



RoomNet, LayoutNet, HorizonNet, HoHoNet

2021. 09. 07

AI융합학부 길다영

CONTENTS

1

RoomNet

<https://github.com/GitBoSun/roomnet>

2

LayoutNet

<https://github.com/sunset1995/pytorch-layoutnet>

3

HorizonNet

<https://github.com/sunset1995/HorizonNet>

4

HoHoNet

<https://github.com/sunset1995/HoHoNet>

5

정리 및 추후 계획

https://github.com/arittung/3D_Room_Reconstruction





1. RoomNet

1 RoomNet



RoomtNet

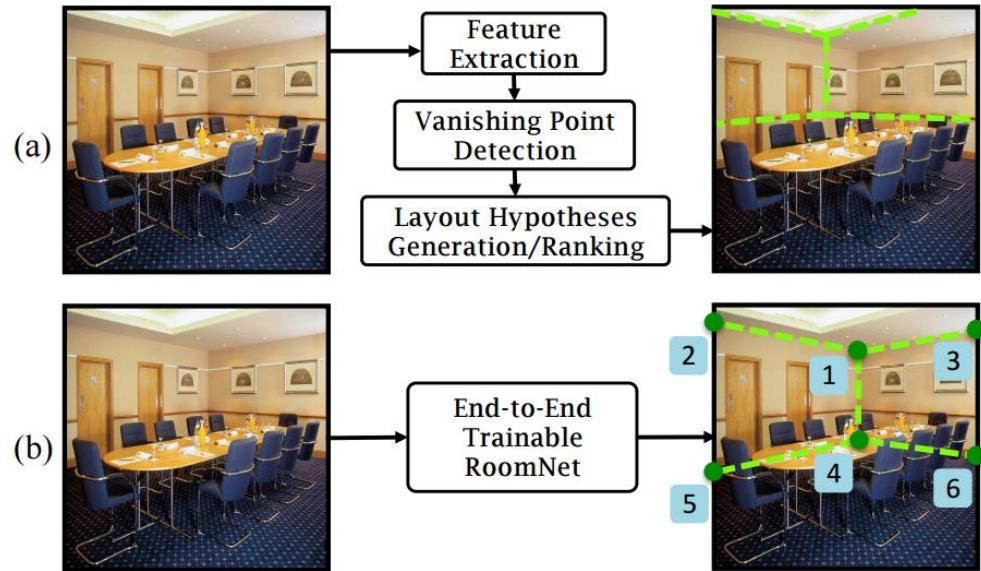
- 단안의 RGB 이미지에서 룸 레이아웃 추정 작업
- end-to-end trainable encoder-decoder 네트워크인 RoomNet을 사용하여 룸 레이아웃 키포인트의 위치를 예측.
- RoomNet 알고리즘은 객체(예: 테이블, 의자, 침대)에 의한 키포인트 폐색에 강함.



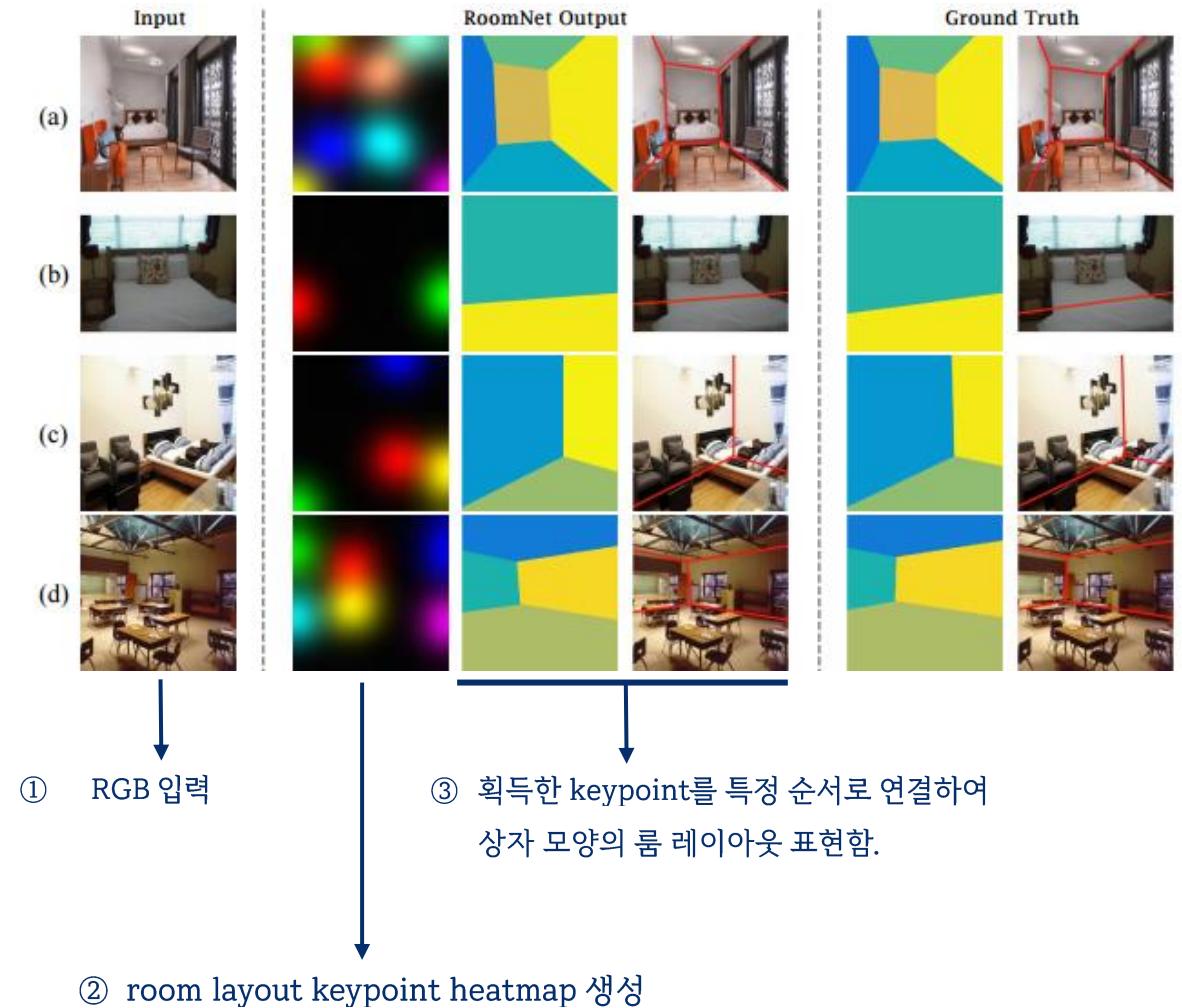
궁금한 점.

- ① RoomNet Architecture에 대한 이해가 부족하다

1 RoomNet - 소개



- ① 컨볼루션 encoder 구조에서 크기 320×320 의 입력 이미지를 처리.
- ② 방 레이아웃 키포인트의 집합을 추출,
- ③ 획득한 키포인트를 방 레이아웃을 그리기 위해 특정 순서로 연결.



1 RoomNet – ① keypoint 기반 룸 레이아웃 표현

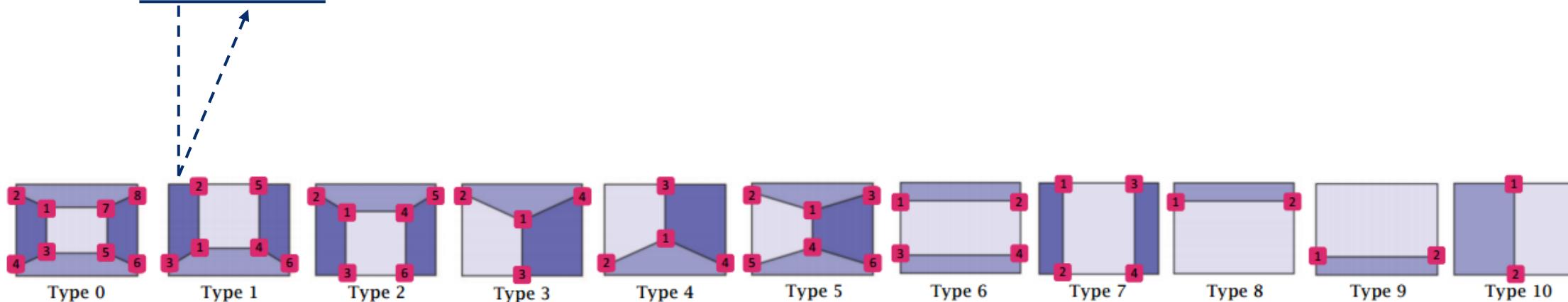
키포인트 기반 표현을 사용하는 또 다른 중요한 특성은 픽셀 기반 표현의 모호성을 제거한다는 것이다

CNN은 종종 다른 표면 정체성을 구별하는 데 어려움을 겪는다.

⇒ ex) CNN은 전면 벽 등급과 우측 벽 등급을 혼동하여 동일한 표면 내에서 불규칙하거나 혼합 픽셀 단위 레이블을 출력할 수 있다. 이 현상은 전체 room layout estimation 성능을 크게 저해한다.

따라서 키포인트 기반 룸 레이아웃 표현을 사용하여 모델을 훈련한다.

훈련된 모델이 관련 객실 유형을 사용하여 정확한 키포인트 위치를 예측하면 이러한 지점을 특정 순서로 연결하여 박스형 객실 레이아웃 표현을 생성할 수 있다.



룸 레이아웃 유형에 대한 정의

형식은 0부터 10까지 색인화 된다. 각 키포인트의 숫자는 실측 정보에 저장된 포인트의 특정 순서를 정의한다. 지정된 룸 유형의 경우 키포인트 순서는 연결성을 지정한다.

이러한 11개의 방 배치는 전형적인 카메라 포즈와 "Manhattan World 가정"[6]에 따른 일반적인 직육면체 표현 하에서 가능한 상황의 대부분을 다룬다.

1 RoomNet – ② Architecture of RoomNet : ① 키포인트 추정

키포인트 추정 프레임 워크 : 특정 객실 유형에 대한 기본 객실 추정 시스템 역할을 함.

RoomNet의 기본 아키텍처는 기본적으로 SegNet과 동일한 컨볼루션 encoder-decoder 네트워크를 채택한다.

실내 장면의 이미지를 가져와서 일련의 2D 룸 레이아웃 키포인트를 직접 출력하여 룸 레이아웃 구조를 복구한다.

각 키포인트 실측값은 출력 계층의 채널 중 하나로 실제 키포인트 위치를 중심으로 한 2D Gaussian heatmap으로 표현된다.

Encoder - decoder 아키텍처는 병목 계층을 통과하는 정보 흐름을 처리하여 룸 레이아웃의 2D 구조를 인코딩하는 키포인트 간의 관계를 암시적으로 모델링한다.

RoomNet의 decoder는 그림 3에서와 같이 전체 해상도 320×320 대신 공간 차원 10×10 에서 40×40 으로 bottleneck layer에서 feature map을 업샘플링한다.

이 “trimmed” decoder sub network를 사용하면 고해상도 컨볼루션의 높은 계산 비용 때문에 훈련과 시험 중 메모리/시간 비용을 크게 줄일 수 있다.

SegNet 프레임 워크

segmentation을 위해 설계된 SegNet 프레임워크는 encoder-decoder sub network로 구성된다.

SegNet의 encoder는 입력 이미지를 낮은 해상도의 feature map에 매핑한 다음 decoder의 역할은 픽셀 단위 분류를 위해 저해상도 인코딩된 feature map을 전체 입력 해상도로 업샘플링하는 것이다.

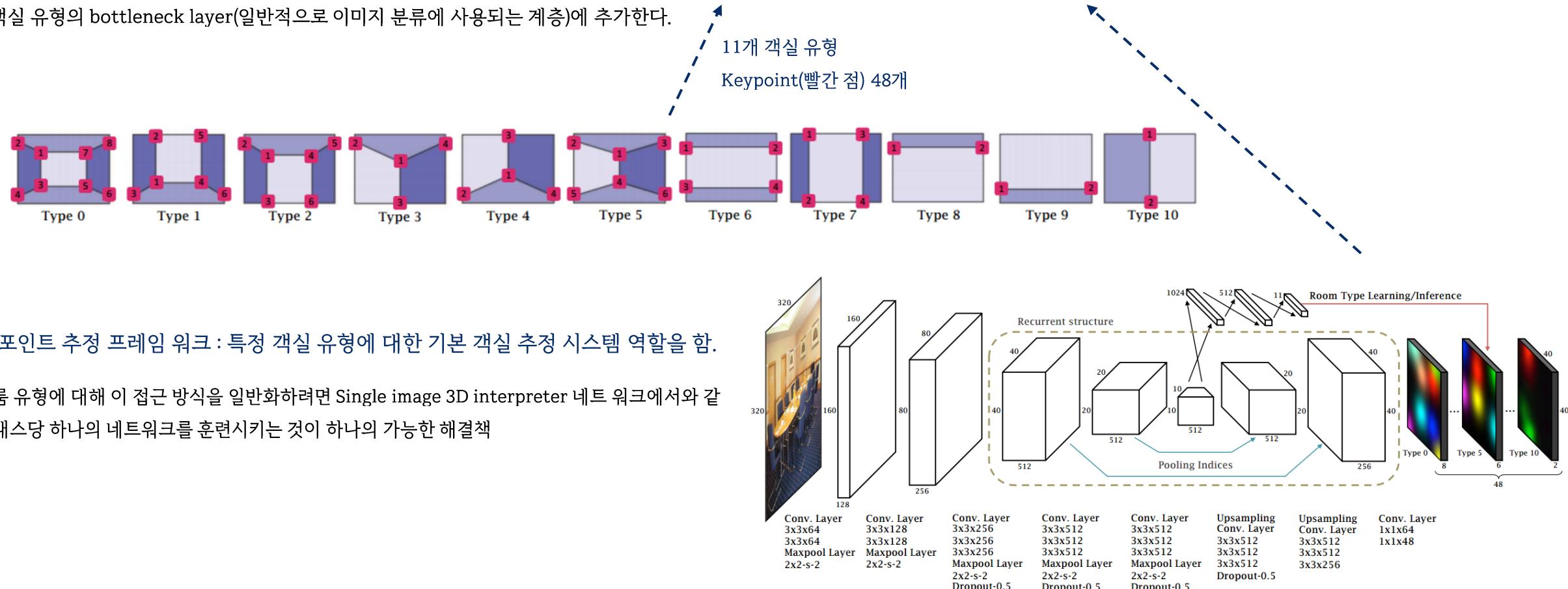
특히 Decoder는 해당 인코더의 최대 풀링 단계에서 계산된 풀링 지수를 사용하여 비선형 업샘플링을 수행한다. 이를 통해 upsampling을 학습할 필요가 없어진다.

업샘플링된 map은 희박하며 고밀도 feature map을 생성하기 위해 훈련 가능한 filter와 컨볼루션 된다.

RoomNet – ② Architecture of RoomNet : ② 여러 개의 룸 유형으로 확장

하나의 evaluation에서 전체 이미지로부터 직접 경계 상자와 클래스 확률을 예측하기 위해 single neural network를 사용하는 최근의 객체 감지 작업 YOLO[37]와 SSD[26]에 고무되어, RoomNet은 마찬가지로 한 번의 forward pass에서 입력 이미지와 관련하여 룸 레이아웃 키포인트와 관련 룸 유형을 예측한다.

출력 계층의 채널 수를 증가시켜 모든 11개 객실 유형의 총 키포인트 수(11개 객실 유형의 총 48개 키포인트 수)를 일치시키고, 또한 그림 3과 같은 룸 타입 예측에 대해 완전히 연결된 레이어를 객실 유형의 bottleneck layer(일반적으로 이미지 분류에 사용되는 계층)에 추가한다.



RoomNet – ② Architecture of RoomNet : ② 여러 개의 룸 유형으로 확장

훈련 단계에서, layout keypoint heatmap regression을 위한 비용 함수로 유클리드 손실을 사용하고, 방 유형 예측을 위해 교차 엔트로피 손실을 사용한다.

키포인트 heatmap regressor φ (decoder 서브 네트워크 출력) 및 룸 유형 분류기 ψ (완전히 연결된 측면 헤드 레이어 출력)를 고려하여 다음 손실 함수를 최적화할 수 있다.

손실함수 \Rightarrow 예측된 방 유형 지수를 사용하여 regressor를 업데이트할 해당 키포인트 heatmap set를 효과적으로 선택

손실 함수

손실 함수의 첫 번째 항(first term)은 예측 heatmap을 각 키포인트에 대해 별도로 합성된 ground-truth heatmap과 비교한다. 각 키포인트 heatmap에 대한 ground-truth는 최근 키포인트 regression 작업에서 일반적인 표준 편차가 5 pixel인 실제 키포인트 위치를 중심으로 한 2D Gaussian이다.

손실 함수의 두 번째 항(second term)은 side head fully-connected layers 가 정확한 객실 유형 등급 라벨에 대해 높은 신뢰도 값을 생성하도록 장려한다.

$$\sum_k \frac{\mathbb{1}_{k,t}^{\text{keypoint}} \|G_k(\mathbf{y}) - \varphi_k(\mathcal{I})\|^2}{G \text{는 } \mathbf{y} \text{에 중심인 Gaussian}} + \lambda \sum_c \frac{\mathbb{1}_{c,t}^{\text{room}} \log(\psi_c(\mathcal{I}))}{\text{실측 자료실 지수 } c \text{가 실측 자료실 유형 } t \text{와 같을 경우}} \quad (1)$$

키포인트 k 가 실측 자료실 유형 t 에 나타나는 경우

실측 자료실 지수 c 가 실측 자료실 유형 t 와 같을 경우

G 는 \mathbf{y} 에 중심인 Gaussian

weight term λ 은 교차 검증에 의해 5로 설정

1 RoomNet – ② Architecture of RoomNet : ③ 키포인트 세분화를 위한 RoomNet 확장

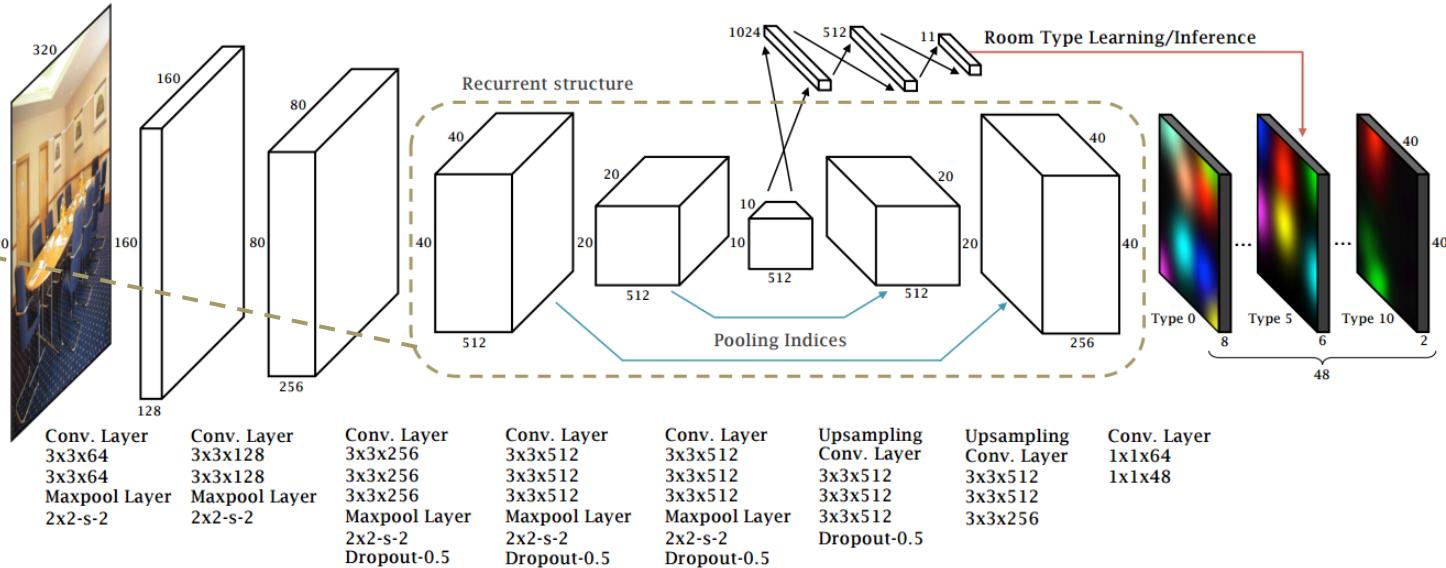
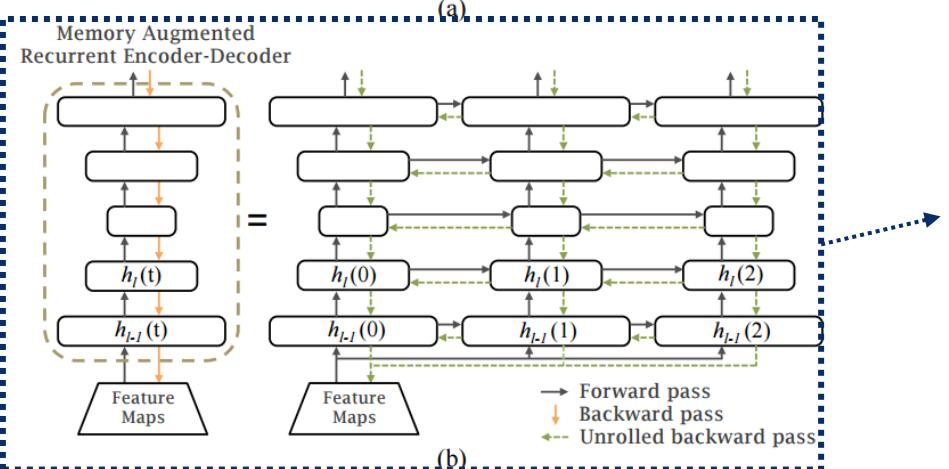
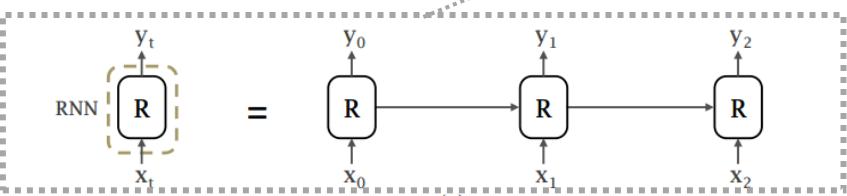
반복 신경 네트워크(RNN)와 그 변형 LSTM[17]은 순차 데이터를 처리할 때 매우 효과적인 모델

최근 CRF-RNN[52]을 사용하는 FCN, 반복 오류 피드백 네트워크[4], 반복 CNN [3], 스택형 encoder-decoder [31] 및 반복 encoder-decoder 네트워크와 같은 2D 정적 입력에 대해 보다 정교한 iterative/recurrent Architecture가 제안되었음.

central encoder-decoder 구성 요소를 반복하여 기본 RoomNet

아키텍처를 확장

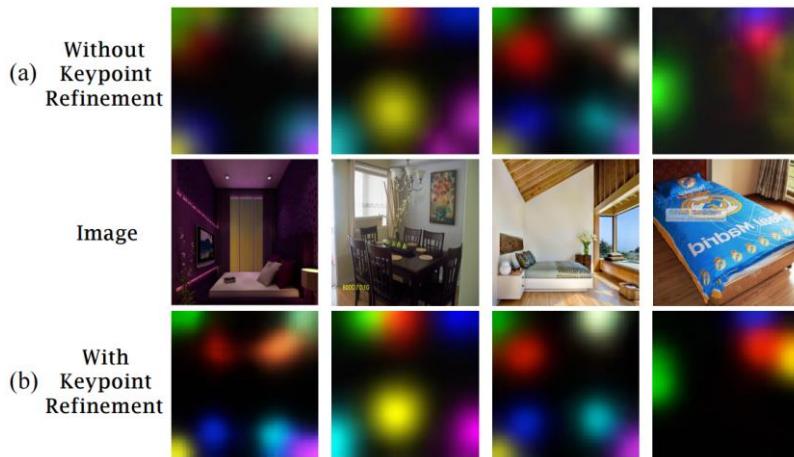
- 반복 구조에 의해 만들어진 인위적인 시간 단계
 - RNN 동작 모방
 - 현재 시간 단계에서 추론을 돋는 이전 활성화를 저장하기 위한 숨겨진 장치가 있다.



- Memory augmented recurrent encoder-decoder(MRED) 구조
 - recurrent encoder-decoder 구조 내에서 깊이와 시간을 통한 forward 및 backward 전파 중 정보 흐름의 전반적인 과정
 - “time” 동안 예측된 키포인트 heatmaps을 개선하기 위해 일반적인 반복 신경망의 동작을 모방하는 것
 - RNN의 동작을 모방하지만 정적 입력을 위해 설계된 증강 반복 encoder-decoder 아키텍처의 unrolled(3회) 버전의 그림
 - 현재 시간 단계에서 추론을 돋는 이전 활성화를 저장하기 위한 숨겨진 장치가 있다.
 - MRED Architecture의 장점
 - ① (recurrent Convolution encoder-decoder 구조에서 탐구되지 않은) hidden/memory unit를 통해 핵심점 간의 상황적 및 구조적 지식을 반복적으로 활용한다.
 - ② recurrent encoder-decoder의 Convolution layer의 가중치 공유는 고정된 수의 매개 변수를 가진 훨씬 더 깊은 네트워크를 초래한다.

1

RoomNet – ② Architecture of RoomNet : ③ 키포인트 세분화를 위한 RoomNet 확장



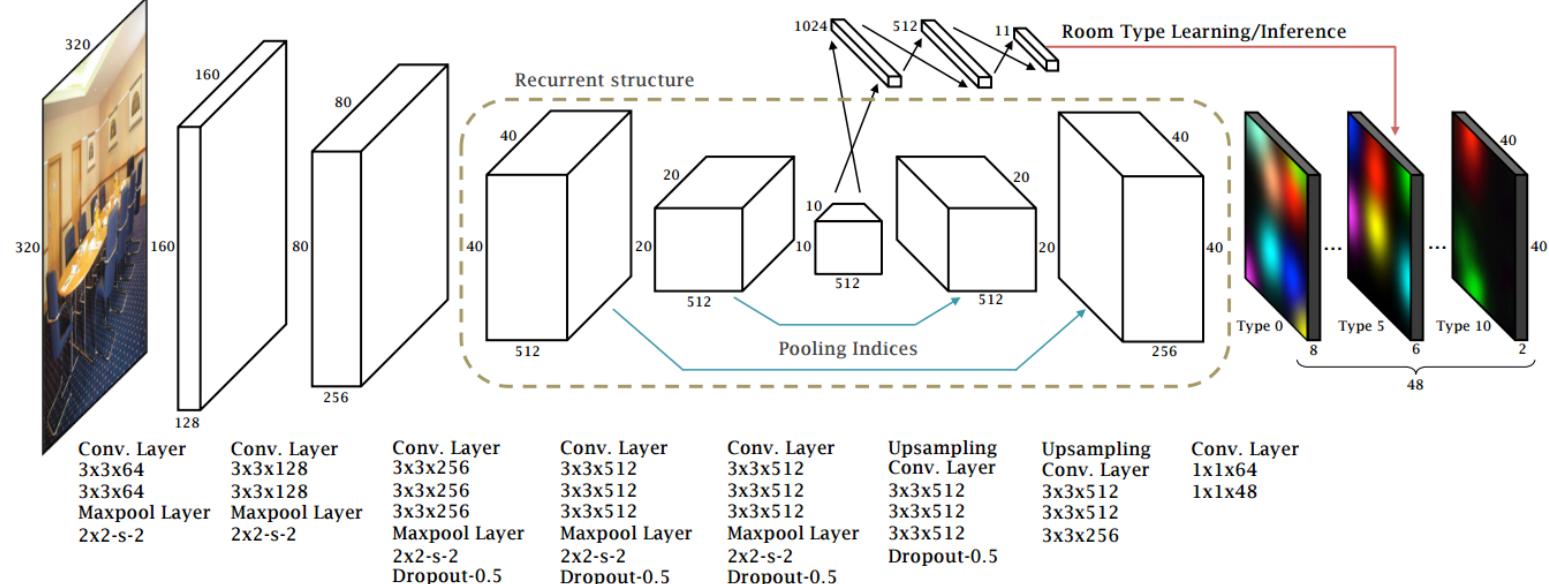
정제되지 않은 단일 이미지(a)와 정제된 단일 이미지(b)에서 룸 레이아웃 키포인트 추정.
여러 채널의 키포인트 heatmap은 시각화 목적으로 컬러 코딩되어 단일 2D 이미지로 표시되는데,
키포인트 개선 단계는 보다 집중적이고 깨끗한 heatmap을 생성하고 일부 잘못된 긍정을 제거한다.

추가 이해 필요

decoder는 encdoer에서 전송된 pooling 지수를 사용하여 입력을 업샘플링하여 희소 feature map을 생성한 다음 feature responses를 밀집시키기 위해 훈련 가능한 필터 뱅크가 있는 여러 컨볼루션 레이어를 생성한다. 최종 decoder output 키포인트 heatmap은 유클리드 손실로 regressor에 공급된다.

완전히 연결된 3개의 레이어가 있는 side head는 병목 계층(bottleneck layer)에 부착되어 룸 유형 클래스 라벨을 교육/예측하는 데 사용되며, 그런 다음 키포인트 heatmap의 관리 세트를 선택하는 데 사용된다.

반복 encoder-decoder(중앙 파선(dashed line) 블록)가 있는 RoomNet의 전체 모델은 그림 4 (b)와 5에 표시된 것처럼 키포인트 개선을 추가로 수행한다.





2. LayoutNet

2 LayoutNet



LayoutNet

- 파노라마와 투시 이미지, 직육면체 레이아웃 및 보다 일반적인 레이아웃(예: "L"자형 룸)에 걸쳐 일반화하는 단일 이미지에서 룸 레이아웃을 예측하는 알고리즘
 - 단일 RGB 등각 파노라마에서 직접 레이아웃을 추정하여 reconstruction을 단순화함.
 - 최종 출력물은 카메라와의 각 벽의 거리, 높이 및 레이아웃 회전으로 매개 변수화된 희박하고 콤팩트한 평면 Manhattan layout 이다.
-
- 소실점을 기준으로 파노라마 이미지를 정렬한 후, 시스템은 심층 네트워크를 사용하여 파노라마 이미지의 경계와 모서리를 직접 예측한다. 이러한 점에서 우리는 심층 네트워크를 사용하여 투시 이미지의 레이아웃 모서리와 가시적인 모서리를 나타내는 레이블을 직접 예측하는 RoomNet과 유사함.
-
- ① 단일 파노라마에서 3D 직육면체 레이아웃을 예측한다.
 - ② 단일 파노라마에서 3D non-cuboid Manhattan layout을 추정한다.
 - ③ 단일 원근 이미지에서 레이아웃을 추정한다.

2 LayoutNet



내용 정리

- ① “Manhattan World” 가정 하에 3D 모델을 만들기 위해 수평 카메라가 촬영한 투시 이미지에 바닥/벽 경계를 맞춘다
- ② 방향 지도(Orientation Maps)를 작성하고, 검출된 선분을 기반으로 레이아웃 가설을 생성하고, 그 중에서 가장 적합한 레이아웃을 선택한다.
- ③ 소실점 3개를 풀고, 소실점과 일치하는 레이아웃을 샘플링하고, 가장자리와 기하학적 맥락 일관성을 기반으로 최적의 레이아웃을 선택하여 입체 레이아웃을 복구한다.
→ 소실점을 기준으로 파노라마 이미지를 정렬한 후, 시스템은 심층 네트워크를 사용하여 파노라마 이미지의 경계와 모서리를 직접 예측한다.



RoomNet에 대한 개선점

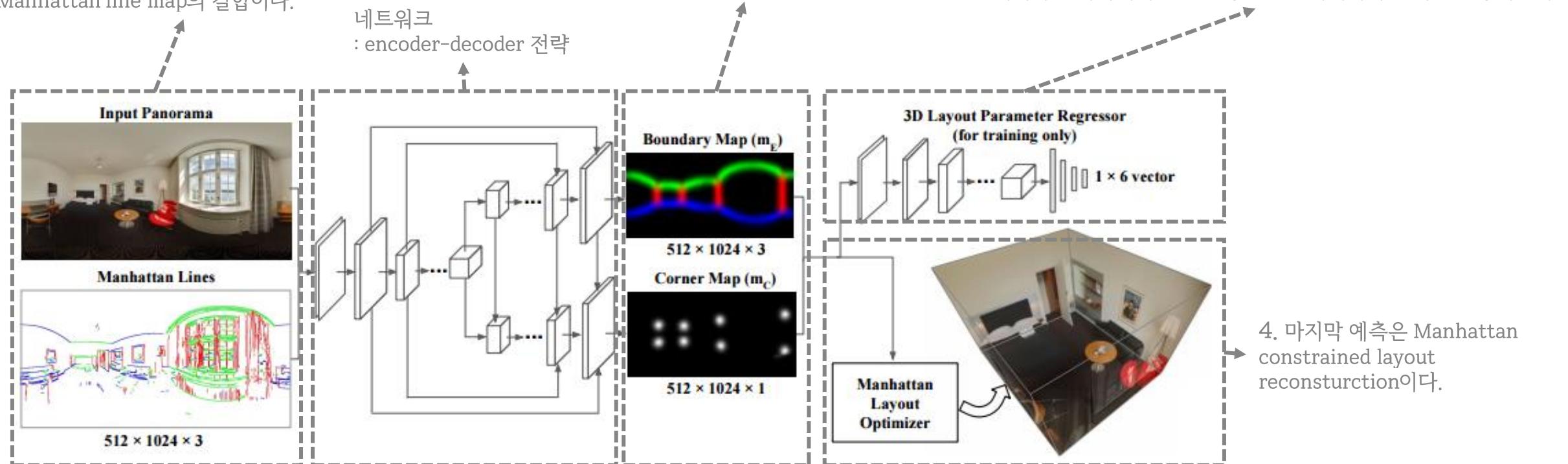
- 소실점을 기반으로 이미지를 정렬하고, 다중 레이아웃 요소(모서리, 경계, 크기 및 변환)를 예측하고, 결과 예측에 제한된 맨해튼 레이아웃을 맞추기 때문에 개선을 보여준다.
- 파노라마 영상에 적용함.
- 정렬단계와 경계, 모서리 및 3D cuboid parameter에 대한 다중 작업 예측에서 다름.
- RoomNet은 RNN을 사용하여 2D 모서리 위치 예측을 세분화하지만 이러한 예측은 3D 입체 배치와 일치하지 않을 수 있다.

2 LayoutNet - 소개

1. 네트워크 입력은 단일 RGB 파노라마와 Manhattan line map의 결합이다.

2. 네트워크는 레이아웃 경계와 모서리 위치를 공동으로 예측한다.

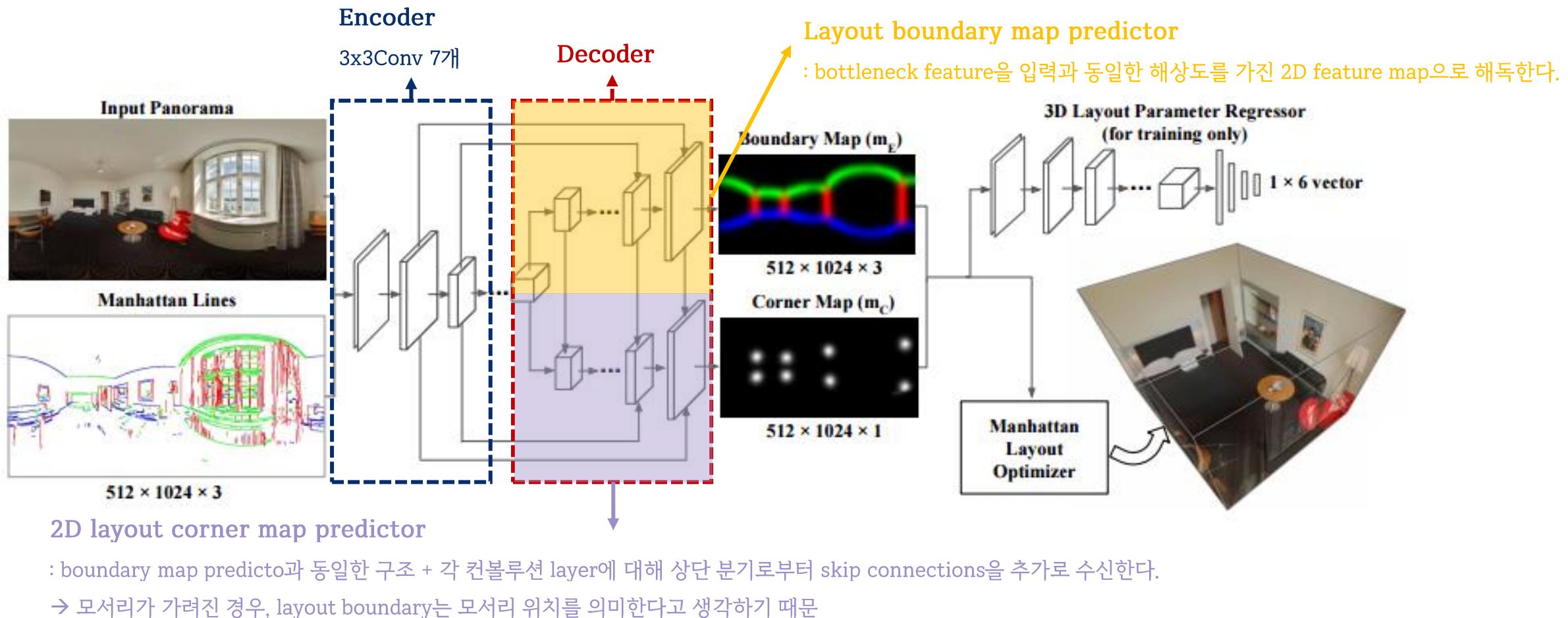
LayoutNet은 encoder-decoder 전략을 따른다.



- ① 시스템이 소실점을 분석하고 이미지를 바닥과 수평으로 정렬한다. 이러한 정렬을 통해 벽면 경계가 수직선임을 보장하고 실험에 따라 오류를 줄일 수 있다.
- ② Encoder-decoder 구조와 skip connections을 가진 CNN을 사용하여 이미지에서 직접 코너(레이아웃 접합)와 경계 확률 맵(corner (layout junctions) and boundary probability maps)을 예측한다. 모서리와 경계는 각각 방 배치를 완벽하게 표현해 준다.
- ③ 3D Layout parameters는 예측된 모서리와 경계에 적합하도록 최적화된다

2 LayoutNet – ① 파노라마 이미지 정렬 + ② corner (layout junctions) and boundary probability maps 예측

- 구형 투영에서 바닥 평면 방향을 추정하여 이미지를 정렬하고 장면을 회전한 다음 2D 등각 투영으로 다시 투영한다
we first align the image by estimating the floor plane direction under spherical projection, rotate the scene, and reproject it to the 2D equirectangular projection.
- 네트워크는 encoder-decoder 전략을 따른다.
- 훈련 방법 : 2D 레이아웃 예측 네트워크의 경우, 먼저 네트워크의 매개 변수를 초기화하기 위해 레이아웃 경계 예측 작업을 훈련한다.

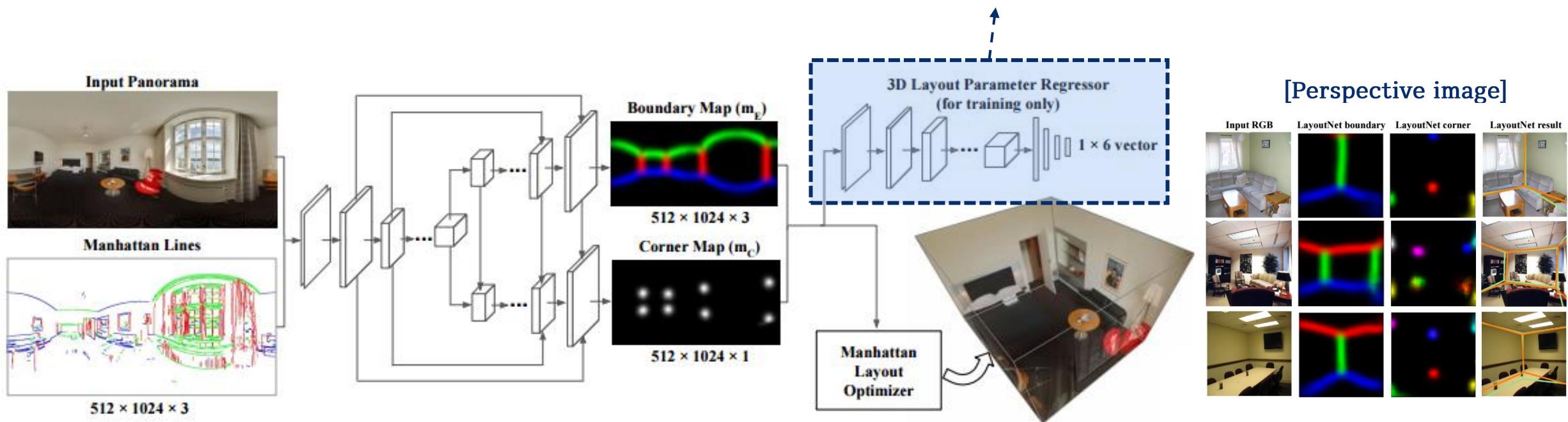


2 LayoutNet – ③ 3D layout Parameter Regressor

- 3D 레이아웃 파라미터에 대한 회귀 분석기를 만드는 것보다는 더 나은 모서리와 경계를 만들기 위해 회귀 분석기를 훈련한다.
- 회귀 분석기는 2D 영상에 다시 투영할 수 있는 3D 레이아웃의 파라미터를 출력하여 end to end 예측 접근 방식을 제시한다.
- 3D 회귀기는 두 예측된 2D map의 연결을 입력으로 얻고 3D 레이아웃의 매개변수를 예측한다.
- 3D 회귀기가 정확하지 않다는 것이 관찰된다. Loss object에 3D 회귀 분석기를 포함하면 네트워크의 예측이 약간 개선되는 경향이 있다.
- 훈련 방법 : 3D 레이아웃 회귀기의 경우, 먼저 지상 실측 레이아웃 경계와 모서리를 입력으로 하여 네트워크를 훈련한 다음 이를 2D 레이아웃 decoder와 연결하고 전체 네트워크 end-to-end를 훈련한다.

Non-cuboid 레이아웃에 대해서는 3D parameter regressor로 훈련하지 않는다.

Perspective image 예측 할 때, 이미지의 모서리와 경계를 직접 예측하는 대신 정렬 및 최적화 단계를 건너뛰고, 또한 regressor branch를 사용하지 않는다.



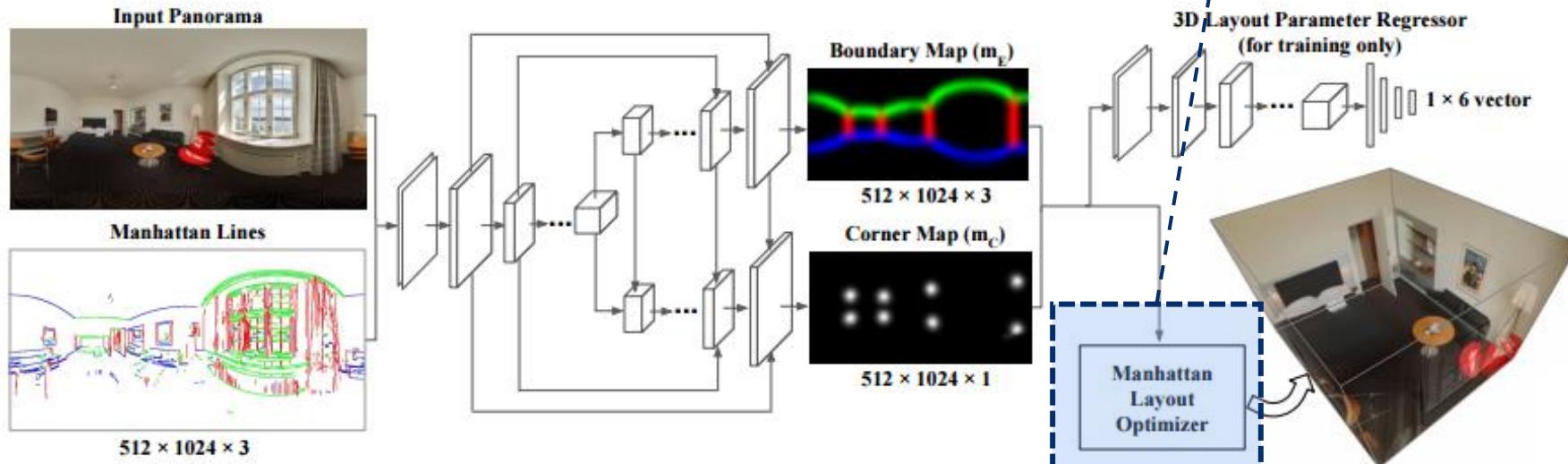
2 LayoutNet – ④ 3D layout Optimazation

- 초기 2D 코너 예측은 네트워크가 출력하는 corner probability maps에서 얻어진다.

- 먼저, 각 열에 대한 반응 합계를 얻기 위해 여러 행에 걸쳐 반응을 합산한다.
- 로컬 최대값은 로컬 최대값 사이의 거리가 최소 20픽셀인 column responses에서 발견된다.
- 선택한 열을 따라 가장 큰 두 개의 봉우리가 표시된다.

이러한 2D 코너는 Manhattan 제약 조건을 충족하지 못할 수 있으므로, 우리는 추정치를 구체화하기 위해 최적화를 수행한다.

- 예측된 모서리 위치가 주어지면, 하단 모서리가 동일한 지면 위에 있고 상단 모서리가 하단 모서리 바로 위에 있다고 가정하여 카메라 위치와 3D 레이아웃을 최대 규모 및 변환까지 직접 복구할 수 있다.
- 또한 배치 형태를 Manhattan으로 제한하여 교차 벽이 수직이 되도록 할 수 있다.





3. HorizonNet

3 HorizonNet



HorizonNet

- 목표 : 360도 파노라마 이미지로부터 Manhattan room layout 추정
- 1D 표현과 효율적인 후처리 절차로 인해 모델의 계산 비용이 매우 낮으며 모델을 쉽게 확장하여 직육면체 또는 L자 이외의 레이아웃으로 복잡한 장면을 처리 가능



한계

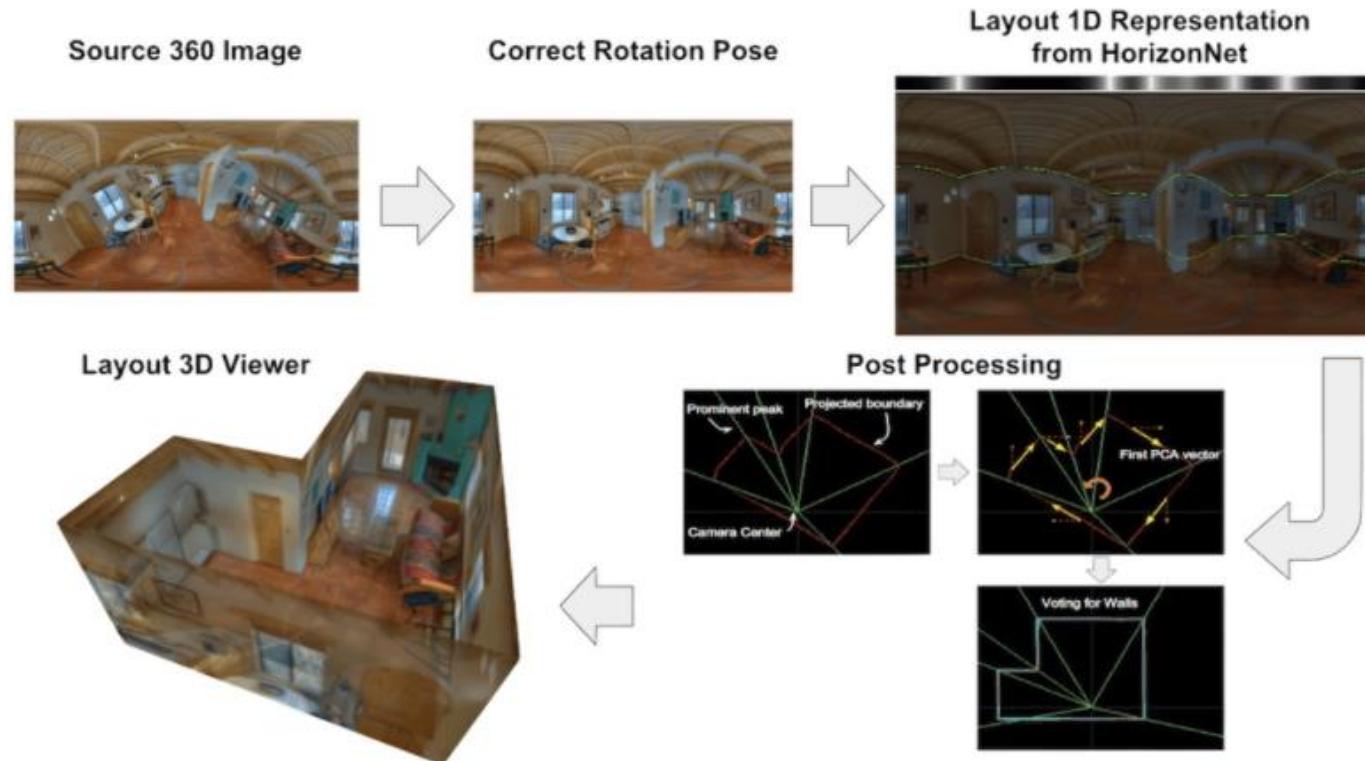
1. 파노라마 이미지를 위한 고품질 room layout 주석(모서리 좌표)을 획득하는 것은 많은 노력이 필요하다. 특히 잘 꾸며진 방의 경우 벽 경계 위치에 대한 모호성으로 인해 다른 사람들이 수행한 주석이 일관되지 않을 수 있다.
2. 현재 사용 가능한 데이터 세트에는 복잡한 룸 레이아웃의 이미지가 더 포함되어 있지 않다.



궁금한 점.

- ① Pano Stretch Data Argumentation은 어디에서 쓰이는가?
- ② 추가 논문해석이 필요해보임.

3 HorizonNet - 소개



1. 수직 보정 전처리를 통해 1D Layout 표현
2. 특징 추출
3. Post-Processing

3

HorizonNet - ① 1D Layout Representation + ② 특징 추출(feature extractor)

- 입력으로 받은 파노라마 사진을 수직 보정하여 소실점과 edge를 찾는다.
- ResNet50과 LSTM을 이용하여 훈련된 모델로 사진의 특징을 추출하여 천장-벽 경계, 바닥-벽 경계, 벽-벽 경계가 표시된 1D Layout을 도출한다.

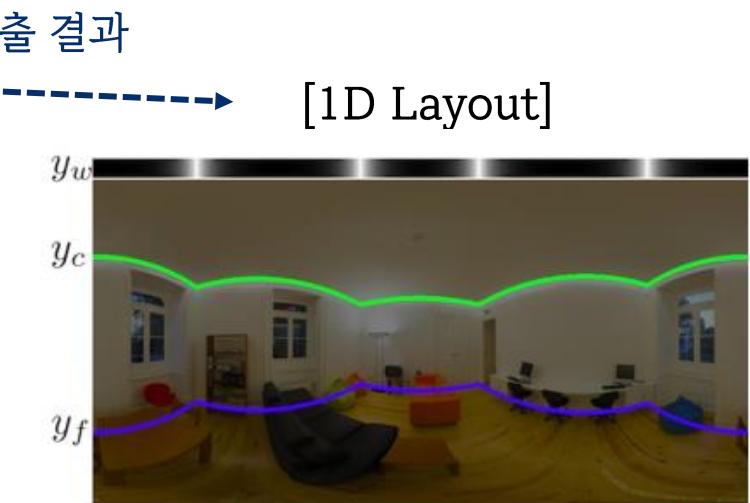
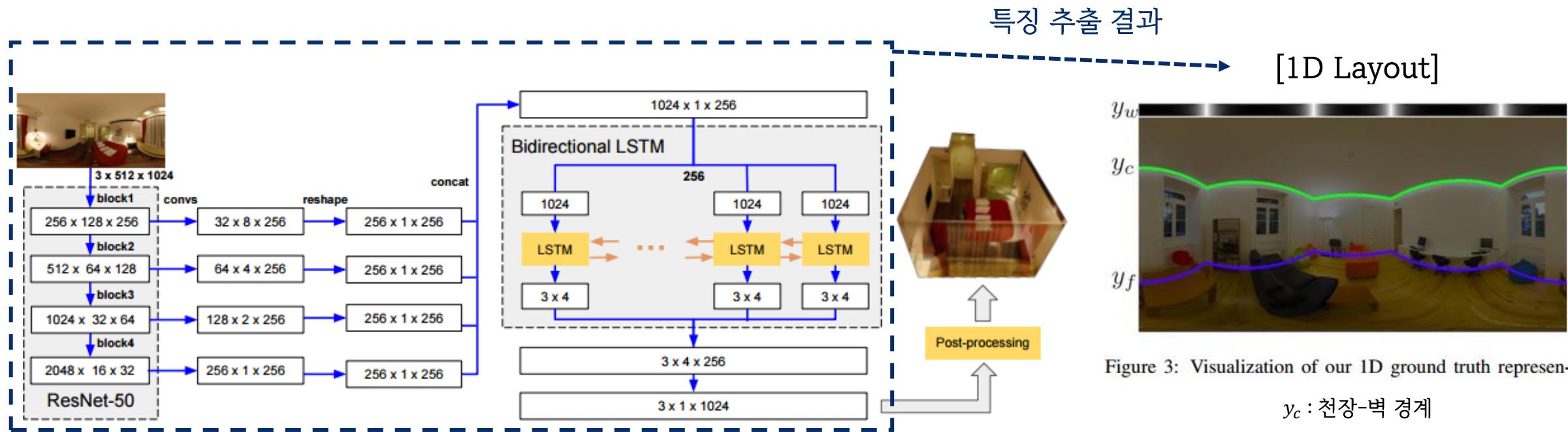


Figure 3: Visualization of our 1D ground truth representation

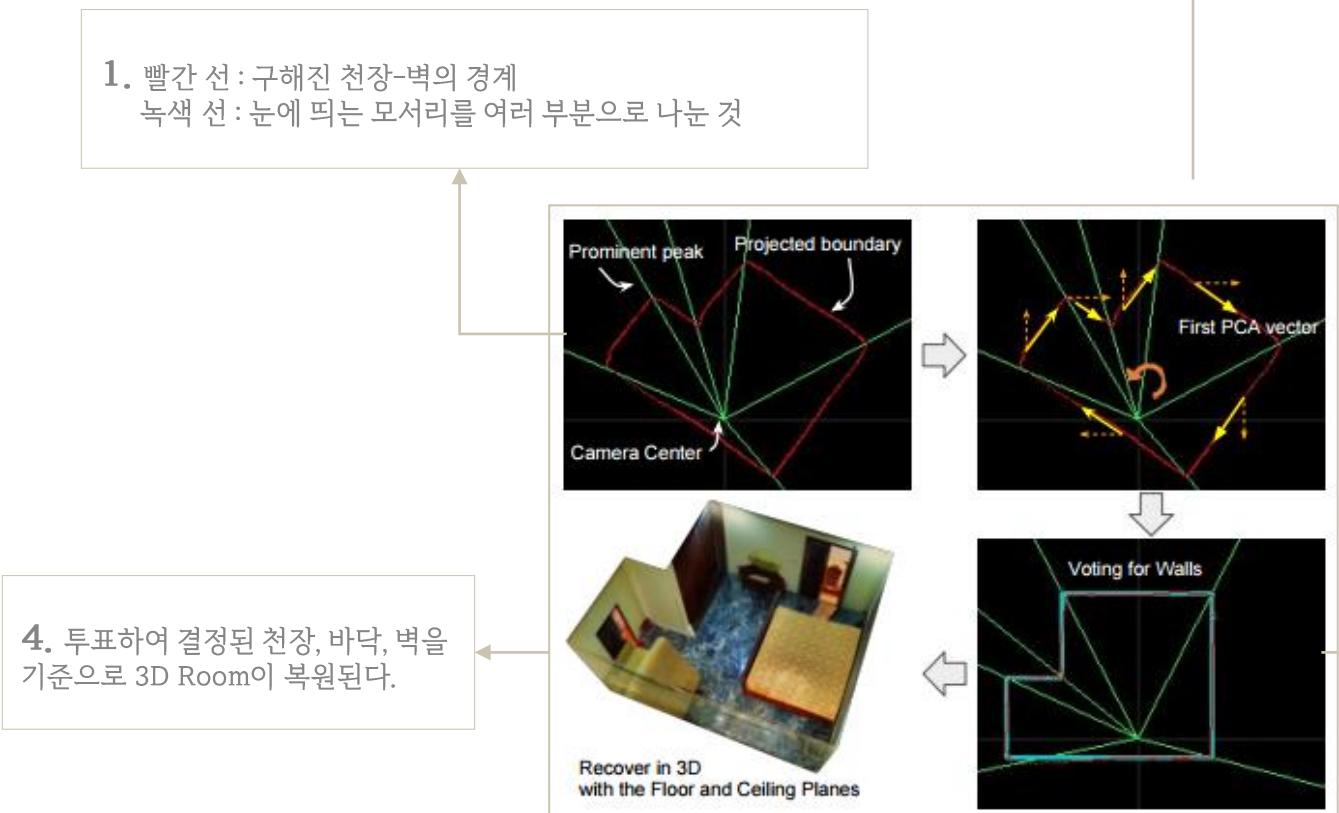
y_c : 천장-벽 경계

y_f : 바닥-벽 경계

y_w : 벽-벽 경계

3 HorizonNet – ③ Post-processing

- Manhattan World 가정으로 바닥 천장, 벽면을 복구.
- 공식을 통해 천장-바닥 거리를 계산한 후 벽면을 복구함.



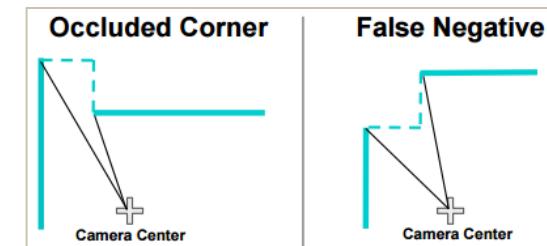
Manhattan World 가정 : 영상에 나타나는 평면들은 3차원상에서 서로 직교하는 평면들로만 이루어져 있다.

Cuboid : 직육면체의 형태

PCA : 분산이 가장 큰 방향이 가장 주된 방향이라고 가정, 다차원의 데이터 분포를 가장 잘 표현하는 성분들을 찾아 주기 위함

- 3. 모든 벽면에 대해서 XZ축과 수직인 벽에 투표를 하여 투표에서 가장 많은 표를 얻은 벽이 선택됨.**

투표를 할 때는 Cuboid 형태의 방과 Non-Cuboid 형태의 방으로 나뉘어 진행됨. → 구조에 따른 벽면의 개수에 차이가 존재하고, Non-Cuboid의 구조에서 두가지의 특이점이 존재하기 때문



『Non-Cuboid 구조에서 두가지 특이점』

- I. 가려져서 보이지 않는 코너가 존재하는 경우
- II. 코너가 존재하지 않고 훈련된 모델이 코너가 없다고 판단한 경우.

이 두 경우에는 기존과 같이 벽에 투표하지 않고 두 개의 두드러진 모서리와 두 개의 벽의 위치에 따라 코너를 추가하는 특별한 방식을 사용한다.

3 HorizonNet – ③ Pano Stretch Data Augmentation

- Pano Stretch Data Augmentation : x축 또는 z축을 따라 이미지와 지상 실측 자료 레이아웃을 확장하여 즉석에서 효과적으로 훈련 데이터를 증강
- Pano Stretch Data Augmentation 절차는 파노라마에서 직접 작동하는 다른 작업(예: 의미론적 분할의 지상 실측 지도, 객체 감지를 위한 경계 상자)에도 사용될 수 있다.
증강 절차는 이러한 작업의 정확도를 높일 수 있다.

$k_x = 1.0, k_z = 1.0$ (original)



$k_x = 2.0, k_z = 1.0$



$k_x = 1.0, k_z = 2.0$



$k_x = 2.0, k_z = 2.0$





4. HoHoNet

4 HoHoNet



HoHoNet

높이 치수가 평평한 잠재 수평 특징 표현(LHFeat)을 통해 레이아웃 구조, dense depth 및 semantic segmentation을 모델링하기 위한 새로운 딥 러닝 프레임워크.



HoHoNet은 두 가지 측면에서 발전했다

- ① 심층 아키텍처는 향상된 정확도로 더 빠르게 실행되도록 재설계됨.
- ② LHFeat의 픽셀당 조밀한 예측을 가능하게 하여 열당 출력 형태 제약을 완화하는 새로운 수평선 대 밀도 모듈을 제안함.



궁금한 점.

- ① 열당 예측과 픽셀당 예측의 차이?

⇒ 열 당 예측으로 layout 을, 픽셀당 예측으로 depth와 semantic을 나오게 하는 것 같다.

- ① Hohonet은 Depth estimation와 semantic sementation이 나오고 그걸 horzonnet처럼 3d room reconstruction을 하는 듯..? → 이걸 코드에서 알 수 있을까?
⇒ 알 수 있다!! HoHoNet/lib/model/modality/에 각 코드가 나와 있음.
- ② 실행 시 room layout이 만들어지지만, 코드 상 depth와 semantic 을 사용한 부분, 그리고 그게 결과로 나타난 부분이 있는지 알아봐야 할 거 같다

4 HoHoNet



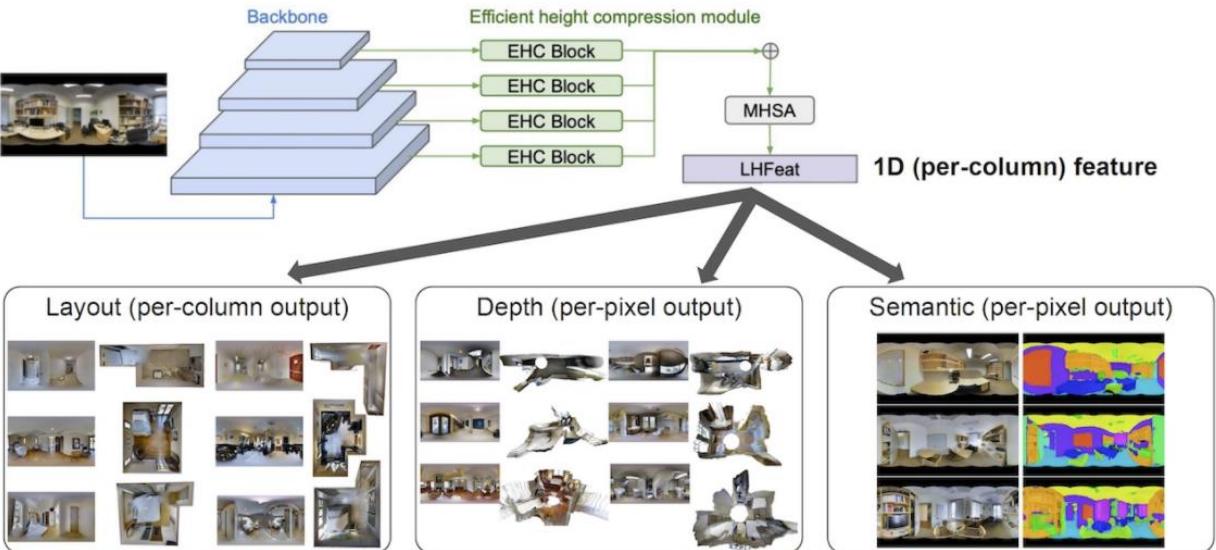
내용 정리

Hohonet의 결과 : layout, depth, semantic

Hohonet은 레이아웃 재구성을 위한 새로운 방법을 설계하는 것이 아님.

Room layout : Hohonet은 resnet34, horzonnet은 resnet50 사용.

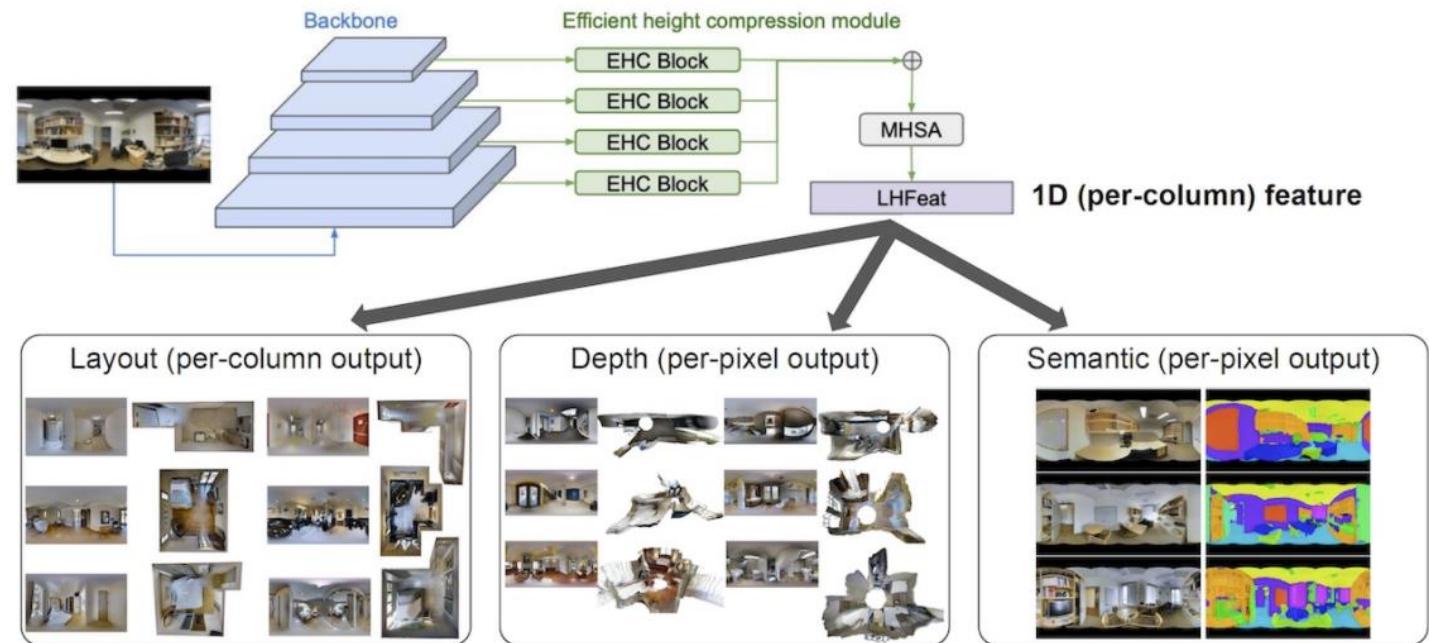
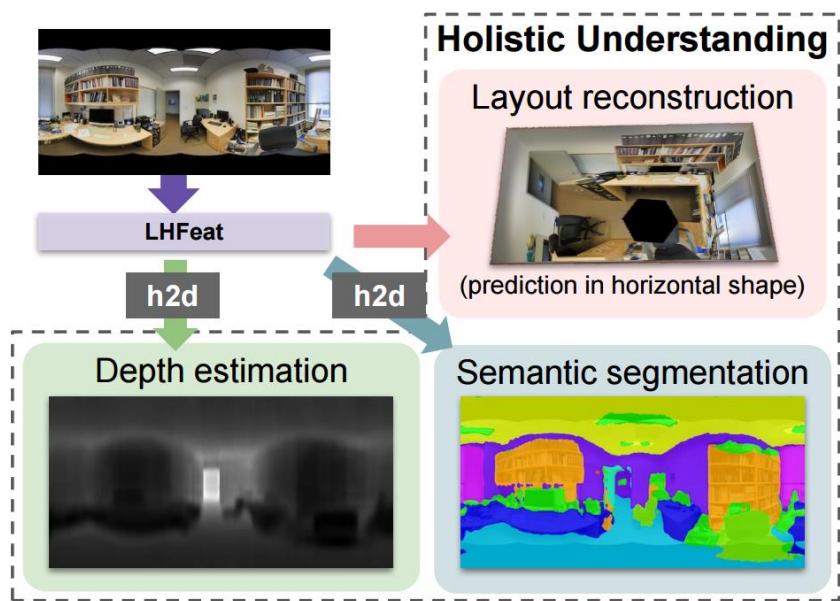
Semantic segmentation : resnet101 사용.



Depth estimation : resnet50 사용

- HoHoNet은 이전의 최첨단 기술인 BiFuse[24]를 큰 폭으로 능가한다는 것을 입증한다. HoHoNet은 장면의 전체적인 구조를 잘 포착한다.
- BiFuse는 ERP와 큐브맵을 모두 모델 입력으로 사용하므로 두 개의 backbone 네트워크가 필요.
- Hohonet 단점 : HoHoNet의 깊이 경계는 BiFuse의 깊이 경계에 비해 더 흐릿하다. 열의 일부 고주파 신호는 HoHoNet에 의해 폐기된다.

4 HoHoNet - 소개



- ① 입력 ERP 이미지는 형상 피라미드 추출을 위해 먼저 CNN backbone을 통과한 다음,
- ② 제안된 효율적인 높이 압축 모듈은 형상 피라미드를 높이 치수가 평평한 잠재 수평 형상 표현(LHFeat)으로 인코딩한다.
- ③ 마지막으로, LHFeat에서 HoHoNet 프레임워크는 최첨단 품질의 열당 및 픽셀당 양식(레이아웃의 모서리 또는 경계)을 모두 제공할 수 있다.



→ (a)처럼 y축이 중력방향으로 정렬되었을 때 이미지 column 구조 정보를 더 잘 압축하여 보관할 수 있음.

(a) Aligned 360.

(b) Roll rotation.

(c) Pitch rotation.

4 HoHoNet - 소개

360 이미지에 대한 깊이 추정.

Depth estimation



360 이미지에 대한 Semantic segmentation

Semantic segmentation



의미론적 분할은 장면 모델링의 기본 작업이다.

DistConv는 ERP 이미지의 조밀한 깊이 및 semantic 예측을 위한 왜곡 인식 변형 가능한 컨볼루션 레이어를 제안 한다.

360 의미 있는 분할을 위한 최근의 대부분의 방법은 정이십면체 mesh와 관련된 표현으로 작동하는 훈련 가능한 층을 설계한다.

그러나 위의 모든 방법은 파노라마 신호에 대해 비교적 낮은 해상도로 실행된다.

탄젠트 이미지는 고해상도 파노라마를 처리하고 미리 훈련된 가중치를 투시 이미지에 배치할 수 있는 세분화된 20면체에 접하는 다중 평면 이미지에 전방위 신호를 투사한다.

탄젠트 이미지와 마찬가지로 HoHoNet도 고해상도 이미지에서 작동할 수 있으며, 이는 더 나은 의미 있는 분할 정확도를 달성하는데 필수적인 요소로 나타났다.

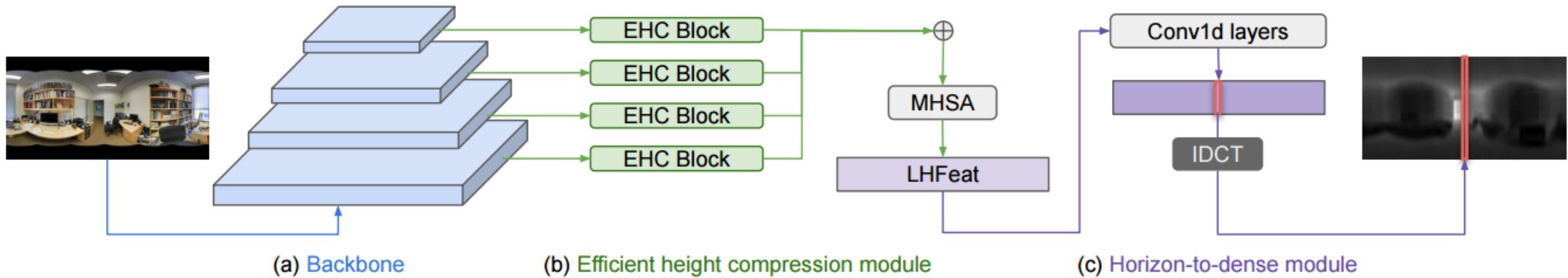
최근의 방법과 대조적으로 HoHoNet은 ERP 이미지에서 직접 실행되며 고도로 최적화된 딥 러닝 라이브러리는 모든 작업을 쉽게 구현할 수 있다.

전방위 이미지의 깊이를 모델링하기 위해, OmniDepth는 ERP 왜곡을 고려하여 encoder-decoder architecture를 설계한다.

PanoPopups는 평면 인식 손실로 360 깊이를 학습하는 것이 합성 환경에 도움이 된다는 것을 보여준다.

계단식 훈련 단계를 가진 여러 백본(backbone)을 사용하는 대부분의 최신 방법과 대조적으로, HoHoNet은 하나의 백본으로만 구성되며 한 단계에서만 훈련된다.

또한, HoHoNet은 소형 LFeat을 통해 밀도가 높은 깊이를 모델링하는 반면 이전의 기술은 기존의 밀도가 높은 특징에서 깊이를 추정한다.



① 고해상도 파노라마는 먼저 backbone(예: ResNet)에 의해 처리된다.

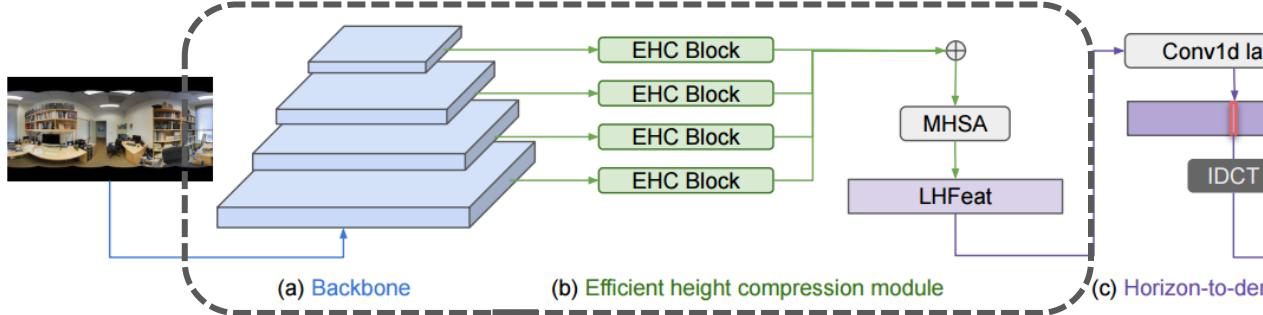
② 형상 피라미드는 제안된 EHC(Efficient Height Compression) 모듈과 정교화를 위한 다중 헤드 자기 주의(MHSA) 모듈에 의해 압착 및 융합된다.

그 결과 LHFeat은 compact되며(예: 입력 이미지가 $R 3 \times 512 \times 1024$ 인 경우 $R 256 \times 1024$), 전체 네트워크가 기존의 인코더-디코더 네트워크보다 훨씬 빠르게 조밀한 기능을 실행할 수 있다는 점에 유의한다.

③ 마지막으로, 최종 예측을 산출하기 위해 1D 컨볼루션 레이어를 사용한다.

DCT 주파수 영역에서 예측이 우수한 결과를 가져오기에 각 열의 예측에 IDCT를 적용한다.

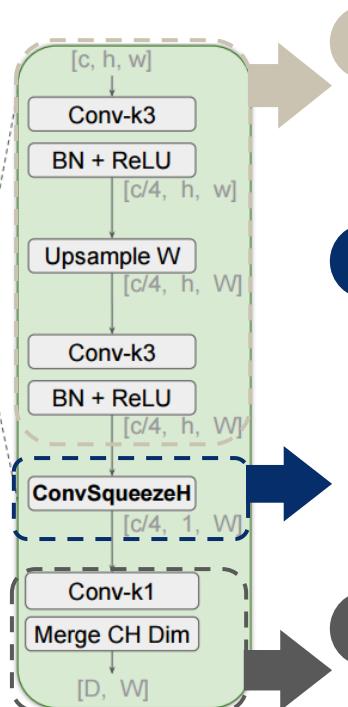
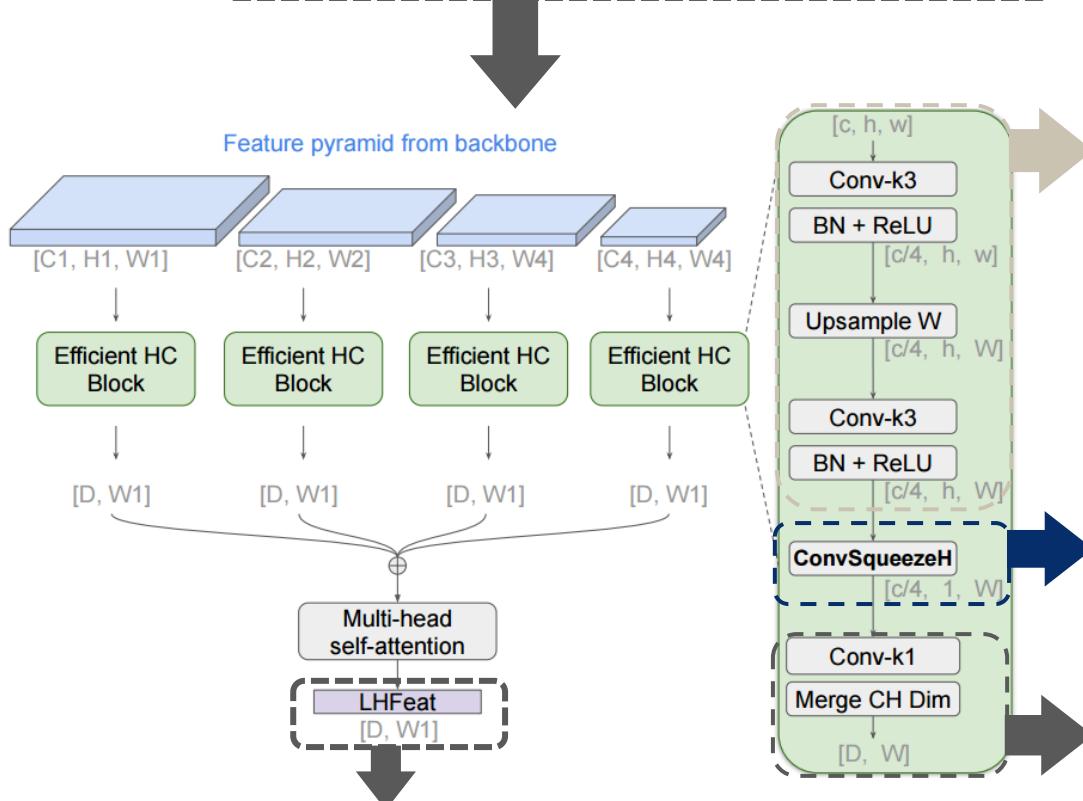
HoHoNet - 조밀한 깊이 추정을 위한 HoHoNet 프레임워크 개요 ▶ ② LHFeat에 대한 EHC(효율 높이 압축) 모듈



먼저 backbone의 피라미드에서 각 2D feature의 높이를 짜내기 위해 EHC 블록을 사용한다.

그 결과 1D 형상이 합산으로 간단히 융합된다.

2D 및 1D 형상의 크기는 각각 $[C, H, W]$ 및 $[C, W]$ 로 표기한다.



EHC 블록 내에서 입력 2D feature은 채널 감소를 위해 먼저 Conv2D 블록에 의해 처리된 다음, 필요한 경우 공간 너비가 W_1 로 upsampling되고 마지막으로 다른 Conv2D 블록이 upsampling된 feature을 개선한다.

2

feature 높이를 1로 효율적으로 줄이기 위해 커널 크기를 $(h, 1)$ 로 설정하여 패딩 없이 전체 feature 높이를 커버하는 깊이 있는 컨볼루션 레이어인 ConvSqueezeH 레이어를 설계한다.

ConvSqueezeH 레이어는 커널 크기가 패딩 없이 이전에 알려진 입력 형상 높이로 설정된 깊이 별 컨볼루션 레이어로 출력 형상 높이 1을 생성한다.

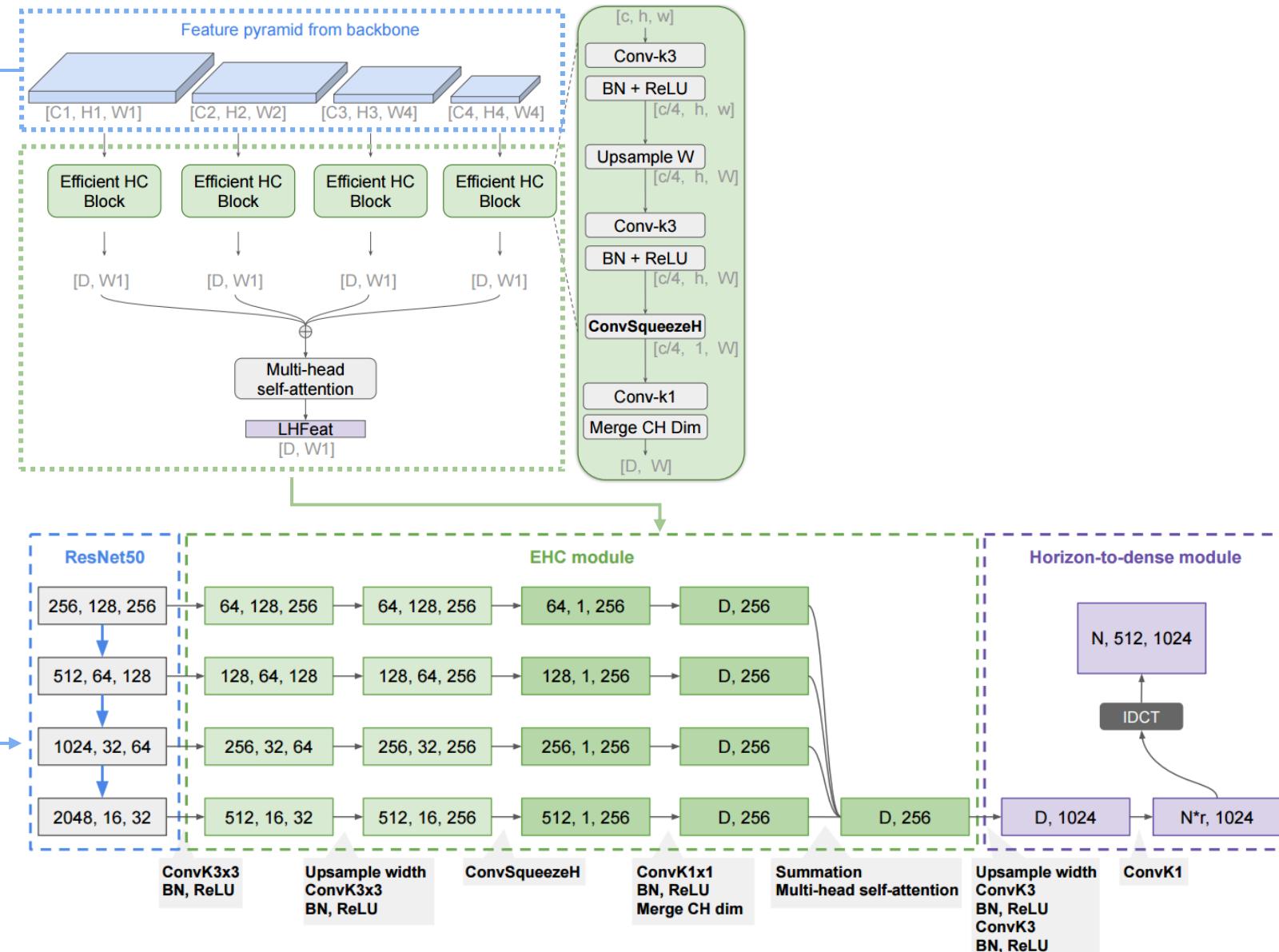
각 EHC 블록의 파라미터 h 는 H_{inp} 가 주어지면 자동으로 사전 계산된다.

3

마지막으로, Conv2D 레이어는 채널 수를 LHFeat의 잠재 크기 D 로 변환하고, 높이 차수는 ConvSqueezeH 레이어에 의해 이미 1로 감소되었기 때문에 간단히 폐기된다.

4

HoHoNet - 조밀한 깊이 추정을 위한 HoHoNet 프레임워크 개요 ▶ ② LHFeat에 대한 EHC(효율 높이 압축) 모듈



입력 파노라마의 높이와 너비는 각각 512와 1024로 가정한다.

D와 E는 하이퍼 파라미터다.

4

HoHoNet - 조밀한 깊이 추정을 위한 HoHoNet 프레임워크 개요 ► ② LHFeat에 대한 EHC(효율 높이 압축) 모듈

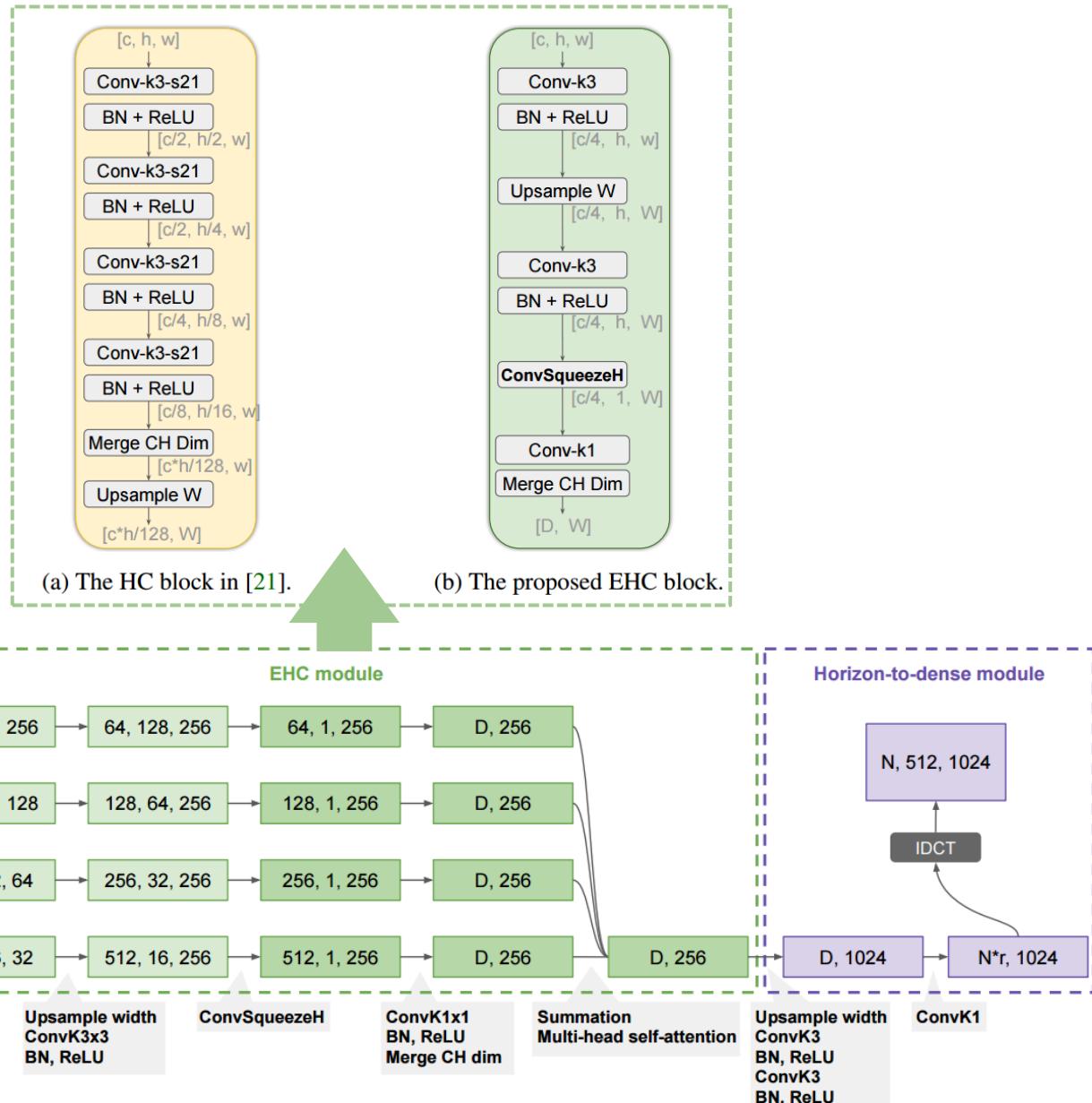
EHC 블록과 HC 블록 비교

높이 압축 블록은 백본에서 2D 형상을 압착하여 1D 수평 형상을 생성하는 것을 목표로 한다.

HC 블록은 일련의 컨볼루션 레이어를 사용하여 채널 수와 높이를 점차적으로 감소시킨다.

반면, EHC 블록은 먼저 채널 감소를 위한 컨볼루션 레이어를 사용한 다음 이중선형 업샘플링 및 ConvSqueezeH 레이어를 사용하여 수평 형상의 형상을 생성한다.

절제 실험에서 HC 블록을 제안된 ECH 블록으로 대체하면 속도와 정확도가 향상된다.



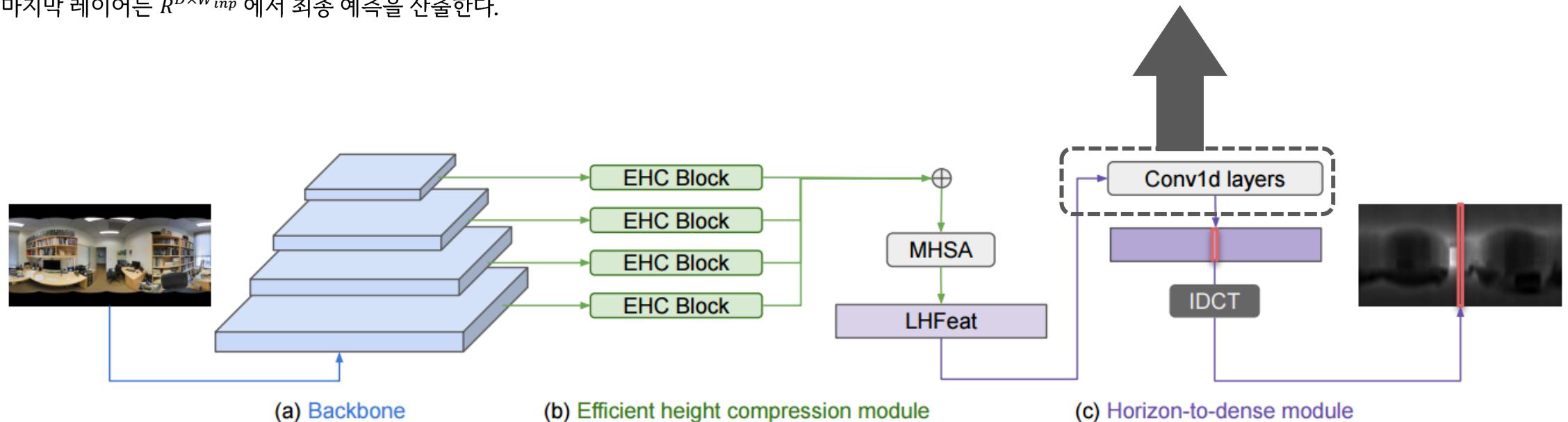
4

HoHoNet - 조밀한 깊이 추정을 위한 HoHoNet 프레임워크 개요 ► ③ horizon-to-dense 모듈, 열 당 1D 양식 예측



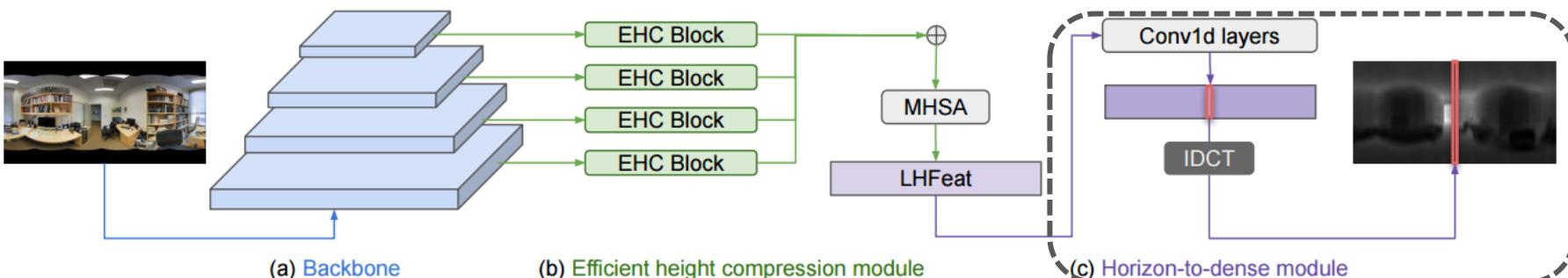
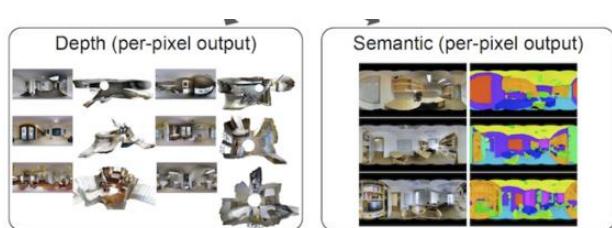
1D 양식을 예측하기 위해 먼저 $R^{D \times W_1}$ 에서 $R^{D \times W_{inp}}$ 로 수평 형상을 upsampling하고 BN, ReLU 사이에 커널 크기 3, 3, 1의 Conv1D 레이어를 각각 적용한다.

마지막 레이어는 $R^{D \times W_{inp}}$ 에서 최종 예측을 산출한다.



4

HoHoNet - 조밀한 깊이 추정을 위한 HoHoNet 프레임워크 개요 ▶ ③ horizon-to-dense 모듈, 픽셀당 2D 양식 예측

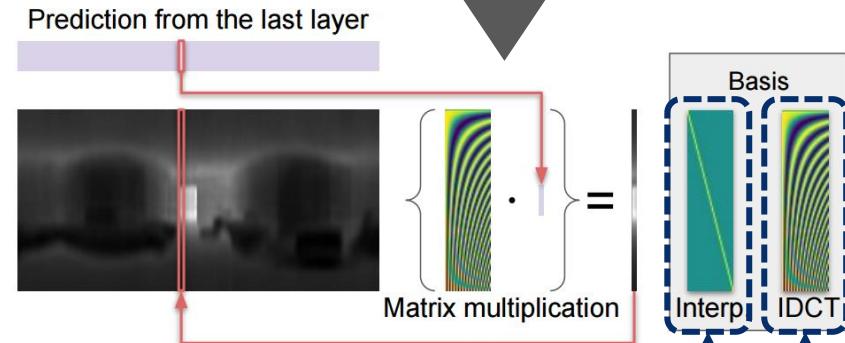


출력 공간을 열당 형식으로 shaping하는 전략은 픽셀당 양식이 포함된 작업에는 적용되지 않는다.

여기서는 compact LHFeat $R^{D \times W_1}$ 에서 조밀한 예측 $R^{N \times H_{inp} \times W_{inp}}$ 를 도출하기 위한 HoHoNet의 수평 대 밀도 모듈을 제시한다.

이 기능은 다양한 애플리케이션에 보다 일반적인 시나리오의 문을 열어준다.

2D 양식 예측을 위한 훈련 가능한 계층은 출력 계층의 채널 수가 $E = N \cdot r$ 로 증강되고 여기서 N 은 작업에 대한 대상 채널의 수이고 r 은 이미지 열에 의해 공유되는 구성 요소의 수라는 점을 제외하면 3.3항에서 소개한 1D 예측을 위한 계층과 거의 동일하다.



생성된 예측은 $R^{E \times W_{inp}}$ 에서 $R^{N \times r \times W_{inp}}$ 로 재구성된다. 예측된 r 값에 할당한 물리적 의미에 따라 각 열에 대해 R^r 를 $R^{H_{inp}}$ 로 복구하기 위한 두 가지 다른 연산을 제시한다.

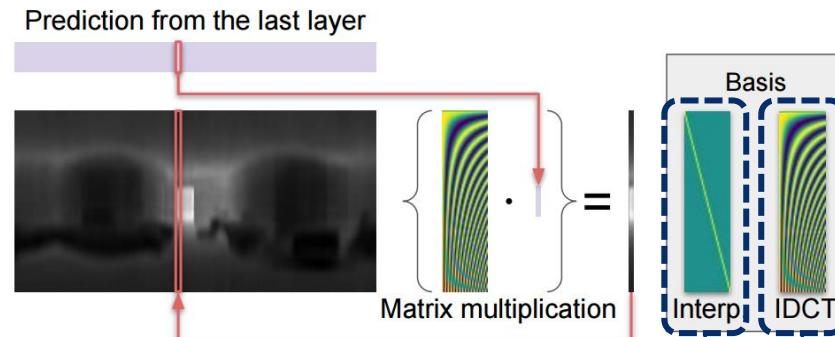
① 보간법 (interpolation)

② 역 이산 코사인 변환 (IDCT, Inverse Discrete Cosine Transform)

제안된 수평 대 밀도(h2d) 모듈은 compact LHFeat에서 밀도 예측을 생성할 수 있다.

Linear interpolation을 IDCT로 대체함으로써 조밀한 예측 결과를 개선할 수 있다.

수평 대 밀도 모듈(h2d)을 통해 효율적으로 인코딩된 LHFeat은 이제 조밀한 양식을 모델링할 수 있다.



각 열의 예측은 기초 M의 성분들의 선형 조합에 대한 가중치로 작용한다.

① HoHoNet은 M이 선형 보간을 구현하는 경우 공간 영역에서 예측하고

① 보간법 (interpolation)

가장 간단한 방법은 잠재 치수 r 을 출력 높이로 보고 선형 보간법을 적용하여 $r < H_{inp}$ 일 경우, H_{inp} 의 r 크기를 조정하는 것이다.

② 역 이산 코사인 변환 (IDCT, Inverse Discrete Cosine Transform)

에너지 압축 특성에 대한 이미지 압축에서 DCT의 적용에 영감을 받아, r 예측 값을 높은 주파수가 잘리는 DCT 주파수 영역에 있는 것처럼 본다. ② M이 IDCT를 구현하는 경우 주파수 영역에서 학습한다.

이 경우 IDCT를 적용하여 low-pass 신호를 원래 신호로 복구할 수 있다.

IDCT가 선형 보간법을 지속적으로 능가한다.

LHFeat은 공간-행 정보를 혼합하므로, 평평한 행이 없는 LHFeat에서 행에 의존하는 밀도 양식을 분리하기 위해 마지막 층을 훈련시키는 것은 문제가 될 것이다.

반대로, 주파수 영역에서 예측하는 법을 배우는 것은 각 열의 원래 행 정보를 전체적으로 특징짓는 의미 있는 공간 주파수를 가진 잘 정의된 기본 함수로부터 이익을 얻을 수 있으므로 행 의존성 문제를 완화시킬 수 있다.

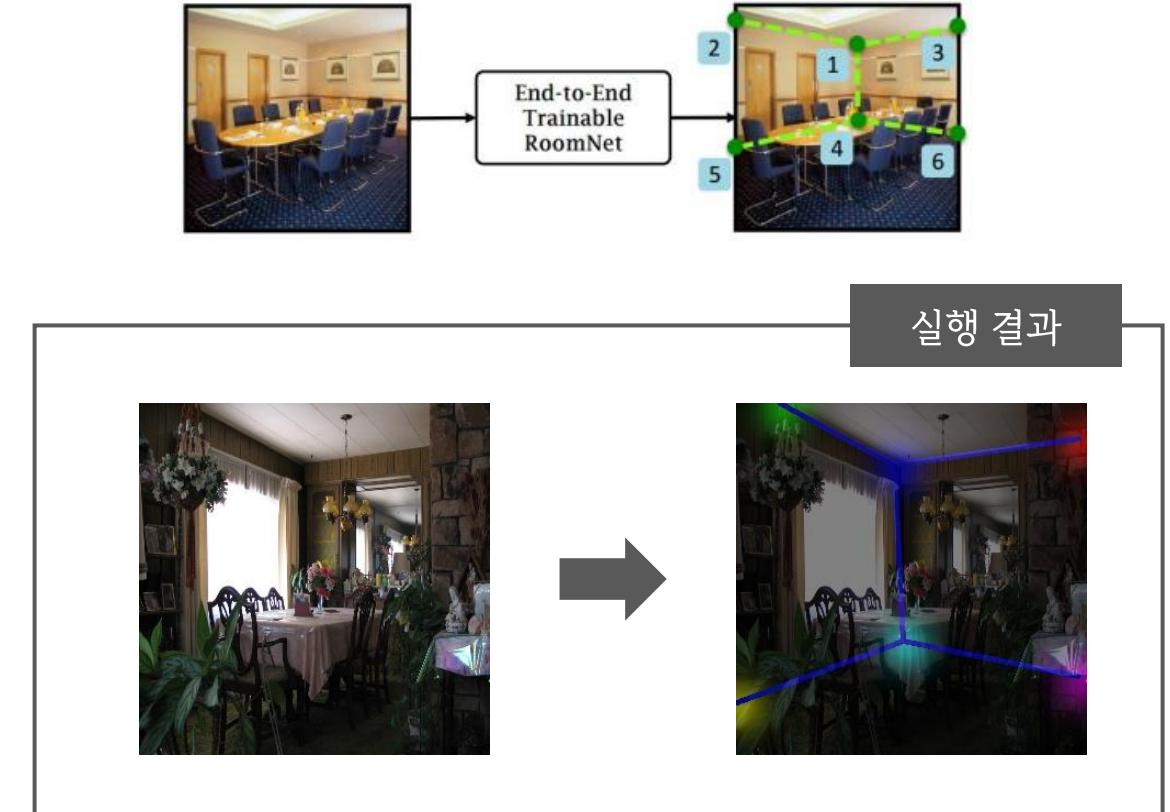
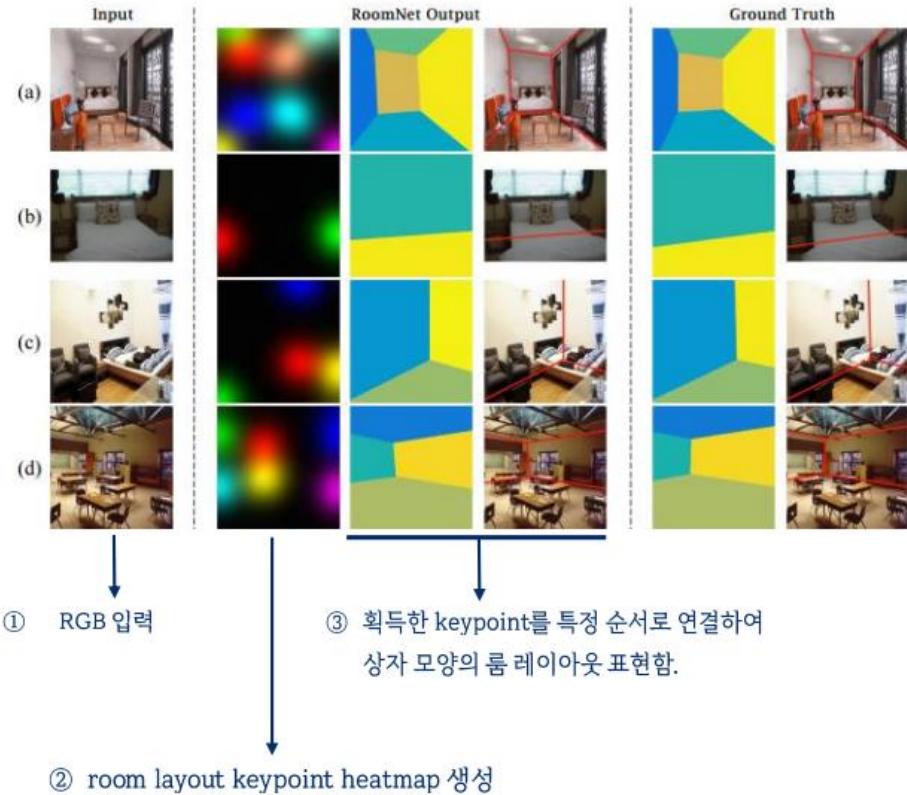


5. 정리 및 추후 계획

5 정리 및 추후 계획 - 정리

1 RoomNet

- 방의 일부 이미지에서 레이아웃을 추정하는 작업
- **end-to-end trainable encoder-decoder Network(RNN)**를 통해 키포인트 집합을 추출하고, 획득한 **keypoint**를 특정 순서로 연결하여 방 레이아웃을 그림.



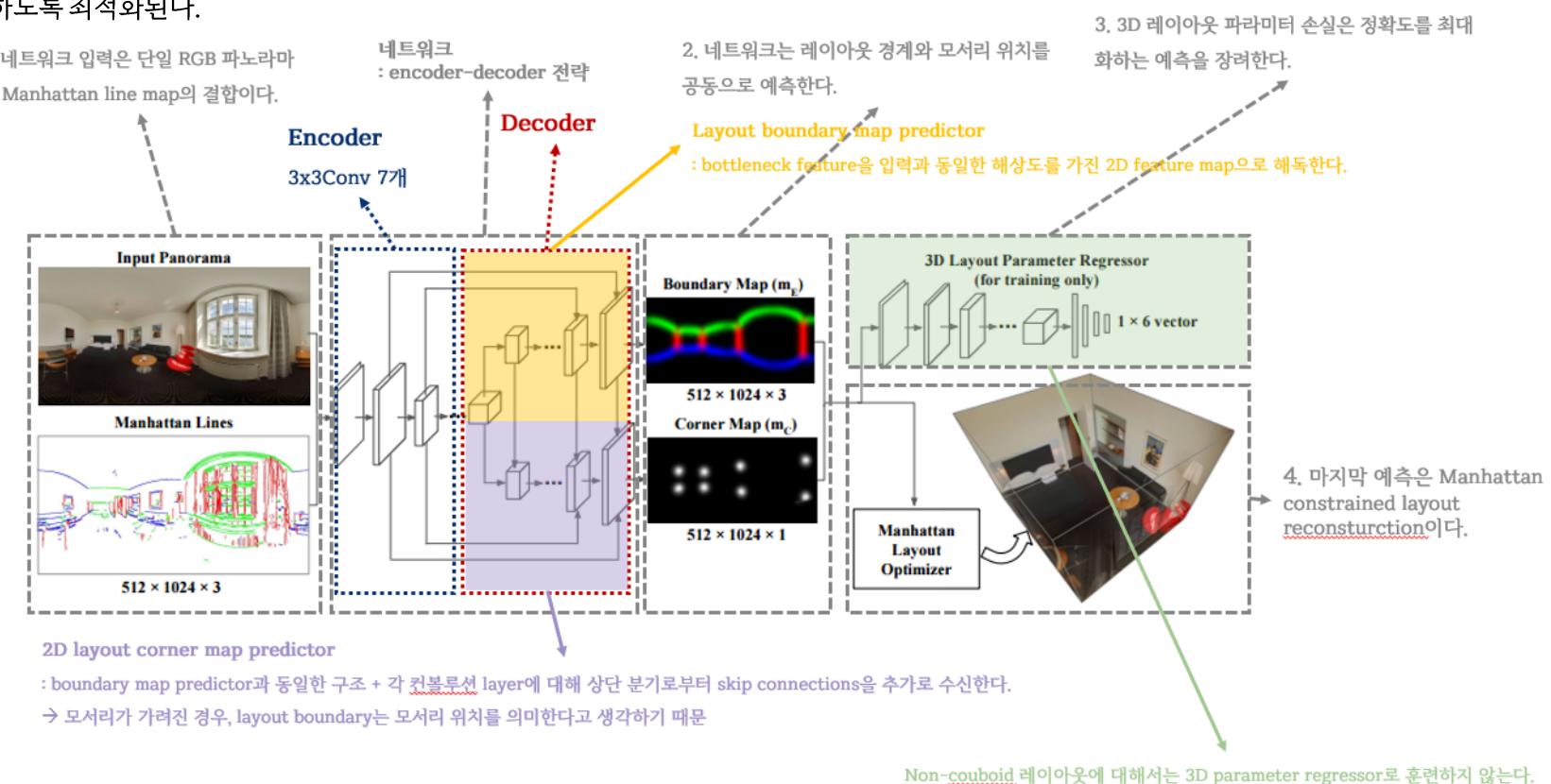
5 정리 및 추후 계획 - 정리

2 LayoutNet

- 소실점을 기준으로 파노라마 이미지를 정렬한 후, 시스템은 심층 네트워크를 사용하여 파노라마 이미지의 경계와 모서리를 직접 예측한다.

방법

- ① 시스템이 소실점을 분석하고 이미지를 바닥과 수평으로 정렬한다. 이러한 정렬을 통해 벽면 경계가 수직선임을 보장하고 실험에 따라 오류를 줄일 수 있다.
- ② **Encoder-decoder 구조와 skip connections**을 가진 CNN을 사용하여 이미지에서 코너(레이아웃 접합)와 경계 확률 맵(corner (layout junctions) and boundary probability maps)을 예측한다. 모서리와 경계는 각각 방 배치를 완벽하게 표현해 준다.
- ③ 3D Layout parameters는 예측된 모서리와 경계에 적합하도록 최적화된다.



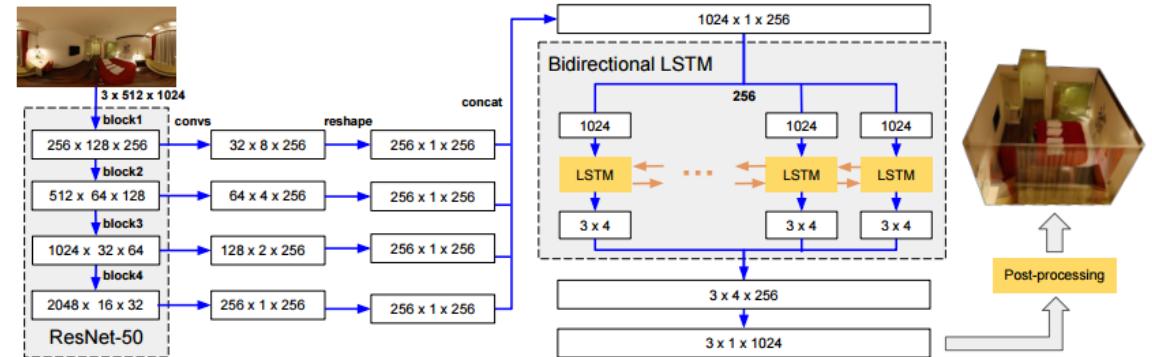
5 정리 및 추후 계획 - 정리

3 HorizonNet

- 목표 : 360도 파노라마 이미지로부터 Manhattan room layout 추정

- 방법

- ① 입력으로 받은 파노라마 사진을 수직 보정하여 소실점과 edge를 찾는다.
- ② ResNet50과 LSTM을 이용하여 훈련된 모델로 사진의 특징을 추출하여 천장-벽 경계, 바닥-벽 경계, 벽-벽 경계가 표시된 1D Layout을 도출한다.
- ③ Manhattan World 가정으로 바닥 천장, 벽면을 복구 : 공식을 통해 천장-바닥 거리를 계산한 후 벽면을 복구함.



[1D Layout]

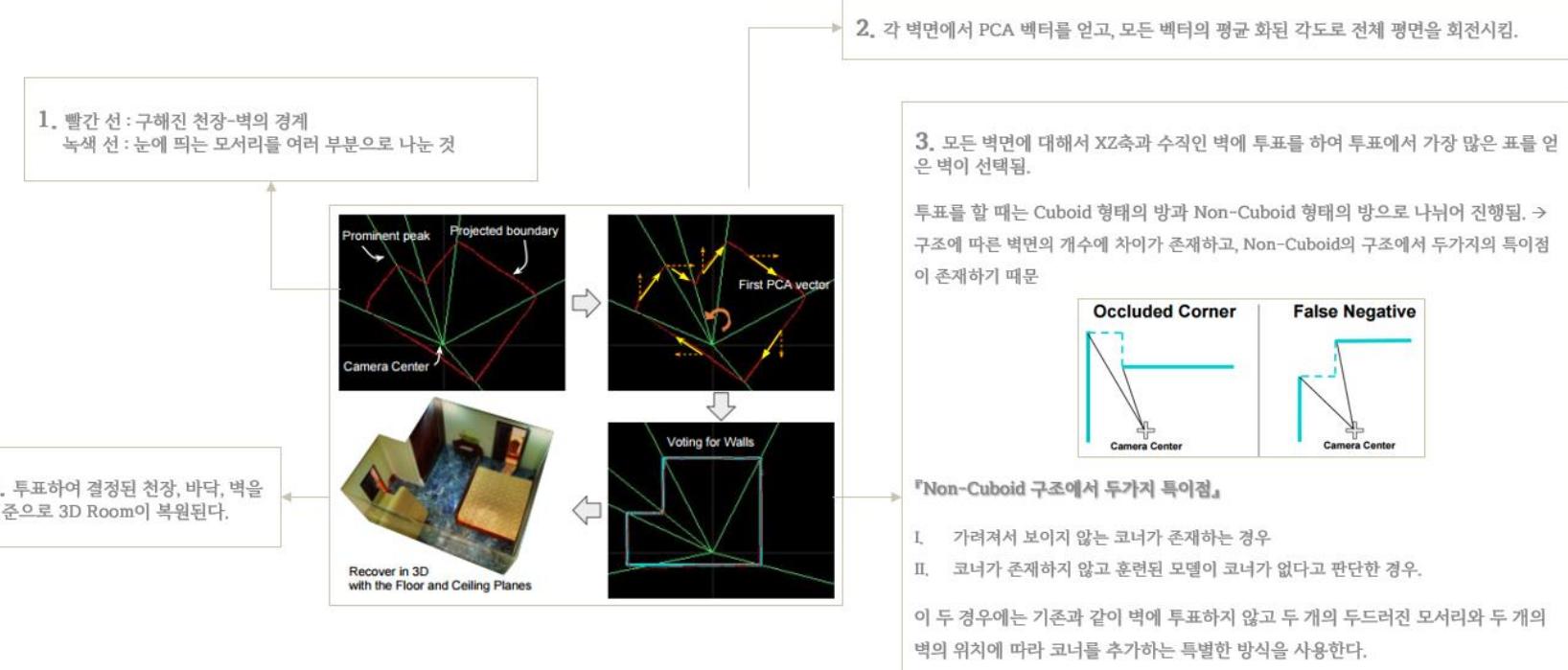


Figure 3: Visualization of our 1D ground truth representation

y_c : 천장-벽 경계

y_f : 바닥-벽 경계

y_w : 벽-벽 경계



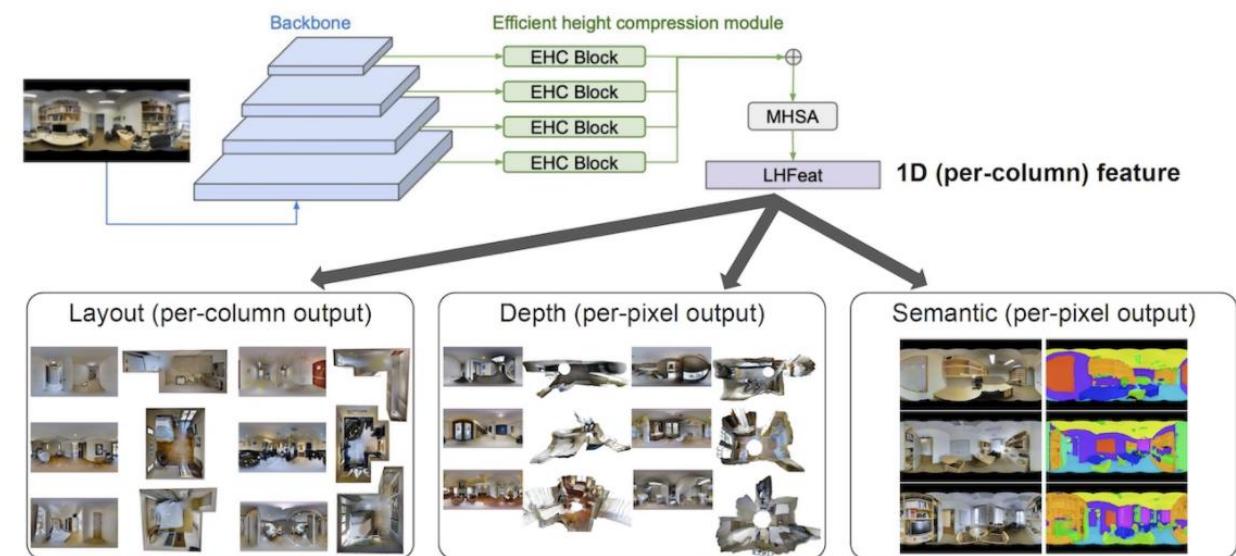
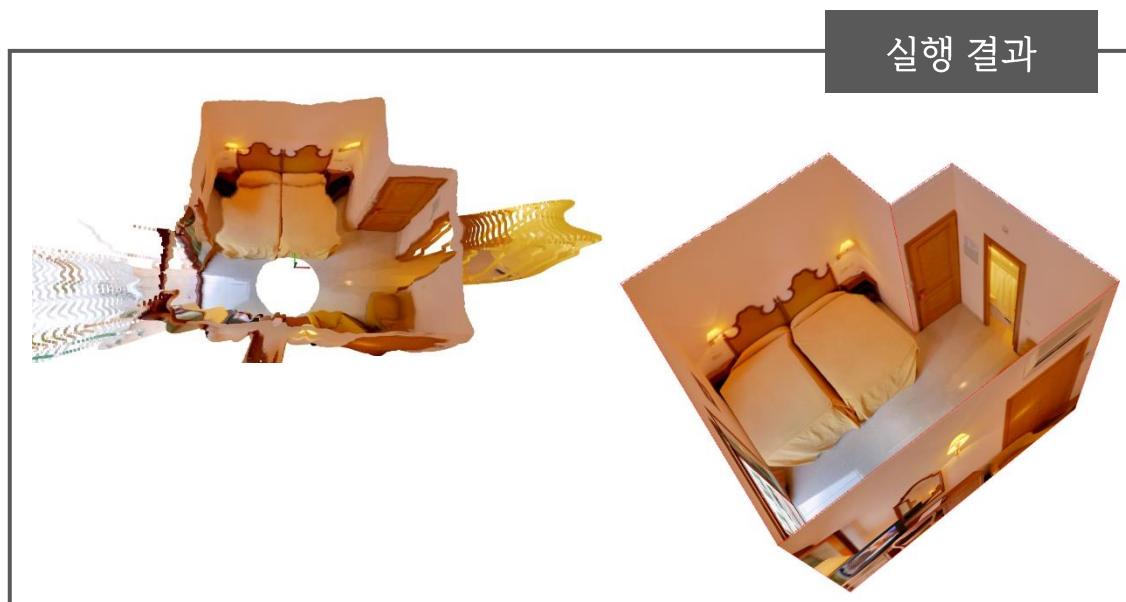
5 정리 및 추후 계획 - 정리

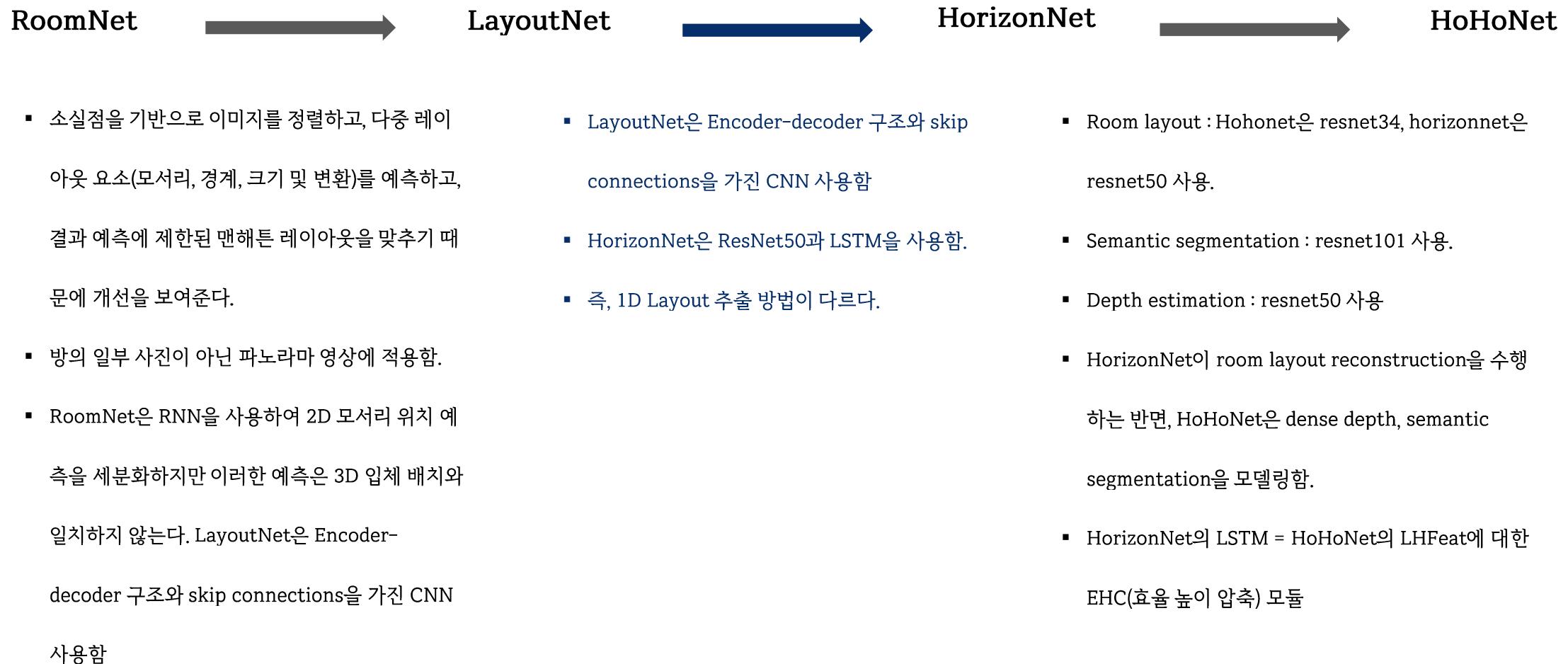
4 HoHoNet

- 목표 : 360° 파노라마를 캡처하는 단일 고해상도 등각 투영(ERP) 이미지에서 전체적인 장면 모델링 문제를 해결하는 것
- 잠재 수평 특징 표현(LHFeat)을 통해 ① 레이아웃 구조, ② dense depth 및 ③ semantic segmentation을 모델링하기 위한 새로운 딥 러닝 프레임워크.
- 레이아웃 재구성을 위한 새로운 방법을 설계하는 것이 아님.

방법

- 고해상도 파노라마는 먼저 **backbone**(예: ResNet)에 의해 처리된다.
- 형상 피라미드는 제안된 EHC(Efficient Height Compression) 모듈과 정교화를 위한 다중 헤드 자기 주의(MHSA) 모듈에 의해 압착 및 융합된다.
그 결과 LHFeat은 compact며(예: 입력 이미지가 $R 3 \times 512 \times 1024$ 인 경우 $R 256 \times 1024$), 전체 네트워크가 기존의 인코더-디코더 네트워크보다 훨씬 빠르게 조밀한 기능을 실행할 수 있다는 점에 유의한다.
- 마지막으로, 최종 예측을 산출하기 위해 1D 컨볼루션 레이어를 사용한다. DCT 주파수 영역에서 예측이 우수한 결과를 가져오기에 각 열의 예측에 IDCT를 적용한다.





5 정리 및 추후 계획 - 추후 계획

- 전체적으로 특정 네트워크나 기술을 통해 어떤 결과가 나오는지는 이해했으나(네트워크 각 구조의 기능은 파악했으나), 구체적인 작동원리를 이해하는데 시간이 오래 걸리고 있다.
 - 『추가로 이해가 필요한 것들』
 - ① RoomNet, LayoutNet : An Encoder-Decoder Based Convolution Neural Network (CNN)의 이해
 - ② RoomNet의 Architecture : MRED Architecture, iterative/recurrent Architecture
 - ③ LayoutNet : bottleneck feature
 - ④ HoHoNet : 역 이산 코사인 변환(IDCT), Semantic segmentation
-

논문 계획

- 서론 : 필요성, 주제(목표) 제시, 전체적인 내용 요약
- 본론 : 알고리즘 제시 – RoomNet, LayoutNet, HorizonNet, HoHoNet의 각 네트워크와 개선된 점
- 결론 : 연구 내용 요약 및 향후 연구 계획 제시



The End