확산 모델 기반 이미지 데이터 증강 기술 동향 분석

정 다 예 나 , 오 시 영 , 최 성 록
(서울과학기술대학교)

I. 서론

딥러닝 기술의 발전으로 컴퓨터 비전은 이미지 분류, 객체 탐지 등 다양한 인식(recognition) 태스크에서 인간을 넘어서는 성능을 달성하며 의료, 자율주행 등 핵심 산업 분야에 성공적으로 적용되고 있다. 그러나 이러한 성과에도 불구하고, 딥러닝 모델의 성능은 학습 데이터의 품질과 양에 절대적으로 의존한다는 근본적인 한계를 가진다. 실제 산업 현장에서는 데이터 편향, 클래스 불균형, 희귀 사례 부족과 같은 문제들이 모델의 일반화 성능을 저해하는 핵심 병목(bottleneck)으로 작용한다.

이러한 데이터의 한계는 특정 도메인에서 더욱 심각한 문제로 이어진다. 자율주행 시스템은 악천후나 야간과 같은 비정형적(out-of-distribution) 시나리오에서 성능이 급격히 저하되며, 의료 영상 AI는 소수의 희귀 질환 데이터를 충분히 학습하지 못해 진단 정확도에 한계를 보인다. 또한, 제조업에서는 미세 결함 데이터 확보의 어려움이, 지능형 감시 분야에서는 예측 불가능한 환경 변화가 각각 모델의 신뢰도를 떨어뜨리는 주요 원인이 된다. 이 모든 문제의 근원은 결국 훈련 데이터가 실제 세계의 무한한 다양성을 충분히 반영하지 못하는 데 있다.

이러한 데이터 부족 문제를 해결하기 위한 가장 보편적인 방법은 데이터 증강(data augmentation)이다. 회전, 자르기, 색상 변환과 같은 전통적인 증강 기법은 데이터의 양을 늘리는 데 효과적이지만, 이는 기존 데이터 분포 내에서의 제한적인 변형에 불과하다. 따라서 실제 수집이 어려운 새로운 환경이나 희귀 사례를 생성하지 못하며, 모델의 근본적인 일반화 성능을 높이는 데에는 명백한 한계가 있다.

이러한 전통적 기법의 한계를 극복할 대안으로, 최근 확산 모델(Diffusion Model) 기반의 생성형 증강이 새로운 해결책으로 부상하고 있다. 확산 모델은 단순한 변형을 넘어, 데이터의 근본적인 분포를 학습하여 세상에 존재할 법한(plausible) 고품질의 이미지를 '창조'한다. 특히 텍스트, 마스크, 스타일 등 다양한 조건을 통해 생성 과정을 정교하게 제어할 수 있다는 점은, 현실에서는 수집이 불가능하거나 매우 드문 시나리오(예: '눈 오는 날의 사막 도로를 주행하는 자율주행차')를 가상으로 생성할 수 있게 한다. 이는 데이터 증강의 패러다임을 바꾸는 혁신적인 잠재력이다.

본고는 이러한 확산 모델 기반 데이터 증강의 최신 연구 동향을 포괄적으로 고찰하고, 실제 산업 응용의 관점에서 그 가능성과 한계를 분석한다. 이를 위해 확산 모델의 기초 원리부터 의미 기반 변형, 개인화, 도메인 적응과 같은 핵심 증강 전략들을 체계적으로 분류하고 설명한다. 나아가 다양한 응용 사례 분석을 통해 기술의 실효성을 검증하고, 현재 기술이 마주한 도전 과제와 미래 연구 방향을 제시하여 독자들에게 깊이 있는 통찰을 제공하고자 한다.

본고의 구성은 다음과 같다. 2장에서는 확산 모델의 기본 원리와 조건부 생성 메커니즘을 소개한다. 3장에서는 다양한 증강 기법을 최신 연구 사례와 함께 심도 있게 분석하고, 실제 다운스트림 태스크에서의 성능을 종합적으로 평가한다. 4장에서는 증강 기법의 효용과 한계를 판단하는 정량적·정성적 평가 지표를 다루며, 5장과 6장에서는 각각 현재 기술의 당면 과제와 미래 연구 방향, 그리고 전체 내용을 요약하는 결론을 제시한다.

II. 확산 모델의 기초

확산 모델(Diffusion Models, DMs)은 이미지 합성 분야에서 가장 주목받는 생성 모델 중 하나로[1,2], 강력한 성능을 바탕으로 다양한 컴퓨터 비전 분야에 폭넓게 적용되고 있다[3,4]. 본 장에서는 확산 모델의 핵심 원리와 주요 변형 모델을 체계적으로 살펴보고, 특히 데이터 증강 관점에서의 이론적 기반을 다룬다. 확산 모델의 기본적인 순방향 및 역방향 과정 파이프라인은 그림 1과 같다.

1. 순방향 확산 과정 (Forward Diffusion Process)

순방향 확산은 원본 이미지에 가우시안 노이즈(Gaussian noise)를 점진적으로 주입하여 데이터를 무작위 분포로 변환하는 과정이다[5]. 초기 이미지 x_0는 데이터 분포 $q(x_0)$로부터 샘플링되며, 시간 단계 $t=1$부터 T까지 반복적으로 노이즈가 추가되어 이미지 시퀀스 x_1, x_2, …, x_T가 생성된다. 각 단계의 조건부 분포는 다음과 같은 가우시안 분포로 정의된다[5].

$$q(x_t \mid x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I\right)$$

여기서 β_t는 각 단계에서 추가되는 노이즈의 강도를 의미하는 분산 파라미터이고, I는 단위 행렬이다.

또한, 재매개변수화 트릭(reparameterization trick)을 활용하면[6], 원본 이미지 x_0로부터 임의의 시간 단계 t의 노이즈 이미지 x_t를 직접 샘플링할 수 있다. α_t를 $1-\beta_t$로 정의하고, 누적 계수인 $\bar{\alpha}_t$는 t까지의 α_i를 곱한 값($\prod_{i=1}^{t}\alpha_i$)으로 정의된다. 이 경우, 다음과 같은 형태로 표현된다[5].

$$q(x_t \mid x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I)$$

따라서, 시간 단계 t에서의 x_t는 다음과 같이 계산할 수 있다[5].

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon,\ \epsilon \sim \mathcal{N}(0,I)$$

이러한 식은 학습 단계에서 모델이 예측해야 하는 노이즈 ϵ에 대한 정답 레이블 역할을 제공한다.

2. 역방향 확산 과정(Reverse Diffusion Process)

순방향 확산에 의해 무작위 노이즈 분포로 전환된 데이터는, 역방향 확산(reverse diffusion; RD) 과정을 통해 원래의 데이터 분포로 복원된다. 이 과정은 디노이징 모델을 활용하여 각 시간 단계에서 노이즈를 점진적으로 제거하는 방식으로 구성된다.

대표적인 접근 방식은 디노이징 확산 확률 모델(Denoising Diffusion Probabilistic Models; DDPM)로, 다음과 같은 모델링을 따른다.

$$p_\theta(x_{0:T}) = p(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1} \mid x_t),$$
$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t,t), \Sigma_\theta(x_t,t))$$

여기서 μ_θ와 Σ_θ는 신경망을 통해 예측되는 평균과 분산이다. U-Net 기반 모델은 다음과 같은 단순한 손실함수로 학습이 가능하다[7].

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{t,x_0,\epsilon}\left[\| \epsilon - \epsilon_\theta(x_t,t) \|_2^2\right]$$

여기서 ϵ는 원본 이미지 x_0에 추가된 실제 노이즈이며, ϵ_θ는 모델이 예측한 노이즈이다.

또한, 스코어 기반 생성 모델(Score-based Generative Models; SGMs)은 데이터의 로그 확률 밀도 함수에 대한 그래디언트, 즉 스코어 함수 $\nabla_{x_t}\log p_\theta(x_t)$를 추정하여 역방향 샘플링을 수행한다[8].

3. 조건부 생성과 유도 기법(Guidance Mechanisms)

조건부 확산 모델은 특정 목표를 향한 이미지 생성을 유도하기 위해 다양한 방식의 조건 정보를 활용한다. 대표적인 방식은 다음과 같다.

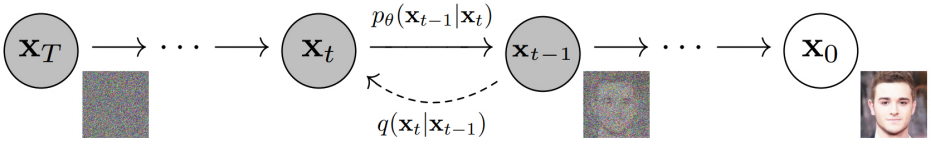

그림 1. 확산 모델의 기본 파이프라인: 순방향 과정에서 점진적으로 노이즈가 주입되고, 역방향 과정에서 단계별로 노이즈가 제거되며 원본 이미지가 복원되는 과정을 보여준다[5].

1) 분류기 유도(Classifier Guidance)

사전 학습된 폐쇄형 분류기를 활용하여, 비조건부로 학습된 확산 모델의 역방향 확산 과정을 원하는 클래스 레이블로 유도하는 방식이 제안되었다[1]. 이때 분류기는 확률 분포 $p_\phi(y \mid x_t)$를 제공하며, 여기서 x_t는 시간 x_t에서의 노이즈 이미지이다. 생성 과정에서 샘플링 경로는 다음 그래디언트에 의해 조정된다.

$$\nabla_{x_t} \log p_\phi(y \mid x_t)$$

이를 기반으로 확산 모델의 평균 예측 $\mu_\theta(x_t, t)$는 다음과 같이 조정된다.

$$\tilde{\epsilon}_\theta(x_t, t) = \epsilon_\theta(x_t, t) - s \cdot \nabla_{x_t} \log p_\phi(y \mid x_t)$$

여기서 s는 유도 강도를 조절하는 하이퍼파라미터이다. 이 방식은 기존 비조건부 모델을 재학습하지 않고도 클래스 조건을 반영한 생성이 가능하다는 장점이 있다.

2) 분류기 없는 유도(Classifier-Free Guidance)

별도의 분류기 없이 조건 유도를 수행하는 분류기 없는 유도(Classifier-Free Guidance) 방식이 제안되었다[9]. 확산 모델은 학습 시 조건부 예측과 비조건부 예측을 모두 번갈아가며 학습한다. 추론 시에는 두 예측 결과를 선형 보간하여 유도 효과를 부여한다. 이때 다음과 같은 방식으로 그래디언트가 조합된다.

$$\epsilon_{\text{guided}} = (1 + w) \cdot \epsilon_\theta(x_t, t, y) - w \cdot \epsilon_\theta(x_t, t)$$

여기서 w는 유도 강도를 조절하는 계수이다. 이 방식은 분류기 없이도 유연하고 강력한 조건 제어가 가능하다는 점에서 최근 많은 모델에서 채택되고 있다.

4. 잠재 공간 기반 확산 모델(Latent Diffusion Models)

기존 확산 모델은 고해상도 이미지에 적용 시 막대한 계산 비용이 발생하는 한계가 있다. 잠재 공간 확산 모델(Latent Diffusion Model, LDM)은 이를 해결하기 위해 제안된 구조이다[2]. LDM은 사전 학습된 오토인코더(autoencoder)를 이용해 고차원 이미지를 저차원 잠재 공간(latent space)으로 압축하고, 이 잠재 공간 안에서 확산 및 복원 과정을 수행한다. 이러한 전체적인 생성 흐름은 그림 2에 자세히 묘사되어 있다.

오토인코더는 다음과 같은 구성으로 동작한다.

- 인코더 E : 입력 이미지 x를 잠재 벡터 $z = E(x)$로 압축
- 디코더 D : z로부터 다시 원래 이미지를 복원 $(D(E(x)) \approx x)$

이러한 인코더-디코더 기반 잠재 공간 표현은 Van Den Oord et al. [10], Agustsson et al. [11] 등의 연구에서 그 정당성이 입증되었으며, LDM은 이를 기반으로 효율적인 고해상도 이미지 생성을 가능하게 한다.

이후, 확산 모델은 잠재 공간 내에서 z를 대상으로 학습되며, 다음과 같은 손실 함수를 따른다.

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{z \sim E(x),\, y,\, \epsilon \sim \mathcal{N}(0,1),\, t}\left[\left\| \epsilon - \epsilon_\theta(z_t, t, c_\theta(y)) \right\|_2^2\right]$$

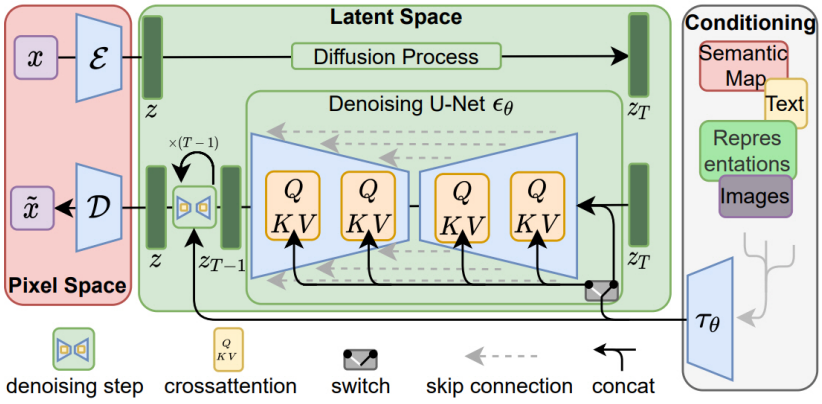

그림 2. 잠재 공간 확산 모델(LDM)의 아키텍처: 픽셀 공간(Pixel Space)의 원본 이미지(x)는 인코더(E)를 통해 저차원 잠재 벡터(z)로 압축된다. 확산 및 디노이징 과정은 계산 효율성이 높은 잠재 공간(Latent Space) 내에서 수행되며, 이때 텍스트 등의 외부 조건(Conditioning)이 교차 주의(cross-attention)를 통해 주입된다. 최종적으로 디코더(D)가 잠재 벡터를 다시 픽셀 공간으로 복원하여 고해상도 이미지를 생성한다[2].

여기서

- z_t는 시간 t에서 노이즈가 추가된 잠재 벡터
- ϵ은 추가된 실제 노이즈이며, ϵ_θ는 네트워크가 예측한 노이즈
- $c_\theta(y)$는 클래스 레이블, 텍스트, 세그멘테이션 마스크 등 다양한 조건 입력을 잠재 벡터로 매핑해주는 함수

잠재 공간 기반 확산 모델은 고해상도 이미지도 빠르게 생성할 수 있다는 장점이 있으며, 다양한 조건 정보와 쉽게 결합되어 실제 응용에서도 널리 활용되고 있다.

5. 트랜스포머 기반 확산 모델(Diffusion Transformers)

기존 확산 모델의 아키텍처가 대부분 U-Net을 기반으로 한 것과 달리, 최근 트랜스포머(Transformer)를 도입한 Diffusion Transformer (DiT) [12]가 제안되며 새로운 전환점이 마련되었다. DiT는 Vision Transformer (ViT) [13]의 핵심 아이디어를 차용하여, 이미지를 여러 개의 작은 패치(patch)로 분할한 뒤 이를 하나의 시퀀스(sequence) 데이터처럼 처리하는 접근법을 사용한다. DiT의 전체적인 아키텍처와 핵심 구성 요소인 DiT 블록의 상세 구조는 그림 3에 나타나 있다.

DiT의 핵심적인 구조는 다음과 같다.

- **패치화(Patchification)**: 입력 이미지를 겹치지 않는 여러 개의 패치로 분할하고, 각 패치를 선형 임베딩(linear embedding)하여 토큰 시퀀스를 생성한다.
- **트랜스포머 블록(Transformer Blocks)**: 패치 토큰 시퀀스는 여러 개의 트랜스포머 블록을 통과한다. 각 블록의 자기 주의(self-attention) 메커니즘은 이미지 전체의 패치 간 관계, 즉 장거리 의존성(long-range dependency)을 효과적으로 학습한다.
- **조건 주입(Conditioning)**: 시간 단계(timestep)와 텍스트 등의 조건 정보는 별도의 임베딩으로 처리된 후, 트랜스포머 블록에 주입되어 생성 과정을 제어한다.
- **디코더(Decoder)**: 마지막 트랜스포머 블록을 통과한 토큰 시퀀스는 최종 디코더를 거쳐 원본 해상도의 이미지로 재구성된다.
- 이러한 DiT 아키텍처는 기존 U-Net 기반 모델 대비 다음과 같은 명확한 이점을 가진다.
- **뛰어난 확장성(Scalability)**: 트랜스포머는 모델과 데이터의 규모가 커질수록 U-Net보다 더 큰 폭의 성능 향상을 보인다. 이는 대규모 데이터셋을 활용한 초거대 모델 학습에 유리하다[12].
- **효과적인 전역 정보 학습**: U-Net의 컨볼루션 연산이 이미지의 지역적 특징(local feature)에 집중하는 반면, DiT의 자기 주의는 이미지 전체의 맥락과 구조를 한 번에 학습하므로 더 정교한 이미지 생성이 가능하다.

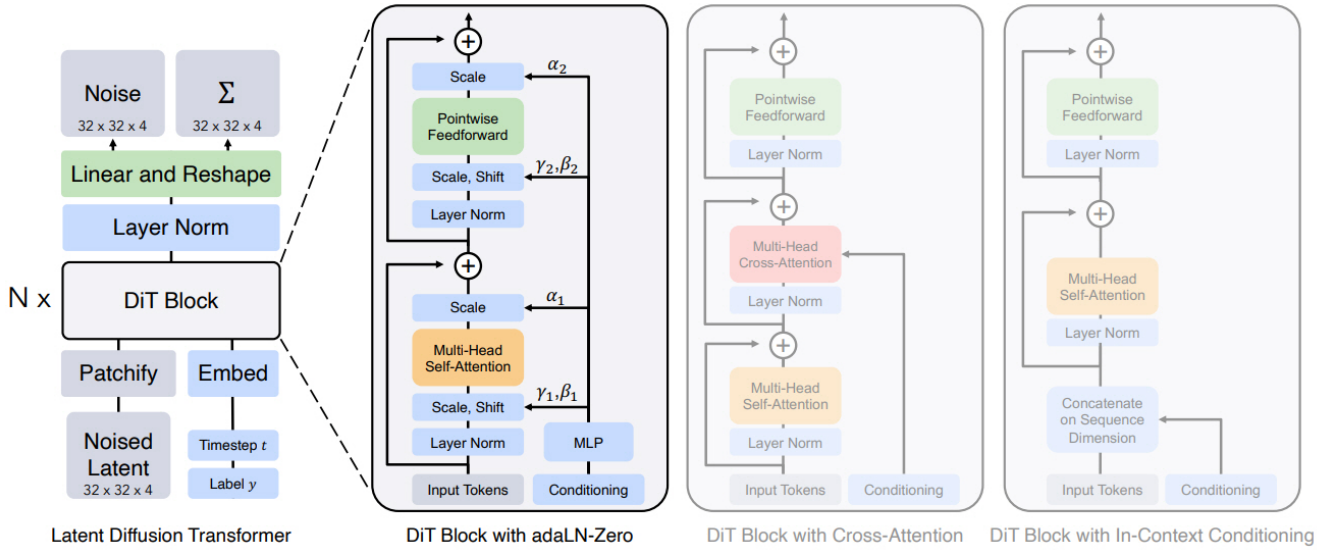

그림 3. Diffusion Transformer (DiT)의 아키텍처: DiT는 U-Net 대신 트랜스포머 구조를 사용한다. 노이즈가 주입된 잠재 벡터(Noised Latent)를 패치화(Patchify)하여 토큰 시퀀스로 변환한 뒤, 여러 개의 DiT 블록을 통과시킨다. 각 DiT 블록은 자기 주의(self-attention)와 adaLN-Zero 기법을 통해 시간 및 조건 정보를 효과적으로 처리하며, 최종적으로 예측된 노이즈를 출력한다[12].

• **유연한 조건 통합**: 텍스트, 클래스 레이블 등 다양한 형태의 조건 정보를 토큰(token) 형태로 쉽게 통합할 수 있어, 다중 모달(multi-modal) 제어에 유리한 구조이다.

물론 DiT는 자기 주의 연산의 계산 복잡도가 시퀀스 길이(n)의 제곱($O(n^2)$)에 비례하기 때문에, U-Net보다 더 많은 연산 자원을 필요로 한다는 단점이 있다. 그럼에도 불구하고, 충분한 규모로 확장된 DiT는 기존 U-Net 기반 모델의 성능을 뛰어넘는 결과를 보여주며[14], 차세대 확산 모델 아키텍처로서의 가능성을 입증했다.

III. 확산 모델 기반 이미지 증강 방법론

1. 의미기반 변형

의미 기반 변형(Semantic Manipulation)은 이미지의 핵심 정체성은 유지하면서 시각적 표현을 선택적으로 제어하는 기법이다. 이 방법론은 텍스트 프롬프트, 개념 임베딩, 시각적 레이아웃 등의 의미 정보를 조건으로 사용하여, 이미지 생성 및 편집 과정에서 높은 수준의 제어 가능성을 부여하는 것을 목표로 한다[22].

1.1 개념 조작

개념 조작(Concept Manipulation)은 원본 이미지의 의미적 맥락을 보존하면서 새로운 객체를 삽입하거나 기존 요소를 대체하는 방식이다. 이 접근법은 주로 사전 학습된 Stable Diffusion (SD) 모델을 기반으로 하며, 대부분 별도의 추가 학습 없이 적용할 수 있다는 장점이 있다[15].

객체 통합 기법

ObjectStitch [15]는 객체의 정체성(identity)과 세부 표현(fine-grained details)을 분리하여 추출한 뒤, 이를 SD 모델에 주입함으로써 새로운 객체를 장면에 매끄럽게 통합하는 기법이다. 이 과정에서 DINOv2 [16]를 통한 ID 특징 추출, 세부 맵 생성, 특징 주입 등의 모듈이 순차적으로 활용되며, 사용자 마스크를 통해 객체의 형태나 위치를 정밀하게 제어할 수도 있다. 다른 접근법으로는 객체의 시각적 특징을 텍스트 임베딩 공간에 매핑하여 SD의 조건으로 활용하는 연구도 제시되었다.

잠재 공간 기반 제어

Composer [17]와 Stable Artist [18]는 잠재 공간(latent space)에서의 연산을 통해 이미지 생성을 정밀하게 제어하는 방법론이다. Composer는 텍스트, 깊이 맵, 스케치 등 다양한 요소를 분해하고 이를 조건으로 이미지를 재조합하는 방식이며, Stable Artist는 SEGA (Semantic Guidance) 기법을 통해 여러 의미 방향으로 노이즈 추정 과정을 조작하여 마스크나 미세조정 없이도 구도와 스타일을 제어한다.

안전성 및 윤리적 고려사항

개념 조작 기술의 높은 자유도는 저작권 침해나 유해 콘텐츠 생성[20]과 같은 윤리적 문제를 야기할 수 있다. 이에 대한 대응책으로 크게 네 가지 접근법이 제시되었다: (1) 생성 후 필터링[21], (2) 추론 과정에서의 유도 제어[22], (3) 인페인팅 기반 제거[23], (4) 모델 미세조정을 통한 개념 소거[24]이다. 예를 들어 Safe Latent Diffusion [21]은 특정 유해 개념에 대한 부정적 프롬프트를 활용하여 해당 콘텐츠의 생성을 억제하는 유도 제어 방식에 해당한다.

이처럼 개념 조작 기법은 대규모 생성 모델의 풍부한 의미 표현력을 바탕으로 별도의 훈련 없이도 객체의 배치, 제거, 변형을 가능하게 한다. 윤리적 문제 해결을 위한 연구가 병행된다면, 향후 이미지 편집 및 생성형 증강 분야에서 그 활용성이 더욱 확대될 전망이다.

1.2 텍스트 기반 편집

텍스트 기반 편집(Text-guided Editing)은 단순히 텍스트로 이미지를 생성하는 것을 넘어, 참조 이미지의 의미를 보존하면서 텍스트 명령에 따라 이를 수정하는 고차원적인 편집 기술이다. 핵심 원리는 참조 이미지를 완벽히 재구성할 수 있는 텍스트 임베딩을 먼저 찾고(inversion), 이를 목표 텍스트와 조합하여 편집 방향을 제어하는 것이다[25].

최적화 기반 접근법

대표적인 연구인 Imagic [25]은 사전 훈련된 확산 모델을 이용해 입력 이미지 재구성에 최적화된 텍스트 임베딩을 학습한다. 이후, 이 임베딩과 목표 텍스트의 CLIP 임베딩을 선형 보간(linear interpolation)하여 최종 조건 벡터를 구성함으로써 별도의 마스크 없이도 정교한 편집을 수행한다.

교차 주의 메커니즘 기반 접근법

교차 주의 맵(Cross-Attention Maps)을 활용하는 기법은 텍스트의 특정 단어와 이미지의 특정 영역 간의 대응 관계를 조절하여 세밀한 편집을 가능하게 한다[26]. Hertz 등[27]은 확산

과정 각 단계에서 단어별 주의 집중 영역을 시각화하여 이미지-텍스트 정렬을 강화했으며, Huang 등[28]은 이미지를 여러 영역으로 분할하고 각 영역을 독립적인 텍스트 명령어로 편집하는 방법을 제안하여 지역적 제어성을 극대화했다.

자연어 지시문 기반 접근법

자연어 지시문(human instructions)을 직접 활용하는 편집 방식 또한 주요 연구 흐름으로 자리 잡고 있다. InstructPix2Pix [29]는 GPT-3로 "말을 타는 여자" → "용을 타는 여자"와 같은 편집 지시문 쌍을 대량 생성하고, 이를 기반으로 이미지를 합성하여 학습 데이터를 구축한다. 이 방식은 인간의 복합적인 의도를 모델이 직접 학습하도록 유도한다는 점에서 큰 의미가 있다.

공간적 제어 기반 접근법

텍스트-이미지 정렬의 공간적 정확도를 높이기 위한 연구도 활발하다. Paint by Word [30]는 사용자가 특정 단어에 해당하는 이미지 마스크를 직접 지정하여 교차 주의 행렬을 조작하는 방식이다. Paint-by-Example [31]은 참조 이미지의 일부를 잘라 붙이는 듯한 예시 기반 편집을 제안하며, 바운딩 박스 내의 참조 임베딩을 CLIP 토큰으로 압축하여 확산 과정에 주입한다.

다중 모달 융합 접근법

X&Fuse [32]와 eDiff-I [33]는 텍스트 조건 외에 참조 이미지 자체를 또 다른 조건으로 활용한다. 두 이미지를 병렬 처리한 후 공유된 신경망 블록에서 특징을 통합하여 시각적 조건을 강화하는 방식이다. 특히 eDiff-I는 각 생성 단계에 특화된 전문가 모델들을 앙상블하여, 다양한 텍스트 표현의 세부 묘사 능력을 크게 향상시켰다.

요컨대, 텍스트 기반 편집 기술은 단순 생성을 넘어 사용자의 복합적인 의도를 해석하고, 의미적 일관성을 유지하며, 이미지의 특정 요소를 정밀하게 제어하는 종합적인 편집 패러다임을 제시한다. 이러한 발전은 생성형 증강(generative augmentation) 기술이 단순 데이터 양적 확대를 넘어, 질적으로 정교한 데이터를 생성하는 단계로 나아가고 있음을 보여준다.

1.3 레이아웃 및 영역 기반 편집

레이아웃 및 영역 기반 제어는 이미지의 공간적 구성과 객체 배치를 정밀하게 조작하는 기술이다. 이 접근법은 텍스트만으로는 제어하기 어려운 공간적 제약을 모델에 직접 부여함으로써, 사용자의 의도를 보다 정확히 반영하는 이미지 생성을 가능하게 한다.

레이아웃 기반 의미적 합성

SceneComposer [34]은 다중 해상도 마스크 피라미드로 각 영역의 형태를 정밀하게 표현하고, 이를 텍스트 임베딩과 결합하는 방식을 사용한다. Layoutdiffusion [35]은 레이아웃 내 객체의 위치와 클래스 정보를 텍스트 토큰으로 변환하여 확산 모델의 조건으로 활용하는 기법을 제시했다.

공간적 제약 조건 활용

공간적 제약 조건을 제어하는 가장 대표적인 방법론은 ControlNet [36]이다. ControlNet의 핵심은 사전 학습된 대규모 확산 모델의 가중치는 그대로 고정(locked)시킨 채, 새로운 공간적 조건을 학습하기 위한 '학습 가능한 사본(trainable copy)' 네트워크를 병렬로 추가하는 독창적인 아키텍처에 있다(그림

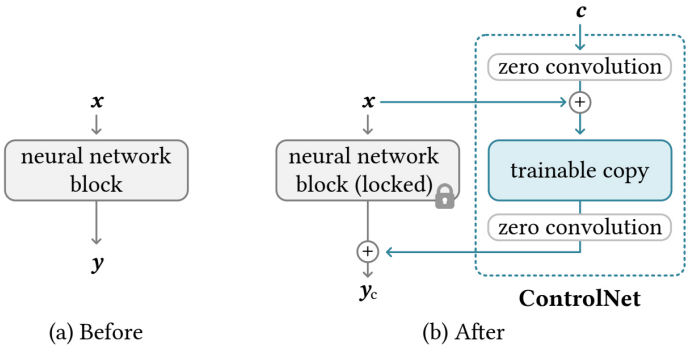

그림 4. ControlNet의 기본 원리: (a) 사전 학습된 신경망 블록. (b) ControlNet은 원본 블록의 가중치를 고정(locked)하고, 새로운 조건(c)을 처리하기 위한 학습 가능한 사본(trainable copy)을 병렬로 추가한다. 제로 컨볼루션을 통해 두 출력이 결합되므로, 원본 모델의 지식을 보존하면서 새로운 제어 능력을 안전하게 주입할 수 있다[36].

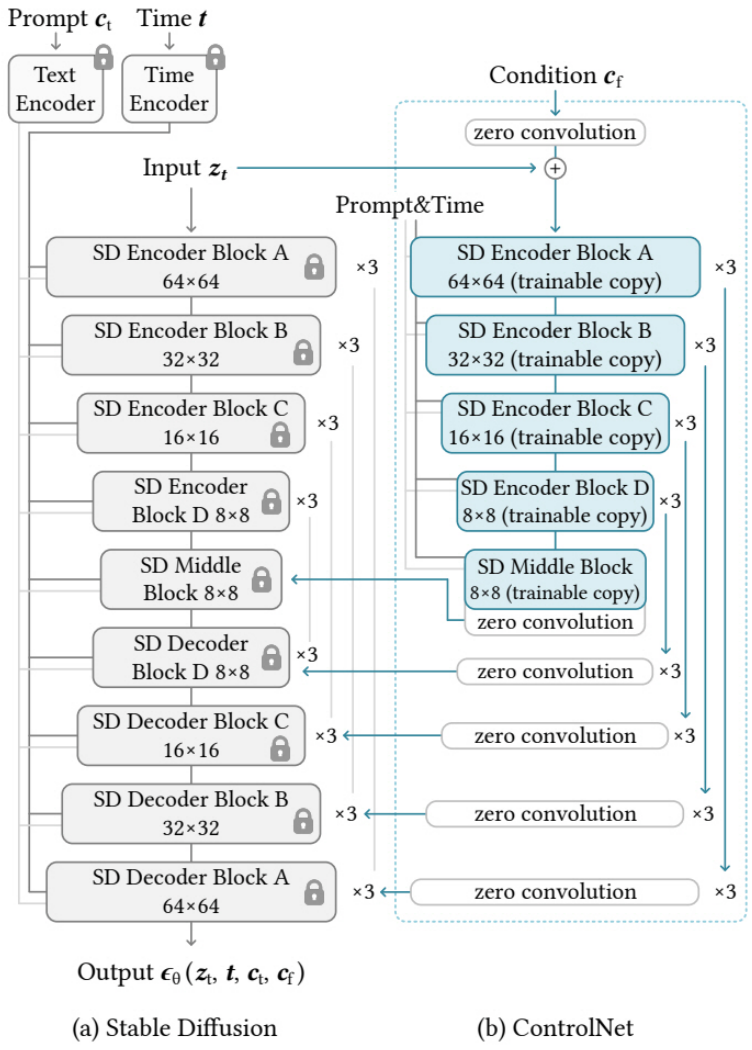

그림 5. Stable Diffusion에 적용된 ControlNet 아키텍처: (a) 기존 Stable Diffusion의 U-Net 구조. (b) ControlNet은 U-Net의 인코더 블록들을 고정하고, 각 블록에 대응하는 학습 가능한 사본을 추가하여 외부 조건(cf)을 처리한다. 이 조건부 특징(feature)은 각 수준(level)에서 원본 특징에 더해져, 전체 생성 과정을 정밀하게 제어한다[36].

4 참조). 그림 5는 이 구조가 Stable Diffusion의 U-Net 아키텍처에 어떻게 적용되는지를 보여준다. 원본 U-Net 인코더의 각 블록은 고정되고, 이에 대응하는 학습 가능한 사본이 추가되어 깊이, 포즈와 같은 외부 조건(condition)을 처리한다.

이때 두 네트워크는 제로 컨볼루션(zero convolution)으로 연결되는데, 이는 추가된 조건 네트워크가 학습 초기에는 원본 모델에 아무런 영향을 주지 않도록 하여 학습 안정성을 확보하는 역할을 한다. 이 구조 덕분에 소규모 데이터셋으로도 효율적인 미세조정이 가능하며, 이후 스케치, 스크리블 등 다양한 공간 제어 연구의 핵심 기반이 되었다.

의미적 레이아웃 제어

Xue 등[37]의 연구는 의미 기반 레이아웃 텍스트를 확산 모델에 주입하는 방법론이다. 이들은 RCA (Rectified Cross-Attention) 모듈을 통해 각 레이아웃 영역에 대응하는 텍스트 의미를 U-Net의 교차 주의 레이어에 정렬시켜, 사용자 정의 레이아웃 기반의 정밀한 이미지 생성을 구현했다.

콜라주 기반 합성

Collage Diffusion [38]은 레이어별 텍스트와 마스크로 구성된 콜라주를 입력받아, 각 객체의 스타일과 위치는 보존하면서 전체 이미지를 조화롭게 합성하는 구조이다. 이 방법은 텍

스트-이미지 교차 주의 메커니즘을 수정하여 사용자가 객체별 디테일을 레이어 단위로 제어할 수 있도록 한다.

영역 단위 인페인팅 접근법

영역 단위 편집을 위한 인페인팅 기반의 방법론 역시 활발히 연구되는 분야이다. SmartBrush [39]는 객체 마스크와 텍스트를 기반으로 특정 영역만 부분적으로 수정하며, FastComposer [40]는 다중 객체의 시각적 특징을 프롬프트에 주입하여 제어한다. Levin과 Fried [41]는 사용자가 직접 정의한 변화 맵(change map)으로 위치별 편집 강도를 조절하는 방식을 제안했다.

이처럼 레이아웃 및 영역 기반 편집 기술은 텍스트, 마스크 등 다중 조건을 결합하여 이미지 생성의 공간 제어성을 한 단계 끌어올린 핵심적인 방법론이다. 이는 텍스트만으로는 불가능했던 정밀한 공간 배치를 구현하며, '의도대로 생성하는' 이미지 기술의 새로운 패러다임을 열고 있다. 향후 더욱 직관적인 인터페이스와 결합하여 전문적인 디자인 도구나 콘텐츠 제작 분야에서 그 활용도가 극대화될 전망이다.

1.4 이미지 대 이미지 변환

이미지 대 이미지(Image-to-Image, I2I) 변환은 입력 이미지를 조건으로 삼아 목표 도메인의 이미지를 생성하는 기술이다. 최근 확산 모델은 정교한 조건부 설계와 새로운 손실 함수를 결합하여 이 분야의 발전을 주도하고 있다. 이러한 기법들의 핵심 과제는 원본의 구조적 특징을 유지하면서 목표하는 시각적 스타일을 효과적으로 구현하는 것이다.

확률적 편집 접근법

SDEdit [42]은 확률적 미분 방정식(SDE)에 기반한 선구적인 방법론이다. 이 기법은 원본 이미지에 노이즈를 주입한 뒤 점진적으로 제거하는 과정을 통해, 원본의 구조는 보존하면서 새로운 스타일을 적용하는 원리이다. 이를 통해 구조적 일관성과 편집 자유도 사이의 이상적인 균형점을 제시했다는 평가를 받는다.

CLIP 기반 의미적 제어

CycleNet [43]과 DiffusionCLIP [44]은 CLIP 기반의 손실 함수를 도입하여 의미적 제어(semantic control)의 정확도를 높인 사례이다. 이 모델들은 생성된 이미지가 목표 텍스트와 시각적으로 일치하도록 확산 모델을 미세 조정함으로써, 다중 모달 정렬(multi-modal alignment)을 통해 의미적 일관성이 뛰어난 고품질 편집을 구현한다.

다중 작업 조건부 확산

Saharia 등이 제안한 모델[45]은 단일 조건부 확산 모델로 색상화, 인페인팅, JPEG 복원 등 다양한 I2I 작업을 수행하는 통합 프레임워크이다. 이 방법은 $p(y|x)$ 분포를 직접 학습하는 방식으로, 입력 이미지와 조건 정보를 U-Net에 함께 입력하여 노이즈를 예측한다. 수천 단계의 점진적 복원 과정을 거쳐 여러 작업에서 일관되게 높은 품질의 결과를 보여주는 것이 특징이다.

제로샷 편집 방법론

pix2pix-zero [46]는 별도의 훈련 없이 텍스트만으로 이미지를 편집하는 제로샷(zero-shot) 방법론이다. 이 기법의 핵심은 DDIM과 BLIP을 이용해 입력 이미지로부터 노이즈 맵을 추출하고, 텍스트 임베딩 간의 의미적 차이를 분석하여 편집 방향을 자동으로 찾아내는 것이다. 여기에 새로운 교차 주의(cross-attention) 정렬 기법을 더해, 원본의 구조를 훼손하지 않으면서도 정확한 편집을 수행한다.

적응적 융합 접근법

DA-Fusion [47]은 사전 훈련된 Stable Diffusion을 기반으로, 참조 이미지를 클래스 라벨 조건에 맞춰 변형하는 적응적 융합 방법론이다. 텍스트 인버전(textual inversion) 방식을 활용해 기존 데이터셋에 없는 새로운 개념을 모델에 주입하거나 확장할 수 있다. 특히 생성 과정의 특정 시점(timestep)에 이미지 정보를 주입(splicing)하여 증강의 강도를 유연하게 조절할 수 있다는 장점이 있다.

구조적 제어 메커니즘

MasaCtrl [48]은 U-Net의 자기 주의(self-attention)를 상호 자기 주의(mutual self-attention)로 대체하여 구조적 제어를 강화한 모델이다. 이를 통해 소스 이미지의 지역 구조와 질감을 참조하면서도, 목표 텍스트와의 일관성을 동시에 확보한다. 또한, 교차 주의 맵에서 배경과 전경 마스크를 추출해 적용함으로써 공간적 정확도를 크게 향상시킨다.

결론적으로, 확산 모델 기반의 이미지 대 이미지 변환 기술은 각기 다른 접근법을 통해 빠르게 발전하고 있다. 이들의 공통된 지향점은 원본의 구조적 무결성을 지키면서 사용자가 의도한 의미적, 시각적 변환을 정교하게 구현하는 것이다. 향후에는 더욱 효율적인 추론 과정과 세밀한 제어 메커니즘의 개발을 통해, 실시간 편집이나 전문가용 도구 등 더욱 폭넓은 응용 분야로 나아갈 것이라 기대된다.

1.5 반사실적 증강

조건부 확산 모델(Conditional Diffusion Model)은 다양한 분야에서 반사실(Counterfactual) 이미지 생성을 위한 핵심 기술로 주목받고 있다.

대표적으로 의료 영상 분야에서는 Sanchez 연구팀[49]이 종양이 있는 뇌 MRI와 종양 마스크를 조건으로 '건강한 뇌' 이미지를 생성함으로써, 의료진의 설명성과 비교 분석을 돕는 사례를 제시했다. 인과 추론(Causal Inference) 분야의 Diff-SCM [50]은 한발 더 나아가, 개입 변수(do(class))를 조건으로 이미지를 변환한다. 예를 들어 do(cat)이라는 개입을 통해 원본 이미지의 배경은 보존하면서 '개'만 '고양이'로 바꾸는 반사실적 이미지를 생성하여 인과 관계를 추론한다.

반사실 이미지는 모델의 해석 가능성(Explainability)과 강건성(Robustness)을 향상시키는 데에도 효과적으로 활용된다. Madaan과 Bedathur [51]의 연구는 반사실 이미지를 통해 예측의 시각적 근거를 제시하고, 외부 요인 변화에 대한 모델의 민감도를 분석하여 강건성을 평가했다.

한편, 데이터 증강(Augmentation)의 관점에서 반사실 기법은 데이터 편향(Bias)과 공정성(Fairness) 문제를 해결하는 대안이 된다. 특정 민감 속성(성별, 인종 등)을 조건으로 이미지를 생성하면, 원본의 맥락은 유지하면서 데이터셋의 분포를 균형 있게 조정할 수 있다[52]. 또한 OOD (Out-of-Distribution) 강건성을 확보하기 위해 객체의 위치나 질감 등 특정 속성만 외삽(Extrapolation)한 반사실 이미지를 생성하여, 모델이 분포에서 벗어난 데이터에 더 잘 대응하도록 훈련 데이터를 확장하는 연구도 활발히 이루어지고 있다[53].

2. 개인화 및 적응

개인화 및 적응은 특정 데이터셋, 작업, 혹은 사용자 선호도에 맞춰 증강 프로세스를 최적화하는 기법이다. 본 절에서는 소수의 데이터를 이용해 생성 모델을 특정 주체나 스타일에 맞게 조정하는 개인화 방법론을 중심으로 살펴본다.

2.1 개인화 방법론

개인화(Personalization)는 사용자의 특정 요구에 맞춰 사전 학습된 확산 모델을 조정하는 과정으로, 주로 소량의 데이터를 이용한 미세조정(fine-tuning) 방식을 따른다.

이 분야의 대표적인 연구인 DreamBooth [54]는 단 3~5장의

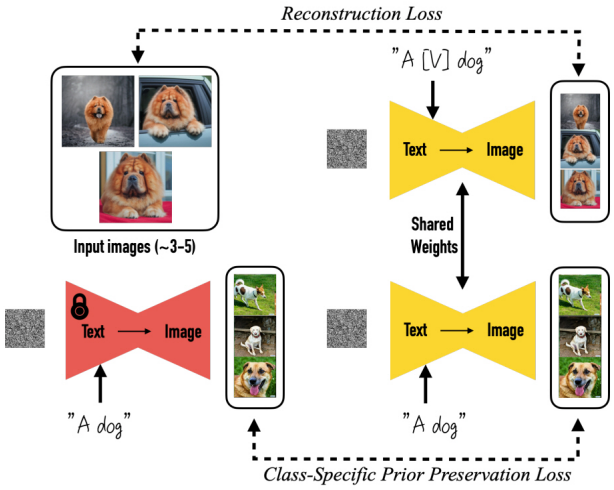

그림 6. DreamBooth의 미세조정 과정: 소수의 피사체 이미지(3~5장)와 'a [V] dog'과 같은 고유 식별자가 포함된 프롬프트를 이용해 텍스트-이미지 확산 모델을 학습시킨다. 이때, 원본 클래스('dog')의 사전 지식을 보존하는 손실 함수를 함께 사용하여 모델이 피사체의 특징을 학습하면서도 기존의 생성 능력을 유지하도록 한다[62].

이미지로 특정 피사체(subject)를 모델에 각인시키는 방법론이다. 그림 6에서 보듯이, 'a [V] dog'과 같이 고유 식별자([V])와 클래스명을 포함한 프롬프트를 사용해 모델을 미세조정한다. 이때, 원본 클래스('dog')의 의미적 사전 지식(semantic prior)이 붕괴되지 않도록 클래스 보존 손실(class-preservation loss)을 함께 적용하여, 피사체의 고유한 특징을 학습하면서도 모델의 기존 생성 능력은 유지시킨다. 그 결과, 그림 7과 같이 기존 생성 모델들이 피사체의 정체성을 유지하지 못하는 것과 달리, DreamBooth는 높은 충실도로 피사체를 보존하며 새로운 맥락에 자연스럽게 합성할 수 있다. 이후 제안된 HyperDreamBooth [55]는 이를 경량화하고 HyperNetwork 및 LoRA를 결합하여 학습 속도와 효율성을 크게 개선했다.

한편, 모델 전체를 미세조정하는 대신, 텍스트 임베딩 공간에서 특정 시각 개념을 나타내는 새로운 '의사 단어(pseudo-word)'를 찾는 Textual Inversion [56] 기법이 제안되었다. 이 접근법은 모델의 가중치를 변경하지 않으면서 새로운 개념을 텍스트로 호출할 수 있게 한다. 이를 확장한 ProSpect [57]는 단일 개념을 넘어 스타일, 질감 등 다양한 속성을 포괄하는 '프롬프트 스펙트럼'을 구성하여 더 풍부한 표현을 가능하게 한다.

여러 개념을 조합하여 새로운 이미지를 생성하려는 시도도 활발하다. Kumari 등[58]은 여러 개념을 각각의 독립적인 텍스트 임베딩으로 학습한 뒤, 이를 조합하여 창의적인 이미지를

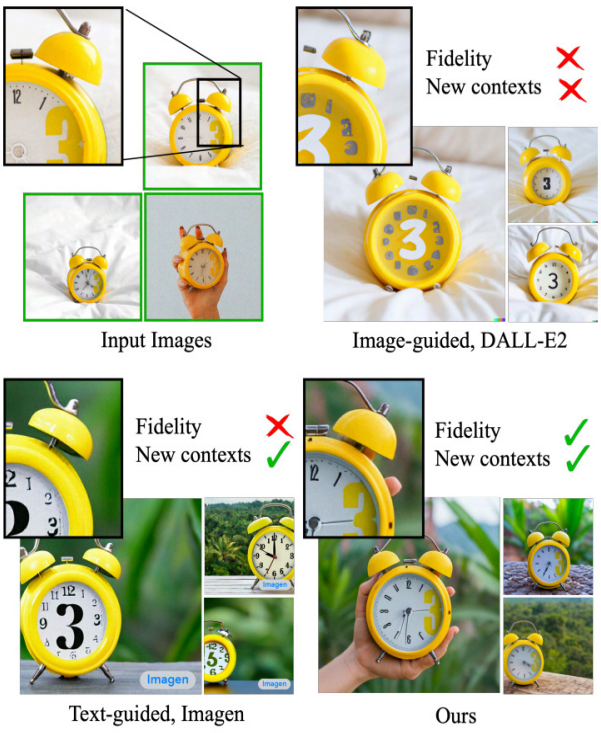

그림 7. 피사체 중심 생성 결과 비교: 기존의 이미지 기반(DALL-E 2) 및 텍스트 기반(Imagen) 생성 방식은 원본 피사체(시계)의 고유한 시각적 특징을 유지하는 데 한계가 있다. 반면, DreamBooth와 같은 개인화 기법은 피사체의 정체성을 높은 충실도로 보존하면서 새로운 맥락에 자연스럽게 합성할 수 있다[62].

생성하는 방법을 제시했다. StyleDrop [59]과 DreamStyler [60]는 특히 스타일에 집중하여, 소수의 이미지만으로 특정 화풍이나 질감을 학습하고 이를 새로운 장면에 적용하는 데 뛰어난 성능을 보인다.

미세조정의 효율성을 높이기 위해 별도의 이미지 인코더를 활용하는 접근법도 활발히 연구되고 있다. ELITE [61]의 연구는 사용자 이미지를 텍스트나 시각적 임베딩으로 변환하는 인코더를 학습하고, 이를 사전 학습된 확산 모델에 조건으로 주입한다. InstantBooth [62]는 한 걸음 더 나아가 이미지 인코더와 어댑터 레이어(adapter layer)만 학습시켜, 확산 모델 자체의 파라미터는 전혀 변경하지 않으면서도 빠른 개인화를 달성한다. Taming Encoder [63]는 긴 최적화 과정 없이 단 한 번의 순전파(forward pass)만으로 사용자 정의 객체를 생성하는 데 초점을 맞춘다.

더 나아가, 피사체의 정체성(identity)과 다른 시각적 속성(스타일, 포즈 등)을 분리하여 제어하려는 분리(disentanglement) 기반 연구도 등장했다. DisenBooth [64]는 정체성을 보존하는 임베딩과 정체성과 무관한 시각적 임베딩을 병렬적으로 활용하여 제어의 유연성을 높인다. HiPer [65] 역시 CLIP 임베딩을 정체성과 의미로 분해하여, 단 한 장의 이미지와 텍스트 쌍만으로도 정교한 이미지 조작을 가능하게 한다.

이 외에도 Perfusion [66]은 'key-locked rank-1 editing'이라는 정교한 편집 기법을 통해 오버피팅을 방지하고 개념의 정체성을 보존하며, SuTI [67]는 여러 전문가 모델을 클러스터링하여 새로운 피사체에 대한 제로샷(zero-shot) 개인화를 구현하는 등, 확산 모델 기반 개인화 기술은 빠르고, 효율적이며, 정교한 방향으로 계속해서 발전하고 있다.

2.2 적응 방법

적응(Adaptation)은 사전 학습된 확산 모델을 새로운 도메인이나 특정 작업에 맞춰 조정하는 방법론으로, 주로 미세조정(fine-tuning)을 기반으로 한다.

핵심 목표 중 하나는 도메인 간극(domain gap)을 줄이는 것이다. DomainFusion [68]은 사전 학습된 LDM을 활용하여, 소스 도메인의 이미지와 타겟 도메인의 속성(텍스트 또는 이미지)을 결합한 중간 형태의 합성 데이터를 생성함으로써 도메인 적응을 유도한다. Wu 등[69]은 디노이징 각 단계에서 소스-타겟 텍스트 임베딩을 가중 결합하는데, 이때 CLIP 기반 손실과 지각 손실(perceptual loss)을 함께 사용하여 목표 속성은 강화하고 나머지 콘텐츠는 보존하도록 최적화한다. 다른 접근법으로는 외부 모델을 활용하여 생성 과정을 제어하는 방식이 있다. Song 등[70]은 목표 도메인에 특화된 CLIP 모델로 확산 과정을 직접 유도하며, 일부 연구에서는 도메인 특화 판별자를 도입하여 부적합한 샘플을 필터링하는 방법도 제안되고 있다[71].

다른 한편으로, 특정 다운스트림 작업(downstream task)에 모델을 직접 적응시키는 연구도 활발하다. 예를 들어, Yu 등[72]은 핵(nuclei) 분할 성능을 높이기 위해, Chowdary 등[73]는 골관절염 심각도 분류를 위해 각각 현미경 이미지와 X-ray 이미지로 확산 모델을 미세조정하여 고품질의 합성 데이터를 생성했다.

미세조정 과정 자체의 효율성과 안정성을 높이려는 연구도 주목받고 있다. DomainStudio [68]는 제한된 데이터 환경에서 쌍별 유사성 손실(pairwise similarity loss)과 고주파 디테일 강

그림 8. DreamBooth와 Textual Inversion의 비교: 동일한 입력 이미지(상단)와 프롬프트를 사용했을 때, DreamBooth(중간 두 줄)가 Textual Inversion(하단)에 비해 피사체의 세부적인 특징을 더 충실하게 재현하고 프롬프트의 지시를 더 정확하게 따르는 결과를 보여준다[62].

화 기법을 통해 도메인 적응 성능을 높인다. Qiu 등[74]이 제안한 직교 미세조정(Orthogonal Finetuning, OFT)은 사전 학습된 가중치의 의미적 표현을 최대한 보존하는 새로운 패러다임을 제시한다. 이 기법은 주의 계층의 가중치에 직교 행렬(orthogonal matrix)을 곱하여 변환하는데, 이 행렬은 학습 중에도 직교성을 유지하여 원본 모델의 지식을 손상시키지 않고 빠르고 안정적인 적응을 가능하게 한다.

마지막으로, DiffuseMix [75]은 프롬프트 기반 편집을 활용해 생성한 증강 이미지로 모델의 강건성(robustness)을 높이는 독특한 적응 방식을 제안했다. 원본과 편집된 이미지를 결합하고 프랙탈 패턴을 혼합하여 과적합을 방지하고 분류 성능을 향상시킨다.

2.3 역변환 기반 방법

역변환(Inversion)은 주어진 실제 이미지를 생성할 수 있는 초기 노이즈 또는 잠재 코드(latent code)를 찾는 과정이다. 이렇게 찾아낸 잠재 코드를 조작함으로써, 기존 확산 모델을 정밀하게 제어하고 고품질의 이미지 편집 및 증강을 수행할 수 있다. 이 기술은 데이터 생성, 스타일 전이, 정밀 편집 등 다양한 분야의 핵심 기반이 된다.

역변환의 가장 직접적인 응용 분야 중 하나는 데이터 증강이다. Zhou 등[76]의 연구는 실제 이미지를 잠재 공간으로 역변환한 뒤, 그 주변에서 새로운 잠재 코드를 샘플링하여 원본과 유사하지만 새로운 합성 데이터를 생성하고, 이를 통해 다운스트림 분류 모델의 성능을 향상시켰다.

역변환 기술은 특히 정밀한 이미지 편집 분야에서 핵심적인 역할을 한다. 대표적으로 Null-Text Inversion [77]은 실제 이미지에 대한 DDIM 역변환 궤적을 먼저 찾은 뒤, 각 시간 단계에서 비조건부 텍스트 임베딩('null-text')만을 최적화하여 재구성 오류를 최소화한다. 이 과정을 통해 얻은 잠재 코드 궤적은 Hertz 등[78]의 Prompt-to-Prompt (P2P)와 같은 기법과 결합하여, 프롬프트 수정만으로 이미지의 특정 요소를 자연스럽게 편집할 수 있는 기반을 제공한다.

스타일 전이(style transfer)에도 역변환이 효과적으로 사용된다. Zhang 등[79]은 참조 예술 작품의 스타일을 표현하는 임베딩을 학습하고, 콘텐츠 이미지로부터 역변환을 통해 얻은 노이즈 맵과 결합하여 새로운 예술 작품을 생성한다. Li 등[80]은 한 걸음 더 나아가, 이미지의 구조를 제어하는 key 임베딩은 고정하고 스타일을 결정하는 value 임베딩만 역변환을 통해 추출함으로써, 구조와 스타일을 분리하여 제어하는 정교한 편집을 가능하게 했다.

역변환 자체의 품질을 개선하려는 연구도 활발하다. 기존의 최적화 기반 역변환은 시간이 오래 걸리고 정확도가 떨어질 수 있다. 이를 해결하기 위해 EDICT [81]는 정방향-역방향 확산 모델을 쌍으로 학습시켜, 더 빠르고 정확한 양방향 변환을 구현했다. LocInv [82]는 공간적 주의 맵을 활용하여 편집 과정에서 이미지의 지역적 구조(local structure)를 더 잘 보존하도록 역변환을 최적화한다.

마지막으로, Asyrp [83]는 비대칭적 역방향 과정을 통해 'h-space'라는 새로운 의미론적 잠재 공간을 발견했다. 이 공간은 시간 단계에 걸쳐 선형성, 조합성 등 안정적인 수학적 특성을 보여, 파괴적 간섭을 줄이고 더욱 안정적인 의미 조작을 가능하게 하는 새로운 길을 열었다.

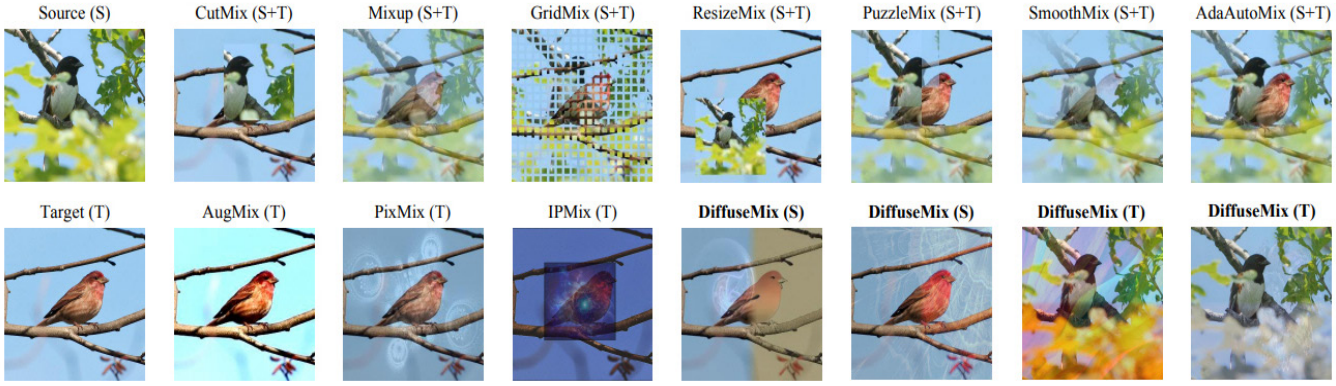

그림 9. 레이블 보존 혼합 증강(DiffuseMix)의 개념: 기존의 믹스업 방식(윗줄)은 두 이미지를 혼합하여 레이블 정보가 모호해지는 단점이 있다. 반면, DiffuseMix와 같은 레이블 보존 방식(아랫줄)은 원본의 핵심 객체는 유지한 채 일부 영역만 생성 및 결합하여, 원본 레이블을 그대로 사용할 수 있는 고품질의 증강 데이터를 만든다[75].

2.4 데이터셋 확장

확산 모델은 데이터 부족 문제를 해결하기 위한 강력한 데이터셋 확장(Dataset Expansion) 도구로 활용된다. 기존 데이터셋의 분포를 학습하여, 통계적으로 유사하지만 새로운 합성 데이터를 대량으로 생성하는 원리이다.

가장 직접적인 방식은 소규모 데이터셋을 기반으로 동종의 데이터를 추가 생성하는 것이다. 클래스 조건부 확산 모델을 학습시켜 각 클래스별 데이터를 생성하고[84], 원본의 의미적 레이아웃을 보존하는 합성 이미지를 만드는 방식[85]이 대표적이다. 한편, Gou 등[86]은 단일 이미지의 내부 패치(internal patch) 분포를 학습하여, 이미지를 암기하지 않으면서도 텍스처 등을 확장하는 새로운 접근법을 제시했다.

생성 데이터의 다양성과 일반화 성능을 높이기 위한 연구도 있다. You 등[87]은 대규모 비라벨 데이터로 사전학습을 수행한 뒤, 소규모 라벨 데이터로 미세조정하는 2단계 전략을 사용했다. Bansal과 Grover [88]는 특징 공간(feature space)에서는 가깝지만 이미지 공간에서는 먼 데이터를 생성하여, 모델이 새로운 시각 개념에 더 잘 일반화되도록 유도했다.

텍스트-이미지(Text-to-Image) 모델의 발전은 데이터셋 확장의 패러다임을 바꾸고 있다. TTIDA [89]는 텍스트-텍스트 모델로 다양한 캡션을 먼저 생성하고, 이를 텍스트-이미지 모델의 입력으로 사용하여 풍부한 주석을 가진 대규모 합성 데이터셋을 자동으로 구축한다. 이 방식은 고품질의 데이터셋을 저비용으로 확보할 수 있게 한다.

최근에는 모델 내부의 지식에만 의존하지 않고, 외부 데이터베이스를 참조하여 생성 품질과 다양성을 높이는 검색 증강(Retrieval-Augmented) 확산 모델이 새로운 방향으로 제시되고 있다. KNN-Diffusion [90]은 추론 시 텍스트 프롬프트와 가장 유사한 k개의 이웃 이미지 임베딩을 검색하여 조건으로 함께 사용한다. 이웃 정보는 텍스트와 이미지 임베딩 간의 분포 차이를 완화하여 생성의 안정성을 높인다. RDM (Retrieval-augmented Diffusion Model) [91]은 여기서 더 나아가, 검색된 시각적 예제를 디코더에 직접 조건으로 제공하여, 모델이 학습 데이터를 모두 암기하지 않고도 외부 메모리를 참조해 새로운 이미지를 생성하도록 한다. 이 구조는 상대적으로 작은 모델로도 높은 성능을 낼 수 있게 하여 모델의 효율성과 유연성을 동시에 향상시킨다.

3. 다운스트림 태스크 성능 향상

확산 모델 기반 데이터 증강은 다양한 다운스트림 태스크(downstream task)의 성능을 실질적으로 향상시키는 핵심 전략으로 자리 잡고 있다. 본 절에서는 분류(Classification), 객체 탐지(Object Detection), 시맨틱 세그멘테이션(Semantic Segmentation)의 세 가지 주요 태스크를 중심으로, 확산 모델이 어떻게 성능 개선에 기여하는지를 구체적인 연구 사례를 통해 분석한다.

3.1 분류 (Classification) 태스크

이미지 분류(Image Classification) 분야에서 확산 모델은 특히 클래스 불균형(class imbalance) 문제를 해결하는 데 효과적으로 활용된다. 데이터가 소수에 불과한 클래스(minority class)의 샘플을 고품질로 증강함으로써, 장꼬리 분포(long-tail

distribution)를 가진 데이터셋에서 모델의 일반화 성능을 크게 향상시킨다.

클래스 균형 증강

Qin 등이 제안한 CBDM (Class-Balancing Diffusion Models) [92]은 클래스별 샘플 수에 따라 확산 과정을 동적으로 조절하는 독창적인 접근법이다. 이 기법은 데이터가 적은 희소 클래스에 대해서는 순방향 확산 과정의 노이즈 스케줄을 완화하여 원본의 고유한 특징이 최대한 보존되도록 하고, 데이터가 많은 다수 클래스에 대해서는 더 강한 노이즈를 주입하여 다양성을 확보한다. 이러한 클래스 재균형 정규화를 통해, CBDM은 ImageNet-LT 데이터셋에서 Top-1 정확도를 7.1%p 향상시키는 등 소수 클래스에 대한 분류 성능을 크게 개선했다.

텍스트 기반 분류 증강

최근에는 대규모 언어 모델(LLM)과 확산 모델을 결합하여 증강의 질을 높이는 연구가 주목받고 있다. DALDA [93]는 텍스트 프롬프트와 시각적 프롬프트(이미지)를 모두 활용하는 적응형 유도(adaptive guidance) 기법을 제안한다. 핵심은 생성 과정에서 CLIP Score를 실시간으로 모니터링하여, 생성된 이미지가 텍스트 설명과 얼마나 일치하는지를 평가하고 이를 바탕으로 유도 강도(λ)를 동적으로 조절하는 것이다. 이 접근법은 Tiny-ImageNet 소수샷(few-shot) 설정에서 4.5%p의 정확도 향상을 달성했다.

자동화된 데이터 증강

어떤 클래스를 얼마나 증강할지 결정하는 최적의 정책을 찾는 것은 어려운 문제이다. AutoGenDA [94]는 이 문제를 해결하기 위해 신경망 구조 탐색(NAS) 기법을 데이터 증강에 도입했다. 진화 탐색(Evolution Search) 알고리즘을 사용하여 각 클래스별로 얼마나 많은 합성 데이터를 추가할지, 즉 최적의 합성 비율(synthesis ratio)을 자동으로 탐색한다. 이 방식을 통해 수동적인 하이퍼파라미터 튜닝의 필요성을 없애고, Places-LT 데이터셋에서 Top-1 정확도를 6.0%p 향상시키는 성과를 거두었다.

3.2 객체 탐지 (Object Detection) 태스크

객체 탐지(Object Detection) 분야에서 확산 모델은 바운딩 박스(bounding box)와 같은 레이블 정보를 보존하면서 객체의 배경, 스타일, 포즈 등을 다양하게 변형시키는 데이터 증강에 주로 활용된다. 이러한 접근법은 특히 소수샷(few-shot) 학습이나 도메인 적응(domain adaptation)과 같이 데이터가 부족한 시나리오에서 효과적이다.

라벨 보존 혼합 증강

객체 탐지 증강의 핵심 과제 중 하나는 원본의 바운딩 박스 레이블을 훼손하지 않으면서 데이터의 다양성을 확보하는 것이다. 기존의 믹스업(Mixup) 방식은 두 이미지를 전역적으로

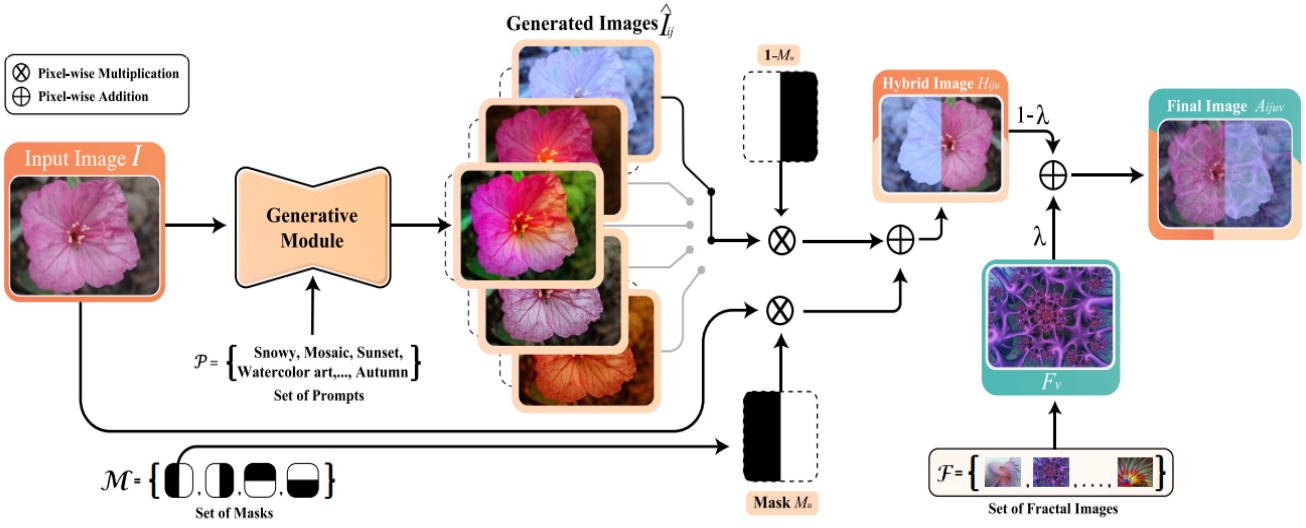

그림 10. DiffuseMix의 구조: 입력 이미지와 조건부 프롬프트를 확산 모델에 입력하여 새로운 시각적 특징을 가진 이미지를 생성한다. 이후, 이진 마스크(binary mask)를 이용해 원본 이미지의 일부와 생성된 이미지의 일부를 결합하여 하이브리드 이미지를 구성하고, 무작위 프랙탈과 혼합하여 최종 증강 이미지를 완성한다[75].

혼합하기 때문에 객체의 경계가 모호해져 레이블 보존이 어렵다. 그림 9는 이러한 단점을 극복하는 레이블 보존 혼합(label-preserving mixup) 방식의 개념을 보여준다.

이러한 원칙을 확산 모델에 적용한 대표적인 사례가 DiffuseMix [75]이다. 그림 10의 구조에서 볼 수 있듯이, DiffuseMix는 확산 모델을 이용해 원본 이미지의 일부 영역만 선택적으로 재생성하고, 이를 이진 마스크(binary mask)로 원본과 결합하여 '하이브리드 샘플'을 만든다. 비록 초기 연구는 이미지 분류에 초점을 맞추었지만, 객체와 같은 핵심 요소는 유지하면서 배경이나 주변 맥락만 효과적으로 변형시키는 이 방법론은 객체 탐지 증강에 중요한 시사점을 제공한다.

조건부 객체 합성

더 나아가, 여러 객체를 포함하는 복잡한 장면(complex scene) 전체를 합성하여 데이터셋을 구축하는 연구도 활발하다. ODGEN [95]은 바운딩 박스 레이아웃과 객체별 텍스트 설명을 조건으로 복합적인 장면 이미지를 생성하는 새로운 패러다임을 제시했다. 이 기법은 개별 객체의 특징과 전체 장면의 맥락적 일관성을 동시에 고려하여 현실적인 탐지 환경을 시뮬레이션한다. 그 결과, 7개 도메인에 걸쳐 평균 mAP를 25.3%p 향상시키는 높은 성능을 달성했다.

3D 객체 탐지

확산 모델은 2D를 넘어 3D 객체 탐지 분야로도 확장되고 있다. 3DiffTection [96]은 단일 RGB 이미지로부터 3D 객체를 탐지하기 위해 '기하학 인지 확산(Geometry-Aware Diffusion)' 기법을 제안했다. 이 방법은 확산 과정에 깊이(depth) 정보와 같은 3D 기하학적 제약 조건을 통합하여, 2D 이미지에 내포된 3차원 정보를 효과적으로 복원하고 추론한다. 이를 통해 Omni3D 데이터셋에서 3D AP 성능을 9.43점 향상시키는 성과를 거두었다.

3.3 시맨틱 세그멘테이션 (Semantic Segmentation) 태스크

시맨틱 세그멘테이션(Semantic Segmentation)은 모든 픽셀에 클래스 레이블을 할당해야 하는 고비용의 어노테이션(annotation) 작업을 요구한다. 확산 모델은 이러한 부담을 줄이고 모델 성능을 높이기 위해, 픽셀 단위의 정밀한 레이블을 보존하면서 사실적인 이미지를 생성하는 데이터 증강 도구로써 주목받고 있다. 특히 약한 감독 학습(weakly-supervised learning)과 소수샷 학습(few-shot learning) 시나리오에서 그 가치가 두드러진다.

약한 감독 학습 강화

가장 활발한 연구 방향 중 하나는 스크리블(scribble)과 같은 저비용의 약한 레이블을 활용하는 것이다[97]. ScribbleGen [98]은 소수의 실제 스크리블 주석만으로 고품질의 이미지와 그에 상응하는 완전한 세그멘테이션 마스크 쌍을 대량 생성하는 방법론으로, 그림 11은 이 전체 과정을 보여준다.

핵심은 ControlNet-Inpaint를 기반으로 스크리블을 정교한 마스크로 변환하고, 이를 다시 이미지 생성의 조건으로 활용

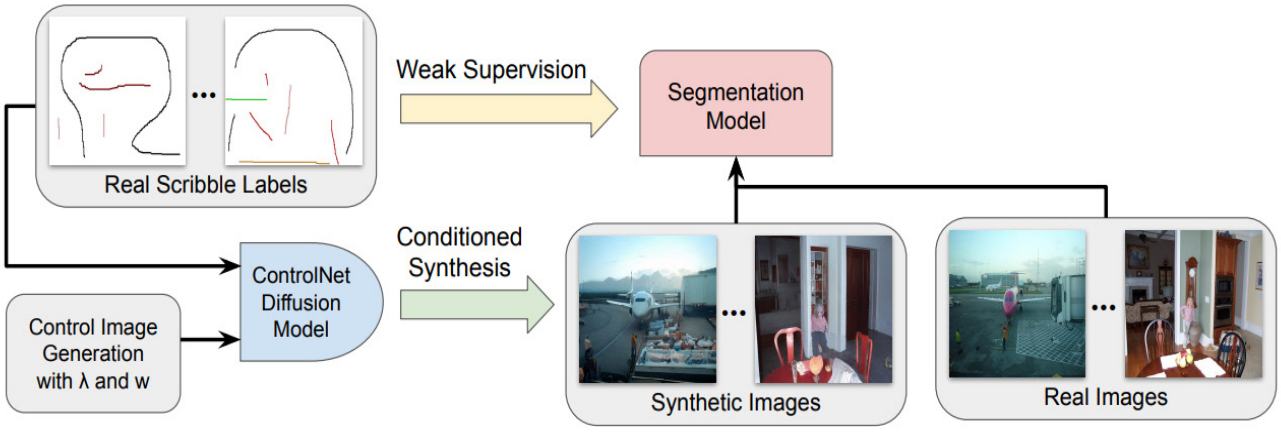

그림 11. 스크리블 주석을 활용한 약한 감독 학습: 제한된 수의 실제 스크리블(scribble)을 조건으로 ControlNet 기반의 확산 모델을 사전학습하여, 고품질의 이미지-마스크 쌍을 생성한다. 인코딩 비율(λ)과 유도 스케일(w) 등의 파라미터를 조절하여 생성 품질을 제어하며, 이렇게 생성된 데이터셋은 세그멘테이션 모델 학습에 직접 활용된다[109].

그림 12. 스크리블 조건부 이미지 생성 예시: 왼쪽은 사용자가 제공한 원본 이미지와 스크리블 레이블이다. 오른쪽은 이를 조건으로 샘플링된 다양한 합성 이미지들로, 원본의 의미적 구조를 유지하면서 사실적인 변형이 생성되었음을 보여준다. 이를 통해 저비용의 스크리블 주석만으로 풍부한 학습 데이터를 확보할 수 있다[109].

하는 것이다. 그림 12에서 볼 수 있듯이, 이 방식을 통해 원본 이미지의 의미적 구조는 유지하면서도 사실적인 시각적 변형이 가미된 다양한 합성 이미지를 생성할 수 있다. 이렇게 생성된 대량의 이미지-마스크 쌍은 세그멘테이션 모델의 학습 데이터로 직접 사용되어, Cityscapes 데이터셋에서 완전 감독 학습(fully-supervised)과의 성능 격차를 크게 줄이며 어노테이션 비용을 획기적으로 절감할 수 있는 가능성을 보여준다.

소수샷 학습을 위한 데이터 증강

데이터가 극히 제한적인 소수샷(few-shot) 환경에서는 증강의 질이 더욱 중요하다. Enhanced GDA [99]는 이러한 시나리오에 특화된 생성형 데이터 증강 시스템이다. 이 기법은 (1) 프롬프트에 클래스 정보를 명시적으로 추가하고(Class-Prompt Appending), (2) 원본 이미지의 시각적 특징을 혼합하며(Visual Prior Blending), (3) 클래스 분포를 균형 있게 조절하는(Class Balancing) 세 가지 전략을 통합한다. 이를 통해 PASCAL VOC 데이터셋의 소수샷 설정에서 기존 세그멘테이션 모델들의 mIoU를 2~7%p 향상시키는 등, 특히 희소 클래스의 성능 개선에 크게 기여했다.

합성 데이터 및 의사 마스크 생성

데이터셋 전체를 합성하는 접근법도 제안되었다. Dataset Diffusion [100]은 사전 학습된 Stable Diffusion의 주의(attention) 메커니즘을 분석하여, 별도의 레이블 없이도 이미지와 그에 상응하는 의사 세그멘테이션 마스크(pseudo-mask)를 동시에 생성한다. 이 방법으로 생성된 합성 데이터셋(synth-VOC, synth-COCO)만으로도 기존 세그멘테이션 모델을 효과적으로 학습시킬 수 있음을 입증했다.

확률적 세그멘테이션

데이터 증강을 넘어, 확산 모델 자체를 세그멘테이션을 위한 추론 프레임워크로 활용하는 연구도 등장했다. SegDiff [101]는 무작위 노이즈로부터 시작하여 점진적으로 세그멘테이션 맵을 정제해나가는 반복적 추론(iterative inference) 방식을 제안한다. 이 접근법은 확산 과정의 특성을 활용하여 세그멘테이션 결과의 불확실성을 자연스럽게 모델링할 수 있으며, 여러 벤치마크 데이터셋에서 최고 수준(state-of-the-art)의 정확도를 달성했다. 이는 확산 모델이 생성뿐만 아니라 인식(recognition) 태스크에도 직접적으로 적용될 수 있음을 보여주는 중요한 사례이다.

4. 특정 응용 분야 증강

확산 모델 기반 데이터 증강은 일반적인 컴퓨터 비전 태스크를 넘어, 특정 도메인의 고유한 문제를 해결하는 데에도 활발하게 적용되고 있다. 본 절에서는 의료, 얼굴 인식, 패션, 농업 등 주요 응용 분야에서의 활용 사례를 살펴본다.

• 의료 영상(Medical Imaging)

의료 영상 분야에서 확산 모델은 데이터 부족 문제를 해결하고 진단 정확도를 높이는 데 핵심적인 역할을 한다. 특히 희귀 질환 데이터 증강에 효과적이다. Akrout 등[102]과 Sagers 등[103]의 연구는 피부 병변 데이터셋에 클래스 조건부 확산 모델을 적용하여, 데이터가 부족한 특정 병변 유형의 합성 이미지를 생성함으로써 분류 모델의 성능을 향상시켰다.

나아가, 다양한 질병 상태를 시뮬레이션하거나 질병이 없는 '건강한' 상태의 반사실적(counterfactual) 이미지를 생성하는 데에도 활용된다. Pinaya 등[104]은 뇌 MRI 이미지를 합성하여 다양한 신경학적 상태를 가진 가상 환자 데이터를 생성했으며, Wolleb 등[105]은 DDIM을 이용해 질병이 있는 환자의 데이터로부터 건강한 상태의 이미지를 생성하여 질병의 영향을 분석하는 데 기여했다. 또한, 흉부 X-ray와 같은 이미지와 방사선 판독 보고서(radiology report)와 같은 텍스

트를 함께 활용하는 다중 모달(multi-modal) 접근법도 제안되었다[106].

• 얼굴 인식 및 편집(Facial Recognition and Editing)

얼굴 인식 및 편집 분야에서는 모델의 정확도를 높이고, 편향을 줄이며, 사실적인 편집을 위해 확산 모델이 사용된다. Boutros 등[107]과 Huang 등[108]의 연구는 데이터셋에 부족한 특정 인종이나 조명 조건의 얼굴 이미지를 고품질로 합성하여, 얼굴 인식 모델의 강건성(robustness)과 공정성(fairness)을 향상시키는 데 중점을 둔다.

• 패션 산업(Fashion Industry)

패션 산업에서는 가상 착용(virtual try-on) 경험을 혁신하는 데 확산 모델이 핵심적인 역할을 한다. Li 등[109]과 PersonaCraft [110]의 연구는 사용자의 체형과 포즈를 조건으로 의류 아이템을 자연스럽게 입혀보는 이미지를 생성한다. 이는 사용자가 온라인 쇼핑 환경에서도 실제와 같은 착용 모습을 시각화할 수 있도록 지원하여, 전자상거래의 패러다임을 바꾸고 있다.

• 농업(Agriculture)

정밀 농업(precision agriculture) 분야에서도 확산 모델의 기여가 크다. 농작물은 성장 단계나 환경에 따라 모습이 크게 달라 데이터 수집이 어렵다. Modak과 Stein [111], Huang 등[112]의 연구는 다양한 조건의 식물 질병 이미지를 합성하거나, 여러 종류의 잡초 이미지를 증강하여 탐지 모델의 정확도를 높이는 데 초점을 맞춘다. 이를 통해 병충해 조기 진단 및 효율적인 잡초 관리 시스템 구축에 기여한다.

IV. 평가 지표

확산 모델 기반 데이터 증강의 실효성을 객관적으로 검증하기 위해서는 체계적인 평가가 필수적이다. 평가는 크게 두 가지 관점에서 이루어진다: (1) 증강된 데이터가 다운스트림 태스크의 성능을 실질적으로 향상시키는지를 측정하는 정량적 평가와, (2) 생성된 이미지의 시각적 품질과 다양성을 분석하는 정성적 평가이다. 본 장에서는 이 두 가지 측면의 핵심 평가 지표를 살펴본다.

1. 정량적 평가

정량적 평가는 크게 두 가지 핵심 질문에 답하는 것을 목표로 한다. 첫째, '증강된 데이터가 최종 목표인 다운스트림 모델의 성능을 얼마나 향상시켰는가?'이며, 둘째, '생성된 이미지 자체는 얼마나 사실적이고 다양한가?'이다.

1) 다운스트림 태스크 성능 평가

증강의 가장 직접적인 효과는 최종 모델의 성능 개선으로 나타난다. 따라서 분류 태스크에서는 정확도(Accuracy), F1 점수 등이, 객체 탐지에서는 mAP (mean Average Precision)가, 그리고 시맨틱 세그멘테이션에서는 mIoU (mean Intersection over Union)가 핵심 평가지표로 사용된다. 표 4는 다양한 태스크와 데이터셋에서 확산 모델 기반 증강이 기존 방식 대비 얼마나 큰 성능 향상을 보이는지를 종합적으로 보여준다.

2) 생성 이미지 품질 및 다양성 평가

다운스트림 성능뿐만 아니라, 생성된 이미지 자체의 품질을 평가하는 것도 중요하다. 이를 위해 다음과 같은 표준 지표들이 널리 사용된다.

- **FID (Fréchet Inception Distance):** 생성된 이미지와 실제 이미지의 특징 분포 간 거리를 측정한다. 사전 학습된 Inception 모델의 특징 공간에서 계산되며, 값이 낮을수록 두 분포가 유사함을 의미하여 고품질의 지표로 간주된다[127].
- **IS (Inception Score):** 생성된 이미지가 얼마나 특정 객체의 모습을 명확하게 띠고 있으며(quality), 얼마나 다양한 클래스의 이미지가 생성되었는지(diversity)를 동시에 측정한다. 값이 높을수록 좋다[128].
- **KID (Kernel Inception Distance):** FID와 유사하게 특징 분포의 차이를 측정하지만, 소규모 샘플에 대해서도 더 안정적인 평가가 가능하다는 장점이 있다[129].
- **지각 기반 유사도(Perceptual Similarity):** LPIPS는 인간의 시각적 인식과 유사한 특징 공간에서 이미지 간 거리를 측정하며, SSIM은 명암, 대비, 구조적 유사도를 평가한다. 주로 원본과 편집된 이미지 간의 구조적 일관성을 측정하는 데 사용된다[130,131].

이처럼 다운스트림 성능과 생성 품질 지표를 함께 사용하면, 제안된 증강 기법의 효과를 다각적이고 신뢰도 높게 평가할 수 있다.

2. 정성적 평가

정량적 지표만으로는 생성된 이미지의 실질적인 유용성을 온전히 평가하기 어렵다. FID 점수가 높아도 문맥에 맞지 않거

표 1. 확산 모델 기반 데이터 증강과 기존 데이터 증강 기법의 정량적 성능 비교.

과제 (Task)	데이터셋	기존 증강 기법	DM 기반 증강 기법	기존 성능 (%)	DM 성능 (%)	성능 향상 (%p)
이미지 분류 (Top-1 정확도)	Aircraft	Guided-SR	Chen et al. [113]	77.38	84.79	+7.41
		Cutmix	Wang et al. [114]	89.44	90.25	+0.81
		CAL-AUG	Michaeli & Fried [115]	81.9	87.4	+5.50
	Caltech101	GuidedAP	Islam et al. [116]	84.32	85.84	+1.52
		RandAugment	Li et al. [117]	58.55	59.17	+0.62
		RandAugment	Fu et al. [118]	60.1	85.3	+25.20
	Cars	RandAugment	Zhang et al. [119]	57.8	65.1	+7.30
		RandAugment	Li et al. [117]	86.55	88.55	+2.00
		RandAugment	Fu et al. [118]	65.5	87.1	+21.60
	CIFAR-100	RandAugment	Zhang et al. [119]	43.2	75.7	+32.50
		Guided-SR	Chen et al. [113]	91.01	93.04	+2.03
		Cutmix	Wang et al. [114]	94.73	95.21	+0.48
		RandAugment + Cutmix	Michaeli & Fried [115]	92.7	93.8	+1.10
		Cutmix	Li et al. [117]	77.56	75.75	−1.81
	ImageNet-LT	RandAugment	Qin et al. [92]	39.6	46.7	+7.1
	Flowers	RandAugment	Li et al. [117]	41.97	45.61	+3.64
		Cutmix	Zhang et al. [119]	83.8	88.3	+4.50
	iWildCam	GuidedMixup	Wang et al. [114]	99.4	99.54	+0.14
	CUB	RandAugment	Rahat et al. [120]	76.78	85.37	+8.59
		Cutmix	Dunlap et al. [121]	77.56	84.87	+7.31
	Pets	RandAugment	Li et al. [117]	57.45	73.71	+16.26
		RandAugment	Fu et al. [118]	61.5	86.5	+25.00
		RandAugment	Zhang et al. [119]	48	73.4	+25.40
의료 이미지 분류 (Top-1 정확도)	OrganMNIST	RandAugment	Zhang et al. [119]	79.6	80.7	+1.10
	PathMNIST	RandAugment	Zhang et al. [119]	79.2	86.9	+7.70
	BreastMNIST	RandAugment	Zhang et al. [119]	68.7	77.4	+8.70
	Shenzhen TB	RandAugment	Fu et al. [118]	75.5	83.5	+8.00
OOD 분류 (OOD 정확도)	Waterbirds	RandAugment	Dunlap et al. [121]	30.32	46.63	+16.31
		Mixup	Chen et al. [113]	72.52	76.16	+3.64
		Cutmix	Wang et al. [114]	71.23	72.47	+1.24
소수샷 분류 (Top-1 정확도)	MS COCO	Standard Aug.	Trabucco et al. [122]	42	47	+5.00
		Standard Aug.	Lingenberg et al. [123]	47	57	+10.00
객체 탐지 (mAP)	PASCAL VOC	RandAugment	Li et al. [117]	78.2	79.1	+0.9
	DIOR-R	CopyPaste + Flip	Tang et al. [124]	38.75	41.69	+2.94
	COCO → CrowdHuman	Original	Chen et al. [125]	41	46.3	+5.3
시맨틱 분할 (mIoU)	PASCAL VOC7	Standard Aug.	Che et al. [126]	46.54	50.27	+3.73
	PASCAL VOC	Standard Aug.	Schnell et al. [98]	78.1	78.9	+0.80

나 의미적으로 왜곡된 이미지는 오히려 모델 성능에 악영향을 미칠 수 있다. 따라서, 인간의 시각적 판단에 기반한 정성적 평가는 증강 기법의 신뢰도를 확보하기 위한 필수적인 과정이다.

1) 시각적 품질 및 일관성 평가

정성적 평가의 첫 단계는 생성된 이미지의 시각적 품질(visual quality)을 직접 확인하는 것이다. 평가자는 주로 다음 세 가지 기준을 종합적으로 판단한다.

- **사실성(Realism):** 생성된 이미지가 실제 사진과 구별하기 어려울 정도로 자연스러운가?
- **세부 정보 보존(Detail Preservation):** 원본 이미지의 중요한 질감, 형태, 색상 등의 세부 정보가 손상되지 않고 잘 보존되었는가?
- **편집 일관성(Editing Consistency):** 이미지에 특정 편집(예: 객체 추가, 스타일 변경)을 적용했을 때, 결과물이 조화롭고 일관성이 있는가?

이러한 평가는 특히 원본의 구조적 변형이 수반되는 증강 기법에서 의미 왜곡 없이 사용자의 의도가 잘 반영되었는지를 검증하는 데 필수적이다[132-134].

2) 문맥 적합성 및 의미 일관성 평가

시각적으로 우수한 이미지라도 특정 도메인의 문맥에 부합하지 않거나 의미적으로 오류가 있다면 데이터 증강에 사용할 수 없다. 따라서 다음과 같은 항목을 추가적으로 평가해야 한다.

- **문맥 적합성(Contextual Appropriateness):** 생성된 이미지가 해당 응용 분야(예: 의료, 자율주행)의 물리적, 논리적 제약 조건에 부합하는가? 예를 들어, 의료 영상에서 해부학적으로 불가능한 구조가 생성되어서는 안 된다.
- **의미 일관성(Semantic Consistency):** 증강된 이미지가 원본의 의미론적 내용(semantic content)과 레이블을 일관되게 유지하는가? 예를 들어, '개' 이미지 증강 시 개의 품종이 바뀌거나 다른 동물처럼 보이는 왜곡이 발생해서는 안 된다.
- **작업 특화 특징 보존(Task-Specific Feature Preservation):** 다운스트림 태스크(예: 객체 탐지)의 성능에 핵심적인 특징(예: 객체의 경계선, 질감)이 증강 과정에서 보존되었는가?

이러한 정성적 평가는 생성된 데이터가 다운스트림 모델의 성능을 저해하는 '독이 든 샘플(poisoning samples)'이 되지 않도록 방지하는 중요한 안전장치 역할을 한다[132,135-138].

3. 전통적 증강 기법과의 비교 분석: 효용과 한계

정량적 및 정성적 평가를 종합하면, 확산 모델 기반 증강은 대부분의 시나리오에서 전통적인 증강 기법(예: Cutout, Mixup)의 성능을 상회한다. 특히 복잡한 장면을 이해하거나 희귀한 샘플을 생성해야 하는 까다로운 조건에서 확산 모델의 장점은 더욱 두드러진다. 전통적 기법이 기하학적 변환이나 픽셀 값 조작에 머무는 반면, 확산 모델은 데이터의 의미적 맥락(semantic context)을 이해하고 이를 바탕으로 새로운 인스턴스를 생성하기 때문이다.

그러나 이러한 높은 효용성은 막대한 계산 비용이라는 명백한 한계와 상충된다.

1) 계산 비용(Computational Cost)

가장 큰 한계는 압도적인 계산 비용이다. 고성능 GPU(H100, RTX 3090 등)를 사용하더라도 이미지 한 장을 생성하는 데 평균 0.43초에서 6.6초가 소요된다. 이는 약 0.008초가 걸리는 전통적 증강 기법과 비교하면 수십에서 수백 배에 달하는 시간이다[139-140]. 대규모 데이터셋 전체에 이를 적용할 경우, 총 증강 시간은 수십 시간에 이를 수 있으며 막대한 저장 공간이 추가로 요구된다.

2) 실패 위험성(Risk of Failure)

확산 모델은 만능이 아니다. 프롬프트의 품질이 낮거나 제어가 실패할 경우, 오히려 모델 성능을 저해하는 저품질의 이미지를 생성할 위험이 있다. 이는 데이터의 노이즈를 증가시켜 오히려 학습을 방해하는 요인으로 작용할 수 있다.

결론적으로, 확산 모델 기반 증강은 강력한 성능 향상을 제공하지만, 현실적인 적용을 위해서는 비용-효과 분석(cost-benefit analysis)이 반드시 선행되어야 한다. 제한된 자원 내에서 최대의 효과를 얻기 위해, 어떤 상황에서 전통적 기법을 사용하고 어떤 상황에서 확산 모델을 적용할지에 대한 전략적 판단이 요구된다[141].

V. 당면 과제 및 향후 연구 방향

확산 모델 기반 데이터 증강은 괄목할 만한 성과를 이루었지만, 실질적인 산업 응용으로 나아가기 위해서는 여러 당면 과제를 해결해야 한다. 본 장에서는 현재 기술의 주요 한계를 짚어보고, 이를 극복하기 위한 향후 연구 방향을 제시한다.

1) 계산 비용 및 추론 속도 문제(Computational Cost and Inference Speed)

확산 모델의 가장 큰 실용적 장벽은 막대한 계산 비용과 느린 추론 속도이다. 수백, 수천 번의 반복적인 디노이징 과정은 실시간 응용을 어렵게 만든다. 이를 해결하기 위해, 모델 경량화(model compression) [142], 지식 증류(knowledge distillation) [143], 양자화(quantization) [144] 등 모델의 효율성을 높이는 연구가 필수적이다. 또한, 반복 횟수를 줄이는 새로운 샘플링 기법[145] 개발 역시 핵심적인 연구 방향으로 남아있다.

2) 제어 가능성 및 해석력 부족(Limited Controllability and Interpretability)

현재 확산 모델은 사용자의 세밀한 의도를 완벽하게 반영하는 데 한계가 있다. 특정 객체의 위치, 크기, 스타일을 정밀하게 제어하거나, 생성 과정의 내부 메커니즘을 해석하는 것은 여전히 어려운 과제이다. 향후에는 주의(attention) 메커니즘을 활용한 국소적 편집(local editing) 기술을 고도화하고, 사용자와 실시간으로 상호작용하며 결과물을 수정해나가는 인간-AI 상호작용(Human-in-the-loop) 인터페이스 설계가 중요해질 것이다[146].

3) 생성 다양성 및 사실성의 한계(Limited Diversity and Realism)

확산 모델이 생성하는 이미지는 때때로 학습 데이터의 '평균적인' 모습(mode collapse-like behavior)에 치우쳐 실제 세계의 무한한 다양성을 온전히 담아내지 못한다. 특히 학습 데이터에 드문 '꼬리(tail)' 분포의 샘플 생성에 어려움을 겪는다. 이를 극복하기 위해, 도메인 적응(domain adaptation), 스타일 전이(style transfer), 그리고 검색 증강(retrieval-augmented) 기법을 결합하여 외부 지식을 활용하는 방향으로 연구가 발전해야 한다.

4) 과적합 및 치명적 망각(Overfitting and Catastrophic Forgetting)

소수의 데이터로 대규모 사전 학습 모델을 미세조정할 경우, 모델이 새로운 데이터에 과적합되어 기존에 학습했던 방대한 일반 지식을 잃어버리는 '치명적 망각' 현상이 발생할 수 있다[147]. 이는 모델의 범용성을 크게 저해하는 문제이다. 이를 방지하기 위해, LoRA와 같은 파라미터 효율적 미세조정(Parameter-Efficient Fine-Tuning, PEFT) 기법을 발전시키고 [148], 지속 학습(continual learning) [149] 및 메타 학습(meta-learning) [150] 접근법을 확산 모델에 맞게 적용하는 연구가 필수적이다.

5) 평가 지표 및 벤치마크의 부재(Lack of Sufficient Metrics and Benchmarks)

현재 널리 사용되는 FID, IS 등의 지표는 생성 이미지의 의미론적 적합성이나 문맥적 일관성까지는 측정하지 못하는 명백한 한계가 있다. 특히 특정 도메인(예: 의료)에서는 정량적 점수가 높아도 실제 사용할 수 없는 이미지가 생성될 수 있다. 따라서, 인간의 인지적 판단과 더 유사한 새로운 평가 지표를 개발하고[151], 다양한 시나리오와 실패 사례를 포괄하는 표준화된 벤치마크 데이터셋을 구축하는 것이 시급한 과제이다.

6) 윤리적 문제와 편향(Ethical Issues and Bias)

대규모 웹 데이터로 학습된 확산 모델은 사회적 편견, 고정관념, 유해 콘텐츠를 무분별하게 학습하고 증폭시킬 위험이 있다. 생성된 데이터의 공정성(fairness), 투명성(transparency), 책임성(accountability)을 확보하는 것은 기술의 신뢰도와 직결되는 매우 중요한 문제이다. 향후 데이터 수집 단계에서의 편향 제거, 모델 학습 과정에서의 공정성 제약 조건 추가, 그리고 생성된 콘텐츠의 출처를 명확히 하는 기술 등 '책임감 있는 AI(Responsible AI)' 원칙에 기반한 연구가 반드시 병행되어야 한다[152-155].

VI. 결론

데이터 증강은 딥러닝 모델의 일반화 성능과 강건성을 확보하기 위한 필수적인 전략이다. 그중에서도 확산 모델(Diffusion Model)은 기존의 기하학적 변환이나 픽셀 조작을 넘어, 데이터의 의미적 맥락(semantic context)을 이해하고 사실적인 이미지를 생성하는 능력 덕분에 데이터 증강의 새로운 패러다임을 제시했다. 텍스트, 레이아웃, 스타일 등 다중 모달(multi-modal) 정보를 조건으로 정교한 제어가 가능하다는 점은 확산 모델의 핵심적인 장점이다.

본고에서는 확산 모델 기반 이미지 증강에 관한 이론적 배경부터 최신 연구 동향까지를 포괄적으로 분석했다. 의미 기반 변형, 개인화 및 적응, 특정 응용 분야에 이르는 다양한 증강 방법론을 체계적으로 분류하고, 각 기법의 핵심 원리와 기

여를 기술했다. 또한, 정량적·정성적 평가 지표와 실제 성능 비교 분석을 통해 현재 기술의 효용과 한계를 명확히 하고, 이를 바탕으로 미래 연구에 필요한 통찰을 제공하고자 했다.

앞으로 확산 모델 기반 데이터 증강 기술이 한 단계 더 도약하기 위해서는 앞서 논의한 당면 과제들을 해결해야 한다. 계산 효율성 증대, 제어 가능성 및 해석력 확보, 생성 다양성 강화는 기술적 성숙도를 높이기 위한 핵심 연구 주제이다. 이와 더불어, 생성물의 품질을 신뢰도 높게 측정할 표준화된 평가 체계를 구축하고, 데이터 편향과 같은 윤리적 문제를 해결하려는 노력이 병행될 때, 확산 모델은 비로소 모든 연구자가 믿고 사용할 수 있는 실용적이고 책임감 있는(responsible) 증강 도구로 자리매김할 것이다.

사사

본 기고는 산업통상자원부(MOTIE)가 지원하고 한국산업기술평가관리원(KEIT)이 주관하는 로봇산업 원천기술개발사업의 “정밀 조립작업 대상 실환경 파라미터가 반영된 로봇용 가상환경 플랫폼 개발” (과제번호: 00419641)의 지원을 받아 수행되었습니다.

참고문헌

[1] **Dhariwal, P., & Nichol, A. (2021).** Diffusion models beat GANs on image synthesis. *Advances in Neural Information Processing Systems*, 34, 8780-8794.

[2] **Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022).** High-resolution image synthesis with latent diffusion models. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10684-10695).

[3] **Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., ... & Norouzi, M. (2022).** Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, *35*, 36479-36494.

[4] **Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022).** Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, *1*(2), 3.

[5] **Ho, J., Jain, A., & Abbeel, P. (2020).** Denoising diffusion probabilistic models. *Advances in neural information processing systems*, *33*, 6840-6851.

[6] **Kingma, D. P., & Welling, M. (2013).** Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

[7] **Ronneberger, O., Fischer, P., & Brox, T. (2015, October).** U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention* (pp. 234-241). Cham: Springer international publishing.

[8] **Song, Y., & Ermon, S. (2019).** Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, *32*.

[9] **Ho, J., & Salimans, T. (2022).** Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598.

[10] **Van Den Oord, A., & Vinyals, O. (2017).** Neural discrete representation learning. *Advances in neural information processing systems*, *30*.

[11] **Agustsson, E., Tschannen, M., Mentzer, F., Timofte, R., & Gool, L. V. (2019).** Generative adversarial networks for extreme learned image compression. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 221-231).

[12] **Peebles, W., & Xie, S. (2023).** Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 4195-4205).

[13] **Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020).** An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

[14] **Pu, Y., Xia, Z., Guo, J., Han, D., Li, Q., Li, D., ... & Li, X. (2024, September).** Efficient diffusion transformer with step-wise dynamic attention mediators. In *European Conference on Computer Vision* (pp. 424-441). Cham: Springer Nature Switzerland.

[15] **Song, Y., Zhang, Z., Lin, Z., Cohen, S., Price, B., Zhang, J., ... & Aliaga, D. (2023).** Objectstitch: Object compositing with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 18310-18319).

[16] **Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., ... & Bojanowski, P. (2023).** Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

[17] **Huang, L., Chen, D., Liu, Y., Shen, Y., Zhao, D., & Zhou, J. (2023).** Composer: Creative and controllable image synthesis with composable conditions. *arXiv preprint arXiv:2302.09778*.

[18] **Brack, M., Schramowski, P., Friedrich, F., Hintersdorf, D., & Kersting, K. (2022).** The stable artist: Steering semantics in diffusion latent space. *arXiv preprint arXiv:2212.06013*.

[19] **Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., ... & Chen, M. (2021).** Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

[20] **Yildirim, A. B., Baday, V., Erdem, E., Erdem, A., & Dundar, A. (2023).** Inst-inpaint: Instructing to remove objects with diffusion

models. *arXiv preprint arXiv:2304.03246*.

[21] **Schramowski, P., Brack, M., Deiseroth, B., & Kersting, K. (2023).** Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 22522-22531).

[22] **Pan, Y., & Bareinboim, E. (2024).** Counterfactual image editing. *arXiv preprint arXiv:2403.09683*.

[23] **Xie, S., Zhang, Z., Lin, Z., Hinz, T., & Zhang, K. (2023).** Smartbrush: Text and shape guided object inpainting with diffusion model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 22428-22437).

[24] **Han, T., Sun, W., Hu, Y., Fang, C., Zhang, Y., Ma, S., ... & Wang, Z. (2024).** Continuous concepts removal in text-to-image diffusion models. *arXiv preprint arXiv:2412.00580*.

[25] **Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., ... & Irani, M. (2023).** Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 6007-6017).

[26] **Liu, B., Wang, C., Cao, T., Jia, K., & Huang, J. (2024).** Towards understanding cross and self-attention in stable diffusion for text-guided image editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7817-7826).

[27] **Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., & Cohen-Or, D. (2022).** Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*.

[28] **Huang, N., Tang, F., Dong, W., Lee, T. Y., & Xu, C. (2023).** Region-aware diffusion for zero-shot text-driven image editing. *arXiv preprint arXiv:2302.11797*.

[29] **Brooks, T., Holynski, A., & Efros, A. A. (2023).** Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 18392-18402).

30] **Andonian, A., Osmany, S., Cui, A., Park, Y., Jahanian, A., Torralba, A., & Bau, D. (2021).** Paint by word. *arXiv preprint arXiv:2103.10951*.

[31] **Yang, B., Gu, S., Zhang, B., Zhang, T., Chen, X., Sun, X., ... & Wen, F. (2023).** Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 18381-18391).

[32] **Kirstain, Y., Levy, O., & Polyak, A. (2023).** X&fuse: Fusing visual information in text-to-image generation. *arXiv preprint arXiv:2303.01000*.

[33] **Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Zhang, Q., ... & Liu, M. Y. (2022).** ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*.

[34] **Zeng, Y., Lin, Z., Zhang, J., Liu, Q., Collomosse, J., Kuen, J., & Patel, V. M. (2023).** Scenecomposer: Any-level semantic image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 22468-22478).

[35] **Zheng, G., Zhou, X., Li, X., Qi, Z., Shan, Y., & Li, X. (2023).** Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 22490-22499).

[36] **Zhang, L., Rao, A., & Agrawala, M. (2023).** Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 3836-3847).

[37] **Xue, H., Huang, Z., Sun, Q., Song, L., & Zhang, W. (2023).** Freestyle layout-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 14256-14266).

[38] **Sarukkai, V., Li, L., Ma, A., Ré, C., & Fatahalian, K. (2024).** Collage diffusion. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 4208-4217).

[39] **Xie, S., Zhang, Z., Lin, Z., Hinz, T., & Zhang, K. (2023).** Smartbrush: Text and shape guided object inpainting with diffusion model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 22428-22437).

[40] **Xiao, G., Yin, T., Freeman, W. T., Durand, F., & Han, S. (2025).** Fastcomposer: Tuning-free multi-subject image generation with localized attention. *International Journal of Computer Vision*, *133*(3), 1175-1194.

[41] **Levin, E., & Fried, O. (2025, April).** Differential diffusion: Giving each pixel its strength. In *Computer Graphics Forum* (p. e70040).

[42] **Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J. Y., & Ermon, S. (2021).** Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*.

[43] **Xu, S., Ma, Z., Huang, Y., Lee, H., & Chai, J. (2023).** Cyclenet: Rethinking cycle consistency in text-guided diffusion for image manipulation. *Advances in Neural Information Processing Systems*, *36*, 10359-10384.

44] **Kim, G., Kwon, T., & Ye, J. C. (2022).** Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2426-2435).

[45] **Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D. J., & Norouzi, M. (2022).** Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, *45*(4), 4713-4726.

[46] **Parmar, G., Kumar Singh, K., Zhang, R., Li, Y., Lu, J., & Zhu, J. Y. (2023, July).** Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 conference proceedings* (pp. 1-11).

[47] **Trabucco, B., Doherty, K., Gurinas, M., & Salakhutdinov, R. (2023).** Effective data augmentation with diffusion models. *arXiv preprint arXiv:2302.07944*.

[48] **Cao, M., Wang, X., Qi, Z., Shan, Y., Qie, X., & Zheng, Y. (2023).** Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF*

international conference on computer vision (pp. 22560-22570).

[49] **Sanchez, P., Kascenas, A., Liu, X., O'Neil, A. Q., & Tsaftaris, S. A. (2022, September).** What is healthy? generative counterfactual diffusion for lesion localization. In *MICCAI workshop on deep generative models* (pp. 34-44). Cham: Springer Nature Switzerland.

[50] **Jeanneret, G., Simon, L., & Jurie, F. (2022).** Diffusion models for counterfactual explanations. In *Proceedings of the Asian conference on computer vision* (pp. 858-876).

[51] **Madaan, N., & Bedathur, S. (2024, April).** Navigating the Structured What-If Spaces: Counterfactual Generation via Structured Diffusion. In *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)* (pp. 710-722). IEEE.

[52] **Hastings Blow, C., Qian, L., Gibson, C., Obiomon, P., & Dong, X. (2025).** Data augmentation via diffusion model to enhance AI fairness. *Frontiers in Artificial Intelligence*, *8*, 1530397.

[53] **Le, T., Lal, V., & Howard, P. (2023).** Coco-counterfactuals: Automatically constructed counterfactual examples for image-text pairs. *Advances in Neural Information Processing Systems*, *36*, 71195-71221.

[54] **Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., & Aberman, K. (2023).** Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 22500-22510).

[55] **Ruiz, N., Li, Y., Jampani, V., Wei, W., Hou, T., Pritch, Y., ... & Aberman, K. (2024).** Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 6527-6536).

[56] **Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G., & Cohen-Or, D. (2022).** An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

57] **Zhang, Y., Dong, W., Tang, F., Huang, N., Huang, H., Ma, C., ... & Xu, C. (2023).** Prospect: Prompt spectrum for attribute-aware personalization of diffusion models. *ACM Transactions on Graphics (TOG)*, *42*(6), 1-14.

[58] **Kumari, N., Zhang, B., Zhang, R., Shechtman, E., & Zhu, J. Y. (2023).** Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1931-1941).

[59] **Sohn, K., Jiang, L., Barber, J., Lee, K., Ruiz, N., Krishnan, D., ... & Castro Chin, D. (2023).** Styledrop: Text-to-image synthesis of any style. *Advances in Neural Information Processing Systems*, *36*, 66860-66889.

[60] **Ahn, N., Lee, J., Lee, C., Kim, K., Kim, D., Nam, S. H., & Hong, K. (2024, March).** Dreamstyler: Paint by style inversion with text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 38, No. 2, pp. 674-681).

[61] **Lee, W., Lee, D., Choi, E., Yu, S., Yousefpour, A., Park, H., ... & Kim, S. (2025).** ELITE: Enhanced Language-Image Toxicity Evaluation for Safety. *arXiv preprint arXiv:2502.04757*.

[62] **Shi, J., Xiong, W., Lin, Z., & Jung, H. J. (2024).** Instantbooth: Personalized text-to-image generation without test-time finetuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8543-8552).

[63] **Jia, X., Zhao, Y., Chan, K. C., Li, Y., Zhang, H., Gong, B., ... & Su, Y. C. (2023).** Taming encoder for zero fine-tuning image customization with text-to-image diffusion models. *arXiv preprint arXiv:2304.02642*.

[64] **Chen, H., Zhang, Y., Wu, S., Wang, X., Duan, X., Zhou, Y., & Zhu, W. (2023).** Disenbooth: Identity-preserving disentangled tuning for subject-driven text-to-image generation. *arXiv preprint arXiv:2305.03374*.

[65] **Han, I., Yang, S., Kwon, T., & Ye, J. C. (2023).** Highly personalized text embedding for image manipulation by stable diffusion. *arXiv preprint arXiv:2303.08767*.

[66] **Tewel, Y., Gal, R., Chechik, G., & Atzmon, Y. (2023, July).** Key-locked rank one editing for text-to-image personalization. In *ACM SIGGRAPH 2023 conference proceedings* (pp. 1-11).

[67] **Chen, W., Hu, H., Li, Y., Ruiz, N., Jia, X., Chang, M. W., & Cohen, W. W. (2023).** Subject-driven text-to-image generation via apprenticeship learning. *Advances in Neural Information Processing Systems*, *36*, 30286-30305.

[68] **Huang, Y., Chen, Y., Liu, Y., Zhang, X., Dai, W., Xiong, H., & Tian, Q. (2024, September).** DomainFusion: Generalizing to unseen domains with latent diffusion models. In *European Conference on Computer Vision* (pp. 480-498). Cham: Springer Nature Switzerland.

[69] **Wu, Q., Liu, Y., Zhao, H., Kale, A., Bui, T., Yu, T., ... & Chang, S. (2023).** Uncovering the disentanglement capability in text-to-image diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1900-1910).

[70] **Song, K., Han, L., Liu, B., Metaxas, D., & Elgammal, A. (2022).** Diffusion guided domain adaptation of image generators. *arXiv preprint arXiv:2212.04473*.

[71] **Yang, L., Qian, H., Zhang, Z., Liu, J., & Cui, B. (2024).** Structure-guided adversarial training of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 7256-7266).

[72] **Yu, X., Li, G., Lou, W., Liu, S., Wan, X., Chen, Y., & Li, H. (2023, October).** Diffusion-based data augmentation for nuclei image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 592-602). Cham: Springer Nature Switzerland.

[73] **Chowdary, P. N., Vardhan, G. V. V., Akshay, M. S., Aashish, M. S., Aravind, V. S., Rayalu, G. V. K., & Aswathy, P. (2023, December).** Enhancing knee osteoarthritis severity level classification using diffusion augmented images. In *Proceedings of the Fourth International Conference on Advances in Computer*

Engineering and Communication Systems (ICACECS 2023) (Vol. 18, p. 266). Springer Nature.

[74] **Qiu, Z., Liu, W., Feng, H., Xue, Y., Feng, Y., Liu, Z., ... & Schölkopf, B. (2023).** Controlling text-to-image diffusion by orthogonal finetuning. *Advances in Neural Information Processing Systems*, *36*, 79320-79362.

[75] **Islam, K., Zaheer, M. Z., Mahmood, A., & Nandakumar, K. (2024).** Diffusemix: Label-preserving data augmentation with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 27621-27630).

[76] **Zhou, Y., Sahak, H., & Ba, J. (2023).** Using synthetic data for data augmentation to improve classification accuracy.

[77] **Mokady, R., Hertz, A., Aberman, K., Pritch, Y., & Cohen-Or, D. (2023).** Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 6038-6047).

[78] **Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., & Cohen-Or, D. (2022).** Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*.

[79] **Zhang, Y., Huang, N., Tang, F., Huang, H., Ma, C., Dong, W., & Xu, C. (2023).** Inversion-based style transfer with diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10146-10156).

[80] **Li, X., Hou, X., & Loy, C. C. (2024).** When stylegan meets stable diffusion: a w+ adapter for personalized image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2187-2196).

[81] **Zhang, G., Lewis, J. P., & Kleijn, W. B. (2024, September).** Exact diffusion inversion via bidirectional integration approximation. In *European Conference on Computer Vision* (pp. 19-36). Cham: Springer Nature Switzerland.

[82] **Tang, C., Wang, K., Yang, F., & van de Weijer, J. (2024).** Locinv: localization-aware inversion for text-guided image editing. *arXiv preprint arXiv:2405.01496*.

[83] **Kwon, M., Jeong, J., & Uh, Y. (2022).** Diffusion models already have a semantic latent space. *arXiv preprint arXiv:2210.10960*.

[84] **Li, Z., Zhang, W., Cechnicka, S., & Kainz, B. (2024, September).** Data-Efficient Generation for Dataset Distillation. In *European Conference on Computer Vision* (pp. 68-82). Cham: Springer Nature Switzerland.

[85] **Lv, Z., Wei, Y., Zuo, W., & Wong, K. Y. K. (2024).** Place: Adaptive layout-semantic fusion for semantic image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 9264-9274).

[86] **Gou, Y., Li, M., Zhang, Y., Zhang, X., & He, Y. (2025).** Multiple One-Shot Image Generation via Deep Structure Reshuffle. *Neural Networks*, 107862.

[87] **You, F., & Zhao, Z.** Transferring Pretrained Diffusion Probabilistic Models.

[88] **Bansal, H., & Grover, A. (2023).** Leaving reality to imagination: Robust classification via generated datasets. *arXiv preprint arXiv:2302.02503*.

[89] **Yin, Y., Kaddour, J., Zhang, X., Nie, Y., Liu, Z., Kong, L., & Liu, Q. (2023).** Ttida: Controllable generative data augmentation via text-to-text and text-to-image models. *arXiv preprint arXiv:2304.08821*.

[90] **Sheynin, S., Ashual, O., Polyak, A., Singer, U., Gafni, O., Nachmani, E., & Taigman, Y. (2022).** Knn-diffusion: Image generation via large-scale retrieval. *arXiv preprint arXiv:2204.02849*.

[91] **Blattmann, A., Rombach, R., Oktay, K., Müller, J., & Ommer, B. (2022).** Retrieval-augmented diffusion models. *Advances in Neural Information Processing Systems*, *35*, 15309-15324.

[92] **Qin, Y., Zheng, H., Yao, J., Zhou, M., & Zhang, Y. (2023).** Class-balancing diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 18434-18443).

[93] **Jung, K., Seo, Y., Cho, S., Kim, J., Min, H. S., & Choi, S. (2024, September).** Dalda: Data augmentation leveraging diffusion model and llm with adaptive guidance scaling. In *European Conference on Computer Vision* (pp. 182-200). Cham: Springer Nature Switzerland.

[94] **Cheung, T. H., & Yeung, D. Y.** AutoGenDA: Automated Generative Data Augmentation for Imbalanced Classifications.

[95] **Zhu, J., Li, S., Liu, Y. A., Yuan, J., Huang, P., Shan, J., & Ma, H. (2024).** Odgen: Domain-specific object detection data generation with diffusion models. *Advances in Neural Information Processing Systems*, *37*, 63599-63633.

[96] **Xu, C., Ling, H., Fidler, S., & Litany, O. (2024).** 3difftection: 3d object detection with geometry-aware diffusion features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10617-10627).

[97] **Zhang, X., Zhu, L., He, H., Jin, L., & Lu, Y. (2024, March).** Scribble hides class: Promoting scribble-based weakly-supervised semantic segmentation with its class label. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 38, No. 7, pp. 7332-7340).

[98] **Schnell, J., Wang, J., Qi, L., Hu, V. T., & Tang, M. (2023).** Scribblegen: generative data augmentation improves scribble-supervised semantic segmentation. *arXiv preprint arXiv:2311.17121*.

[99] **Che, Q. H., Le, D. T., Pham, B. N., Lam, D. K., & Nguyen, V. T. (2024).** Enhanced Generative Data Augmentation for Semantic Segmentation via Stronger Guidance. *arXiv preprint arXiv:2409.06002*.

[100] **Nguyen, Q., Vu, T., Tran, A., & Nguyen, K. (2023).** Dataset diffusion: Diffusion-based synthetic data generation for pixel-level semantic segmentation. *Advances in Neural Information Processing Systems*, *36*, 76872-76892.

[101] **Amit, T., Shaharbany, T., Nachmani, E., & Wolf, L. (2021).** Segdiff: Image segmentation with diffusion probabilistic models. *arXiv preprint arXiv:2112.00390*.

[102] **Akrout, M., Gyepesi, B., Holló, P., Poór, A., Kincső, B., Solis, S., ... & Fazekas, I. (2023, October).** Diffusion-based data augmentation for skin disease classification: Impact across original medical datasets to fully synthetic images. In *International conference on medical image computing and computer-assisted intervention* (pp. 99-109). Cham: Springer Nature Switzerland.

[103] **Sagers, L. W., Diao, J. A., Melas-Kyriazi, L., Groh, M., Rajpurkar, P., Adamson, A. S., ... & Manrai, A. K. (2023).** Augmenting medical image classifiers with synthetic data from latent diffusion models. *arXiv preprint arXiv:2308.12453*.

[104] **Pinaya, W. H., Tudosiu, P. D., Dafflon, J., Da Costa, P. F., Fernandez, V., Nachev, P., ... & Cardoso, M. J. (2022, September).** Brain imaging generation with latent diffusion models. In *MICCAI workshop on deep generative models* (pp. 117-126). Cham: Springer Nature Switzerland.

[105] **Wolleb, J., Bieder, F., Sandkühler, R., & Cattin, P. C. (2022, September).** Diffusion models for medical anomaly detection. In *International Conference on Medical image computing and computer-assisted intervention* (pp. 35-45). Cham: Springer Nature Switzerland.

[106] **Zhan, C., Lin, Y., Wang, G., Wang, H., & Wu, J. (2024).** Medm2g: Unifying medical multi-modal generation via cross-guided diffusion with visual invariant. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11502-11512).

[107] **Boutros, F., Grebe, J. H., Kuijper, A., & Damer, N. (2023).** Idiff-face: Synthetic-based face recognition through fizzy identity-conditioned diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 19650-19661).

[108] **Huang, Z., Chan, K. C., Jiang, Y., & Liu, Z. (2023).** Collaborative diffusion for multi-modal face generation and editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 6080-6090).

[109] **Li, Z., Wei, P., Yin, X., Ma, Z., & Kot, A. C. (2023).** Virtual try-on with pose-garment keypoints guided inpainting. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 22788-22797).

[110] **Kim, G., Jeon, S. Y., Lee, S., & Chun, S. Y. (2024).** PersonaCraft: Personalized Full-Body Image Synthesis for Multiple Identities from Single References Using 3D-Model-Conditioned Diffusion. *arXiv e-prints*, arXiv-2411.

[111] **Modak, S., & Stein, A. (2024, September).** Enhancing weed detection performance by means of genai-based image augmentation. In *European Conference on Computer Vision* (pp. 252-266). Cham: Springer Nature Switzerland.

[112] **Huang, J., Wang, Z., Xu, M., Ma, L., Wu, W., & Cao, J. (2024, November).** Enhancing Few-Shot Plant Disease Classification with Diffusion Model. In *2024 China Automation Congress (CAC)* (pp. 2036-2040). IEEE.

[113] **Chen, X., Huang, L., Liu, Y., Shen, Y., Zhao, D., & Zhao, H. (2024).** Anydoor: Zero-shot object-level image customization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 6593-6602).

[114] **Wang, Z., Wei, L., Wang, T., Chen, H., Hao, Y., Wang, X., ... & Tian, Q. (2024).** Enhance image classification via inter-class image mixup with diffusion model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 17223-17233).

[115] **Michaeli, E., & Fried, O. (2024).** Advancing fine-grained classification by structure and subject preserving augmentation. *Advances in Neural Information Processing Systems*, *37*, 22316-22349.

[116] **Islam, K., Zaheer, M. Z., Mahmood, A., Nandakumar, K., & Akhtar, N. (2024).** Genmix: effective data augmentation with generative diffusion model image editing. *arXiv preprint arXiv:2412.02366*.

[117] **Li, B., Xu, X., Wang, X., Hou, Y., Feng, Y., Wang, F., ... & Che, W. (2024, March).** Semantic-guided generative image augmentation method with diffusion models for image classification. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 38, No. 4, pp. 3018-3027).

[118] **Fu, Y., Chen, C., Qiao, Y., & Yu, Y. (2024).** Dreamda: Generative data augmentation with diffusion models. *arXiv preprint arXiv:2403.12803*.

[119] **Zhang, Y., Zhou, D., Hooi, B., Wang, K., & Feng, J. (2023).** Expanding small-scale datasets with guided imagination. *Advances in neural information processing systems*, *36*, 76558-76618.

[120] **Rahat, F., Hossain, M. S., Ahmed, M. R., Jha, S. K., & Ewetz, R. (2025, February).** Data augmentation for image classification using generative ai. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (pp. 4173-4182). IEEE.

[121] **Dunlap, L., Umino, A., Zhang, H., Yang, J., Gonzalez, J. E., & Darrell, T. (2023).** Diversify your vision datasets with automatic diffusion-based augmentation. *Advances in neural information processing systems*, *36*, 79024-79034.

[122] **Trabucco, B., Doherty, K., Gurinas, M., & Salakhutdinov, R. (2023).** Effective data augmentation with diffusion models. *arXiv preprint arXiv:2302.07944*.

[123] **Lingenberg, T., Reuter, M., Sudhakaran, G., Gojny, D., Roth, S., & Schaub-Meyer, S. (2024).** Diagen: diverse image augmentation with generative models. *arXiv e-prints*, arXiv-2408.

[124] **Tang, D., Cao, X., Wu, X., Li, J., Yao, J., Bai, X., ... & Meng, D. (2025).** AeroGen: Enhancing remote sensing object detection with diffusion-driven data generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference* (pp. 3614-3624).

[125] **Chen, S., Sun, P., Song, Y., & Luo, P. (2023).** Diffusiondet: Diffusion model for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 19830-19843).

[126] **Che, Q. H., Le, D. T., Pham, B. N., Lam, D. K., & Nguyen, V. T. (2024).** Enhanced Generative Data Augmentation for Semantic Segmentation via Stronger Guidance. *arXiv preprint arXiv:2409.06002*.

[127] **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. *Advances in neural information processing systems*, *30*.

[128] **Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016).** Improved techniques for training gans. *Advances in neural information processing systems*, *29*.

[129] **Bińkowski, M., Sutherland, D. J., Arbel, M., & Gretton, A. (2018).** Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*.

[130] **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* (pp. 586-595).

[131] **Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004).** Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, *13*(4), 600-612.

[132] **Hartwig, S., Engel, D., Sick, L., Kniesel, H., Payer, T., Poonam, P., ... & Ropinski, T. (2025).** A survey on quality metrics for text-to-image generation. *IEEE Transactions on Visualization and Computer Graphics*.

[133] **Aziz, M., Rehman, U., Danish, M. U., & Grolinger, K. (2025).** Global-local image perceptual score (glips): Evaluating photorealistic quality of ai-generated images. *IEEE Transactions on Human-Machine Systems*.

[134] **Chen, E. M., Holalkere, S., Yan, R., Zhang, K., & Davis, A. (2023).** Ray conditioning: Trading photo-consistency for photo-realism in multi-view image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 23242-23251).

[135] **Mersha, M. A., Yigezu, M. G., Tonja, A. L., Shakil, H., Iskander, S., Kolesnikova, O., & Kalita, J. (2025).** Explainable AI: XAI-guided context-aware data augmentation. *Expert Systems with Applications*, *289*, 128364.

[136] **Doğan, O., Altıntaş, A., Yücetürk, B., Aydın, D., Soygazi, F., & Kılıçaslan, Y. (2025).** Exploring Semantic Consistency in Generative Artificial Intelligence via Text-to-Image and Image-to-Text Transformation. *Journal of Artificial Intelligence and Data Science*, *5*(1), 53-62.

[137] **Gong, C., Wang, D., Li, M., Chandra, V., & Liu, Q. (2021).** Keepaugment: A simple information-preserving data augmentation approach. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1055-1064).

[138] **De Gaspari, F., Hitaj, D., & Mancini, L. V. (2024, September).** Have you poisoned my data? defending neural networks against data poisoning. In *European Symposium on Research in Computer Security* (pp. 85-104). Cham: Springer Nature Switzerland.

[139] **Zheng, H., Nie, W., Vahdat, A., & Anandkumar, A. (2023).** Fast training of diffusion models with masked transformers. *arXiv preprint arXiv:2306.09305*.

[140] **Mumuni, A., & Mumuni, F. (2022).** Data augmentation: A comprehensive survey of modern approaches. *Array*, *16*, 100258.

[141] **Agia, C., Sinha, R., Yang, J., Cao, Z. A., Antonova, R., Pavone, M., & Bohg, J. (2024).** Unpacking failure modes of generative policies: Runtime monitoring of consistency and progress. *arXiv preprint arXiv:2410.04640*.

[142] **Kim, G., Kim, B., Park, E., & Cho, S. (2024).** Diffusion Model Compression for Image-to-Image Translation. In *Proceedings of the Asian Conference on Computer Vision* (pp. 2105-2123).

[143] **Huang, T., Zhang, Y., Zheng, M., You, S., Wang, F., Qian, C., & Xu, C. (2023).** Knowledge diffusion for distillation. *Advances in Neural Information Processing Systems*, *36*, 65299-65316.

[144] **Li, X., Liu, Y., Lian, L., Yang, H., Dong, Z., Kang, D., ... & Keutzer, K. (2023).** Q-diffusion: Quantizing diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 17535-17545).

[145] **Zheng, H., Nie, W., Vahdat, A., Azizzadenesheli, K., & Anandkumar, A. (2023, July).** Fast sampling of diffusion models via operator learning. In *International conference on machine learning* (pp. 42390-42402). PMLR.

[146] **Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2022).** Lora: Low-rank adaptation of large language models. *ICLR*, *1*(2), 3.

[147] **Zhang, B., Luo, C., Yu, D., Li, X., Lin, H., Ye, Y., & Zhang, B. (2024, March).** Metadiff: Meta-learning with conditional diffusion for few-shot learning. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 38, No. 15, pp. 16687-16695).

[148] **Liang, J., Zhong, J., Gu, H., Lu, Z., Tang, X., Dai, G., ... & Yang, Q. (2024, September).** Diffusion-driven data replay: A novel approach to combat forgetting in federated class continual learning. In *European Conference on Computer Vision* (pp. 303-319). Cham: Springer Nature Switzerland.

[149] **Pan, J., Gao, H., Wu, Z., Hu, T., Su, L., Huang, Q., & Li, L. (2024).** Leveraging catastrophic forgetting to develop safe diffusion models against malicious finetuning. *Advances in Neural Information Processing Systems*, *37*, 115208-115232.

[150] **Ni, F., Hao, J., Mu, Y., Yuan, Y., Zheng, Y., Wang, B., & Liang, Z. (2023, July).** Metadiffuser: Diffusion model as conditional planner for offline meta-rl. In *International Conference on Machine Learning* (pp. 26087-26105). PMLR.

[151] **Peng, Y., Cui, Y., Tang, H., Qi, Z., Dong, R., Bai, J., ... & Xia, S. T. (2024).** Dreambench++: A human-aligned benchmark for personalized image generation. *arXiv preprint arXiv:2406.16855*.

[152] **Birhane, A., Kalluri, P., Card, D., Agnew, W., Dotan, R., & Bao, M. (2022, June).** The values encoded in machine learning research. In *Proceedings of the 2022 ACM conference on fairness,*

accountability, and transparency (pp. 173-184).

[153] **Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March).** On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610-623).

[154] **Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P. S., ... & Gabriel, I. (2021).** Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

[155] **Shelby, R., Rismani, S., Henne, K., Moon, A., Rostamzadeh, N., Nicholas, P., ... & Virk, G. (2023, August).** Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 723-741).

정 다 예 나

2021년~2025년 서울과학기술대학교 인공지능응용학과 (학사)

2025년~현재 서울과학기술대학교 국방인공지능응용학과 (석사)

오 시 영

2021년~2025년 서울과학기술대학교 컴퓨터공학과 (학사)

최 성 록

2001년~2006년 서울대학교 기계항공공학과 (공학사)

2006년~2008년 KAIST 로봇공학학제전공 (공학석사)

2014년~2019년 KAIST 로봇공학학제전공 (공학박사)

2008년~2020년 ETRI 지능로보틱스연구본부 (선임연구원)

2021년~현재 서울과학기술대학교 컴퓨터공학과 (조교수)