석사학위논문

단일 참조 이미지 기반 확산 모델의 효율적인 영상 생성 방법론

2025년

한 성 대 학 교 대 학 원

A I 응 용 학 과

A I 응 용 학 전 공

김 지 수

HANSUNG
UNIVERSITY

석 사 학 위 논 문
지도교수 오희석

단일 참조 이미지 기반 확산 모델의 효율적인 영상 생성 방법론

Efficient Image Synthesis Method Using Single Reference Image-Based Diffusion Model

2024년 12월 일

한 성 대 학 교 대 학 원

A I 응 용 학 과

A I 응 용 학 전 공

김 지 수

석 사 학 위 논 문
지도교수 오희석

단일 참조 이미지 기반 확산 모델의 효율적인 영상 생성 방법론

Efficient Image Synthesis Method Using Single Reference Image-Based Diffusion Model

위 논문을 공학 석사학위 논문으로 제출함

2024년 12월 일

한 성 대 학 교 대 학 원

A I 응 용 학 과

A I 응 용 학 전 공

김 지 수

김지수의 공학 석사학위 논문을 인준함

2024년 12월 일

심사위원장 김 명 선 (인)

심 사 위 원 오 희 석 (인)

심 사 위 원 이 웅 희 (인)

국 문 초 록

단일 참조 이미지 기반 확산 모델의 효율적인 영상 생성 방법론

한 성 대 학 교 대 학 원
A I 응 용 학 과
A I 응 용 학 전 공
김 지 수

최근 이미지 생성 연구는 표현 학습의 발전에 힘입어 놀라운 성장을 이루었으나, 대부분의 혁신적인 모델들은 대량의 학습 데이터에 크게 의존한다는 한계가 있다. 본 연구는 단일 참조 이미지만으로 고품질의 다양한 이미지를 생성할 수 있는 새로운 확산 모델을 제안한다. 제안된 방법은 웨이블릿 도메인에서의 주파수 분해와 계층적 스케일 구조를 결합하여, 기존 방식들이 가진 아티팩트 누적 문제를 해결하였다. 특히, 제한된 수용 영역을 갖는 U-Net 아키텍처를 도입하여 전역적 정보의 과적합을 방지하고, 웨이블릿 변환을 통해 주파수 도메인에서의 효과적인 특징 학습을 가능하게 하였다. 실험을 통해 제안된 모델이 기존 단일 이미지 생성 방법들과 비교하여 우수한 시각적 품질과 다양성을 보임을 확인하였으며, 이미지 조화와 같은 실제 응용에서도 효과적으로 활용될 수 있음을 검증하였다. 또한 제안된 모델은 기존 확산 모델의 긴 샘플링 시간을 대폭 단축하여 계산 효율성을 개선하였으며, 임의의 해상도에서도 안정적인 이미지 생성이 가능한 장점을 보였다. 본 연구는 제한된

데이터 환경에서의 이미지 생성 문제에 대한 새로운 해결책을 제시하며, 다양한 실용적 응용 가능성을 보여준다. 향후 연구에서는 본 방법론을 확장하여 더 빠른 추론 속도와 향상된 이미지 품질을 달성하는 것을 목표로 한다.

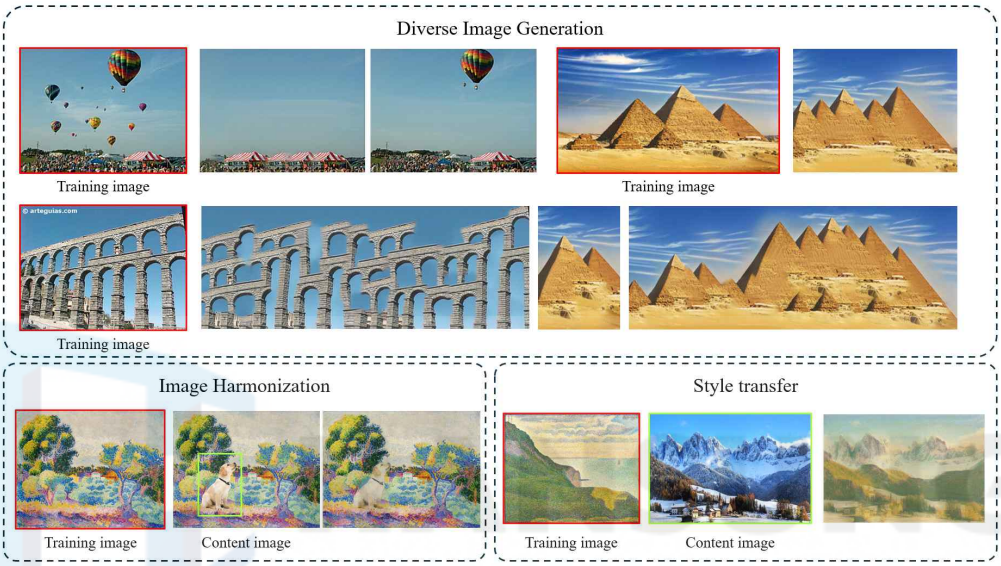
【주요어】 확산 모델, 단일 이미지 생성, 웨이블릿 변환

목 차

표 목 차

그 림 목 차

Ⅰ. 서 론



[그림 1-1] 제안 모델의 이미지 생성 및 조작 결과. 본 연구는 단일 이미지 학습을 통해 해상도 제약없이 사실적이고 다양한 이미지를 생성할 수 있으며, 다양한 하위 응용 분야에 적용 가능하다

최근 이미지 생성 연구는 이미지 다양성과 충실도를 효과적으로 균형 잡는 표현 학습(representation learning)의 발전에 힘입어 놀라운 성장을 이루었다. 특히 확산 모델(Diffusion model)[1]의 등장으로, 이미지 생성 모델들은 사용자의 의도를 반영하면서도 고수준의 사실적인 이미지를 생성하는 데 획기적인 성과를 거두었다.

더불어 이미지 생성 응용 분야의 다양화로 고품질 이미지 생성뿐만 아니라 임의의 이미지 크기를 다룰 수 있게 되었으며 이미지 복원, 완성[2] 등 다양한 응용 분야에서도 주목할 만한 수준의 적용이 가능해졌다. 심지어 텍스트, 마스크, 깊이 등 복잡한 프롬프트를 조건으로 이미지를 생성할

수 있게 되었지만[3], 이러한 혁신적인 프로세스들은 대부분 대량의 학습 데이터에 크게 의존한다. 그러니 실무 환경에서는 양질의 학습 데이터를 충분히 확보하는 것이 현실적인 제약으로 작용한다[4].

이를 해결하기 위해 최근 연구들은 다양한 생성 모델을 제안했다. 대규모 데이터셋 학습과는 달리, 단일 이미지의 지역적 통계를 포착하여 다양한 샘플을 생성하는 단일 이미지 생성이라는 새로운 분야가 등장했다. 이 접근법은 더 많은 학습 데이터 없이도 이미지 생성이 가능하며, 단일 이미지만으로 학습하여 복잡한 구조와 텍스처를 효과적으로 처리할 수 있다. 결과적으로 초기 학습된 이미지의 스타일을 계승하면서 다양한 이미지를 생성할 수 있게 되었다. 하지만 기존의 단일 이미지 생성 방식은 다중 스케일 방식으로 이미지를 점진적으로 생성하는데, 이는 하위 스케일에서 축적된 세부 디테일의 붕괴로 인해 최종 이미지에서 눈에 띄는 특징적인 아티팩트가 발생하는 단점이 있다.

우리는 다중 스케일 구조를 사용하면서도 단일 확산 모델을 활용하여 아티팩트 오류없이 단일 이미지 학습을 하는 연구를 진행하였다. 또한 다중 스케일 구조에서 하위 스케일(대략적인 정보)에서 상위 스케일(세밀한 정보)로 전이되는 과정 사이에 웨이블릿 도메인 서브밴드를 활용하여 정보의 연속성을 보존하는 체계적인 접근 방식을 제안한다.

제안된 방법에서는 학습 이미지를 공간 및 스펙트럴 특성을 포함하는 서브밴드로 분해하여, 주파수 영역에서의 변환된 도메인을 기반으로 전방향 및 역방향 확산 프로세스를 구축한다. 웨이블릿 서브밴드의 잠재 공간 조작을 통해, 각 스케일에서 독립적으로 정의된 전방향 확산 프로세스는 이미지 생성 단계에서 해당 주파수 특성을 반영하는 일관된 궤적으로 수렴된다.

아울러 본 연구에서는 제한된 수용 영역을 갖는 U-Net[5] 아키텍처를 도입하여 전역적 정보의 과적합을 방지하였으며, 입력 이미지에 대한 무작위 크롭 전략을 적용함으로써 생성 결과의 다양성과 품질을 제고하였다. 이러한 설계는 모델의 연산 효율성 또한 개선하는 효과를 수반하였다.

Ⅱ. 연구 배경

2.1 단일 이미지 생성

단일 이미지 생성 연구는 이미지 내 패치 분포나 학습 이미지의 고유 특징을 파악하여 다양한 기법을 통해 새로운 이미지를 생성하는 데 초점을 맞추고 있다. 선구적인 연구인 SinGAN[6]은 계층적 생성 적대 신경망(Generative Adversarial Network)[11]을 도입하여 다운샘플링된 이미지로부터 단계적으로 세부 사항을 생성하는 방식이며 여러 개의 생성자를 사용한다. SinGAN은 다중 스케일 구조를 통해 이미지의 전역적 구조와 지역적 텍스처를 모두 학습할 수 있으며, 각 스케일에서 독립적인 생성자와 판별자를 활용하여 이미지 생성의 세밀도를 점진적으로 향상시킨다. 이를 통해 단일 학습 이미지로부터 고품질의 다양한 샘플을 생성할 수 있게 되었다. 그러나 하위 스케일에서 축적된 세부 디테일의 붕괴가 상위 스케일까지 이어지면 최종 이미지에서 눈에 띄는 특징적인 아티팩트가 발생하는 단점이 있다.

ConSinGAN[7]은 SinGAN의 학습 과정을 개선하여 여러 단계를 순차적으로 동시에 학습할 수 있도록 발전시켰다. 특히 ConSinGAN은 학습 시간을 크게 단축시켰으며, 더 적은 수의 스테이지를 사용하면서도 유사하거나 더 나은 품질의 결과물을 생성할 수 있게 되었다. InGAN[8]은 각 자연 이미지가 고유한 패치 분포를 가진다는 가정 하에 단일 이미지에서 학습을 진행하여 다양한 크기의 새로운 자연 이미지를 생성할 수 있게 되었다. InGAN의 주요 특징은 기하학적 변환에 불변하는 특성을 가지도록 설계되었다는 점이며, 이를 통해 입력 이미지의 내부 구조를 유지하면서 다양한 변형이 가능하게 되었다. 하지만 자연 이미지에 국한하여 이미지를 효과적으로 생성해낸다는 문제점이 있었고, 인공적인 패턴이나 구조적 이미지에 대해서는 성능이 제한적이었다.

GPNN[9]은 기존의 패치 기반 접근 방식을 수정하여 패치 통계의 양방향

유사성을 강화했고, 이를 통해 단일 자연 이미지로부터 고품질의 시각적 결과물을 합성할 수 있게 되었다.

2.2 확산 모델

확산 모델은 데이터에 점진적으로 노이즈를 추가하고 제거하는 과정을 통해 새로운 데이터를 생성하는 생성 모델이다. 이는 노이즈가 충분히 적을 때, 노이즈를 더하는 과정의 역방향을 신경망으로 복원할 수 있다는 원리를 기반으로 한다. 확산 모델의 프로세스는 크게 두 가지로 구분된다.

2.2.1 순방향 과정(Forward process)

순방향 과정은 확산 과정으로 타임스텝 $t \in \{0,1,...,T\}$에서 분산 스케줄 $\{\beta_1,...,\beta_T\}$에 따라 원본 이미지에 단계적으로 가우시안 노이즈인 ϵ를 추가하는 마르코프 연쇄에 따라 사후 확률분포 $q(x_1,...,x_T)$에 대한 수식은 다음과 같이 계산된다.

$$q(x_t|x_{t-1}) = N(x_t;\sqrt{1-\beta_t}x_{t-1}, \beta_t I), \quad (1)$$

여기서 I는 단위행렬이고, N은 가우시안 분포를 나타내며 β_t는 타임스텝에 대해 사전 정의된 작은 상수이다. 이 지속적인 노이즈 추가를 통해 최종적으로는 이미지가 순수한 가우시안 노이즈 상태가 된다. 이 과정은 데이터의 분포를 점진적으로 단순화하는 역할을 한다.

2.2.2 역방향 과정(Reverse process)

실질적인 데이터 생성이 이루어지는 역방향 과정에서는 순수한 가우시안 노이즈에서 시작하여 점진적으로 노이즈를 제거하면서 최종 이미지를

생성한다. 일반적으로 약 1000회의 반복적인 노이즈 제거 과정을 거쳐 깨끗한 출력 이미지를 얻을 수 있다. $\alpha_t := 1-\beta_t$, $\overline{\alpha_t} = \Pi_{s=1}^{t}\alpha_s$로 t시점의 노이즈한 이미지 x_t를 다음과 같은 수식으로 추정할 수 있다.

$$q(x_t|x_0) = N\left(x_t; \sqrt{\overline{\alpha_t}}\,x_0, (1-\overline{\alpha_t})I\right), \quad (2)$$

$$x_t = \sqrt{\overline{\alpha_t}}\,x_0 + \sqrt{1-\overline{\alpha_t}}\,\epsilon,\ \epsilon \sim N(0, I). \quad (3)$$

확산 모델의 구현은 U-Net 아키텍처를 기반으로 이루어진다. 학습 과정에서는 학습 데이터셋에 가우시안 노이즈를 추가한 이미지를 U-Net 기반 모델에 입력하고, L2 손실 함수를 사용하여 추가된 노이즈를 예측하도록 학습한다. 이를 통해 U-Net이 효과적으로 노이즈를 식별하고 제거할 수 있게 된다.

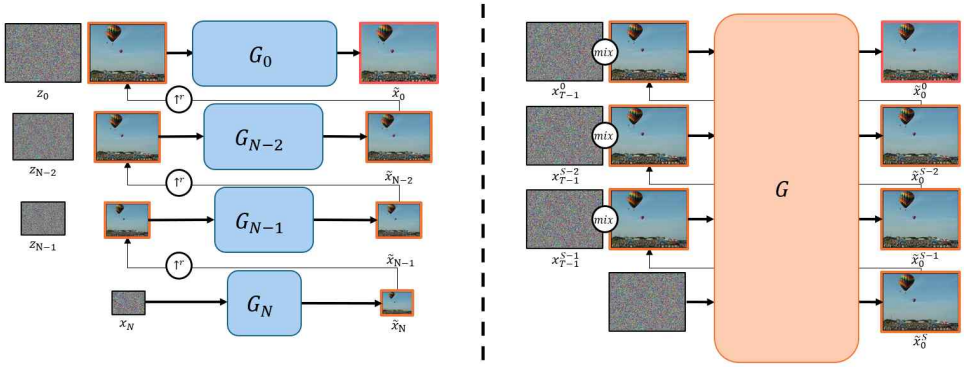
확산 모델은 다양한 이미지 처리 및 생성 작업에서 성공적으로 적용되고 있으며 특히 Guided-diffusion[12]의 등장으로 조건부 생성이 가능해졌으며 이는 노이즈 이미지 분류기의 그래디언트를 활용하여 조건에 맞는 이미지를 생성하는 방식으로 구현된다. 확산 모델은 실제 데이터와 유사한 고품질의 데이터 생성이 가능하며 안정적인 학습 과정을 동반한다는 장점을 가지고 있다. 그러나 U-Net 모델 전체를 활용하여 약 1000회의 노이즈 제거 과정을 거치기 때문에 다른 생성 모델에 비해 상대적으로 느린 샘플링 속도를 지니고, 학습을 위해서는 다른 생성 기법보다 높은 컴퓨팅 리소스를 요구한다는 단점이 있다.

Ⅲ. 본 론

3.1 계층적 네트워크와 수용영역 축소 모델

SinGAN은 이미지 피라미드를 구축하고 이 계층적 네트워크를 통해 여러 스케일에서 점진적으로 해상도를 증가시키며 GAN을 학습한다. 이 때, 하위의 작은 스케일에서 발생한 세부적인 오류가 해상도가 증가하면서 누적되어 최종 결과물에 특징적인 아티팩트를 발생시킨다[10].

다중 스케일을 활용하는 주된 이유는 하위 스케일의 이미지가 상위 스케일의 이미지보다 저주파 특성을 가지고 있어 상대적으로 더 다양한 이미지 생성이 가능하기 때문이다[13]. 이러한 특성을 활용하면 하위 스케일에서는 다양성 있는 이미지를 생성하고, 상위 스케일로 진행하면서 점진적으로 세부 사항을 추가하여 고품질의 이미지를 생성할 수 있다. 특히 확산 모델을 활용할 경우, 다중 스텝(타임스텝)에 따라 하위 스케일의 저주파 이미지를 상위 스케일에서 점진적으로 처리함으로써 오류의 누적 없이 효과적인 이미지 생성이 가능하다. 그래서 본 논문은 다양성 있는 이미지 생성을 하면서도 고해상도의 이미지를 생성하기 위해 SinGAN에서의 계층적 네트워크를 활용하면서 모든 계층에서 동일한 사이즈이지만 다른 주파수의 이미지를 활용하여 품질적으로 보증된 이미지를 생성할 수 있게 하였다. 또한 단일 확산 모델이 학습되어 오류의 축적을 방지한다. 저주파 이미지에서부터 고주파 이미지까지의 확산 과정을 다루는 구체적인 학습 과정은 다음 절에서 다룬다.



(a) SinGAN Multiple model Training process (b) Proposed method Single model Training process

[그림 3-1] 제안하는 모델과 SinGAN의 생성 프레임워크 비교. $\uparrow^r$은 업샘플링 연산을 의미하고 mix는 하위 스케일과 현재 스케일과의 이미지 보간을 의미한다

그림 3-1에서는 본 논문에서 제안하는 방법의 학습 과정을 SinGAN과 비교하였다. SinGAN은 단일 이미지로부터 다양성 있는 이미지들을 생성해내기 위해 이미지 피라미드를 구축하고 이 계층적 네트워크를 통해 여러 스케일에서 점진적으로 해상도를 증가시키며 GAN을 학습한다. 각 스케일 n에서 이전 스케일의 이미지 $\tilde{x}_{n+1}$을 업샘플링하여 노이즈 z_n와 함께 PatchGAN에 입력 된다.

$$\tilde{x}_n = G_N(\alpha_n(\tilde{x}_{n+1})\uparrow^r + (1-\alpha_n)z_n). \quad (4)$$

α_n은 스케일 n이 감소할수록 같이 감소하는 보간값으로 0과 1사이의 값을 가진다. $\uparrow$는 업샘플링 연산을 나타내며 G_N은 n번째 스케일의 생성자를 의미한다. 이렇게 될 경우 작은 스케일에서 발생한 세부 오류가 해상도가 증가하면서 누적된다. 그러나 확산 모델을 활용하면 다중 스텝 과정에서 생성 프로세스를 수행하기 때문에 중간에서 생기는 아티팩트들을 자연스레 수정하며 더 높은 품질을 달성할 수 있게 된다. 이 때, 이미지 피라미드의 가장 하위 스케일인 S에서의 t시점의 이미지는 다음과 같이 나타낼 수 있다.

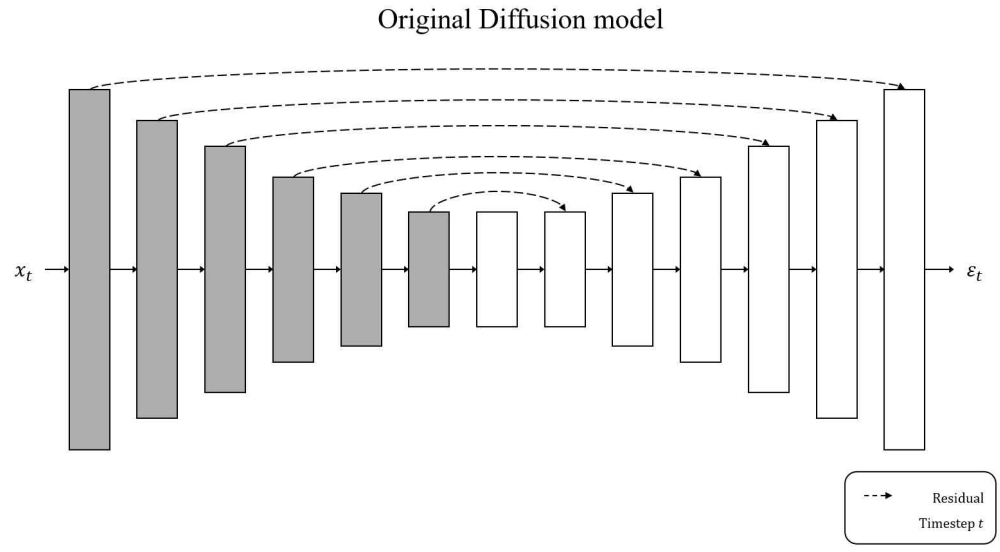
$$\tilde{x}_t^{s,mix} = \tilde{x}_t^s. \quad (5)$$

그리고 하위 스케일에서 생성된 이미지를 상위 스케일인 $s = S-1$의 t시점에서 보간할 때 다음과 같이 수식할 수 있다.

$$\tilde{x}_t^{s,mix} = \left(\gamma_t^s \tilde{x}_0^s\right) + \left(1-\gamma_t^s\right)x_t^s, \quad (6)$$
$$\tilde{x}_t^s = G\left(\tilde{x}_t^{s,mix}\right). \quad (7)$$

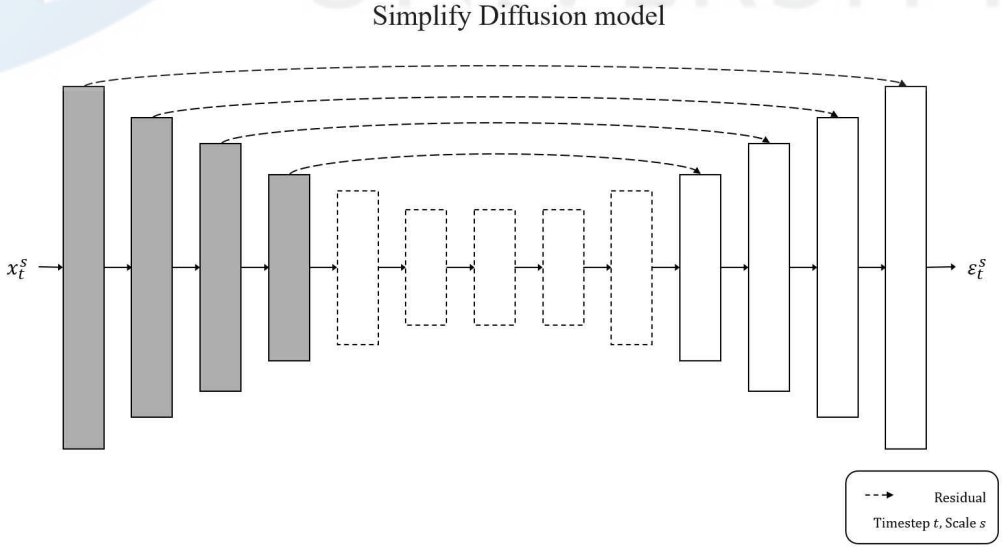
이렇게 하위 스케일에서 생성된 이미지를 상위 스케일인 s에서 γ만큼 보간하여 믹스하게 된다. 이 때, γ은 0부터 0.55까지 단조적으로 증가하는 상수값으로 타임스텝 t와 스케일 s에 따른 보간 계수이며 G는 확산 모델의 디노이징 네트워크이다. 결론적으로 하위 스케일의 생성 이미지와 그 바로 상위 스케일의 $T-1$시점의 노이즈가 보간된 이미지를 해당 스케일의 확산 모델 역과정 입력이 되게 된다. 이렇게 하게 되면 가장 하위 스케일에서는 전반적인 이미지의 구조들을 생성하게 되고 상위 스케일로 갈수록 세밀한 디테일들을 추가하며 최종 이미지를 생성하게 된다.

여기에 우리는 전 이미지 피라미드에 걸쳐 동일한 사이즈의 이미지를 사용했기에 다양성 있는 이미지를 생성하기 위해서는 추가적인 방안이 필요하였다. 확산 모델의 수용 영역(receptive filed)이 작을수록 더 다양한 이미지를 생성하게 된다. 하지만 수용 영역이 지나치게 작을 경우, 세부적인 패치 분포를 학습하게 되어 생성된 이미지들이 학습 이미지의 내부 구조를 보존하지 못하게 된다[10]. 그렇기에 본 연구는 확산 모델에서 각 스케일에서 이미지의 수용영역을 축소시키며 내부 구조를 잘 보존하기 위해 U-Net 아키텍처를 수정하였다.



[그림 3-2] 기본 확산 모델에서 사용하는 U-Net 모델의 아키텍처

그림은 기본 확산 모델에서 사용하는 U-Net의 아키텍처이다. 기본 확산 모델의 수용영역은 학습 이미지 전체를 포함하게 되는데 이는 학습 이미지와 동일한 이미지만을 생성하게 되는 결과를 초래한다.

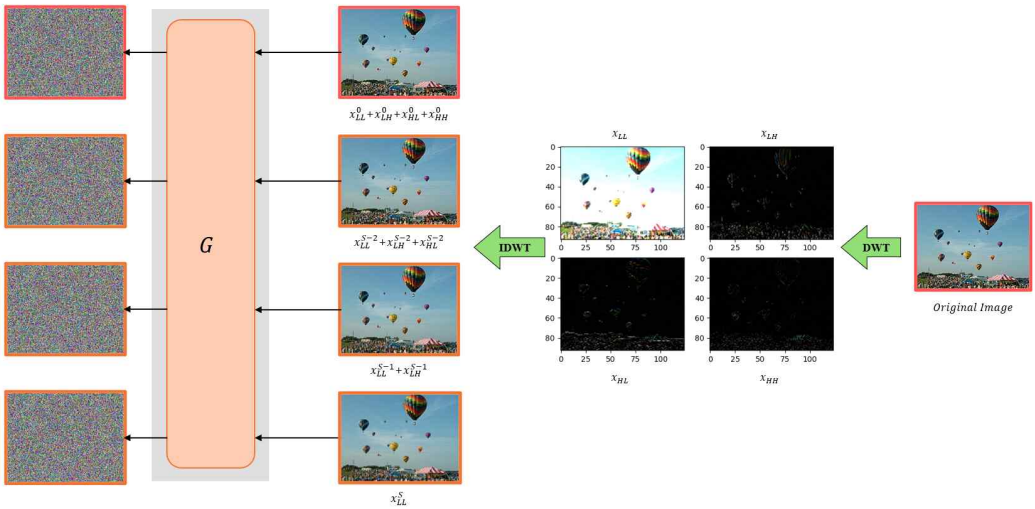


[그림 3-3] 제안하는 모델에서 사용하는 U-Net 모델의 아키텍처

반면 제안된 단순 확산 모델은 U-Net의 다운샘플링과 업샘플링 연산을 줄임으로써 U-Net 네트워크의 깊이를 감소시켰는데 이는 기본 확산 모델의 수용 영역이 이미지 전체를 커버하는 것을 방지하였다. 나아가 기본 확산 모델의 깊은 층에서 사용되던 어텐션 층이 자연스럽게 제거 되었다. 이로 인해 임의의 이미지 사이즈 생성에 적용 가능한 완전 합성곱 네트워크가 되었으며 동시에 모델이 가벼워짐에 따라 연산 효율성이 향상되었다.

3.2 웨이블릿 변환 기반 학습

기존 단일 이미지 생성 연구들은 축소 이미지를 점진적으로 업샘플링하는 설계 방법들이 많다. 바로 이러한 방법론이 하위 스케일에서의 오류 아티팩트가 최종 이미지에 영향을 미치는 설계 방법이 된다. 각 스케일에서 점진적으로 고주파 성분이 보증되어야 하위 스케일에서 오류 아티팩트가 확산 모델의 역과정을 거치며 수정된다. 본 연구에서 제안하는 학습 알고리즘은 크게 학습 과정(순방향)과 샘플링 과정(역방향)으로 구성되는데 학습 과정에서 업샘플링한 이미지를 활용하지 않고 웨이블릿 변환을 통한 서브 밴드인 LL, LH, HL, HH를 활용하여 각 스케일마다 점진적으로 활용하였다. LL, LH, HL, HH는 각각 수평과 수직 방향에 대한 주파수 성분을 의미한다. LL은 수평과 수직 방향이 모두에서 저주파 성분을 포함하며 이미지의 전반적인 구조와 대략적인 표현을 잘 포착한다. HH는 양방향으로 고역 통과 필터를 적용하여 획득한 성분으로 대각선 방향의 에지와 고주파 세부 정보를 강조한다.



[그림 3-4] 제안하는 모델의 확산 과정을 활용한 다중 스케일 구조. G는 확산 모델을 나타내며 DWT는 이산 웨이블릿 변환 $IDWT$는 이산 웨이블릿 역변환을 의미한다. x_{LL}, x_{LH}, x_{HL}, x_{HH}는 각각 입력 이미지 x에 대한 수평-수직 저주파, 수평 고주파-수직 고주파, 수평 저주파-수직 고주파, 대각선 고주파 성분을 나타낸다

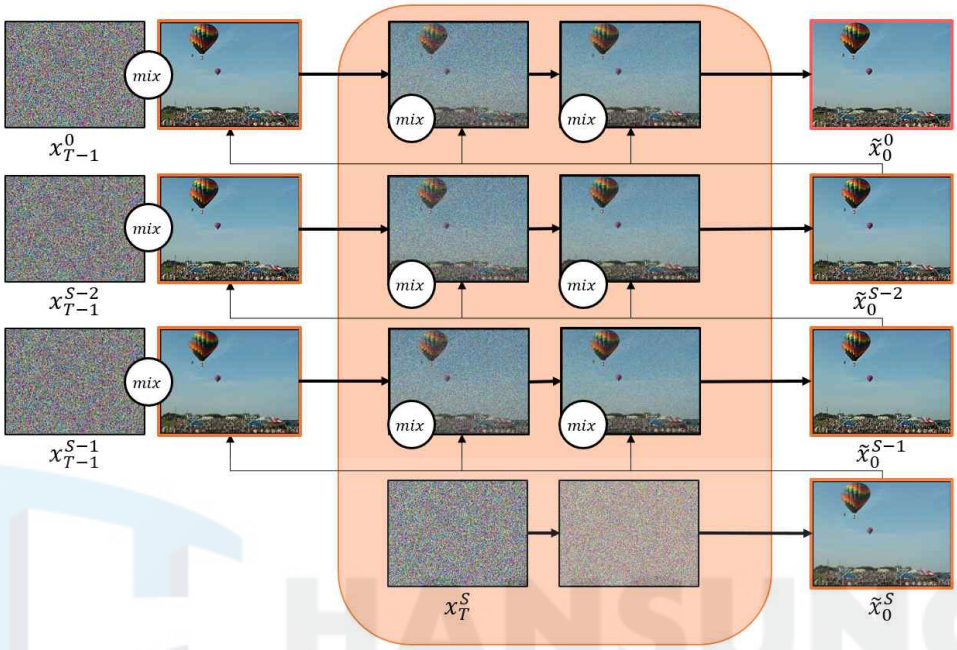
Algorithm 1 Training process

repeat
  $s \sim \text{Uniform}(\{S, S-1, \ldots, 0\})$
  $t \sim \text{Uniform}(\{0, \ldots, T\})$
  $\epsilon \sim \mathcal{N}(0, I)$
  $x_{ll}, x_{hl}, x_{lh}, x_{hh} = \text{DWT}(x)$ ▷ $x \in \mathbb{R}^{3\times H\times W}$
  if $s = S$ **then**
    $\widetilde{x}_0^s = \text{IDWT}(x_{ll})$ ▷ $\widetilde{x}_0^s \in \mathbb{R}^{3\times H\times W}$
  else if $s = S-1$ **then**
    $\widetilde{x}_0^s = \text{IDWT}(x_{ll} + x_{lh})$ ▷ $\widetilde{x}_0^s \in \mathbb{R}^{3\times H\times W}$
  else if $s = S-2$ **then**
    $\widetilde{x}_0^s = \text{IDWT}(x_{ll} + x_{lh} + x_{hl})$ ▷ $\widetilde{x}_0^s \in \mathbb{R}^{3\times H\times W}$
  else
    $\widetilde{x}_0^s = x$ ▷ Original image for remaining scales
  end if
  $\widetilde{x}_t^s = \sqrt{\bar{\alpha}_t}\widetilde{x}_0^s + \sqrt{1-\bar{\alpha}_t}\epsilon$
  Update model ϵ_θ by taking a gradient descent step on:
  $\nabla_\theta \|\epsilon - \epsilon_\theta(\widetilde{x}_t^s, t, s)\|_1$
until converged

[표 3-1] 제안하는 모델의 학습 알고리즘. s는 스케일, t는 타임스텝, ϵ은 가우시안 노이즈, x는 입력 이미지, θ는 모델 파라미터를 의미하며 ∇_θ는 θ에 대한 그래디언트를 나타낸다

그림3-4는 본 연구에서 제안하는 확산 모델의 순방향 과정이다. 그림에서 보여지듯이 순방향 과정은 동일한 사이즈의 다른 해상도로 이루어진 이미지 피라미드를 구성하는 것으로 시작된다. 순방향 과정의 학습 단계는 동일 차원의 동일 해상도 이미지 피라미드 구축으로 시작되며, 최하위 스케일의 다운샘플링된 이미지는 이산 웨이블릿 변환(Discrete-Wavelet Transform)을 통해 저주파 정보만을 포함하는 LL 성분으로 대체된다. 스케일 s는 가장 큰 값 S부터 시작하여 0까지 단계적으로 감소하며 각 스케일마다 타임스텝 t가 T부터 0까지 변화한다. 학습 시 가우시안 노이즈 ϵ이 이미지에 점진적으로 추가되며 각 스케일에서 초기 이미지 $\tilde{x}_0^s$로 표기된다. 이러한 웨이블릿 도메인에서의 주파수 분해와 업샘플링 프로세스의 생략은 이미지 리사이징 과정에서 발생하는 정보 손실을 최소화하는 효과를 가진다. 상위 스케일로의 전이 과정에서는 저주파-고주파, 고주파-저

주파 성분이 순차적으로 통합되어 고주파 특성의 점진적 보존이 이루어진다.



[그림 3-5] 제안하는 모델의 이미지 생성을 하는 샘플링 과정. x_T는 각 스케일에서의 초기 노이즈를 나타내며 x_0는 생성된 최종 이미지를 의미한다. mix는 하위 스케일에서 생성된 이미지와 현재 스케일의 이미지를 결합하는 과정이다. 각 스케일에서 점진적인 노이즈 제거를 통해 이미지를 생성하며, 하위 스케일의 결과가 상위 스케일의 생성 과정에 반영된다

하위 스케일에서 생성된 다양성 이미지의 구조적 다양성을 보존하기 위해, 매 타임스텝에서 하위 스케일의 생성된 이미지를 상위 스케일인 s에서 γ 비율로 보간한 후, 상위 스케일의 특성과 결합하는 프로세스를 적용한다.

Algorithm 2 Sampling process

for all $s = S, S-1, \ldots, 0$ **do**
  $x_T^s \sim \mathcal{N}(0, I)$
  for $t = T, \ldots, 1$ **do**
    $\epsilon_\theta = \epsilon_\theta(\widetilde{x}_t^s, t, s)$
    $\widetilde{x}_t^s = \frac{1}{\sqrt{\alpha_t}}(x_t^s - \sqrt{1-\alpha_t}\epsilon_\theta(x_t^s, t, s))$ ▷ Predicted x_0
    if $s = S$ **then**
      $\widetilde{x}_t^s = \widetilde{x}_t^s$
    else
      $\widetilde{x}_t^{s,mix} = \gamma_t^s \widetilde{x}_0^{s+1} + (1-\gamma_t^s)\widetilde{x}_t^s$
    end if
    $\widetilde{x}_{t-1}^s = \sqrt{\alpha_{t-1}} f_\theta^{(t)}(\widetilde{x}_t^{s,mix}) + \sqrt{1-\alpha_{t-1}}\epsilon_\theta(\widetilde{x}_t^s, t, s)$ ▷ DDIM update
  end for
end for

[표 3-2] 제안하는 모델의 생성 알고리즘

타임스텝 T에서의 랜덤 노이즈는 점진적 디노이징 프로세스를 통해 $t=0$에서 정제된 이미지로 수렴되며 DDIM(Denoising Diffusion Implicit Model)[14]의 업데이트 규칙에 따라 스케줄링 파라미터가 업데이트 된다. 본 과정은 최하위 스케일 S에서 저주파 컴포넌트의 생성을 통해 이미지의 글로벌 구조가 확립된다. 상위 스케일에서는 가우시안 노이즈와 하위 스케일에서 생성된 저주파 정보의 결합을 통해 역확산 과정이 수행되며, 이러한 하위 스케일 정보의 통합은 각 타임스텝에서 반복적으로 이루어진다. 본 연구에서 제안하는 계층적 스케일 아키텍처는 하위 스케일에서 저주파 특성을, 상위 스케일에서 고주파 특성을 점진적으로 합성하는 방식을 채택하고 있다. 이는 하위 레벨에서 시맨틱 구조의 학습을, 상위 레벨에서 세부적 특징의 학습을 가능하게 함으로써, 생성 이미지의 다양성과 품질을 동시에 최적화할 수 있다. 또한, 기존의 업샘플링 과정에서 발생하는 정보 손실과 그에 따른 품질 저하 문제를 해결하기 위해, 웨이블릿 도메인에서의 주파수 분해를 통한 업샘플링 프로세스의 생략을 제안한다. 이를 통해 매크로 구조와 마이크로 디테일이 조화롭게 보존된 고품질 이미지의 생성이 가능하다.

Ⅳ. 실 험

4.1 데이터셋과 실험 설정

본 연구의 실험은 10개의 단일 이미지를 대상으로 수행되었다. 제안하는 모델의 아키텍처는 GELU(Gaussian Error Linear Unit) 활성화 함수를 적용한 fully convolutional network를 기반으로 구성하였다. 계산 효율성을 고려하여 어텐션 연산을 배제하였으며, 임베딩 모듈에서는 스케일 s와 타임스텝 t의 정보를 효과적으로 인코딩한다. 구체적으로, 사인파 위치 임베딩(Sinusoidal Positional Embedding)을 통해 스케일과 타임스텝 정보를 통합하고, 이를 GELU 활성화 함수가 포함된 선형 레이어에 통과시켜 특징을 추출하였다. 모델은 기본 구조는 그림 3-3에서 제시된 U-Net을 사용한다. 실험은 NVIDIA GeForce RTX 3090 GPU 환경에서 수행되었으며, 각 스케일은 기존 확산 모델의 $T=1000$ 타임스텝이 아닌 $T=100$ 타입스텝을 사용하였다. 학습은 Adam 옵티마이저를 사용하여 70,000~80,000 스텝 동안 학습되어 모든 스케일 수준에서 충분한 학습을 보장하였다.

4.2 정량 평가

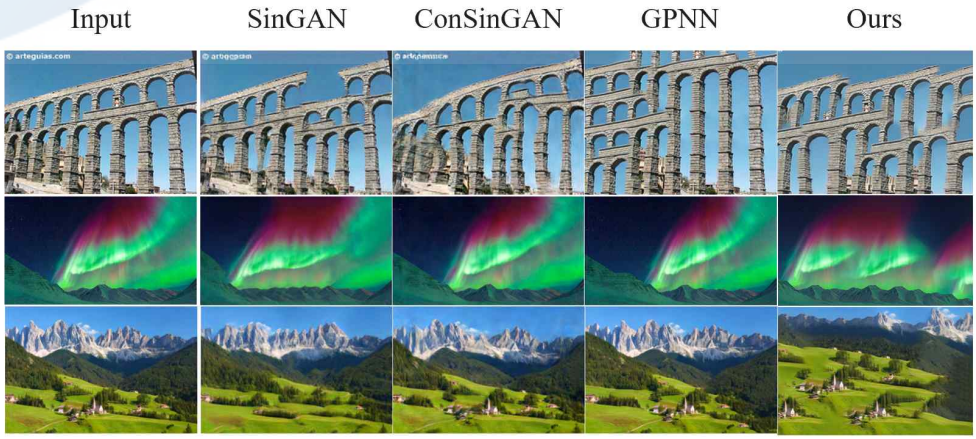
Type	Metric	SinGAN	ConSinGAN	GPNN	Ours
Diversity	LPIPS Div.↑	0.18±0.07	0.15±0.07	0.1±0.07	0.52 ± 0.06
No reference IQA	NIQE↓	7.3±1.5	6.4±0.9	7.7±2.2	6.1 ± 2.5
	NIMA↑	5.6± 0.5	5.5±0.6	5.6±0.7	4.3 ± 0.3
	MUSIQ↑	43± 9.1	45.6±9	52.8±10.9	54.7 ± 9.1
Patch Distribution	SIFID↓	0.15± 0.05	0.09±0.05	0.05±0.04	0.49 ± 0.27

[표 4-1] 조건 없는 이미지 생성에 대한 정량적 평가

표4-1은 12개의 이미지에 대해 다른 단일 이미지 생성 모델들과의 정

량적 비교 결과를 보여준다. 제안하는 모델의 성능은 생성된 이미지의 시각적 품질과 다양성 두 가지 측면에서 평가되었다. 시각적 품질 평가를 위해 SinGAN에서 사용된 단일 이미지 Frechet Inception Distance[19](SIFID)을 평가 지표로 채택하였다. SIFID는 FID와 유사한 방식으로, 생성된 이미지와 실제 이미지의 패치 단위 특징 분포 간의 편차를 측정한다. 또한 모델의 지각적 품질 우수성을 입증하기 위해 NIQE(Natural Image Quality Evaluator)[15], NIMA(Neural Image Assessment)[16], MUSIQ(Multi-scale Image Quality Transformer)[17]와 같은 참조 없는 이미지 품질 평가 지표들을 활용하였다. 한편, 생성된 이미지의 다양성은 다중 모달 생성 결과들 사이의 LPIPS(Learned Perceptual Image Patch Similarity)[18] 지표로 측정된 평균 거리를 계산하였다. 이 때, 생성 모델 연구에서는 LPIPS의 값이 클수록 다양성 있는 이미지가 생성되었다고 평가하는데 표4-1에서 확인할 수 있듯이 이미지의 품질적인 측면에서도 상응하면서 다양성 있는 이미지를 생성하는 것을 확인할 수 있다.

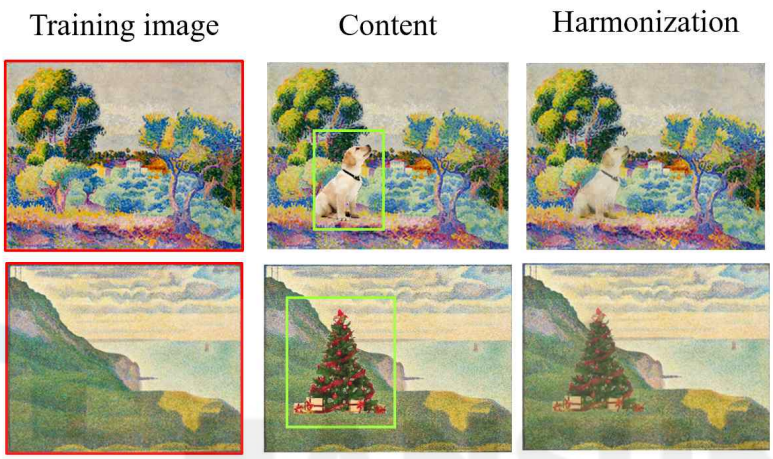
4.3 정성 평가



[그림 4-1] 조건 없는 이미지 생성 결과

본 연구에서는 제안하는 모델과 다른 단일 이미지 생성 모델들을 조건 없

는 이미지 생성 측면에서 정성적으로 비교하였다. 결과에서 볼 수 있듯이, 우리의 모델은 품질과 일반화 측면에서 다른 모델들과 최소한 동등한 수준의 성능을 보여주면서 동시에 아티팩트 누적 오류없이 이미지 생성이 된 것을 확인할 수 있다.



[그림 4-2] 이미지 조화(Harmonization)결과

이미지 조화(Harmonization)는 단순히 붙여 넣은 객체가 있는 이미지를 중간 스케일과 타임스텝에 주입함으로써, 해당 객체의 외관을 학습 이미지의 스타일과 일치시키는 이미지 조작 작업이다. 결과를 보면 알 수 있듯이 객체가 블러하게 조화되지 않고 형태를 유지하면서 학습 이미지와 조화롭게 생성되는 것을 확인할 수 있다.

V. 결 론

우리는 확산 모델의 강력한 이미지 생성 능력과 웨이블릿 도메인을 활용하여 다중 스케일 구조를 결합한 단일 이미지 생성 모델을 연구하였다. 추가적으로 본 연구는 다양한 이미지 조작 기법을 가능하게 하여 효과성을 입증하였다.

본 논문의 주요 목적은 주어진 단일 이미지로부터 고품질의 다양성 있는 이미지를 생성하는 것에 있으며 이는 제한된 데이터셋 환경에서의 이미지 생성, 3D 생성, 참조 기반 편집, 맞춤형 생성 등 다양한 작업에 유용하게 활용될 수 있다. 이러한 특성으로 인해 실제 응용에서 폭넓게 활용될 수 있는 가능성을 보여준다. 우리는 추후 본 기술을 발전시켜 더 짧은 샘플링 시간으로 단일 이미지 편집을 할 수 있는 단일 이미지 생성을 목표로 연구 중이다.

참 고 문 헌

1. 국외문헌

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015. Lecture Notes in Computer Science, Vol. 9351, pp. 234-241, 5.

Dhariwal, P., & Nichol, A. (2021). Diffusion models beat gans on image synthesis. In Advances in Neural Information Processing Systems, Vol. 34, 3.

Granot, N., Feinstein, B., Shocher, A., Bagon, S., & Irani, M. (2022). Drop the gan: In defense of patches nearest neighbors as single image generative models. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 9.

Hinz, T., Fisher, M., Wang, O., & Wermter, S. (2021). Improved techniques for training single-image gans. In Proceedings- 2021 IEEE Winter Conference on Applications of Computer Vision, WACV 2021, 7.

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. In Advances in Neural Information Processing Systems, Vol. 33, 1.

Kawar, B., Elad, M., Ermon, S., & Song, J. (2022). Denoising diffusion restoration models. In Advances in Neural Information Processing Systems, Vol. 35, pp. 23593-23606, 2.

Kulikov, V., Yadin, S., Kleiner, M., & Michaeli, T. (2023). Sinddm: A single image denoising diffusion model. In International Conference on Machine Learning, pp. 17920-17930, 4.

Shaham, T. R., Dekel, T., & Michaeli, T. (2019). Singan: Learning a generative model from a single natural image. In Proceedings of the IEEE International Conference on Computer Vision, 6.

Shocher, A., Bagon, S., Isola, P., & Irani, M. (2019). Ingan: Capturing and retargeting the 'dna' of a natural image. In Proceedings of the IEEE International Conference on Computer Vision, 8.

Wang, W., Bao, J., Zhou, W., Chen, D., Chen, D., Yuan, L., & Li, H. (2022). Sindiffusion: Learning a diffusion model from a single natural image, 10.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2020). Generative adversarial networks. Communications of the ACM, 63(11), 11.

P. Dhariwal, A. Nichol, Diffusion models beat gans on image synthesis, In: Advances in Neural Information Processing Systems, Vol. 34, 2021, 12.

S. Lee, H. Chung, J. Kim, J. C. Ye, (2022). Progressive Deblurring of Diffusion Models for Coarse-to-Fine Image Synthesis. NeurIPS 2022 SBM workshop, arXiv:2207.11192v2, 13.

J. Song, C. Meng, S. Ermon. (2020). Denoising Diffusion Implicit Models, In: International Conference on Learning Representations (ICLR), 14.

Mittal, A., Soundararajan, R., & Bovik, A. C. (2013). Making a 'completely blind' image quality analyzer. IEEE Signal Processing Letters, Vol. 20(3), pp. 209-212, 15.

Talebi, H., & Milanfar, P. (2018). NIMA: Neural Image Assessment. IEEE Transactions on Image Processing, Vol. 27(8), pp. 3998-4011, 16.

Ke, J., Wang, Q., Wang, Y., Milanfar, P., & Yang, F. (2021). MUSIQ: Multi-scale Image Quality Transformer. In Proceedings of the

IEEE International Conference on Computer Vision (ICCV), pp. 5148-5157, 17

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 586-595, 18

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In Advances in Neural Information Processing Systems (NeurIPS), Vol. 30, pp. 6626-6637, 19.

ABSTRACT

Efficient Image Synthesis Method Using Single Reference Image-Based Diffusion Model

Kim, Ji-Soo

Major in Applied Artificial Intelligence

Dept. of Applied Artificial Intelligence

The Graduate School

Hansung University

Recent advances in representation learning have led to remarkable progress in image synthesis. However, most innovative models heavily rely on large-scale training datasets, limiting their practical applications. This paper presents a novel diffusion model that can generate diverse, high-quality images from a single reference image. Our proposed method combines wavelet domain frequency decomposition with a hierarchical scale structure to address the artifact accumulation problem inherent in existing approaches.

Specifically, we introduce a U-Net architecture with restricted receptive fields to prevent overfitting to global information, while enabling effective feature learning in the frequency domain through wavelet transforms. Through extensive experiments, we demonstrate that our

model achieves superior visual quality and diversity compared to existing single-image generation methods, and validates its effectiveness in real-world applications such as image harmonization. Furthermore, our proposed model significantly reduces the lengthy sampling time of conventional diffusion models, improving computational efficiency while enabling stable image generation at arbitrary resolutions.

Our research presents a novel solution to the image generation problem in limited data environments and demonstrates various practical applications. The quantitative and qualitative evaluations show that our method outperforms previous approaches across multiple metrics, including LPIPS diversity scores and no-reference image quality assessments. Future work will focus on extending this methodology to achieve faster inference speeds and enhanced image quality through advanced optimization techniques.

【Key words】 Diffusion Models, Single Image Generation, Wavelet Transform