다중 스케일 고주파 오류 맵을 활용한 확산 모델 기반 이미지 초해상화*

서준혁O, 이동규
경북대학교 인공지능학과
ssam2s@knu.ac.kr, dglee@knu.ac.kr

Diffusion Model Based Image Super-Resolution with Multi-Scale High Frequency Error Maps

Jun-Hyeok SeoO, Dong-Gyu Lee
Department of Artificial Intelligence, Kyungpook National University
ssam2s@knu.ac.kr, dglee@knu.ac.kr

요 약

본 논문은 다중 스케일 고주파 오류 맵을 활용한 확산 모델 기반 이미지 초해상화 프레임워크에 대한 연구이다. 기존의 확산 모델 기반 이미지 초해상화 기법들은 저해상도 이미지의 복원 과정에서 이미지에 존재하는 세부 정보를 보존하는데 어려움이 있었다. 본 논문에서는 모델이 복원한 이미지와 실제 고해상도 이미지를 다양한 스케일로 변환하고 두 이미지 사이의 오류를 고주파 영역에서 분석하여 모델이 이미지의 세부 정보를 보다 정확하게 복원할 수 있도록 한다. 제안하는 프레임워크는 벤치마크 데이터셋을 이용한 기존 기법과의 성능 비교에서 우수한 성능을 보임으로써 제안한 프레임워크가 이미지 초해상화에 효과적임을 보여준다.

1. 서 론

이미지 초해상화(Super-Resolution, SR)는 저해상도 이미지를 고해상도 이미지로 변환하는 기술로 이미지 복원 등 다양한 목적[1]을 위해 활용되고 있다. 초기의 연구들은 합성곱 신경망(Convolutional Neural Networks, CNN)을 이용하여 주로 제안되었으며, 이미지 생성 분야에서 적대적 생성 신경망(Generative Adversarial Networks, GAN)[2]이 등장한 이후 이를 이용한 연구들이 많이 제안되었다. GAN은 생성자와 판별자 구조로 이루어져 학습이 진행될수록 판별자가 실제 데이터인지 생성된 데이터인지 구분할 수 없도록 생성자가 더 좋은 결과를 생성하는 방식으로, 이미지 생성에 있어 큰 성과를 이루었다. 뒤이어 자연어 처리 분야에서 사용되던 트랜스포머(Transformer) 모델을 비전 분야에 적용한 ViT(Vision Transformer)[3] 기반의 연구들도 좋은 성능을 보였다. ViT는 이미지를 패치로 나누고 각 패치를 단어처럼 처리함으로써 이미지의 전반적인 맥락(Global context)을 이해하고 자연스러운 이미지를 생성하였다.

* 본 연구는 과학기술정보통신부의 재원으로 한국 연구재단 (No. 2021R1C1C1012590), (No. 2022R1A4A1023248)과 과학기술정보통신부 및 정보통신기획평가원의 대학ICT 연구센터지원사업의 연구결과로 수행되었음 (IITP-2024-2020-0-01808), (IITP-2024-RS-2024-00437718)

최근 이미지 생성 분야에서 확산 모델을 이용한 연구들이 주목할만한 성과[4]를 거두면서 확산 모델이 이미지 초해상화 분야에서도 활용되고 있다. 확산 모델은 이미지에 노이즈를 반복적으로 추가하고 반대로 추가한 노이즈를 제거하는 과정을 반복 학습하면서 이미지를 효과적으로 생성할 수 있게 된다. 이 과정은 이미지의 세부 정보를 더욱 정밀하게 생성할 수 있도록 해 트랜스포머 기반의 연구들을 능가하는 결과[5]를 보여주었다. 하지만 학습 및 추론 과정에서 많은 연산량이 요구된다는 단점이 있어 컴퓨팅 리소스가 제한된 상황에서는 효과적으로 이미지 초해상화를 수행할 수 없다는 문제가 있다. 이러한 문제를 해결하기 위해 잠재 확산 모델(Latent Diffusion Model, LDM)[6]이 도입되었다. LDM은 확산 과정을 이미지 레벨이 아닌 잠재 공간(Latent space)에서 수행하여 기존 모델들의 성능은 뛰어넘으면서 연산량은 획기적으로 감소시켰다. 하지만 LDM의 경우 그림 1과 같이 이미지의 세부 정보와 질감과 같은 고주파 성분들이 왜곡되어 표현되는 문제점이 존재한다.

본 논문에서는 기존 이미지 초해상화 분야에서 사용되던 LDM을 그대로 사용함으로써 연산의 효율성은 유지하되, 초해상화 과정에서 발생하는 왜곡을 최소화할 수 있는 프레임워크를 제안한다. 제안하는 프레임워크는 LDM을 그대로 채택하여 효율적인 연산이 가능할 뿐만 아니라, 이미지의 세부 정보와 질감과 같은 고주파 성분에 대해 다양한 스케일에서 상대적인 가중치를 부여하여

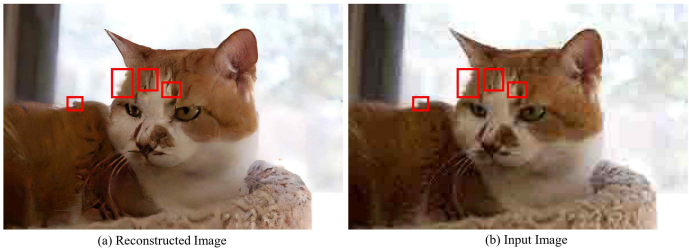

그림 1. 왜곡 문제에 대한 예시 이미지

모델이 이를 중점적으로 학습하도록 한다. 이미지에 존재하는 고주파 성분 분석에서는 이산 웨이블릿 변환(Discrete Wavelet Transform, DWT)을 사용하며, 다양한 데이터셋에 대한 실험을 통해 제안하는 프레임워크가 효과적임을 증명한다.

2. 본 론

본 논문에서는 효과적인 초해상화를 위한 새로운 프레임워크를 제안한다. 제안하는 프레임워크는 LDM의 학습 과정에서 이미지의 세부 정보와 질감을 효과적으로 보존할 수 있도록 설계되었다. 이 프레임워크에서는 변환하고자 하는 저해상도 이미지(Low Resolution, LR)와 실제 고해상도 이미지(High Resolution, HR)를 입력으로 받아, 초해상화 결과 이미지(SR)와 HR 이미지 사이의 손실을 계산하며 학습이 진행된다. 이 과정에서 픽셀 사이의 오차만 고려할 뿐 아니라, SR 이미지와 HR 이미지를 다양한 크기로 증강하고 각 스케일별 영상에 대해 주파수 대역별 성분을 DWT를 이용해 추출한다. 이 중 이미지의 세부 정보와 질감을 대부분 포함하고 있는 고주파 성분만을 비교하여 두 이미지 사이의 오류 맵을 생성한다. 생성된 오류 맵은 앞에서 생성한 각 스케일별 SR 이미지와 HR 이미지에 곱해져 오류 영역을 강조시키며, 두 종류의 이미지들은 사람의 인지적인 측면에서 느껴지는 왜곡을 최소화하기 위해 특징 벡터 간의 손실을 계산하는 LPIPS(Learned Perceptual Image Patch Similarity) [7] 손실 함수 계산에 사용된다.

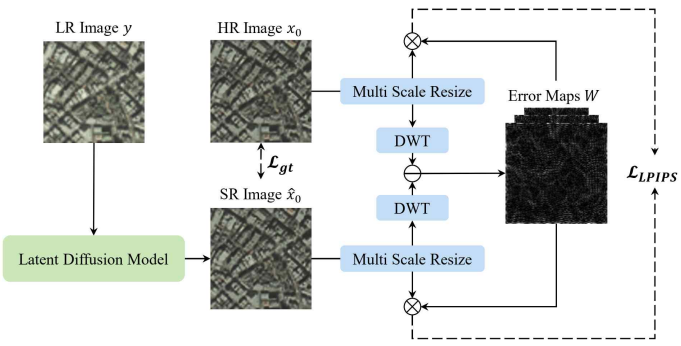

그림 2. 제안하는 프레임워크의 구조

또한 각 스케일에서 계산된 LPIPS 손실은 높은 스케일에서 계산된 값일수록 더 많은 세부 정보를 고려한 것이기에 각 스케일마다 가중치를 비례하게 곱하여 모두 더하는 방식으로 최종 손실 함수를 계산하였다. 스케일은 1, 0.5, 0.25, 0.125로 하였으며 이에 대응하는 가중치 W_i는 0.4, 0.3, 0.2, 0.1로 설정하였다. 최종 손실 함수 L_{total}의 계산 과정은 아래 수식과 같다.

$$L_{total} = \sum_{i=1}^{4} W_i \cdot L_{LPIPS}(S_i). \quad (1)$$

3. 실험 결과

본 논문에서 제안한 프레임워크의 객관적인 평가를 위해 LDM 기반의 기존 연구들과의 비교 실험을 진행한다. 유사한 기존 연구 SinSR[8]을 베이스라인으로 설정해 실험을 진행하였으며 학습 데이터셋과 실험 데이터셋 또한 동일하게 설정하였다. 성능 지표는 CLIP을 활용한 이미지 품질 측정 방식인 CLIPIQA[9]와 Multi-Scale Transformer 모델을 활용해 이미지 품질을 측정하는 방식인 MUSIQ[10]를 사용하였다.

Methods	ImageNet		RealSR		RealSet65	
	CLIPIQA	MUSIQ	CLIPIQA	MUSIQ	CLIPIQA	MUSIQ
LDM[6]	0.572	50.895	0.383	49.317	0.427	47.488
ResShift[11]	0.603	53.897	0.595	59.873	0.653	61.330
SinSR(baseline)	0.611	53.357	0.688	**61.582**	0.715	62.169
Ours	**0.648**	**54.212**	**0.695**	61.465	**0.718**	**63.216**

표 1. 벤치마크 데이터셋에 대한 성능 결과 비교

실험 결과 RealSR 데이터셋의 MUSIQ 측정 결과에서 베이스라인에 대해 근소하게 낮은 성능을 보인 것을 제외하고는 본 논문에서 제안한 방법이 기존 방법들보다 우수한 성능을 보이는 것을 확인할 수 있었다.

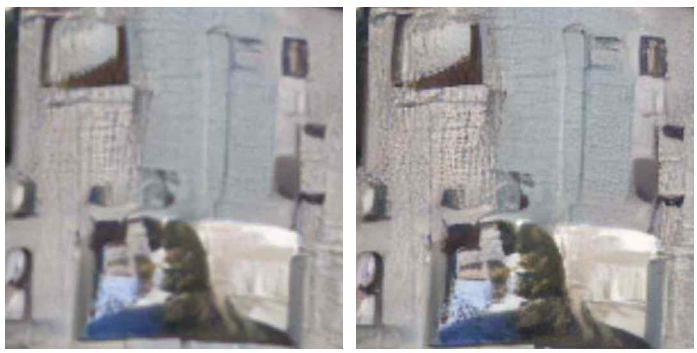

그림 3. 베이스라인(좌)과 제안 방법(우)의 결과 비교

실제 결과 영상에 대한 시각적 비교를 수행한 결과에서도, 제안된 방법이 베이스라인 대비 영상 내 세부 정보와 질감을 더욱 선명하게 표현한다는 점을 확인할 수 있었다. 이를 통해 평가 지표에 더불어 시각적으로도 제안된 방법이 고해상도 복원 및 세부 구조 보존 측면에서 우수한 성능을 보이는 것을 알 수 있었다.

4. 결 론

본 논문에서는 기존의 이미지 초해상화 기법들에 존재하던 왜곡 문제를 효과적으로 해결하기 위해 새로운 프레임워크를 제안한다. 제안하는 프레임워크는 추론 이미지와 정답 이미지 간의 고주파 성분을 다중 스케일로 추출하고 비교하여 오류 맵을 구성함으로써 이미지의 고주파 성분에 존재하는 세부 정보와 질감을 최대한 보존하며 왜곡 문제를 최소화한다. 세 가지 벤치마크 데이터셋에 대한 실험은 제안하는 프레임워크가 최신 연구들보다 높은 성능을 보이며 이미지 초해상화에 효과적임을 증명한다.

참 고 문 헌

[1] Wang, Zhihao, Jian Chen, and Steven CH Hoi. "Deep learning for image super-resolution: A survey." IEEE transactions on pattern analysis and machine intelligence 43.10 (2020): 3365-3387.
[2] Goodfellow, Ian, et al. "Generative adversarial nets." Advances in neural information processing systems 27 (2014).
[3] Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).
[4] Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models." Advances in neural information processing systems 33 (2020): 6840-6851.
[5] Saharia, Chitwan, et al. "Image super-resolution via iterative refinement." IEEE transactions on pattern analysis and machine intelligence 45.4 (2022): 4713-4726.
[6] Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.
[7] Zhang, Richard, et al. "The unreasonable effectiveness of deep features as a perceptual metric." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.
[8] Wang, Yufei, et al. "SinSR: diffusion-based image super-resolution in a single step." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.
[9] Wang, Jianyi, Kelvin CK Chan, and Chen Change Loy. "Exploring clip for assessing the look and feel of images." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 37. No. 2. 2023.
[10] Ke, Junjie, et al. "Musiq: Multi-scale image quality transformer." Proceedings of the IEEE/CVF international conference on computer vision. 2021.
[11] Yue, Zongsheng, Jianyi Wang, and Chen Change Loy. "Resshift: Efficient diffusion model for image super-resolution by residual shifting." Advances in Neural Information Processing Systems 36 (2024).