조건 기반 배경 이미지 생성과 희귀 객체 합성에 관한 연구

박상영*, 정성윤, 김병현, 정남훈, 조규태
LIG넥스원
{sangyoung.park*, seongyun.jeong, byunghyun.kin, namhoon.jung, kyutae.cho2}@lignex1.com

A Study for Rare Object Synthesis with Conditioned Background Image Generation

Sangyoung Park*, Seongyun Jeong, Byunghyun Kim, Namhoon Jung, Kyutae Cho
LIG Nex1

요 약

본 연구는 다양한 환경에서의 희귀 객체를 포함한 2-stage 이미지 생성 방법을 제안한다. 기존 딥러닝 기반 이미지 생성 모델들은 대부분 일반적인 객체에 대한 학습에 집중되어 있어, 희귀 객체에 대해서는 생성 성능이 떨어지는 한계가 있다. 또한, 이러한 객체에 대해 인식 및 분류 등의 성능을 향상 시키기 위해서는 다양한 환경에서 해당 객체가 포함 되어있는 많은 이미지가 필요하며, 목적에 따라 특정 환경에서의 객체 이미지가 필요한 경우도 존재한다. 본 연구는 이러한 문제를 해결하기 위해 사용자가 지정한 위치에 생성할 물체를 text로 입력하여 이미지를 생성하는 생성 모델을 기반으로 배경 이미지를 생성한 뒤, 희귀 객체를 합성하는 방식으로 이미지를 생성한다. 객체 삽입 후 diffusion 모델을 이용해 객체의 구조는 보존하되 디테일의 다양성을 추가함으로써 희귀 객체에 대한 학습 데이터를 효율적으로 생성할 수 있으며, 보다 강건하고 일반화된 인식 및 분류 모델 학습에 기여할 수 있다.

Ⅰ. 서 론

딥러닝 기반 이미지 생성 기술은 최근 다양한 분야에서 활용되고 있으며, 특히 Generative Adversarial Networks (GAN)와 Diffusion 모델은 고해상도의 사실적인 이미지를 생성하는 데 있어 뛰어난 성능을 보이고 있다. 하지만 이러한 생성 모델은 일반적인 객체에 비해, 더 많은 세부 구조와 디테일을 요구하는 희귀 객체를 생성하는 데 한계가 존재한다. 이는 실제 응용 분야에서 희귀 객체의 인식 및 분류가 시스템의 성능과 안정성에 중대한 영향을 미친다는 점에서 중요한 문제로 작용한다.

더불어, 희귀 객체에 대한 데이터는 수집 자체가 어렵고, 시간, 날씨, 배경 등 다양한 환경 조건을 반영한 이미지를 확보하는 것은 더욱 큰 과제로 남아 있다. 이와 같은 한계점을 극복하기 위해 생성 모델을 통해 희귀 객체 이미지를 생성하는 다양한 연구 [1], [2]가 진행되고 있지만, 아직 근본적인 문제를 완전히 해결하지는 못한 상황이다. 특히 텍스트 기반 생성 모델의 경우, 단순한 프롬프트만으로 객체의 정확한 위치를 지정하는 데 어려움이 있어 실제 응용 가능한 정제된 데이터 생성을 제한하는 요인이 된다.

이러한 한계를 극복하기 위해, 본 연구에서는 사용자가 원하는 객체와 위치를 직접 지정할 수 있도록 지원하는 GLIGEN [3] 기반의 2-stage 이미지 합성 방법을 제안한다.

첫 번째 단계에서는 GLIGEN을 활용해 사용자가 지정한 위치에 생성할 물체 및 option prompt를 정해 배경 이미지를 생성하여 다양한 배경에서의 희귀 객체 이미지를 얻을 수 있도록 한다. 두 번째 단계에서는 희귀 객체를 해당 위치에 삽입한 뒤, 삽입된 객체의 patch를 diffusion 모델에 통과시켜 객체의 구조적 특성을 유지하면서 디테일을 자연스럽게 보완할 수 있도록 한다. 이와 같은 2-stage 이미지 합성 방식을 통해 학습 모델의 도메인 차이를 줄여 높은 일반화 성능을 기대할 수 있다.

Ⅱ. 본 론

본 논문에서는 비어있는 이미지 내에 bounding box를 지정하고 그 부분에 생성할 물체 및 option prompt를 입력하여 배경 이미지를 생성하는 “Background Generation” 및 사용자가 가지고 있는 Object 이미지를 생성한 배경 이미지 내 원하는 부분에 합성하는 “Synthesis Process” 과정을 통해 이미지 합성을 진행한다. 전체적인 생성 과정을 [그림 1]를 통해 나타내었다.

1-stage: Background Generation

해당 과정에서는 GLIGEN을 활용하여 희귀 객체가 배치될 배경 이미지를 만든다. 사용자는 배경 이미지 내의 특정 위치를 지정하여 해당 위치에 배치될 물체 및 option(환경)을 직접 설정할 수 있고 물체가 복수인 경우에도 적용 가능하다. 다만 위치를 겹치게 설정하는 경우 생성 품질이 떨어질 수 있다.

바다, 숲속, 도시 등과 같아 특정 환경 내에 존재하는 객체 이미지가 필요한 경우 유용하게 사용할 수 있으며, 조건에 맞게 랜덤하게 생성되는 배경 이미지들을 활용하여 다양한 환경에 희귀 객체를 합성할 수 있도록 배경 이미지를 생성한다.

Option의 경우 물체로 표현되기 어려운 형용사 및 배경 이미지의 스타일을 prompt 형태로 적어 생성하고자 하는 이미지에 디테일을 추가한다. 그 후, 생성된 배경 이미지 내에 Region Of Interest (ROI)를 지정하여 희귀 객체가 합성될 위치를 특정한다.

2-stage: Synthesis Process

해당 과정에서는 만들어진 배경 이미지에 희귀 객체를 합성하여 최종 이미지를 생성한다. 우선, 희귀 객체 이미지의 배경을 제거해 객체만 남긴 후 ROI 내에 해당 객체를 붙여 넣어 patch를 생성한다.

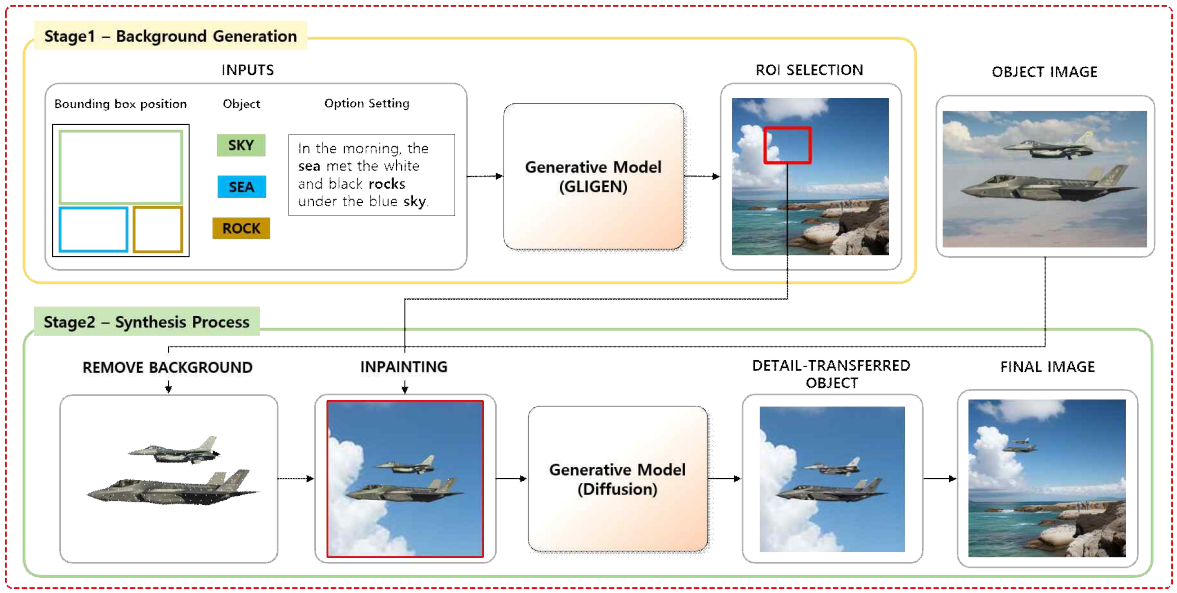

그림 1 전체 이미지 생성 과정

그 후, 해당 patch에 diffusion을 적용하여 객체의 특성을 유지하면서도 디테일이 변경될 수 있도록 한다. 이를 위해 원본 patch를 VAE에 통과시켜 latent space로 인코딩한 뒤, diffusion 과정을 적용할 때 활용한 후, 디코딩을 진행한다. [그림 2]에 원본 patch와 함께 diffusion 및 디코딩을 완료한 sample patch 2개의 확대 이미지를 나타내었다. 객체의 구조는 보존한 체 무늬, 색과 같은 디테일이 변경된 것을 확인할 수 있다.

마지막으로 해당 patch를 배경 그림의 ROI 위치에 붙여 넣어 최종 이미지를 생성한다.

그림 2 원본 patch 및 diffusion 과정 진행 후의 patch (확대)

Diffusion 과정 중 노이즈 제거량을 너무 높이게 되면 객체의 구조 자체가 변형되어 객체의 구조 보존의 의미가 사라지고 너무 낮춰 생성하면 객체의 디테일 변화가 거의 없어지기 때문에 적절한 hyperparameter를 설정하는 것이 중요하다. 또한, 과정 중에 사용되는 positive prompt 및 negative prompt는 공란으로 설정하였다.

해당 과정을 통해 매번 동일한 object를 붙여 넣는 경우와 비교해 모델이 다양한 representation을 학습할 수 있도록 하여 높은 일반화 성능을 기대할 수 있다.

Ⅲ. 실험 세팅

이미지 합성에는 다음과 같은 환경을 사용하였다.

Background Generation

diffusion model: endlessreality_v11

VAE: vae-ft-mse-840000-ema-pruned

Sampling: steps-12, cfg-12, euler sampler, normal echeduler

LoRA: add-detail-xl

Synthesis Process

diffusion model: DreamShaperXL_Lightning

VAE: DreamShaperXL_Lightning base VAE

Sampling: steps-5, cfg-2, euler ddpm_sde, karras scheduler

LoRA: add-detail-xl

Ⅳ. 결 론

본 논문에서는 2-stage 이미지 생성 및 희귀 객체 합성 방법에 대해 기술 하였다. Bounding box와 그 안에 생성될 물체, option prompt를 설정하여 배경 이미지를 쉽게 생성할 수 있어 원하는 환경에서의 객체 이미지를 획득할 수 있으며, 배경 이미지에 객체를 합성하는 경우에도 별도의 diffusion process를 거쳐 학습모델이 객체의 다양한 representation을 학습할 수 있도록 구성하였다.

응용 분야의 이미지가 일반적이고 획득이 어렵지 않은 경우라면 사진을 촬영 하거나 [4]와 같이 생성 모델을 이용하는 등의 방법을 통해 추가적인 성능 향상을 기대할 수 있지만, 군사·의료 등과 같이 한정된 domain이나 fine-grained class에 대한 학습이 필요한 경우와 같이 해당 방법이 제한되는 경우에 사용할 수 있을 것으로 기대한다.

참 고 문 헌

[1] Pan, Z., et al., “FineDiffusion: scaling up diffusion models for fine-grained image generation with 10,000 classes,” Applied Intelligence, vol. 55, no. 5, p. 309, 2025.

[2] Monsefi, A. K., et al., “TaxaDiffusion: Progressively Trained Diffusion Model for Fine-Grained Species Generation,” arXiv preprint arXiv:2506.01923, 2025.

[3] Li, Y., et al., “Gligen: Open-set grounded text-to-image generation,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 22511 - 22521, 2023.

[4] K. Islam et al., “DiffuseMix: Label-Preserving Data Augmentation with Diffusion Models,” arXiv preprint arXiv:2405.14881, 2024.