Journal of The Institute of Electronics and Information Engineers Vol.58, NO.6, June 2021 https://doi.org/10.5573/ieie.2021.58.6.41
ISSN 2287-5026(Print) / ISSN 2288-159X(Online)

논문 2021-58-6-5

시각 인지 특성을 반영한 생성적 이미지 인페인팅 기술

(Generative Image Inpainting Method Reflecting Human Perceptual Characteristic)

오 희 석*, 최 원 석*

(Heeseok Oh and Wonsuk Choi©)

요 약

이미지 인페인팅은 의도적 혹은 비의도적으로 발생한 손실 영역을 자동으로 복원하는 기술이며, 특히 편집 관점에서는 사용자가 직접 특정 객체를 제거하기 위한 목적으로 널리 활용된다. 고전적인 확산 및 패치 기반 접근과 더불어, 최근에는 타 컴퓨터 비전 분야와 마찬가지로 심층신경망을 활용한 기술들이 소개되었다. 하지만 기존 CNN(convolutional neural network) 및 GAN(generative adversarial network) 기반 인페인팅 결과는 블러나 시각적 결함(artifact)과 같은 화질 저하 요소를 빈번히 포함한다. 이는 인페인팅 시나리오에 있어 전역적 의미 추론을 위해 CNN을 통해 특징 공간에 임베딩된 표현이, 손실 영역 복원을 위한 구조적 정보와 고주파 성분으로 동시에 디코딩되기 어려움을 의미한다. 본 논문에서는 이러한 한계를 극복하고자 인간의 시각 인지 특성을 반영한 두 단계의 인페인팅 과정을 제안한다. 제안 방법은 두 개의 GAN으로 구성되어 있으며, 첫 번째 생성망은 시각 정보의 일차시각피질 전달 과정을 모방한 MSCN(mean subtracted contrast normalized) 계수를 생성한다. 두 번째 생성망은 이미지 완성을 위해 생성된 MSCN 계수를 사전 구조 정보로 이용함으로써 손실 영역이 복원된 RGB 이미지를 생성한다. 가중치를 공유하지 않는 두 독립적인 생성망은 각각의 판별망에 적대적으로 end-to-end 학습되며, 두 단계에 걸친 추론을 통해 손실 영역의 단순한 구조 뿐 아닌 미세 성분까지 효과적으로 복원이 가능하다. CelebA 데이터셋을 활용한 실험을 통해 제안하는 시각 인지 특성을 반영한 생성적 인페인팅 방법이 넓은 비규칙적 손실 영역에 대해서도 기존 기술 대비 정성적/정량적으로 우수한 성능을 나타냄을 확인하였다.

Abstract

Image inpainting represents a technique of recovering the content in missed spatial area, and which is widely being employed to discard an unwanted object in terms of image adjustment. In recent, beyond the conventional diffusion- and patch-based approaches, various deep learning-based schemes have been introduced as the other computer vision tasks. However, inpainting methods with utilizing a CNN (convolutional neural network) and GAN (generative adversarial network) generally resulted in the blurred image or local artifacts. That is, previous methods cannot effectively decode the embedded features for abstracting global semantic representation to RGB image domain implying contextual structure and fine-details simultaneously. To cope with this problem, we propose a two-stage inpainting scheme reflecting the human's perceptual characteristic. The proposed method consists of two GANs. The first generative network aims for generation of MSCN (mean subtracted contrast normalized) coefficients which resembles a visual conveyance process from the eyes to the primary visual cortex (area V1) in the brain. The second generative network completes RGB image which refers the generated MSCN coefficients as a prior structural information. Non-shared two generative networks with each corresponding discriminator are trained in end-to-end manner, and the result argues that the proposed two stage method can sufficiently recover both structure and high-frequencies of the missed regions. The experiments are performed on CelebA dataset, and the quantitative results show that the proposed generative inpainting scheme reflecting perceptual characteristic outperforms the previous methods regarding visual quality even in a large irregular hole scenario.

Keywords : Image inpainting, perceptual characteristic, MSCN coefficient, generative adversarial network

*정회원, 한성대학교 IT융합공학부(Department of IT Convergence Engineering, Hansung University)
©Corresponding Author(E-mail : wonsuk@hansung.ac.kr)
※ 이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2020R1G1A1100674)

Received ; February 21, 2021 Revised ; March 11, 2021 Accepted ; March 18, 2021

Ⅰ. 서 론

이미지 인페인팅(inpainting)은 획득 및 전송 시, 비의도적 원인으로 야기된 이미지 내 손상 혹은 사용자가 의도적으로 삭제하고자 수동으로 지정한 불필요 객체를 손실 영역(missing region)이라 정의하고 해당 손실 영역을 자동으로 채우는 것을 목표로 하는 이미지 복원 분야의 일종이다. 이미지 인페인팅 기술은 360° 파노라마 영상 합성, 양안식/가상현실 디스플레이를 위한 다시점 영상 생성, 또는 신뢰성 높은 3차원 복원을 위한 전처리 등 각종 응용 분야에서도 활발히 이용되고 있다[1, 2]. 특히 영상의 편집 관점에서 사용자가 의도하지 않은 객체를 임의로 지정하고 삭제함으로써 손실 영역을 이미지에 강제한 뒤 해당 손실 영역을 배경으로 새롭게 복원하는 형태로 활용되고 있으며, 해당 시나리오에서 자동으로 복원된 영역은 단순한 패턴의 복사가 아닌 인간 시각 및 의미론적 관점에서 충분한 화질의 보장이 뒷받침되어야 한다. 하지만 손실 영역에 내재해야 할 의미적(semantic) 모호함과 자연 영상이 갖는 구조적 복잡성으로 인해 그 난이도가 높으며, 이에 이미지 인페인팅은 여전히 컴퓨터비전 내에서도 도전적인 분야로 여겨진다.

전통적으로 이미지 인페인팅 기술은 두 가지 접근 방법을 통해 실현되었다. 첫 번째 방법은 확산(diffusion) 기반의 텍스쳐 합성에 관한 접근으로, 손실 영역 주변특정 범위 내 배경부 텍스쳐 정보를 확장하거나 픽셀 간 보간을 통해 이루어진다[3, 4]. 두 번째 접근 방법은 패치(patch) 기반으로 이루어지며, 수집한 자료 영역(source region)에서 가장 높은 유사도를 나타내는 패치를 찾아내고 이를 복사함으로써 손실 영역을 복구하는 형태이다[5, 6]. 상기 두 가지 접근들은 손실 영역 인근 혹은 특정 자료 영역에서의 픽셀 성분과 복원될 영역의 픽셀 성분이 서로 의미적 상관관계가 존재한다는 가정하에 이루어지나, 대부분의 실제 이미지는 비규칙적 구조를 가짐으로 인해 전통적 접근을 통해서는 손실 영역 복원에서의 한계가 명확하였다.

근래에는 심층신경망의 눈부신 발전에 힘입어 여타의 컴퓨터비전 응용 분야와 마찬가지로 콘볼루션 신경망(CNN; convolutional neural network)을 이용해 이미지 인페인팅을 수행하고자 하는 연구가 다수 소개되었으며 비약적인 성능의 발전을 달성하였다[7, 8]. CNN 기반 인페인팅은 손실 영역 주변의 지역적(local) 특징만을 고려하는 것이 아닌, 이미지 전체의 전역적(global) 의미를 이해하고 대량의 데이터셋으로부터 이미지 텍스쳐 및 구조의 분포를 학습하여 이를 통해 손실 영역을 추론하고자 하는 시도로 해석할 수 있다. CNN 기반의 인페인팅은 빠른 속도로 이미지의 전역적 의미와 일관된(coherent) 구조의 복원이 수행 가능하나, 결과물이 블러(blur)하거나 종종 중대한 결함(artifact)을 유발한다는 시각적 화질 관점에서의 단점이 존재한다. 이는 CNN 기반 인페인팅 기술이 차용하는 인코더-디코더 형태의 심층신경망 구조에서 기인하며, 단순히 전역적 정보 해석을 위한 수용장(receptive field) 확장과 저차원 특징으로의 압축 및 임베딩이 손실된 이미지 내 고주파 성분 복원에 효율적이지 않음을 방증한다.

본 논문에서는 손실 영역의 미세(fine detail) 성분을 효과적으로 복원하고자 인간 시각의 인지적 특성을 반영한 두 단계에 걸친 생성 기반의 이미지 인페인팅 기술을 제안한다. 제안하는 모델은 두 개의 적대적생성신경망(GAN; generative adversarial network)으로 구성되어있으며 첫 번째 신경망은 손실 영역에 대한 구조적 정보 생성을 목표로 MSCN(mean subtracted contrast normalized) 계수를 생성한다. MSCN 계수는 인간의 시각 체계를 반영한 공간적 인지 특성을 수치화한 것으로, 이를 활용한 영상처리 기법은 이미지 분류 혹은 화질 평가 분야에서 우수한 결과를 도출하여왔다[9, 10]. 두 번째 신경망은 생성된 MSCN 계수를 토대로 손실 영역의 RGB 픽셀 값을 복원하는 이미지 완성(completion) 신경망이며, MSCN 계수가 표현하는 이미지의 구조 정보를 토대로 공간적 고주파 정보를 포함하는 보다 고품질의 색상 및 텍스쳐의 복원을 목표로 동작한다. 가중치를 공유하지 않는 두 개의 독립적인 생성망은 end-to-end로 학습되며 추론 시에는 MSCN 계수 및 복원 이미지를 동시에 생성한다. 실험에서는 제안하는 두 개의 생성망을 통해 MSCN 계수를 거쳐 복원된 손실 영역이 시각적으로 향상된 화질을 보장함을 실험적으로 확인하였다.

Ⅱ. 선행 연구 및 한계

1. 전통적 인페인팅 기술

고전적인 확산 기반의 인페인팅 기술은 손실 영역 부근의 이웃 픽셀로부터 타겟으로의 구조적 정보를 전파함으로써 이미지를 복원함이 목적이었으며 이미지 공간의 거리변환(distance transform)과 같은 수학적 해석이 도입되기도 하였다[3, 11]. Esedoglu 등은 오일러의 탄성 이론과 같은 역학적 특성을 영상처리 분야에 응용하여

인페인팅을 위한 손실 영역으로의 외형(appearance) 확산을 시도하였다[4]. 하지만 단순히 주변 정보에 의존하는 기술 특성상 이러한 방법은 색상이나 텍스쳐 변화가 상대적으로 적은 영상에서의 국소적인 손실 영역만을 효율적으로 다룰 수 있는 한계가 있었다.

패치 기반의 인페인팅은 영상 내 비손실 영역 내에서 복원을 위한 적절한 패치를 반복적으로 찾아내는 방법이다. 이 과정에서 복사된 픽셀 간 공간적 불연속성(discontinuity)의 제거를 위해 자료 영역에서 수집된 다수 정보가 블렌딩 과정을 거치기도 한다[5, 6]. 패치 기반의 방법은 상기 과정의 반복적 계산 수행에 기인하여 복원에 요구되는 연산량이 높으며 처리 속도 또한 매우 느리다. 연산 복잡도의 단점을 해결하고자 Barnes 등은 자료 영역에서의 적절한 패치 검색 속도를 향상시킨 PatchMatch 방법을 제안하기도 하였다[12]. 하지만 그럼에도 패치 기반의 방법은 전체적인 이미지의 의미와는 무관한 패치 단위의 선택을 수행하는 까닭에 고품질의 복원 결과를 도출하는데 한계가 존재한다.

2. 학습 기반 인페인팅 기술

이미지 인페인팅 분야에서도 심층신경망의 출현과 함께 CNN 및 GAN을 활용한 손실 영역 복원의 연구가 활발히 수행되었다. Pathak 등은 오토인코더(autoencoder)와 유사한 구조의 인페인팅 신경망을 제안하였다[13]. 인코더는 손실 영역을 포함하는 입력을 저차원의 잠재 특징으로 표현함으로써 전체적인 이미지의 맥락(context)을 학습한다. 디코더는 임베딩된 특징 벡터로부터 맥락에 부합하는 손실 영역을 생성하며, 본 연구를 통해 GAN 기반 인페인팅의 가능성이 제시되었다. Yang 등은 이러한 GAN 기반 인페인팅 결과를 바탕으로 자료 영역의 텍스쳐 정보와 손실 영역 간 특징 공간에서의 거리를 줄이기 위한 사전 학습 신경망을 추가하여 복원 시의 성능 향상을 보였다[14]. Iizka 등은 dilated 콘볼루션을 이용한 전연결층이 없는 생성망을 구성하였으며, 전역 정보와 지역 정보에 따른 적대적 손실 함수를 위한 판별 과정에 두 개의 신경망을 별도로 구성함으로써 고품질의 인페인팅을 수행하고자 하였다[15].

상기 기존 인코더-디코더 기반 인페인팅 기술의 문제점은 자료 영역으로부터 손실영역에 대한 의미론적 구조를 추론함과 동시에 텍스쳐 정보를 한 번에 구성하려 함으로써 발생하는 블러한 결과물과 각종 시각적 결함이었다. 이러한 한계를 극복하고자 Nazeri 등은 두 개의 생성망을 두어 손실 영역의 구조 추론과 이미지 완성 과정을 분리하였다[16]. 첫 번째 생성망은 손실 영역의 엣지(edge) 성분을 추론하며, 두 번째 생성망은 구조를 의미하는 엣지라는 사전(prior) 정보를 활용하여 손실 영역을 채우는 역할을 수행한다. Ren 등은 역시 이와 유사하게 두 단계에 걸친 손실 영역의 복원을 수행하였으며, 이미지 완성을 위한 사전 정보로 엣지가 아닌 평활화된 형태의 구조 정보 흐름(flow)을 추론한다는 점에 차이가 있다[17].

이렇듯 GAN 기반 인페인팅의 품질을 향상시키려는 시도들이 존재하였으나, 시각적 화질 관점에서 엣지 혹은 블러 영상은 미세 수준의 텍스쳐 디테일을 복원하기 위한 구조적 단서로써 충분하지 않은 경향을 보인다. 이를 극복하고자 본 논문에서는 인지적 특성을 반영한 MSCN 계수를 사전 정보로 활용하고, 이를 통해 이미지 내 시각적으로 민감한 부분의 고주파 성분을 보다 효과적으로 복원하는 기술을 제안한다.

Ⅲ. 제안 기술

본 연구에서 제안하는 이미지 인페인팅 방법의 전체적인 프레임워크는 그림 1과 같다. 가중치를 공유하지

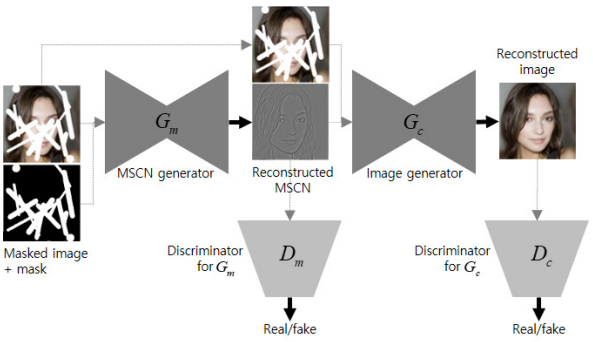

그림 1. 제안하는 이미지 인페인팅 프레임워크
Fig. 1. Overall framework of the proposed image inpainting scheme.

않는 두 개의 독립적인 GAN이 상호 보완적으로 학습되며, 따라서 전체 구성은 총 4개의 신경망으로 구성되어있다. MSCN 계수 생성망 G_m은 손실된 이미지와 손실 영역에 대한 마스크를 입력으로 받아 MSCN 계수 이미지를 복원한다. G_m은 신뢰성 있는 MSCN 계수 생성을 위해 학습 과정에서 판별망 D_m에 적대적으로 MSCN 계수 데이터셋의 분포를 학습한다. G_m을 통해 생성된 MSCN 계수 이미지는 복원되어야 할 손실 영역의 구조적 정보를 포함하고 있다. 손실된 이미지와 함

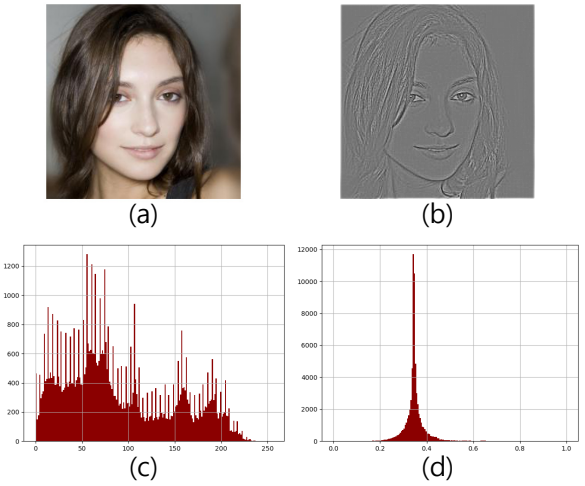

그림 2. (a) RGB 이미지 (b) MSCN 계수 이미지 (c) 이미지 밝기 성분의 히스토그램 (d) MSCN 계수 이미지 히스토그램

Fig. 2. (a) RGB image. (b) MSCN coefficient image. (c) Histogram of image brightness. (d) Histogram of MSCN coefficient image.

께 다시 이미지 완성 신경망 G_c로 입력되며 G_c 역시 판별망 D_c와 적대적으로 학습된다. 최종적으로 G_c를 통해 손실 영역이 복원된 완성 이미지를 획득 가능하다.

1. Mean Subtracted Contrast Normalized (MSCN) 계수

MSCN 계수는 인간 시각의 지역적 인지 특성을 반영한 정규화된 계수이다. 인간의 시각 인지는 안구의 광학적 특성을 통해 망막에 맺힌 정보를 전달함으로써 시작된다. 대부분의 이미지들은 자연 장면 통계(NSS; natural scene statistics) 특성을 띄며, 이는 각 픽셀이 국소적으로 인근에 있는 픽셀들과 높은 상관성을 가짐을 의미한다[9]. 인간 시각 체계는 인지적 판단을 위해 망막에 맺힌 이미지의 지역 정보를 시교차 후 LGN을 통해 일차시각피질(primary visual cortex) V1 영역으로 전달한다. 이 과정에서 시각 체계는 상기 서술한 NSS 특성을 띄는 지역적 이미지 영역에서의 중복을 제거하고 인지적으로 가치있는 정보만을 전달하고자 하며, 이는 일종의 생체인지적 지역 정규화 과정으로 설명 가능하다[18]. 이러한 시각적 특성의 수학적 모델링을 위해 MSCN 계수가 제안되었으며, 실제 영상 화질 향상 및 이미지 인식과 같은 다양한 분야에서 MSCN 계수 이용을 통해 성능의 향상을 이루었다[9, 10, 19].

픽셀 좌표 (x, y)에 대해 이미지 도메인 I로부터 MSCN 계수로의 변환은 지역 평균 μ와 대비도(contrast) σ의

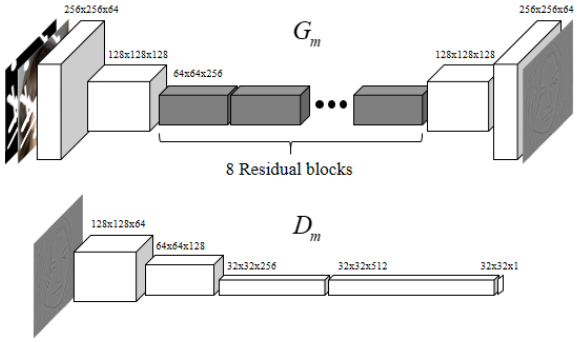

그림 3. 손실된 이미지와 마스크를 입력받아 MSCN 계수를 생성하는 생성망 G_m(위)와 적대적 학습을 위한 판별망 D_m(아래)

Fig. 3. MSCN coefficient generator G_m (above) and whose corresponding discriminator D_m (below) for adversarial training.

계산을 통해 이루어진다[20].

$$\mu(x,y)=\sum_{k=-K}^{K}\sum_{l=-L}^{L} w_{k,l} I(x+k,y+l), \quad (1)$$

$$\sigma(x,y)=\sqrt{\sum_{k=-K}^{K}\sum_{l=-L}^{L} w_{k,l}[I(x+k,y+l)-\mu(x,y)]^2}. \quad (2)$$

여기서 $w=\{w_{k,l} \mid k=-K,\ldots,K, l=-L,\ldots,L\}$은 2차원 원형(circularly) 대칭 가우시안 가중치로써, $K=L=3$을 이용하였다[21]. μ와 σ를 이용해 최종적으로 MSCN 계수 이미지 C를 계산 가능하다.

$$C(x,y)=\frac{I(x,y)-\mu(x,y)}{\sigma(x,y)+1}. \quad (3)$$

그림 2 (a)는 RGB 이미지를 나타내며 (b)는 (1)-(3)을 통해 계산된 (a)의 MSCN 계수 이미지이다. 단순한 엣지와는 다르게 시각적 인지를 위해 중요한 구조 정보를 효과적으로 표현함을 가시적으로 확인할 수 있다. (c)와 (d)는 각각 RGB 이미지의 밝기 성분과 MSCN 계수 이미지의 히스토그램을 가시화 하였으며, NSS의 특징 중 하나인 GGD(generalized Gaussian distribution)을 따름을 확인할 수 있다[9].

2. MSCN 계수 생성망

MSCN 계수 생성망 G_m과 적대적 학습을 위한 판별망 D_m의 구조는 Nazeri[16]의 모델과 유사한 구조를 이용하였으며 그림 3에 도식화하였다. 인코더 부분의 첫 번째 콘볼루션 레이어는 7×7 크기 커널을 이용하며,

이후에는 4x4 커널의 stride를 2로 하여 특징맵의 공간적 크기를 1/2씩 감소하였다. 특징맵의 공간적 크기가 1/4로 감소된 후에는 ResNet[22]의 residual block을 8번 통과함으로써 특징맵을 정제하였다. 이후에는 인코더와 대칭적으로 디코더 부분에서 4×4 커널을 갖는 전치(transposed) 콘볼루션 연산을 통해 특징맵의 공간적 크기를 4배로 확장하며, 최종적으로 7×7 커널을 통해 1채널의 MSCN 계수 이미지를 생성한다. G_m에서 특징 채널 별 정규화는 instance normalization을 통해 수행되었으며 활성화 함수로 ReLU(rectified linear unit)를 이용하였고, 콘볼루션 가중치의 정규화를 위한 spectral normalization[23] 기법을 추가적으로 적용하였다. MSCN 계수 생성을 위한 G_m는 손실 영역을 나타내는 마스크 M과 손실 영역을 포함하는 이미지 $\tilde{I}= I\odot(1-M)$가 연결(concatenate)된 4 채널의 텐서를 입력받으며, $\odot$은 Hadamard 곱을 의미한다. G_m은 손실 영역이 복원된 MSCN 계수 이미지 C_{pred}를 생성하며 이를 다음과 같이 나타낼 수 있다.

$$C_{pred} = G_m(\tilde{I}, M). \tag{4}$$

판별망 D_m은 G_m을 통해 생성된 MSCN 계수 이미지 C_{pred}와 ground-truth인 C_{gt}의 진위 여부를 구분하도록 학습되며 4x4 커널을 이용한 5개의 콘볼루션 레이어로 설계되었다. 처음 3개 레이어에서는 stride를 2로하여 특징의 공간적 크기를 1/2로 감소하며, 마지막 레이어를 통과한 1채널 특징맵의 평균값을 출력한다. 활성화 함수로는 LReLU(Leaky ReLU)를 이용하였으며 역시 콘볼루션 레이어의 가중치 정규화를 위한 spectral normalization이 적용되었다. D_m의 학습을 위한 loss는 생성망과 판별망의 상호 안정적 학습을 보장하고자 hinge loss를 이용하였으며[24], 이에 G_m의 학습을 위한 adversarial loss L_{adv,G_m}는 다음과 같이 표현할 수 있다.

$$\begin{aligned} L_{adv,G_m} = & -E_{C_{gt}}[\min(0, -1+D(C_{gt}))] \\ & -E_{C_{pred}}[\min(0, -1-D(C_{pred}))] \end{aligned}. \tag{5}$$

또한, G_m이 생성한 C_{pred}의 품질 향상을 위한 복원(reconstruction) loss L_{rec,G_m}를 추가하였다.

$$L_{rec,G_m} = E\left[\parallel C_{gt} - C_{pred} \parallel_1\right]. \tag{6}$$

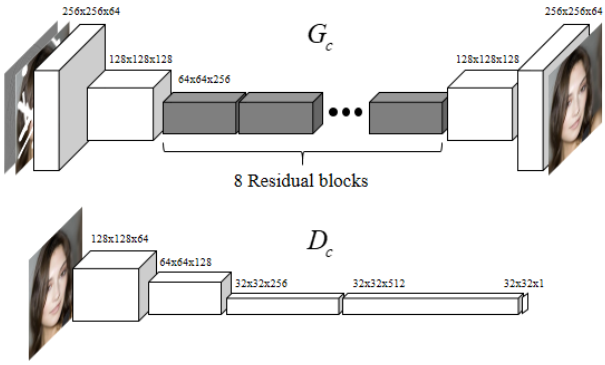

그림 4. 손실된 이미지와 생성된 MSCN 계수 이미지를 입력받아 손실 영역을 복원하는 이미지 완성망 G_c(위)와 적대적 학습을 위한 판별망 D_c(아래)

Fig. 4. Image completion generator G_c (above) and whose corresponding discriminator D_c (below) for adversarial training.

L_{adv,G_m}와 L_{rec,G_m}를 통해 MSCN 계수 생성망 G_m의 목적(objective) 함수는 다음과 같이 나타낼 수 있다.

$$\min_{G_m}\max_{D_m} L_{G_m} = \min_{G_m}\left(\max_{D_m} L_{adv,G_m} + \lambda L_{rec,G_m}\right). \tag{7}$$

여기서 λ는 복원 loss를 위한 가중치이며, 본 논문에서는 실험적으로 10을 이용하였다.

3. 이미지 완성망

MSCN 계수 생성망 G_m을 통해 생성된 MSCN 계수 이미지 C_{pred}로부터 손실 영역이 복원된 완전한 RGB 이미지 I_{pred} 생성을 위한 G_c를 이미지 완성망이라 명명하였으며 신경망의 구조는 그림 4와 같이 G_m과 동일하게 구성하였다. G_c은 $\tilde{I}$와 C_{pred}가 연결된 4채널의 텐서를 입력받으며 이를 통해 RGB 3채널의 I_{pred}를 생성한다.

$$I_{pred} = G_c(\tilde{I}, C_{pred}). \tag{8}$$

생성망 G_c와 판별망 D_c의 학습을 위한 adversarial loss L_{adv,G_c}는 ground-truth RGB 이미지인 I와 hinge loss를 통해 다음과 같이 표현 가능하다.

$$\begin{aligned} L_{adv,G_c} = & -E_I[\min(0, -1+D(I))] \\ & -E_{I_{pred}}[\min(0, -1-D(I_{pred}))] \end{aligned}. \tag{9}$$

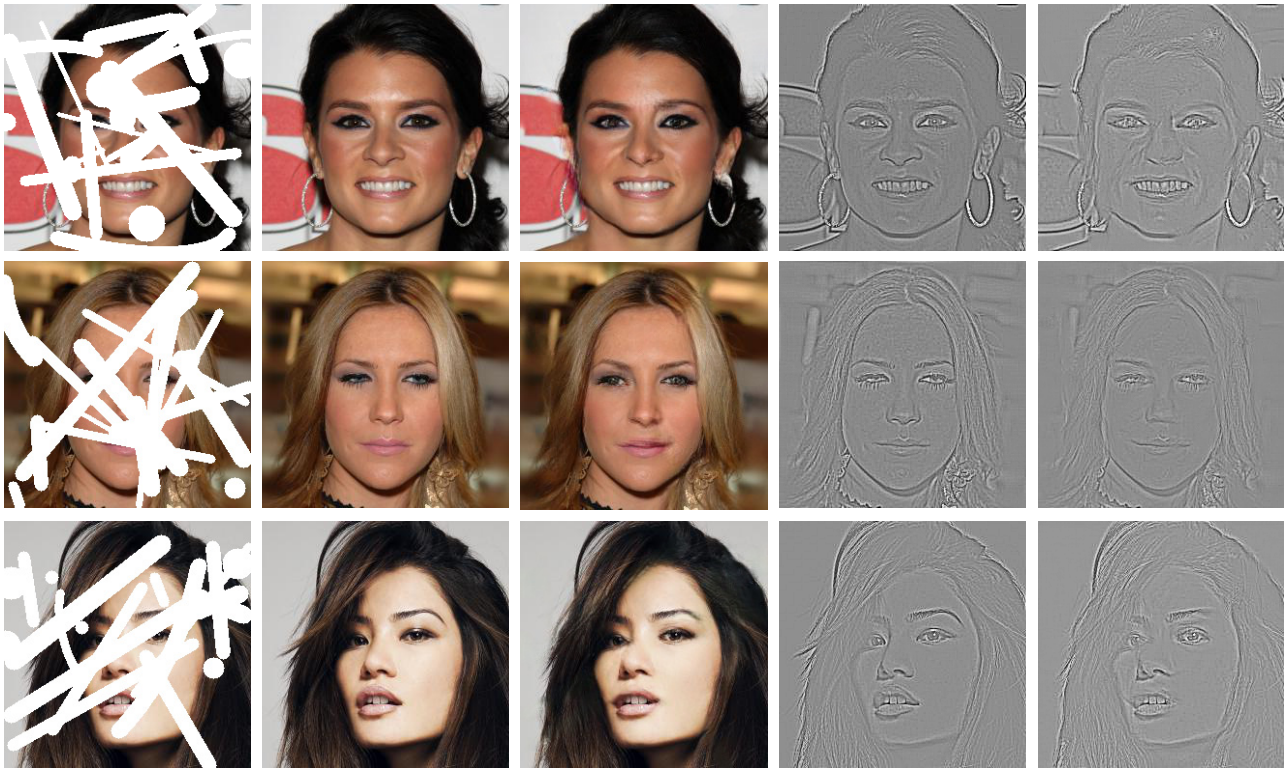

그림 5. 인페인팅 실험 결과: (a) 손실된 이미지 (b) 원본 이미지 (c) 제안하는 알고리즘을 통해 복원된 이미지 (d) 원본 이미지의 MSCN 계수 (e) 복원된 MSCN 계수

Fig. 5. Experimental results. (a) Masked image having missing regions. (b) Original image. (c) Reconstructed image by using the proposed method. (d) MSCN coefficient of original image. (e) Reconstructed MSCN coefficient.

MSCN 계수 이미지 생성망 G_m과 마찬가지로 G_c를 통해 생성된 I_{pred}와 ground-truth I를 직접적으로 비교하는 복원 loss L_{rec,G_c}를 추가하였다.

$$L_{rec,G_c} = E\left[\| I - I_{pred} \|_1\right]. \tag{10}$$

최종적으로 L_{adv,G_c}과 L_{rec,G_c}를 통해 이미지 완성망 G_c의 최종적인 목적 함수를 구성할 수 있다.

$$\min_{G_c}\max_{D_c} L_{G_c} = \min_{G_c}\left(\max_{D_c} L_{adv,G_c} + \lambda L_{rec,G_c}\right). \tag{11}$$

식 (7)과 (11)은 두 개의 GAN을 학습시키기 위한 목적 함수이며, 본 논문에서는 G_m과 G_c의 상호 보완적 동작을 위해 모든 신경망을 순차적 학습이 아닌 end-to-end로 학습하였다.

Ⅳ. 실 험

1. 신경망 학습

본 논문에서는 이미지 인페인팅 실험을 통한 결과 확인을 위해 CelebA 데이터셋을 이용하였다. CelebA 데이터셋은 256×256 해상도를 갖는 총 30000장의 이미지로 구성이 되어 있으며, 본 논문에서는 70%를 학습, 15%를 검정(valid), 15%를 테스트 데이터셋으로 활용하였다. 배치(batch) 크기는 8로 설정하였으며, 75번의 epoch 학습을 수행하였다. 신경망의 모든 학습 가능한(trainable) 파라미터는 정규 분포로부터 초기화되었다. 네 개의 신경망 학습을 위한 학습률(learning rate)은 1e-4로 동일하게 설정하였고, 학습을 위한 손실 영역 마스크는 불규칙적(irregular) 형태로 랜덤하게 생성되었다.

2. 실험 결과

제안하는 알고리즘을 이용한 이미지 인페인팅의 정성적 결과를 그림 5에 가시화하였다. (a)-(e)는 각각 손실 영역을 포함하는 이미지, 원본 이미지, 최종 복원 이미지, 원본 이미지의 MSCN 계수, 그리고 G_m을 통해 복원된 MSCN 계수를 나타낸다. G_m이 출력하는 MSCN 계수가 원본 이미지의 MSCN 계수와 유사하게 생성됨으로써 이를 입력받아 G_c가 손실 영역의 구조적 정보

그림 6. 기존 인페인팅 기술과의 결과 비교: (a) 손실된 이미지 (b) Pathak 등[13] (c) Nazeri 등[16] (d) Zheng 등[25] (e) 제안 방법 (e) 원본 이미지

Fig. 6. Qualitative comparisons with previous inpainting methods. (a) Masked image. (b) Pathak *et al.*[13]. (c) Nazeri *et al.*[16]. (d) Zheng *et al.*[25]. (e) Proposed method. (e) Original image.

뿐만 아닌 미세 영역까지 우수하게 복원하였음을 확인 가능하다. 이는 곧 제안하는 두 단계의 인페인팅 방법에 도입된 시각 인지를 모방한 MSCN 계수가 최종 RGB 이미지 복원을 위한 구조적 특징을 표현하는 강건한 사전 정보로 활용될 수 있음을 확인할 수 있다.

정성적 결과뿐만이 아닌 정량적 지표를 통한 성능의 향상을 확인하기 위해 본 논문에서는 Pathak 등[13], Nazeri 등[16], Zheng 등[25]의 기존 심층신경망 기반 인페인팅 방법들과 비교를 수행하였다. 성능의 지표는 PSNR(peak signal-to-noise ratio)와 SSIM(structural similarity)[26]를 이용하였다. 표 1에서는 제안하는 방법과 기존 방법들의 PSNR 및 SSIM 수치를 나타내었다. 제안하는 방법이 정성적 화질뿐만이 아닌 정량적 수치 측면에서도 기존 방법 대비 손실 영역 복원에 있어 우수한 성능을 나타냄을 확인할 수 있다. 특히 SSIM의 경우에는 구조적 유사도를 측정함으로써 인지적 특징을 수치에 반영하는 특성을 가짐에 따라, 제안하는 MSCN 계수 생성을 통한 인페인팅이 이러한 인지적 구조 정보 및 미세 성분에 있어 타 기술 대비 효과적으로 복원함을 객관적으로 확인 가능하다.

그림 6에서는 표 1에서의 기존 심층신경망 기반 이미지 인페인팅의 결과들과 제안하는 방법의 결과를 가시화하여 비교하였다. 손실 영역의 복원에 있어 MSCN 계수로부터의 구조적 정보를 먼저 생성한 후 이미지를 완성하는 방식의 본 제안 방법이 미세 성분의 표현에 있어서도 기존 기술 대비 사실적인 결과물 도출이 가능함을 확인하였다. 특히 손실 영역이 비규칙적이며 넓게 분포해 있음에도 효과적인 복원이 수행 가능하다는 부분에서 본 연구의 추가적인 우수성을 증명할 수 있다.

표 1. CelebA 데이터셋을 활용한 복원 성능의 정량적 비교

Table1. The quantitative comparisons over the CelebA dataset.

	PSNR(dB)	SSIM
Pathak 등[13]	24.67	0.806
Nazeri 등[16]	25.83	0.819
Zheng 등[25]	26.14	0.837
제안 기술	**26.40**	**0.844**

본 논문에서 제안한 MSCN 계수 기반 인페인팅 알고리즘에서 MSCN 계수 생성망을 제거하고 이미지 완성망 G_c만을 학습한 결과, 즉 이미지 인페인팅에 한 개의 GAN만을 학습하여 이용하였을 때의 이미지 복원 결과를 ablation study로써 확인하였다. 그림 7에서는 원본 영상 (a)를 (b)와 같이 손상시켰을 경우, 제안하는

그림 7. Ablation study 결과: (a) 원본 이미지 (b) 손상 이미지 (c) G_m 없이 학습 결과(SSIM 0.810) (d) 제안 방법(SSIM 0.913)

Fig. 7. Ablation study. (a) Original image. (b) Masked image (c) Without G_m training (SSIM 0.810). (d) Proposed method (SSIM 0.913).

기술에서의 MSCN 계수 생성망 G_m을 제거했을 때의 결과를 나타내었다. (c)에서 볼 수 있듯, 한 개의 생성망 G_c만을 이용한 직접적인 이미지 인페인팅은 손실 영역 복원을 위한 사전 정보의 부재로 인해 제안하는 방법의 결과인 (d) 대비 복원된 영상의 화질이 급격하게 저하되었음을 확인할 수 있다. 이는 본 연구에서 활용한 MSCN 계수가 미세 영역의 효과적 복원을 위한 고품질의 이미지 인페인팅에 있어 중요한 구조적 사전 정보를 제공하며 두 단계에 걸친 생성이 시각적 화질 측면에서 효과적인 방법임을 의미한다.

V. 결 론

본 논문에서는 기존 이미지 인페인팅 기술이 도출하던 화질적 저해 요소 발생을 최소화하기 위한 두 단계의 생성적 인페인팅 기술을 소개하였다. 특히 손실 영역의 복원을 위한 맥락적 구조를 내포하는 MSCN 계수를 사전 정보로 생성함으로써 이미지 완성 단계에서 미세 성분 향상시킨 보다 고품질의 인페인팅이 가능하도록 학습 프레임워크를 설계하였고, 실험 결과를 통해 제안하는 방법의 정성적/정량적 우수성을 입증하였다. 이러한 접근은 인간의 시각 인지적 특성이 저수준(low-level)의 컴퓨터 비전 분야에 효율적으로 반영될 수 있음을 시사하며, 추후 여타 초해상도, 디헤이징 혹은 디노이징과 같은 영상 처리 분야에서도 성능 향상의 도모를 기대할 수 있다.

REFERENCES

[1] A. Levin, A. Zomet, S. Peleg, and Y. Weiss, “Seamless image stitching in the gradient domain,” European Conf. Compuer Vision (ECCV), 2004.

[2] E. Park, J. Yang, E. Yumer, D. Ceylan, and A. C. Berg, “Transformation-grounded image generation network for novel 3D view synthesis,” IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2017.

[3] C. Ballester, M. Bertalmio, V. Caselles, G. Sapiro, and J. Verdera, “Filling-in by joint interpolation of vector fields and gray levels,” IEEE Trans. on Image Process. (TIP), vol. 10, no. 8, pp. 1200-1211, 2001.

[4] S. Esedoglu and J. Shen, “Digital inpainting based on the Mumford-Shah-Euler image model,” European J. Appl. Math., vol. 13, no. 4, pp. 353-370, 2002.

[5] S. Darabi, E. Shechtman, C. Barnes, D. B. Goldman, and P. Sen, “Image melding: combining inconsistent images using patch-based synthesis,” ACM Trans. Graphics (TOG), vol. 31, no. 4, July 2012.

[6] J. B. Huang, S. B. Kang, N. Ahuja, and J. Kopf, “Image completion using planar structure guidance,” ACM Tran. Graphics (TOG), vol. 33, no. 4, July 2014.

[7] N. Wang, J. Li, L. Zhang, and B. Du, “MUSICAL: multi-scale image contextural attention learning for inpainting,” Int’l Joint Conf. Artificial Intelligent (IJCAI), 2019.

[8] S. Iizuka, E. Simo-Serra, and H. Ishikawa, “Globally and locally consistent image completion,” ACM Tran. Graphics (TOG), vol. 36, no. 4, July 2017.

[9] A. Mittal, A. K. Moorthy, and A. C. Bovik, “No-reference image quality assessment in the spatial domain,” IEEE Tran. Image Process. (TIP), vol. 21, no. 12, pp. 4695-4708, Dec. 2012.

[10] S. B. Kotsiantis, D. Kanellopoulos, and P. E. Pintelas, “Data preprocessing for supervised learning,” Int. J. Comput. Sci., vol. 1, no. 1, pp. 111-117, 2006.

[11] A. Telea, “An image inpainting technique based on the fast marching method,” J. Graphics Tools, vol. 9, no. 1, pp. 23-34, 2004.

[12] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman, “Patchmatch: a randomized correspondence algorithm for structural image editing,” ACM Tran. Graphics (TOG), vol. 28, no. 3, 2009.

[13] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: feature learning by inpainting," IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2016.
[14] C. Yang, X. Lu, Z. Lin, E. Shechtman, O. Wang, and H. Li, "High-resolution image inpainting using multi-scale neural patch synthesis," IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2017.
[15] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," ACM Tran. Graphics (TOG), vol. 36, no. 4, 2017.
[16] K. Nazeri, E. Ng, T. Joseph, F. Z. Qureshi, and M. Ebrahimi, "EdgeConnect: generative image inpainting with adversarial edge learning," Int'l Conf. Computer Vision (ICCV), 2019.
[17] Y. Ren, X. Yu, R. Zhang, T. H. Li, S. Liu, and G. Li, "StructureFlow: image inpainting via structure-aware appearance flow," Int'l Conf. Computer Vision (ICCV), 2019.
[18] N. Pinto, D. D. Cox, and J .J. DiCarlo, "Why is real-world visual object recognition hard?," PLoS Comput. Biol., vol.4, no. 1, Jan. 2008.
[19] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun, "What is the best multi-stage architecture for object recognition?," Int'l Conf. Computer Vision (ICCV), 2009.
[20] H. Oh, S. Ahn, J. Kim, and S. Lee, "Blind deep S3D image quality evaluation vial local to global feature aggregation," IEEE Trans. on Image Process. (TIP), vol. 26, no. 10, pp. 4923-4936, Oct. 2017.
[21] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," IEEE Signal Process. Lett., vol. 20, no. 3, pp. 209-2012, March 2013.
[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2016.
[23] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," Int'l Conf. Learning Representation (ICLR), 2018.
[24] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Free-form image inpainting with gated convolution," Int'l Conf. Computer Vision (ICCV), 2019.
[25] C. Zheng, T-J. Cham, and J. Cai, "Pluralistic image completion," IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2019.
[26] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment from error visibility to structural similarity," IEEE Tran. Image Process. (TIP), vol. 13, no. 4, pp. 600-612, April 2004.

저 자 소 개

오 희 석(정회원)
2017년 연세대학교 전기전자 공학과 박사 졸업
2017년～2017년 삼성전자 DMC 연구소 책임연구원
2017년～2020년 한국전자통신 연구원 선임연구원
2020년～현재 한성대학교 IT융합공학부 조교수
<주관심분야 : 영상처리, 컴퓨터비전, 혼합현실, 심층생성모델 등>

최 원 석(정회원)
2018년 고려대학교 정보보호 대학원 박사 졸업
2018년～2020년 고려대학교 정보 보호연구원 연구교수
2020년～현재 한성대학교 IT융합 공학부 조교수
<주관심분야 : 센서 보안, 자동차 보안, 암호 프로토콜>