확산 모델기반 자동 큐레이션을 통한 도메인 특화 데이터셋 생성

윤동식, 김종은
HDC LABS

{kevinds1106, JongeunKim}@hdc-labs.com

요 약

본 연구는 확산 모델을 활용하여 도메인 특화 합성 데이터셋을 생성하는 자동화 파이프라인을 제시하며, 이를 통해 사전 학습 모델과 실제 배포 환경 간의 분포 변화 문제를 해결한다. 제안하는 프레임워크는 먼저 제어된 인페인팅을 통해 도메인 특화 배경 내에 목표 객체를 합성한다. 생성된 출력물은 객체 탐지, 미적 평가, 그리고 비전-언어 정렬을 통합한 다중 모델 평가를 통해 검증된다. 본 파이프라인은 광범위한 실제 데이터 수집에 대한 의존도를 줄이면서 고품질의 배포 가능한 데이터셋을 효율적으로 구축할 수 있도록 한다.

Ⅰ. 서 론

딥러닝의 급속한 발전과 함께 대규모 데이터셋과 이에 대응하는 사전 학습 모델이 광범위하게 활용 가능해졌다. 그러나 공개 코퍼스 기반 사전 학습 모델은 학습 데이터와 실제 배포 환경 간의 분포 불일치로 인해 성능 저하를 겪는 경우가 빈번하다. 공개 벤치마크가 특정 지리적 환경, 비정형적 촬영 시점, 특수 객체 범주 등 도메인 특화 요소를 충분히 반영하지 못하기 때문에 사전 학습 모델의 직접적인 전이 가능성은 제한적이다. 더욱이 이러한 도메인 특화 데이터 (특정 장소, 비정형 카메라 각도, 특수 객체 등) 수집은 범용 데이터셋 구축에 비해 월등히 복잡하며, 수집, 큐레이션, 주석 작업에 상당한 비용을 요구한다.

기존 연구들은 이러한 문제를 해결하기 위해 크게 세 가지 방향에서 접근해왔다. 첫째, 도메인 적응 기법은 특징 분포 정렬이나 도메인 불변 표현 학습을 통해 소스-타겟 분포 격차를 완화한다. 둘째, few-shot 미세 조정 방법은 소량의 라벨링된 타겟 샘플을 활용하여 사전 학습 모델을 특정 환경에 적응시킨다. 셋째, 확산 기반 생성 모델은 희귀한 장면, 시점, 객체 범주를 포함하는 타겟 정렬 데이터셋을 합성하여 기존 코퍼스를 증강한다.

최근 확산 모델의 발전은 합성 데이터셋 구축 비용을 획기적으로 절감하였으며, 생성된 합성 데이터는 공개 벤치마크 강화에 적극 활용되고 있다. 본 연구는 이러한 확산 기반 생성 모델링의 흐름 속에서, 수집이 어려운 시점과 객체 범주의 이미지를 합성하고 이를 고품질 데이터셋으로 전환하는 자동화된 큐레이션 파이프라인 개발을 목표로 한다.

제안하는 방법론은 사용자 지정 시점과 타겟 객체를 기반으로 확산 생성기를 통해 후보 이미지를 생성하며, 이후 텍스트-이미지 정렬 검증 과정을 거친다. 확산 모델이 종종 입력 프롬프트에서 벗어난 객체를 합성하는 고질적 한계를 극복하기 위해, VLM 이 생성한 이미지 캡션과 원본 프롬프트를 비교하여 정렬도를 평가한다. 이러한 필터링을 통해 프롬프트 불일치 이미지를 제거하고 의도된 시점 및 객체 범주에 부합하는 합성 샘플만을 선별함으로써, 실제 데이터 수집의 높은 비용 부담 없이 타겟 환경에 최적화된 합성 데이터셋을 효율적으로 구축한다.

Ⅱ. 본론

본 파이프라인은 도메인 특화 배경 이미지와 타겟 객체라는 두 가지 핵심 입력을 기반으로 작동한다. 도메인 특화 배경은 지하 주차장, 엘리베이터 CCTV 화면, 홈 보안 카메라 영상과 같이 수집이 본질적으로 제한적이거나 특정 위치에 종속된 장면을 의미한다. 이는 실제 배포 환경의 시각적 맥락을 반영하므로 데이터셋 구축의 출발점이 되지만, 획득 가능한 샘플 수는 제한적일 수밖에 없다. 타겟 객체는 해당 도메인 내에서 합성되어야 할 대상으로, 일반적으로는 흔하게 관찰되지만 공개 데이터셋에서는 특정 장면과의 자연스러운 조합이 희소한 경우를 다룬다. 예를 들어 화재가 발생한 지하 주차장, 엘리베이터 내부의 개, 홈 보안 영상 속 로봇 청소기 등이 이에 해당한다.

이미지 생성 단계에서는 Stable Diffusion, Midjourney, FLUX 와 같은 최신 확산 기반 생성기[1-2]를 활용하여 도메인 배경에 타겟 객체를 삽입한다. 먼저 도메인 장면과 타겟 객체를 명시하는 프롬프트 P 를 정의하며, 이를 기반으로 사용자 정의 관심 영역(ROI) 내에서 인페인팅을 수행한다. 마스크는 ROI 내에서 무작위로 샘플링되며, 그 크기는 사용자가 지정한 객체의 예상 크기를 반영한다. 그림 2

생성된 이미지의 품질을 보장하기 위해 객체 탐지기[3], 미적 평가 모델[4], 비전-언어 모델(VLM)[5]로 구성된 삼중 검증 프로토콜을 적용한다. 객체 탐지기는 타겟 클래스에 대한 신뢰도 S_{det}와 바운딩 박스 B_{det}를 출력하며, B_{det}와 인페인팅 마스크 영역 M 간의 IoU 를 계산하여 공간적 정확성을 확인한다. 높은 IoU 값은 합성 객체가 의도된 위치와 크기 범위 내에 적절히 배치되었음을 의미한다. 미적 평가 모델은 S_{aes} 점수를 통해 이미지의 전반적인 시각적 완성도와 지각적 품질을 측정하며, 이는 탐지기의 신뢰도를 보완하는 독립적 품질 지표로 기능한다. VLM 은 프롬프트-이미지 정렬을 검증하는 역할을 담당한다.

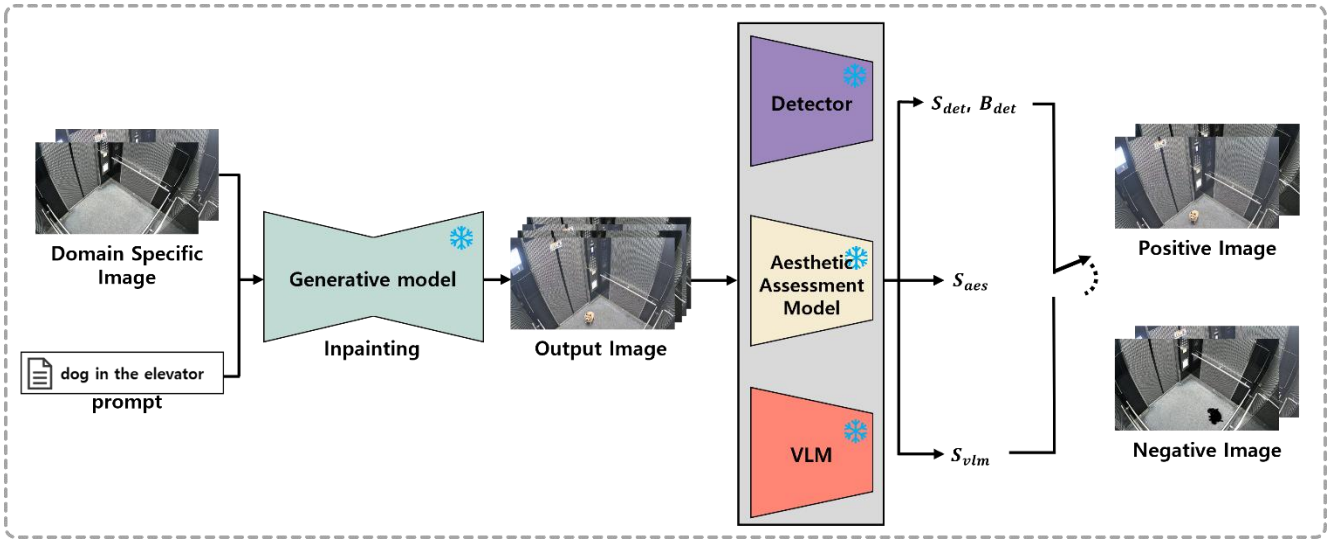

그림 1. 제안하는 확산 기반 데이터셋 생성 및 자동 큐레이션 파이프라인 개요

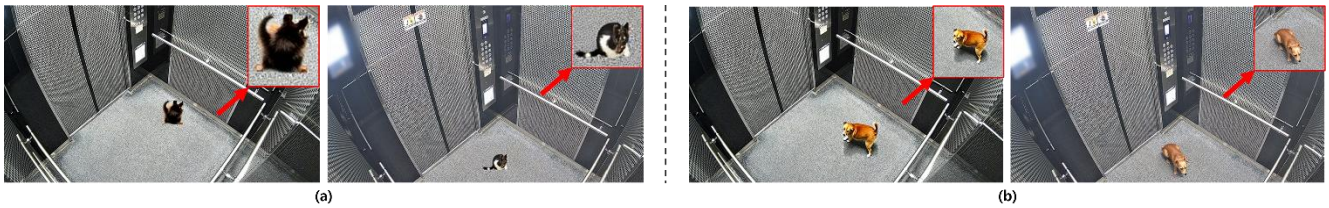

그림 2. (a) 낮은 탐지 및 미적 점수로 인해 객체 합성에 실패한 이미지, (b) 도메인 특화 배경 내에서 타겟 객체 합성을 성공적으로 수행하여 통과한 성공 이미지

생성된 이미지가 주어지면 VLM 은 이미지 전체를 서술하는 캡션 C 를 생성하고, 문장 임베딩 $\phi(\cdot)$ 를 사용하여 원본 프롬프트 P 와의 다음의 코사인 유사도를 계산한다.

$$S_{vlm} = \cos\big(\phi(C), \phi(P)\big) \tag{1}$$

이러한 세 가지 모델의 통합적 활용을 통해 합성 이미지가 의도된 도메인 맥락에서 타겟 객체를 충실하게 표현하는지 체계적으로 평가하고 타겟 환경에 적합한 고품질 합성 데이터셋을 구축한다. 최종적으로 자동 필터링을 통과한 이미지는 주석자의 수동 검수를 거쳐 데이터셋 포함 여부가 결정된다.

Ⅲ. 실험 설정

본 연구에서 다룬 사례 연구(엘리베이터 내 개 생성 시나리오)에서는 다음의 임계값을 적용하였다. 탐지기 신뢰도 S_{det} > 0.75, 미적 점수 S_{aes} > 5, 공간적 충실도 IoU(B_{det}, M) > 0.8, VLM 정렬도 S_{vlm} > 0.8. 이러한 임계값은 사용되는 사전 학습 모델의 선택에 따라 조정 가능하며, 특정 도메인의 요구사항에 맞춰 최적화될 수 있다.

그림 1, 2는 제안된 파이프라인 및 이를 통해 필터링된 이미지와 합성 결과를 나타낸다. 그림 2(a)는 제안된 파이프라인에 의해 제외된 부적합한 이미지를 보여주며, 그림 2(b)는 도메인 특화 배경에 타겟 객체가 자연스럽게 합성된 결과를 나타낸다.

Ⅳ. 결론

본 연구는 도메인 특화 데이터셋 생성 및 자동화 큐레이션을 위한 확산 기반 파이프라인을 제시하였다. 본 프레임워크는 객체 탐지기, 미적 평가기, 그리고 프롬프트-이미지 정렬을 위한 비전-언어 모델을 통합한다. 이러한 구성요소들의 결합을 통해 본 접근법은 체계적인 필터링을 통해 생성된 데이터셋의 품질을 보장하면서 도메인 특화 학습 데이터 획득의 비용과 복잡성을 감소시킨다.

참 고 문 헌

[1] Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." In *CVPR*. 2022.

[2] Batifol, Stephen, et al. "FLUX. 1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space." *arXiv e-prints* (2025): arXiv-2506.

[3] Tianheng Cheng and Lin et al. Song. 2024. Yolo-world: Real-time open-vocabulary object detection. In CVPR. 16901-16911.

[4] Christoph Schuhmann. 2022. LAION-Aesthetics Predictor v2. https://github.com/christophschuhmann/improved-aesthetic-predictor

[5] Andrés Marafioti and Orr Zohar et al. 2025. SmolVLM: Redefining small and efficient multimodal models. arXiv preprint arXiv:2504.05299 (2025).