비전 언어 모델을 사용한 텍스트 조건부 평면도 검색 기법

한성민, 심종화, 김은빈, 황인준
고려대학교 전기전자공학과
{tjdals2038, indexlibrorum3822, gichanac, ehwang04}@korea.ac.kr

Text-Conditional Floor Plan Retrieval Scheme Using Vision Language Model

Seongmin Han, Jonghwa Shim, EunBeen Kim and Eenjun Hwang
School of Electrical Engineering, Korea University

요 약

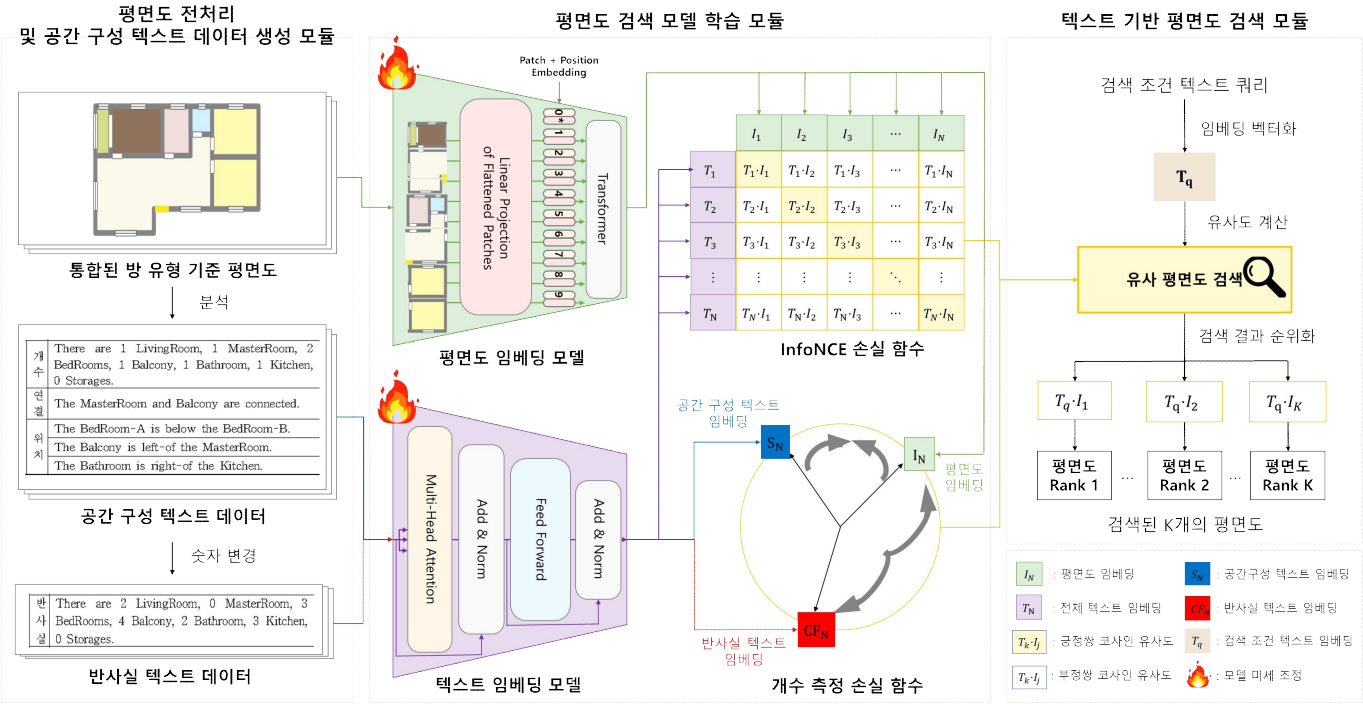
건축 설계의 초기 단계에서는 발주자와 설계자 간의 의사소통을 돕고 설계 업무의 효율성을 증대시키기 위해 요구사항이 유사한 평면도를 검색한다. 기존의 평면도 검색 시스템은 방의 유형, 개수, 면적과 같이, 제한적으로 입력된 메타데이터의 일치 여부만으로 평면도를 검색하기 때문에, 방 간의 연결 관계나 위치 관계와 같은 공간의 의미론적인 특성을 반영하여 검색하지 못한다. 이러한 문제를 해결하기 위해, 본 논문은 평면도를 분석하여 방의 개수, 연결 관계, 위치 관계를 포함한 공간 구성 텍스트 데이터를 자동으로 생성하고, 이미지 임베딩 모델과 텍스트 임베딩 모델을 미세 조정하여, 공간 구성 텍스트의 의미론적인 특성을 반영할 수 있는 텍스트 조건부 평면도 검색 기법을 제안한다. 또한 방의 개수와 같은 수량 정보에 대한 검색 성능을 향상시키기 위해 임베딩 모델을 미세 조정하는 단계에서 개수 측정 손실 함수를 함께 적용하였다. 비교 실험 및 절제 연구 결과, 제안 기법은 기존 방법보다 검색 성능이 최대 54% 향상되었으며, 개수 측정 손실 함수를 적용한 경우에는 검색 성능이 최대 8.41% 추가로 개선되었다.

1. 서 론

건축 설계의 초기 단계에서는 발주자가 요구하는 공간적인 요구사항을 정확하게 파악하고, 이를 설계에 반영하기 위해 이전 프로젝트들의 평면도를 검색하여 참고한다. 기존의 평면도 검색 시스템은 평면도를 등록할 때 함께 입력한 방의 유형, 개수, 면적 등과 같은 메타데이터의 매칭을 기반으로 검색하는 원리이다. 그러나 이러한 매칭 기반 평면도 검색 시스템은 사전에 저장할 수 있는 메타데이터의 양이 제한되어 있기 때문에 평면도에 대한 모든 정보를 담을 수 없고, 단순히 메타데이터의 일치 여부를 기준으로 검색하기 때문에 방 간의 연결 관계나 위치 관계와 같은 의미론적인 특성을 반영하여 검색할 수 없다는 한계가 있다.

최근 이러한 매칭 기반 검색 시스템의 한계를 극복하기 위해, 이미지 또는 텍스트의 의미론적인 특성을 이해하고 검색에 반영[1]할 수 있는 벡터 기반 검색 기술이 주목받고 있다. 벡터 기반 검색 기술은 검색 대상과 검색 조건 쿼리를 각각 임베딩 벡터로 변환한 뒤, 두 벡터 간의 유사도를 학습해두고, 새로운 검색 조건 쿼리가 입력되면 가장 높은 유사도를 갖는 검색 대상을 반환하는 기술이다. 이를 활용하면 사용자의 검색 의도를 보다 명확하게 반영할 수 있기 때문에 필요하다. 하지만 평면도 검색 분야에서는 방 개수, 연결 관계, 위치 관계 등 평면도의 의미론적인 특성을 설명하는 텍스트 데이터셋이 부재하기 때문에 활용이 미진하다.

따라서 본 논문은 평면도를 분석하여 방의 개수, 연결 관계, 위치 관계를 표현하는 공간 구성 텍스트 데이터를 생성하고, 이를 활용하여 벡터 기반 이미지 임베딩 모델과 텍스트 임베딩 모델을 미세 조정함으로써, 공간 구성 텍스트의 의미론적인 특성을 반영하여 검색을 할 수 있는 텍스트 조건부 평면도 검색 기법을 제안한다. 제안하는 기법은 평면도 전처리 및 공간 구성 텍스트 데이터 생성, 평면도 검색 모델 학습, 텍스트 기반 평면도 검색 세 가지 단계로 구성된다. 먼저, 평면도 전처리 및 공간 구성 텍스트 데이터 생성 단계에서는 용도가 유사한 방의 유형을 통합하고, 통합된 방의 유형을 기준으로 색상을 부여하여 새로운 평면도를 생성한다. 이후 새로운 평면도를 분석한 정보를 텍스트 형태로 표현하여 평면도-텍스트 쌍 데이터셋을 구축한다. 평면도 검색 모델 학습 단계에서는 이전 단계에서 생성한 학습 데이터셋을 활용하여 벡터 기반 검색을 수행하는 평면도 검색 모델의 평면도 임베딩 모델과 텍스트 임베딩 모델을 학습한다. 이때, 방의 개수와 같은 수량 정보에 대한 검색 성능을 향상시키기 위해 개수 측정 손실 함수[2]를 적용한다. 마지막으로, 텍스트 기반 평면도 검색 단계에서는 사용자가 입력한 검색 조건 텍스트 쿼리와 데이터베이스에 저장된 모든 평면도의 임베딩 간의 유



(그림 1) 제안하는 기법의 전체 구성도

사도를 계산하여 의미적으로 일치하는 평면도의 순위를 매긴다. 본 논문의 2장에서는 제안 기법에 대해서 설명하고, 3장에서는 실험 및 결과에 대해 기술한다. 마지막으로 4장에서는 결론을 제시한다.

2. 제안 기법

제안하는 기법의 전체적인 구성은 그림 1과 같이 총 세 단계로 이루어져 있다. 2.1절은 평면도 전처리 및 공간 구성 텍스트 데이터 생성에 대해 다루고, 2.2절은 평면도 검색 모델 학습, 2.3절은 텍스트 기반 평면도 검색에 관해 서술한다.

2.1. 평면도 전처리 및 공간 구성 텍스트 데이터 생성

본 절에서는 평면도 검색 모델이 공간 구성에 대한 의미론적 검색 능력을 훈련할 수 있도록 평면도-텍스트 쌍 데이터셋을 구축하며, 이를 위해 평면도 전처리 단계와 공간 구성 텍스트 데이터 생성 단계를 순차적으로 수행한다. 먼저, 평면도 전처리 단계에서는 기존의 방 유형을 비슷한 용도를 기준으로 재분류하여 표 1과 같이 통합된 방 유형으로 재구성하였다. 이후 통합된 방 유형을 기준으로 색상을 부여하여 새로운 평면도를 생성한다.

공간 구성 텍스트 데이터 생성 단계에서는 새롭게 생성된 평면도의 픽셀을 순차적으로 탐색하면서, 서로 연결된 픽셀 그룹을 하나의 방 영역으로 묶고 고유한 라벨을 부여함으로써 각 방마다 개수를 산출하였다. 이때, 존재하지 않는 방 유형에 대해서도 개수를 '0'으로 명시하고, 모든 수량 정보는 숫자 형태로 통일하여 텍스트 데이터의 일관성을 유지하였다. 방의 연결 관계는 두 방 사이에 문이 존재할 때 '연결됨 (Connected)'으로 표현하였으며, 동일한 유형의 방이 두 개 이상 존재할 경우 탐색된 순서에

<표 1> 기존의 방 유형 통합과 통합된 방의 색상 표시

번호	기존 방 유형	통합된 방 유형	통합된 방 색상
1	LivingRoom	LivingRoom	Beige
2	MasterRoom	MasterRoom	Brown
3	Kitchen	Kitchen	Pale Pink
4	Bathroom	Bathroom	Light Blue
5	DiningRoom	**Kitchen**	Light Pink
6	ChildRoom	**BedRoom**	Light Yellow
7	StudyRoom	**BedRoom**	Light Yellow
8	SecondRoom	**BedRoom**	Light Yellow
9	GuestRoom	**BedRoom**	Light Yellow
10	Balcony	Balcony	Olive Green
11	Entrance	**LivingRoom**	Beige
12	Storage	Storage	Light Orange
13	Wall-in	**Storage**	Light Orange
14	ExternalArea	ExternalArea	White
15	ExteriorWall	**Wall**	Dark Gray
16	FrontDoor	FrontDoor	Yellow
17	InteriorWall	Wall	Dark Gray
18	InteriorDoor	InteriorDoor	White

따라 이름 뒤에 '-A', '-B' 등과 같은 접미사를 추가하여 독립된 객체로 구분하였다. 마지막으로 위치 관계는 중점 공식을 통해 라벨링된 방들의 중심 좌표를 계산하고, 이를 기반으로 상대적인 위치를 계산하여 상하좌우 대각선과 같은 팔방위로 표현하였다. 구체적인 팔방위 표현은 다음과 같다. "왼쪽 위에(Left-above)", "왼쪽에(Left-of)", "왼쪽 아래(Left-below)", "아래에(Below)", "오른쪽 아래(Right-below)", "오른쪽에(Right-of)", "오른쪽 위에(Right-above)", "위에(Above)", "내부에(Inside)", "둘러싸고 있는(Surrounding)". 이때 위치 관계는 필연적으로 방의 인접 관계를 내포하므로 별도의 인접 관계 표현 없이 위치 관계로 통합하였다.

2.2. 평면도 검색 모델 학습

평면도 검색 모델 학습 단계에서는 이전 단계에서 구축한 평면도-텍스트 쌍 데이터셋을 사용하여, 공간 구성에 대해 의미론적인 검색을 할 수 있도록, 평면도 임베딩 모델과 텍스트 임베딩 모델을 학습한다. 평면도 임베딩 모델은 입력된 평면도의 시각적 특징들을 임베딩 벡터로 변환하고, 텍스트 임베딩 모델은 입력된 공간 구성 텍스트 데이터를 임베딩 벡터로 변환한다. 이후 각 임베딩 모델을 통해 변환된 벡터들을 내적하여 공유 임베딩 공간으로 매핑하고, 의미적으로 일치하는 긍정쌍의 코사인 유사도는 최대가 되도록, 의미적으로 일치하지 않는 부정쌍 간의 코사인 유사도는 최소가 되도록 InfoNCE 손실 함수(1)를 사용하여 대조 학습[3]을 수행한다.

$$L_{InfoNCE} = -\frac{1}{N}\sum_{k=1}^{N}\log\frac{\exp(e_t^k \cdot e_{fp}^k)}{\sum_{j=1}^{N}\exp(e_t^k \cdot e_{fp}^j)} \quad (1)$$

여기서, $L_{InfoNCE}$ 는 텍스트 쿼리 중심으로 평면도를 검색하기 위한 단방향 손실 함수이며, N은 배치 크기, e_t^k는 k번째 텍스트 임베딩, e_{fp}^k는 k번째 평면도 임베딩, $e_t^k \cdot e_{fp}^k$는 정답 임베딩 쌍의 유사도, $e_t^k \cdot e_{fp}^j$는 텍스트 임베딩과 전체 평면도 임베딩 쌍의 유사도이다.

또한, 대조 학습을 수행할 때 방의 개수와 같은 수량적인 정보를 반영한 검색 성능을 향상시키기 위해 개수 측정 손실 함수(2)를 추가로 적용하였다. 개수 측정 손실 함수(Count Loss)는 원본 텍스트와 원본에서 숫자만 다르게 변경한 반사실(Counterfactual) 텍스트를 함께 사용하여, 정확한 수량 정보를 가진 텍스트와 해당 평면도의 임베딩 백터 간 거리를 서로 가까운 벡터 공간으로, 잘못된 숫자를 가진 텍스트와 평면도의 임베딩 백터는 먼 벡터 공간으로 매핑함으로써 모델이 숫자의 차이를 명확하게 구별하도록 학습을 유도하였다.

$$L_{count} = -\frac{1}{N}\sum_{k=1}^{N}\log\frac{\exp(e_t^k \cdot e_{fp}^k)}{\exp(e_t^k \cdot e_{fp}^k)+\exp(e_{t-cf}^k \cdot e_{fp}^k)} \quad (2)$$

여기서, L_{count}는 텍스트에 표기된 수량이 평면도와 일치하는지 판단하는 손실 함수이며, e_{fp}^k는 평면도 임베딩, e_t^k는 정답 텍스트 임베딩, e_{t-cf}^k는 반사실 텍스트 임베딩이다. 결과적으로, 평면도 검색 모델을 학습할 때 사용한 최종 손실 함수는 식 (3)과 같이 InfoNCE 손실 함수와 개수 측정 손실 함수를 더한 형태이다.

$$L = L_{InfoNCE} + L_{Count} \quad (3)$$

2.3. 텍스트 기반 평면도 검색

텍스트 기반 평면도 검색 방법은 임베딩 벡터화, 유사 평면도 검색, 검색 결과 순위화 세 가지 단계로 구성된다. 먼저, 임베딩 벡터화 단계는 이전 단계에서 학습한 텍스트 임베딩 모델을 활용하여 사용자가 입력한 검색 조건 텍스트 쿼리를 벡터화한다. 다음 유사 평면도 검색 단계에서는 데이터 베이스에 저장되어 있는 모든 평면도를 평면도 임베딩 모델을 사용하여 벡터화하고, 입력된 텍스트와 평면도 간의 코사인 유사도를 측정한다. 마지막으로 검색 결과 순위화 단계에서는 측정된 코사인 유사도를 내림차순으로 정렬하여 유사한 평면도들의 순위를 매긴다.

3. 실험 및 결과

3.1. 평면도 검색 모델 및 학습 데이터셋 구성

본 실험은 평면도 임베딩 모델로 잔차 신경망(ResNet)[4]과 비전 트랜스포머(Vision Transformer)[5] 인코더, 텍스트 임베딩 모델로 트랜스포머(Transformer)[6] 인코더를 사용하였다. 학습 데이터는 약 80,000장의 아시아 주거용 평면도로 구성된 RPLAN[7] 데이터셋을 활용하여, 재구성된 평면도 이미지와 공간 구성 텍스트 데이터를 생성하고, 훈련, 검증, 테스트 세트 6:2:2 비율로 분할하여 평면도-텍스트 학습 데이터셋을 구축하였다.

3.2. 실험 방법 및 환경

평면도 임베딩 모델의 학습 성능 향상을 위해, 전체 평면도 이미지에 대한 평균과 분산을 사전에 계산하여 평면도 벡터화 모델에 반영하였다. 텍스트 데이터는 방의 개수를 명시하는 문장 한 줄을 기본으로 구성하였으며, 모델의 일반화 성능을 확보하기 위해 매 에폭마다 연결 관계 및 위치 관계 문장을 1~4개 랜덤하게 추가하였다. 또한 개수 정보를 명확하게 학습시키기 위해, 개수 측정 손실 함수를 사용할 때 방 개수 문장의 모든 숫자를 0~4 사이의 랜덤한 값으로 변경한 반사실 텍스트 데이터를 생성하여 학습에 활용하였다. 마지막으로, 사용자가 "There are 3 BedRooms, 2Bathroom." 같이 일부 정보만으로도 정확한 검색 결과를 얻을 수 있도록, 전체 텍스트 데이터의 문장 길이를 랜덤하게 조정하여 학습하였다. 실험은 Intel i7-9700 CPU와 NVIDIA TITAN RTX(24GB) GPU 환경에서 진행되었다. 학습에 사용한 주요 하이퍼 파라미터는 배치 크기(Batch size)를 64로 설정하였으며, 총 300 에폭(Epoch)동안 학습하였다. 학습률(Learning Rate)은 5e-6, 가중치 감소(Weight Decay)는 1e-2로 설정하였으며, 초반 학습의 안정화를 위해 웜업(Warm-up) 에폭 수를 5로 설정하였다.

3.3. 평가 지표

본 실험은 공간 구성 텍스트를 입력하여 의미적으로 유사한 평면도를 검색하는 작업을 수행한다. 평면도 검색 모델의 성능 평가는 이미지 검색 분야에서 일반적으로 사용하는 Recall at K(R@K) 지표를 사용하였다. Recall at K는 전체 쿼리(Query) 중에서 상위 K 개의 검색 결과 내에 최소 하나 이상의 정답 이미지를 포함하는 쿼리의 비율로 정의되며, 이는 사용자가 상위 K개의 결과만을 주로 확인하는 실질적인 검색 환경을 고려한 지표로 식 (4)과 같다.

$$Recall\ at\ K = \frac{1}{N}\sum_{i=1}^{N}\frac{|R_i \cap S_i^K|}{|R_i|} \quad (4)$$

<표 2> InfoNCE와 개수 측정 손실 함수를 사용한 평면도 검색 모델의 텍스트 조건부 평면도 검색 성능

Loss Function	Model	Recall at 10 (R@10)	Recall at 20 (R@20)	Recall at 30 (R@30)	Recall at 50 (R@50)
InfoNCE	RN50	32.30	40.45	45.84	52.95
Ours	**RN50**	**36.96**	**46.93**	**53.08**	**61.12**
InfoNCE	ViT-B/16	33.80	42.23	47.58	54.07
Ours	**ViT-B/16**	**37.77**	**48.13**	**54.54**	**62.48**

여기서 N은 전체 텍스트 쿼리의 개수, R_i는 i번째 쿼리에 대한 정답 이미지 집합, S_i^K는 i번째 쿼리에 대해 상위 K개의 검색 결과 이미지 집합, $|\cdot|$는 집합의 크기이다. 각 쿼리에는 하나 이상의 정답 이미지가 존재할 수 있으며, Recall at K는 이 중 최소 하나의 정답 이미지가 상위 K개의 검색 결과 내에 존재하는지 여부만을 측정한다. 최종적으로 모든 쿼리에 대해 Recall at K를 계산한 뒤 평균을 취하여 성능 평가의 지표로 활용하였다.

3.4. 실험 결과

학습된 평면도 검색 모델의 의미론적인 검색 성능과 개수 측정 손실 함수의 영향을 평가하기 위해 비교 실험 및 절제 연구를 진행하였다. 비교 실험에서는 잔차 신경망과 비전 트랜스포머 두 가지 이미지 임베딩 모델의 성능을 비교하였으며, 절제 연구에서는 기본적인 InfoNCE 손실 함수만 적용한 경우와 개수 측정 손실 함수를 함께 적용한 경우로 나누어 성능을 평가하였다.

먼저, 잔차 신경망(RN50)에 InfoNCE 손실만 적용하여 학습한 경우, 방의 연결 관계 및 위치 관계를 전혀 이해하지 못한 사전 학습된 벡터 기반 검색 모델보다 R@10 기준으로 32% 검색 성능이 향상되어 공간 구성에 대한 의미론적인 이해도가 향상되었다. 또한 제시된 표 2의 결과에 따르면, 모든 평가에서 개수 측정 손실을 추가한 모델이 일관되게 더 우수한 성능을 나타냈으며, K값이 증가할수록 성능 향상 폭도 함께 증가하는 경향을 보였다. 구체적으로 R@10의 경우 32.30%에서 36.96%(+4.66%p), R@20은 40.45%에서 46.93%(+6.48%p), R@30은 45.84%에서 53.08% (+7.24 %p)로 성능이 향상되었다. 또한 비전 트랜스포머 모델(ViT-B/16)에서도 잔차 신경망과 동일한 경향성을 보였으며, 특히 비전 트랜스포머의 R@50은 54.07%에서 62.48% (+8.41%p)로 전체 실험 중 가장 큰 성능 향상 폭을 기록하였다. 이는 개수 측정 손실이 모델의 수량적 관계 학습을 촉진하여, 방의 수량 정보가 포함된 텍스트 쿼리에 대해 의미적 매칭을 수행하고, 보다 정밀한 평면도 검색이 가능해졌음을 시사한다.

4. 결론

본 논문은 기존의 메타데이터 기반 평면도 검색 시스템의 한계를 극복하고자, 벡터 기반의 검색 방식을 활용한 텍스트 조건부 평면도 검색 기법을 제안하였다. 방의 개수, 연결 관계 및 위치 관계를 표현하는 공간 구성 텍스트 데이터를 자동 생성하여 벡터 기반의 평면도 검색 모델을 학습하였으며, 수량 정보에 대한 이해도를 높이고 검색 성능을 향상시키기 위해서 개수 측정 손실 함수를 추가하였다. 실험 결과, 제안된 기법은 모든 평가에서 성능 향상을 기록하였다. 이는 제안 기법이 공간 구성 텍스트에 대한 의미론적인 특성을 반영한 검색 성능을 향상시킴으로써, 발주자 및 설계자의 의도를 보다 명확하게 반영한 검색 결과를 제공하고, 이를 통해 설계 업무의 효율성과 생산성 증대에 기여할 것으로 기대된다. 향후 연구에서는 제안하는 기법의 성능 고도화를 위해, 평면도 검색 모델을 학습할 때 한 개의 텍스트 데이터와 매칭되는 다수의 평면도 이미지가 존재하는 느슨한 상관관계(Loose correlation)를 개선하는 방안을 적용할 예정이다.

사사문구

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. RS-2023-00252257, No. RS-2024-00397293).

참고문헌

[1] Shim, Jonghwa, et al., "FloorDiffusion: Diffusion model-based conditional floorplan image generation method using parameter-efficient fine-tuning and image inpainting," Journal of Building Engineering, vol. 95, pp. 110320, 2024.

[2] Paiss, Roni, "Teaching CLIP to Count to Ten," Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2023, pp. 3170- 3180.

[3] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PmLR, 2021.

[4] He, Kaiming. "Deep Residual Learning for Image Recognition." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 770-778.

[5] Dosovitskiy, Alexey, et al., "An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale," Proceedings of the International Conference on Learning Representations (ICLR), Virtual Event, 2021.

[6] Vaswani, Ashish, et al., "Attention is All You Need," Advances in Neural Information Processing Systems, vol. 30, Long Beach, CA, USA, 2017.

[7] rPLAN Dataset, https://paperswithcode.com/dataset/rplan