Free-Form Image Inpainting with Gated Convolution

김형범

2021.11.15

Introduction

- **Image inpainting 이란?**

- 이미지에서 누락된 영역을 시각적으로 현실적이고 의미론적으로 정확하도록 수정하는 작업
- 이를 통해 이미지에서 방해물을 제거하거나 원하지 않는 부분을 수정할 수 있다.

Introduction

- **컴퓨터 비전에서는 image inpainting에 대한 두가지 접근법이 존재한다.**

1. low level image feature를 사용한 패치 매칭
   - 정적인 texture는 합성할 수 있지만 복잡한 장면, 얼굴, 정지 상태가 아닌 물체의 경우 심각하게 성능이 저하된다.

2. deep convolution 네트워크를 가진 feed-forward 생성 모델
   - 대규모 데이터 셋에서 학습한 semantics를 활용하여 비정형 이미지의 contents를 최신의 방법으로 합성할 수 있다.
   - 하지만 vanilla convolution을 기반으로 하는 deep generative model은 convolution 필터가 모든 입력 픽셀(feature)를 동일한 유효 픽셀로 취급하기 때문에 이미지 구멍 채우기에 적합하지 않다

이미지 구멍(image hole)이란?
이미지에서 누락된 부분 또는 수정을 원하는 부분

Introduction

ㅇ 이미지 구멍 채우기의 경우 각 레이어에 대한 입력은 구멍 외부의 유효한 영역과 내부의 유효하지 않는(무효) 픽셀로 구성된다.

ㅇ vanilla convloution은 모든 유효 픽셀과 무효 픽셀에 동일한 필터를 적용하므로 free form mask로 test 했을 때 색상 불일치, 블러효과, edge effect 와 같은 시각적 아티팩트가 발생한다.

Origin input output

→ 이 한계를 해결하기 위해 convolution이 마스크되고 유효한 픽셀에서만 조건화 되도록 정규화된 최신 partial convolution이 제안되었다.

→ Partial convolution은 모든 입력 위치를 무효하거나 유효한 것으로 분류하고 0 또는 1 마스크를 모든 계층의 입력에 곱한다.

Vanilla convolution

ㅇ 바닐라 컨볼루션이 free-form image inpainting에 적합하지 않은 이유

- 입력이 C 채널, 출력이 C′ 채널이라 가정하고 출력 map의 (y,x)에 위치한 각 픽셀은 다음과 같이 계산된다.(K는 kernel size를 의미하고 K′는 (K-1)/2 이다)

$$O_{y,x} = \sum_{i=-k'_h}^{k'_h} \sum_{j=-k'_w}^{k'_w} W_{k'_h+i,k'_w+j} \cdot I_{y+i,x+j},$$

- 위 식은 모든 공간 위치 (y,x)에 동일한 필터를 적용하여 출력을 생성한다는 것을 보여준다.

- 이는 입력 이미지의 모든 픽셀이 유효한 픽셀로 간주되기 때문에 전체 이미지의 feature를 처리하는 것이기 때문에 이미지 분류 및 객체 탐지와 같은 작업이 적합하다.

Vanilla convolution

ㅇ 바닐라 컨볼루션이 free-form image inpainting에 적합하지 않은 이유

- 입력이 C 채널, 출력이 C′ 채널이라 가정하고 출력 map의 (y,x)에 위치한 각 픽셀은 다음과 같이 계산된다.(K는 kernel size를 의미하고 K′는 (K-1)/2 이다)

$$O_{y,x} = \sum_{i=-k'_h}^{k'_h} \sum_{j=-k'_w}^{k'_w} W_{k'_h+i,k'_w+j} \cdot I_{y+i,x+j},$$

- 위 식은 모든 공간 위치 (y,x)에 동일한 필터를 적용하여 출력을 생성한다는 것을 보여준다.

- 이는 입력 이미지의 모든 픽셀이 유효한 픽셀로 간주되기 때문에 전체 이미지의 feature를 처리하는 것이기 때문에 이미지 분류 및 객체 탐지와 같은 작업이 적합하다.

그러나 이미지 인페인팅에 경우 입력은 유효 픽셀과 무효 픽셀로 구성되어야 하므로 vanilla convolution을 사용하면 시각적 아티팩트가 발생한다.

Partial convolution

ㅇ 따라서 vanilla convolution을 개선하기 위해 최근 연구에서는 partial convolution이 제안되어 masking, re-normalization step을 통해 컨볼루션이 유효 픽셀에만 적용되게 한다.

$$O_{y,x} = \begin{cases} \sum\sum W \cdot (I \odot \frac{M}{sum(M)}), & \text{if sum(M)} > 0 \\ 0, & \text{otherwise} \end{cases}$$

- M은 binary mask를 뜻하고 1은 유효하다는 것을 의미하고 0은 픽셀이 유효하지 않다는 것을 의미한다.(하드 게이트 마스크)

- partial 컨볼루션 연산 후에는 다음 규칙을 사용하여 M(마스크)를 갱신한다.

$$m'_{y,x} = 1, \text{iff sum(M)} > 0.$$

Partial convolution

ㅇ Partial convolution은 inpainting의 품질을 개선하지만 여전히 문제가 남아있다.

1. 모든 Spatial locations에 대해 픽셀이 유효한지 않은지를 경험적으로 분류를 진행한다.
→ 다음 레이어의 마스크는 이전 레이어의 필터 범위에 의해 커버되는 픽셀 수에 관계 없이 하나로 설정된다.
→ 정해진 rule에 따라서 M을 갱신하다 보면 어떤 feature는 1개의 valid pixel을 커버하여 생성되고 또 다른 하나는 filter가 9개의 valid pixel를 cover하여 생성된다. 두 픽셀은 유효한 정도가 차이가 있지만 동일하게 1의 mask 값으로 갱신된다.

2. User-guided image inpainting이 불가능하다.
→ 마스크가 정해진 룰에 의해 정해지므로 불가능하다.

Partial convolution

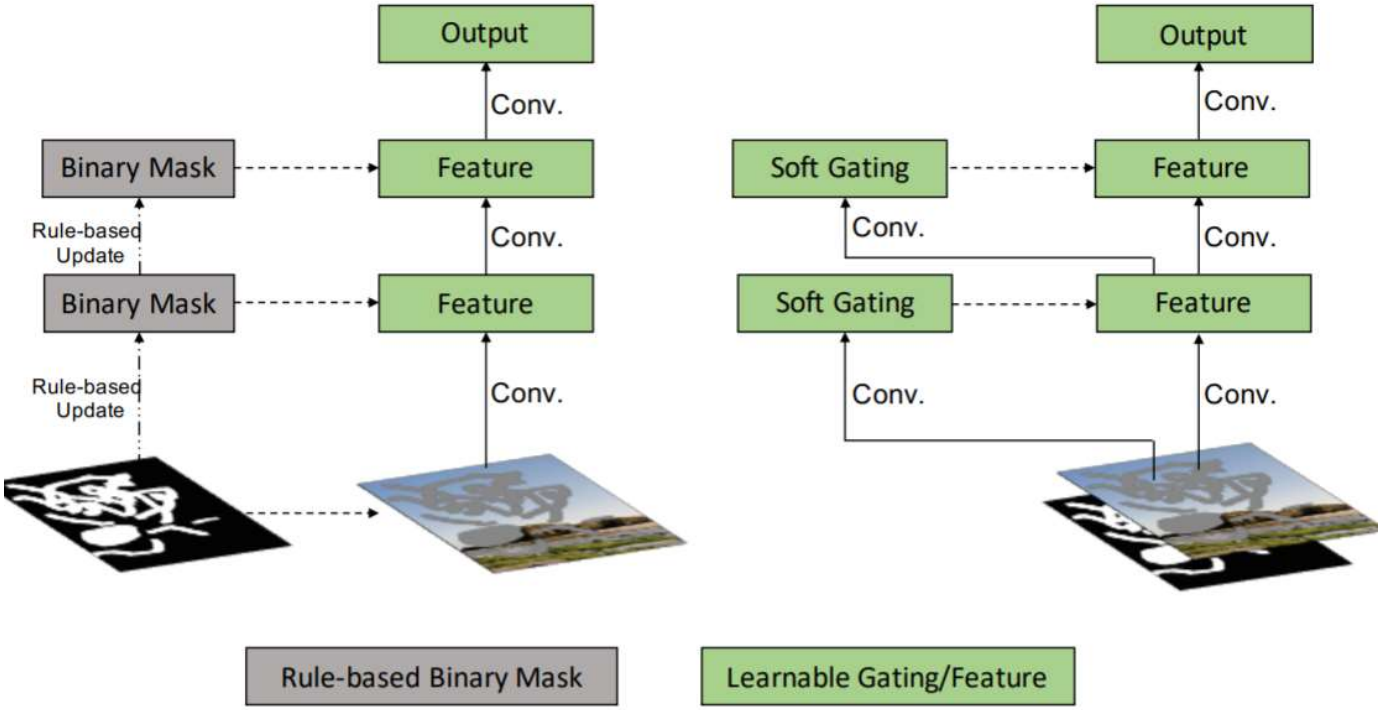
ㅇ Partial convolution은 inpainting의 품질을 개선하지만 여전히 문제가 남아있다.

3. Layer가 깊어짐에 따라 유효하지 않는 픽셀이 점점 줄어들고 결국 모든 mask 값들은 1로 변환된다.
→ 비교적 적은 비율의 유효하지 않는 픽셀이 레이어가 깊어질 수록 없어진다

4. 각 layer의 모든 channel는 같은 mask를 공유한다.
→ 이렇기 때문에 유연한 결과 이미지를 생성하는데 어려움이 있다.

모든 문제를 개선하고 inpainting의 결과를 개선하기 위해서 Gated convolution을 제안한다.

Gated convolution



Partial convolution

Gated convolution

Gated convolution

$$Gating_{y,x} = \sum\sum W_g \cdot I$$

$$Feature_{y,x} = \sum\sum W_f \cdot I$$

$$O_{y,x} = \phi(Feature_{y,x}) \odot \sigma(Gating_{y,x})$$

- $Gating_{y,x}$, $Feature_{y,x}$는 각각 다른 weights에 대한 convolution 연산 결과이다.
- Activation function과 sigmoid, pixel wise 곱을 사용한다.
- Feature $_{x,y}$를 추출하고 해당 이미지에서 soft mask인 $Gating_{y,x}$를 얻어 element-wise product를 진행하여 valid feature일 확률이 강한 feature에는 강한 attention을 주도록 연산이 된다.

이를 논문에서는 dynamic feature selection mechanism for each channel and each spatial localtion을 학습한다고 한다.

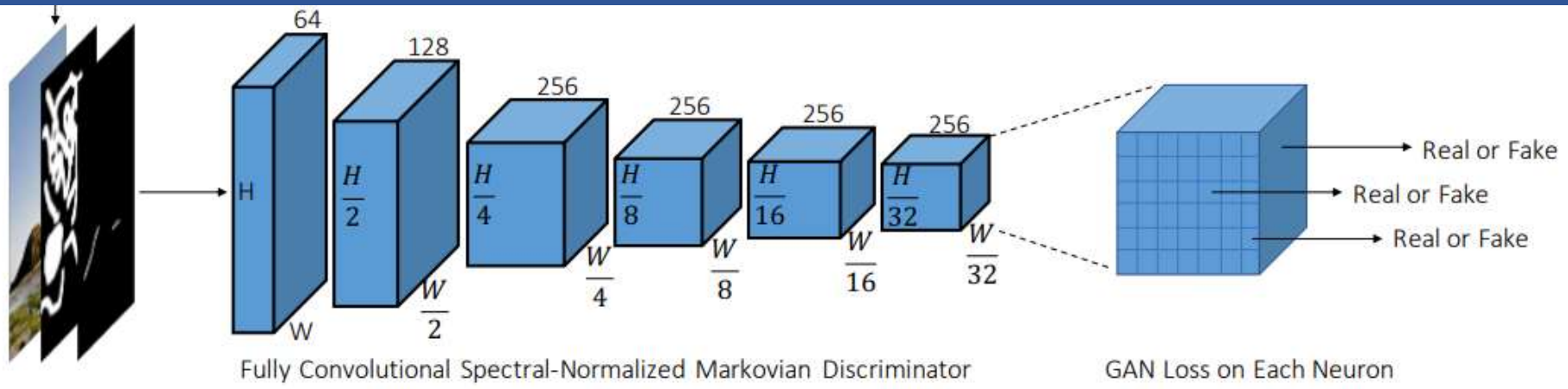
SN-PatchGAN

ㅇ 본 논문은 free-form image inpainting이 목적이다.
- free-form image inpainting이란 임의의 위치와 임의의 모양을 가진 다수의 holes를 자연스럽게 채우는 것을 의미한다.

ㅇ 단일 직사각형 모양의 hole을 채우려는 이전의 inpainting 네트워크의 경우 masking된 직사각형 영역(즉 hole)에 추가적인 local GAN을 사용하여 결과를 개선해왔다.
- 하지만 임의의 위치와 임의의 모양을 처리해야 하는 free-form image inpainting에는 적합하지 않는 방법이다.

free-form image inpainting 네트워크를 훈련하기 위한 단순하고 효과적인 GAN 손실과 SN-PatchGAN을 제안한다.
SN=spectral normalized

SN-PatchGAN

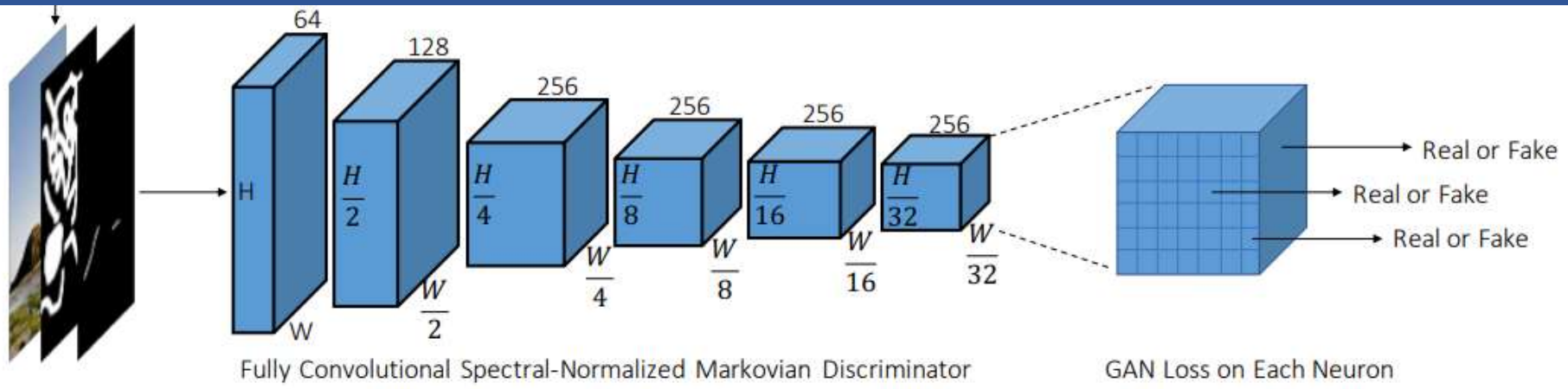


ㅇ CNN으로 이루어진 이 discriminator는 이미지,마스크,guidance channels을 입력으로 받아 3D-feature of shape R^{HxWxC}를 출력한다.

ㅇ 그림에서 볼 수 있듯이 6개의 strided convolution을 쌓아 Patches의 feature statistics를 capture하여 Real 인지 Fake인지 판별한다.

ㅇ 그리고 학습을 안정화하기 위해 최근 제안된 spectral normalization을 채택했다.
- weight normalization 기술 중 하나로 gradient에 제약을 주어 학습을 용이하게 한다.

SN-PatchGAN



- 입력이 진짜인지 가짜인지 구별하기 위해 힌지 loss를 목적함수로 사용한다.

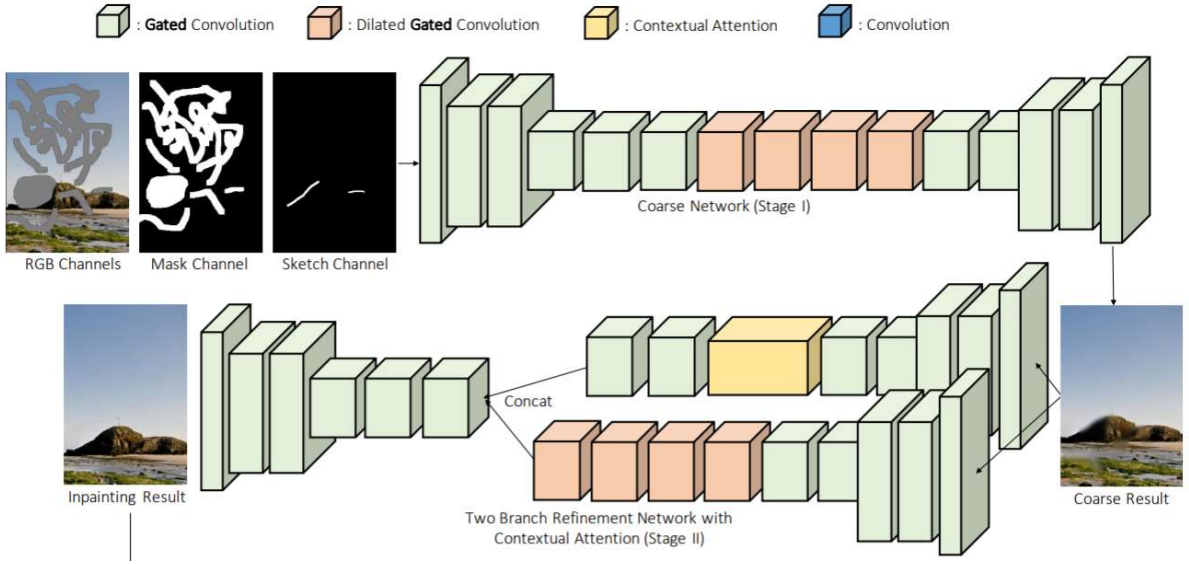
$$loss = max\{0, 1 - (y' \times y)\}$$

- hinge loss는 학습 데이터 각각의 class를 구분하면서 두 데이터 분포 간 거리가 가장 먼 decision boundary를 찾기 위해 고안된 loss 함수다.

For generator, $\mathcal{L}_G = -\mathbb{E}_{z\sim\mathbb{P}_z(z)}[D^{sn}(G(z))]$ For discriminator,

$$\mathcal{L}_{D^{sn}} = \mathbb{E}_{x\sim\mathbb{P}_{data}(x)}[ReLU(1 - D^{sn}(x))] + \mathbb{E}_{z\sim\mathbb{P}_z(z)}[ReLU(1 + D^{sn}(G(x)))]$$

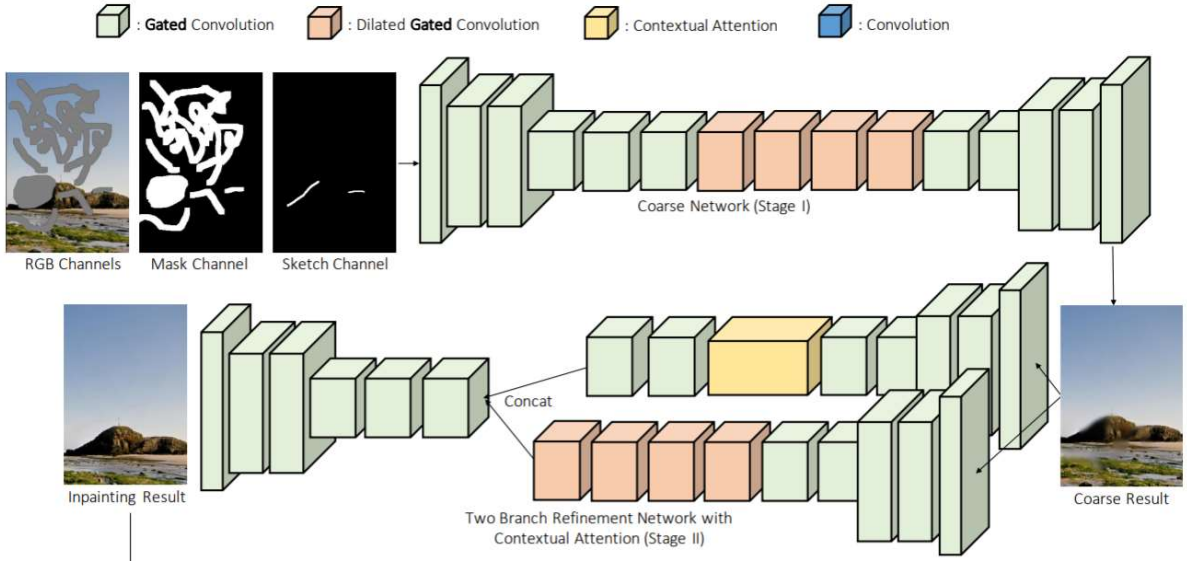
SN-PatchGAN



- inpainting network(generator)는 다음과 같이 coarse and refinement network 구조를 가진다.

- Partial convolution에서 사용한 U-net 구조를 사용하지 않고 encoder-decoder 구조를 사용했다.
  - U-net 구조에서 사용되는 skip connection이 Hole의 경계면의 디테일한 색상과 texture 정보를 전달하지 못하는 것을 발견했기 때문이다.

SN-PatchGAN



ㅇ 각 레이어는 모두 Gated convolution으로 이루어져 있다.

ㅇ Coarse Network에서는 입력한 이미지에 Mask를 적용하여 inpainting한 결과를 만들어내고 Refinement network에서는 시각적으로 자연스럽게 만들어준다.

Free-Form Mask Generation

ㅇ Free-form mask을 생성하는 것은 중요한 일이다. Free-form mask를 생성할 때에는 다음 네가지 요건을 지켜야 한다.

1. 실제 사용 사례에서 그린 마스크와 유사해야 함
2. 과적합(Overfitting)을 방지하기 위해 다양하게 생성되어야 함.
3. 계산 및 저장에 효율적이어야 함
4. 제어 가능하고 유현해야 함

ㅇ 이전 방법은 연속된 두 영상 프레임 사이의 occlusion estimation 방법에서 고정된 불규칙 마스크 세트를 수집한다.

- 다양성을 증가시키기 위해 확장, 회전 및 자르기 기능이 추가되지만 다른 요견을 충족하지 못한다.

→ train 중에 무작위 free-form 마스크를 자동으로 생성하는 알고리즘을 도입하였다.

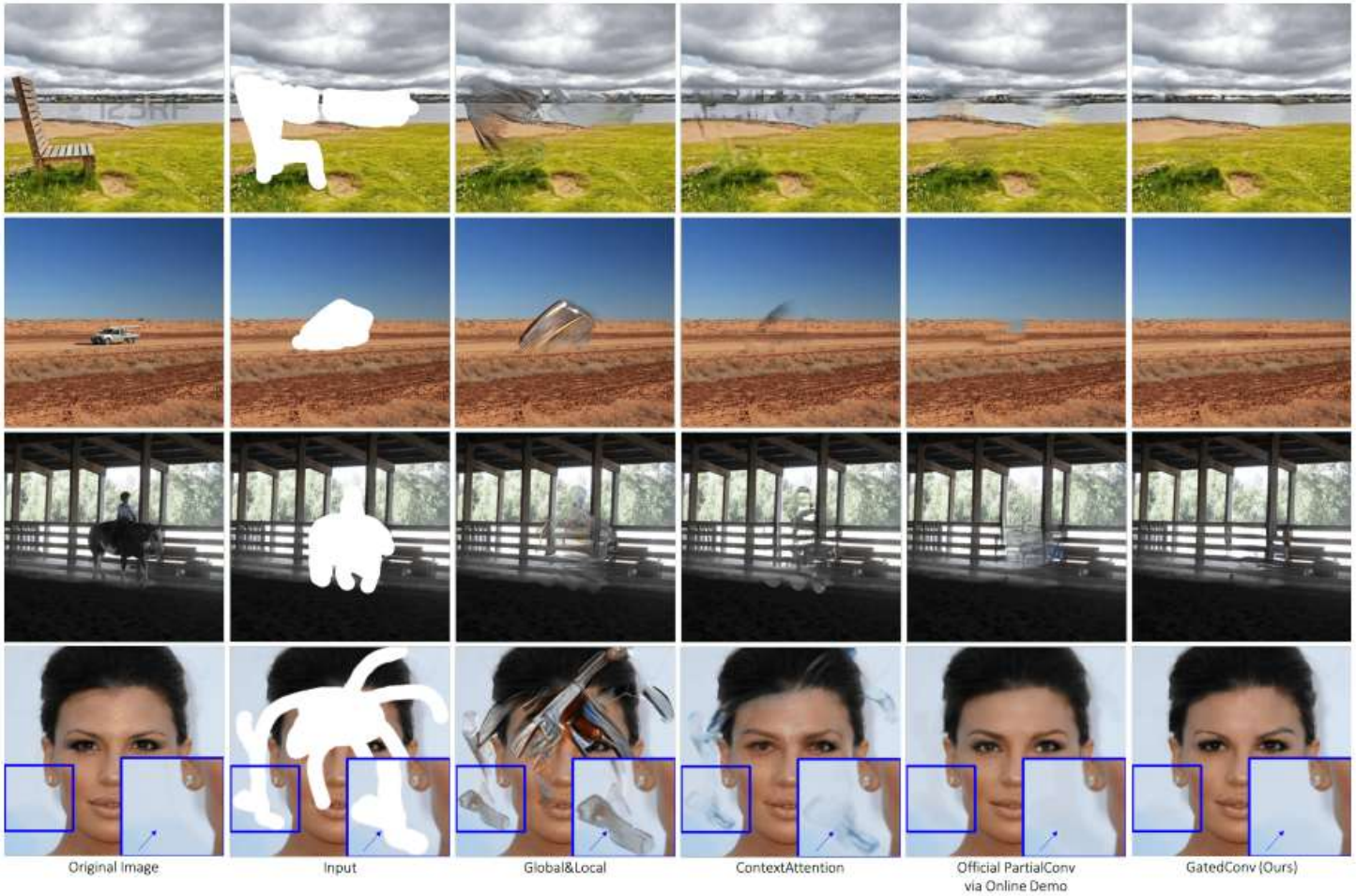
Results

ㅇ 이미지 inpainting에는 좋은 정량적 평가 지표가 부족하다.
- 그럼에도 불구하고 이미지 중심의 직사각형 마스크와 free-form 마스크를 사용 했을 때 Places2 데이터 셋에 대한 평균 L1 loss 및 L2 loss를 측정하였다.

	rectangular mask		free-form mask	
Method	ℓ_1 err.	ℓ_2 err.	ℓ_1 err.	ℓ_2 err.
PatchMatch [3]	16.1%	3.9%	11.3%	2.4%
Global&Local [15]	9.3%	2.2%	21.6%	7.1%
ContextAttention [49]	**8.6%**	2.1%	17.2%	4.7%
PartialConv* [23]	9.8%	2.3%	10.4%	1.9%
Ours	**8.6%**	**2.0%**	**9.1%**	**1.6%**

ㅇ 직사각형 마스크, free-form mask 둘 다 우리의 모델에서 가장 낮은 loss를 기록했다.

Results


Original Image
Input
Global&Local
ContextAttention
Official PartialConv via Online Demo
GatedConv (Ours)

Results

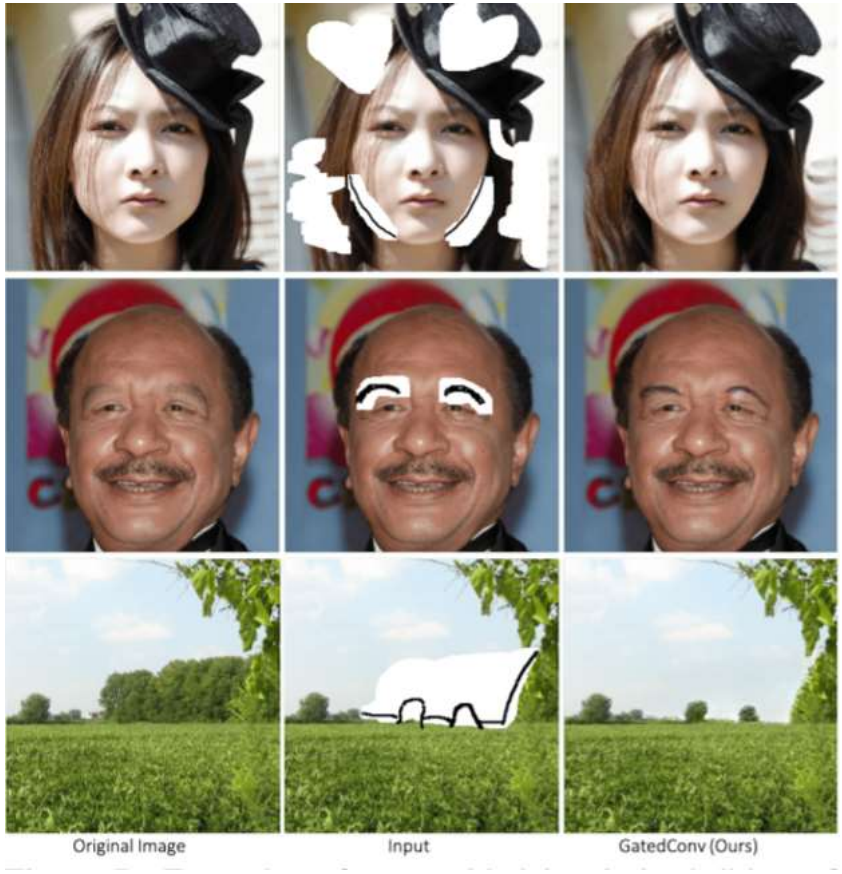
ㅇ 결과 이미지를 이전의 최첨단 방법들과 비교하였다.

ㅇ vanilla convolution의 경우 결과 이미지에서 명백한 시각적 아티팩트와 hole 가장자리에서 edge effect를 관찰할 수 있다.

ㅇ Partial GAN은 더 나은 결과를 생성하지만 관찰 가능한 색상 불일치를 보여준다.

ㅇ Gated convolution에 기초한 우리의 방법은 눈에 띄는 색상 불일치 없이 시각적으로 만족스러운 결과를 얻는다.

Results



ㅇGated convolution을 사용하여 이미지를 creative하게 편집하는 것도 가능하다.

ㅇ 사용자가 입력으로 그린 사용자 스케치에 따라 그림을 편집한다.

Results

ㅇ 결과 이미지를 inpainting의 품질과 얼마나 자연스러운가에 대해 human study를 진행하였다.

GT	Ours	Re-implemented Partial Conv	Official Partial Conv
9.89	7.72	7.07	6.54

ㅇ 본 논문의 결과와 이전의 inpainting 최신 기술인 Partial conv와 비교했을 때 human study 결과 역시 높게 측정되었다.

Conclusion

ㅇ Free-form inpainting을 위한 SN-PatchGAN으로 훈련된 게이트 컨볼루션 네트워크를 제시하였다.

ㅇ 게이트 컨볼루션은 free-form 마스크와 inpainting 결과를 크게 개선한다는 것을 다른 모델의 결과와 비교함으로써 입증하였다.

ㅇ 정량적, 정성적인 결과 비교 및 human study는 제안된 방법의 free-form inpainting 시스템의 우수성을 입증하였다.