일반논문 (Regular Paper)
방송공학회논문지 제30권 제2호, 2025년 3월 (JBE Vol.30, No.2, March 2025)
https://doi.org/10.5909/JBE.2025.30.2.179
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

적응적 FreeU: Diffusion 기반 이미지 생성 모델의 무학습 성능 향상 방법

유 장 현[a)]*, 조 현 동[a)]*, 최 승 미[a)], 배 성 준[b)], 김 휘 용[a)‡]

Adaptive FreeU: Improving Performance of Diffusion-based Image Generation Models without Training

Janghyun Yu[a)]*, HyunDong Cho[a)]*, Seungmi Choi[a)], Seong-Jun Bae[b)], and Hui Yong Kim[a)‡]

요 약

본 논문에서는 최근 우수한 생성 영상의 성능을 보이는 확산 모델의 U-Net 구조를 개선하기 위한 새로운 방법인 적응적 FreeU를 제안한다. 기존의 FreeU는 백본 특징과 스킵 특징에 입력과 무관한 고정된 스케일링 팩터를 일괄 적용하여 생성 품질을 높였으나, 입력에 따른 중요도를 반영하지 못한다는 한계가 있었다. 이를 보완하기 위해 적응적 FreeU는 채널별 엔트로피 분석을 통해 백본 특징을 동적으로 증폭하고, 스킵 특징에서 저주파수 정보의 에너지를 정량화하여 과도한 저주파수 성분을 억제함으로써 오버스무딩 현상을 방지한다. Stable Diffusion 2.1과 Stable Diffusion XL으로 실험한 결과, 기존 FreeU 대비 더욱 개선된 생성 영상의 품질을 보였고, 객관적인 평가 지표(FID, CLIP-Score)에서도 성능 향상을 달성하였다.

Abstract

In this paper, we propose an adaptive FreeU method designed to improve the U-Net architecture of diffusion models, which have recently demonstrated excellent performance in image generation. While FreeU enhances overall generation quality by applying a single fixed scaling factor to both backbone and skip features regardless of the input, it fails to account for variations in importance across different inputs. To address this limitation, the proposed adaptive FreeU dynamically amplifies the backbone features through channel-wise entropy analysis and quantitatively assesses the energy of low-frequency information in the skip features to suppress excessive low-frequency components, thereby preventing oversmoothing. Experimental results on Stable Diffusion 2.1 and Stable Diffusion XL show that our method yields higher-quality generated images compared to FreeU, and also achieves performance gains on objective evaluation metrics such as FID and CLIP-Score.

Keyword : Diffusion, Stable Diffusion, FreeU, U-Net, Text-to-Image Generation

a) 경희대학교 컴퓨터공학과(Kyung Hee University, Department of Computer Engineering)
b) 한국전자통신연구원(ETRI)
*) 공동 1저자
‡ Corresponding Author : 김휘용(Hui Yong Kim)
E-mail: hykim.v@khu.ac.kr
Tel: +82-31-201-3760
ORCID: https://orcid.org/0000-0001-7308-133X
※ 이 논문은 한국전자통신연구원 내부연구개발사업 차세대 미디어 부호화 및 전송 표준 원천기술 개발(24BC1200/24ZC1500)의 논문입니다.
※ 이 논문은 2025년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.RS-2022-00155911, 인공지능융합혁신인재양성(경희대학교)).
· Manuscript January 1, 2025; Revised February 13, 2025; Accepted February 17, 2025.

Ⅰ. 서 론

확산 모델(Diffusion Model)은 최근 몇 년간 이미지 생성, 텍스트-이미지 변환, 텍스트-비디오 변환 등에서 뛰어난 생성 능력을 보여왔다. DDPM(Denoising Diffusion Probabilistic Models)[3]을 시작으로 다양한 변형 모델이 등장하였다. DDPM 이후로 Stable Diffusion[4][5], DreamBooth[6], ModelScope[7] 등 텍스트-이미지와 텍스트-비디오 변환에서 우수한 생성 능력을 보이는 모델이 다수 등장하였다. 이들 모델의 핵심 구조로는 U-Net[1]이 널리 활용되며, 특히 가우시안 노이즈를 제거(Denoising)하는 데 있어서 중요한 역할을 맡고 있다. 하지만 U-Net에서 백본 특징(Backbone Feature)과 스킵 특징(Skip Feature) 간의 정보 전달 특성 때문에, 필요 이상의 고주파수 정보가 강조되거나, 저주파수 정보가 부족해지는 문제가 종종 발생하여 확산 모델의 생성 품질에 악영향을 미칠 수 있다. 이러한 문제를 개선하기 위해, FreeU[2]는 추가 학습 없이 확산 모델의 성능을 개선할 수 있는 방법을 제안하였다. FreeU는 백본 특징을 강화하기 위해 백본 특징 스케일링 팩터를 사용하고, 백본 특징과 스킵 특징 사이에서 중복되는 저주파수 정보의 양을 조절하기 위해 스킵 특징 스케일링 팩터를 적용하여 고주파수와 저주파수 정보를 균형 있게 조절한다. 이를 통해 추가 학습이나 추가 파라미터 없이 기존 확산 모델의 성능을 높였다.

그러나 FreeU는 두 스케일링 팩터를 입력과 무관하게 네트워크 레이어 전체에 일괄적으로 적용하며 스케일링 팩터의 값을 실험적으로 설정해야 하는 한계가 있다. 이는 입력에 따라 정보량이나 중요도가 달라질 수 있는 실제 특성을 충분히 반영하지 못하고, 최적의 값을 찾기 위한 많은 실험을 해야만 하는 단점이 있다. 본 논문에서는 이러한 한계를 개선하기 위해 적응적 FreeU를 제안한다. 적응적 FreeU는 먼저 채널별 엔트로피를 계산하여 각 채널의 정보량에 비례하도록 백본 특징 스케일링 팩터를 적응적으로 결정하여 백본 특징을 증폭한다. 또한 스킵 특징의 전체 주파수 정보량 대비 저주파수 정보량을 고려하여 중복되는 저주파수 정보를 억제함으로써 오버스무딩 현상을 방지한다. 이를 통해 추가 학습 없이 확산 모델의 성능을 높이는 FreeU의 기본 원칙을 따르면서, FreeU보다 더 우수한 생성 품질을 객관적, 주관적으로 달성하였다.

본 논문의 Ⅱ장에서는 확산 모델의 핵심 구조인 U-Net을 소개하고, 추가 학습 없이 기존 확산 모델의 성능을 개선한 FreeU에 대해 설명한다. Ⅲ장에서는 FreeU를 개선하여 제안한 적응적 FreeU의 동작 원리와 구현 방법을 설명한다. Ⅳ장에서는 제안 방법의 유효성을 검증하고, 객관적 성능 지표와 주관적 품질을 제시한다. 마지막으로 Ⅴ장에서는 본 논문의 결과를 요약하고 향후 연구 방향에 대해 논의한다.

Ⅱ. 관련 연구

1. U-Net[1]

확산 모델은 일반적으로 가우시안 노이즈를 점진적으로 추가하고 이를 디노이징하는 과정을 통해 고품질 영상을 생성하는 방식을 따른다. 이러한 모델에서 U-Net 구조는 가우시안 노이즈를 점진적으로 제거하여, 텍스트 정보를 반영한 영상을 생성하는 데 핵심적인 역할을 한다.

U-Net은 입력을 다운샘플링하여 특징을 추출하는 인코더와 추출 특징을 업샘플링하여 입력 크기로 복원하는 디코더로 구성된다. 이 때 인코더와 디코더 사이의 스킵 연결을 통해 다운샘플링 과정에서 손실되는 정보를 보완한다. U-Net은 인코더와 디코더, 스킵 연결을 통해 다양한 해상도에서 특징을 추출하고 결합할 수 있는 구조를 갖추어 디노이징을 효과적으로 수행할 수 있기 때문에 확산 모델에서 주로 사용된다. 이 때 디코더에서 생성하는 특징을 백본 특징, 스킵 연결을 통해 전달되는 특징을 스킵 특징이라고 할 수 있다.

확산 모델은 DDPM(Denoising Diffusion Probabilistic Models)[3]을 필두로 하여 다양한 변형 모델이 등장하였다. DDPM 이후로 Stable Diffusion[4][5], DreamBooth[6], ModelScope[7] 등 텍스트-이미지와 텍스트-비디오에서 우수한 생성 능력을 보이는 모델이 다수 등장하였다. 특히 Stable Diffusion은 잠재 공간에서 확산 과정을 수행함으로써 고해상도 이미지를 효율적으로 생성하였다. 본 논문에서는 Stable Diffusion의 두 가지 버전[4][5]에 대해 실험하였다.

2. FreeU[2]

FreeU는 확산 모델의 성능을 향상시키기 위해 U-Net(그림 1 (a))의 백본 특징(그림 1 (b)의 Backbone features)과

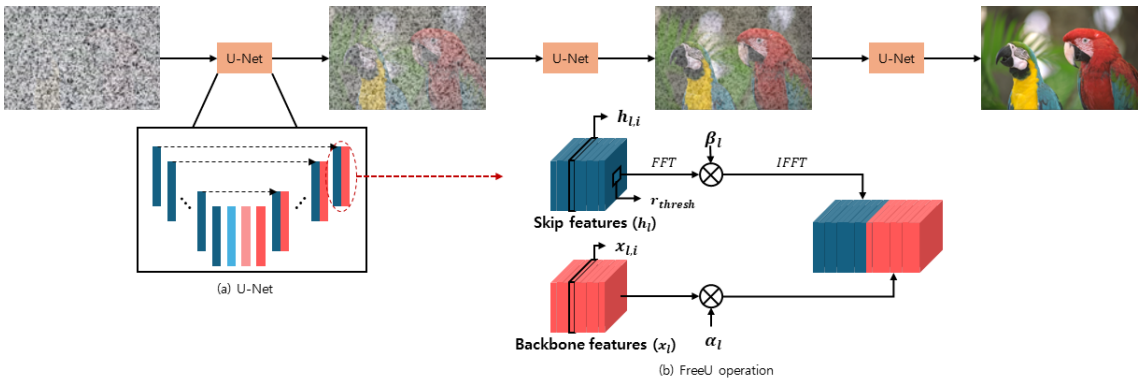


그림 1. 확산 모델에서 (a) U-Net 구조와 (b) FreeU에서 스케일링 팩터를 백본 특징과 스킵 특징에 각각 적용한 동작 구조[2]
Fig. 1. In the diffusion model: (a) the U-Net architecture, and (b) the operation structure in FreeU where the scaling factor is applied to both backbone features and skip features[2]

스킵 특징(그림 1 (b)의 Skip features)의 기여도를 조정하는 방식을 제안했다. 백본 특징은 주로 저주파수 정보를 포함하여 디노이징에 중요한 역할을 하며, 스킵 특징은 고주파수 정보를 중심으로 생성 영상의 세밀한 디테일을 살리는 역할을 한다. 이를 기반으로 FreeU는 백본 특징을 증폭하여 디노이징의 효과를 강화하고, 오버스무딩 현상을 방지하기 위해 스킵 특징의 저주파수 정보를 억제하여 중복된 정보를 줄이는 동시에 중요한 고주파수 정보만을 전달하도록 설계되었다. 이러한 방법을 통해, FreeU는 추가 학습이나 추가 파라미터 없이도 추론 과정만으로 생성 영상의 품질을 효과적으로 향상시켰다.

식 (1)은 백본 특징을 증폭하기 위한 식을 의미한다. $x_{l,i}$는 U-Net 디코더의 l번째 블록의 백본 특징의 i번째 채널(c)를 의미한다. 즉, 채널의 절반에 대해 하이퍼 파라미터로 정한 백본 스케일링 팩터 α_l을 $x_{l,i}$에 픽셀 간 곱($\odot$)을 하여 백본 특징을 증폭한다.

$$x'_{l,i} = \begin{cases} x_{l,i} \odot \alpha_l & if\ i < c/2 \\ x_{l,i} & otherwise \end{cases} \quad (1)$$

식 (2), 식 (3), 식 (4), 식 (5)는 스킵 특징의 저주파수 정보를 억제하는 식을 의미한다. $h_{l,i}$는 U-Net 디코더의 l번째 블록의 스킵 특징의 i번째 채널(c)의 값을 의미한다. FFT와 IFFT는 각각 푸리에 변환과 역변환을 의미하며, $\beta_{l,i}$는 식 (5)에서 정의하여 $h_{l,i}$의 주파수 영역의 값에 픽셀 간 곱($\odot$)을 위해 쓰인다. 식 (5)에서 r은 특정 픽셀 위치에서 주파수 공간 내 거리(반지름)을 나타내며, r_{thresh}는 하이퍼 파라미터로 정의한 임계치 값이다. 즉, 주파수 공간에서의 반지름 r이 하이퍼 파라미터로 정의한 반지름 임계치 r_{thresh}를 초과한 경우, 하이퍼 파라미터로 정의한 스킵 스케일링 팩터 s_l을 $h_{l,i}$의 주파수 영역의 값에 픽셀 단위로 곱하여 값을 조정한다.

$$\mathcal{F}(h_{l,i}) = FFT(h_{l,i}) \quad (2)$$

$$\mathcal{F}'(h_{l,i}) = \mathcal{F}(h_{l,i}) \odot \beta_{l,i} \quad (3)$$

$$h'_{l,i} = IFFT\left(\mathcal{F}'(h_{l,i})\right) \quad (4)$$

$$\beta_{l,i}(r) = \begin{cases} s_l & if\ r < r_{thresh} \\ 1 & otherwise \end{cases} \quad (5)$$

하지만, FreeU는 백본 특징을 증폭하고 스킵 특징의 저주파수 정보를 억제하기 위한 값을 입력과 무관하게 일괄적으로 적용한다는 한계가 있다. 구체적으로, 백본 특징은 모든 레이어의 채널 절반에 대해 동일한 값을 통해 증폭되며, 스킵 특징은 주파수 영역에서 특정 저주파수 영역에 대해 모든 레이어에 동일한 값을 통해 억제된다. 이러한 접근 방식은 입력의 특성과 중요도를 전혀 고려하지 않기 때문에, 채널별로 정보량이나 중요도가 다를 수 있는 실제 상황을 제대로 반영하지 못한다. 결과적으로, 고정된 값을 사용하는 것은 다양한 입력에 최적화된 성능을 발휘하기 어려울 수 있으며, 더 나아가 최적의 값을 찾기 위해 많은 실험적 노력을 필요로 한다는 단점이 존재한다.

Ⅲ. 제안 방법

1. 제안 방법 개요

본 논문에서 제안하는 적응적 FreeU는 기존 FreeU[2]와 마찬가지로 U-Net[1]에서 백본 특징 스케일링 팩터와 스킵 특징 스케일링 팩터를 각각 조정한다. 그러나, 고정된 값을 사용하는 대신, 각 특징이 가진 정보량에 따라 동적으로 스케일링 팩터를 할당하는 방법을 제안한다. 이를 통해 입력의 특성과 중요도를 반영할 수 있도록 한다. 제안 방법의 구조는 그림 2와 같다.

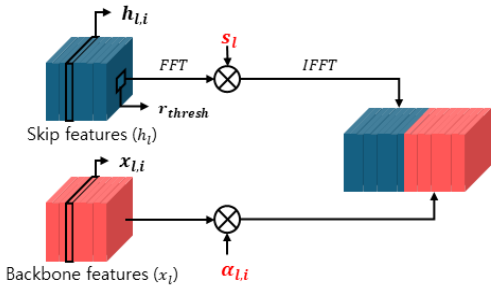


그림 2. 제안 방법의 구조
Fig. 2. The architecture of proposed method

2. 백본 특징 조절

정보량이 많은 채널은 중요한 특징을 더 잘 표현할 가능성이 높다. 따라서 채널별로 정보의 양을 정량적으로 측정하고, 이를 기반으로 각 채널에 가중치를 부여함으로써 채널의 중요도를 반영할 수 있다. 우리는 데이터 분포의 불확실성을 나타내는 엔트로피를 활용하여 채널의 중요도를 계산하였다. 구체적으로, 네트워크 각 레이어의 출력 채널에 대해 엔트로피를 계산하고, 이를 통해 해당 채널의 가중치를 동적으로 결정하였다. 엔트로피의 값이 클수록 다양한 정보를 포함하고 있기 때문에 해당 채널에는 더 높은 가중치를 부여하여 중요한 특징이 더욱 강조되도록 설계하였다. 실제 구현 시에는 채널 내의 픽셀 값의 분포를 확률적 해석하기 어렵기 때문에, 픽셀 값의 히스토그램 기반으로 근사한 확률을 사용하였다.

$$p_{l,i}[k] = \frac{x_{l,i}[k]}{\sum_{j=1}^{N} x_{l,i}[j]}, for\ k = 1,2,\dots,N \tag{6}$$

$$H_{l,i} = -\sum_{k=1}^{N} p_{l,i}[k] \times \log_2(p_{l,i}[k]) \tag{7}$$

$$\alpha_{l,i} = 1 + \sigma(H_{l,i}) \times (b-1) \tag{8}$$

$$x'_{l,i} = x_{l,i} \odot \alpha_{l,i}, \quad for\ all\ channels \tag{9}$$

식 (6)에서는 U-Net 디코더의 l번째 블록에서 각 i번째 채널에 대해 픽셀 값의 히스토그램을 계산하고, 이를 기반으로 확률 값을 계산한다. 이 과정에서 N개의 빈(bin)을 사용하여 히스토그램을 계산하게 된다. 식 (7)에서는 식 (6)에서 구한 확률 값을 기반으로 엔트로피를 계산하고, 이를 1에서 2 사이의 값으로 정규화하기 위해 식 (8)을 사용하여 추가적으로 처리하였다. 과도하게 백본 특징의 값이 증폭되어 오버스무딩되는 현상을 방지하고자 최댓값 b를 하이퍼 파라미터로 정의하였다. σ는 시그모이드 함수를 의미한다. 최종적으로 식 (9)와 같이 백본 특징의 모든 채널에 대해 $\alpha_{l,i}$만큼 증폭시킨다.

정리하면, 각 채널의 중요도를 계산하여 중요도에 따라 동적으로 채널에 가중치를 할당한 후, 값의 최댓값을 제한함으로써 과도하게 채널 값이 증폭되는 것을 방지하고, 이를 통해 백본 특징을 조절하였다.

3. 스킵 특징 조절

[2]에 따르면, 오버스무딩 현상을 방지하기 위해 스킵 특징에서 중요한 고주파수 정보 위주로 전달하고, 저주파수 정보를 억제하는 방법을 제안하였다. 이에 따라 본 논문에서는 스킵 특징의 저주파수 정보의 양을 정량적으로 측정하고, 이를 스케일링 팩터에 반영하는 방식을 추가적으로 도입하였다.

식 (10)에서는 U-Net 디코더의 l번째 스킵 특징의 i번째 채널의 값을 의미하는 $h_{l,i}$를 주파수 변환하여, 특정 반지름 내의 값을 LF(Low Frequency, 저주파수)라고 정의한다. 저주파수 영역의 에너지(LF Energy)는 식 (10)에서 정의한 LF 값들의 진폭의 합을 의미하며, 이때 계산되는 픽셀의 범위는 식 (10)에서 정의한 반지름 내의 값으로 제한된다. 전체 에너지(Total Energy)는 스킵 특징의 모든 픽셀에 대한 진폭의 합을 의미한다. 이를 통해 FreeU에서는 하이퍼 파라미터로 정의된 식 (5)의 s_l을 식 (13)에서 볼 수 있듯이 동적인 값으로 재정의하였다.

$$LF = \mathcal{F}(h_{l,i}), if\ r < r_{thresh} \tag{10}$$

$$LF\ Energy = \sum_{all\ pixels} |LF| \tag{11}$$

$$Total\ Energy = \sum_{all\ pixels} |\mathcal{F}(h_{l,i})| \tag{12}$$

$$s_l = 1 - \frac{LF\ Energy}{Total\ Energy} \quad (13)$$

구체적으로, 스킵 특징 $h_{l,i}$의 주파수 영역에서의 값을 기반으로 각 주파수 영역의 에너지를 계산하고, 저주파수 영역의 에너지의 비율을 계산한다. 이 비율을 스킵 스케일링 팩터로 활용하여 저주파수 정보의 기여도를 조절함으로써, 스킵 특징 내의 중복되는 저주파수 정보를 억제하면서, 고주파수 정보를 보다 효과적으로 전달할 수 있도록 한다.

Ⅳ. 실험

1. 실험 상세 사항

적응적 FreeU의 성능을 평가하기 위해, Stable Diffusion의 두 가지 버전[4][5]을 대상으로 실험을 진행하였다. Stable Diffusion은 잠재 공간에서 텍스트를 이미지로 변환하는 확산 모델로, 고해상도의 우수한 품질의 영상을 생성하는 데 특화되어 있다. 우리는 FreeU와 동일하게 추가 학습이나 추가 파라미터 없이 영상을 생성하며, FreeU에서 제시한 방법(백본 특징과 스킵 특징을 위한 스케일링 팩터)과 비교하였다.

Stable Diffusion 2.1[4] (SD 2.1)과의 비교에서는 백본 특징 스케일링 팩터의 최댓값(b)으로 각각 1.2를 사용하였고, Stable Diffusion XL[5] (SDXL)과의 비교에서는 백본 특징 스케일링 팩터의 최댓값으로 1.2를 사용하였다. 히스토그램을 계산하기 위한 빈(bin)의 개수(N)은 256으로 설정하였다.

성능 지표로는 FID[8], CLIP-Score[9]를 활용하였으며, MS-Coco 데이터셋[10]에서 랜덤하게 추출한 1,000장의 이미지를 사용하여 평가를 진행하였다.

2. 생성 영상의 객관적 성능 비교

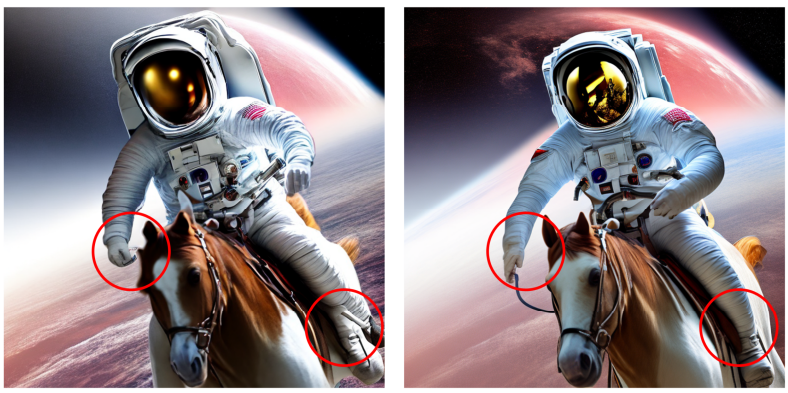
표 1에서는 FreeU와 본 논문에서 제안한 방법을 정량적인 지표를 통해 비교 분석하였다. SD 2.1과 SDXL 모델을 사용한 실험 결과에서, FID와 CLIP-Score 지표를 기준으로 할 때 제안 방법이 일관되게 더 우수한 성능을 나타내었다. 이는 제안된 방법이 입력 이미지의 특성과 중요도를 효과적으로 반영하여 생성함으로써, 결과적으로 객관적인 성능 지표에서도 높은 평가를 받은 것으로 해석할 수 있다.

표 1. FreeU와 제안 방법의 정량적 지표 (FID, CLIP Score) 비교
Table 1. Quantitative Comparison of FreeU and the Proposed Method (FID, CLIP Score)

Base	Method	CLIP Score (↑)	FID (↓)
SD 2.1	FreeU	33.7557	80.738
	Proposed	**33.9955**	**79.095**
SD-XL	FreeU	34.4139	93.952
	Proposed	**34.5388**	**89.651**

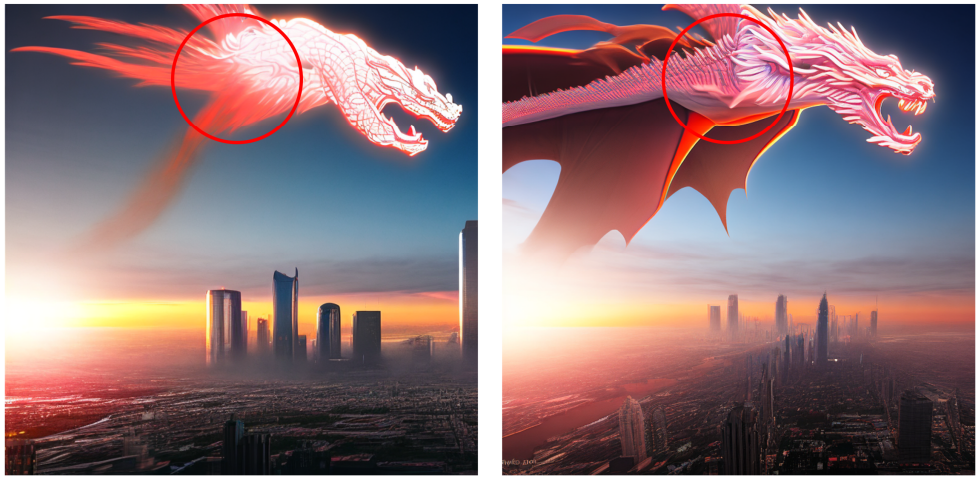
3. 생성 영상의 주관적 성능 비교 - SD 2.1[4]

그림 3, 그림 4, 그림 5는 제안 방법(b)과 FreeU(a)를



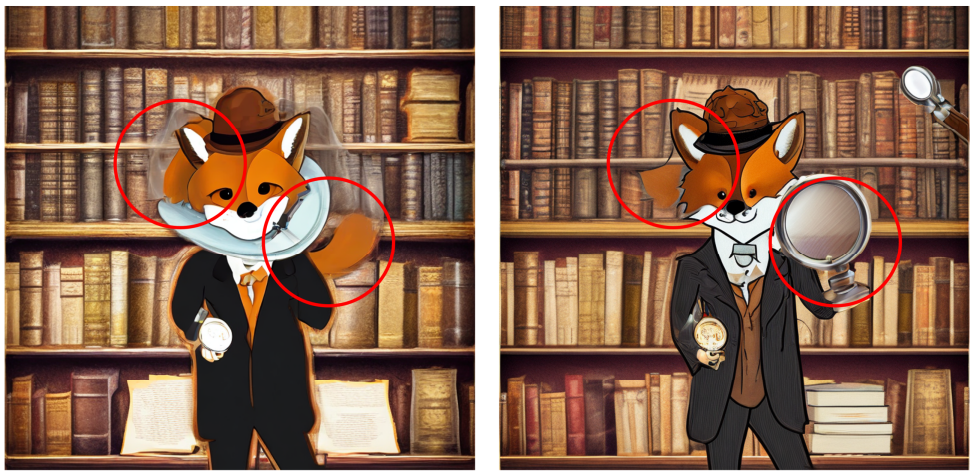
(a) FreeU (b) Proposed

그림 3. FreeU와 제안 방법의 생성 영상 비교 (SD 2.1)
Fig. 3. Comparison of Generated Images between FreeU and the Proposed Method (SD 2.1)
(Text: An astronaut is riding a horse in the space in a photorealistic style)



(a) FreeU (b) Proposed

그림 4. FreeU와 제안 방법의 생성 영상 비교 (SD 2.1)
Fig. 4. Comparison of Generated Images between FreeU and the Proposed Method (SD 2.1)
(Text: A dragon made of clouds flying over a futuristic city with neon lights, during sunset)



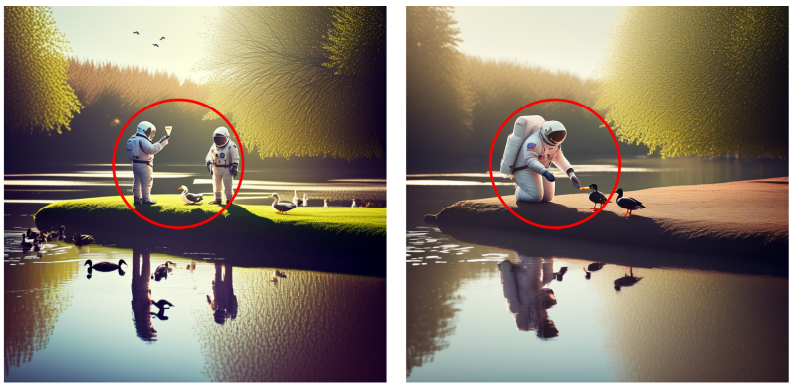
(a) FreeU (b) Proposed

그림 5. FreeU와 제안 방법의 생성 영상 비교 (SD 2.1)
Fig. 5. Comparison of Generated Images between FreeU and the Proposed Method (SD 2.1)
(Text: A fox wearing a detective hat and magnifying glass, solving mysteries in a Victorian-style library)

Stable Diffusion 2.1[4]를 사용해 실험한 결과를 보여준다. 각 그림에서 공통적으로 제안 방법이 주요 정보를 더욱 정확히 표현하며, 특히 객체를 집중적으로 잘 묘사하는 모습을 확인할 수 있다. 그림 3에서는 제안 방법의 사람과 말의 형태가 정확하게 생성된 것을 볼 수 있다. 그림 4에서는 용의 모습이 제안 방법이 더 선명하게 생성되었고, 그림 5에서는 여우와 돋보기가 더 잘 생성되었다. 이는 제안 방법이 채널별 중요도를 고려해 스케일링을 적용하고, 스킵 특징의 저주파수 정보를 적절히 조정한 결과로, 주요 정보를 효과적으로 강조한 것으로 해석된다.

4. 생성 영상의 주관적 성능 비교 - SDXL[5]

그림 6, 그림 7, 그림 8은 제안 방법(b)과 FreeU(a)를

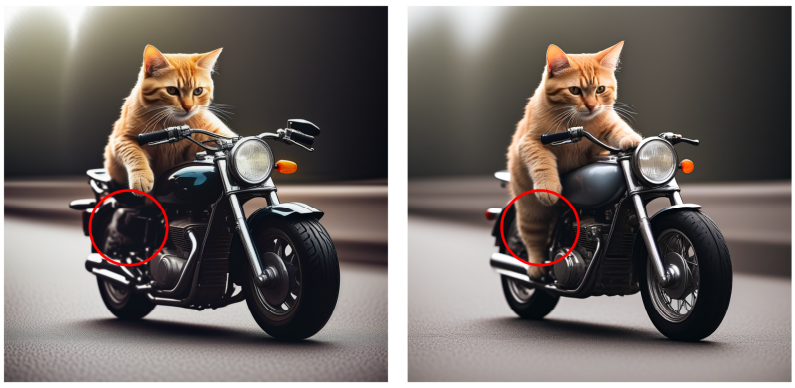


(a) FreeU (b) Proposed

그림 6. FreeU와 제안 방법의 생성 영상 비교 (SDXL)
Fig. 6. Comparison of Generated Images between FreeU and the Proposed Method (SDXL)
(Text: An astronaut feeding ducks on a sunny afternoon, reflection from the water)

(a) FreeU (b) Proposed

그림 7. FreeU와 제안 방법의 생성 영상 비교 (SDXL)
Fig. 7. Comparison of Generated Images between FreeU and the Proposed Method (SDXL)
(Text: A drone view of celebration with Christmas tree and fireworks, starry sky – background)



(a) FreeU (b) Proposed

그림 8. FreeU와 제안 방법의 생성 영상 비교 (SDXL)
Fig. 8. Comparison of Generated Images between FreeU and the Proposed Method (SDXL)
(Text: A cat riding a motorcycle)

Stable Diffusion XL을 사용하여 실험한 결과를 보여준다. 이들 그림에서도 SD 2.1[4] 실험 결과와 마찬가지로, 제안 방법이 객체를 집중적으로 더욱 세밀하게 표현하는 것을 확인할 수 있다. 그림 6에서는 우주 비행사의 수와 우주 비행사가 먹이를 주는 모습이 정확하게 생성되었다. 그림 7에서는 크리스마스를 축하하는 사람들의 모습이 생성되었고, 그림 8에서는 고양이의 다리가 정확하게 생성되었다. 이 또한 제안 방법이 의도한 대로 채널별 중요도를 반영하고, 스킵 특징의 저주파수 정보를 적절히 조정함으로써 주요 정보를 효과적으로 강조한 결과로 해석된다.

Ⅴ. 결 론

본 논문에서는 기존 FreeU[2]의 한계였던 백본 특징 스케일링 팩터와 스킵 특징 스케일링 팩터가 고정된 값을 사용하는 한계를 해결하기 위해, 입력 데이터의 특성과 중요도에 따라 스케일링 값을 동적으로 조정하는 '적응적 FreeU'를 제안했다. 제안된 방법은 채널별로 엔트로피를 계산해 각 채널이 포함한 정보의 양을 파악하고, 이를 바탕으로 백본 특징을 더 강조할 수 있도록 설계되었다. 또한, 스킵 특징에서는 저주파수 정보의 비중을 계산해 불필요한 저주파수 정보를 억제함으로써 오버스무딩 현상을 줄이는 방식을 적용했다.

Stable Diffusion 2.1[4] (SD2.1)와 Stable Diffusion XL[5] (SDXL) 모델에서 제안 방법의 성능을 측정한 결과, FreeU보다 FID[8]와 CLIP-Score[9]같은 정량적 지표에서 더 나은 성능을 보였다. 또한, 생성된 이미지에서 객체의 형태나 세부 디테일이 더 뚜렷하게 표현하는 등 정성적인 측면에서도 개선된 결과를 확인할 수 있었다. 이는 입력 데이터의 특성과 중요도를 반영해 스케일링 값을 조정한 제안 방법이 효과적인 것을 보여준다.

앞으로는 제안된 방법을 다양한 생성 모델에 적용하여 효과를 검증함으로써 모델의 범용성을 확보하고, 백본 특징 스케일링 팩터의 최댓값을 자동으로 설정하는 방법을 연구하여 성능을 더욱 향상시키고자 한다. 이를 통해 다양한 상황에서도 일관적으로 우수한 성능을 발휘할 수 있는 방법을 개발할 계획이다. 추가적으로 복잡도 대비 어느 정도 성능 향상이 있는지 기존 FreeU와 비교 분석을 할 예정이고, 제안하는 방법들에 대해 ablation study를 진행할 예정이다.

참 고 문 헌 (References)

[1] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. Springer International Publishing, 2015.
doi: https://doi.org/10.1007/978-3-319-24574-4_28

[2] Si, Chenyang, et al. "Freeu: Free lunch in diffusion u-net." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.
doi: https://doi.org/10.1109/CVPR52733.2024.00453

[3] Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models." Advances in neural information processing systems 33 (2020): 6840-6851.
doi: https://doi.org/10.48550/arXiv.2006.11239

[4] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In CVPR, 2023.
doi: https://doi.org/10.48550/arXiv.2304.08818

[5] Podell, Dustin, et al. "Sdxl: Improving latent diffusion models for high-resolution image synthesis." arXiv preprint arXiv:2307.01952 (2023).
doi: https://doi.org/10.48550/arXiv.2307.01952

[6] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In CVPR, 2023.
doi: https://doi.org/10.1109/CVPR52729.2023.02155

[7] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. VideoFusion: Decomposed diffusion models for high-quality video generation. In CVPR, 2023.
doi: https://doi.org/10.1109/CVPR52729.2023.00984

[8] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium, 2018.
doi: https://doi.org/10.48550/arXiv.1706.08500

[9] Hessel, Jack, et al. "Clipscore: A reference-free evaluation metric for image captioning." arXiv preprint arXiv:2104.08718 (2021).
doi: https://doi.org/10.48550/arXiv.2104.08718

[10] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In ECCV, pages 740–755. Springer, 2014.
doi: https://doi.org/10.48550/arXiv.1405.0312

저 자 소 개

유 장 현

- 2023년 2월 : 경희대학교 전자공학과 학사
- 2023년 2월 ~ 현재 : 경희대학교 컴퓨터공학부 석사
- ORCID : https://orcid.org/0009-0009-1818-4315
- 주관심분야 : 비디오 부호화, 딥러닝, 컴퓨터비전, 표준화

조 현 동

- 2023년 8월 : 경희대학교 소프트웨어융합 학사
- 2023년 8월 ~ 현재 : 경희대학교 컴퓨터공학부 석사
- ORCID : https://orcid.org/0009-0002-8939-7624
- 주관심분야 : 비디오 부호화, 딥러닝, 컴퓨터비전, 표준화

최 승 미

- 2022년 8월 : 경희대학교 컴퓨터공학과 학사
- 2024년 2월 : 경희대학교 컴퓨터공학부 석사
- 2024년 2월 ~ 현재 : 경희대학교 컴퓨터공학부 박사
- ORCID : https://orcid.org/0000-0002-6402-7785
- 주관심분야 : 영상처리, 인공지능, 디지털 홀로그램

배 성 준

- 1997년 2월 : 고려대학교 전자공학과
- 1999년 2월 : KAIST 전자전산학과 석사
- 2004년 8월 : KAIST 전자전산학과 박사
- 2004년 8월 ~ 2005년 10월 : 하나로텔레콤 기술전략팀
- 2005년 10월 ~ 현재 : 한국전자통신연구원 방통미디어연구본부 책임연구원
- ORCID : https://orcid.org/0009-0004-8030-1563
- 주관심분야 : 2D/3D 비디오 객체 및 장면 부/복호화, 렌더링 및 서비스 기술

김 휘 용

- 1994년 8월 : KAIST 전기및전자공학과 공학사
- 1998년 2월 : KAIST 전기및전자공학과 공학석사
- 2004년 2월 : KAIST 전기및전자공학과 공학박사
- 2003년 8월 ~ 2005년 10월 : ㈜애드팍테크놀러지 멀티미디어팀 팀장
- 2005년 11월 ~ 2019년 8월 : 한국전자통신연구원(ETRI) 실감 AV연구그룹 그룹장
- 2013년 9월 ~ 2014년 8월 : Univ. of Southern Califonia (ISC) Visiting Scholar
- 2019년 9월 ~ 2020년 2월 : 숙명여자대학교 전자공학전공 부교수
- 2020년 3월 ~ 현재 : 경희대학교 컴퓨터공학과 부교수
- ORCID : https://orcid.org/0000-0001-7308-133X
- 주관심분야 : 비디오 부호화, 딥러닝 영상처리, 디지털 홀로그램