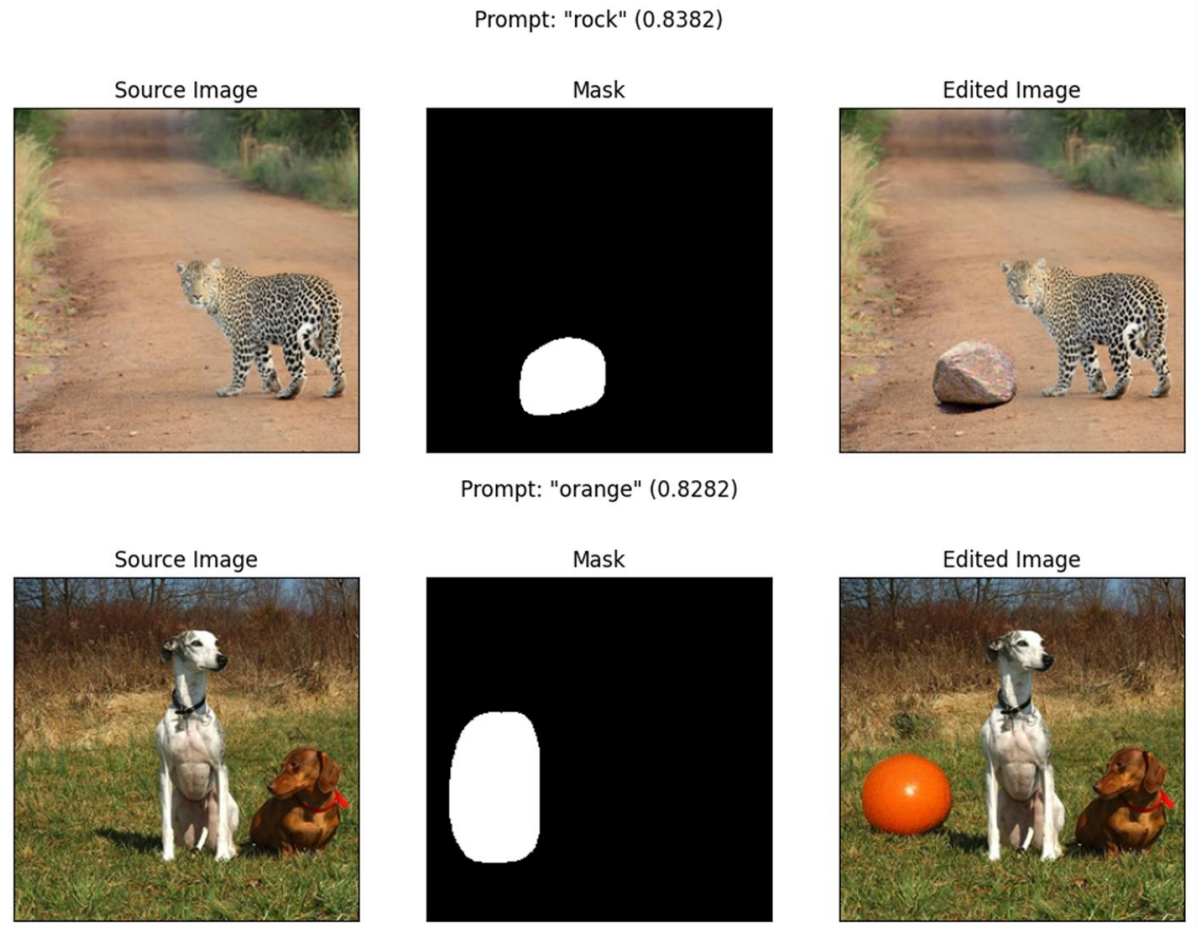
python main.py -p "rock" -i "input_example/img.png" --mask "input_example/mask.png" --output_path "output"

생성 결과는 output/racked 폴더에 CLIP 유사도 순위에 따라 정렬되어 저장됩니다. 최상의 결과를 얻기 위해서는 가능한 한 많은 수의 결과를 생성하고 (저자는 최소 64를 권장), 그 중 가장 좋은 것을 선택하는 것이 좋습니다.



CLIP 유사도에 따라 정렬된 그림의 예시: 왼쪽이 가장 높은 랭크이며, 오른쪽으로 갈수록 낮은 유사도에 따라 정렬.

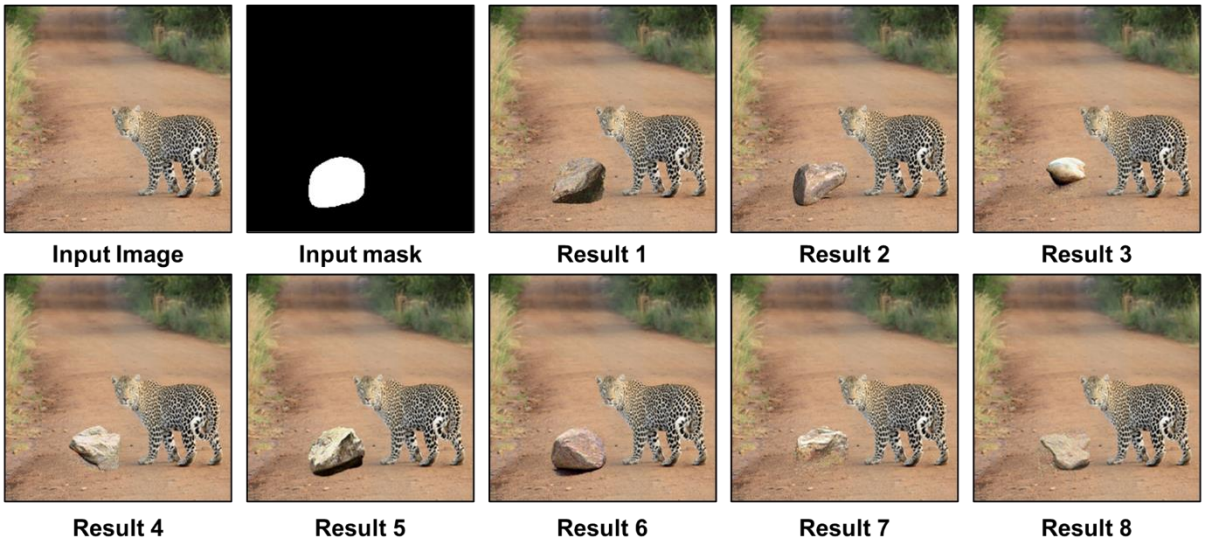
단일 확산 프로세스에서 다중 결과를 생성하기 위해 배치 처리를 활용했습니다. CUDA out of memory오류가 발생하면 먼저 --batch_size 1을 설정하여 배치 크기를 줄여보시기 바랍니다.



텍스트 기반 이미지 합성 예시: 원본 이미지, 이진 마스크, 텍스트 프롬프트 “rock”의 결과 (first row) 및 원본 이미지, 이진 마스크, 텍스트 프롬프트 “orange”의 생성 결과 (second row).

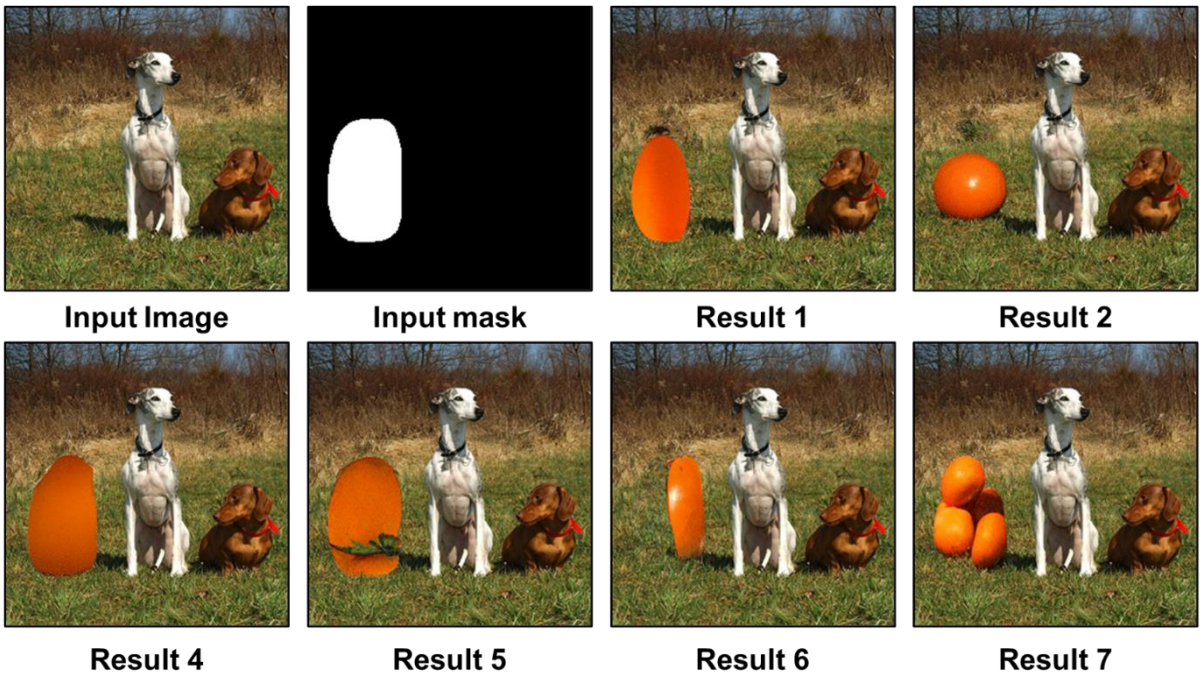
Applications

Multiple synthesis results for the same prompt



Adding a new object (multiple results for the same input): Given the input image, mask, and text description "rock", proposed model is able to generate multiple plausible results.

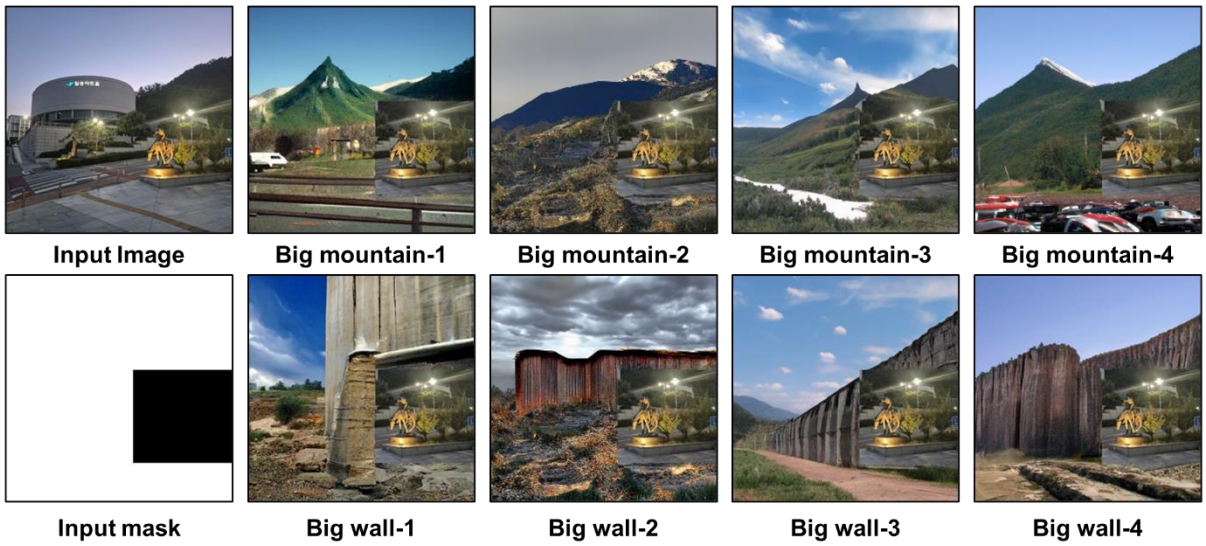
텍스트 프롬프트 "rock"과 예시 이미지, 예시 마스크를 사용하여 생성한 여러 결과 예시.



Adding a new object (multiple results for the same input): Given the input image, mask, and text description "orange", proposed model is able to generate multiple plausible results.

텍스트 프롬프트 "orange"와 예시 이미지, 예시 마스크를 사용하여 생성한 여러 결과 예시.

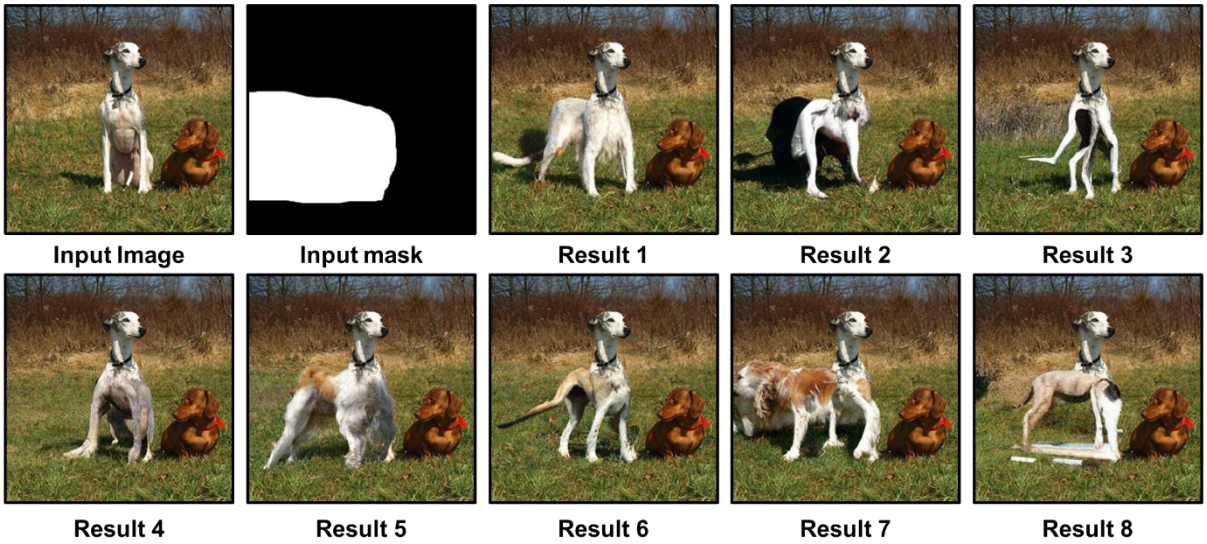
Background replacement



Background replacement: 입력 이미지 (일송아트센터), 제작한 이진 마스크, 텍스트 프롬프트 "Big mountain" 과 "Big wall" 의 생성 결과. 매끄럽지 못한 마스크의 형태 때문인지, 어색한 결과가 생성되는 것을 확인.

예시 그림이 아닌, 실제 이미지에서 (일송아트센터)의 생성 결과 확인을 위해 직접 제작한 마스크를 활용해 배경 교체를 시도했으나, 매끄럽지 못한 마스크의 형태 때문인지 (보존하려고 한 동상 이미지에 국한된 마스크를 생성해야 했으나 실패함.) 전경과 배경이 전혀 어울리지 않는 이미지를 생성하는 것을 확인할 수 있습니다.

Altering part of an existing object



기존 전경 객체의 일부를 변경: 입력 이미지 및 마스크를 입력으로 받아, 텍스트 "body of a standing dog"에 해당하는 전경 객체를 변경하는 것을 목표로 함. 여러 가지 가능한 결과가 생성되며, 일부 결과는 다른 결과보다 더 그럴듯한 결과를 생성.

저자들의 예시 재현을 위해, 마스크를 제작하여 동일한 텍스트 프롬프트로 시도한 생성 결과입니다. 권장하는 대로 높은 수의 이미지를 생성하지 않고, default 세팅으로 시도했기 때문에 Result 1을 제외하고는 좋은 결과를 보여주지 못하는 모습입니다.

Overall Discussions

본 보고서에서는 *Blended Diffusion for Text-driven Editing of Natural Images* 논문의 구현 코드를 재현해보았습니다. 저자들이 방법을 비교적 상세히 설명해두었기 때문에, 일부 사소한 오류를 제외하고 전체적인 실행에는 큰 문제가 없었습니다.

다만, 제공된 예시 이미지와 마스크가 각 2장뿐이었고, 마스크를 직접 제작하는 과정이 다소 까다로워 다양한 실험을 진행하는 데 제약이 있었습니다. 또한, 저자들이 권장한 이미지 생성 수(최소 64장)를 컴퓨팅 자원 문제로 충분히 시도하지 못해, 일부 결과물의 텍스트 프롬프트 반영도가 다소 낮게 나타났습니다.

아울러, 논문과 GitHub의 Applications 항목에서 소개된 다양한 활용 사례(Scribble-guided editing, Text-guided extrapolation, Composing several applications 등)를 실습해보지 못한 점은 아쉬운 부분입니다. 이는 앞서 언급한 컴퓨팅 자원 한계 및 마스크 제작의 어려움 때문이었습니다.