KOREA INSTITUTE OF CONSTRUCTION ENGINEERING AND MANAGEMENT

스마트 건설기술을 위한 새로운 도전: Autodistill 기반 데이터 라벨링 자동화


이슬비 인천대학교 도시건축학부 조교수, sblee@inu.ac.kr
윤태관 인천대학교 건축학과 석사과정, x8333x@inu.ac.kr

1. 서론

드론이 아파트 외벽의 균열을 탐지하고, 로봇이 건설현장을 순찰하는 사례는 더 이상 놀라운 일이 아니게 되었다. 건설산업 생산성 혁신을 위해 앞다투어 개발된 스마트 건설기술들은 이제 건설현장의 모습을 변화시키고 있으며, 일부 유망 기술들은 상용화를 통해 시장 선점을 본격화하고 있다. 일례로 지난 2024년 현대건설은 경기도 용인시에서 일본 오사카 현장까지 장거리 원격 조종이 가능한 타워크레인 타와레모(TawaRemo) 시연회를 개최하였고, 삼성물산은 국토교통부가 주최한 스마트 건설 챌린지에서 철골 볼트 조임 자동화 로봇을 선보였다. 한국건설산업연구원의 「2025년 건설 분야 AI 기술 적용과 미래 전망」 보고서에 따르면, 건설산업의 품질, 안전 관련 이슈 대응을 위해 스마트 건설기술 개발은 앞으로도 활발하게 이어질 전망이다.

스마트 건설기술은 건설산업의 인력난 해결에도 중요한 역할을 할 수 있을 것으로 기대되고 있다. 건설현장에 도입된 스마트 건설기술이 기존에 사람이 하던 상황 인식, 판단, 실행의 과정을 대신하며 인력 부족을 보완하고 있기 때문인데(Niu et al., 2016), 이때 상황 인식 분야를 이끄는 핵심기술이 머신비전(Machine Vision)이다. 머신비전은 카메라를 이용해 획득한 디지털 이미지로부터 의미 있는 정보를 추출하는 인공지능의 한 분야로, 사람의 눈과 뇌를 대체하는 기술로 여겨진다(Davies 2004). 건설현장에 설치된 CCTV를 통해 작업자의 안전모 착용 여부를 감지하는 것도 머신비전 기술 중 하나인데, 이를 구현하기 위해서는 사람이 직접 디지털 이미지에 등장한 작업자를 '작업자'로 안전모를 '안전모'로 표시해주는 데이터 라벨링이 선행되어야 한다. 앞서 소개한 아파트 외벽 균열 탐지 드론의 경우 기술개발에 총 24,641장의 외벽 이미지가 수집되었다고 하니(Jeong et al. 2024), 전체 이미지의 70%를 학습에 이용했다고 가정하면 약 17,000장 이상의 이미지에 대한 데이터 라벨링이 수행되었을 것으로 예상할 수 있다.

데이터 라벨링은 '디지털 시대의 인형 눈알 붙이기'라고도 비유되며 단순노동 취급되기도 한다. 그러나 동시에 인공지능이 공부하기 위한 학습자료를 생성한다는 측면에서 결국 머신비전 모델의 성능을 결정짓는 중요한 작업이다. 아이러니하게도 건설현장 단순 인력 대체를 위한 스마트 건설기술의 개발 이면에는 또 다른 수많은 단순 인력이 숨겨져 있는 것이다. 일론 머스크가 설립한 인공지능 스타트업 xAI에서 언어 라벨링을 위한 전문인력의 인건비를 시간당 약 5만 원에서 9만 4,000원 사이로 책정했다는 것을 생각하면 비용효율적인 측면에서도 현재와 같이 전적으로 사람에 의존하는 데이터 라벨링은 올바른 방향이 아닐 수 있다. 이러한 배경에서, 본 고에서는 데이터 라벨링 자동화를 위한 제로 샷 학습(Zero-shot learning) 모델 중 하나인 Autodistill을 소개하고, 건설산업에서의 적용성을 높이기 위한 개선 방안을 연구 사례와 함께 제시하고자 한다.

2. 데이터 라벨링 자동화: Autodistill

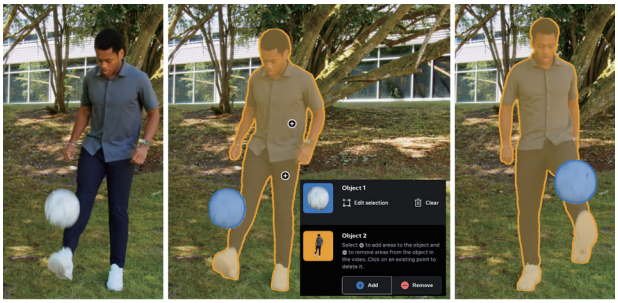
사전에 별도로 학습되지 않은 새로운 객체를 인식하기 위한 학습 방법을 제로 샷 학습이라고 한다(Lampert et al., 2013). 이 학습 방식은 다양한 작업을 수행할 수 있도록 미리 구축

되어있는 사전 모델의 속성 정보를 활용하여 새로운 객체를 해석하는데, 예를 들어 '긴 코를 가진 큰 회색 동물'이라는 설명이 주어지면 이를 기반으로 '코끼리'를 유의미하게 식별할 수 있도록 하는 것이다(Lampert et al., 2013). 2023년 6월 Roboflow가 처음 소개한 Autodistill은 이러한 제로 샷 학습을 활용하여 라벨이 없는 상태에서도 컴퓨터 비전 모델을 구축할 수 있도록 하는 라이브러리이다. 구체적으로, Autodistill은 Base model을 사용하여 라벨링되지 않은 이미지로부터 학습 데이터 세트를 생성한 후, 이를 활용하여 Target model을 훈련시켜 최종적으로 Distilled model을 생성한다. 이번 장에서는 Autodistill에서 Base model이 학습 데이터 세트를 생성하는 과정을 이해하기 위해, 대표적인 Base model 중 하나인 Grounded SAM을 구성하는 기본 기술들과 Caption Ontology에 대하여 설명한다.

2.1. Grounded SAM

1) SAM (Segment Anything Model)

페이스북의 모회사 메타는 2023년 4월 이미지 분할을 위한 새로운 AI 모델인 SAM(Segment Anything Model)을 발표했다 (Kirillov et al. 2023). SAM의 주요 특징은 사용자가 프롬프트를 통해 객체 분할을 할 수 있다는 것인데, 〈그림 1〉과 같이 이미지 내 특정 객체를 클릭하면 자동으로 객체의 마스크가 생성된다. 이미지에서 객체를 분할할 수 있도록 개발된 SAM은 2024년 8월 비디오에서도 같은 기능을 수행할 수 있는 SAM2로 확장되었다. 메타에서는 관련 분야 연구 활성화를 위해 SAM2와 함께 개발에 사용된 51,000개의 비디오 데이터를 포함하는 SA-V 데이터 세트를 오픈소스로 제공하고 있다.



(a) 원본 비디오 (b) 프롬프트 입력 과정: Point방식 (c) 객체 분할 결과

그림 . SAM2의 구현 예시

2) Grounding Dino

SAM2는 비디오에서 객체를 분할해낼 수는 있지만, 해당 객체가 무엇인지에 대한 정보는 포함하지 않는다. Grounding Dino는 이러한 SAM2의 한계를 보완할 수 있는 기술로, 텍스트 프롬프트를 통해 객체를 식별하는 Open-world 객체 탐지 모델이다(Ren et al. 2024). 천만 개 이상의 이미지-텍스트 쌍을 학습하여 개발된 Grounding Dino는 사용자가 입력한 텍스트에 해당하는 객체의 경계 상자(Bounding box)를 라벨과 함께 표시하기 때문에, SAM2의 분할 작업을 위한 프롬프트 역할을 하게 된다. 예를 들어, 앞서 설명한 〈그림 1〉의 비디오와 함께 사용자가 '공'과 '사람'이라는 텍스트를 입력하면 Grounding DINO가 이에 해당하는 영역에 대한 경계 상자를 생성하고 각각 '공'과 '사람'으로 라벨을 부여한다. 이후 SAM2는 Grounding DINO가 생성한 경계 상자 내부에서 객체 분할을 수행함으로써, 분할의 효율성이 향상될 뿐만 아니라 객체의 의미도 알 수 있게 된다.

2.2. Caption Ontology

Autodistill에서 Base model이 라벨이 없는 이미지를 어떻게 이해하고 라벨을 생성해야 하는지 구체적으로 명시해주는 것이 Ontology이다. 특히 텍스트 프롬프트를 사용하는 경우를 Caption Ontology라고 한다. Caption Ontology는 보통 {"프롬프트": "라벨"}의 형식을 취하는데, 프롬프트는 객체에 대한 자연어 설명을 나타내고 라벨은 해당 객체의 클래스 이름으로 사용된다. Caption Ontology는 Autodistill의 최종 결과물을 결정하는 핵심 요소이기 때문에 객체를 가장 잘 표현할 수 있는 프롬프트를 작성하는 것이 무엇보다 중요하다. 선행연구에서는 이미지에서 바지를 분할하기 위해 프롬프트로 'pants'를 사용하였을 때와 'slacks' 또는 'trousers'를 사용하였을 때의 최종 데이터 세트의 정확도가 달라지는 것을 확인하였다(Abluton and Portinale 2024).

3. 건설산업에서의 적용

건설산업에서 Autodistill의 활용 가능성을 평가하기 위해, 본 연구진은 건설현장 주변에서 흔히 발견되는 여러 객체에 대해 Autodistill이 제시한 방법론을 적용하여 수동 데이터 라벨링 없이 학습 데이터를 생성하였다. 이 장에서는 연구 사례에 대한 간략한 소개와 연구 결과를 제시하고, Autodistill의 적용성을 높이기 위한 향후 연구 방향을 논의한다.

표 1. 건설현장 주변 객체에 대한 Caption Ontology

라벨	프롬프트
안전펜스 (Safety fence)	What blue, fabric-like barrier, explicitly designed to shield or guide pedestrians near construction zones, is visible in this image?
보행로 (Sidewalk)	What type of pedestrian pathway, found alongside roads or within construction zones, composed of materials like nonwoven fabric, sand, bricks, or asphalt, is visible in this image?
자동차 (Car)	What motorized vehicle, designed for passenger or cargo transport, commonly seen in urban or road environments, is visible in this image?
트래픽콘 (Traffic cone)	What small, brightly colored cone-shaped object, specifically designed to redirect traffic or highlight construction hazards, is visible in this image?

3.1. 건설현장 주변 보행자 안전관리 플랫폼

건설현장 주변에 적치된 자재, 불법 주차 차량 등으로 인한 보행자 안전사고가 빈번하게 발생하고 있음에도 불구하고, 건설현장 주변 보행로는 건설현장 내부와 비교하여 상대적으로 소홀하게 관리되고 있다. 이러한 배경에서 본 연구진은 건설현장 주변의 보행 안전성을 평가하는 영상 기반 안전관리 플랫폼을 개발하는 것을 최종 목적으로 Autodistill 기반의 비전 데이터 자동 학습 모델을 개발하였다. 모델 개발을 위해 국내 26개 건설현장을 방문하여 총 1시간 25분 길이의 보행로 주변 영상을 수집하였으며, 영상에서 추출된 4,959장의 이미지를 사용하여 시범적으로 안전펜스, 보행로, 자동차, 트래픽콘의 네 가지 객체에 대한 학습 데이터 세트를 생성하였다. Base Model은 위에서 소개한 바와 같이 Grounded SAM을 사용하였고, Caption Ontology는 각각 다음 〈표 1〉과 같이 정의하였다.

프롬프트의 의미적 유연성과 일반화 성능을 확보하기 위해 Caption Ontology는 GPT-4o 모델을 사용하여 추론 형식으로 정의되었다. 최종적으로 〈그림 2〉와 같이 네 가지 객체에 대한 자동 학습 데이터 세트를 생성하는 데 소요된 시간은 평균 500초로, 기존의 수동 라벨링 방식과 비교하여 상당한 시간 절감을 달성할 수 있었다. 그러나 생성된 학습 데이터를 수동으로 재검토한 결과, 정확도는 현저히 낮은 것으로 나타났다. 예를 들어, 보행로의 경우 자동으로 생성된 4,955장 중 3,043장이 실제 보행로를 포함하였으며, 트래픽콘의 경우 4,136장 중 단 676장만이 정확하게 검출되었다. 평균적으로 네 개의 객체에 대한 정확도는 43.6%로 추정되었다. 제로 샷 학습을 사용하여 Text-to-Image를 구현한 선행 모델의 정확도는 54.3%에서 76.7% 사이에 분포하므로(Yuan et al. 2021), 본 연구 결과는 Autodistill을 활용한 자동 라벨링이 건설 현장에 적용되기 위해서는 정확도 향상을 위한 추가적인 개선이 필요함을 시사한다.

3.2. 라벨링 정확도 향상을 위한 방안

Autodistill 기반의 비전 데이터 자동 학습 모델의 라벨링 정확도를 높이기 위해, 본 연구진은 다음과 같은 두 가지 개선

(a) 안전 펜스 (b) 보행로 (c) 자동차 (d) 트래픽콘

그림 . SAM2의 구현 예시

방안을 고려하고 있다. 첫째, 일관된 용어 사용과 명확한 설명은 모델이 정확한 라벨을 생성하는 데 필수적이므로 Caption Ontology에 사용되는 프롬프트를 더욱 정교하게 생성해야 한다. 이를 위해 다양한 표현 방식을 실험하여 건설현장 주변의 객체를 잘 표현할 수 있는 최적의 프롬프트를 도출하는 과정을 추가할 수 있도록 Florence-2의 Vision-Language Model을 활용할 예정이다. 둘째, 생성된 학습 데이터 세트를 추가로 필터링하는 기능을 포함해야 한다. 이를 위해 이미지 사이의 유사성을 클러스터 단위에서 평가하는 비지도 학습 모델인 SPICE의 적용을 시도하고 있다(Niu et al. 2022).

4. 맺음말

먼 미래로만 인식되던 스마트 건설기술이 어느새 건설산업의 핵심기술로 자리매김하는 상황에서, Autodistill을 활용한 데이터 라벨링 자동화는 스마트 건설기술 개발을 위한 데이터 준비 단계에서의 자원 소모를 크게 줄여줄 것이다. 그러나, 아직 이러한 제로 샷 학습 모델에는 정확도라는 큰 약점이 존재하며, 객체의 다양성과 배경의 복잡성을 특징으로 하는 건설현장 환경에서 자동으로 생성되는 양질의 데이터는 상상으로만 가능한 것일 수도 있다. 그럼에도 불과 몇 년 전까지 로봇과 함께하는 건설현장이 현실이 됨을 예상할 수 없었듯이 건설현장을 위한 데이터 라벨링도 곧 만족할만한 자동화 수준을 달성할 수 있을 것으로 전망된다. 앞으로의 많은 혁신 AI 기술들이 스마트 건설기술의 효율성과 신뢰성을 높이고, 더불어 건설산업 전반의 생산성과 안전성 향상을 견인해줄 것을 기대한다.

참고문헌

1. Abluton, A., and Portinale, L. (2024). "Knowledge Distillation for a Domain-Adaptive Visual Recommender System." Proceedings of the International FLAIRS Conference, The Florida AI Research Society, 37(1).
2. Davies, E. R. (2004). "Machine vision: theory, algorithms, practicalities." Morgan Kaufmann Publishers, Elsevier, Amsterdam.
3. Jeong, Y., Lee, T., Bae, J., and Lee, K. (2024). "A case study of AI model development and on-site application for exterior wall of concrete building crack detection." Magazine of Korea Concrete Institute, Korea Concrete Institute, 36(2), pp. 74-77.
4. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollar, P., and Girshick, R. (2023). "Segment anything." Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4015-4026.
5. Lampert, C. H., Nickisch, H., and Harmeling, S. (2013). "Attribute-based classification for zero-shot visual object categorization." IEEE transactions on pattern analysis and machine intelligence, IEEE, 36(3), pp. 453-465.
6. Niu, C., Shan, H., & Wang, G. (2022). "Spice: Semantic pseudo-labeling for image clustering." IEEE Transactions on Image Processing, IEEE, 31, pp. 7264-7278.
7. Niu, Y., Lu, W., Chen, K., Huang, G. G., and Anumba, C. (2016). "Smart construction objects." Journal of Computing in Civil Engineering, ASCE, 30(4), 04015070.
8. Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., Zeng, Z., Zhang, H., Li, F., Yang, J., Li, H., Jiang, Q., and Zhang, L. (2024). "Grounded sam: Assembling open-world models for diverse visual tasks." arXiv preprint, arXiv, 2401.14159.
9. Yuan, L., Chen, D., Chen, Y. L., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C., Liu, C., Liu, M., Liu, Z., Lu, Y., Shi, Y., Wang, L., Wang, J., Xiao, B., Xiao, Z., Yang, J., Zeng, M., Zhou, L., and Zhang, P. (2021). "Florence: A new foundation model for computer vision." arXiv preprint, arXiv, 2111.11432.