일반논문 (Regular Paper)
방송공학회논문지 제31권 제4호, 2026년 7월 (JBE Vol.31, No.4, July 2026)
https://doi.org/10.5909/JBE.2026.31.4.793
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

의미 및 구조 경로 조건화를 이용한 확산 기반 초해상화

김 태 환[a)], 이 진 영[a)‡]

Diffusion-Based Super-Resolution Using Semantic and Structural Path Conditioning

Taehwan Kim[a)] and Jin Young Lee[a)‡]

요 약

단일 영상 초해상화는 저해상도 입력 영상에서 손실된 세부 질감과 구조 정보를 복원하는 문제이며, 확산 모델은 반복적인 노이즈 제거 과정을 통해 선명한 고주파 성분을 생성할 수 있다. 그러나 저해상도 조건이 복원 과정에 충분히 반영되지 않을 경우, 경계 위치, 반복 패턴, 객체 윤곽이 원본과 어긋나는 구조적 왜곡이 발생할 수 있다. 본 논문에서는 저해상도 조건을 전역 문맥을 담당하는 의미 경로와 공간 구조를 담당하는 구조 경로로 나누어 조건화하는 확산 기반 초해상화 모델을 제안한다. 의미 경로는 저해상도 영상의 장면 문맥과 질감 분포를 토큰 형태로 요약하여 병목 구간의 교차 어텐션에 사용하며, 구조 경로는 공간적으로 정렬된 특징 맵을 생성하여 복원 단계의 특징 맵 변조에 활용한다. 이를 통해 확산 모델의 생성 능력을 유지하면서도 저해상도 입력에 포함된 구조적 단서를 안정적으로 반영하도록 한다. 실험 결과, Set5, Set14, BSD100 및 Urban100 데이터셋을 이용한 4배 초해상화 실험에서 제안 모델은 비교 모델보다 구조적 안정성과 지각적 화질 간의 균형을 보였다.

Abstract

Diffusion models can generate sharp high-frequency details for single-image super-resolution through iterative denoising. However, when low-resolution conditions are not sufficiently reflected during restoration, the generated images may contain structural distortions such as misaligned boundaries, inconsistent repetitive patterns, or altered object contours. This paper proposes a diffusion-based super-resolution model using semantic and structural path conditioning, in which low-resolution information is separated into semantic and structural paths. The semantic path summarizes scene-level context and texture distributions into tokens and injects them through cross-attention at the bottleneck stage. The structural path preserves spatially aligned features and applies them as feature-map modulation during reconstruction. This design enables the model to retain the generative capability of diffusion models while more reliably preserving structural cues from the low-resolution input. Experiments on Set5, Set14, BSD100 and Urban100 for 4× super-resolution demonstrate that the proposed model achieves a favorable balance between structural stability and perceptual quality compared to baseline models.

Keyword : Diffusion, Super-Resolution

Ⅰ. 서 론

단일 영상 초해상화는 하나의 저해상도 영상으로부터 고해상도 영상을 복원하는 문제이다[1]. 이 과정에서는 저해상도 영상에서 손실된 세부 질감과 경계 정보를 추정해야 하므로, 복원 결과의 선명도뿐 아니라 원본 구조와의 정합성도 함께 고려되어야 한다. 초해상화는 의료 영상, 위성 및 항공 영상, 감시 영상, 모바일 영상 처리와 같이 제한된 해상도에서 더 많은 시각 정보를 요구하는 분야에서 중요하게 활용된다[2,3].

기존 회귀 기반 초해상화 모델은 픽셀 단위 오차를 최소화하는 방식으로 높은 PSNR과 SSIM을 달성해 왔으나, 복원 가능한 다양한 고해상도 후보들의 평균적인 결과를 생성하는 경향으로 인해 세부 질감이 손실되고 복원 결과가 흐릿해질 수 있다[4]. 반면 확산 기반 초해상화 모델은 반복적인 노이즈 제거 과정을 통해 선명한 고주파 질감을 생성할 수 있지만, 저해상도 조건 정보가 충분히 반영되지 않을 경우 경계 위치, 반복 패턴, 객체 윤곽 등이 원본 구조와 어긋나는 왜곡이 발생할 수 있다[5].

본 논문은 이러한 문제를 완화하기 위해 저해상도 조건 정보를 의미 경로와 구조 경로로 나누어 복원 과정에 반영하는 확산 기반 초해상화 모델을 제안한다. 의미 경로는 영상 전체의 문맥과 질감 분포를 복원 과정에 제공하고, 구조 경로는 경계와 윤곽 등 공간적으로 정렬된 정보를 전달한다. 두 조건은 하나의 위치에서 단순히 결합되지 않고, 노이즈 제거 백본의 서로 다른 단계에 주입되어 전역 문맥과 공간 구조가 각각의 역할에 맞게 반영되도록 한다.

제안 구조는 확산 모델의 생성 능력을 활용하면서도 저해상도 입력에 포함된 구조적 단서가 복원 과정에서 약화되는 문제를 완화하는 것을 목표로 한다. 이를 검증하기 위해 본 논문에서는 Set5, Set14, BSD100, Urban100 데이터셋을 이용한 4배 초해상화 실험을 수행하고, PSNR, SSIM, LPIPS와 같은 정량적 성능 평가와 MOS를 이용한 주관적 화질 평가 결과도 함께 분석한다.

a) 세종대학교 AI로봇학과(Sejong University)
‡ Corresponding Author: 이진영(Jin Young Lee)
E-mail: jinyounglee@sejong.ac.kr
Tel: +82-2-6935-2672
ORCID: https://orcid.org/0000-0002-4384-2633
※ 본 논문은 과학기술정보통신부의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (RS-2022-00156345, RS-2026-25544647)
· Manuscript received July 3, 2026; Revised July 6, 2026; Accepted July 6, 2026.

Ⅱ. 관련 연구

1. 단일 영상 초해상화

단일 영상 초해상화 연구는 저해상도 영상과 고해상도 영상 사이의 매핑을 학습하는 합성곱 신경망 기반 방법에서 출발하였다[1]. 이후 깊은 네트워크의 학습 안정성을 높이기 위해 잔차 연결을 활용한 모델이 제안되었으며[6], 밀집 연결과 계층적 특징 융합을 통해 여러 계층의 정보를 효과적으로 재사용하는 방법도 연구되었다[7]. 또한 채널 어텐션 구조를 활용하여 중요한 특징 채널을 강조하는 방식이 초해상화 성능 개선에 사용되었다[8]. 최근에는 윈도우 기반 셀프 어텐션과 계층적 특징 표현을 활용하는 트랜스포머 기반 모델들이 등장하여 넓은 문맥 정보를 효과적으로 반영하고 있다[9].

이러한 회귀 기반 초해상화 모델들은 PSNR과 SSIM과 같은 왜곡 기반 지표에서 높은 성능을 달성할 수 있으나, 평균 제곱 오차나 절대 오차와 같은 픽셀 중심 손실을 주로 사용하기 때문에 세부 질감이 부드럽게 평균화되는 경향을 보인다. 특히 창문, 난간, 외벽, 격자 구조와 같이 반복 패턴이 많은 영상에서는 세부 구조가 흐려질 경우 사람이 인식하는 화질 저하가 크게 나타날 수 있다.

지각적 화질을 개선하기 위해 적대적 생성 네트워크, 지각 손실, 질감 손실 등을 활용하는 연구도 수행되었다[4]. 이러한 방법은 회귀 기반 모델보다 시각적으로 선명한 결과를 생성할 수 있지만, 생성된 세부 질감이 실제 정답 구조와 일치하지 않거나 특정 영역에서 불안정한 패턴을 생성할 수 있다. 따라서 초해상화에서는 지각적 화질 향상뿐 아니라, 저해상도 입력에 남아 있는 구조 정보를 보존하면서 자연스러운 세부 표현을 복원하는 것이 중요하다.

2. 확산 모델 기반 초해상화

확산 모델은 데이터에 점진적으로 노이즈를 추가하는 순방향 과정과 노이즈가 포함된 샘플에서 원본에 가까운 데이터를 복원하는 역방향 과정을 학습하는 생성 모델이다[10]. 초해상화에서는 저해상도 영상을 조건으로 사용하여 노이즈 상태의 고해상도 영상을 단계적으로 복원한다. 이러한 방식은 복원 과정에서 다양한 고주파 세부 정보를 생성할 수 있기 때문에, 기존 회귀 기반 모델보다 자연스럽고 선명한 결과를 보이는 경우가 많다.

SR3는 확산 모델을 초해상화에 적용한 대표적인 연구 중 하나로, 저해상도 조건을 활용하여 반복적인 노이즈 제거 과정을 수행하였다[5]. 이후 확산 기반 초해상화 연구는 복원 화질 향상, 추론 효율 개선, 실제 열화 영상 대응이라는 방향으로 확장되어 왔다. ResShift는 저해상도 영상과 고해상도 영상 사이의 잔차 이동을 활용하여 보다 효율적인 확산 복원 과정을 구성하였으며[11], SinSR은 이를 단일 단계 추론으로 단순화하여 확산 기반 초해상화의 높은 추론 비용을 완화하고자 하였다[12]. 한편 StableSR과 DiffBIR은 사전 학습된 대규모 확산 모델의 생성 사전 지식을 활용하여 실제 열화 영상에서의 지각적 화질을 향상시키는 방향으로 연구되었다[13,14].

확산 기반 초해상화의 장점은 복잡한 질감과 고주파 성분을 자연스럽게 생성할 수 있다는 점이다. 그러나 이러한 생성 능력은 동시에 구조적 왜곡의 원인이 될 수 있다. 저해상도 입력에 존재하는 경계, 선의 방향, 반복 패턴, 물체 윤곽이 복원 과정에서 충분히 반영되지 않으면, 결과 영상은 선명해 보이더라도 원본 구조와 다른 패턴을 포함할 수 있다. 특히 건축물이나 도시 영상처럼 직선 구조와 반복 패턴이 중요한 영상에서는 이러한 왜곡이 주관적 화질 평가에서 불리하게 작용할 수 있다.

따라서 확산 기반 초해상화에서는 저해상도 조건을 복원 과정에 어떻게 반영하는지가 핵심 설계 요소가 된다. 저해상도 영상을 단순히 업샘플링하여 입력에 결합하는 방식은 위치 정보를 제공할 수 있으나, 복원 과정 전체에서 전역 문맥과 공간 구조를 구분하여 제어하기에는 한계가 있다. 반대로 전역 조건만 강하게 주입할 경우 전체적인 질감은 자연스러워질 수 있으나, 세부 경계나 반복 구조의 위치 정합성이 약해질 수 있다.

본 논문은 이러한 한계를 고려하여 저해상도 조건을 전

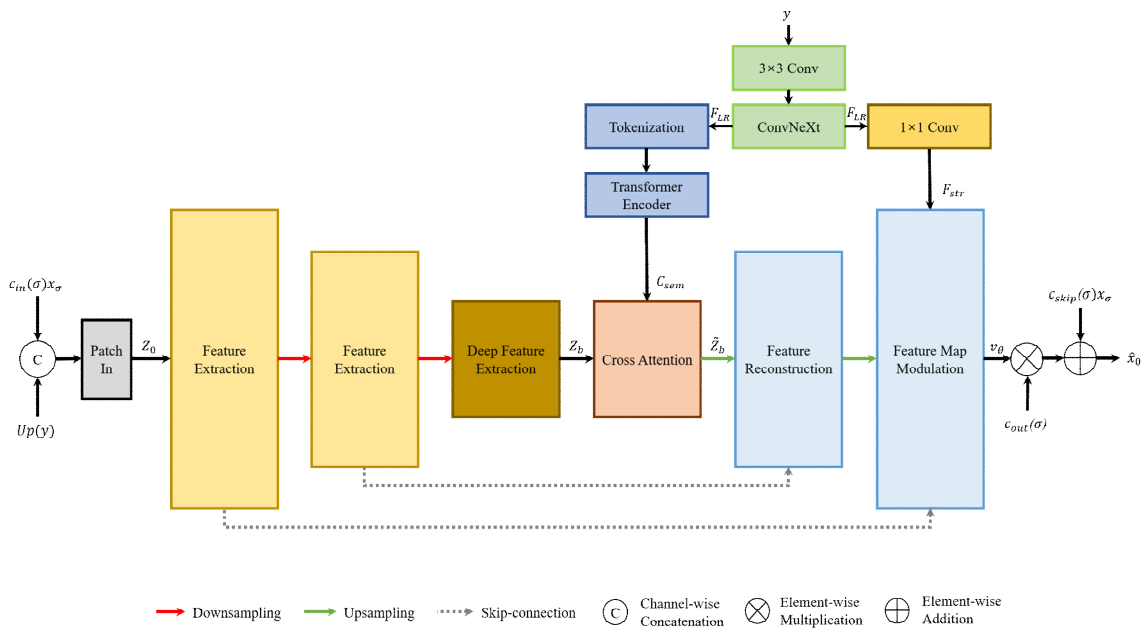

그림 1. 제안 모델의 전체 구조
Fig. 1. Overall architecture of the proposed model

역 문맥 정보와 공간 구조 정보로 나누고, 이를 복원 과정의 서로 다른 단계에 반영하는 조건화 방식을 제안한다. 이를 통해 확산 모델의 생성 능력을 유지하면서도, 저해상도 입력이 제공하는 구조적 단서를 복원 과정에 보다 직접적으로 반영하고자 한다.

Ⅲ. 제안 방법

본 논문은 사전 조건화 기반 확산 복원 프레임워크를 바탕으로 단일 영상 초해상화를 수행한다[15]. 제안 모델은 기존 조건부 확산 백본의 노이즈 제거 구조를 기반으로 하며, 저해상도 조건 정보를 단일 특징으로만 사용하지 않고 의미 경로와 구조 경로로 분리하여 복원 과정에 반영한다. 이를 통해 전역 문맥 정보와 공간 구조 정보가 각각 병목 구간과 복원 단계에서 역할에 맞게 활용되도록 한다. 그림 1은 제안 모델의 전체 구조를 나타낸다.

1. 제안 모델의 전체 구조

그림 1에서 초기 입력 특징 Z_0은 고해상도 정답 영상 x_0에 노이즈가 적용된 x_σ와 저해상도 입력 영상 y를 업샘플링한 $Up(y)$를 채널 방향으로 결합한 뒤, 패치 입력 계층을 통해 생성한다. 이는 식 (1)과 같이 나타낼 수 있다.

$$Z_0 = PatchIn([c_{in}(\sigma)x_\sigma \oplus_c Up(y)]) \tag{1}$$

식 (1)에서 $c_{in}(\sigma)$는 사전 조건화 입력 계수이고, $\oplus_c$는 채널 방향 결합을 의미하며, $PatchIn(\cdot)$은 결합된 입력을 패치화하여 입력 투영 계층을 거쳐 초기 입력 특징 Z_0로 변환하는 연산이다. 이를 통해 노이즈가 포함된 복원 상태와 저해상도 정보를 초기 단계부터 함께 고려할 수 있다.

그림 1의 Feature Extraction 모듈은 초기 입력 특징 Z_0로부터 얕은 특징을 추출하여 국소 패턴을 포착하고, Deep Feature Extraction 모듈은 계층적으로 추출된 특징으로부터 넓은 문맥 정보를 반영한 병목 특징 Z_b를 생성하며, 얕은 계층에서 추출된 특징들은 스킵 연결을 통해 복원 단계로 전달되어 경계와 세부 구조 보존을 보조한다. 교차 어텐션 (Cross Attention) 모듈은 백본의 병목 구간에서 1회 적용되며, 의미 경로 (Semantic Path) 조건 C_{sem}을 Key와 Value로, 병목 특징 Z_b를 Query로 사용하여 조건화된 병목 특징 $\tilde{Z}_b$를 식 (2)와 같이 생성한다.

$$\tilde{Z}_b = CrossAttn(Z_b, C_{sem}) \tag{2}$$

Feature Reconstruction 모듈에서는 $\tilde{Z}_b$를 기반으로 특징 해상도를 높이고, 대응되는 스킵 특징과 결합하여 고해상도 복원에 필요한 특징을 재구성한다. 재구성된 복원 특징은 Feature Map Modulation 모듈에서 구조 경로 (Structural Path) 특징 F_{str}에 의해 위치별로 보정되고[16], 출력 계층을 통해 모델의 예측값 v_θ로 변환된다. 최종적으로 복원 영상 추정값 $\hat{x}_0$은 식 (3)과 같이 모델 예측값 v_θ와 현재 노이즈 상태 x_σ를 사전 조건화 계수를 이용하여 결합함으로써 계산된다.

$$\hat{x}_0 = c_{skip}(\sigma)x_\sigma + c_{out}(\sigma)v_\theta \tag{3}$$

$c_{skip}(\sigma)$와 $c_{out}(\sigma)$는 각각 현재 노이즈 상태와 모델 출력 항의 스케일을 조정하는 사전 조건화 계수이다. 따라서 복원 영상 추정값은 노이즈 수준에 따라 현재 노이즈 상태와 모델 예측값이 스케일링되어 선형 결합된 형태로 계산된다.

학습 단계에서는 모델 예측값 v_θ와 목표값 v^* 사이의 L1 손실을 최소화한다. v^*는 선택한 예측 파라미터화에 따라 정답 영상 x_0와 노이즈 상태 x_σ로부터 정의되는 목표값을 의미한다. 본 논문에서 사용한 학습 손실은 식 (4)와 같다.

$$L = E_{x_0,y,\sigma}[\|v_\theta - v^*\|_1] \tag{4}$$

식 (4)에서 $E_{x_0,y,\sigma}$는 학습 데이터 x_0, y와 노이즈 수준 σ에 대한 기댓값을 의미한다. 추론 단계에서는 저해상도 입력 영상 y를 조건으로 사용하여 초기 노이즈 상태로부터 반복적인 노이즈 제거를 수행한다. 이때 식 (3)의 복원 추정 과정은 각 샘플링 단계에서 수행되며, 본 연구에서는 총 30단계의 샘플링을 통해 최종 복원 영상을 생성한다.

2. 의미 및 구조 경로 조건화

제안 모델은 저해상도 입력 영상 y로부터 의미 조건 C_{sem}과 구조 특징 F_{str}을 분리하여 구성된다. 의미 조건은 영상의 전역 문맥과 질감 분포를 요약한 토큰이며, 구조 특징은 경계, 윤곽, 반복 패턴과 같은 공간적 단서를 유지하는 특징 맵이다. 저해상도 조건 인코더는 먼저 공통 저해상도 특징 F_{LR}을 추출한 뒤, 의미 경로와 구조 경로로 분기한다. 저해상도 특징 F_{LR}은 저해상도 입력 영상 y를 저해상도 인코더 $E_{LR}(\cdot)$에 통과시켜 생성한다. $E_{LR}(\cdot)$은 합성곱 계층과 ConvNeXt 블록[17]으로 구성되며, 저해상도 입력의 국소 질감과 구조 정보를 추출한다. ConvNeXt 블록은 합성곱 기반의 국소 특징 추출 능력을 유지하면서도 경계와 질감 정보를 효과적으로 표현할 수 있으므로, 저해상도 조건 특징을 추출하기 위해 사용하였다.

구조 경로는 F_{LR}에 1×1 합성곱 기반의 투영 연산 $P_{str}(\cdot)$을 적용하여 공간적으로 정렬된 구조 특징 F_{str}을 생성한다. 의미 경로는 F_{LR}을 토큰 형태로 변환한 뒤, 위치 임베딩 P와 Transformer Encoder $E_T(\cdot)$[18]를 통해 의미 조건 C_{sem}을 생성한다. 본 연구에서는 2개의 Transformer Encoder 계층과 8개 헤드의 셀프 어텐션을 사용하였다.

$$\begin{aligned} &F_{LR}=E_{LR}(y) \quad F_{str}=P_{str}(F_{LR}),\\ &C_{sem}=E_T(Tok(F_{LR})+P) \end{aligned} \tag{5}$$

식 (5)에서 $Tok(\cdot)$는 특징 맵의 공간 차원을 펼쳐 토큰 시퀀스로 변환하는 연산이다.

C_{sem}은 병목 구간의 교차 어텐션에서 Key와 Value로, 병목 특징은 Query로 사용되어, 전역 문맥과 질감 분포를 병목 표현에 반영한다. F_{str}은 디코더 특징 맵 해상도에 맞게 보간된 뒤 채널 방향으로 분리되어 위치별 스케일 항과 바이어스 항을 이루고, 이를 통해 디코더 특징 맵을 위치별로 변조하여 경계와 반복 패턴 등 공간적 단서를 복원 특징에 반영한다. 이처럼 두 경로는 공통 특징을 공유하되 전역 문맥은 병목에, 공간 구조는 복원 단계에 각각 주입된다. 이를 통해 제안 구조는 확산 모델의 생성 능력을 유지하면서 저해상도 입력과의 공간적 일관성을 높인다.

Ⅳ. 실험 결과 및 분석

1. 실험 설정

제안 모델의 성능을 검증하기 위해 Set5[19], Set14[20], BSD100[21] 및 Urban100[22] 데이터셋을 사용하였다. Set5와 Set14는 단일 영상 초해상화 분야에서 널리 사용되는 표준 평가 데이터셋으로, 다양한 자연 영상과 객체 영상을 포함한다. BSD100은 다양한 자연 영상을 포함하는 대표적인 초해상화 평가 데이터셋이며, Urban100은 건물 외벽, 창문, 격자 구조와 같이 반복 패턴을 포함하여 구조적 복원 능력 평가에 적합하다. 모든 실험은 4배 초해상화 설정에서 수행하였으며, 저해상도 입력 영상은 고해상도 영상에 Bicubic 다운샘플링을 적용하여 생성하였다.

성능 비교에는 회귀 기반 초해상화 모델인 SwinIR[9], 확산 기반 초해상화 모델인 DiT-SR[23], SinSR[12], StableSR[13]을 사용하였다. SwinIR은 윈도우 기반 트랜스포머 구조를 활용하여 복원 영상을 직접 예측하는 대표적인 회귀 기반 모델이다. DiT-SR은 확산 트랜스포머 구조를 초해상화에 적용한 방법이며, SinSR은 지식 증류를 통해 확산 기반 초해상화의 추론 단계를 줄인 방법이다. StableSR은 사전 학습된 확산 모델의 생성 사전 지식을 활용하여 초해상화 복원 성능을 향상시키는 방법이다. 제안 모델은 확산 기반 복원 방식을 사용한다는 점에서 이들 확산 기반 모델과 관련되지만, 저해상도 조건 정보를 의미

경로와 구조 경로로 분리하여 백본의 서로 다른 단계에 주입한다는 점에서 차별화된다.

정량적 성능 평가를 위해 PSNR과 SSIM을 사용하였고, 이는 각각 원본 영상과 복원 영상 간의 픽셀 단위 유사도와 구조적 유사성 평가에 널리 사용되는 지표이다. 그러나 이러한 왜곡 기반 지표는 인간이 인지하는 지각적 화질을 충분히 반영하지 못할 수 있기 때문에, 복원 영상의 지각적 유사도를 평가하기 위해 LPIPS를 함께 사용하였다[24]. LPIPS는 사전 학습된 신경망의 특징 공간에서 두 영상 간의 차이를 측정하는 지표로, 값이 낮을수록 원본 영상과 지각적으로 더 유사한 복원 결과를 의미한다. 또한 실제 사용자가 인식하는 주관적 화질을 평가하기 위해 MOS 기반의 주관적 평가를 추가로 수행하였다[25].

2. 정량적 성능 평가

정량적 성능 평가 결과는 표 1과 표 2에 나타내었다. 표 1은 Set5 및 Set14 데이터셋에서의, 표 2는 BSD100 및 Urban100 데이터셋에서의 4배 초해상화 결과를 보여준다. 실험 결과, 회귀 기반 초해상화 모델인 SwinIR은 모든 데이터셋에서 가장 높은 PSNR과 SSIM을 기록하였다. 이는 픽셀 단위 오차를 최소화하도록 학습되는 회귀 기반 복원 방식이 원본 영상과의 수치적 유사도 측면에서 강점을 가짐을 보여준다. 그러나 SwinIR은 LPIPS에서 상대적으로 높은 값을 보였으며, 이는 높은 PSNR 및 SSIM이 반드시 지각적 유사도의 향상으로 이어지지 않음을 의미한다.

제안 모델은 SwinIR보다 PSNR 및 SSIM은 낮았으나, 모든 데이터셋에서 가장 낮은 LPIPS 값을 기록하여 지각적 유사도 측면에서 우수한 성능을 보였다. 이는 왜곡 기반 지표와 지각 기반 지표 사이의 상충 관계와 관련된다. 픽셀 단위 오차를 최소화하는 회귀 기반 방식은 PSNR 및 SSIM에서 유리하지만 세부 질감이 평균화되어 지각적 자연스러움이 저하될 수 있는 반면[26], 확산 기반 복원 방식은 자연스러운 질감과 고주파 세부 정보를 생성하는 데 강점을 가져 LPIPS와 같은 지각 기반 지표에 유리하다. 두 방식은 최적화 목표가 다르므로 단일 순위로 비교하기보다 상충 관계 위에서 각 방법의 균형점을 함께 고려하는 것이 적절하며, 이러한 관점에서 제안 모델은 회귀 기반 방법의 강점

표 1. Set5 및 Set14에서의 PSNR, SSIM, LPIPS 정량적 결과
Table 1. Quantitative results on Set5 and Set14 using PSNR, SSIM, and LPIPS

Dataset	Set5			Set14		
Model	PSNR(↑)	SSIM(↑)	LPIPS(↓)	PSNR(↑)	SSIM(↑)	LPIPS(↓)
SwinIR	**30.89**	**0.8915**	0.16637	**27.03**	**0.7773**	0.26635
DiT-SR	25.68	0.7563	0.14002	23.54	0.6451	0.19908
SinSR	25.69	0.7582	0.12753	23.35	0.6223	0.18728
StableSR	25.88	0.7743	0.17778	23.33	0.6606	0.24372
Proposed	27.72	0.8153	**0.10488**	24.78	0.6976	**0.15500**

표 2. BSD100 및 Urban100에서의 PSNR, SSIM, LPIPS 정량적 결과
Table 2. Quantitative results on BSD100 and Urban100 using PSNR, SSIM, and LPIPS

Dataset	BSD100			Urban100		
Model	PSNR(↑)	SSIM(↑)	LPIPS(↓)	PSNR(↑)	SSIM(↑)	LPIPS(↓)
SwinIR	**26.60**	**0.7309**	0.35331	**25.88**	**0.8138**	0.18402
DiT-SR	23.79	0.6126	0.24757	22.08	0.6779	0.18048
SinSR	23.47	0.5963	0.22459	21.92	0.6563	0.18325
StableSR	23.80	0.6506	0.29585	20.61	0.6585	0.23387
Proposed	24.37	0.6470	**0.19050**	24.47	0.7656	**0.12841**

인 PSNR과 SSIM을 일정 수준 유지하면서 지각적 유사도에서 가장 우수한 위치를 차지한다.

제안 모델은 DiT-SR, SinSR, StableSR과 같은 확산 기반 비교 모델들과 비교했을 때, 대부분의 데이터셋에서 더 높은 PSNR 및 SSIM을 보였으며, 모든 데이터셋에서 가장 낮은 LPIPS를 기록하였다. 특히 Set5, Set14, Urban100에서는 확산 기반 비교 모델 대비 PSNR, SSIM, LPIPS가 모두 개선되었고, BSD100의 경우 StableSR보다 SSIM은 근소하게 낮았으나 PSNR과 LPIPS에서는 더 우수한 성능을 보였다. 이는 제안 모델이 지각적 유사도를 높이기 위해 왜곡 기반 성능을 희생한 것이 아니라, 확산 기반 비교 모델들과의 비교에서 대부분의 왜곡 기반 지표와 지각 기반 지표를 함께 개선하는 경향을 보였음을 의미한다. 즉 제안 모델이 의미 정보와 구조적 단서를 복원 과정에 함께 반영함으로써 전반적인 복원 품질 개선에 기여한 것으로 해석할 수 있다.

그림 2에서는 각 모델의 4배 초해상화 결과에서 특정 영역을 확대하여 세부 구조와 질감 표현을 분석하였다. SwinIR은 전반적인 윤곽과 구조를 안정적으로 복원하지만, 확대 영역에서는 고주파 세부 질감이 충분히 복원되지 않아 결과가 다소 부드럽고 흐릿하게 나타났다. 특히 얇은 선이나 미세한 경계 영역에서 세부 구조가 뚜렷하게 표현되지 않는 경향이 관찰되었다. DiT-SR, SinSR, StableSR과 같은 확산 기반 모델은 SwinIR보다 더 선명한 질감을 생성하는 경향을 보였으나, 복잡한 반복 구조에서는 원본과 다른 형태의 패턴이 생성되는 구조적 왜곡이 나타났다.

제안 모델은 세부 질감과 구조적 일관성을 비교적 안정적으로 유지하는 결과를 보였다. 의미 경로는 전역 문맥과

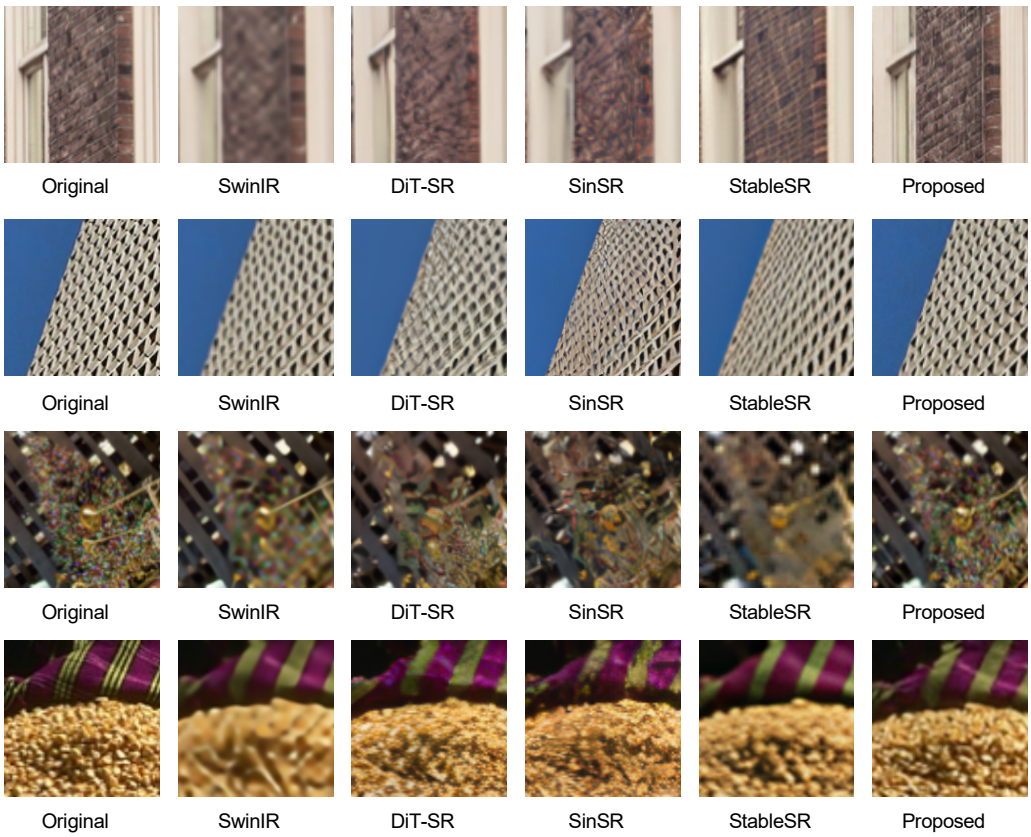

그림 2. 시각적 비교 결과
Fig. 2. Visual comparison results

질감 분포를 반영하여 흐릿함을 완화하고, 구조 경로는 경계와 공간 정보를 보존하여 구조적 왜곡을 줄였다. 그 결과 반복 패턴과 선형 구조에서도 원본과 유사한 형태를 유지하며 자연스러운 복원 결과를 생성하였다. 이러한 시각적 평가 결과는 제안 모델이 지각적 화질 개선과 구조적 왜곡 완화에 효과적으로 작용했음을 보여준다.

3. 주관적 화질 평가

PSNR, SSIM, LPIPS와 같은 정량적 성능 평가만으로는 실제 사용자가 인식하는 주관적 화질을 충분히 설명하기 어렵기 때문에, 본 논문에서는 MOS 기반의 주관적 화질 평가를 추가로 수행하였다[25]. MOS 평가는 15명의 참여자를 대상으로 블라인드 방식으로 진행되었다. BSD100 및 Urban100 데이터셋에서 다양한 특성을 지닌 10개의 영상을 선정하였으며, 각 참가자는 모델 정보가 제거된 상태에서 복원 결과를 관찰하고 선명도, 자연스러움, 전반적 화질 등을 종합적으로 고려하여 매우 나쁨을 의미하는 1점에서 매우 좋음을 의미하는 5점까지의 척도로 평가하였다. 각 모델의 MOS는 모든 참가자 평가 점수의 평균으로 계산하였으며, 95% 신뢰구간은 참가자별 평가 점수에 대해 t-분포를 적용하여 산출하였다.

표 3은 MOS 평가에 사용한 10개 영상에 대한 평균 PSNR, SSIM, LPIPS 및 MOS 결과를 함께 보여준다. 분석 결과, SwinIR은 가장 높은 왜곡 기반 지표를 기록하였으나 MOS는 제안 모델보다 낮게 나타났다. 이는 픽셀 단위 유사도가 높더라도 세부 질감이 부드럽게 복원될 경우 주관적 화질이 제한될 수 있음을 보여준다. 반면 제안 모델은 확산 기반 비교 모델들보다 높은 PSNR 및 SSIM을 유지하면서 가장 낮은 LPIPS와 가장 높은 MOS를 기록하였다. 다만 MOS는 선명도, 자연스러움, 구조적 안정성 등을 종합적으로 반영하는 주관적 평가이므로 모델별 순위가 LPIPS와 완전히 일치하지는 않았다. 그럼에도 제안 모델은 LPIPS와 주관적 화질 평가 모두에서 가장 우수한 결과를 보였다. 이는 제안 모델이 확산 기반 복원의 지각적 장점을 유지하면서 구조적 왜곡을 완화한 결과로 해석된다.

표 3. MOS 평가 영상에 대한 정량 지표와 MOS 결과
Table 3. Quantitative metrics and MOS results on the images used for MOS evaluation

Model	PSNR(↑)	SSIM(↑)	LPIPS(↓)	MOS(↑)
SwinIR	**22.01**	**0.7371**	0.27763	2.57±0.15
DiT-SR	20.01	0.6497	0.21766	2.43±0.17
SinSR	19.76	0.6319	0.19742	2.00±0.16
StableSR	19.37	0.6458	0.24779	2.50±0.14
Proposed	21.39	0.7098	**0.16115**	**3.73**±0.16

4. Ablation Study

의미 경로와 구조 경로가 복원 성능에 미치는 영향을 분석하기 위해, BSD100 데이터셋을 대상으로 의미 경로를 제거한 경우와 구조 경로를 제거한 경우를 각각 제안 모델과 비교하였으며, 그 결과를 표 4에 정리하였다. 모든 모델은 동일한 4배 초해상화 설정에서 PSNR, SSIM, LPIPS로 평가하였다. 의미 경로를 제거하고 구조 경로만 사용한 경우, 공간적으로 정렬된 구조 특징이 주로 반영되어 PSNR 및 SSIM은 가장 높았으나, LPIPS 역시 가장 높게 나타났다. 이는 구조 경로만으로는 지각적 질감 표현을 충분히 개선하기 어려움을 보여준다. 반대로 구조 경로를 제거하고 의미 경로만 사용한 경우에는 PSNR 및 SSIM은 낮아졌지만 LPIPS는 개선되었으며, 이는 전역 문맥 및 질감 조건이 지각적 유사도 개선에 기여함을 보여준다.

표 4. BSD100에서의 의미 경로와 구조 경로에 대한 절제 실험 결과
Table 4. Ablation study of the Semantic Path and Structural Path on BSD100

Semantic Path	Structural Path	PSNR(↑)	SSIM(↑)	LPIPS(↓)
X	O	**24.64**	**0.6748**	0.21829
O	X	24.41	0.6678	0.21365
O	O	24.37	0.6470	**0.19050**

두 경로를 모두 사용한 제안 모델은 PSNR과 SSIM에서는 두 절제 모델보다 낮은 값을 보였으나, LPIPS는 0.19050으로 가장 낮은 값을 달성하였다. 이는 제안 모델

이 왜곡 기반 지표를 최대화하는 방향보다는 지각적 유사도 개선 측면에서 더 뚜렷한 효과를 보였음을 의미한다. 두 경로를 함께 사용할 경우 구조 경로가 제공하는 공간적으로 정렬된 구조 단서와 의미 경로가 제공하는 전역 문맥 및 질감 조건이 복원 과정에서 함께 반영되며, 그 결과 단일 경로만 사용하는 경우보다 지각적 유사도가 개선되었다.

V. 결 론

본 논문은 확산 기반 초해상화에서 발생할 수 있는 구조적 왜곡과 회귀 기반 모델의 흐릿함 문제를 완화하기 위해, 저해상도 조건 정보를 전역 문맥을 담당하는 의미 경로와 공간 구조를 담당하는 구조 경로로 나누어 조건화하는 확산 기반 초해상화 모델을 제안하였다. 의미 경로는 영상 전체의 문맥과 질감 정보를 복원 과정에 제공하고, 구조 경로는 경계와 공간 패턴을 반영하여 저해상도 입력과의 구조적 일관성을 보조한다. 두 경로를 함께 사용함으로써 제안 모델은 지각적 화질과 구조적 안정성 간의 균형을 향상시키고자 하였다. 또한 의미 조건은 병목 구간의 교차 어텐션에 주입하여 전역 문맥 정보를 반영하고, 구조 특징은 복원 단계의 특징 맵 변조에 주입하여 위치별 구조 정보를 반영하도록 하였다.

Set5, Set14, BSD100 및 Urban100 데이터셋을 사용한 4배 초해상화 실험에서 제안 모델은 비교 확산 모델 대비 대부분의 데이터셋에서 더 높은 PSNR 및 SSIM을 기록하였으며, 모든 비교 모델 중 가장 낮은 LPIPS를 달성하였다. 또한 MOS에서도 가장 높은 점수를 기록하여, 제안 모델이 확산 기반 초해상화에서 지각적 화질 개선에 효과적으로 작용하며, 구조적 안정성과 지각적 화질 간의 균형을 개선할 수 있음을 확인하였다.

참 고 문 헌 (References)

[1] C. Dong et al., “Image Super-Resolution Using Deep Convolutional Networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.38, No.2, pp.295-307, February 2016. doi: https://doi.org/10.1109/TPAMI.2015.2439281

[2] K. Nasrollahi and T. B. Moeslund, “Super-Resolution: A Comprehensive Survey,” *Machine Vision and Applications*, Vol.25, No.6, pp.1423-1468, August 2014. doi: https://doi.org/10.1007/s00138-014-0623-4

[3] J. D. Joshua et al., “DMFASR: Dense Multi-Feature Aggregation-Based Super-Resolution for Brain MR Images,” *Image and Vision Computing*, Vol.161, Art. no.105652, pp.1-12, September 2025. doi: https://doi.org/10.1016/j.imavis.2025.105652

[4] C. Ledig et al., “Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network,” *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.4681-4690, 2017. doi: https://doi.org/10.1109/CVPR.2017.19

[5] C. Saharia et al., “Image Super-Resolution via Iterative Refinement,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.45, No.4, pp.4713-4726, September 2022. doi: https://doi.org/10.1109/TPAMI.2022.3204461

[6] B. Lim et al., “Enhanced Deep Residual Networks for Single Image Super-Resolution,” *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp.1132-1140, 2017. doi: https://doi.org/10.1109/CVPRW.2017.151

[7] Y. Zhang et al., “Residual Dense Network for Image Super-Resolution,” *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.2472-2481, 2018. doi: https://doi.org/10.1109/CVPR.2018.00262

[8] Y. Zhang et al., “Image Super-Resolution Using Very Deep Residual Channel Attention Networks,” *Proceedings of the European Conference on Computer Vision*, pp.294-310, 2018. doi: https://doi.org/10.1007/978-3-030-01234-2_18

[9] J. Liang et al., “SwinIR: Image Restoration Using Swin Transformer,” *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp.1833-1844, 2021. doi: https://doi.org/10.1109/ICCVW54120.2021.00210

[10] J. Ho et al., “Denoising Diffusion Probabilistic Models,” arXiv:2006.11239, 2020. doi: https://doi.org/10.48550/arXiv.2006.11239

[11] Z. Yue et al., “ResShift: Efficient Diffusion Model for Image Super-Resolution by Residual Shifting,” *Advances in Neural Information Processing Systems*, Vol.36, pp.13294-13307, 2023. doi: https://doi.org/10.52202/075280-0583

[12] Y. Wang et al., “SinSR: Diffusion-Based Image Super-Resolution in a Single Step,” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.25796-25805, 2024. doi: https://doi.org/10.1109/CVPR52733.2024.02437

[13] J. Wang et al., “Exploiting Diffusion Prior for Real-World Image Super-Resolution,” *International Journal of Computer Vision*, Vol.132, No.12, pp.5929-5949, December 2024. doi: https://doi.org/10.1007/s11263-024-02168-7

[14] X. Lin et al., “DiffBIR: Toward Blind Image Restoration with Generative Diffusion Prior,” *Proceedings of the European*

Conference on Computer Vision, pp.430-448, 2024.
doi: https://doi.org/10.1007/978-3-031-73202-7_25
[15] T. Karras et al., "Elucidating the Design Space of Diffusion-Based Generative Models," *Advances in Neural Information Processing Systems*, Vol.35, pp.26565-26577, 2022.
doi: https://doi.org/10.52202/068431-1926
[16] E. Perez et al., "FiLM: Visual Reasoning with a General Conditioning Layer," *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol.32, No.1, 2018.
doi: https://doi.org/10.1609/aaai.v32i1.11671
[17] Z. Liu et al., "A ConvNet for the 2020s," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.11976-11986, 2022.
doi: https://doi.org/10.1109/CVPR52688.2022.01167
[18] A. Vaswani et al., "Attention Is All You Need," arXiv:1706.03762, 2017.
doi: https://doi.org/10.48550/arXiv.1706.03762
[19] M. Bevilacqua et al., "Low-Complexity Single-Image Super-Resolution based on Nonnegative Neighbor Embedding," *Proceedings of the British Machine Vision Conference*, 2012.
doi: https://doi.org/10.5244/C.26.135
[20] R. Zeyde et al., "On Single Image Scale-Up Using Sparse-Representations," *Proceedings of the International Conference on Curves and Surfaces*, pp.711-730, 2012.
doi: https://doi.org/10.1007/978-3-642-27413-8_47
[21] D. Martin et al., "A Database of Human Segmented Natural Images and its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics," *Proceedings of the Eighth IEEE International Conference on Computer Vision*, Vol.2, pp.416-423, 2001.
doi: https://doi.org/10.1109/ICCV.2001.937655
[22] J. B. Huang et al., "Single Image Super-Resolution from Transformed Self-Exemplars," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.5197-5206, 2015.
doi: https://doi.org/10.1109/CVPR.2015.7299156
[23] K. Cheng et al., "Effective Diffusion Transformer Architecture for Image Super-Resolution," *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol.39, No.3, pp.2455-2463, 2025.
doi: https://doi.org/10.1609/aaai.v39i3.32247
[24] R. Zhang et al., "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.586-595, 2018.
doi: https://doi.org/10.1109/CVPR.2018.00068
[25] ITU-R, "Methodology for the Subjective Assessment of the Quality of Television Pictures," Recommendation ITU-R BT.500-13, 2012, Retrieved from https://www.itu.int/rec/R-REC-BT.500
[26] Y. Blau and T. Michaeli, "The Perception-Distortion Tradeoff," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.6228-6237, 2018.
doi: https://doi.org/10.1109/CVPR.2018.00652

저 자 소 개

김 태 환

- 2024년 : 세종대학교 지능기전공학부 스마트기기공학 학사
- 2026년 : 세종대학교 AI로봇학과 석사
- 2026년 ~ 현재 : 세종대학교 AI로봇학과 석사연구원
- ORCID : https://orcid.org/0009-0003-8714-8875
- 주관심분야 : 영상처리, 컴퓨터비전, 딥러닝

이 진 영

- 2006년 : 성균관대학교 정보통신공학부 학사
- 2008년 : KAIST 전기및전자공학부 석사
- 2018년 : KAIST 전기및전자공학부 박사
- 2008년 ~ 2018년 : 삼성전자 책임연구원
- 2018년 ~ 현재 : 세종대학교 AI로봇학과 교수
- ORCID : https://orcid.org/0000-0002-4384-2633
- 주관심분야 : 영상처리, 비디오압축