

적대적 손실 기반 하이브리드 디퓨전 모델의 컬러 이미지 생성 성능 개선

이 중 현¹ · 최 웅^{2*}

¹강남대학교 ICT융합공학부 학사과정

²강남대학교 컴퓨터공학부 교수

Improvement of Color Image Generation Performance of Adversarial Loss-Based Hybrid Diffusion Models

Jung-Hyeon Lee¹ · Woong Choi^{2*}

¹Bachelor's Degree, School of ICT Convergence Engineering, Kangnam University, Yongin 16979, Korea

²Professor, Department of Computer Engineering, Kangnam University, Yongin 16979, Korea

[요 약]

최근 디퓨전 확률 모델을 활용한 이미지 생성 연구가 활발하나, 전통적인 평균 제곱 오차(Mean Squared Error, MSE) 손실 함수 기반의 훈련은 고주파 영역의 미세 텍스처가 묻개지는 블러 현상을 야기한다. 선행 연구는 흑백 이미지 환경에서 적대적 손실을 결합해 이를 개선했으나, 복잡한 컬러 데이터셋에 단순 적용 시 수렴 불안정성으로 인해 성능이 저하되는 한계를 보였다. 본 연구에서는 이러한 제약을 극복하기 위해 역과정의 타임 임베딩 차원을 재조정하고 텐서 디바이스 매칭 구조를 교정한, 디퓨전 확률 모델(Denoising Diffusion Probabilistic Models, DDPM)과 생성적 적대 신경망(Generative Adversarial Network, GAN)을 결합한 개량형 하이브리드 아키텍처를 제안한다. 나아가 컬러 환경 확장에 따른 판별자의 과도한 패널티를 완화하도록 적대적 가중치를 최적화했다. 실험 결과, 제안 모델은 기존 DDPM 대비 더 낮은 FID(Fréchet Inception Distance) Score를 기록하였으며, 적대적 손실 가중치의 적절한 조정이 컬러 이미지 생성 품질 향상에 기여함을 확인하였다.

[Abstract]

Recently, denoising diffusion probabilistic models (DDPMs) have advanced remarkably; however, traditional mean squared error (MSE) training causes blurring in high-frequency textures. Previous studies integrated adversarial loss in grayscale environments but failed to scale to multi-channel color datasets due to convergence instability. To overcome this limitation, this study proposes an enhanced hybrid framework that combines a generative adversarial network (GAN) and a DDPM by recalibrating time-embedding dimensions and resolving device-indexing mismatches in the reverse process. Furthermore, we optimize the adversarial weight to alleviate excessive discriminator penalties. The results demonstrate that our proposed model successfully overcomes previous convergence failures, achieving Fréchet Inception Distance (FID) reductions of 2.6% (CIFAR-10) and 4.0% (anime-face) over the baseline DDPM.

색인어 : 디퓨전 모델, DDPM, 적대적 손실 함수, 컬러 데이터셋, 이미지 생성 품질

Keyword : Diffusion Models, DDPM, Adversarial Loss, Color Dataset, Image Generation Quality

<http://dx.doi.org/10.9728/dcs.2026.27.9.2623>



This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Received 12 June 2026; **Revised** 16 July 2026

Accepted 05 August 2026

***Corresponding Author, Woong Choi**

Tel: +82 31-280-3753

E-mail: wchoi@kangnam.ac.kr

I. 서 론

다양한 형태의 이미지나 콘텐츠를 창작하는 생성형 모델의 발전 과정에서 디퓨전 확률 모델(Diffusion Probabilistic Models)은 최근 이미지 생성 분야에서 FID 등 분포 적합도 지표에서 경쟁력 있는 결과가 보고되며 대표적 생성 모델로 활용되고 있다[1]. 기존 생성 모델의 양대 축이었던 생성적 적대 신경망(Generative Adversarial Network, GAN)[2]은 세밀한 표현에 유리하나 훈련 과정이 불안정하고 데이터 분포의 일부를 놓치는 모드 붕괴(Mode Collapse) 한계를 지니며, 변분 오토인코더(VAE)[3]는 안정적인 수렴을 보장하지만 픽셀 평균치 학습 특성으로 인해 결과물이 흐릿해지는 고질적인 문제를 안고 있다. 반면 디퓨전 모델은 순방향 과정에서 주입된 노이즈를 역과정을 통해 단계적으로 제거해 나가는 마르코프 체인을 활용함으로써[1], GAN의 학습 불안정성과 VAE의 평균화에 따른 블러 문제를 동시에 완화하는 것으로 보고되었다. 이에 본 연구에서는 디퓨전 모델의 대표적인 구조인 DDPM(Denoising Diffusion Probabilistic Models)을 기반 모델로 선정하고, 적대적 손실 함수를 결합한 하이브리드 구조를 통해 생성 품질 향상 가능성을 분석하고자 한다. 한편, DDPM 발표 이후에도 잠재 공간에서 연산을 수행하여 효율성을 높인 latent diffusion 모델[4], 별도의 분류기 없이 조건부 생성 품질을 향상시키는 classifier-free guidance 기법[5], 단일 단계 샘플링을 가능케 하는 consistency model[6] 등 후속 연구가 활발히 이루어지고 있다. 그러나 DDPM 역시 몇 가지 한계를 가진다. 예측된 노이즈와 실제 노이즈 간의 평균 제곱 오차(Mean Squared Error, MSE)를 주 목적으로 학습되기 때문에, 데이터의 픽셀 단위 통계적 평균치에 안전하게 수렴하려는 성질을 띤다. 이러한 특성은 전체적인 이미지 구조를 안정적으로 복원하는 데는 유리하지만, 윤곽선이나 텍스처와 같은 고주파 성분의 표현을 약화시켜 결과 이미지가 흐릿해지는 블러(Blurring) 현상을 유발할 수 있다.

이를 극복하기 위해 선행 연구에서는 흑백 이미지(MNIST) 환경에서 판별자(Discriminator)를 통한 적대적 손실(adversarial loss)을 결합하는 하이브리드 프레임워크를 제안하였다[7]. 그러나 선행 연구는 다채널 컬러 이미지 데이터셋인 CIFAR-10 환경에 동일 구조를 적용했을 때, 판별자의 과도한 패널티와 학습 불안정성으로 인해 오히려 이미지 생성 품질이 저하되는 한계점을 남겼다. 본 연구에서는 컬러 이미지 생성 환경에서 적대적 손실 함수가 학습 안정성에 미치는 영향을 분석하고, 적대적 손실 가중치 최적화를 통해 DDPM과 GAN 기반 손실의 균형점을 탐색하였다. 특히 핵심 하이퍼파라미터인 적대적 손실 가중치(λ_{adv})의 정밀한 미세 조정(Fine-tuning)을 통해, DDPM의 안정적인 데이터 분포 학습 능력과 GAN의 고주파 텍스처 복원 특성을 동시에 활용하는 최적의 균형점을 탐색하고, 그 효과를 정량적 평가 지표인 FID(Fréchet Inception Distance)[8]를 통해 검증

하고자 한다.

한편, 디퓨전 모델의 학습 목적함수에 적대적 손실을 선형 결합하는 접근 자체는 본 연구 이전에도 여러 형태로 보고된 바 있다. 대표적으로 Adversarial Diffusion Distillation[9]은 대규모 사전학습(pre-trained) teacher 디퓨전 모델의 지식을, 적은 샘플링 스텝만으로도 고품질 생성이 가능한 student 모델로 증류(distillation)하기 위해 teacher-student 증류 손실과 적대적 손실을 $L_{ADD} = L_{adv} + \lambda \cdot L_{distill}$ 의 형태로 결합하였다. 또한 Denoising Diffusion GAN[10]은 역확산 과정의 각 전이(transition) 단계를 다중 모드(multimodal) 조건부 GAN으로 근사하여 샘플링에 필요한 스텝 수 자체를 수 단계 수준으로 대폭 줄이는 데 목적을 두었으며, Diffusion-GAN[11]은 판별자의 입력단에 확산 기반의 점진적 노이즈 주입을 적용하여 GAN 자체의 학습 붕괴를 방지하는 데 초점을 두었다. 이들 선행 하이브리드 연구는 공통적으로 대규모 사전학습 teacher 모델을 전제로 한 증류 설정이거나, 샘플링 스텝 수 축소 또는 GAN 자체의 학습 안정화를 목적으로 한다. 아울러 손실 가중치 λ 는 대체로 고정값 또는 휴리스틱으로 설정될 뿐, 데이터셋 특성에 따른 민감도 분석은 다루지 않는다는 공통점을 지닌다.

이에 반해 본 연구는 teacher 네트워크 없이 DDPM을 처음부터(from-scratch) 학습시키는 표준 학습 설정을 유지한 채, 노이즈 예측기의 픽셀 단위 복원 손실(L_{simple})에 판별자 기반 적대적 손실만을 결합한다는 점에서 증류 기반 선행 연구들과 문제 설정 자체가 다르다. 아울러 본 연구의 목적은 샘플링 스텝 수 축소가 아니라 동일한 스텝 수 하에서의 고주파 텍스처 복원 품질 개선에 있다. 이러한 차이를 전제로, 본 연구의 기여는 새로운 손실 결합 구조를 제안하는 데 있지 않다. 대신 클래스 다양성이 큰 다중 클래스 자연영상(CIFAR-10)과, 구조적 유사성이 높은 단일 도메인 영상(애니메이션 얼굴)이라는 성격이 상반된 두 데이터셋에 대해 적대적 손실 가중치(λ_{adv})를 동일한 실험 범위 내에서 정량적으로 비교·분석하였다. 이를 통해 데이터셋의 분포 복잡도에 따라 최적 λ_{adv} 및 학습 안정성이 어떻게 달라지는지를 실증적으로 규명했다는 데 본 연구의 기여가 있다. 즉 본 연구는 선행연구[7]에서 흑백 환경에 한정되어 있던 하이브리드 구조의 실험적 검증 범위를 컬러 다중 채널 환경으로 확장하고, 그 과정에서 데이터셋 특성별 λ_{adv} 민감도에 대한 실증적 근거를 제시한 경험적 분석 연구로 자리매김한다.

II. 관련 연구

2-1 확산 모델 기반 이미지 생성의 발전

Ho et al.은 2020년 DDPM을 제안하여 마르코프 연쇄 기반의 순방향 노이즈 추가 과정과 역방향 복원 과정을 통해

GANs에 필적하는 수준의 이미지 생성 품질을 달성하였다 [1]. 순방향 과정에서는 원본 이미지 x_0 에 분산 일정표 $T(\beta_t)_{t=1}^T$ 에 따라 가우시안 노이즈를 점진적으로 추가하며, 역방향 과정에서 신경망 ϵ_θ 가 각 타임스텝 t 의 노이즈를 추정하여 이미지를 복원한다.

이때 역과정에서 각 타임스텝(t)마다 잔존하는 노이즈를 정밀하게 예측하는 핵심 신경망 백본으로는 대칭형 스킵 커넥션(Skip Connection) 구조를 지닌 U-Net 아키텍처[12]가 표준적으로 채택되어 연산을 수행한다. 학습 목적함수는 다음과 같이 단순화된 MSE로 정의된다.

$$Loss_{MSE} = E_{t, x_0} [\| \epsilon - \epsilon_\theta(x_t, t) \|^2] \quad (1)$$

그러나 이 MSE 목적함수는 픽셀 분포를 평균화하는 특성으로 인해 고주파 세부 텍스처가 손실되고 블러(Blurring) 현상이 발생하는 구조적 한계를 지닌다. DDPM은 CIFAR-10에서 FID 3.17을 달성하였으나, 수천 스텝의 샘플링이 필요하여 추론 속도가 느리다는 문제도 지적되었다.

이를 개선하기 위해 Song et al.은 DDIM(Denoising Diffusion Implicit Models)을 제안하였다[13]. DDIM은 동일한 학습 목적함수를 유지하면서 비마르코프적(non-Markovian) 결정론적 역과정을 도입하여, DDPM 대비 10배에서 50배 빠른 샘플링을 가능하게 하였다. 이는 샘플링 스텝 수를 1,000에서 수십 수준으로 줄이면서도 생성 품질을 유지한다는 점에서 실용적인 기여가 크다. 그러나 DDIM 역시 MSE 기반 학습을 그대로 사용하므로 블러 현상의 근본적 해결에는 한계가 있다.

2-2 생성적 적대 신경망과 훈련 불안정성

Goodfellow et al.이 제안한 GAN은 생성자(Generator)와 판별자(Discriminator)의 적대적 학습을 통해 MSE 기반 모델이 재현하기 어려운 고주파 텍스처를 효과적으로 생성한다 [2]. 그러나 GAN의 훈련 과정은 본질적으로 불안정하며, 크게 세 가지 문제를 수반한다. 첫째, 모드 붕괴(mode collapse)로 생성자가 전체 데이터 분포를 포괄하지 못하고 일부 모드만 반복적으로 생성하는 현상이다. 둘째, 비수렴(non-convergence)으로 생성자와 판별자가 내쉬 균형(Nash equilibrium)에 도달하지 못하고 진동하는 현상이다. 셋째, 그래디언트 소실(vanishing gradient)로 판별자가 과도하게 강력해질 경우 생성자로 전달되는 학습 신호가 소멸하여 훈련이 정체되는 현상이다. 이러한 문제는 단채널 흑백 이미지 환경에서도 발생하지만, R·G·B 세 채널을 동시에 처리하는 컬러 이미지 환경에서는 판별자가 더 많은 피쳐 공간에서 패널티를 부과하게 되어 불안정성이 심화된다.

2-3 초기 하이브리드 GAN-DDPM 모델의 한계 분석 및 개선 방향

기존 DDPM은 노이즈 예측 기반의 안정적인 확률 분포 학습이 가능하다는 장점이 있으나, 평균 제곱 오차(MSE)를 중심으로 최적화되기 때문에 이미지의 고주파 성분 복원에는 구조적인 한계를 가진다. 본 연구에서는 이러한 문제를 해결하기 위해 적대적 손실을 추가한 하이브리드 구조를 설계하였다.

그러나 초기 구현 단계에서 단순히 GAN 기반 판별자의 피드백을 DDPM 학습 과정에 결합하는 방식은 컬러 이미지 환경에서 여러 문제점을 발생시켰다. 초기 모델은 Binary Cross Entropy 기반 적대적 손실을 사용하였으며, 판별자가 실제 이미지와 생성 이미지 사이의 미세한 차이를 빠르게 학습하면서 생성자보다 우세해지는 현상이 나타났다.

특히 RGB 컬러 이미지 데이터는 단일 채널 데이터보다 더 많은 특징 공간을 포함하기 때문에 판별자가 색상 정보와 공간적 특징을 동시에 활용할 수 있다. 이 과정에서 생성자는 안정적인 노이즈 복원보다 판별자를 속이기 위한 방향으로 과도하게 업데이트되었으며, 일부 생성 결과에서는 색상 왜곡 및 비현실적인 고주파 패턴이 발생하였다.

이러한 문제를 해결하기 위해 본 연구에서는 단순한 적대적 손실 결합 방식에서 벗어나 세 가지 방향의 구조 개선을 수행하였다. 첫째, 데이터셋별 적대적 손실 가중치(λ_{adv})를 조절하여 DDPM의 복원 능력과 GAN의 세부 표현 능력 사이의 균형점을 탐색하였다. 둘째, Spectral Normalization 기반 판별자 구조를 적용하여 판별자의 과도한 업데이트를 제한하였다. 셋째, Exponential Moving Average(EMA)를 적용하여 Generator 파라미터 변화에 따른 생성 결과의 불안정성을 감소시켰다.

2-4 이미지 생성 품질 평가 지표

생성 모델의 성능 평가에는 Heusel et al.이 2017년 제안한 FID(Fr chet Inception Distance)가 사실상의 표준 지표로 사용된다[8]. FID는 InceptionV3 네트워크의 잠재 공간에서 실제 이미지와 생성 이미지의 특징 벡터 분포를 각각 다변량 가우시안으로 모델링하고, 두 분포 간의 프레스 거리를 계산한다.

$$FID = \|\mu_r - \mu_g\|_2^2 + \text{tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2}) \quad (2)$$

여기서 μ_r , Σ_r 는 실제 이미지 특징 분포의 평균과 공분산이며, μ_g , Σ_g 는 생성 이미지 특징 분포의 평균과 공분산을 의미한다. FID 값이 낮을수록 생성 이미지의 특징 분포가 실제 데이터 분포와 가까움을 의미하며, 이는 보다 사실적인 이미지 생성 품질을 나타낸다. 기존 DDPM 기반 연구에서는 주로 MSE 기반 노이즈 예측 오차를 활용하여 모델을 최적화

하지만, MSE는 생성 이미지의 픽셀 복원 정확도만 반영할 뿐 실제 이미지 분포와의 유사성을 직접적으로 평가하지 못한다.

특히 본 연구에서 목표로 하는 고주파 텍스처 및 세부 구조 개선 효과는 단순 픽셀 오차만으로 판단하기 어렵기 때문에, 생성 데이터의 분포 적합성을 평가할 수 있는 FID를 주요 성능 평가 지표로 선정하였다.

III. 제안 방법

3-1 컬러 최적화 아키텍처

본 연구에서는 컬러 이미지 환경에서 발생하는 학습 불안정성을 완화하기 위해 DDPM 기반 구조에 적대적 손실을 결합한 하이브리드 아키텍처를 설계하였다.

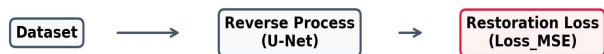


그림 1. 오리지널 DDPM 구조도

Fig. 1. Original DDPM structure

그림 1은 오리지널 DDPM의 기본 연산 구조를 나타낸다. 기존 DDPM은 데이터셋으로부터 입력된 이미지에 대해 역확산 과정을 수행하고, 복원 손실(MSE)을 최소화하는 방식으로 학습된다. 그러나 이러한 구조는 데이터 분포 학습에는 효과적이지만 고주파 텍스처 복원에는 한계를 가진다.

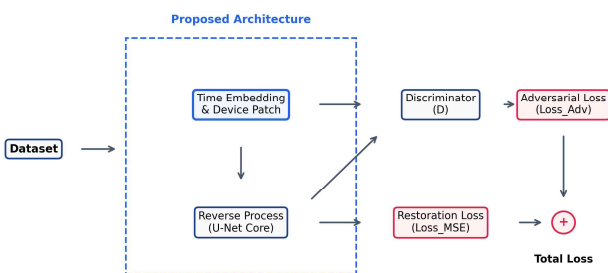


그림 2. 제안된 하이브리드 GAN-DDPM 구조도

Fig. 2. Proposed hybrid GAN-DDPM architecture

그림 2는 본 연구에서 제안한 하이브리드 GAN-DDPM 구조를 나타낸다. 기존 DDPM의 역확산 과정에 적대적 손실을 추가하여 복원 손실과 적대적 손실을 동시에 최소화하도록 설계하였다. 또한 Time Embedding 및 Device Patch 모듈을 통해 컬러 데이터셋 환경에서 발생하는 학습 불안정성을 완화하였다.

3-2 구조적 최적화 메커니즘

해상도 및 컬러 채널 확장에 따른 단일 GPU의 연산 병목 및 VRAM 초과(OOM) 문제를 통제하기 위해 아키텍처 내에 효율적인 최적화 메커니즘을 내장하였다. 고해상도 연산 구간(16×16 초과 부위)에서는 거대한 연산 자원을 요구하는 Self-Attention 블록의 연산을 생략하고 우회하도록 U-Net 로직을 개량하였으며, 대신 GroupNorm 및 SiLU 활성화 함수 기반의 컨볼루션 스택을 경유하게 하여 속도와 메모리 효율을 확보하였다. 또한 추론용 모델의 가중치가 급격히 튀는 현상을 막기 위해 지수 이동 평균(EMA) 가중치 보존 로직을 결합하였다.

3-3 Generator 구조

Generator는 대칭형 스킵 커넥션을 갖는 U-Net으로 구성되며, 채널 폭은 [64, 128, 256, 512]의 4단계로 확장·축소된다. 각 해상도 단계(Down/Up)는 GroupNorm(8그룹)-SiLU-Conv(3×3)로 구성된 정규화-활성화-합성곱 블록(NormActConv) 2개(num_layers=2)를 순차 배치한 잔차(Residual) 구조이며, 다운샘플링은 stride=2 Conv2d(kernel 4), 업샘플링은 stride=2 ConvTranspose2d(kernel 4)로 수행한다. 타임스텝 임베딩은 사인·코사인 기반 위치 임베딩(128 차원)을 2-layer MLP(Linear-SiLU-Linear)로 1차 투영한 뒤, 각 블록 내에서 SiLU-Linear로 채널 수에 맞게 재투영되어 특성 맵에 브로드캐스트 형태로 더해진다. 고해상도 연산 구간(16×16 초과 부위)에서는 연산 비용이 큰 Self-Attention 블록을 생략하고 GroupNorm/SiLU/Conv 스택으로 우회하도록 구성하여 VRAM 사용량을 제어하였으며, 16×16 이하 해상도 구간에서만 Self-Attention(Multi-Head, 4-head)을 선택적으로 적용하였다.

3-4 판별자(Discriminator) 아키텍처 구조

제안하는 하이브리드 프레임워크에서 디퓨전 역과정의 생성기(U-Net) 피드백을 담당하는 판별자(Discriminator)는 시간 조건부(time-conditional) CNN으로 설계되었다. 입력 이미지에 대해 5단계 합성곱(64 → 128 → 256 → 512 → 1 채널, kernel 4×stride 2×padding 1, LeakyReLU(0.2))을 거쳐 최종적으로 4×4 특성 맵을 1×1 스칼라 판별 점수로 축약한다. U-Net이 각 타임스텝에서 점진적으로 노이즈를 제거하여 추론한 예측 이미지 분포와 원본 컬러 이미지 분포를 판별자에 교차 입력하여 적대적 손실을 산출하며, 이때 각 해상도 채널(64/128/256/512)에 타임스텝 임베딩을 투영하는 4개의 보조 MLP 분기(time_mlp1~4)를 두어 시간 조건 정보를 반영하였다.

특히 다채널 이미지 환경에서는 판별자가 생성기보다 과도하게 빠르게 수렴하여 학습 균형이 붕괴되는 현상이 빈번하

게 발생한다. 이렇게 도출된 적대적 오차는 기존 DDPM의 목적 함수에 규제 항(Regularization term)으로 더해져 역전과 된다.

초기 모델에서는 Binary Cross Entropy 기반 GAN 손실을 사용하였으나, 디퓨전 모델의 노이즈 예측 과정과 결합될 경우 판별자가 빠르게 최적화되어 생성자에게 불안정한 gradient를 전달하는 문제가 발생하였다.

이를 해결하기 위해 개선 모델에서는 Spectral Normalization 기반 Convolution Layer를 적용하였다. 구체적으로, 5개 합성곱 계층 전체와 4개의 보조 MLP 분기(time_mlp1~4)에 이르기까지 판별자 내 모든 Conv2d 및 Linear 계층에 Spectral Normalization을 예외 없이 적용하여 판별자의 립시츠(Lipschitz) 상수를 전역적으로 제한하였다. 이를 통해 판별자의 업데이트 폭이 제어되어 생성자와 판별자 사이의 학습 균형이 유지된다.[14]

또한 기존 32×32 입력 환경에서 설계된 판별자를 64×64 컬러 이미지 환경에 맞게 확장하여 4단계 Feature Extraction 구조를 구성하였다.

3-5 하이브리드 손실 함수

제안 모델의 전체 최적화 목적 함수는 디퓨전의 확률 분포 안정성을 보존하는 복원 손실과 텍스처 선명도를 보완하는 적대적 피드백 손실을 선형 결합하여 아래 식 (3)와 같이 구성된다.

$$Loss_{Total} = Loss_{MSE} + \lambda_{adv} \times Loss_{Adv} \quad (3)$$

다차원 컬러 공간에서 과도한 적대적 규제 패널티는 생성기에 무리한 고주파 복원을 강제하여 지엽적 픽셀 왜곡(Artifact)을 일으킨다. 이에 본 연구에서는 판별자의 규제 강도를 정밀하게 완화하여 FID를 최소화하는 λ_{adv} 영역을 탐색하는 가중치 제어 전략을 수립하였다. 하이퍼파라미터 탐색 실험 결과, 일반 목적 컬러 데이터셋인 CIFAR-10 환경에서는 $\lambda_{adv} = 0.001$ 일 때 가장 안정적인 수렴을 보였으며, 본 연구의 핵심 표적인 고해상도 애니메이션 캐릭터 얼굴 데이터셋 환경에서는 $\lambda_{adv} = 0.0002$ 에서 본 실험 범위 내 최저 FID(70.51)를 기록하였다.

3-6 EMA 기반 추론 안정화

GAN-DDPM 구조에서는 Generator와 Discriminator가 동시에 업데이트되기 때문에 학습 과정에서 parameter oscillation이 발생할 수 있다. 이를 억제하기 위해 두 네트워크는 각각 독립적인 Adam($\beta_1=0.5$, $\beta_2=0.999$) 옵티마이저로 학습되며, 초기 학습률은 Generator $lr_G=1 \times 10^{-4}$, Discriminator $lr_D=2 \times 10^{-5}$ 로 서로 다르게 설정하여 판별자를 생성기보다 상대적으로 느리게 학습시킴으로써 판별자

우위로 인한 학습 붕괴를 억제하였다.

아울러 본 연구에서는 Generator의 추론 안정성을 높이기 위해 Exponential Moving Average(EMA)를 적용하였다. EMA 연산은 훈련 단계의 현재 가중치 θ_G 와 이전 스텝의 EMA 가중치 θ_{EMA} 를 활용하여 아래 식 (4)와 같이 지속적으로 업데이트된다[1].

$$\theta_{EMA} = \beta \theta_{EMA} + (1 - \beta) \theta \quad (4)$$

따라서 최종 이미지 생성 및 정량 평가(FID 측정)는 학습 중 직접 업데이트된 Generator가 아닌 EMA 기반 Generator weight를 사용하여 보다 안정적인 샘플링 결과를 확보하였다.

본 실험에서는 파라미터 감쇠계수(Decay Rate)인 β 를 0.9999의 고정값으로 설정하였다. 이를 통해 판별자의 일시적인 과잉 패널티로 인해 훈련 중 생성기 가중치가 국소적으로 손상되더라도, 최종 이미지 생성 과정에서는 누적된 EMA 기반 가중치를 사용하여 보다 안정적이고 품질 변동이 적은 샘플링 결과를 확보하였다.

3-7 구현 세부사항 및 학습 환경

재현성 확보를 위해 본 연구의 구현 세부사항 및 학습 환경을 다음과 같이 명시한다. Generator와 Discriminator는 모두 Adam($\beta_1=0.5$, $\beta_2=0.999$) 옵티마이저로 학습하였으며, 초기 학습률은 Generator $lr_G=1 \times 10^{-4}$, Discriminator $lr_D=2 \times 10^{-5}$ 로 서로 다르게 설정하여 판별자를 생성기보다 상대적으로 느리게 학습시킴으로써 판별자 우위로 인한 학습 붕괴를 억제하였다. 두 옵티마이저에는 각각 독립적으로 CosineAnnealingLR 스케줄러(T_{max} =총 epoch 수, $n_{min}=1 \times 10^{-6}$)를 적용하였다.

학습은 배치 크기 32, 최대 80 epoch로 진행하였다. 학습 중 MSE 손실이 최저치를 갱신할 때마다 Best 체크포인트를 저장하였으며, 이와 별도로 10 epoch 주기마다 Generator, EMA-Generator, Discriminator, Optimizer, Scheduler의 상태를 포함한 최신 체크포인트를 저장하여 학습 중단 후에도 재개가 가능하도록 구성하였다. 디퓨전 과정의 타임스텝 수는 $T=1,000$ 으로 설정하였고, β_t 는 $\beta_{start}=1 \times 10^{-4}$ 에서 $\beta_{end}=0.02$ 까지 선형적으로 증가하는 스케줄을 사용하였다.

입력 이미지는 64×64(애니메이션 얼굴) 해상도로 Resize한 뒤 RandomHorizontalFlip($p=0.5$)을 적용하고 $[-1, 1]$ 범위로 정규화하는 전처리 과정을 거쳤다. 전체 구현은 PyTorch 및 torchvision을 기반으로 하였으며, 단일 GPU(NVIDIA RTX 4070 Ti Super) 환경에서 학습을 수행하였다.

IV. 실험

4-1 실험 환경 및 데이터셋

제안하는 하이브리드 아키텍처의 성능 검증을 위해, 본 연구에서는 개량된 알고리즘을 2종의 컬러 데이터셋 환경에서 독립적으로 훈련 및 평가하였다. 베이스라인 성능 분석을 위한 일반 목적 데이터셋으로는 CIFAR-10(50,000장, 32×32)[15]을 사용하였으며, 본 연구의 주 표적인 고해상도 도메인 확장을 위해 Kaggle에서 제공하는 애니메이션 캐릭터 얼굴 데이터셋[16](63,565장, 64×64 해상도)을 활용하였다. CIFAR-10은 10개 클래스에 걸친 다양한 객체(동물, 차량 등)를 포함하는 다중 클래스 자연영상 데이터셋으로 클래스 간 형태·색상 편차가 크다. 반면 애니메이션 캐릭터 얼굴 데이터셋은 얼굴이라는 단일 도메인 내에서 구조적 유사성이 높은 이미지로 구성되어 있어, 두 데이터셋은 생성 난이도와 FID 값의 절대적 수준에서 근본적인 차이를 가진다. 모델 학습을 위한 최적화 알고리즘으로는 Adam[17]을 사용하였으며, 학습률 및 기타 하이퍼파라미터는 표 1과 같이 설정하였다. 또한 이미지 생성 과정에서는 DDIM 샘플링 기법[13]을 사용하여 추론을 수행하였다.

4-2 적대적 손실 가중치 분석

제안하는 하이브리드 아키텍처에서 적대적 손실 가중치 λ 는 DDPM의 복원 손실과 GAN 기반 적대적 손실 간의 균형을 결정하는 핵심 하이퍼파라미터이다. λ_{adv} 가 지나치게 작을 경우 적대적 손실의 영향이 충분히 반영되지 않아 고주파 텍스처 복원 효과가 제한될 수 있으며, 반대로 λ_{adv} 가 과도하게 클 경우 판별자의 영향력이 증가하여 학습 안정성이 저하될 수 있다. 따라서 본 연구에서는 데이터셋 특성에 따른 최적의 λ 값을 탐색하기 위하여 다양한 λ 설정에 대해 실험을 수행하였으며, 생성 성능은 FID를 기준으로 평가하였다.

표 2에서 확인할 수 있듯이 적대적 손실 가중치의 설정은 생성 성능에 직접적인 영향을 미쳤다. CIFAR-10 데이터셋에서는 $\lambda=0.001$ 일 때 가장 낮은 FID(206.20)를 기록하였으며, $\lambda=0.0005$ 에서는 FID가 206.20에서 208.92로 약 1.3%(절대 +2.72) 상승하였다. 반면 Anime Face 데이터셋에서는 $\lambda=0.0002$ 에서 가장 우수한 성능(FID=70.51)을 기록하였으며, $\lambda=0.0005$ 및 $\lambda=0.001$ 에서는 FID가 각각 81.33, 81.81로 증가하였다. 이는 λ 값에 따른 FID 변화 추이로 미루어 볼 때, 적대적 손실이 일정 수준까지는 세부 텍스처 복원에 기여하나 과도한 가중치는 학습 불안정성과 관련이 있을 것으로 추정된다. 또한 데이터셋마다 최적 λ 값이 상이하게 나타난 것은 이미지 복잡도와 데이터 분포 특성에 따라 적대적 손실의 영향력이 달라질 수 있음을 보여준다. 구체적으로, CIFAR-10과 같이 클래스가 다양하고 배경이 복잡한 데이터셋은 판별자가 학습해야 할 특징 공간이 넓어 적은 가중

표 1. 실험 설정

Table 1. Experimental setup

Parameter	Value
GPU	NVIDIA RTX 4070 Ti Super
Original DDPM Learning Rate	1×10^{-4}
Proposed Model Generator Learning rate	1×10^{-4}
Proposed Model Discriminator Learning rate	2×10^{-5}
Optimizer	Adam ($\beta_1=0.5$, $\beta_2=0.999$)
Batch Size	32
Epoch	80
EMA decay (β)	0.9999
Total Timesteps	1000
Time Embedding Dimension	128

표 2. 적대적 손실 가중치(λ)에 따른 FID 변화

Table 2. Effect of adversarial loss weight (λ) on FID

Dataset	λ	FID
CIFAR-10	0.001	206.2002
	0.0005	208.9215
Anime Face	0.001	81.8119
	0.0005	81.3289
	0.0002	70.5139

치만으로도 판별자 피드백이 과도하게 작용할 수 있는 반면, 애니메이션 캐릭터 얼굴 데이터셋은 단일 도메인의 정교한 디테일(눈, 머리카락 텍스처 등)을 포착하기 위해 상대적으로 낮은 가중치에서부터 민감하게 반응하는 것으로 판단된다. 따라서 적대적 손실 기반 하이브리드 디퓨전 모델에서는 데이터셋 특성에 맞는 λ 최적화 과정이 필수적임을 확인하였다.

4-3 정량적 성능 평가 및 대조 분석

제안 모델의 생성 품질을 정량적으로 평가하기 위해 생성형 모델의 표준 지표인 FID Score를 측정하였다. FID 수치가 낮을수록 원본 데이터셋의 실제 분포와 유사한 고품질의 이미지가 생성되었음을 의미한다. 복원 손실 함수($Loss_{MSE}$)만을 사용하는 오리지널 DDPM과, 선행 연구의 과도한 적대적 규제 조건, 그리고 본 연구에서 제안한 데이터셋별 정밀 가중치 제어 전략을 적용한 최종 모델의 성능 비교 결과는 아래 표 3과 같다.

실험 결과, 오리지널 DDPM은 데이터 분포의 거시적 형태

표 3. 데이터셋별 FID 성능 대조

Table 3. Comparison of FID scores across datasets

Model Architecture	CIFAR-10 (FID)	Anime Character Faces (FID)
Original DDPM (MSE Only)	211.7633	73.4550
Proposed Model	206.2002	70.5139

를 안정적으로 구조화하였으나 고주파 디테일 복원에는 한계를 보였다. 반면, 본 연구에서 제안한 하이브리드 모델은 미세 텍스처와 윤곽선 등의 고주파 성분을 정밀하게 포착함으로써 CIFAR-10에서는 FID가 211.76에서 206.20으로 감소하여 약 2.6% 개선되었다. 애니메이션 얼굴 데이터셋에서는 73.45에서 70.51로 감소하여 약 4.0% 개선되었다. 이는 적대적 가중치 조정이 FID 감소에 기여했음을 보여준다. 다만 본 절에서 제시한 FID 수치는 컴퓨팅 자원의 제약으로 인해 각 λ 설정당 단일 실행(single-run) 결과이며, λ 에 따른 경향성은 예비적(preliminary) 분석 결과로 해석될 필요가 있다. 반복 실험 기반의 통계적으로 엄밀한 검증은 향후 연구로 남긴다.

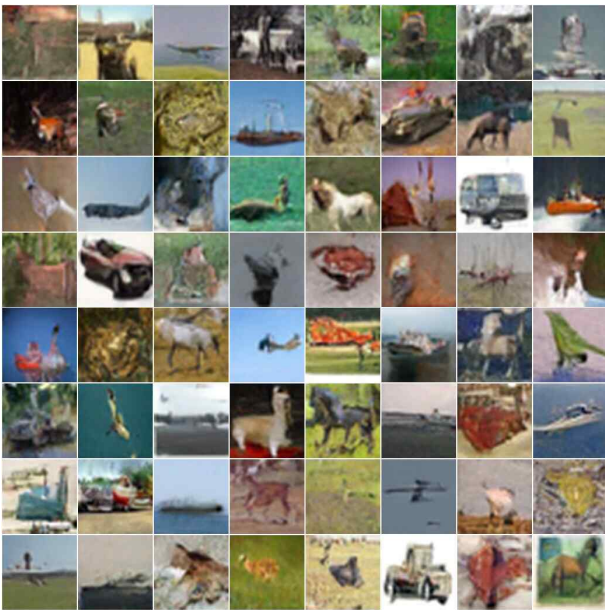


그림 3. 제안 하이브리드 모델 기반 CIFAR-10 이미지 생성 결과 (8×8 grid)

Fig. 3. Generation results of CIFAR-10 images using the proposed hybrid model (8×8 grid)

V. 고찰 및 분석

5-1 데이터셋별 적대적 가중치(λ)의 영향

표 2의 결과는 두 데이터셋에서 상반된 경향을 보인다. CIFAR-10에서는 $\lambda=0.001$ (FID 206.20)이 $\lambda=0.0005$ (FID 208.92)보다 우수하여 시험 구간에서 가중치가 클수록 성능이 향상되었다. 반면 애니메이션 얼굴에서는 λ 를 0.001에서 0.0002로 낮출수록 FID가 단조 감소하였다. 즉, 두 데이터셋 모두 시험 구간의 경계점에서 최저 FID가 관측되었으며 내부 극소점은 포착되지 않았다. 따라서 본 연구의 λ 값은 ‘시험 범위 내 최적값’으로 해석하는 것이 정확하다.



그림 4. 제안 하이브리드 모델 기반 애니메이션 캐릭터 얼굴 생성 결과(4×4 grid)

Fig. 4. Generation results of anime character faces using the proposed hybrid model (4×4 grid)

이러한 데이터셋 의존성은 판별자가 받는 학습 신호의 차이로 설명할 수 있다. CIFAR-10은 다중 클래스가 혼재되어 판별자의 식별 신호가 분산되므로 비교적 큰 가중치를 견딘다. 반면 단일 도메인(얼굴)인 애니메이션 데이터셋은 구조가 균질하여 판별자가 빠르게 강해지므로, 더 낮은 λ 에서 균형이 형성된 것으로 해석된다.

5-2 절대 FID 수준과 자원 제약의 해석

본 실험의 절대 FID 수치(CIFAR-10 약 206, 애니메이션 약 70)는 대규모 분산 학습이 적용된 최신 디퓨전 모델들의 보고치보다는 상대적으로 높다. 이는 단일 GPU(RTX 4070 Ti Super) 자원 환경 하에서 배치 크기 32, 80 Epoch의 제한된 학습 예산과 메모리 효율을 위한 16×16 초과 구간의 Self-Attention 생략 등 자원 절약형 아키텍처 설정에 기인한 것으로 판단된다. 따라서 본 연구의 정량적 수치는 타 최신 기법과의 절대 비교가 아닌, 동일한 제약 조건 하에서 오퍼리널 DDPM과 제안하는 하이브리드 모델을 비교하는 상대적 성능 개선의 관점에서 해석되어야 한다.

5-3 개선폭의 데이터셋 간 차이

제안 모델의 상대적 개선폭은 애니메이션 얼굴 도메인(4.0%)이 CIFAR-10 도메인(2.6%)보다 크게 나타났다(표 3). 적대적 손실이 고주파 텍스처 복원을 보강한다는 전제하에, 구조가 균질한 단일 도메인인 애니메이션 얼굴은 복원 대

상인 텍스처(머리카락, 눈동자 윤곽 등)가 일관되어 적대적 신호가 효과적으로 작용한 것으로 보인다. 반면 해상도가 낮고 다중 클래스가 혼재된 CIFAR-10은 복원 목표가 분산되어 개선 여지가 상대적으로 작았던 것으로 해석할 수 있다.

VI. 결론 및 향후 연구

본 연구에서는 다채널 컬러 이미지 생성 환경에서 기존 복원 손실 함수 기반 디퓨전 모델(DDPM)이 직면하는 고주파 디테일 영역의 선명도 저하 및 블러(Blurring) 현상을 극복하기 위해, 판별자의 적대적 피드백을 결합한 개량형 하이브리드 아키텍처 및 데이터셋별 적대적 가중치(λ) 최적화 전략을 제안하였다. 훈련 과정에서 다차원 컬러 공간 확장에 따라 발생하던 런타임 행렬 차원의 비대칭성과 디바이스 인덱싱 결함을 제어하는 텐서 정렬 메커니즘을 정립하여 학습의 연속성을 확보하였으며, 효율적인 하이브리드 목적 함수 설계를 통해 디퓨전 고유의 분포 수렴 성능과 GAN의 미세 텍스처 복원력을 양립시켰다. 다만 본 연구에서 제시한 적대적 손실과 텍스처 복원 간의 관계는 FID 수치 변화에 근거한 해석으로, 텍스처 복원 정도를 직접 측정할 수 있는 지표나 LPIPS 등의 정성적·정량적 근거는 포함하지 않았다. 이에 대한 시각적·정량적 검증은 향후 연구 과제로 남긴다.

정량적 실험 결과, 제안 모델은 일반 목적 컬러 CIFAR-10 데이터셋에서는 기존 DDPM의 FID 211.76 대비 206.20을 기록하여 생성 품질이 개선됨을 확인하였다. 아울러 구조적 복잡도가 높은 고해상도 애니메이션 캐릭터 얼굴 데이터셋 도메인 확장 실험에서도 기존 성능을 유의미하게 상회하는 최종 FID 70.5139를 달성함으로써, 시각적 왜곡(Artifact)이 제어된 정밀한 안면 이목구비 생성 품질 개선 효과를 실증적으로 규명하였다. 본 연구는 도메인 특성에 따른 차별화된 가중치 밸런싱이 생성 품질 향상에 직결됨을 입증했다는 점에서 학술적 의의를 지닌다.

다만, 본 실험은 CIFAR-10 및 특정 애니메이션 캐릭터 얼굴 데이터셋만을 대상으로 수행되었기에, 복잡도가 극대화된 임의의 대규모 자연 이미지(Natural Images) 데이터셋에 대한 범용적 검증은 전방위적으로 수행하지 못하였다는 한계를 지닌다. 또한 단일 GPU(NVIDIA RTX 4070 Ti Super) 자원 환경 하에서 하이퍼파라미터 탐색 및 다중 시드(multi-seed) 반복 학습을 제한적으로 진행하였으므로, 대규모 다중 노드 분산 학습 환경에서의 일반화 가능성과 함께 반복 실험을 통한 평균·표준편차 기반의 통계적 유의성 검증에 대해서도 추가적인 확인이 필요하다. 아울러 본 연구는 GAN 단독 성능과의 직접 비교보다 DDPM 대비 개선 효과 검증에 초점을 두었으며, GAN 단독 모델과의 정량적 비교는 향후 과제로 남긴다. 또한 적대적 손실 가중치(λ)의 탐색은 시험된 범위(0.0002~0.001) 내에서 데이터셋별로 상이한 경향이 관찰된 결과이며, 보다 세밀하고 넓은 범위의 탐색은 향후 과제로 남긴다.

감사의 글

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(RS-2022-NR069955 (2022R1A2C1092178)).

참고문헌

- [1] J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," in *Proceedings of Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, Vancouver, Canada, pp. 6840-6851, December 2020.
- [2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, ... and Y. Bengio, "Generative Adversarial Nets," in *Proceedings of Advances in Neural Information Processing Systems 27 (NeurIPS 2014)*, Montreal, Canada, pp. 2672-2680, December 2014.
- [3] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," in *Proceedings of International Conference on Learning Representations (ICLR)*, Banff, Canada, April 2014.
- [4] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans: LA, pp. 10684-10695, Jun. 2022.
- [5] J. Ho and T. Salimans, "Classifier-Free Diffusion Guidance," in *Proceedings of NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, December 2021.
- [6] Y. Song, P. Dhariwal, M. Chen, and I. Sutskever, "Consistency Models," in *Proceedings of the 40th International Conference on Machine Learning (ICML)*, Honolulu: HI, pp. 32211-32252, July 2023.
- [7] J. H. Lee and W. Choi, "Improving MNIST Image Generation Quality by Combining Denoising Diffusion Probabilistic Models (DDPM) with Adversarial Learning," in *Proceedings of the 2026 International Conference on Next-generation Convergence Technology (ICNCT 2026)*, Macau, pp. NC26-046, January 2026.
- [8] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium," in *Proceedings of Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, Long Beach: CA, pp. 6629-6640, December 2017.
- [9] A. Sauer, D. Lorenz, A. Blattmann, and R. Rombach,

“Adversarial Diffusion Distillation,” in *Proceedings of the 18th European Conference on Computer Vision (ECCV)*, Milan, Italy, pp. 87-103, October 2024.

- [10] Z. Xiao, K. Kreis, and A. Vahdat, “Tackling the Generative Learning Trilemma with Denoising Diffusion GANs,” in *Proceedings of International Conference on Learning Representations (ICLR)*, April 2022.
- [11] Z. Wang, H. Zheng, P. He, W. Chen, and M. Zhou, “Diffusion-GAN: Training GANs with Diffusion,” in *Proceedings of International Conference on Learning Representations (ICLR)*, Gothenburg, Sweden, May 2023.
- [12] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” in *Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Munich, Germany, pp. 234-241, October 2015.
- [13] J. Song, C. Meng, and S. Ermon, “Denoising Diffusion Implicit Models,” in *Proceedings of International Conference on Learning Representations (ICLR)*, May 2021.
- [14] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, “Spectral Normalization for Generative Adversarial Networks,” in *Proceedings of International Conference on Learning Representations (ICLR)*, Vancouver, Canada, April 2018.
- [15] A. Krizhevsky, Learning Multiple Layers of Features from Tiny Images, University of Toronto, Toronto, Canada, Technical Report TR-2009, 2009.
- [16] S. Churchill. Anime Face Dataset [Internet]. Available: <https://www.kaggle.com/datasets/splcher/animefacedataset>.
- [17] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” in *Proceedings of International Conference on Learning Representations (ICLR)*, San Diego: CA, May 2015.

이중현 (Jung-Hyeon Lee)



2020년 ~ 현 재: 강남대학교 ICT융합공학부 학사과정
 ※ 관심분야: 생성형 AI, 컴퓨터 비전

최웅 (Woong Choi)



2000년 : 조선대학교(공학석사)
 2005년 : Tokyo Institute of
 Technology(공학박사-지능시
 스템과학)

2005년 ~ 2010년: Ritsumeikan University
 2010년 ~ 2022년: National Institute of Technology, Gunma
 College
 2022년 ~ 현 재: 강남대학교 컴퓨터공학부 교수
 ※ 관심분야: 생성형 AI, 컴퓨터 비전