Journal of Next-generation Convergence Technology Association ISSN 2508-8270 EISSN : 2765-4133
Vol.7, No.12, pp. 1-000, 2023 https://doi.org/10.33097/JNCTA.2023.07.12.000

영상 복원을 위한 다차원 윈도우 기반 비전 트랜스포머 네트워크

김성주[1], 최재영[2*]
[1]한국외국어대학교 컴퓨터공학과 석사과정, [2]한국외국어대학교 컴퓨터공학과 부교수

Multiscale Window based Vision Transformer for Image Inpainting

Seong-Joo Kim[1], Jae-Young Choi[2*]
[1]Master Student, Division of Computer Engineering, Hankuk University of Foreign Studies
[2]Associate Professor, Division of Computer Engineering, Hankuk University of Foreign Studies

요 약 효과적인 영상 복원을 달성하려면 모델이 맥락정보(contextual information)를 파악하는 것이 중요하다. 합성곱 신경망(CNN) 기반 알고리즘을 사용한 이전 연구는 장거리 종속성이 부족하여 모델이 상황별 정보를 캡처할 수 없다는 한계에 직면했다. 이를 해결하기 위해 본 논문에서는 고품질 영상 복원을 위한 다중 차원의 윈도우(Multi-Scale Window) 기반 비전 트랜스포머 모델을 제안한다. 다중 차원 기반 비전 트랜스포머를 도입함으로써 다양한 창 크기의 영향을 반영하고 이에 따른 맥락정보를 얻을 수 있다. 또한 손실 마스크 업데이트 모듈을 적용하여 효율적인 연산방식을 도입하였다. 실험 결과에 기반하면 본 논문의 제안 모델은 누락된 영역을 효과적으로 복원하고 벤치마크 데이터셋에서 FID 수치가 다른 최신 복원 모델들의 결과에 비해 약 50%이상 낮은 것으로 확인 되었다. 이는 제안 복원 모델의 우수성을 검증한다고 할 수 있다.

주제어 : 영상 복원, 맥락정보, 비전 트랜스포머, 다중차원 윈도우, 딥러닝

Abstract To achieve successful image inpainting, it is crucial for a model to understand contextual information. Previous research using Convolutional Neural Network (CNN)-based algorithms have limitations in capturing context-specific information due to a lack of long-range dependencies. To address this issue, we proposes a Multiscale Window based Vision Transformer (MWVT) model for high-quality image completion. By introducing a multi-dimensional vision transformer, the proposed MWVT can capture context information by using various window sizes, resulting in the reflectance of their respective influences. Furthermore, our MWVT incorporates a mask update module to reduce the computational cost. According to our experiments, the proposed model in this paper effectively restores the missing area and the FID value in the benchmark dataset is about 50% lower than the results of other state-of-the-art restoration models.

Key Words : Image Inpainting, Contextual Information, Vision Transformer, Multiscale Window, Deep Learning

* This research was supported by the "Development of technology using analysis of ocean satellite images (20210046)" funded by the Korea Institute of Marine Science & Technology Promotion (KIMST).
This research was supported by Hankuk University of Foreign Studies Research Fund.

* Corresponding Author : Jae Young Choi (jychoi@hufs.ac.kr)

Received November 02, 2023 Revised December 19, 2023
Accepted December 20, 2023 Published December 28, 2023

Ⅰ. 서론

영상 복원은 손상된 영상의 비어 있는 영역을 타당한 내용으로 채워 넣어 복원하는 것을 목표로 한다. 영상 복원은 사진 편집(photo editing)[1], 영상복원(Image restoration)[2]과 같이 영상을 복원하거나 새로운 영상으로 편집하는 기술에 적용될 수 있다. 또한 객체 제거(object removal)[3] 와 같은 영상 내의 원하지 않는 객체를 삭제하는 기술로 응용이 가능하다.

효과적인 영상 복원에 있어서 딥러닝을 적용한 많은 알고리즘들이 제시되고 있다. 효과적인 영상 복원을 위해서는 맥락정보(contextual information)를 모델이 파악하는 것이 매우 중요하다. 여기서 맥락정보란 영상 내 손실된 영역, 즉 채워야 할 영역의 주변 화소(pixel)들로부터 추론된 정보들을 의미한다. 이러한 맥락정보들이 충분하게 제공이 되어야 손실된 영역의 형상, 구조 내지는 텍스쳐(texture) 등을 잘 파악할 수 있기 때문이다. 따라서 모델은 손실된 영역을 복원할 때 손실 영역의 맥락정보를 추론하며 이를 적절하게 사용하여 복원한다. 이전의 연구들[4,5] 은 합성곱 신경망 기반의 인코더-디코더(encoder-decoder) 구조를 통해 복원을 위한 적절한 주변 정보들을 저차원부터 고차원까지 파악하여 복원을 수행한다. 그러나 이러한 심층 합성곱 신경망 기반 알고리즘들은 작은 손실 영역에는 효과적으로 복원이 되지만, 큰 손실 영역에서는 그렇지 못하다. 이는 심층 합성곱 신경망의 특징 맵 생성에서의 본질적 특성인 '제한된 수용 영역(receptive field)'으로 인해 전역 특징 및 즉 장거리 의미는 학습할 수 없게 된다는 문제에서 발생하게 된다.[6]

위의 문제점을 해결하기 위해 모든 레이어에 self-attention이 기본 구성요소로 있는 비전 트랜스포머[7] 를 기반으로 한 모델[8,9]이 도입되었다. 비전 트랜스포머를 기반으로 한 영상 복원 알고리즘은 심층 합성곱 신경망 기반의 알고리즘 보다 전역 정보들을 잘 파악하며 이를 통해 보다 유효한 맥락정보를 활용할 수 있다. 대부분의 ViT[7] 기반의 모델들은 입력 영상을 하나의 패치 사이즈로 나누어 나온 토큰으로 self-attention을 진행하거나, 패치화 된 토큰들을 윈도우로 나누어 해당 윈도우 내에서 self-attention을 수행하는 Swin-transformer[10] 기반의 모델이다. Swin-transformer 기반의 모델들은 대부분 한 트랜스포머 블록 당 하나의 고정된 윈도우 사이즈를 사용하여 self-attention을 진행하게 된다. 이는 다양한 윈도우 사이즈에 따른 영향을 반영하지 못하며, 고정된 윈도우 사이즈로 패치를 나누어 영상 내의 크기가 다양한 객체들을 적절하게 처리하기 어렵다. 따라서 객체의 크기에 따른 다차원 정보를 얻는 데에 제한되기 때문에 맥락정보 역시 한정적이다.[11]

따라서 본 연구에서는 큰 손실 영역까지 복원가능한 다중 차원 윈도우 기반 비전 트랜스포머 네트워크(Multscale Window based Vision Transformer, MWVT)를 제안한다. MWVT는 트랜스포머 블록을 통해 다중 차원의 윈도우를 사용하여 병렬적으로 self-attention을 수행하게 하였다. 이를 통해 더욱 다양한 맥락정보를 확보하고자 하였다.

Ⅱ. 제안 모델

2.1 전체 모델 구조

본 연구에서 제안한 전체 모델 구상도는 Fig. 1과 같다. 입력으로 손실 영상과 해당 영상의 이진 마스크가 들어가게 되며, 손실 영상이 복원되는 것이 본 연구의 목표이다. 본 모델은 크게 영상이 복원되는 MWVT 블록과 이진 마스크가 업데이트 블록으로 이루어져 있으며, 업데이트된 마스크는 비전 트랜스포머의 입력으로 영상과 같이 들어가 MCA(Multi-Head Contextual Attention)를 수행하게 된다. 업데이트된 마스크를 통해 유효한 토큰들 간의 attention을 진행할 수 있게 되는 것이다. 먼저 입력으로 들어간 영상은 합성곱 신경망층을 통해 토큰화 되고, 이어서 5개의 트랜스포머 블록을 통과하게 된다. 앞서 합성곱 신경망층을 통과하는 이유는 초기에 local inductive prior을 활용하게 되면 더 나은 형상(representation)이 나올 수 있기 때문이다.[13] 또한 합성곱 신경망을 사용해 원래 영상을 다운샘플링(downsampling) 하여 컴퓨팅 연산량을 줄이고자 하였다. 5개의 트랜스포머 블록의 입력 차원은 패치 병합(patch merging)과 패치 업샘플링(patch upsampling)을 통해 다르게 들어간다. 패치 병

합을 통해서 해당 레이어의 특징의 차원은 1/2씩 줄어들게 되며, 그에 따라 채널 차원은 두 배씩 증가하게 된다. 패치 업샘플링의 경우 패치 병합과 반대로 계산되게 되어 특징 차원은 두 배씩 증가하고 채널 차원은 1/2씩 줄어든다. 따라서 마지막 트랜스포머 블록을 통과한 결과는 합성곱 신경망층을 다시 통과하여 원래의 입력 사이즈로 복원이 된다. 더 나아가 마지막 레이어로 U-Net[14] 을 사용하여 심층 합성곱 신경망 국소 영역 텍스처(local texture)를 개선하는 기능을 통해 영상의 세밀한 질감을 향상하게 하였다.

2.2 다차원 윈도우 기반 비전 트랜스포머 (MWVT)

본 연구에서는 큰 손실 영역을 잘 복원하는 영상 복원 모델을 위한 새로운 트랜스포머 블록 구조를 제시하고자 한다. MWVT의 경우 기존의 비전 트랜스포머 구조들과 달리 다중 차원의 윈도우로 패치들을 분할하여 self-attention을 진행했다. 이를 통해 다중차원의 윈도우에 기반한 정보들을 포착하여 활용하고자 하였다. Fig. 2를 보면, 한 트랜스포머 블록에는 크게 세 가지의 모듈인 MCA와 FC(Fully Connected) layer, MLP로 구성되어 있다. 이를 수식화 하면 다음과 같다.

$$\begin{aligned} x^{l+1} &= \mathrm{MLP}(\mathrm{FC}(\hat{x}^{l})) \\ \hat{x}^{l} &= \mathrm{concat}(\{\mathrm{MCA}_i(x_i^{l})\}) \end{aligned} \tag{1}$$

수식 (1)에서 concat은 입력 데이터를 결합(concatenation)하는 함수이다. 입력 특징이 $x \in \mathbb{R}^{H\times W\times C}$ 와 같고 윈도우의 개수는 n_{win}으로 가정하면, 입력 특징 x는 n_{win}개의 특징으로 나누어진다. 이 때 x_i^l는 l번째 트랜스포머 블록의 i번째 윈도우로 나누어진 특징을 의미한다. 이는 MCA를 통과 후 $\mathrm{H}\times\mathrm{W}\times\frac{\mathrm{C}}{n_{win}}$ 차원으

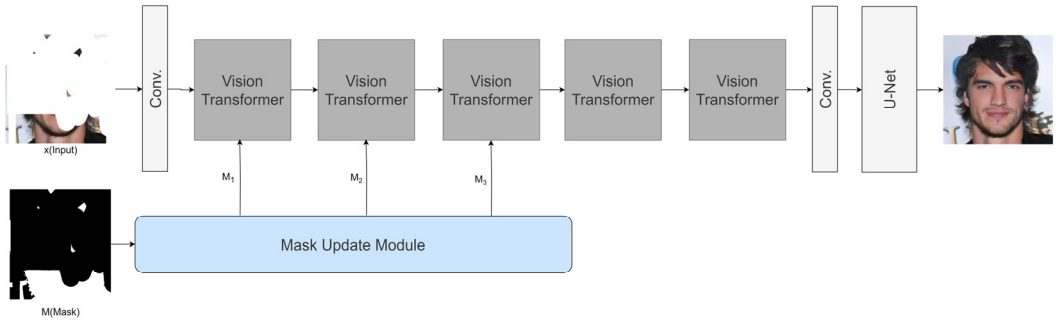

Fig. 1. Overall structure of the proposed model, named MWVT. (Conv. block refers to a convolution operation, and the T block refers to a vision transformer.)

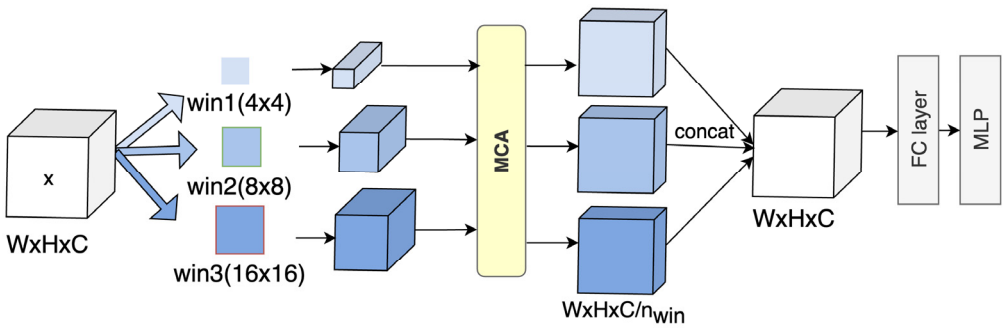

Fig. 2. Detailed structure of the vision transformer block composed of the proposed model MWVT.

로 복원되며, 복원된 특징들은 채널 차원으로 결합되어 나온 출력은 $\hat{x}^l$이다. 이는 곧 FC 레이어와 MLP를 통과하여 최종 출력 x^{l+1}로 나와 다음 트랜스포머 블록의 입력으로 들어간다. 본 논문에서 $n_{win}=3$으로 설정하였으며, 각 윈도우의 사이즈 $win_{i(i=1,2,3)}=\{4,8,16\}$으로 수행하였다.

2.2.1 Multi-head Contextual Attention(MCA) 모듈

영상 복원의 경우 유효한 토큰과 그렇지 않은 토큰으로 이루어진 마스킹 된 영상을 다루게 된다. 이 경우 기존의 self-attention을 사용하게 되면, 유효하지 않은 토큰의 정보도 연산하기 때문에 연산의 비효율성이 유발된다. 또한 유효한 토큰의 정보를 약화시켜 손실된 픽셀 정보들을 채우는 데에 한계로 작용할 수 있다. 따라서 마스크에 따른 유효한 토큰만을 이용하여 self-attention을 진행하는 MCA[12]를 도입해 더욱 고품질의 복원의 결과를 내려 하였다. MCA의 수식은 다음과 같다.

$$\mathrm{Att}(\mathbf{Q},\mathbf{K},\mathbf{V})=\sigma\left(\frac{\mathbf{Q}\mathbf{K}^{\mathrm{T}}+\mathbf{M}}{\sqrt{d_k}}\right)\mathbf{V} \tag{2}$$

여기서 Q, K, V는 각 query, key, value이고 d_k의 경우 차원 맞춤을 위한 스케일 팩터이다. M의 경우 마스크를 의미하며, 이는 해당 토큰이 유효한 경우(손실 영역이 아닌 경우) 0으로 설정되고 유효하지 않은 경우(손실 영역인 경우) -3000으로 설정된다. 따라서 MCA는 토큰이 유효한지 유효하지 않은지에 따라 self-attention 값을 살릴지 말지를 결정하게 된다. MCA는 윈도우 기반 어텐션 모듈이기 때문에 shifted window[9] 기법을 같이 사용하여 윈도우들 간의 정보 공유를 유지할 수 있게 하였다.

2.2.2 손실 마스크 업데이트 모듈

마스크는 바이너리로 유효한 토큰은 1을 가지고 그렇지 않은 토큰은 0이다. 마스크 토큰에 따라 해당 토큰의 위치에 있는 값이 MCA 연산이 적용될지 말지가 결정된다. 초기 마스크의 경우 입력 영상의 마스크를 그대로 사용하며, 다음 트랜스포머 블록으로 넘어갈 때 마스크가 업데이트 된다. 목표는 영상의 빈 영역을 모두 채우는 것이기 때문에 업데이트 과정은 모든 토큰이 유효한 토큰으로 변환되는 과정이라고 볼 수 있다.

마스크는 한 트랜스포머 블록 마다 업데이트가 된다. 먼저 가장 작은 윈도우 사이즈로 마스크를 분할 한 후 한 윈도우 내에 유효한 토큰이 하나라도 있을 경우 해당 윈도우 내에 있는 모든 토큰을 유효한 토큰으로 업데이트 한다. 만약 윈도우 내에 유효한 토큰이 하나도 없을 경우 해당 윈도우 내의 토큰들은 업데이트하지 않는다. 몇 번의 업데이트 과정을 거치게 되면, 마스크는 모두 유효한 토큰으로 업데이트 된다.

III. 손실 함수

영상 생성의 품질을 향상하기 위해서 본 논문에서는 적대적 손실(adversarial loss)[15]과 지각 손실(perceptual loss)[16], 스타일 손실(style loss)[17]을 사용하여 모델을 훈련하였다.

3.1 적대적 손실함수

영상 생성에서 가장 널리 사용되는 GAN을 안정적으로 훈련하기 위해서 본 논문에서는 비포화 적대적 손실(non-saturating adversarial loss)[15]을 도입하였다.

$$\begin{aligned}\mathcal{L}_G &= -E_{\hat{x}}[\log(D(\hat{x}))] \\ \mathcal{L}_D &= -E_x[\log(D(x))] - E_{\hat{x}}[\log(1-D(\hat{x}))]\end{aligned} \tag{3}$$

위의 수식에서 x는 정답 영상이고, $\hat{x}$는 생성기 G가 만들어낸 가짜 영상이다. 따라서 생성기와 판별기 D는 각 손실함수에 따라 최적화된다.

3.2 지각 손실함수

지각 손실함수는 미리 훈련된 VGG-16[18] 네트워크를 사용하여 구한다. 이는 생성된 영상의 특징맵과 실제 영상의 특징맵 간의 차이를 비교하여 더욱 사실적인 영상으로 만들어준다.

$$\mathcal{L}_{perc}=\sum_{i=1}^{N}\eta_i \parallel \phi_i(\hat{x})-\phi_i(x) \parallel_1 \tag{4}$$

ϕ_i은 VGG16 네트워크의 i번째 레이어의 활성함수를 의미한다. 그리고 η_i의 경우 스케일 계수를 의미한다.

3.3 스타일 손실함수

보다 더 세밀한 질감을 복원하기 위해 스타일 손실함수를 추가하였다. 스타일 손실은 지각 손실함수와 마찬가지로 미리 훈련된 VGG-16 레이어의 특징맵을 사용한다. 특징맵과 정답 영상의 그람 행렬(Gram matrix)을 구한 후 두 그람 행렬은 평균제곱오차(MSE)로 계산된다. 해당 과정은 아래 수식과 같다.

$$\mathcal{L}_{style} = \sum_{i=1}^{N} \eta_i \| \phi_i(\hat{x})(\phi_i(\hat{x}))^T - \phi_i(x)(\phi_i(x))^T \|_1 \quad (5)$$

따라서 최종 생성기의 전체 손실함수는 다음과 같이 구성된다.

$$\mathcal{L} = \mathcal{L}_G + \lambda_1 R_1 + \lambda_2 \mathcal{L}_{perc} + \lambda_3 \mathcal{L}_{style} \quad (6)$$

λ_1와 λ_2, λ_3은 손실함수의 매개변수로 본 실험에서는 $\lambda_1 = 10$, $\lambda_2 = 0.1$, $\lambda_3 = 0.1$로 설정하였다.

Ⅳ. 실험

4.1 실험 데이터셋 및 조건

본 논문에서 실험한 데이터셋은 영상 복원 검증에 가장 많이 활용되는 CelebA-HQ[19]이다. 영상 해상도는 256x256으로 실험하였다. 해당 데이터셋은 훈련에는 27,007장이 사용되고 검증에는

2,993장이 사용되었다. 훈련과 검증을 위한 마스크 생성은 DeepFill v2[6]를 참고하여 free-form 마스크로 생성하였다. 마스크의 손실 영역 비율 범위의 경우 [0.0,0.9]로 설정하여 이 범위 내에서 무작위 샘플링을 진행하여 생성하였다. 위의 마스크 손실 비율 범위는 타 모델과의 결과 비교를 위해 이전 연구의 인페인팅 모델[11]을 참고하여 생성했다.

실험조건의 경우 배치사이즈는 16에 학습율은 0.001로 설정하였다. 실험은 2 parallel NVIDIA Geforce RTX 3090이 장착된 PC로 수행하였다. 제안 네트워크 모델을 구성하는 계층 구조(layer)에 대한 세부정보는 Table 1을 통해 알 수 있다.

Table 1. Structure of proposed network by representing the type of each layer, number of parameters, and output shape of each layers.

Layer	Output Shape	Param #
conv_1	[8, 180, 256, 256]	6660
enc_conv_1	[8, 180, 128, 128]	291780
enc_conv_2	[8, 180, 64, 64]	291780
transformer_1	[8, 4096, 180]	584640
transformer_2	[8, 1024, 180]	876420
transformer_3	[8, 256, 180]	1169280
transformer_4	[8, 1024, 180]	876420
transformer_5	[8, 4096, 180]	876420
dec_conv_1	[8, 180, 128, 128]	876243
dec_conv_2	[8, 180, 256, 256]	876243
encoder_block (256x256)	[8, 128, 256, 256]	148608
encoder_block (128x128)	[8, 256, 128, 128]	885248
encoder_block (64x64)	[8, 512, 64, 64]	3539968
encoder_block (32x32)	[8, 512, 32, 32]	4719616
encoder_block (16x16)	[8, 512, 16, 16]	4719616
decoder_block (32x32)	[8, 512, 32, 32]	7081989
decoder_block (64x64)	[8, 512, 64, 64]	7081989
decoder_block (128x128)	[8, 256, 128, 128]	3344645
decoder_block (256x256)	[8, 128, 256, 256]	1229957

4.2 실험결과

본 논문에서 제안하는 모델인 MWVT와 영상 복원 최신 모델들과 비교하여 실험 및 분석하고자 한다. 정성적 분석과 정량적 분석으로 나누어 진행하며 정량적 분석에서는 FID[20], P-IDS[21], U-IDS[22]를 이용하여 총 세 가지 지표로 비교하고자 한다. FID(Fréchet Inception Distance)의 경우 모델을 통해 복원된 영상의 품질을 평가하는 데에 사용되며, 생성된 영상과 실

제 영상 간의 데이터 분포의 거리를 계산한다. P-IDS (Paired Inception Discriminative Score)는 생성된 샘플와 진짜 샘플이 주어질때 가짜 이미지가 진짜 이미지보다 더 진짜라고 판단될 확률을 구한다. U-IDS(Unpaired Inception Discriminative Score)는 생성된 이미지 집합과 진짜 이미지 집합에 대해 따로 확률을 계산한다. PSNR과 SSIM[23]의 경우 많은 연구[24]에서 고해상도의 영상 퀄리티를 평가할 때 인간의 시각과는 상당한 차이를 보여 부정확한 평가를 내리는 것으로 나타나 본 연구에서는 제외하였다.

4.2.1 정량적 평가

본 실험에서는 제안 모델 MWVT와 여러 최신기술(SOTA)과 비교를 진행하였다. 공정한 비교를 위해 훈련과 테스트는 모두 같은 데이터셋을 사용하였으며, 테스트에 사용되는 손실 마스크 역시 동일한 데이터로 실험하였다. 비교를 위한 최신기술 모델로는 CoModGAN[21], DeepFillv2[6], MADF[25]이다. 정량적 비교 결과 Table 2와 같이 본 논문이 제안한 모델이 좋은 성능을 내는 것을 확인할 수 있다. FID의 경우 낮은 수치가 좋은 성능을 의미하며, U-IDS, P-IDS의 경우 높은 수치일수록 정답과 가까운 결과를 의미한다. 세 지표 모두 제안모델 MWVT가 가장 성능이 높게 나온 것을 확인할 수 있다.

4.2.2 정성적 평가

결과에 대해 시각적 비교를 위해서 실험 데이터셋 CelebA-HQ의 테스트 결과를 정답 영상과 함께 비교하였다. Fig. 3을 보면 제안 모델인 MWVT의 결과 영상과 총 3개의 비교 모델의 결과 영상을 볼 수 있다. 모든 영상은 256x256 사이즈의 영상으로 왼쪽부터 정답영상, 입력으로 들어가는 손실 영상, 모델을 통한 복원 영상들 순이다. 다른 모델의 결과들 보다 제안 모델

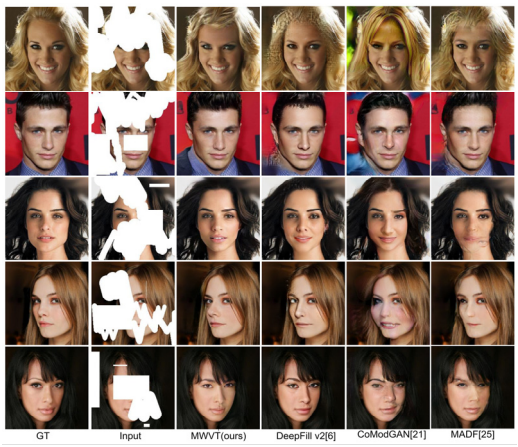

Fig. 3. Results of the models for the quantitative evaluation of this experiment.

의 결과 영상이 정답 영상과 가장 흡사하게 시각적으로 채워져 있는 것을 확인 할 수 있다. 또한 영상 자체의 잡음이 가장 적은 것을 확인할 수 있다.

Ⅴ. 결론

본 논문에서는 고품질 영상 복원을 위한 네트워크로 다중차원 윈도우 기반 트랜스포머를 제안하였다. 다양한 맥락 정보를 통합하기 위해 다차원 윈도우을 통해 병렬적으로 셀프 어텐션을 수행하여 기존 모델에 비해 우수한 성능을 달성했다. 또한 MCA를 통해 업데이트를 통한 이진 마스크에 따른 유효한 토큰만 이용하여 복원을 진행하였다. 이를 통해 연산의 효율성을 확보할 수 있었으며, 마스크 없이 self-attention을 진행하는 것 보다 실질적인 맥락정보를 확보할 수 있게 된다. 실험을 통해 제안모델의 FID는 5.30으로 타모델들의 FID에 비해 약 50% 이상 낮은 것으로 확인하였다. 또한 P-IDS와 U-IDS 각 8.13 및 18.23으로 타모델

Table 2. Quantitative comparison between modelson the Celeba-hq dataset

Model	MWVT (proposed Model)	CoModGAN[21]	DeepFill v2[6]	MADF[25]
FID	5.30	14.56	12.79	11.13
P-IDS(%)	8.13	6.41	5.47	0.82
U-IDS(%)	18.23	13.8	3.07	3.17

들과 비교했을 때 가장 높은 것으로 확인된다. 이에 따라 정성적 평가에서도 제안모델의 결과 이미지가 타모델의 결과들보다 현실적으로 빈 영역을 채운 것으로 평가된다. 따라서 정량적 평가와 정성적 평가 모두에서 다른 모델에 비해 우수한 성능을 보이는 것을 관찰하였다. 향후 연구에서는 여러 데이터셋을 추가로 활용하여 제안 모델의 성능 및 복원 성능의 안정성을 추가적으로 검증할 예정이다.

REFERENCES

[1] Y. Jo & J. Park. (2019). Sc-fegan: Face editing generative adversarial network with user's sketch and color. *In ICCV*, 1745-1753. DOI : 10.1109/ICCV.2019.00183

[2] Z. Wan, B. Zhang, D. Chen, P. Zhang, D. Chen, J. Liao & F. Wen. (2020). *Old photo restoration via deep latent space translation*. arXiv preprint arXiv:2009.07047. DOI : 10.1109/TPAMI.2022.3163183

[3] R. Shetty, M. Fritz & B. Schiele. (2018). Adversarial scene editing: Automatic object removal from weak supervision. *In Proc NeurIPS*, 7717-7727. DOI : 10.1109/ACCESS.2021.3106124

[4] Y. Zhaoyi. et al. (2018). Shift-net: Image inpainting via deep feature rearrangement. *Proceedings of the European Conference on Computer Vision(ECCV)*, 1-17. DOI : 10.1007/978-3-030-01264-9_1

[5] Z. Yanhong. et al. (2019). Learning pyramid-context encoder network for high-quality image inpainting. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. DOI : 10.1109/CVPR.2019.00158

[6] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu & T. S. Huang. (2018). Generative image inpainting with contextual attention. *In CVPR*, 5505-5514. DOI : 10.1109/CVPR.2018.00577

[7] D. Alexey. et al. (2020). *An image is worth 16x16 words: Transformers for image recognition at scale*. arXiv preprint arXiv:2010.11929. DOI : 10.48550/arXiv.2010.11929

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser & I. Polosukhin. (2017). Attention is all you need. *In NIPS*, 5998-6008. DOI : 10.48550/arXiv.1706.03762

[9] Z. Wan, J. Zhang, D. Chen & J. Liao. (2021). *High-fidelity pluralistic image completion with transformers*. arXiv preprint arXiv:2103.14031. DOI : 10.1109/iccv48922.2021.00465

[10] L. Ze. et al. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision*. DOI : 10.1109/iccv48922.2021.00986

[11] R. Pengzhen. et al. (2022). Beyond fixation: Dynamic window visual transformer. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. DOI : 10.1109/cvpr52688.2022.01168

[12] L. Wenbo. et al. (2022). Mat: Mask-aware transformer for large hole image inpainting. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. DOI : 10.1109/cvpr52688.2022.01049

[13] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang & A. Dosovitskiy. (2021). *Do vision transformers see like convolutional neural networks?* arXiv preprint arXiv:2108.08810. DOI : 10.48550/arXiv.2108.08810

[14] R. Olaf, P. Fischer & T. Brox. (2015). U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015: 18th International Conference, Proceedings, Part III 18*. Springer International Publishing. DOI : 10.1007/978-3-319-24574-4_28

[15] W. T. Dumoulin, V. Arulkumaran K. Sengupta, B. Bharath & A. A. Creswell. (2018). Generative

adversarial networks: An overview. *In IEEE Signal Process. Mag*, 53-65. DOI : 10.20944/preprints202212.0191.v1

[16] J. Johnson, A. Alahi & L. Fei-Fei. (2016). Perceptual losses for real time style transfer and super-resolution. *In ECCV*, 694-711. DOI : 10.1007/978-3-319-46475-6_43

[17] A. S. Ecker, M. Bethge & L. A. Gatys. (2019). Image style transfer using convolutional neural networks. *In CVPR*, 1745-1753. DOI : 10.1109/cvpr.2016.265

[18] K. Simonyan & A. Zisserman. (2014). *Very deep convolutional networks for large-scale image recognition*. arXiv preprint arXiv: 1409.1556. DOI : 10.48550/arXiv.1409.1556

[19] T. Karras, T. Aila & J. Lehtinen. (2018). Progressive growing of gans for improved quality, stability, and variation. *In ICLR*. DOI : 10.48550/arXiv.1710.10196

[20] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler & S. Hochreite. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NIPS*. DOI : 10.18034/ajase.v8i1.9

[21] S. Zhao, J. Cui, Y. Sheng, Y. Dong, X. Liang, E. Chang & Y. Xu. (2020). Large scale image completion via co-modulated generative adversarial networks. *In ICLR*. DOI : 10.48550/arXiv.2103.10428

[22] R. Zhang, P. Isola, A. A. Efros, E. Shechtman & O. Wang. (2018). The unreasonable effectiveness of deep features as a perceptual metric. *In CVPR*, 586-595. DOI : 10.1109/cvpr.2018.00068

[23] A. C. Bovik, H. R. Sheikh, E. P. Simoncelli & Z. Wang. (2004). Image quality assessment: From error visibility to structural similarity. *In IEEE TIP*, 600-612. DOI : 10.1109/tip.2003.819861

[24] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang. et al. (2017). Photorealistic single image super-resolution using a generative adversarial network. *In CVPR*, 4681-4690. DOI : 10.1109/CVPR.2017.19

[25] M. Zhu, D. He, X. Li, C. Li, F. Li, X. Liu, E. Ding & Z. Zhang. (2021). Image inpainting by end-to-end cascaded refinement with mask awareness. *In IEEE TIP*, 4855-4866. DOI : 10.1109/tip.2021.3076310

저 자 약 력

김성주(Kim, Seong Joo) [회원]



- 2021년 8월 : 한국외국어대학교 컴퓨터공학과 학사
- 2021년 3월~현재 : 한국외국어대학교 컴퓨터공학과 석사과정

E-Mail : seongjookim@hufs.ac.kr

<관심분야>
컴퓨터비전, 이미지 인페인팅, 이미지 복원, 딥러닝

최재영(Choi, Jae Young) [회원]



- 2011년 8월 : KAIST 전기및전자공학과 박사
- 2008년 11월~2012년 5월 : 토론토대학 연구원
- 2012년 7월~2013년 6월 : 펜실베니아대학 연구원
- 2013년 9월~2014년 8월 : 삼성전자 책임연구원
- 2014년 9월~2016년 8월 : 중원대학교 의료공학과 조교수
- 2023년 2월~2024년 2월 : 브리티시컬럼비아대학 컴퓨터공학과 방문교수
- 2016년 9월~현재 : 한국외국어대학교 컴퓨터공학과 부교수

E-Mail : jychoi@hufs.ac.kr

<관심분야>
딥러닝, 머신러닝, 패턴인식, 영상처리