2024년 대한건축학회 추계학술발표대회
제44권 제2호(통권 제82집) 2024. 10. 23 ~ 10. 25

2024년 추계학술발표대회 : 일반부문

건축 설계 AI 적용을 위한 건축 도면 데이터셋 구축의 현황 분석

An analysis of the current state of architectural drawing datasets for AI applications in architectural design

○허 민 지* 구 형 모** 김 근 재** 추 승 연***
Heo, Minji Gu, Hyeongmo Kim, Geunjae Choo, Seungyeon

Abstract

The purpose of this study is to analyze the current status of AI datasets utilized in the field of architectural design, and to suggest future directions for architectural drawing datasets. The current status was analyzed by focusing on domestic and foreign architectural drawing datasets, and the architectural drawings were limited to plan data. Based on the analysis, each dataset was compared, and finally, the deficiencies of the current datasets and the datasets that can complement them were proposed.

키워드 : 건축 설계, 건축 도면 데이터셋, 건축 평면 데이터셋, 벤치마크 데이터셋
Keywords : Architectural Design, Architectural Drawing Dataset, Floor Plan Dataset, Benchmark Dataset

1. 서론

전 산업 분야에 걸쳐 AI를 도입하고자 하는 시도가 가속화됨에 따라 건축 설계 분야 또한 AI를 업무에 적용하고자 하는 움직임이 늘어나고 있다. 지난해 통계청이 실시한 2023 인공지능 실태조사에 따르면 국내 기업 중 23.8%가 AI 기술을 도입하여 실제 업무에 활용 중인 것으로 나타났으나, 이 중 건설업 분야의 AI 기술 도입 비율은 7%에 불과한 것으로 나타났다. 이는 타 분야에 비해 건축 분야에서 활용할 수 있는 AI 학습용 데이터가 부족하여 연구 진행에 어려움이 있기 때문으로 판단된다.

AI 기술의 건축 설계 분야 적용 활성화를 위해서는 AI 학습용 데이터 확보가 필수적이며, 데이터의 수량과 질에 따라 AI의 성능 또한 달라진다. 이에 건축 설계 분야 AI 연구가 활발해지며 다양한 건축 관련 데이터셋의 구축으로 이어졌다. 그러나 공개된 데이터의 대부분이 건축물의 이미지 형상 정보만을 대상으로 구축되어 있다. 건축 설계 분야에서 필요로 하는 건축 도면을 대상으로 한 데이터셋은 그 수가 현저히 적으며, 그마저도 일부만 공개된 실정이다.

* 경북대 대학원 석사과정
** 경북대 대학원 박사수료
*** 경북대 건축학부 교수, 공학박사(Dr.-Ing.)
(Corresponding author : School of Architecture, Kyungpook National University, choo@knu.ac.kr)

본 연구는 국토교통부/국토교통과학기술진흥원의 2024년도 지원으로 수행되었음(과제번호 RS-2021-KA163269). 본 연구는 한국연구재단의 지원으로 수행되었음(NO. RS-2024-00349586).

따라서 본 연구는 현재 건축 설계 분야의 AI 활용 연구에서 사용되고 있는 국내·외 건축 도면 데이터셋의 구축 현황을 분석하고, 이를 통해 향후 건축 도면 데이터셋의 구축 방향을 제안하고자 한다. 건축 도면 데이터셋은 AI를 건축 설계 분야에 도입하기 위한 연구 중에서도 특히 활발한 자동 평면 설계에서 필요한 평면 데이터로 한정하였다.

2. 국내·외 건축 도면 데이터셋 현황

본 연구에서는 데이터 규모, 클래스 규모, 원천 데이터의 형식 등을 고려하여 건축 도면 데이터셋을 표1과 같이 살펴보았으며, 이 중 최대 규모 데이터, 클래스 규모, CAD 및 다중 유닛 데이터의 원천 데이터 사용 등의 특징을 가진 데이터셋을 비교·분석 대상으로 선정하였다.

2.1 AI 허브 건축 도면 데이터

국내 AI 통합 플랫폼인 AI 허브를 통해 얻을 수 있는 건축 도면 데이터는 국내 최대 규모의 건축 도면 데이터셋이다. 해당 데이터셋은 한국에서 흔히 볼 수 있는 형태의 주택유형별(아파트, 연립 다세대, 단독주택) 도면(평면도, 단면도, 입면도, 구조도)인 48,033장의 원천 데이터(PNG)와 도면 내 주요 객체의 위치 및 크기를 식별할 수 있는 라벨이 포함된 라벨링 데이터(JSON)로 구성되어 있다.

원천 데이터는 아파트 도면이 전체의 80%로 가장 많은 분포를 차지하고 있으며, 라벨링 데이터는 구조 8종, 공간 12종, 객체 5종의 총 25개 클래스로 구분되어 있다.

표1. 건축 도면 데이터셋 제공 현황

No.	데이터셋	공개년도	제공 여부*
1	FPLAN-POLY	2010	△
2	SESYD	2010	○
3	CVC-FP	2015	○
4	Rent3D(R3D)	2015	△
5	SydneyHouse	2016	○
6	R-FP-500	2017	△
7	ROBIN	2017	△
8	Raster-to-Vector(R2V)	2017	○
9	BRIDGE	2019	○
10	Cubicasa5K	2019	○
11	RPLAN	2019	○
12	BTI	2020	×
13	LIFULL	2021	△
14	ZSCVFP	2021	×
15	REP	2021	×
16	RuralHomeData	2021	×
17	FloorPlanCAD	2021	○
18	AI허브 건축 도면 데이터	2022	○
19	MLSTRUCT-FP	2023	○
20	Modified Swiss Dwellings(MSD)	2024	○

*△ : 제공하는 것으로 표기되어 있으나 확인할 수 없음

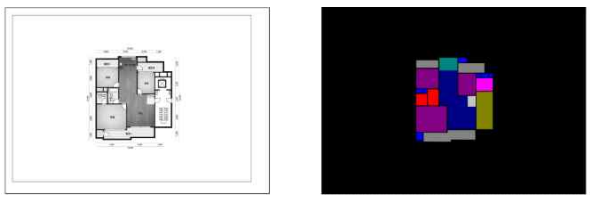

그림1. AI 허브 건축 도면 데이터,
(Semantic Segmentation) 원천 데이터-라벨링 데이터

2.2 CubiCasa5K

CubiCasa5K 데이터셋은 벽이나 방과 같은 건축물의 구성 요소뿐 아니라, 가구나 설비 같은 객체를 포함한 방대한 클래스 규모가 특징으로, 핀란드의 15,000개 대규모 평면도 이미지 세트에서 수집하고 검토한 5,000개의 평면도로 구성되어 있다. 평면도는 다양한 규모의 단일 유닛으로 구성되어 있으며, 레스터화 된 도면 이미지를 원천 데이터(PNG)와 벡터 그래픽 형식의 라벨링 데이터(SVG)를 제공한다. 데이터의 클래스는 주거 건물의 평면도에 일반적으로 나타나는 방, 벽, 문, 창문 등의 주요 객체와 구조를 중심으로 구분되어 있고, 화장실의 변기와 같은 비품 객체와 소파나 테이블과 같은 가구의 라벨 등을 포함하여 총 80개 이상의 평면도 객체 라벨을 제공한다.

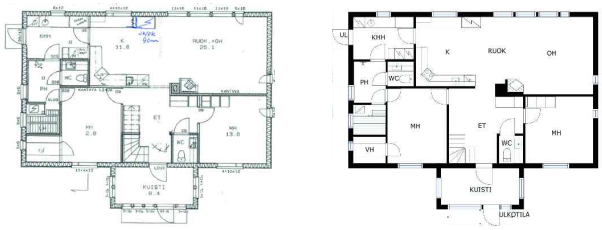

그림2. CubiCasa5K, 원천 데이터-라벨링 데이터

2.3 RPLAN

RPLAN 데이터셋은 건축 도면 관련 데이터셋 중 최대 규모의 데이터셋이며, 다양한 크기와 형태를 가진 8만 개 이상의 평면도 데이터로 구성되어 있다. 각 평면도는 18m×18m의 정사각형 영역 내에서 벡터 그래픽 형식으로 표현되었으며, 인공지능 학습에 최적화된 사이즈인 256×256으로 변형되었다.

해당 데이터셋은 원천 데이터가 포함되어 있지 않고, 라벨링 데이터(PNG)로만 구성되어 있으며, 라벨링 데이터인 각 평면도는 4채널 이미지로 표현되어 있다. 첫 번째 채널은 건물 외벽이나 주 출입구와 같은 건물 외곽선과 관련된 정보를 나타내며, 두 번째 채널은 거실, 주방, 욕실과 같은 방의 유형을 나타낸다. 세 번째 채널은 라벨이 같은 서로 다른 방을 구분하는 데 사용하는 정숫값이 기록되어 있고, 네 번째 채널은 외부 영역과 내부 영역을 구분하는 정보를 담고 있다.

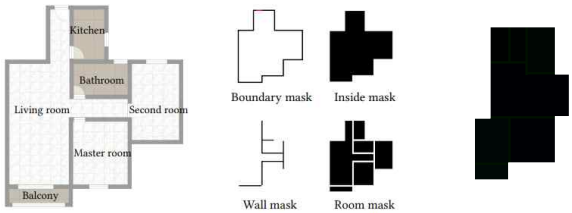

그림3. RPLAN, 원천 데이터-채널 추상화-라벨링 데이터

2.4 FloorPlanCAD

FloorPlanCAD 데이터셋은 원천 데이터로 여러 산업 분야의 실제 건축 프로젝트에서 수집한 대규모 CAD 도면 데이터를 이용하여 건축 도면 이미지 데이터를 원천 데이터로 사용하는 보편적인 데이터셋과 차이가 있다. 데이터는 100개 이상의 주거 건물, 학교, 병원, 대형 쇼핑몰 등 다양한 용도의 프로젝트로부터 확보한 2,500개의 평면도의 레이어를 분리하는 전처리 과정을 거쳤다.

데이터셋은 CAD 도면을 레스터화 한 이미지 데이터(PNG)와 벡터 그래픽 형식의 라벨링 데이터(SVG), COCO (Common Objects in Context) 데이터셋 형식에 맞춘 시각화 결과물 데이터(PNG)로 구성되어 있다. 원천 데이터는 CAD 도면의 전체 이미지가 아닌 도면의 일부를 잘라 가로×세로 1,000픽셀에 맞춘 사이즈로 변형시켰다. 클래스는 벽, 창, 문, 가구 등을 포함한 총 35개의 객체 클래스로 구분되어 있으며, 파노픽 심볼 스포팅(Panoptic Symbol Spotting)을 위한 라벨을 포함하고 있어 문, 창문, 전기 기구, 계단 등의 위치와 형상을 정확하게 인식할 수 있는 것이 주요 특징이다. 이는 벽이나 기둥 같은 구조뿐만 아니라 평면도에 포함된 다양한 종류의 기호를 구분하고, 각 기호와 주변 요소의 관계를 학습하여 복잡한 도면을 분석하는 작업에 최적화 되어있다.

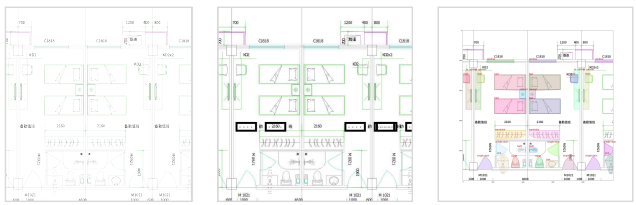

그림4. FloorPlanCAD, 원천 데이터-라벨링 데이터-COCO 데이터

2.5 MLSTRUCT-FP

MLSTRUCT-FP 데이터셋은 대규모 다중 유닛 평면도를 원천 데이터로 사용한 최초의 데이터셋으로, 165개의 칠레 주거용 건물 프로젝트에서 얻은 954개의 고해상도 평면도 이미지로 구성되어 있다. 해당 데이터셋은 특히, 각 유닛의 벽과 슬래브가 폴리곤 형태로 라벨링 되어 있어 건축 평면도의 구조적 요소를 자동으로 인식하고 분석하는 연구에 적합하다. 또한, 단일 유닛의 평면도가 아닌 다중 유닛 평면도로 구성되어 있어 다른 데이터셋보다 복잡한 벽 구조와 복도나 주차장 등과 같은 공용 구역에 대한 요소가 포함되어 있어 대규모 건물의 구조 분석에 유용하다.

데이터셋에 포함된 원천 데이터(PNG)는 평균적으로 9,450픽셀 너비와 6,700픽셀 높이의 범위 내인 다양한 크기로 구성되어 있고, 배경을 제거하여 추가적인 분석 작업에 용이하게 했다. 라벨링 데이터(JSON)는 벽과 슬래브 두 가지 클래스에 대한 라벨이 포함되어 있으며, 각각의 벽과 슬래브는 각 객체의 기하학적 형상과 위치가 상세히 기록되어 좌표 위치, 길이, 두께, 각도와 같은 정보를 포함하고 있다. 데이터셋에 포함된 클래스는 기본적으로 벽과 슬래브 두 가지 클래스만을 제공하고 있지만, 필요에 따라 문, 창문, 가구 등의 클래스 및 라벨을 수동으로 추가할 수 있는 확장성을 가지고 있다.

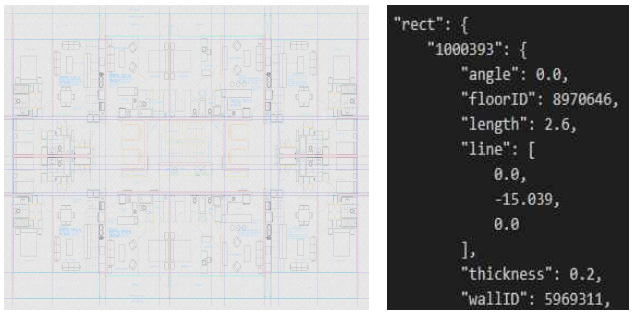

그림5. MLSTRUCT-FP, 원천 데이터-라벨링 데이터일부

2.6 Modified Swiss Dwellings(MSD)

MSD 데이터셋은 다세대 주택의 레이아웃이 포함된 대규모 다중 유닛 평면도를 원천 데이터로 사용한 데이터셋이다. MSD는 스위스 주택 데이터베이스(Swiss Dwellings)에서 파생되었으며, 가구나 계단과 같은 요소를 제거하는 정제 과정과 함께 주거용 이외의 도면, 같은 단지 내에 위치하여 동일한 배치를 가진 도면, 일정 규모 미만의 도면, 단일 유닛 도면 등을 제거하는 필터링을 거쳐 얻은 총 5,372개의 평면도에 대한 라벨링 데이터로 구성되어 있다. 라벨링 데이터는 평면도 이미지(NPY)와 그래프(PICKLE), 필수 구조 요소 이미지(NPY)의 세 가지 유형 데이터를 포함하고 있다.

해당 데이터셋의 클래스는 유형과 구역의 범주에서 두 가지 유형으로 구분되어 있으며, 두 가지 유형 모두 구조와 개구부를 포함하고 있다. 유형에 기반한 클래스는 침실이나 거실과 같은 방을 기준으로 하며, 구역에 기반한 클래스는 Khodabakhshi, Khaghani, & Garmaroodi[1]에 따라 개인 공간, 공용공간, 서비스 공간, 외부 공간으로 나뉜다. 해당 데이터셋은 소규모의 단일 주거 평면도보다 중대형의 다세대 아파트 평면도를 생성하기 위한 학습에 최적화되어있다.

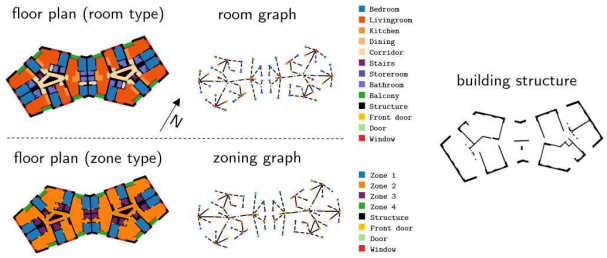

그림6. Modified Swiss Dwellings 라벨링 데이터, 평면도 이미지-그래프-필수 구조 요소

3. 국내·외 건축 도면 데이터셋 비교 및 분석

건축 도면 데이터는 여전히 데이터의 규모가 작다는 문제점을 갖고 있다. AI 활용이 활발한 다른 분야에서는 수십만 개 이상의 데이터를 학습에 활용하는 반면, 건축 도면 데이터는 만 개 이상의 데이터를 확보하기조차 어려운 상황이다. 공개된 데이터 또한 그 수가 적어 쉽게 접근하고 연구에 활용하기 힘들다. 건축 도면 데이터 중 가장 많은 데이터를 제공하는 RPLAN 데이터셋은 원천 데이터를 제공하지 않는다는 아쉬움이 있다.

데이터의 대상이 되는 건축물 용도의 경우, 공개되어 있는 대부분의 데이터셋은 주거용 건축을 대상으로 하고 있으며, 데이터셋의 대부분이 단일유닛에 국한되어 있다. MLSTRUCT-FP 데이터셋과 MSD 데이터셋의 경우 다중 유닛을 대상으로 하고 있지만, MLSTRUCT-FP 데이터셋의 경우 1,000개 미만의 데이터를 제공하고 있고, 가장 최신 데이터라고 할 수 있는 MSD 데이터셋도 5,000개가량의 데이터를 제공하여 타 분야와 비교하였을 때, 규모가 크다고는 볼 수 없다.

원천 데이터의 경우에는 대부분의 데이터셋에서 PNG 형식의 건축 도면 이미지를 원천 데이터로 제공하고 있다. 이 중 FloorPlanCAD 데이터셋은 CAD 도면인 DWG 파일을 가공하여 원천 데이터를 생성했다는 차이점이 있다. 하지만 DWG 파일은 데이터셋에 포함되지 않고, DWG 파일을 가공한 PNG 파일을 기준으로 데이터 라벨링이 이루어져 CAD 파일에 대한 이점은 크게 두드러지지 않는다.

원천 데이터가 대부분 PNG 파일로 일관된 것에 반해 라벨링 데이터는 다양한 포맷으로 제공되고 있다. 이는 건축 도면 데이터에서 어떤 부분에 중점을 두고 연구를 진행하는가에 따라 라벨링 유형이 달라지기 때문으로 보인다. 데이터 라벨링 유형은 시멘틱 세그멘테이션이 많이 사용되었으며, 최근에는 폴리곤 라벨링도 사용되고 있다. 클래스의 경우, 대부분의 데이터셋이 건축 구조와 방의 유형을 중점으로 하였다. CubiCasa5K와 FloorPlanCAD 데이터

1) Khodabakhshi, K., Khaghani, S., & Andaji Garmaroodi, A. (2022). A Procedural Approach for Configuration of Residential Activities Based on Users' Needs and Architectural Guidelines. Nexus Network Journal, 24(3), pp. 787-808

표2. 건축 도면 데이터셋 비교

구분	AI 허브 건축 도면 데이터	CubiCasa5K	RPLAN	FloorPlanCAD	MLSTRUCT-FP	Modified Swiss Dwellings(MSD)
데이터 규모 (도면 수량)	48,033	5,000	80,787	15,663	954	5,372
대상 건축물 용도	아파트, 연립다세대, 단독주택	아파트, 단독주택	주거용 건축물	주거용 건축물, 학교, 병원, 대형 쇼핑몰 등	다중 유닛(주거 이외 용도 포함)	다세대 주택 (다중 유닛)
원천 데이터 포함 여부	○	○	×	△[2]	○	△[3]
원천 데이터 형식	PNG	PNG	-	DWG, PNG	PNG	PNG
라벨링 유형	Bounding Box, Semantic Segmentation	Heatmap, Semantic Segmentation	Semantic Segmentation	Panoptic Segmentation	Polygon, Semantic Segmentation	Polygon
라벨링 데이터 형식	JSON	SVG	PNG	SVG	JSON	NPY, PICKLE
클래스 규모(수량)	25	12+ α	18	35	2	17
피인용 수(회)	-	122	213	40	3	1

셋은 가구와 같은 객체 요소를 클래스 종류로 추가하여 클래스의 규모를 키웠으며, MLSTRUCT-FP 데이터셋의 경우에는 벽과 슬라브의 구조 요소만 클래스로 선정한 것이 특징이다.

또한, 대부분의 데이터가 건축 도면 내 객체에 중점을 두고 라벨링 데이터를 제작한 것으로 보인다. 일반적으로 라벨링 데이터 내에는 객체의 위치 정보가 포함되어 있지만, 이러한 정보만으로는 객체들의 연결성이나 도면 내의 동선을 설명하기 어렵다. 이에 MSD 데이터셋은 다른 데이터셋에는 포함되지 않는 그래프 파일을 데이터셋에 포함시켜 이를 해결했다. 다른 데이터셋과 비교하여 MSD 데이터셋은 각 공간 간의 연결을 표현한 그래프 데이터를 추가하는 것으로 건축 도면에서 중요하게 여겨지는 공간의 연결성에 더 집중하게 만들었다.

4. 결론

AI 기술의 발전이 가속화됨에 따라 AI 기술에 필연적인 데이터셋에 대한 수요 또한 증가하고 있다. 그러나 건축 도면과 관련한 분야에서 사용하는 데이터의 경우, 데이터의 다양성과 그 규모에 있어 미흡함을 확인하였다. 본 연구에서 분석한 데이터셋의 대부분이 평면 자동 생성 분야와 자동 분석 및 인식 분야에서 사용되고 있지만, 해당 분야에서 앞서나가고 있는 연구들의 대부분이 대규모 데이터셋을 필요로 하여 RPLAN 데이터를 사용하고 있음을 고려하면 데이터의 규모 문제는 해결 방안이 시급할 것으로 판단된다. 또한 현재까지 공개된 대부분의 건축 도면 데이터셋은 건축 도면에서 중요하게 여겨지는 요소들 간의 관계성을 포함하는 데이터에 대한 고려가 미흡하다고 판단되어 추후, 건축 도면 내 요소들 간의 관계성에 더 중점을 둔 데이터셋의 추가적인 구축이 필요할 것으로 사료된다.

2) DWG 파일 미제공
3) Swiss Dwellings 데이터베이스에서 확인

참고문헌

1. 한국지능정보사회진흥원, 인공지능 학습용 데이터 품질 관리 가이드라인 및 구축 안내서 v3.0, 2023
2. 조용현, Mask R-CNN 기반 심층학습을 이용한 개체영상의 인공지능 학습데이터 구축, 2022
3. Ahti, K., Juha, Y., Markus, H., Antti, K., & Juho, K. (2019). CubiCasa5K: A Dataset and an Improved Multi-task Model for Floorplan Image Analysis. In M. Felsberg, P-E. Forssén, I-M. Sintorn, & J. Unger (Eds.), Image Analysis, pp. 28-40
4. Wu, W., Fu, X., Tang, R., Wang, Y., Qi, Y. & Liu, L. (2019). Data-driven interior plan generation for residential buildings, ACM Transactions on Graphics (TOG), Volume 38, Issue 6, pp. 1-12
5. Fan, Z., Zhu, L., Li, H., Zhu, S., & Tan, P. (2021). FloorPlanCAD: A Large-Scale CAD Drawing Dataset for Panoptic Symbol. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)
6. Pizarro, P. N., Hitschfeld, N., & Sipiran, I. (2023). Large-scale multi-unit floor plan dataset for architectural plan analysis and recognition. Automation in Construction, 156, 105132
7. Engelenburg, C., Mostafavi, F., Kuhn, E., Jeon, Y., Franzen, M., Standfest, M., Gemert, J. & Khademi, S. (2024). MSD: A Benchmark Dataset for Floor Plan Generation of Building Complexes, arXiv:2407.10121v3