특집논문 (Special Paper)
방송공학회논문지 제29권 제6호, 2024년 11월 (JBE Vol.29, No.6, November 2024)
https://doi.org/10.5909/JBE.2024.29.6.888
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

혼합 열화 영상 복원을 위한 2단계 U-Net 트랜스포머

김 태 환[a)], 김 문 철[a)‡]

Two-stage U-Net Transformer for Image Restoration with Mixed Degradation

Taehwan Kim[a)] and Munchurl Kim[a)‡]

요 약

최근 인공 신경망 기술의 발전으로 이를 응용하는 영상 복원 성능이 크게 향상되었다. 특히, 트랜스포머 기반 모델들은 영상 복원에 획기적인 성능을 보인다. 그러나 이들은 단일 열화에만 초점을 맞추어 더 복잡한 혼합 열화 영상 복원 문제는 다루지 않았다. 이를 해결하기 위해 본 연구에서는 Two-stage U-Net Transformer(TUT)를 제안한다. TUT는 혼합 열화 영상을 효과적으로 복원하기 위해 복원 과정을 두 단계로 나눈다. 1단계는 공간 열화, 2단계는 색상 열화를 처리하여 복원된 영상의 품질을 더욱 향상시킨다. 또한, 효과적인 특징은 강조하고 도움이 되지 않는 특징은 감소하는 목적의 조절자를 공간 및 채널 방향의 트랜스포머에 적용하여 복원한다. 끝으로, 혼합 열화 복원에 적합한 학습 전략과 손실 함수를 적용하여 복원 과정을 최적화한다. 다양한 데이터에서의 실험 결과, TUT는 혼합 열화 영상을 복원하는 데 있어 정량적 및 정성적으로 기존의 트랜스포머 기반 영상 복원 모델들을 능가하는 성능을 보였다. 본 연구의 결과는 TUT가 복잡한 영상 복원 작업의 실질적 응용에 새로운 방향성을 제시할 수 있음을 시사한다.

Abstract

Recent advancements in neural networks have markedly improved the performance of image restoration tasks. Transformer-based models have particularly distinguished themselves by achieving state-of-the-art results. However, despite their impressive capabilities, these models have predominantly focused on images with single types of degradation, leaving the more complex issue of mixed degradation largely unaddressed. In response to this gap, we introduce a novel approach: Two-stage U-Net Transformer (TUT). TUT is specifically designed to tackle the intricacies of images with mixed degradations by strategically dividing the restoration process into two stages where (i) spatial and (ii) color degradations are remedied subsequently. This division not only simplifies the problem but also enhances the restoration quality. Also, our model employs both spatial-wise and channel-wise Transformers, enhanced by modulators to amplify useful features while suppressing irrelevant ones. Finally, we optimize TUT with learning strategies and loss functions tailored to mixed degradation restoration. Extensive experiments show that our TUT significantly outperforms existing state-of-the-art models in restoring images affected by mixed degradations both qualitatively and quantitatively. The findings of this study highlight TUT as a robust and comprehensive solution for complex image restoration tasks, offering a new direction for future research and practical applications in this evolving field.

Keyword : Image Restoration, Mixed Degradation, Transformer, Computer Vision, Deep Learning

Ⅰ. 서 론

영상 획득 및 처리를 통한 TV 방송 및 실시간 스트리밍 기술의 발전으로 누구나 고품질의 영상을 즐길 수 있는 시대에 살고 있다. 이러한 고품질 영상 시청 증가로 인해 저품질 영상을 고품질 영상으로 변환하기 위한 영상 처리 기술에 대한 관심과 연구가 촉진되었다. 특히, 인공 심층 신경망 기반의 방법들이 ImageNet[1] 대회와 같은 고차원 비전 작업에 성공적으로 적용된 이후, 고품질 영상 변환에 대한 연구가 크게 활성화되어 왔다. 그 결과, 오래된 사진이나 열화가 포함된 영상을 고품질로 변환하려는 시도가 활발해지고 있다. 이러한 저품질 영상을 고품질로 변환하는 과정을 영상 복원이라고 하며, 이는 컴퓨터 비전 분야에서 중요한 연구 주제로 자리잡고 있다.

저품질 영상에는 다양한 열화 요인이 존재한다. 예를 들어, 잡음, 흐릿함, 손실 압축 등은 저품질 영상에서 흔히 나타나는 열화 요인이다. 많은 경우, 이러한 다양한 열화 요인들이 결합되어 저품질 영상을 형성하게 된다. 영상 복원의 궁극적인 목표는, 여러 열화 요인이 혼재된 저품질 영상에 대해 열화 요인들을 일관되게 제거하면서 고품질 영상으로 변환할 수 있는 기술을 개발하는 것이다. 이러한 복합적인 혼합 열화를 처리하는 것은 매우 어려운 작업이며, 이는 컴퓨터 비전 및 영상 처리 분야에서 지속적인 연구가 요구되는 이유이기도 하다.

최근에는 복원 성능을 극대화하기 위해 트랜스포머(Transformer)[2] 기반의 방법들이 제안되고 있다. 트랜스포머는 처음에 자연어 처리(Natural Language Processing, NLP)에서 제안된 기법으로, 셀프 어텐션(Self-Attention) 메커니즘을 통해 문맥 간의 관계를 전역적으로 탐색할 수 있어, 자연어 처리에서 강력한 도구로 자리잡았다. 또한, 영상 복원 분야에서도 트랜스포머의 셀프 어텐션 메커니즘을 활용하여 영상 내의 모든 화소 단위 영역을 전역적으로 탐색하고 최적의 특징을 추출하여 고품질 영상으로 변환하는 연구가 활발히 진행되어 왔다[3,4]. 더불어, 이러한 트랜스포머 기반 영상 복원 기법들이 최근 최고(State-of-the-Art, SOTA) 성능을 보이면서, 많은 연구자들로부터 주목받고 있으며, 다양한 응용 가능성을 열어주고 있다.

트랜스포머 기반 영상 복원의 잠재력은 매우 크지만, 현재까지의 연구는 주로 단일 열화 요인을 처리하는 데 초점이 맞춰져 있었다. 예를 들어, 잡음만 포함된 저품질 영상을 복원하는 모델이나 흐릿함만 포함된 영상을 복원하는 모델의 경우 등을 포함하여 각각의 특정 단일 열화 요인에만 대응할 수 있다. 이러한 방법들은 일반적으로 복합적인 열화 요인이 혼합된 상황에서는 성능이 보장되지 않는다. 또한, 실제 세계의 저품질 영상은 다양한 열화 요인이 복합적으로 포함된 경우가 많기 때문에, 단일 열화 요인에만 초점을 맞춘 모델들은 실제 열화 영상의 열화 상태를 적절히 반영하지 못한다는 문제가 있다. 이와 같이 실제 응용에서 기존 트랜스포머 기반 모델[3,4]의 성능이 제한될 수 있으며, 이를 극복하기 위한 연구가 필요하다.

이 문제를 해결하기 위해, 본 논문은 단일 모델을 통해 실제 상황과 유사한 혼합 열화를 처리할 수 있는 트랜스포머 기반의 새로운 영상 복원 방법을 제안한다. 제안된 방법은 두 단계의 U-Net 트랜스포머(Two-stage U-Net Transformer, TUT) 구조로 되어 있으며 혼합 열화가 포함된 영상을 효과적으로 처리할 수 있도록 설계되었다. TUT는 잡음, 흐릿함, JPEG 압축 열화와 같은 공간적 열화뿐만 아니라, 색 영역의 축소, 변화된 색조, 밝기, 채도, 대비와 같은 색상 열화까지도 복원할 수 있는 보다 범용적 영상 복원 모델이다.

본 논문의 주요 특징은 다음과 같다:

- 본 논문은 복잡한 혼합 열화 영상 복원 문제를 두 단계로 나누어 해결하는 2단계 U-Net 트랜스포머(Two-stage U-Net Transformer, TUT) 구조를 제안한다. 첫 번째 단계에서는 공간 열화를 처리하는 공간 트랜스

a) 한국과학기술원 전기및전자공학부(Dept. of Electrical Engineering, KAIST)
‡ Corresponding Author : 김문철(Munchurl Kim)
E-mail: mkimee@kaist.ac.kr
Tel: +82-42-350-7419
ORCID: https://orcid.org/0000-0003-0146-5419
※ 이 논문은 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No. 2017-000419, 스마트 방송 미디어를 위한 지능형 고실감 영상처리 연구).
※ 본 논문은 2024년 한국과학기술원(KAIST) 김태환의 석사학위 논문 “Transformer 기반 2단계 U-Net을 이용한 혼합 열화 영상 복원[40]”을 기반으로 작성됨.
※ 본 논문은 2024년 한국방송·미디어공학회 하계학술대회에서 발표된 “TUT: 혼합 열화 영상 복원을 위한 2단계 U-Net 기반 Transformer[41]” 논문의 확장 연구임.
· Manuscript September 5, 2024; Revised October 14, 2024; Accepted October 15, 2024.

포머 기반 U-Net과 두 번째 단계에서는 색상 열화를 처리하는 채널 트랜스포머 기반의 U-Net을 제안하여 효과적인 혼합 열화 복원을 가능하게 한다.

- 각 단계에 존재하는 트랜스포머 블록은 기존 트랜스포머 모델을 기반으로 설계되어 있지만, 성능 향상을 위해 각각의 트랜스포머 구조에 적합한 공간 및 채널 조절자(modulators)를 추가하여 설계하였다.
- 제안된 TUT는 혼합 열화 요인이 적용된 다양한 평가 데이터셋에서 기존의 트랜스포머 기반 영상 복원 모델들보다 정량적 및 정성적으로 우수한 성능을 보인다.

Ⅱ. 관련연구

1. 비전 트랜스포머를 이용한 영상 복원

다양한 비전 트랜스포머(Vision Transformer, ViT)[6] 변형들이 고차원 비전 문제를 해결하기 위해 개발됨에 따라, 영상 복원 분야에서도 여러 향상된 ViT 모델들[3,4]이 등장하여 기존의 Convolutional Neural Network(CNN) 기반 모델들[7]의 성능을 크게 능가하고 있다. 그 중 하나인 SwinIR[8]은 영상 복원 작업을 위해 설계된 Swin Transformer[9] 기반 모델로서, 트랜스포머[2]를 효과적으로 변형한 모델이다. SwinIR은 Swin Transformer의 이동 윈도우(Window) 내 셀프 어텐션 메커니즘을 활용한다. 비슷하게, Uformer[10]는 잘 알려진 U-Net[5] 구조를 기반으로 U자형 구조와 이동 윈도우 기반 셀프 어텐션을 결합하여 모델을 구성하였다.

이동 윈도우 기반 셀프 어텐션을 넘어, 또 다른 다양한 연구들은 효율적인 셀프 어텐션 메커니즘을 가진 트랜스포머 블록을 개발해왔다. FFTformer[11]는 Fast Fourier Transform(FFT)를 활용하여 신호를 공간 도메인과 주파수 도메인 간에 변환하는 방법을 사용하며, 주파수 도메인에서의 셀프 어텐션을 효과적으로 수행한다. 이는 푸리에 도메인에서의 주파수 성분간 곱셈이 공간 도메인에서의 컨볼루션과 동일하다는 원리를 이용한 것이다. Stripformer[4]는 영상을 수평 및 수직 스트립(Strip)으로 나눈 후, 각 스트립 내에서 1차원 로컬(Local) 셀프 어텐션을 수행하고, 그런 다음 각 스트립을 하나의 토큰(Token)으로 사용하여 글로벌(Global) 셀프 어텐션을 실행한다. 또 다른 주목할 만한 모델인 Restormer[3]는 전치(Transposed) 셀프 어텐션을 통해 채널(Channel) 간의 전역적인 상관관계를 탐색하며, 이를 통해 연산 복잡도를 크게 줄였고, 모션 디블러링(Motion Deblurring)과 같은 영상 처리에서 최고 성능을 입증했다. NAFNet[12]은 영상 복원의 기준을 제시하는 복잡한 어텐션 메커니즘을 제거하여 간소화된 모델이지만, 여전히 다양한 트랜스포머 모델 변형에서 사용된 특정 레이어를 사용한다. 최근에 DAT[13]는 채널별 셀프 어텐션과 공간별 셀프 어텐션을 순차적으로 결합한 모델을 도입하여 초해상화에서 최고 성능을 보였다. 이와 같은 연구들은 트랜스포머 기반 연구가 영상 복원에 있어 매우 중요하며, 눈에 띄는 성능 향상을 가져왔음을 보여주고 있다.

트랜스포머 기반 영상 복원에 대한 광범위한 연구에도 불구하고, 혼합 열화가 적용된 영상을 복원하는 것에 초점을 맞춘 연구는 아직 없었다. 각 트랜스포머 모델은 단일 유형의 열화를 처리하였기 때문에, 혼합 열화가 포함된 영상에 대해서는 최적의 성능을 발휘하지 못하는 결과를 초래한다.

2. 혼합 열화 영상 복원

트랜스포머의 잠재적 능력에도 불구하고, 혼합 열화에 대한 트랜스포머 기반 영상 복원 연구는 거의 이루어지지 않았다. 반면, CNN 기반 기존 일부 모델들은 이러한 복합 열화 문제를 다루었다. Liu[14]는 각기 다른 열화 유형을 목표로 하는 여러 분기를 통해 영상을 재귀적으로 복원하는 스템 네트워크(Stem Network)를 제안하였다. 그러나 이 접근법은 명확하게 구분되지 않는 복잡한 혼합 열화 문제를 처리하는 데 한계를 보인다. Kim[15]은 영상의 다른 영역에 다양한 유형의 열화가 결합된 새로운 데이터셋을 제시하고, Mixture of Experts(MoE)[16] 방식을 적용하여 공간적으로 이질적인 왜곡을 해결하는 모델을 제안하였다. 이 모델은 여러 CNN 기반 네트워크를 병렬로 사용한다. 이와 유사하게, Shin[17]은 CNN 기반의 공간적 특징 변환(Spatial Feature Transform, SFT) 레이어[18]를 활용하여 공간적으로 이질적인 열화에 대한 복원 강도를 조절하는 방식을 사용하였다.

그러나 이러한 CNN 기반 방법들은 지역적으로만 특징

을 추출하기 때문에 전역적인 특징을 추출하는 트랜스포머의 높은 잠재력을 활용하지 못한다는 한계를 가지고 있다. 또한, 이들 접근법은 주로 공간적 열화에만 초점을 맞추고 있어, 색상 열화, 명암 왜곡과 같은 실제 세계의 복잡한 열화 시나리오를 완전히 반영하지 못하며, 이에 따라 복원 성능에 한계를 보인다.

Ⅲ. 제안하는 2단계 U-Net 트랜스포머 (Two-stage U-Net Transformer)

1. 전체 구성도

기존의 트랜스포머[2] 기반 모델을 혼합 열화 영상 복원에 적용할 경우, 다양한 열화 유형의 복잡성으로 인해 성능이 저하되는 경우가 많다. 이를 해결하기 위해, 본 논문에서는 새로운 트랜스포머 기반 아키텍처인 두 단계 U-Net 트랜스포머(Two-stage U-Net Transformer, TUT)를 제안한다.

그림 1에서 볼 수 있듯이, 제안된 TUT는 두 개로 직렬 연결된 트랜스포머 기반 U-Net 아키텍처로 구성되어 있다. U-Net[5] 구조는 일반적으로 인코더(Encoder) 단계에서 공간 크기를 줄이고 채널 수를 늘린 특징을 추출하고, 디코더(Decoder) 단계에서 이를 복원하여 영상 복원에 필수적인 특징을 학습하는 구조이다. 기존의 기본 U-Net과 달리, 제안된 TUT는 디코더 단계에서 업샘플링(Up-sampling) 시 체커보드(Checker Board) 현상을 최소화하기 위해 픽셀 셔플 레이어(Pixel-shuffle layer)[20]를 사용한다. 단일 트랜스포머 기반 U-Net이 열화 정도가 심한 영상을 복원하는 데 제약이 있음을 고려하여, 제안된 TUT는 트랜스포머 기반 U-Net을 2개 직렬 연결하여 두 단계 방식을 채택하여 복잡한 열화 문제를 보다 효과적으로 해결하고자 한다.

TUT는 입력된 RGB 영상을 YCbCr 색 공간으로 변환하여, Y 채널(흑백 영상)과 Cb 및 Cr 채널(색상 정보)로 분리한다. Y 채널은 1단계 트랜스포머 기반 U-Net인 Gray-scale U-Net(GUN)에 의해 처리되며, 흑백 영상의 공간적 정보를 복원한다. GUN의 출력은 열화가 여전히 존재하는 Cb 및 Cr 채널과 결합된 후, 1단계 트랜스포머 기반 U-Net인 Color-scale U-Net(CUN)이 이러한 결합된 채널을 사용하여 색상 정보를 복원한다. 마지막으로, 영상을 YCbCr 색 공간에서 RGB로 다시 변환하여 완전히 복원된 영상을 얻는다. 이 두 단계 U-Net 프레임워크(Framework)는 혼합 열화 영상에서 색상과 공간적 열화를 효과적으로 복원한다.

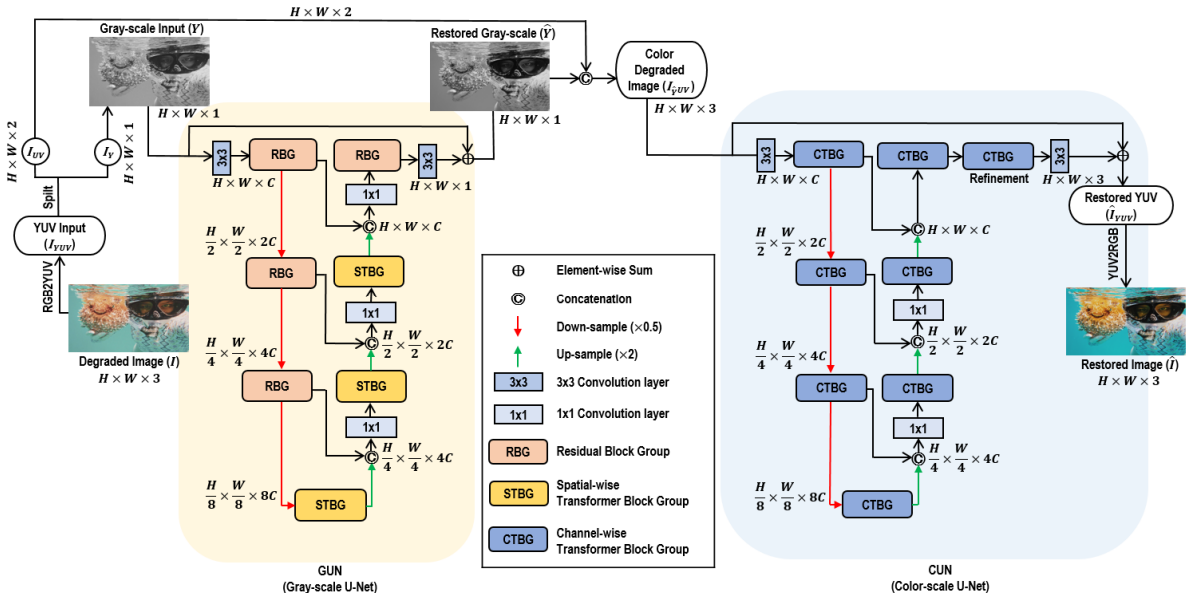

그림 1. 제안하는 2단계 U-Net 트랜스포머의 전체 구성도
Fig. 1. Overall Architecture of Two-stage U-Net Transformer

2. 공간적 열화 처리를 위한 Gray-scale U-Net

흑백 영상 복원을 위해 설계된 트랜스포머 기반 U-Net인 Gray-scale U-Net(GUN)은 제안된 TUT의 첫번째 단계의 트랜스포머 기반 U-Net 구조로서, 영상의 공간 정보를 추출하고 복원하는 데 중점을 두며, 영상의 구조 정보(Image Structural Information)에 열화가 있을 경우 인간의 시각 시스템이 색상 채널인 Cb와 Cr보다 흑백 채널인 Y의 변화에 더 민감하기 때문에 매우 중요하다[21]. 이를 반영하기 위해, 제안 모델은 인코더 및 디코더에서 Residual Block Group

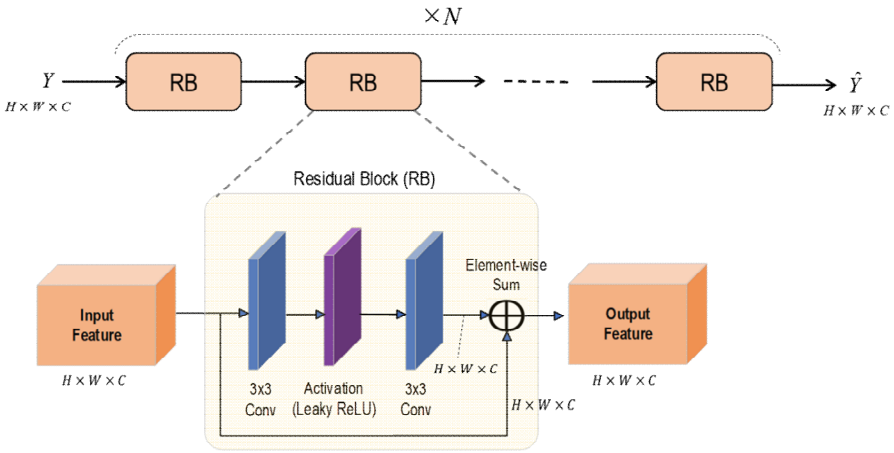

그림 2. GUN 인코더의 구성요소인 Residual Block Group (RBG) 구조
Fig. 2. Structure of Residual Block Group (RBG), components of GUN Encoder

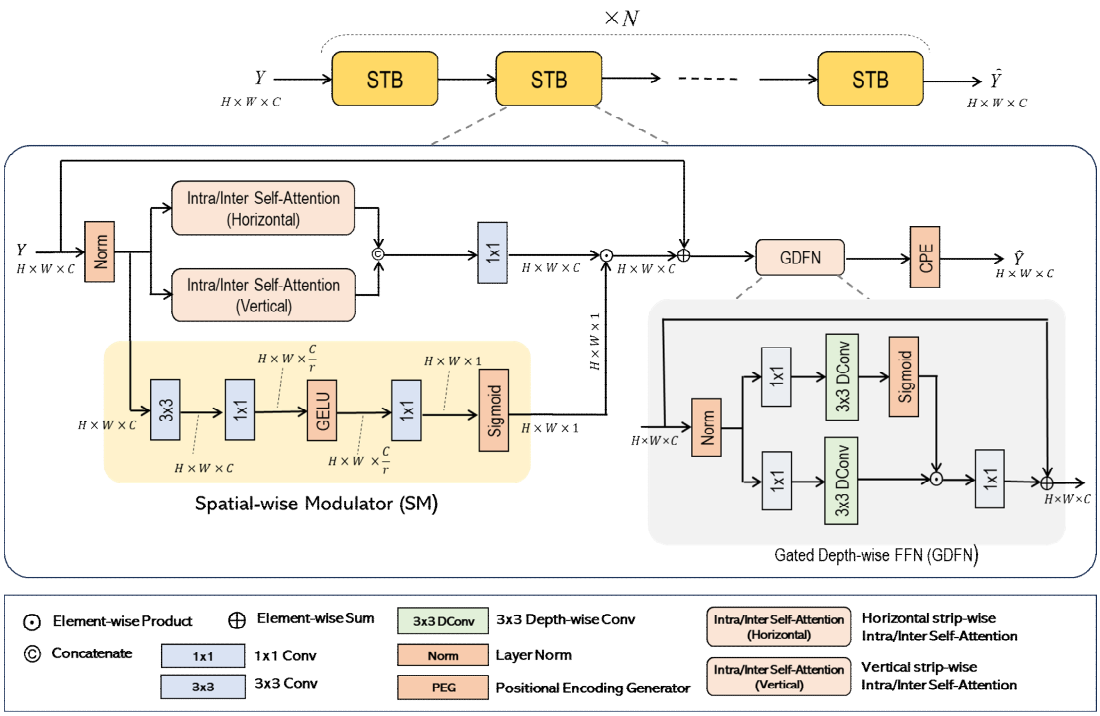

그림 3. GUN 디코더의 구성요소인 Spatial-wise Transformer Block Group (STBG) 구조
Fig. 3. Structure of Spatial-wise Transformer Block Group (STBG), components of GUN Decoder

(RBG)와 Spatial-wise Transformer Block Group(STBG)를 효과적으로 연결하여 복잡도 대비 복원 화질 성능 간의 균형을 맞추었다. RBG는 여러 개의 Residual Block(RB)으로 직렬 연결된 구조를 가지고 있고(그림 2), STBG는 여러 개의 Spatial-wise Transformer Block(STB)이 순차적으로 연결된 구조를 가진다(그림 3).

인코더 단계에서 GUN은 RBG를 사용하며, RBG는 그림 2와 같이 컨볼루션 레이어[7]와 잔차 연결[20]로 구성된 RB 여러 개를 직렬 연결한 구조이다. STBG도 사용할 수 있지만, 본 논문에서는 RBG가 인코더 단계에서 더 효율적이고 성능이 우수하다는 것을 실험적으로 확인하였다. 최근 연구에서도 영상 공간/구조 정보에 대한 특징 추출을 위한 U-Net 인코더에서는 트랜스포머 블록을 사용하지 않는 경향이 있다[11]. 그 이유는 두 가지이다. 첫째, 본 논문에서의 U-Net 인코더는 영상 복원을 위한 최적의 특징을 추출하는 것을 목표로 한다. 모든 화소 간의 상관관계를 계산하는 트랜스포머의 공간적 셀프 어텐션 메커니즘은 얕은 계층의 인코더(Shallow-layer Encoder)를 이용하면 깨끗한 특징 값 추출이 어렵기 때문에 인코더에 적용하는 것이 효과적이지 않고, 대신 깊은 계층에서 깨끗한 특징 값이 추출되어 존재할 가능성이 높기 때문에 디코더에 적용하는 것이 효과적인 것으로 알려져 있다[11]. 따라서, 제안된 GUN의 인코더는 RBG만으로 설계되었다. 둘째, RBG는 셀프 어텐션의 높은 연산 복잡도를 줄여주어, 더 빠른 학습과 추론 속도를 제공하고 메모리 사용량도 감소시킨다. 또한, U-Net 디코더의 마지막 단계에서도 특징지도(Feature Map)의 공간 해상도가 영상의 입력 해상도와 동일하여 셀프 어텐션 수행 시에 연산량이 크게 증가하는 것을 피하기 위해 RBG를 사용하였다.

GUN의 디코더 단계에서는 영상의 구조 정보에 해당하는 공간적 열화를 효과적으로 제거하기 위해 Spatial-wise Transformer Group(STBG)를 사용하였다. 여기서 낮은 연산량 복잡도에도 높은 성능을 발휘하는 Stripformer[4]의 스트립 단위 셀프 어텐션을 사용하였는데, 이는 제안 모델에 Swin Transformer[9]의 분할된 윈도우 영역 내 2차원 셀프 어텐션 보다 실험적으로 더 높은 성능을 보였기 때문이다. 그러나 스트립 단위 셀프 어텐션은 영상을 가로 또는 세로 방향으로 나누어 1차원 셀프 어텐션을 스트립 내에서 수행한 후, 스트립 간의 셀프 어텐션을 수행하는 방식이므로 서로 다른 스트립 내의 모든 토큰(Token)들의 전역적인 관계를 고려하는 데 한계가 있어 성능 저하로 이어질 수 있다. 따라서 특징 추출을 더욱 향상시키기 위해, Transformer에 Spatial-wise Modulator(SM)을 통합하였다. 이전 연구들[13,22,23]에서와 같이, GUN의 SM 구조(그림 3)에는 3×3 컨볼루션, 채널 수를 줄이기 위한 두 개의 1×1 컨볼루션, GELU 활성화 함수[24] 그리고 가중치를 생성하는 시그모이드 함수를 포함한다. 이러한 가중치는 셀프 어텐션의 출력을 조정하여(식 1), 유용한 특징을 강화하고 불필요한 특징을 감소시킨다.

$$SM(Y) = Strip\_SA(Y) \odot s\left(W'' g\left(W' Y\right)\right) \tag{1}$$

식 (1)에서, Y는 계층 정규화 이후의 입력 특징지도, $Spatial_SA(\bullet)$는 스트립 단위 셀프 어텐션, $s(\bullet)$는 Sigmoid, $g(\bullet)$는 GELU, $W'(\bullet)$는 1×1 conv, $W''(\bullet)$는 1×1 conv를 나타낸다.

그림 3에 보인 바와 같이, 제안된 전체 STBG의 구조는 Restormer[3]에서 사용된 Gated Depth-wise Feed-Forward Network(GDFN)과 Spatial Modulator(SM)을 결합한 여러 개의 STB가 직렬 연결화된 형태이다. GDFN은 게이팅(Gating) 기능과 1×1 컨볼루션 그리고 Depth-wise Convolution[25]을 사용하여 지역적으로 특징을 선택한다. STBG의 각 STB의 처리 과정에는 Layer 정규화[26], 가로/세로 스트립 단위 1차원 셀프 어텐션, 1×1 컨볼루션을 통한 채널 혼합 그리고 SM을 통한 스케일링이 포함된다(식 1). 이후에는 잔차 연결과 GDFN을 통한 특징 강화가 이루어진다. 마지막으로, 각 스트립 간의 위치 정보를 잃지 않기 위해 Depth-wise Convolution 층을 활용한 Conditional Positional Encoding(CPE)[27]를 통해 공간적 위치 정보를 저장한다. 이러한 간결한 접근 방식은 이전 방법들의 한계를 해결하면서, 흑백 영상 복원에 효과적이다.

3. 색 열화 복원을 위한 Color-scale U-Net

제안한 TUT의 Color-scale U-Net(CUN)은 2단계 트랜스포머 기반 U-Net 구조로서, Y 채널을 통한 복원된 영상 구

조 정보와 열화가 남아있는 색상 정보(Cb, Cr 채널)를 함께 사용하여 색상 열화 복원 및 영상 구조정보 세부조정을 위한 학습에 중점을 둔다. 색상 정보를 효율적으로 복원하기 위해, 복원된 Y 채널과 열화를 처리하지 않은 Cb와 Cr 채널을 합칠 경우, 공간 정보의 열화는 적기 때문에 Spatial-wise Transformer Block Group(STBG) 대신 Channel-wise Transformer Block Group(CTBG)를 사용한다. CTBG는 복원된 Y와 열화를 처리하지 않은 Cb, Cr 채널 간의 관계를 효과적으로 분석하여 색상 영상 복원을 위한 필수적인 특징을 추출한다.

STBG와 달리, CTBG는 공간적 상관관계를 고려하지 않기 때문에 CUN의 인코더 및 디코더의 모든 단계와 혼합 열화를 최종적으로 정제하기 위한 정제(Refinement) 블록에서 모두 사용된다. CTB는 Restormer[3]의 Multi-Head Transposed Self-Attention을 통합하여, 셀프 어텐션 과정에서 Query, Key, Value 프로젝션(Projection)을 전치(Transpose)하여 채널 간 유사성을 전역적으로 비교한다. Multi-Head Transposed Self-Attention은 낮은 연산 복잡도의 장점을 가지지만, 단점으로는 셀프 어텐션을 헤드(Head) 수에 따라 나누어진 각 채널 그룹(Group) 내에서만 수행하기 때문에, 효과적인 특징 추출에 제한이 있다. 이러한 한계를 해결하기 위해, 제안된 CTBG의 CTB에는 STB에서의 조절자와 유사하게 Channel-wise Modulator(CM)를 식 2와 같이 추가하였다.

$$CM(I) = Transposed\_SA(I) \odot s\left(W'' g\left(W'(GAP(I))\right)\right) \quad (2)$$

식 2에서, I는 계층 정규화 이후의 입력 특징지도, $Transposed_SA(\bullet)$는 다중 헤드 전치 셀프 어텐션, $GAP(\bullet)$는 Global Average Pooling, $s(\bullet)$는 Sigmoid, $g(\bullet)$는 GELU, $W'(\bullet)$는 1×1 conv, $W''(\bullet)$는 1×1 conv을 나타낸다.

CM은 전치 셀프 어텐션의 결과를 채널별 가중치로 조정한다. 이 과정은 3×3 컨볼루션과 Global Average Pooling[28]을 적용한 후, 두 개의 1×1 컨볼루션과 GELU 활성화 함수를 거쳐, 마지막으로 시그모이드 함수를 사용하여 가중치를 생성하는 방식으로 진행된다. 이러한 가중치는 셀프 어텐션 결과를 조정하여 유용한 채널을 강화하고, 유용하지

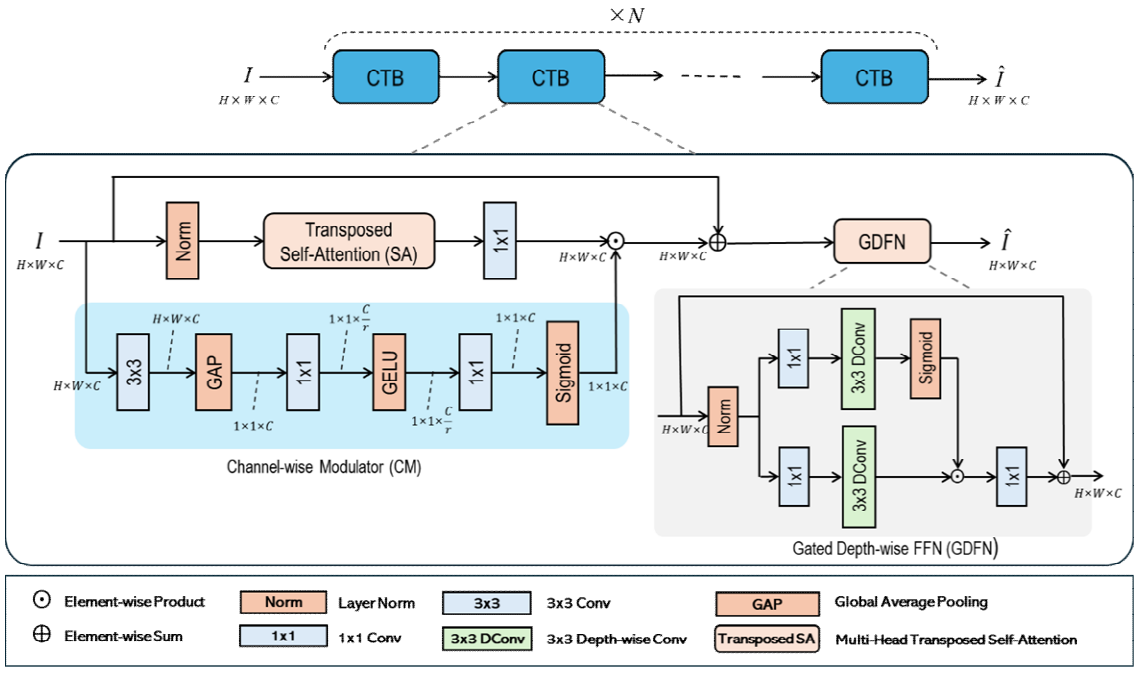

그림 4. CUN의 구성요소인 Channel-wise Transformer Block Group (CTBG) 구조
Fig. 4. Structure of Channel-wise Transformer Block Group (CTBG), components of CUN

않은 채널은 약화시킨다. 또한, Global Average Pooling을 통하여 전역 공간 정보를 단일 화소 정보로 축약하고 채널 간의 상관 관계를 학습하기 때문에 제한된 전역적 공간적 특징 추출 성능을 보완한다.

그림 4에서 볼 수 있듯이, 전체 CTBG 구조는 여러 개의 CTB를 직렬 연결하여 구성된다. 각 CTB는 Layer 정규화, 셀프 어텐션, 1×1 컨볼루션 레이어와 잔차 학습을 포함하며, 최종 출력을 생성하기 위해 Restormer의 Gated Depth-ise Feed-Forward Network(GDFN)를 사용하여 특징을 더욱 정제한다. STBG와 달리, CTBG는 채널 위치가 고정된 순서를 따르지 않으며, 중요하지 않기 때문에 STBG의 STB에서 사용하였던 위치 인코딩(Positional Encoding)을 포함하지 않는다.

전체적으로, GUN에서의 공간적인 열화를 제거하는 STBG에서는 스트립 단위의 셀프 어텐션 메커니즘의 단점을 공간 조절자가 보완하여 성능을 극대화하고, CUN의 CTBG에서는 열화된 색상 정보를 복원하면서 영상의 구조 정보를 정제하기 위해서 그룹 단위의 채널 방향의 셀프 어텐션의 단점을 채널 조절자가 보완하여 복원 능력을 향상시킨다.

4. TUT의 학습 전략과 손실 함수

제안된 TUT는 성능을 극대화하기 위해 두 단계 학습 전략을 채택한다. 첫 번째 단계에서는 GUN이 사전 학습되며, 두 번째 단계에서는 CUN이 학습하는 동안 GUN은 사전 학습된 가중치를 미세 조정(Fine-tuning)한다. 이러한 미세 조정은 특정 열화가 공간 열화와 색상 열화를 모두 포함하기 때문에 필요하다. 이 두 단계 접근 방식을 채택함으로써, TUT는 다른 트랜스포머 기반 영상 복원 모델들보다 우수한 성능을 달성하며, 다양한 열화 유형을 학습하여 처리하게 되므로 더욱 효과적인 복원 모델을 제공한다.

혼합 열화 영상 복원에 최적화된 TUT를 구성하기 위해, 각 단계에서 특정 손실 함수가 적용된다. 이러한 맞춤형 손실 함수는 복원 품질을 크게 향상시켜, 각 학습 단계가 효과적으로 진행되도록 한다. GUN을 사전 학습시키는 첫 번째 단계에서는 L1 손실 $\mathscr{L}_1^{1st}$(식 4)을 통해 흑백 영상(Y-Channel)을 전체적으로 복원하고 2D Fast Fourier Transform(FFT)을 사용하여 주파수 도메인으로 변환한 결과의 실수부와 허수부를 연결하여 L1 손실을 계산하는 FFT 손실 $\mathscr{L}_{FFT}^{1st}$(식 5)[11]을 적용하며 필요한 주파수 정보를 복원하며 5×5 Laplacian 커널과 Charbonnier 손실을 이용하여 Edge 손실 $\mathscr{L}_{edge}^{1st}$(식 6)[4,29]을 계산하여 세부 정보를 복원한다. 첫 번째 단계의 총 손실 함수(식 6)는 L1 손실, FFT 손실, Edge 손실의 가중합인 $\mathscr{L}_{1st}$로서 다음과 같다.

$$\mathscr{L}_{1st} = \mathscr{L}_1^{1st} + \alpha \mathscr{L}_{FFT}^{1st} + \beta \mathscr{L}_{edge}^{1st} \tag{3}$$

여기서, α와 β는 가중치 계수로서 $\alpha = 0.1$, $\beta = 0.1$를 사용하였다.

$$\mathscr{L}_1^{1st} = \frac{1}{HW} \sum_{i=1}^{HW} |Y_{GT} - \hat{Y}_{GUN}| \tag{4}$$

$$\mathscr{L}_{FFT}^{1st} = \frac{1}{HW} \sum_{i=1}^{HW} |concat(FFT_{real}(Y_{GT}), FFT_{imag}(Y_{GT})) - concat(FFT_{real}(\hat{Y}_{GUN}), FFT_{imag}(\hat{Y}_{GUN}))| \tag{5}$$

$$\mathscr{L}_{edge}^{1st} = \frac{1}{HW} \sum_{i=1}^{HW} \sqrt{(|\Delta(Y_{GT}) - \Delta(\hat{Y}_{GUN})|)^2 + \mathscr{E}^2} \tag{6}$$

여기서, H와 W는 입력 영상의 가로 및 세로 화소수를 각각 나타내며 Y_{GT}와 $\hat{Y}_{GUN}$는 정답 원본 영상의 Y 채널과 GUN을 통해 복원된 Y 채널을 각각 나타낸다. 또한, 식 6에서, Δ는 5×5 커널의 라플라시안 연산자, $\mathscr{E}$는 10^{-3}이다.

GUN의 미세조정 학습과 CUN의 학습을 동시에 수행하는 두 번째 단계 학습에서는, 첫 번째 단계의 손실 함수에 곱셈 가중치를 적용하여 추가적인 미세 조정을 진행하며, RGB 색상 공간에서 색상 영상을 복원하기 위해 RGB의 총 3채널간의 L1 손실(식 7), FFT 손실(식 8) 그리고 Edge 손실 함수(식 9)를 통한 비교가 추가된다. 따라서, 두 번째 단계에서의 총 손실 함수(식 10)는 1단계 손실 함수들과, 2단

계의 L1 손실, FFT 손실, Edge 손실의 가중합으로서 $\mathscr{L}_{2nd}$으로 정의되며 다음과 같이 구성된다.

$$\mathscr{L}_{2nd}= \mathscr{L}_1+\alpha\mathscr{L}_{FFT}+\beta\mathscr{L}_{edge}+\gamma\mathscr{L}_{1st} \tag{7}$$

여기서, 가중치 계수는 $\alpha= 0.1$, $\beta= 0.1$, $\gamma= 0.1$으로 설정하였다.

$$\mathscr{L}_1= \frac{1}{3HW}\sum_{i=1}^{3HW}|I_{GT}- \hat{I}_{CUN}| \tag{8}$$

$$\mathscr{L}_{FFT}= \frac{1}{3HW}\sum_{i=1}^{3HW}|concat(FFT_{real}(I_{GT}), FFT_{imag}(I_{GT})) - concat(FFT_{real}(\hat{I}_{CUN}), FFT_{imag}(\hat{I}_{CUN}))| \tag{9}$$

$$\mathscr{L}_{edge}= \frac{1}{3HW}\sum_{i=1}^{3HW}\sqrt{(|\Delta(I_{GT})-\Delta(\hat{I}_{CUN})|)^2+\mathscr{E}^2} \tag{10}$$

여기서, H와 W는 입력 영상의 가로 및 세로 화소수를 각각 나타내며 I_{GT}와 $\hat{I}_{CUN}$는 RGB 도메인에서의 정답 원본 영상과 CUN을 통해 최종 복원된 RGB 도메인에서의 영상을 각각 나타낸다. 또한, 식 10에서, Δ는 5×5 커널의 라플라시안 연산자, $\mathscr{E}$는 10^{-3}이다.

TUT의 2단계 학습 전략 및 학습에 사용된 손실 함수의 조합은 혼합 열화 복원 목적에 맞게 효율적이고 효과적인 학습을 유도하여, 우수한 혼합 열화 영상 복원 성능을 발휘할 수 있게 한다.

Ⅳ. 실험 결과

1. 데이터셋과 구현 세부사항

본 논문에서 제안한 방법 및 비교 방법들에 대한 실험을 수행하기 위해, 실제 복합열화를 포함하고 있는 대용량의 과거(방송) 영상(legacy data)들을 획득하는 것이 현실적으로 어렵기 때문에, 딥러닝 연구를 위해 공개된 영상 데이터셋을 사용하여 다양한 전처리 기법을 통해 실제 영상 열화를 시뮬레이션 하여 복합 열화 데이터셋을 만들고, 이를 학습 및 평가에 사용하였다. 그림 5는 일반적인 단순 열화 과정과 본 논문에서 적용한 복합 열화 전처리 과정을 도식으로 비교하였다. 일반적인 단순 열화 과정은 간단한 3가지의 열화만 이용하는 합성 열화 전처리 과정에 사용해 왔지만[30], 본 실험에서는 총 9가지의 다양한 합성 열화를 생성하여 현실세계의 열화와 유사하도록 재현하였다. 먼저 좁아

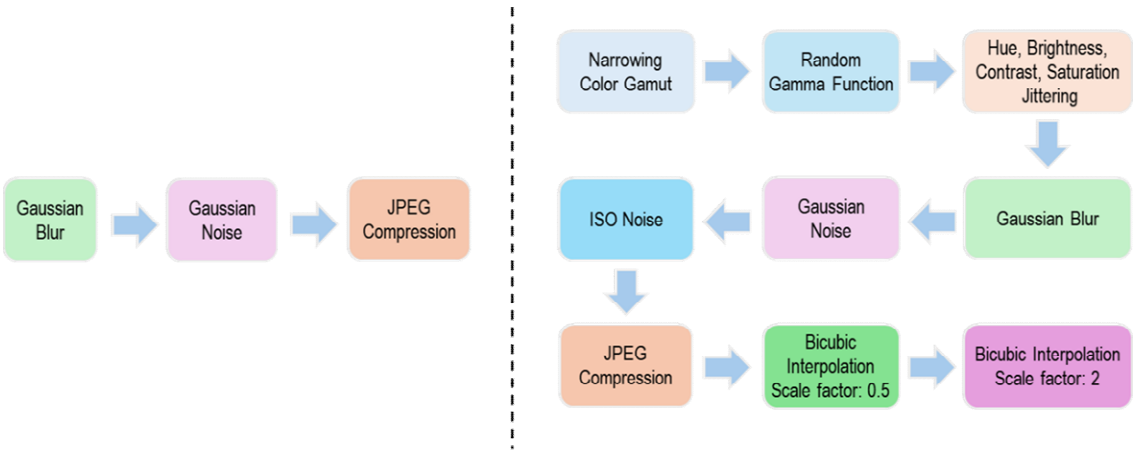

그림 5. 일반적인 단순 합성 열화 생성 과정 (좌측)과 본 실험에 사용한 혼합 열화 생성 과정 (우측)
Fig. 5. Typical synthetic degradation pipeline (left) / mixed degradation pipeline used in the experiment (right)

진 색 영역을 재현하기 위하여 sRGB에서 표준 D50 화이트 포인트를 가진 삼각형의 세 꼭짓점을 xyY 색 공간에서 (0.64, 0.33), (0.30, 0.60), (0.15, 0.06)으로 변환하여 색 영역을 좁히는 작업을 진행하였고, 그 이후, 열화 생성 도구 Albumentations[31]를 사용하여 감마 함수 조정, 색조/밝기/대비/채도 조절, 가우시안 흐릿함 및 잡음 생성, ISO 잡음, JPEG 압축 열화를 합성하였다. 마지막으로 쌍입방 보간(Bicubic Interpolation) 해상도 축소(축소 크기 0.5배)와 그 후에 쌍입방 보간 해상도 확대(확대 크기 2배)를 사용하여 세부 사항을 제거하였다.

사용한 데이터셋으로는 학습 시에는 DIV2K[32] 데이터셋을 사용하였고, 평가 시에는 Set5[33], Set14[34], Urban100[35], BSD100[36] 그리고 DIV2K 검증 데이터셋을 사용하였다. 실험은 RTX 4090 GPU 단일 장치를 활용하였으며, 배치 크기는 12로 설정하고, 64×64 크기의 패치를 무작위로 뒤집고 잘라서 학습에 사용하였다.

그림 1의 제안 모델에 사용된 하이퍼 파라미터(Hyperarameter)는 다음과 같다: GUN 인코더 단계에서는 각 단계의 RBG에는 [3, 3, 3]개의 RB를 사용하였다(그림 2). GUN 디코더 단계에서의 STBG에는 단계별로 [6, 4, 3]개의 STB를 [4, 3, 2]개의 헤드와 함께 사용하였다(그림 3). GUN 디코의 마지막 단계에서는 3개의 RB로 이루어진 RBG를 사용하였다(그림 2). SM의 채널 축소 비율은 16으로 설정하였다(그림 3). CUN의 각 단계에서 CTBG는 [4, 6, 6, 8, 6, 6, 4]개의 CTB를 사용하였으며(그림 4), 정제(Refinement) 블록에서의 CTBG는 4개의 CTB를 사용하였다(그림 4). 이 때, 각 단계의 헤드 수는 [1, 2, 4, 8, 4, 2, 1]이었고, 정제 블록에서는 2개의 헤드를 사용하였다. CM의 채널 축소 비율은 8로 설정하였다. 그림 1의 제안 모델 내부 U-Net의 채널 구성은 GUN에서 [1→48, 96, 192, 384, 192, 96, 48→1], CUN에서 [3→48, 96, 192, 384, 192, 96, 96→3]로 설정하였다. 학습률(Learning Rate)은 첫 번째 단계에서 2e-4로 설정되었으며, [140, 160, 180, 190, 195] 에포크(Epoch)에서 1/2 비율로 선형 감소하도록 구성하였다. 두 번째 단계에서는 GUN의 학습률은 에포크에 상관없이 1.25e-5로 고정되었으며, CUN의 학습률은 2e-4로 설정되고 [140, 160, 180, 190, 195] 에포크에서 1/2 비율로 선형 감소하도록 설정하였다. Adam[37] 옵티마이저(Optimizer)의 파라미터는 두 단계 모두에서 $\beta_1 = 0.9$, $\beta_2 = 0.999$로 설정되었다.

데이터 전처리 과정에서의 Albumentations 하이퍼 파라미터는 다음과 같다: 임의의 감마(Random Gamma): gamma_limit = (90, 110), 색 조절(Color Jittering): brightness = 0.1, contrast = 0.1, saturation = 0.1, hue = 0.025, 가우시안 흐릿함(Gaussian Blur): blur_limit = (3, 9), sigma_limit = 0, 가우시안 잡음(Gaussian Noise): var_limit = (10.0, 50.0), mean = 0, per_channel = True, ISO 잡음(Noise): color_shift = (0.01, 0.05), intensity = (0.1, 0.5), 영상 압축(Image Compression): quality_lower = 70, quality_upper = 95, compression_type = 0, always_apply = True.

2. 정량적 분석

본 논문에서는 제안된 TUT 모델의 효과성을 입증하기 위해, DAT[13], NAFNet[12], Restormer[3], Stripformer[4], SwinIR[8], Uformer[10]를 포함한 최고 성능의 최신 트랜스포머[2] 기반 모델들과 정량적으로 성능을 비교하였다. 사전 학습된 기존의 공개 모델들은 혼합 열화 영상에 대하여 학습을 진행하지 않았기 때문에, 공정한 비교를 위해 합성된 혼합 열화 영상 데이터셋을 사용하여 재학습을 진행하였다. 재학습 시에, 모든 모델에 대해 옵티마이저, 학습률, 학습률 감소, 총 학습 반복 횟수를 통일하였으며, 손실 함수와 블록의 수는 각 모델의 공식 문서에 명시된 대로 유지하였다. DAT의 학습은 GPU 메모리 제약으로 인해 배치 크기 10으로 1개의 NVIDIA DGX A100에서 진행되었다. 학습시에 모델 가중치는 DIV2K 검증 데이터셋에서 PSNR(Peak Signal-to-Noise Ratio)이 높은 기준으로 5 에포크마다 저장되었다. 추론 시에는 PSNR과 SSIM(Structural Similarity Index Measure)을 RGB 색상 공간에서 측정하여 성능 평가 지표로 사용하였다.

표 1에서 볼 수 있듯이, 영상에 심한 혼합 열화가 적용되어 전체 평가 데이터셋에서 평균 PSNR이 21.43dB, SSIM이 0.634으로 낮은 경우에도 TUT는 비교 모델들의 복원 결과보다 PSNR과 SSIM이 수치적으로 능가함을 보였다. 특히, 전역적으로 반복되는 패턴을 포함하고 있어 트랜스포머 기반 영상 복원에 가장 신뢰할 수 있는 평가 데이터셋으로 간주되는 Urban100 데이터셋에서 TUT는 뛰어난 성

표 1. 평가 데이터셋에 대한 정량적 분석 결과 (빨강: 최고 성능 모델, 파랑: 두번째로 최고 성능 모델)
Table 1. Qualitative results on test datasets (RED: Best Model, BLUE: Second-Best Model)

	Set5[33] (#5)		Set14[34] (#14)		Urban100[35] (#100)		BSD100[36] (#100)		DIV2K[32] Val. (#100)		Total Dataset (#319)	
Metric ↑ / Model	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Degraded Image (LQ)	22.71	0.679	22.05	0.661	20.22	0.624	22.66	0.640	21.25	0.631	21.43	0.634
DAT[13] (ICCV 2023)	25.79	0.825	25.25	0.791	24.92	0.810	24.45	0.771	25.13	0.791	25.18	0.791
NAFNet[12] (ECCV 2022)	26.28	0.827	24.65	0.773	23.95	0.778	25.23	0.761	24.51	0.767	24.59	0.770
Restormer[3] (CVPR 2022)	26.19	0.846	25.08	0.788	25.09	0.812	25.45	0.772	25.09	0.791	25.22	0.793
Stripformer[4] (ECCV 2022)	26.63	0.835	25.00	0.770	24.14	0.774	25.40	0.754	24.79	0.763	24.81	0.765
SwinIR[8] (ICCVw 2021)	26.37	0.830	24.64	0.779	23.73	0.786	25.16	0.756	24.29	0.773	24.43	0.773
Uformer[10] (CVPR 2022)	25.92	0.832	24.60	0.773	23.63	0.781	24.94	0.753	24.05	0.767	24.25	0.768
TUT (Ours)	27.46	0.852	25.79	0.792	25.54	0.825	25.66	0.777	25.38	0.797	25.57	0.800

능을 보였다. 이 데이터셋에서 TUT는 PSNR 25.54dB, SSIM 0.825를 달성하여, Stripformer(PSNR: 24.14dB, SSIM: 0.774)와 Restormer(PSNR: 25.09dB, SSIM: 0.812)보다 유의미하게 높은 성능을 기록하였다. 전체 평가 데이터셋에서는 TUT(PSNR: 25.57dB, SSIM: 0.800)는 두 번째로 우수한 모델인 Restormer를 PSNR에서 0.35dB, SSIM에서 0.007 차이로 능가하였다. 이러한 결과를 통하여 TUT의 정량적으로 우수함을 입증하며, 혼합 열화 복원 문제에 대하여 공간 및 색상의 2단계 복원으로 나누어 처리하는 방식이 효과적임을 확인 가능하다.

3. 정성적 분석

그림 6과 그림 7은 제안된 TUT 모델과 최신 트랜스포머 기반 영상 복원 모델들을 비교한 정성적 결과를 나타낸다. 정성적 분석에서도 정량적 실험에서 사용된 모델과 하이퍼파라미터를 동일하게 사용하였으며 비교를 용이하게 하기 위해, 복원 결과는 정답 원본 영상에서 빨간색 사각형으로 표시된 확대된 부분에 초점을 맞추어 제공된다.

그림 6의 Urban100 데이터셋은 복잡한 전역적 패턴과 세부 사항이 포함되어 있어, 트랜스포머 기반의 복원 모델 성능을 평가하는 데 있어 신뢰할 수 있는 데이터셋으로 간주된다. 그림 6에서 볼 수 있듯이, Urban100 데이터셋의 96번째 영상 결과에서 다른 모델들은 반복적인 쌍입방 보간(Bicubic Interpolation) 해상도 축소 및 확대 과정으로 인해 건물의 네모난 유리나 창틀과 같은 부분에서 심각한 에일리어싱(Aliasing) 현상을 보였다. 이러한 현상은 특히 복잡한 선이나 패턴이 반복되는 부분에서 더욱 두드러지게 나타나며, 이는 영상의 품질을 크게 저하시킨다. 그러나 TUT는 에일리어싱이 거의 발생하지 않았고, 깨끗하고 선명한 선을 복원하여 정답 영상과 매우 유사한 결과를 제공하였다. 이러한 결과는 TUT가 공간 정보를 복원하는 데 있어 뛰어난 성능을 발휘하며, 특히 제안된 Gray-scale U-Net(GUN)이 공간적 열화가 혼합된 상황에서도 효과적으로 열화를 처리할 수 있는 능력을 지니고 있음을 보여준다.

그림 7은 DIV2K 검증 데이터셋에서 얻은 결과를 보여주며, 이 데이터셋의 25번째 영상은 복원이 어려운 복잡한 디테일을 포함하고 있다. 이 영상은 작은 글자나 미세한 패턴이 많아 대부분의 모델들이 이러한 복잡한 세부 사항을 복

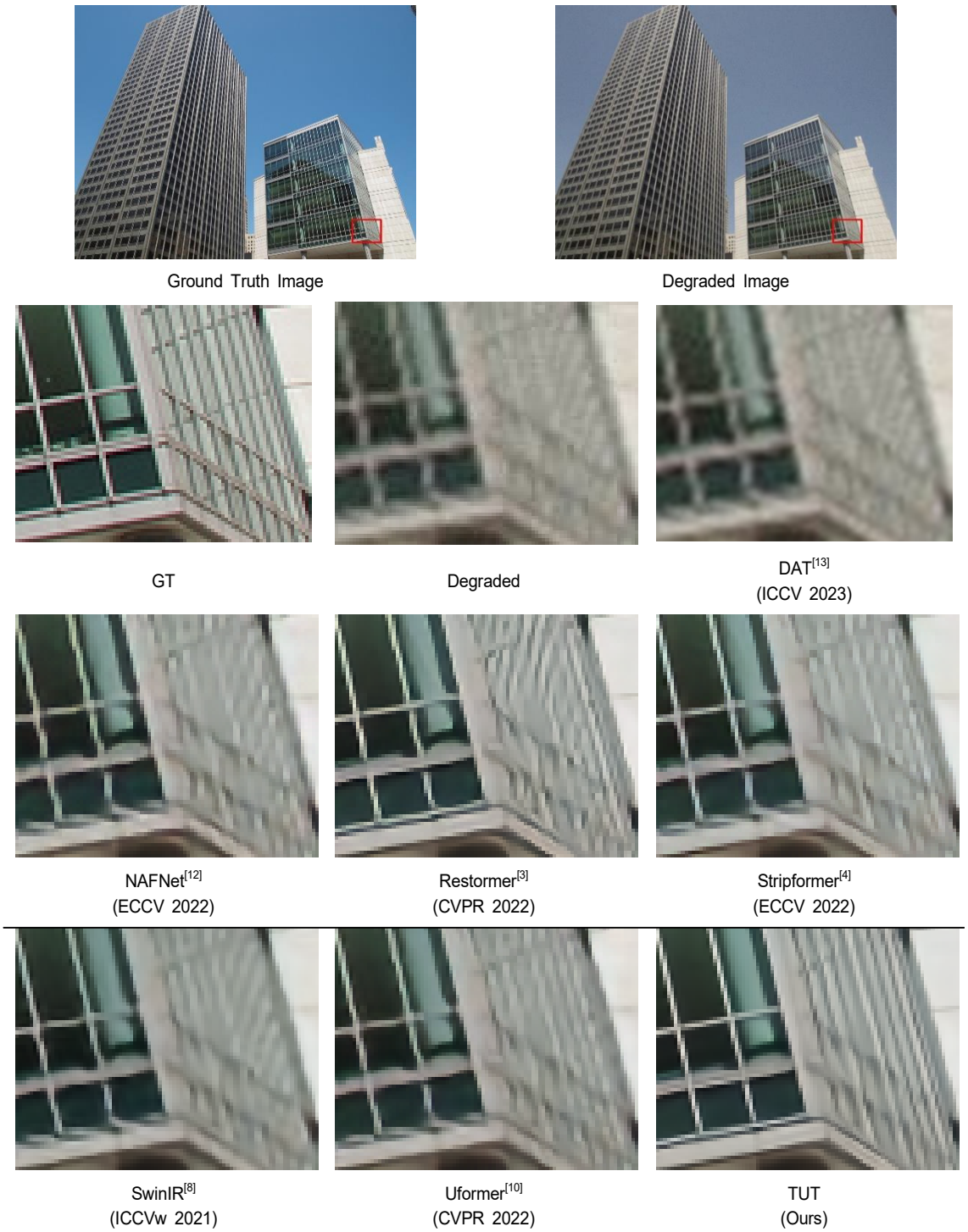

그림 6. Urban100[35] 데이터셋의 96번째 영상에 대한 정성적 결과 (확대를 통해 비교 가능)
Fig. 6. Qualitative results on the 96th image from the Urban100[35] set (Best viewed in zoom)

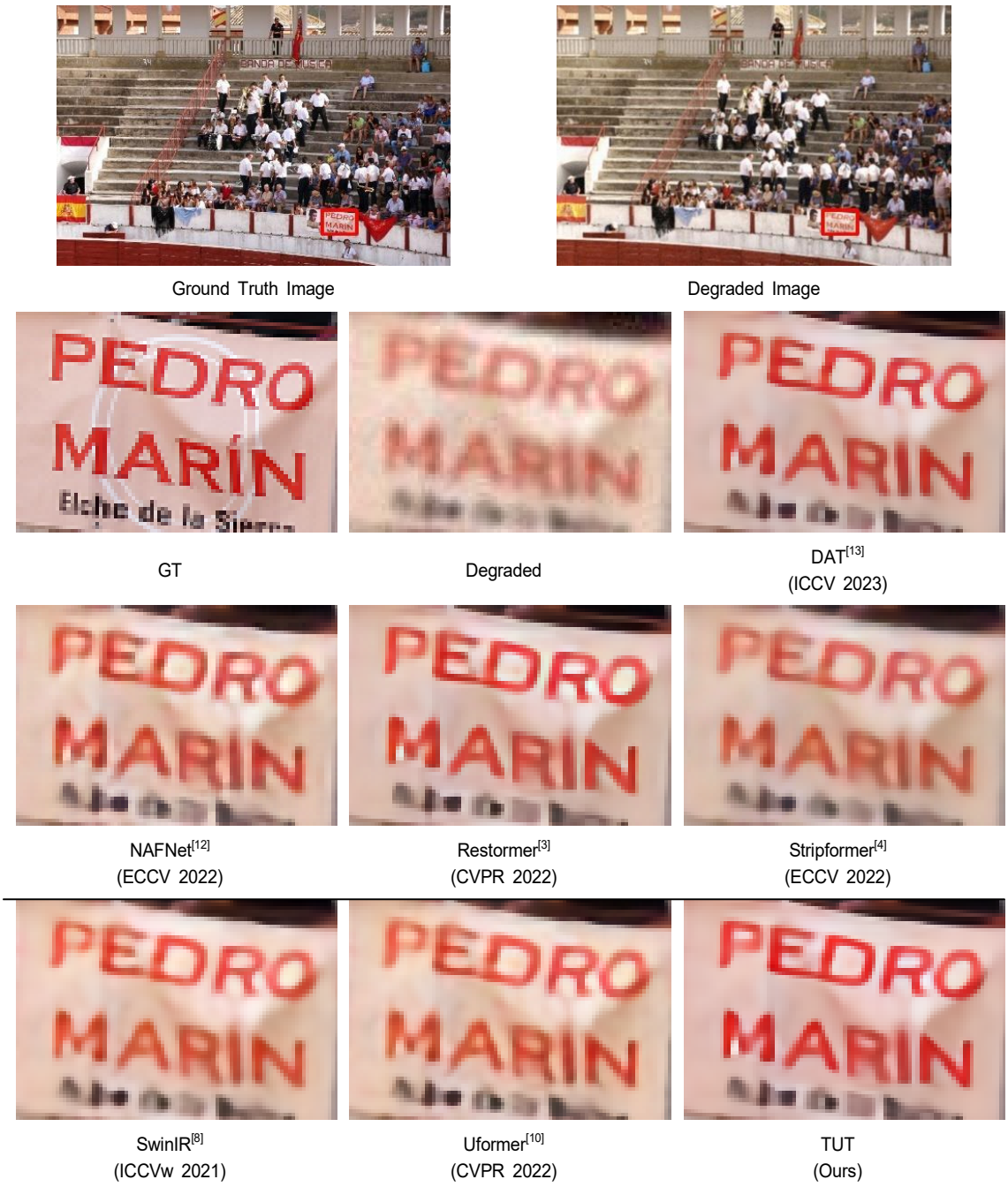

그림 7. DIV2K[32] 검증 데이터셋의 25번째 영상에 대한 정성적 결과 (확대를 통해 비교 가능)
Fig. 7. Qualitative results on the 25th image from the DIV2K[32] val. set (Best viewed in zoom)

원하는 데 어려움을 겪고 있으며 복원된 텍스트의 색상이 달라지는 경향이 있었다. 그러나 TUT는 빨간색 글자를 순수한 빨간색으로 성공적으로 복원하며, 이는 TUT가 색상 복원을 매우 효과적으로 수행할 수 있음을 나타낸다. 반면,

다른 모델들은 텍스트를 연한 빨간색이나 주황색으로 복원하는 경향을 보였으며, 이는 이들 모델이 색상 열화와 좁아진 색 영역에 대한 복원 성능이 부족함을 보여준다. 이러한 결과는 TUT의 Color-scale U-Net(CUN)이 색상 열화가 혼합되어 존재하는 영상에서도 효과적임을 보인다.

4. 구성 요소 소거 실험(Ablation Study)

TUT의 각 모듈의 효과를 입증하기 위해, 각 구성 요소를 모두 제거한 상태에서 제안한 모듈을 하나씩 추가하여 모델을 학습한 후 DIV2K[32] 검증 데이터셋에서의 PSNR을 측정하는 실험을 진행하였다. 먼저, CUN만을 사용한 단일 단계 학습을 하였다. 이후, L1 손실 함수와 함께 Edge 손실[4,29] 및 FFT 손실[11] 함수를 추가하여 성능을 평가하였다. 다음으로, 모든 단계가 STBG로 이루어진 GUN을 추가하여 두 단계 학습의 효과를 평가하였고 GUN 인코더에서 STBG를 RBG로 대체하여 RBG의 효율성도 입증하였다. 마지막으로, 각 트랜스포머에 공간 및 채널 조절자를 추가해 모든 구성 요소가 있을 때의 성능을 평가하였다.

표 2의 결과에 따르면, TUT는 제안하는 모든 구성 요소가 존재할 때(CUN + Edge/FFT loss + GUN w/ RBG & STBG + SM/CM), PSNR 25.38dB의 최고 성능을 달성하였다. 같은 손실 함수(Edge/FFT loss 추가) 조건에서 비교하였을 때는, 두 단계의 U-Net 구조(CUN + Edge/FFT loss + GUN w/ only STBG: 25.21dB)와 CUN만으로 구성된 단일 U-Net 구조(CUN + Edge/FFT loss: 25.14dB)를 사용할 때의 성능 차이는 0.07dB였다. 또한, 조절자를 사용하지 않았을 때(CUN + Edge/FFT loss + GUN w/ RBG & STBG: 25.22dB)와 SM 및 CM을 모두 사용했을 때(CUN + Edge/FFT loss + GUN w/ RBG & STBG + SM/CM: 25.38dB)의 성능 차이는 0.16dB로, 조절자가 포함된 셀프 어텐션 메커니즘의 효과가 두드러졌다. 더욱이, GUN 인코더에서 RBG를 사용했을 때(CUN + Edge/FFT loss + GUN w/ RBG & STBG: 25.22dB)는 GUN의 인코더에 STBG를 사용했을 때(CUN + Edge/FFT loss + GUN w/ only STBG: 25.21dB) 보다 성능이 0.01dB 소폭 향상되었음을 보여주었다. 이는 계산 비용과 성능 측면에서 U-Net의 인코더 단계에서 STBG 대신에 상대적으로 연산 복잡도가 낮은 RBG를 사용하는 것이 복잡한 공간적 열화를 처리하는 데 효율적인 접근 방식임을 입증한다. 또한, 동일 조건의 모델 구성에서 Edge 및 FFT 손실 함수를 L1 손실 함수와 결합하여 사용했을 때(CUN + Edge/FFT loss: 25.14dB)는 L1 손실 함수만 사용했을 때(CUN in Single-stage: 25.09dB)보다 성능이 0.05dB 향상되어, 추가 손실 함수들이 혼합 열화의 세부 정보를 복원에 효과적임을 나타냈다.

그림 8에서는 각 구성 요소를 모두 제거한 후, 차례대로 하나씩 추가하였을 때, DIV2K[32] 검증 데이터셋의 12번째 영상에 대한 모델의 정성적인 결과를 보여준다. 그림 8의 원본 영상의 빨간 부분을 확대한 영상에서 볼 수 있듯이 건물의 각 기둥에 있는 무늬는 직선을 띄어야 한다. 하지만, 열화 영상에서는 기둥의 무늬가 심하게 휘어져 있는 것을 볼 수 있다. CUN만을 이용한 단일 모델로 이 영상을 복원하였을 경우, 기둥을 제외한 전체적인 선들은 복원이 가능했지만 기둥의 무늬는 여전히 흐릿하게 보인다. 하지만, 추가적인 Edge/FFT loss를 사용하고 GUN을 추가하여 2단계 모델로 설계하였을 때는 점차 큰 기둥의 직선 무늬가 반듯하게 복원됨을 볼 수 있다. 특히, 2단계 모델에 SM/CM을 추가하였을 때는 큰 기둥과 더불어 작은 기둥의 무늬 또한

표 2. DIV2K[32] 검증 데이터셋에 대한 각 구성요소의 필요성 실험 결과 (빨강: 최고 성능 모델)
Table 2. Ablation study on DIV2K[32] val. dataset (RED: Best performance)

CUN (Single-stage)	Edge/FFT loss (Use Additional Loss functions)	GUN w/ only STBG (Two-stage)	GUN w/ RBG & STBG (Two-stage)	SM/CM (Add Modulators)	PSNR↑(dB)
✓					25.09
✓	✓				25.14
✓	✓	✓			25.21
✓	✓		✓		25.22
✓	✓		✓	✓	25.38

그림 8. DIV2K[32] 검증 데이터셋의 12번째 영상에 대한 구성 요소 소거 실험의 정성적 결과 (확대를 통해 비교 가능)
Fig. 8. Ablation study of qualitative results on the 12th image from the DIV2K[32] val. set (Best viewed in zoom)

적절하게 복원하고 있음을 알 수 있다. 이를 통하여 정성적으로도 Edge/FFT loss, RBG를 포함한 2단계 U-Net 기반의 트랜스포머 구조 그리고 SM/CM가 효과적임을 보여준다.

V. 결 론

1. 한계점과 추후 연구

제안한 TUT는 혼합 열화 영상을 복원하기에는 우수한 복원 능력을 지녔지만 두 가지의 단점이 존재한다. 먼저, TUT는 2단계의 U-Net 구조로 구성되어 있기 때문에 계산 자원이 다소 많이 필요하다는 단점이 있다. 표 3에서는 각 모델의 Floating Point Operations(FLOPs)와 PSNR 성능을 비교하였다. FLOPs는 256×256 영상에서 측정되었으며, PSNR은 DIV2K[32] 검증 데이터셋에서 측정되었다. TUT의 경우 전처리와 후처리 과정에 해당하는 RGB2YCbCr 및 YCbCr2RGB 변환은 FLOPs 계산에서 제외하였다. 표 3에 의하면, TUT는 SwinIR[8] 및 DAT[13]와 같은 모델들보다 적은 FLOPs로 더 우수한 성능을 발휘한다. 특히, DAT는

TUT와 유사하게 공간적 셀프 어텐션 블록과 채널 셀프 어텐션 블록을 하나의 네트워크에서 순차적으로 반복하여 사용하는 구조를 가지고 있지만, 두 단계로 나누어진 U-Net인 TUT는 FLOPs가 3배 가까이 적으면서도 더 나은 성능을 달성하였다. 그러나 Stripformer[4]나 Restormer[3]와 같은 기본 모델들에 비해 TUT는 여전히 많은 계산 자원을 요구하고 있어, 향후에는 낮은 계산 자원을 지니는 TUT에 대한 연구가 추가로 필요하다.

표 3. 모델 별 연산량과 성능 분석 결과 (빨강: 최고 성능 모델, 파랑: 두번째로 최고 성능 모델)
Table 3. Analysis on Computational Costs and Performance (RED: Best Model, BLUE: Second-Best Model)

Metric / Model	FLOPs ↓ (G)	PSNR ↑ (dB)
DAT[13] (ICCV 2023)	1848	25.13
NAFNet[12] (ECCV 2022)	126	24.51
Restormer[3] (CVPR 2022)	280	25.09
Stripformer[4] (ECCV 2022)	296	24.79
SwinIR[8] (ICCVw 2021)	1492	24.29
Uformer[10] (CVPR 2022)	76	24.05
TUT (Ours)	746	25.38

또한, 본 논문에서는 합성 열화 데이터셋을 사용하였기 때문에, 실제 열화된 영상을 복원하는 것은 여전히 어려운 과제로 남아 있다. 이는 합성 학습 데이터셋과 실제 평가 데이터셋 사이의 도메인 차이 때문이다. 향후 연구에서는 [38, 39]에서와 같은 도메인 적응 기법(Domain Adaptation)을 탐구하여, TUT가 오래된 사진이나 저품질 방송 영상과 같은 실제 열화된 영상을 복원하는 능력을 향상시키고자 한다.

2. 최종 결론

본 논문은 혼합 열화와 같은 복잡한 시나리오에서 트랜스포머를 이용한 영상 복원 문제를 처음으로 다루었다. 구체적으로, 혼합 열화가 적용된 영상을 복원하기 위하여 새로운 트랜스포머[2] 기반의 모델인 2단계 U-Net 트랜스포머(Two-stage U-Net Transformer, TUT)를 제안하였다. TUT는 혼합 열화를 공간 열화와 색상 열화로 나누어 효과적으로 영상을 복원 가능하였다. 또한, 공간 조절자(SM) 및 채널 조절자(CM)가 추가된 셀프 어텐션 메커니즘을 도입하여 기존 Transformer 기반의 모델들[3,4]의 한계를 극복하였으며 다양한 합성 혼합 열화 평가 데이터셋에서 정량적 및 정성적으로 트랜스포머 기반 영상 복원 모델들을 능가하는 성능을 보였다.

참 고 문 헌 (References)

[1] Deng, Jia, et al. “ImageNet: A Large-Scale Hierarchical Image Database.” 2009 IEEE Conference on Computer Vision and Pattern Recognition, June 2009.
doi: https://doi.org/10.1109/cvpr.2009.5206848

[2] Vaswani, Ashish, et al. “Attention is all you need.” Advances in neural information processing systems 30 (2017).

[3] Zamir, Syed Waqas, et al. “Restormer: Efficient Transformer for High-Resolution Image Restoration.” 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2022.
doi: https://doi.org/10.1109/cvpr52688.2022.00564

[4] Tsai, Fu-Jen, et al. “Stripformer: Strip Transformer for Fast Image Deblurring.” Computer Vision – ECCV 2022, 2022, pp. 146 – 62.
doi: https://doi.org/10.1007/978-3-031-19800-7_9

[5] Ronneberger, Olaf, et al. “U-Net: Convolutional Networks for Biomedical Image Segmentation.” Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, 2015, pp. 234 – 41.
doi: https://doi.org/10.1007/978-3-319-24574-4_28

[6] Dosovitskiy, Alexey, et al. “An image is worth 16x16 words: Transformers for image recognition at scale.” arXiv preprint arXiv:2010.11929 (2020).

[7] Krizhevsky, Alex, et al. “ImageNet Classification with Deep Convolutional Neural Networks.” Communications of the ACM, vol. 60, no. 6, May 2017, pp. 84 – 90.
doi: https://doi.org/10.1145/3065386

[8] Liang, Jingyun, et al. “SwinIR: Image Restoration Using Swin Transformer.” 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Oct. 2021, pp. 1833 – 44.
doi: https://doi.org/10.1109/iccvw54120.2021.00210

[9] Liu, Ze, et al. “Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows.” 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Oct. 2021, pp. 9992 – 10002.
doi: https://doi.org/10.1109/iccv48922.2021.00986

[10] Wang, Zhendong, et al. “Uformer: A General U-Shaped Transformer for Image Restoration.” 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2022, pp. 17662 – 72.

doi: https://doi.org/10.1109/cvpr52688.2022.01716
[11] Kong, Lingshun, et al. “Efficient Frequency Domain-Based Transformers for High-Quality Image Deblurring.” 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2023.
doi: https://doi.org/10.1109/cvpr52729.2023.00570
[12] Chen, Liangyu, et al. “Simple Baselines for Image Restoration.” Computer Vision – ECCV 2022, 2022, pp. 17–33.
doi: https://doi.org/10.1007/978-3-031-20071-7_2
[13] Chen, Zheng, et al. “Dual Aggregation Transformer for Image Super-Resolution.” 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Oct. 2023.
doi: https://doi.org/10.1109/iccv51070.2023.01131
[14] Liu, Xing, et al. “Restoring images with unknown degradation factors by recurrent use of a multi-branch network.” arXiv preprint arXiv:1907.04508 (2019).
[15] Kim, Sijin, et al. “Restoring Spatially-Heterogeneous Distortions Using Mixture of Experts Network.” Computer Vision – ACCV 2020, 2021, pp. 185–201.
doi: https://doi.org/10.1007/978-3-030-69532-3_12
[16] Shazeer, Noam, et al. “Outrageously large neural networks: The sparsely-gated mixture-of-experts layer.” arXiv preprint arXiv:1701.06538 (2017).
[17] Shin, Wooksu, et al. “Exploiting Distortion Information for Multi-Degraded Image Restoration.” 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), June 2022, pp. 536–45.
doi: https://doi.org/10.1109/cvprw56347.2022.00069
[18] Wang, Xintao, et al. “Recovering Realistic Texture in Image Super-Resolution by Deep Spatial Feature Transform.” 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 2018, pp. 606–15.
doi: https://doi.org/10.1109/cvpr.2018.00070
[19] Shi, Wenzhe, et al. “Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network.” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016, pp. 1874–83.
doi: https://doi.org/10.1109/cvpr.2016.207
[20] He, Kaiming, et al. “Deep Residual Learning for Image Recognition.” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016.
doi: https://doi.org/10.1109/cvpr.2016.90
[21] Fairchild, Mark D. Color Appearance Models. June 2013.
doi: https://doi.org/10.1002/9781118653128
[22] Park, Jongchan, et al. “Bam: Bottleneck attention module.” arXiv preprint arXiv:1807.06514 (2018).
[23] Woo, Sanghyun, et al. “CBAM: Convolutional Block Attention Module.” Computer Vision – ECCV 2018, 2018, pp. 3–19.
doi: https://doi.org/10.1007/978-3-030-01234-2_1
[24] Hendrycks, Dan, and Kevin Gimpel. “Gaussian error linear units (gelus).” arXiv preprint arXiv:1606.08415 (2016).
[25] Howard, Andrew G., et al. “Mobilenets: Efficient convolutional neural networks for mobile vision applications.” arXiv preprint arXiv:1704.04861 (2017).
[26] Ba, Jimmy Lei, Jamie Ryan Kiros, and Geoffrey E. Hinton. “Layer normalization.” arXiv preprint arXiv:1607.06450 (2016).
[27] Chu, Xiangxiang, et al. “Conditional positional encodings for vision transformers.” arXiv preprint arXiv:2102.10882 (2021).
[28] Lin, Min, Qiang Chen, and Shuicheng Yan. “Network in network.” arXiv preprint arXiv:1312.4400 (2013).
[29] Zamir, Syed Waqas, et al. “Multi-Stage Progressive Image Restoration.” 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2021.
doi: https://doi.org/10.1109/cvpr46437.2021.01458
[30] Wang, Xiaohong, et al. “Mixed Distortion Image Enhancement Method Based on Joint of Deep Residuals Learning and Reinforcement Learning.” Signal, Image and Video Processing, vol. 15, no. 5, Jan. 2021, pp. 995–1002.
doi: https://doi.org/10.1007/s11760-020-01824-y
[31] Buslaev, Alexander, et al. “Albumentations: Fast and Flexible Image Augmentations.” Information, vol. 11, no. 2, Feb. 2020, p. 125.
doi: https://doi.org/10.3390/info11020125
[32] Agustsson, Eirikur, and Radu Timofte. “NTIRE 2017 Challenge on Single Image Super-Resolution: Dataset and Study.” 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), July 2017, pp. 1122–31.
doi: https://doi.org/10.1109/cvprw.2017.150
[33] Bevilacqua, Marco, et al. “Low-Complexity Single-Image Super-Resolution Based on Nonnegative Neighbor Embedding.” Procedings of the British Machine Vision Conference 2012, 2012, pp. 135.1-135.10.
doi: https://doi.org/10.5244/c.26.135
[34] Zeyde, Roman, et al. “On Single Image Scale-Up Using Sparse-Representations.” Curves and Surfaces, 2012, pp. 711–30.
doi: https://doi.org/10.1007/978-3-642-27413-8_47
[35] Huang, Jia-Bin, et al. “Single Image Super-Resolution from Transformed Self-Exemplars.” 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015, pp. 5197–206.
doi: https://doi.org/10.1109/cvpr.2015.7299156
[36] Martin, D., et al. “A Database of Human Segmented Natural Images and Its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics.” Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001, vol. 2, pp. 416–23.
doi: https://doi.org/10.1109/iccv.2001.937655
[37] Kingma, Diederik P., and Jimmy Ba. “Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980 (2014).
[38] Yang, Yanchao, and Stefano Soatto. “FDA: Fourier Domain Adaptation for Semantic Segmentation.” 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020, pp. 4084–94.
doi: https://doi.org/10.1109/cvpr42600.2020.00414
[39] Tzeng, Eric, et al. “Adversarial Discriminative Domain Adaptation.” 2017 IEEE Conference on Computer Vision and Pattern Recognition

(CVPR), July 2017, pp. 2962 – 71.
doi: https://doi.org/10.1109/cvpr.2017.316

[40] Taehwan Kim, Image Restoration of Mixed Degradation using Transformer-based Two-stage U-Nets, Master's Thesis of Korea Advanced Institute of Science and Technology, Daejeon, Korea, 2024, https://library.kaist.ac.kr/search/ctlgSearch/posesn/view.do?bibctrlno=1097280&se=t0&ty=B&_csrf=32e6c1ee-5c0f-4849-beef-0d8ab8b4e497

[41] Taehwan Kim, Munchurl Kim. "TUT: Two-stage U-Net-based Transformer for Image Restoration with Mixed Degradation." Proceedings of the Korean Institute of Broadcast and Media Engineers Summer Conference. 2024.

저 자 소 개

김 태 환

- 2021년 8월 : 광운대학교 소프트웨어학부 학사
- 2024년 2월 : 한국과학기술원 전기및전자공학부 석사
- 2024년 2월 ~ 현재 : 한국과학기술원 전기및전자공학부 박사과정
- ORCID : https://orcid.org/0009-0000-2543-2852
- 주관심분야 : 컴퓨터 비전, 딥러닝, 영상 복원, 초해상화, 생성 모델

김 문 철

- 1996년 8월 : University of Florida Electrical and Computer Engineering 박사
- 2001년 2월 ~ 2009년 3월 : 한국정보통신대학교 공학부 조교수/부교수
- 2009년 3월 ~ 현재 : 한국과학기술원 전기및전자공학부 부교수/정교수
- ORCID : https://orcid.org/0000-0003-0146-5419
- 주관심분야 : 컴퓨터 비전, 딥러닝, 영상 처리, 비디오 코딩