이미지 인페인팅을 위한 윤곽 유지 및 내부 문맥 개선법

구윤회, 이한솔, 이창화, 정민우, 하선, 조윤성, 정해선, 박한수, 김은서, 백승렬
울산과학기술원
{kks6716, hansollee, changhwalee, minu0122, seonha, yscho710, hss9594, hansupark, eskim, srbaek}@unist.ac.kr

요 약

인페인팅 (inpainting) 분야는 많은 연구자의 관심을 받으며 매년 성장하고 있다. 인페인팅 분야는 크게 이미지 인페인팅과 비디오 인페인팅으로 나누어지며, 본 논문에서는 이미지 인페인팅 분야의 최첨단 기술의 문제점과 그에 대한 해결 방법을 제시하고자 하였다. 윤곽 유지 및 내부 문맥 개선이 그것이며, 윤곽 손실 함수와 내부 맥락 고려 모듈을 제안하여 해결책을 제시하고자 하였으며 각각의 방법론에 대해 기존 모델보다 PSNR 과 SSIM 성능 지표에서 상승한 결과를 보였다.

1.서론

인페인팅 (inpainting)은 사람이 눈치채기 어려울 정도로 이미지의 손상된 부분을 복원하는 과정을 의미한다. 이는 과거부터 많은 연구자가 관심을 가졌던 연구 분야이지만, 오늘날에도 완벽한 인페인팅 알고리즘을 만들기는 여전히 어렵다. 인페인팅 분야는 크게 이미지 인페인팅과 비디오 인페인팅으로 나누어질 수 있다.

이미지 인페인팅은 해결 방법에 따라 크게 두 부분으로 나눌 수 있다. 하나는 비 학습 기반 (non-learning based) 방법이라고도 하는 고전적 방법과 학습 기반 (learning based) 방법이라고도 하는 딥러닝 방법이다. 과거부터 고전적 방법에 대한 논문들이 있었지만 [1, 2], 고전적 방법의 한계는 명확하다. 우선, 일부 방법들은 손상된 부위 주변의 정보를 이용하기 때문에 손상된 부위가 너무 크면 인페인팅이 불가능하다. 둘째, 또 다른 일부 방법들은 손상되지 않은 부분의 패치 (patch) 정보를 이용하는데, 만약 이미지의 패치에는 없는 특별한 개체가 손상된 부분에 있는 경우 만족할만한 결과물을 얻기 힘들다. 이러한 이유로, 본 논문에서는 학습 기반 이미지 인페인팅 방법에 대해서 집중할 것이다.

다른 분야인 비디오 인페인팅은 일반적으로 이미지 인페인팅보다 어려운 작업으로 알려져 있다. 비디오의 기본 단위인 프레임 (frame) 이라고 하는 연속된 이미지에 이미지 인페인팅 기법을 적용할 수 있으므로 더 쉬운 작업이라고 생각할 수 있으나, 자연스러운 움직임을 복원하기 위해서는 비디오 인페인팅에서 프레임 간의 시간적, 공간적 정렬에 대한 고려가 추가로 필요하다. 해당 분야도 이미지 인페인팅처럼 학습 기반과 비학습 기반의 두 가지 방법이 있다.

각 분야의 인페인팅 논문들을 설명한 후에, 소개한 이미지 인페인팅의 한계를 개선하는 방법을 설명하며, 측정된 성능들을 보여주고자 한다.

2.관련 연구

2.1 이미지 인페인팅

가장 먼저 소개할 인페인팅 논문은 Pathak *et al.* 의 방법이다 [3]. 전체적인 구조는 자동 부호기 (auto-encoder) 와 유사하지만, 해당 구조만으로는 모델이 이미지에서 의미 있는 특징들을 학습하는 것이 불가능하다고 생각하여, 저자들은 이 문제를 해결하기 위해 두 가지 손실 함수를 추가했었다. 하나는 L2 손실이라고 불리는 재구성 손실이고, 다른 하나는 적대적 생성 신경망 [4]에서 사용되는 적대적 손실이다. 재구성 손실은 이미지의 맥락을 고려해 손상된 영역의 전체적인 구조를 유지시키며, 적대적 손실은 모델이 복구한 부분이 사실적인 이미지처럼 보이도록 지역적 구조를 유지시킨다.

Iizuka *et al.* [5]가 제안한 모델은 하나의 완성 네트워크와 두 개의 판별기 네트워크로 구성되어 있다. 완성 네트워크는 [3]에서 적용되는 완전 연결 네트워크가 아닌 확장된 합성 곱을 적용하여, 많은 계산량을 줄임과 동시에 수용 필드의 시야를 확대한다. 또한 두 판별기 중, 전체적 판별기는 전체 이미지를 입력으로 받고 지역적 판별기는 전체적 판별기와 다른 입력의 크기로 완성된 영역을 받는다. 이를 통해 전체적 판별기는 복구된 이미지가 전체적인 일관성을 가질 수 있도록 하고, 지역적 판별기는 지역적인 일관성을 가지도록 하여 사실적인 결과물을 만든다.

다른 방법인 generative image inpainting with contextual attention [6]은 거친 (coarse) 네트워크와

개선 (refinement) 네트워크를 활용한다. 우선 손상된 이미지를 입력으로 받아, L1 재구성 손실을 이용하여 거친 이미지를 만든다. 개선 네트워크는 거친 네트워크의 결과물을 입력으로 받아, 재구성 손실과 전체적 및 지역적 WGAN-GP [7] 손실을 사용하여 최종 결과물을 완성한다. 또한 이전 모델에서 사용된 합성 곱 신경망은 손상된 부분과 관련 있는 먼 패치의 정보를 반영하기 어렵기 때문에, 이를 문맥 주의 층 (Contextual Attention Layer) 으로 대체하여 모델이 개선 네트워크에서 확장된 합성 곱만 사용할 때보다 더 나은 결과를 얻는다.

Liu *et al.* 이 제시한 방법에서 [8], 저자들은 이전 이미지 인페인팅 방법들의 문제를 개선하려고 노력했었다. 우선 이전 방법들은, 초기 구멍 문제 (initial hole problem) 로 인해 모델이 좋지 않은 텍스처, 심각한 대비 및 인공적인 모양을 생성하는 문제가 있었다. 또한, 일반적으로 이미지의 정중앙에 직사각형의 손상된 영역만을 가정하는 이미지 인페인팅 방법들이 많았었다. 그래서 저자들은 부분적 합성 곱 층 (Partial Convolutional Layer) 을 도입하여 문제를 해결했다. 이미지의 손상된 부분을 복구하기 전에, 부분적 합성 곱 층은 손상된 부분에 대한 재 정규화된 몇 가지 계산을 적용함으로써, 어떠한 손상된 부분의 모양에 대해서도 완벽한 결과를 얻었다.

그러나 Yu *et al.* 의 논문에서는 [9] 각 채널에서 가중치가 없고 0과 1의 경우만 있으며 또한 동일한 손상된 부분을 공유한다는, 부분적 합성 곱이 *hard-gating* 이라는 문제가 있음을 밝혔었다. 또한 부분적 합성 곱은 사용자가 제시한 마스크 (user-guided mask) 에 대해서 결과를 얻을 수 없었다. 그래서 그들은 부분적 합성 곱을 학습할 수 없는 단일 채널 특징 하드 게이팅 (unlearnable single-channel feature hard-gating) 이라고 부르고 각 채널 및 각 공간 위치에 대한 동적 특성 선택 메커니즘을 학습할 수 있는 새로운 합성 곱, 게이트된 합성 곱 (gated convolution) 으로 이 문제를 해결했었다. 또한 SN-PatchGAN (Spectral-Normalized Markovian Discriminator) 을 도입하여 불규칙한 모양과 다양한 수의 마스크를 그리는 작업을 성공시킬 수 있었다. 이러한 종류의 작업을 통해 이미지 인페인팅 및 사용자 안내 인페인팅 (user-guided inpainting) 에서 결과물을 성공적으로 만들 수 있었다.

또 다른 방법인 전경 인식 이미지 인페인팅 (foreground-aware image inpainting) [10]의 전체 구조는 세 가지 모듈인 윤곽 감지 모듈, 윤곽 완성 모듈 및 이미지 완성 모듈로 구성되며, 이미지 인페인팅을 구조 추론과 이미지 완성이라는 두 가지 작업으로 분할시킨다. 먼저 윤곽 검출 모듈에서 손상된 입력을 처리하여 불완전한 윤곽을 만들며, 윤곽 완성 모듈은 이 출력을 이용하여 완전한 윤곽을 만들어 구조 추론을 마친다. 마지막으로 이미지 완성 모듈은 앞선 출력을 이용해 거친 결과물을 만들고, 이를 다시 개선해 최종 결과물을 만든다. 해당 논문의 방식은 윤곽이라는 이미지의 구조적인 정보를 고려하여 기존 방식보다 더 선명한 결과를 도출할 수 있었다.

Xiong *et al.* 의 실험 [10]과 유사하게 EdgeConnect [11]는 윤곽 생성 네트워크와 이미지 완성 네트워크, 총 2 단계로 구성되어 있으며, 앞선 논문과 같이 해당 논문에서는 구조 정보가 주어질 때 이미지 인페인팅의 성능이 더 좋다는 사실을 이용한다. 각 네트워크에는 하나의 생성기와 판별기가 있으며 윤곽 생성기는 회색조 이미지, 회색조 이미지에서 얻은 불완전한 윤곽 이미지와 그에 해당하는 마스크를 입력으로 받아 불완전한 윤곽 이미지의 빈 영역을 완성한다. 이는 손상된 입력 이미지와 결합되어 이미지 완성 네트워크에 의해 최종적으로 완벽한 인페인팅 결과를 만든다.

한편 Ren *et al.* 의 방법 [12]은 윤곽 추론을 이용하여 이미지를 완성하는 이전과 유사한 모델 구조이지만, EdgeConnect [11]의 윤곽 이미지의 한계를 해결했다. 해당 논문에서는 윤곽이 보존된 부드러운 이미지 (Edge-preserved smooth image) 를 이미지의 구조 정보로 사용한다. 이때, 해당 이미지를 얻기 위해 사용되는 Edge-preserved Smooth 방법 [13, 14]은 두드러진 윤곽과 저주파 구조를 유지하면서 고주파 텍스처를 제거하는 방법이며, 이에 따라 구조 생성기가 인페인팅과 관련 없는 픽셀의 방해 없이 이미지의 전체적인 구조를 쉽게 복구할 수 있게 되고, 제거된 고주파 정보는 완성 네트워크인 텍스처 생성기에서 최종적으로 복구하여 복원된 이미지의 질을 높인다.

이러한 좋은 방법에도 불구하고 손상된 부분의 세부적인 부분을 복원하는 것은 어려웠었으며, 연구원들은 이미지 인페인팅을 위한 시각적 구조의 점진적 재구성 (progressive reconstruction of visual structure for image inpainting) [15]에서 누락된 영역의 크기를 줄이기 위해 VSR (Visual Structure Reconstruction) 층을 도입했었다. VSR 층을 사용하여 모델은 윤곽 이미지의 일부를 완성할 수 있으며, 다음 층에서는 이전 층의 출력을 입력으로 받아 이전 VSR 층보다 완성도 높은 윤곽 이미지를 만들 수 있다. 이러한 반복적인 과정으로 인해, 최종적으로 만들어진 윤곽 이미지를 이용해 이미지의 특징을 성공적으로 추출할 수 있게 되었고, 결과적으로 모델이 점진적이며 합리적인 인페인팅 결과를 만들 수 있게 되었다.

Jingyuan *et al.* 의 방법 [16]에서 저자는 영역 식별, 이미지 특징 추론 및 이미지 특징 병합의 세 가지 작업으로 구성되며 다른 네트워크에도 설치할 수 있는 RFR-Net (Recurrent Feature Reasoning Network) 이라는 순환 추론 모듈을 도입했었다. 이는 한 번에 인페인팅을 하는 one-shot fill 방법론의 단점과 점진적으로 인페인팅을 진행하지만, 결과를 반복적으로 유추하지 못하는 기존 방식의 단점에 대한 해결책으로, 모듈은 반복적으로 추론할 수

있으며 KCA (Knowledge Consistent Attention) 가 도입되어 각 반복에 대한 주의 점수 (attention score)를 공유, 최종적으로 모델이 일관된 결과와 우수한 성능을 동시에 생성할 수 있게 되었다.

마지막으로 Xiefan *et al.* 의 방법은 두 개의 하위 작업을 만들어 서로의 정보를 이용한다 [17]. 하나는 구조가 제한된 텍스처 합성에 대한 작업이고, 다른 하나는 텍스처를 기반으로 둔 구조 재구성 작업이다. 전체 모델은 두 작업을 합친 스트림을 병렬 연결하여 재구성된 이미지의 텍스처와 구조를 향상시킨다. 이 과정에서, Bi-GFF (Bi-directional Gated Feature Fusion) 모듈과 CFA (Contextual Feature Aggregation) 모듈이 생성기 끝에 추가되어 거친 이미지를 만들고 이를 정제한다. 또한, 판별기에는 생성기와 유사하게 두 개의 분기가 생성된 텍스처와 구조의 타당성을 별도로 확인한다.

2.2비디오 인페인팅

Ya-Liang *et al.* 의 모델 [18]은 불규칙한 마스크 영역에 참여하는 게이트된 합성 곱 [9]이 충분하지 않기 때문에, 더 나은 결과를 얻을 수 있도록 인접 프레임의 정보를 활용할 수 있는 3D 컨볼루션을 결합한 3D 게이트된 합성 곱으로 생성기가 구성되었다. 또한, 적대적 손실과 마스크의 모양이 정해져 있지 않은 이미지 인페인팅 문제 사이의 균형을 해결할 수 있는 SN-patchGAN [9]이 고화질 비디오 완성을 위한 필수 요소인 프레임 간의 시간적 일관성을 해결할 수 없으므로, 저자들은 판별기에 Temporal PatchGAN 을 이용하여 문제를 해결했었다.

제안 기반 비디오 완성 (proposal-based video completion) [19]에서 모델은 입력 비디오와 마스크를 여러 3D 게이트된 합성 곱[18] 층이 있는 부호기-복호기 구조인 3D 인페인팅 네트워크의 입력으로 받아 거친 질감의 비디오를 만든다. 이후, 제안 생성 네트워크 (proposal generation network) 를 통해 거친 결과를 개선하기 위한 제안 기능을 만들며, 이때 제안을 찾기 위하여 지역적 프레임에서 전역적 프레임 모두를 살펴본다. 마지막으로 모든 제안을 시간에 걸쳐 융합하여 최종 인페인팅 결과를 생성한다.

Chen *et al.* [20]은 광학 흐름과 윤곽 정보를 결합하여 좋은 결과를 만드는 모델을 제안한다. 해당 논문에서 손상된 입력 비디오를 이용하여 손상된 영역이 있는 광학 흐름을 추출한다. 이 출력을 이용하여 캐니 윤곽 검출기 [21]는 불완전한 윤곽 이미지를 만들 수 있으며 EdgeConnect [11]를 기반으로 하는 윤곽 완성 네트워크에 의해 불완전한 부분을 채울 수 있게 된다. 마지막으로 손상된 광학 흐름을 미세한 광학 흐름으로 만드는 과정으로 인하여, 모델은 문제없이 비디오를 성공적으로 복원할 수 있다.

Chen *et al.* 의 방법과 유사한 [20], Rui 의 방법 [22] 또한 광학 흐름을 사용한 모델을 제안했으나, 이 모델의 복원 과정은 거친 결과물을 만들고 그것을 개선하는 방식과 2 스트림 전파 방식으로 구성되어있다. 이때, DFC-Net (Deep Flow Completion Network) 은 손상된 광학 흐름을 복원시키는 역할을 하며, 거친 결과물을 개선하여 최종 광학 흐름을 달성시키기 위한 총 3 개의 DFC-S (Deep Flow Completion Subnetworks) 으로 구성되어 있다. 각 DFC-S 에서, 광학 흐름은 올바른 광학 흐름을 얻기 위해 앞뒤로 전파된다.

Yanhong *et al.* [23]은 비디오 인페인팅 문제를 손상된 입력 비디오의 빈 영역을 그리기 위한 핵심 정보로 전체적 및 지역적 프레임을 사용하는 다중-다중 문제로 간주하기 위해 변환기 (transformer) [24]의 구조를 사용한다. 이 과정에서 모델은 공간적 차원과 시간적 차원을 고려하여 패치의 유사한 맥락을 찾는데, 이는 다양한 크기를 가진 패치를 기반으로 한 주의 모듈 (attention module) 에 의해 수행되며, 다중-헤드 자가 주의 (Multi-head Self Attention) 모듈 덕분에 모델은 서로 다른 크기의 패치 간의 유사성을 파악할 수 있다.

3. 최첨단 모델 개선

우리는 이미지 인페인팅 방법의 최첨단 방법의 하나인 CR-Fill [25]을 기반으로, 최첨단 방법의 문제점을 해결하여 해당 모델의 성능을 개선하려고 한다.

3.1 CR-Fill

적대적 생성 신경망의 접근법을 사용하고 있는 CR-Fill [25]의 생성기 구조는 그림 1 과 같으며, 거친 결과물을 만들고 이를 개선하는 방법을 사용하고 있다. 해당 모델에서는 이전 방법과 마찬가지로 수용 필드의 크기를 확대하기 위해 확장된 합성 곱을 적용하고, 마스크의 불규칙한 모양을 처리하기 위해 게이트된 합성 곱을 추가했다 [9]. 원본 이미지에서 생성된 마스크와 손상된 이미지는 흐릿한 이미지를 만들기 위해 거친 네트워크에 대한 입력이 되며, 해당 결과물은 결과의 품질을 향상시키기 위해 정제 네트워크로 전달된다.

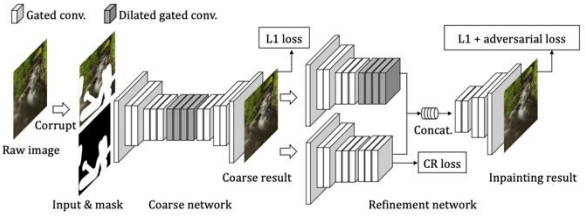

그림 1. 생성기 네트워크의 전체적인 구조

DeepFillv2 [9]와 유사한 구조인 생성기에는 맥락별 주의 층 (Contextual Attention layer)이 없고

대신 맥락별 재구성 손실 (Contextual Reconstruction Loss)이 있으며, 이는 CR-Fill 의 핵심이다. 맥락별 주의 층이 잘못된 패치를 짝지어서 인공적인 패턴을 만드는 경우가 있으므로, 맥락별 재구성 손실은 L1 손실과 결과의 적대적 손실의 영향을 줄임으로써 최적의 결과물을 생성시킨다. 그림 2 는 맥락별 재구성 손실의 개요를 보여준다.

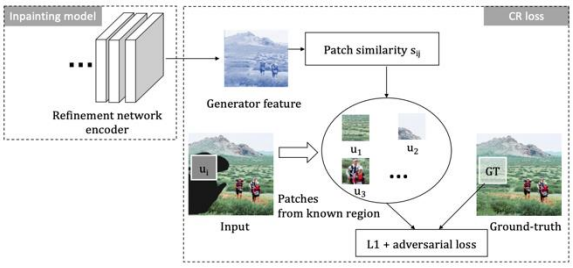

그림 2. 맥락 재구성 손실의 개요

또한 CR-Fill 은 PatchGAN [26]에 스펙트럼 정규화가 적용된 SN-PatchGAN 판별기 [9]를 사용하고 있다.

3.2 한계점

CR-Fill 에는 두 가지 문제가 있으며, 우리는 이 문제들에 집중했다. 첫째, CR-Fill 은 선명한 윤곽 정보를 만드는 데 문제가 있다. 그림 3 의 빨간색 원은 사람을 마스킹한 왼쪽 이미지가 입력으로 주어졌을 때 인페인팅된 결과이다. 자세히 보면 개의 귀 근처의 윤곽이 흐릿하며, 이처럼 인페인팅 결과물의 윤곽을 뚜렷하게 하지 못하는 경우가 있다.

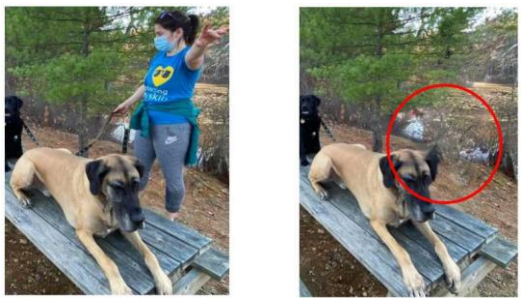

그림 3. 윤곽이 뚜렷하지 않은 문제의 경우

둘째, CR-Fill 이 이미지의 전체적 맥락에 비해 일관된 인페인팅 결과를 만들지 못하는 경우가 있다. 그림 4 의 빨간색 원을 보면, 일부 픽셀은 근처의 픽셀들과 조화되지 않는 색을 가지고 있음을 확인할 수 있다.

그림 4. 일관성 없는 색 문제의 경우

3.3 개선 방법

이러한 심각한 문제들을 해결하기 위한 해결 방법을 소개하고 세부 사항을 구체적으로 설명한다.

3.3.1 윤곽 손실 함수

윤곽 손실 함수는 구조 정보를 생성하는 일부 모델에서 영감을 받았다 [10, 11, 12, 15]. 먼저 모델은 기존 CR-Fill 에서 캐니 윤곽 검출기 [21]를 이용하여 원본 이미지의 윤곽선을 감지한다. 이 상황에서 문제가 되는 경우는 하나의 윤곽선을 기준으로 양옆의 픽셀들의 값 차이가 적은 경우이며, 해당 경우에는 이미지의 윤곽선이 흐릿하게 보일 것이다. 따라서 우리는 해당 문제를 두 부분의 픽셀값 차이가 일정 이상이 되도록 손실 함수를 이용하여 경계를 회복시킨다. 윤곽 손실 함수의 예는 그림 5 에 나와 있다.

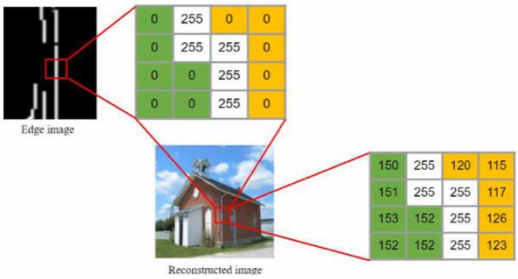

그림 5. 윤곽 손실 함수의 예제

윤곽 손실의 계산은 이미지에서 윤곽 및 픽셀값을 가져오는 것으로 시작된다. 영역에 대한 윤곽의 픽셀값으로 윤곽 부분은 255 로 채워지고 나머지 부분은 0 으로 채워진다고 가정한다. 이를 통해 윤곽 부분과 윤곽이 아닌 부분을 알 수 있고, 윤곽선으로 인해 두 영역으로 구분된다. 두 영역으로 분할되었으면, 인페인팅한 이미지에서 사용할 수 있도록 좌표를 저장한다. 그다음 두

영역의 각 좌표와 재구성된 이미지의 영역을 사용하여 각각 두 하위 영역에 대한 평균 픽셀 값을 구한 후, 두 값의 차이가 임계 값보다 크도록 학습시킨다. 윤곽 손실 함수 방정식은 다음과 같다:

$$Edge\ loss\ = \frac{1}{N}\sum ReLu(|threshold - x_i - x_j|_1) \quad (1)$$

$$x_i = \frac{\Sigma_{t\epsilon region\ A}\ pixel}{number\ of\ pixel\ in\ region\ A} \quad (2)$$

$$x_j = \frac{\Sigma_{t\epsilon region\ B}\ pixel}{number\ of\ pixel\ in\ region\ B} \quad (3)$$

아래의 그림 6 은 윤곽 손실 함수가 추가된 CR-Fill 의 구조이다.

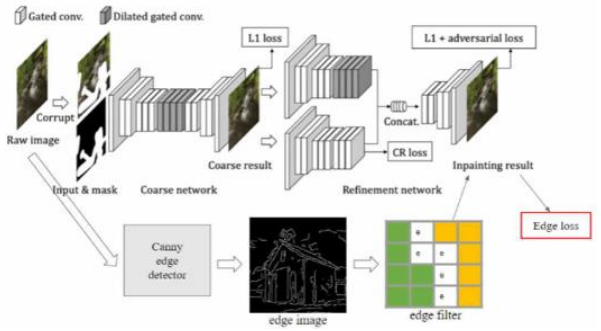

그림 6. 윤곽 손실 함수가 추가된 구조

3.3.1.1 전체적 판별기

앞서 설명한 것처럼, CR-Fill 에 의해 인페인팅된 이미지가 지역적으로 그럴 듯하더라도 전체적인 맥락과 맞지 않을 수 있다. 이 문제는 이미지에서 손상된 부분의 비율이 증가할수록 발생하기 쉽다. 이는 CR-Fill 이 지역적 판별기만을 사용하여 부분적인 맥락만 고려하기 때문에 발생하는 문제이며, 이 문제를 해결하기 위해 기존 CR-Fill 에 전체적 판별기를 추가하고 두 판별기에 가중치를 다르게 하여 지역적인 맥락과 전체적인 맥락 모두를 고려하여 성능을 향상시키도록 했다. 이에 대한 방정식은 다음과 같다:

$$L_D = 0.9L_{LD} + 0.1L_{GD} \quad (4)$$

3.3.2 내부 문맥 고려

내부 문맥 고려는 앞의 3.2 절에서 설명했던 내용들이 인페인팅된 부분의 문맥을 고려하지 않아 생긴 결과라고 생각하여, 이를 해결하기 위해 내부 문맥을 고려할 수 있는 pix2pixHD [27]의 방법론을 이용하여 문제를 해결하려 한다. 이는 조건이 있는 적대적 생성 신경망 (condition Generative-adversarial network, cGAN)을 활용하여 추가적인 윤곽 정보를 이용해 고해상도의 이미지를 유지하게 만드는 방법이다. 또한 크기가 작고 큰 영역을 모두 잘 알기 위해, 생성기의 경우 특징 피라미드 구조를 다운 샘플링 (down-sampling)과 잔차 블록 (residual block)을 사용하며, 이러한 방법론을 개선 네트워크로 활용함으로 성능을 향상시키려고 했다. 해당 방법론이 추가된 개요는 그림 7 에서 확인할 수 있다.

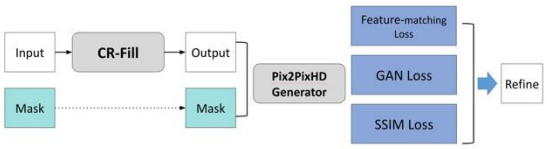

그림 7. 내부 맥락 모델이 추가된 모델 개요

CR-Fill 을 거쳐서 나온 인페인팅 결과와 마스크 이미지는 마스크 부분만 pix2pixHD 생성기의 입력으로 들어가 향상된 결과물로 나오는데, 이에 patchGAN [26] 손실함수와 특징 매칭 손실함수를 논문 [27]과 같게 사용했으며, 이에 구조적 유사성 손실함수를 추가했다.

4. 실험 결과

4.1 윤곽 손실 함수 결과

해당 실험에서 사용되는 이미지는 Places2 [29]데이터 셋의 이미지이며, 무작위로 300 장을 뽑아 모델을 훈련시켰다. 훈련시킬 때 같이 들어가는 마스크 이미지는 인터넷에서 사람의 모습이 들어가 있는 사진을 선택하여, 사람의 형상만 따로 추출하여 생성했다.

최종적으로 모델에 들어가는 이미지는 그림 8 의 왼쪽 두 이미지인 원본 이미지와 마스킹 이미지가 합쳐진 오른쪽 이미지가 입력으로 들어가게 된다. 이때, 사용된 Places2 이미지와 마스킹 이미지의 크기는 512x512 의 정사각형 이미지이다.

그림 8. 데이터 생성 과정 예시

4.1.1 정성적 평가

그림 9 는 CR-Fill 모델에 윤곽 손실 함수만을 추가한 인페인팅 결과이다. 이는 자연경관뿐만 아니라 기존 CR-Fill 모델에서는 잘 작동하지 않던 도시나 공장 등의 환경 이미지에도 잘 적용됨을 확인할 수 있었다. 특히 윤곽이 존재하는 부분에 마스크를 적용해도 잘 인페인팅 되는 결과를 얻을 수 있었다.

(a) Input (b) CR-Fill (c) Ours

그림 9. 윤곽 손실 함수를 추가한 결과

그림 10 은 CR-Fill 모델에 전체적 판별기만을 추가하여 훈련시킨 후 실험한 결과이다. 모델은 마스킹된 부분을 복구할 때 전체 이미지의 컨텍스트와 일치하지 않는 CR-Fill 문제를 해결할 수 있었으며, 특히 마스킹된 부분이 주변 환경에 잘 조화로운 상태인 결과를 얻을 수 있었다.

(a) Input (b) CR-Fill (c) Ours

그림 10. 전체적 판별기를 추가한 결과

그림 11 은 CR-Fill 모델에 윤곽 손실 함수와 전체적 판별기를 모두 추가하여 훈련시킨 후 실험한 결과이다. 윤곽 손실 함수만 적용하였을 때 윤곽이 강조되어 주변 맥락과 맞지 않는 결과가 나왔으나, 이러한 문제는 전체적인 맥락을 고려한 전체적 판별기를 적용함으로써 어느 정도 완화될 수 있었다.

(a) Input (b) CR-Fill + E.L (c) Both

그림 11. 윤곽 손실 함수, 전체적 판별기를 추가한 결과

4.1.2 정량적 평가

해당 실험에서는 최대 신호 대 잡음비 (Peak Signal-to-Noise Ratio)와 구조적 유사도 지수 (Structural Similarity Index Map)를 이용하여 각 모델의 성능을 측정했다. CR-Fill 에 윤곽 손실 함수만 적용하면 PSNR 이 기본 CR-Fill 에 비해 약 0.02 가 낮아졌지만, SSIM 은 약 0.01 이 높아졌다. PSNR 은 손실 정보가 최소일 때 큰 값을 가지므로, 윤곽을 강조하면서 복구를 수행시키면 PSNR 값이 낮아질 수 있지만 윤곽이 조금 더 선명해져서 SSIM 값이 올랐다.

CR-Fill 에 전체적 판별기만 적용하면 PSNR 이 기본 CR-Fill 에 비해 약 0.02 증가하며, SSIM 은 0.01 정도 상승했다. 이는 전체 맥락을 기반으로 손상된 부분을 복구하기 때문에 더 높은 품질의 출력을 가질 수 있기 때문이다.

마지막으로 윤곽 손실 함수와 전체적 판별기 모두를 CR-Fill 에 적용했을 때, 윤곽 손실 함수에 의해 PSNR 은 기본 CR-Fill 의 결과보다 약간 낮은 값을 가졌지만, SSIM 은 전체적 판별기로 인하여 윤곽 손실 함수만 적용했을 때보다 약간 큰 값을 가졌다.

표 1. 윤곽 손실 함수에 대한 PSNR, SSIM 결과

방법	PSNR	SSIM
CR-Fill	36.1745	0.8587
CR-Fill + 윤곽 손실 함수	36.1540	0.8623
CR-Fill + 전체적 판별기	36.1910	0.8628
CR-Fill + 윤곽 손실 함수 + 전체적 판별기	36.1696	0.8626

4.2 내부 문맥 고려 결과

해당 실험은 Paris Street [28] 이미지를 10 만 장을 기본으로 훈련했으며, 공장 이미지를 잘 인페인팅하지 못하는 점을 고려하여 5088 장의

이미지를 추가했다. 추가된 이미지에는 공장 관련 이미지가 3548 장, 사무실 관련 이미지는 750 장, 자연 관련 이미지는 800 장이 추가되었다. 모델의 입력에 관한 정보는 4.1 절에서 설명했던 윤곽 손실 함수의 입력과 같다.

4.2.1 정성적 평가

그림 12 는 원본 이미지와 내부 문맥을 고려할 때에 대한 결과를 보여준다. 원본 이미지에 비해서 흠이 없는 결과물을 보여주고 있다.

그림 12. 원본 이미지(왼쪽)와 내부 문맥 고려의 결과(오른쪽)

4.2.2 정량적 평가

해당 실험은 4.1.2 절과 같은 측정 기준을 가지고 실험했으며, 결과는 내부 문맥을 고려한 모델이 기본 CR-Fill 의 최대 신호 대 잡음비보다 약 0.02 높은 성능을 얻었고, 구조적 유사도 지수도 0.01 만큼 향상된 성능을 얻었다. 이는 4.1 절에서 확인할 수 있는 성능을 포함하여, 가장 좋은 성능을 보여주었다.

표 2. 내부 문맥 고려의 PSNR, SSIM 결과

방법	PSNR	SSIM
CR-Fill	36.1745	0.8587
CR-Fill + 내부 문맥 고려	36.1988	0.8645

5. 결론

본 논문에서는 이미지 인페인팅과 비디오 인페인팅의 최신 기술을 살펴보고, 최신 기술 중 하나를 기반으로 성능 향상 방법을 제안하고 측정된 성능을 공개했다.

감사의 글

본 연구는 울산과학기술원의 CCTV 내 현장 작업자의 고성능, 실시간 검출 및 소거 기술 연구 과제 (2.210894.01)의 지원을 받아 수행되었습니다.

참고문헌

[1] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester. “Image inpainting” In SIGGRAPH, 2000.
[2] Bertalmio, L. Vese, G. Sapiro, and S. Osher. “Simultaneous structure and texture image inpainting” IEEE transactions on image processing, 2003.
[3] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A Efros. “Context encoders: Feature learning by inpainting” In CVPR, 2016.
[4] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. “Generative adversarial nets” In NIPS, 2014.
[5] S. Iizuka, E. Simo-Serra, and H. Ishikawa. “Globally and locally consistent image completion” ACM Transactions on Graphics, 2017.
[6] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S Huang. “Generative image inpainting with contextual attention” In CVPR, 2018.
[7] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville. “Improved training of wasserstein gans” arXiv preprint arXiv:1704.00028, 2017.
[8] G. Liu, F. A Reda, K. J Shih, T. Wang, A. Tao, and B. Catanzaro. “Image inpainting for irregular holes using partial convolutions” In ECCV, 2018.
[9] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S Huang. “Free-form image inpainting with gated convolution” In ICCV, 2019.
[10] W. Xiong, J. Yu, Z. Lin, J. Yang, X. Lu, C. Barnes, and J. Luo. “Foreground-aware image inpainting” In CVPR, 2019.
[11] K. Nazeri, E. Ng, T. Joseph, F. Qureshi, and M. Ebrahimi. “Edgeconnect: Generative image inpainting with adversarial edge learning” In ICCVW, 2019.
[12] Y. Ren, X. Yu, R. Zhang, T.H. Li, S. Liu and G. Li. “StructureFlow: Image inpainting via structure-aware appearance flow” In ICCV, 2019.
[13] L. Xu, C. Lu, Y. Xu, and J. Jia. “Image smoothing via l0 gradient minimization” In SIGGRAPH, 2011.
[14] L. Xu, Q. Yan, Y. Xia, and J. Jia. “Structure extraction from texture via relative total variation” ACM Transactions on Graphics, 2012.
[15] J. Li, F. He, L. Zhang, B. Du, and D. Tao. “Progressive reconstruction of visual structure for image inpainting” In ICCV, 2019.
[16] J. Li, N. Wang, L. Zhang, B. Du, and D. Tao. “Recurrent feature reasoning for image inpainting” In CVPR, 2020.
[17] X. Guo, H. Yang, and D. Huang. “Image inpainting via conditional texture and structure dual generation” In ICCV, 2021.
[18] Y. Chang, Z. Yu Liu, K. Lee, and W. Hsu. “Free-form video inpainting with 3D gated convolution and temporal patchgan” In ICCV, 2019.
[19] Y. Hu, H. Wang, N. Ballas, K. Grauman, and A. G Schwing. “Proposal-based video completion” In ECCV, 2020.
[20] C. Gao, A. Saraf, J. Huang, and J. Kopf. “Flow-edge guided video completion” In ECCV, 2020.

[21] C. John. “A Computational Approach to Edge Detection.” IEEE Transactions on Pattern Analysis and Machine Intelligence, 1986
[22] R. Xu, X. Li, B. Zhou, and C. Change Loy. “Deep flow-guided video inpainting” In CVPR, 2019.
[23] Y. Zeng, J. Fu, and H. Chao. “Learning joint spatial-temporal transformations for video inpainting” In ECCV, 2020.
[24] V. A. Shazeer, N. Parmar, N. Uszkoreit, J. Jones, L. Gomez, A.N. Kaiser, L. Polosukhin, I. “Attention is all you need” In NIPS, 2017
[25] Y. Jiahui, L. Zhe, Y. Jimei, S. Xiaohui, L. Xin, and H. S. Thomas. “Generative image inpainting with contextual attention” In CVPR, 2018.
[26] P. Isola, J.Y. Zhu, T. Zhou, and A. A. Efros. “Image-to-image translation with conditional adversarial networks” In CVPR, 2017.
[27] M. Berning, K. M. Boergens, and M. Helmstaedter. “SegEM: Efficient Image Analysis for High-Resolution Connectomics” Neuron, 2015.
[28] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. Efros. “What makes paris look like paris?” ACM Transactions on Graphics, 2012.
[29] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. “Places: A 10 million image database for scene recognition” IEEE transactions on pattern analysis and machine intelligence, 2017.