One-Step Diffusion 기반 화재 장면 생성을 통한 시뮬레이터 데이터 개선 및 화재 인식 성능 분석

김가현, 박민호, 강동오

한국전자통신연구원

{gahyeon, roger618, dongoh}@etri.re.kr

Improving Simulation Data via One-Step Diffusion-Based Fire Scene Generation for Fire Perception Performance Analysis

Kim Ga Hyeon, Park Min Ho, Kang Dong Oh

Electronics and Telecommunications Research Institute (ETRI)

요 약

최근 로봇 및 자율 시스템의 학습과 검증을 위해 3차원 가상환경 시뮬레이터를 활용한 시뮬레이션 기반 데이터 생성(sim2real, real2sim)이 활발히 연구되고 있다. 특히 화재와 같은 위험 상황은 실제 환경에서의 데이터 수집이 어렵고 높은 비용과 안전상의 제약을 수반하므로, 가상환경에서 현실감 있는 이미지 생성 및 편집 기술의 중요성이 더욱 커지고 있다. 본 논문에서는 GazeboSim 기반 3차원 가상환경 시뮬레이터에서 획득한 데이터를 대상으로 one-step diffusion 기반 Image-to-Image 변환 기술을 적용하여 개선된 화재 장면을 생성하는 방법을 제안한다. SwiftEdit와 같은 이미지 편집 모델을 시뮬레이터 내 에이전트 로봇의 시점 이미지에 적용함으로써 real2sim 파이프라인으로서의 활용 가능성을 분석한다. 또한 탐지 모델인 Grounding DINO와 멀티모달 임베딩 모델인 CLIP을 활용하여 데이터 개선 전후의 효과를 정성 및 정량적으로 평가하고, 이를 통해 시뮬레이터 데이터의 시각적 실제성과 의미 정합성 향상을 검증한다.

Ⅰ. 서 론

최근 확산 모델(diffusion model)은 고품질 이미지 생성 성능으로 인해 컴퓨터 비전과 로보틱스 분야에서 널리 활용되고 있다. 기존 확산 모델은 다단계 반복 샘플링을 통해 이미지를 생성하기 때문에 높은 품질을 제공하는 반면, 추론 시간이 길어 응용에는 한계가 있었다. 이러한 문제를 해결하기 위해 SD-Turbo[1]와 같은 one-step diffusion 계열의 모델이 제안되었으며, 이는 학습 단계에서 distillation 기법을 활용하여 단 한 번의 추론으로 이미지를 생성할 수 있도록 설계되었다. One-step diffusion의 등장은 로봇 시뮬레이션 및 디지털 트윈 환경에서 새로운 가능성을 제공한다. 특히 화재와 같은 비정상 상황은 실제 데이터 수집이 제한적이며, 다양한 조건(연기, 조명, 시야 가림 등)을 포괄하기 어렵다. 따라서 가상환경 이미지에 대해 텍스트 가이드를 활용하여 화재 장면을 즉시 생성하거나 편집할 수 있다면, 위험 상황 대응 로봇 및 인공지능 모델의 학습 데이터를 효율적으로 확장할 수 있다.

본 연구에서는 이러한 one-step diffusion 기반 Image-to-Image 변환 기술을 GazeboSim[2] 기반 물류센터 환경 시뮬레이터에 적용한다. 시뮬레이터 내 이동 로봇의 카메라 뷰에 화재 장면을 자연스럽게 주입함으로써, 실제 물류센터에서 발생할 수 있는 화재 상황을 가상환경에서 재현한다. 이는 향후 화재 탐지, 경로 계획, 자율 대응 로봇 시스템의 사전 학습 및 검증 단계에서 활용될 수 있으며, 실제 물류센터 운영 시 안전성과 대응 속도를 향상시키는 데 기여할 수 있다.

Ⅱ. 본론

Ⅱ-1. One-step Diffusion 기반 Image-to-Image 변환 모델

기존 확산 모델(diffusion model)은 확률적 역과정을 반복적으로 근사하는 다단계 noise - denoise 구조를 통해 고품질 이미지를 생성하지만, 이러한 반복 샘플링 특성으로 인해 추론 시간이 길다는 한계를 가진다. 이를 완화하기 위해 최근에는 다단계 확산 과정을 단일 함수로 근사하려는 distillation 및 adversarial 학습 기반의 one-step diffusion 계열 모델이 제안되었으며, 이는 확산 모델의 생성 품질을 유지하면서도 추론 시 한 번의 forward pass만으로 이미지 생성 및 편집을 가능하게 한다. 본 연구에서는 해당 접근법을 기반으로 한 텍스트 가이드 Image-to-Image 편집 모델인 SwiftEdit[3]을 사용하였다. SwiftEdit은 입력 이미지와 텍스트 프롬프트를 직접 조건으로 활용하여 별도의 diffusion inversion 과정 없이 의미적으로 일관된 편집 결과를 생성하고, 화재나 연기와 같은 국소적 시각 변화를 입력 이미지의 기하 구조를 유지한 채 삽입할 수 있다. 실험 결과, SwiftEdit은 NVIDIA A100 GPU에서 512×512 입력 기준 평균 0.8 - 1.0초/frame의 지연 시간을 보여 GazeboSim 기반의 시뮬레이터 파이프라인에 적용 가능함을 확인하였다.

Ⅱ-2. GazeboSim 기반 가상환경 데이터 수집 및 표현 한계

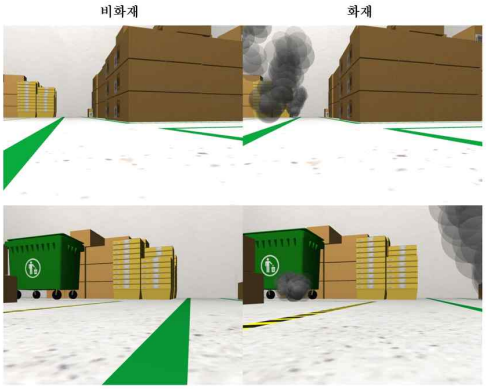

그림 1. GazeboSim 가상 환경 시뮬레이터 수집 데이터 예시

화재 장면 생성을 위한 기초 데이터로 GazeboSim 기반 3차원 가상환경 시뮬레이터를 활용하였다. 물류센터를 모사한 환경에서 화재/비화재 상태, 다양한 카메라 각도 및 시점 조건을 포함한 scene 데이터를 수집하였다. 해당 데이터는 Image-to-Image 변환 모델의 적용 및 성능 평가를 위한 학습 및 검증 데이터셋으로 활용된다. 다만 그림 1에서와 같이 GazeboSim은 부자연스러운 연기 표현이나 실제 불꽃의 시각적 특성을 사실적으로 재현하는 데 한계가 있어, 고현실성 화재 장면 생성을 위해 추가적인 이미지 변환 기법이 요구된다.

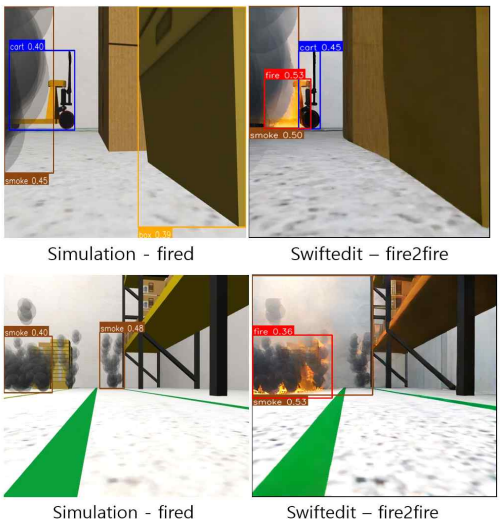

그림 2. Grounding DINO 기반 화재 탐지 성능 평가 예시

Ⅱ-3. 데이터 개선 효과에 대한 정성 및 정량 평가

시뮬레이터 환경 이미지에 SwiftEdit를 적용하여 화재 장면을 생성한 후, 데이터 개선 효과를 정성, 정량적으로 평가하였다. 정량 평가로 Grounding DINO[4]를 사용하였으며, bounding box 기준의 인식 정확도(hit score)를 지표로 활용해 데이터 개선 전후의 화재 탐지 성능을 비교하였다. 그 결과, 그림 3에서처럼 GazeboSim에서 생성된 원본 화재 데이터 대비 SwiftEdit를 적용하여 화재 장면을 사실적으로 보강한 데이터셋에서 가장 높은 화재 탐지 정확도를 보였다. 이는 화재 영역의 시각적 표현이 보다 명확해짐에 따라 bounding box 기반 탐지 과정에서 객체 경계와 특징이 안정적으로 인식되었다고 해석할 수 있다. 정성 평가는 CLIP[5]을 활용하여 텍스트 프롬프트와 이미지 간 코사인 유사도를 계산하고, t-SNE 시각화를 통해 임베딩 분포의 변화를 분석하였다. 프롬프트로는 “a photo of a fire”를 사용하였으며, 원본 데이터와 개선된 데이터를 각각 CLIP 임베딩 공간에 시각화하였다. 그 결과, SwiftEdit를 적용한 데이터는 전반적으로 텍스트 임베딩 방향으로 이동하는 경향을 보였으나, 샘플별 코사인 유사도 변화에 차이가 존재했다. 특히 그림 2처럼 불꽃이 명확하게 생성된 샘플의 경우 텍스트 임베딩과의 유사도가 뚜렷하게 향상되는 경향을 확인하였다. 이는 SwiftEdit를 통해 개선된 화재 이미지가 텍스트와의 의미 정합성 측면에서 가장 안정적인 특성을 보이며, 화재 현장의 시각적 특징이 멀티모달 표현 공간에서 보다 명확하게 구분함을 보인다.

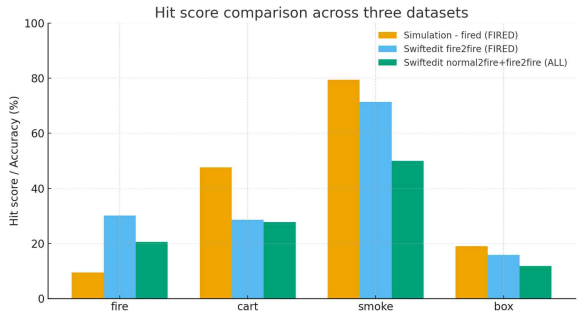

그림 3. Grounding DINO 기반 화재 탐지 hit score 데이터 간 비교

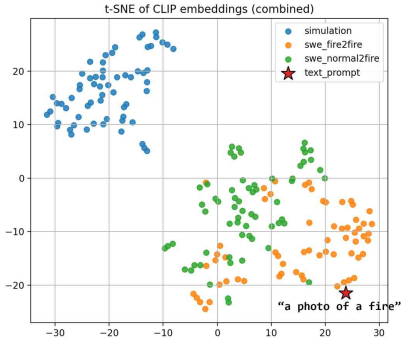

그림 4. 텍스트 프롬프트-이미지 간 CLIP 임베딩 공간 t-SNE 시각화

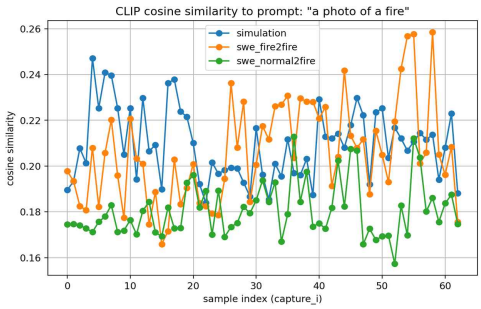

그림 5. 데이터 샘플 별 텍스트 프롬프트-이미지 간 코사인 유사도 비교

Ⅲ. 결론

본 논문에서는 GazeboSim 기반 3차원 가상환경 시뮬레이터에 one-step diffusion 기반 Image-to-Image 변환 기술을 적용하여, 화재 장면을 생성하는 방법을 제안하였다. SwiftEdit와 같은 모델을 활용함으로써, 시뮬레이터 데이터의 실제성과 다양성을 효과적으로 확장할 수 있음을 확인하였다. 또한 Grounding DINO와 CLIP 기반 평가를 통해 데이터 개선 효과를 정성, 정량적으로 검증하였다. 본 연구는 향후 물류센터 화재 대응 로봇 및 자율 시스템의 학습 데이터 구축과 실환경 적용을 위한 실질적인 기반 기술로 활용될 수 있을 것으로 기대된다.

ACKNOWLEDGMENT

이 논문은 2026년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.RS-2022-II220907, (2세부) AI Bots 협업 플랫폼 및 자기조직 인공지능 기술개발).

참 고 문 헌

[1] Sauer, A. et al., “Adversarial Diffusion Distillation,” arXiv preprint arXiv:2311.17042, 2023. (SD-Turbo)

[2] Koenig, N., and Howard, A., “Design and Use Paradigms for Gazebo, an Open-Source Multi-Robot Simulator,” IROS, 2004.

[3] Nguyen, T. et al., “SwiftEdit: Lightning-Fast Text-Guided Image Editing via One-Step Diffusion,” CVPR, 2025.

[4] Liu, S. et al., “Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection,” ECCV, 2024.

[5] Radford, A. et al., “Learning Transferable Visual Models From Natural Language Supervision,” ICML, 2021.