Im2Struct: Recovering 3D Shape Structure from a Single RGB Image

2023.05.19

- Deep convolutional neural networks의 성공으로 이미지 기반 학습의 성능이 향상됨

- 그러나 기존의 모델은 3D 형상의 체적 표현 출력을 목표로 했음. 이러한 모델은 기본적으로 입력 2D 이미지를 3D 이미지(3D 볼륨에서 3D 형상의 복셀 점유)에 매핑하는 방법을 학습함.

- 이미지 간 매핑을 학습하는 데 있어 딥 모델의 높은 용량을 활용하는 반면, 이러한 방법으로 재구성된 3D 볼륨은 3D 형상의 중요한 정보인 형상 토폴로지 또는 부품 구조를 잃게 됨.

- 3D 형상이 체적 표현으로 변환된 후에는 재구성된 볼륨에 위상학적 결함이 있는 경우 형상의 위상 및 구조를 복구하기가 어려움

- 3D 형상(표면 또는 체적 모델)에 대한 부품 분할을 추론하는 것은 어려우며, 분할이 주어지더라도 연결, 대칭, 병렬성 등과 같은 부품 관계를 추론하는 것은 여전히 어려운 과제임.

-> 구조 마스킹 네트워크와 구조 복구 네트워크, 두 개의 네트워크를 학습하고 통합하는 방법을 제안

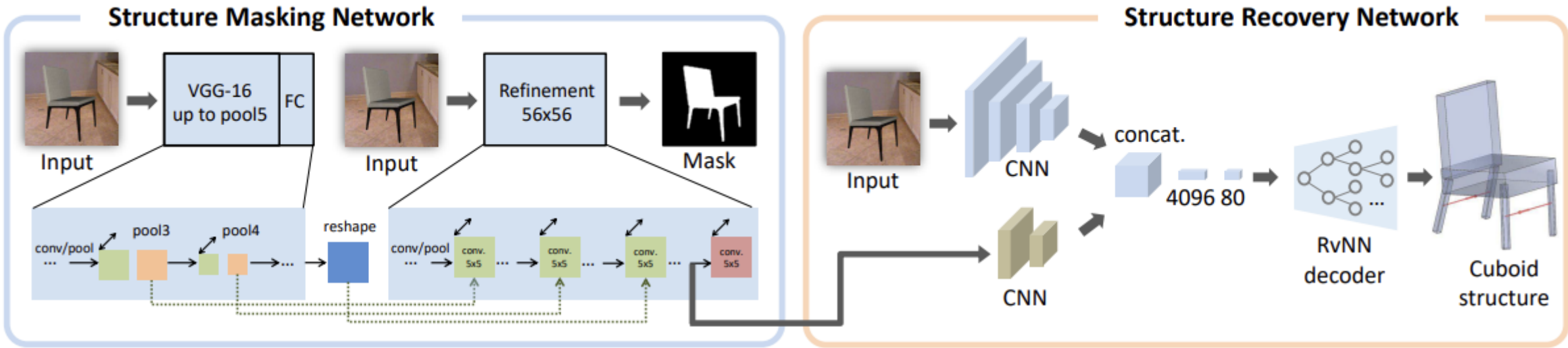
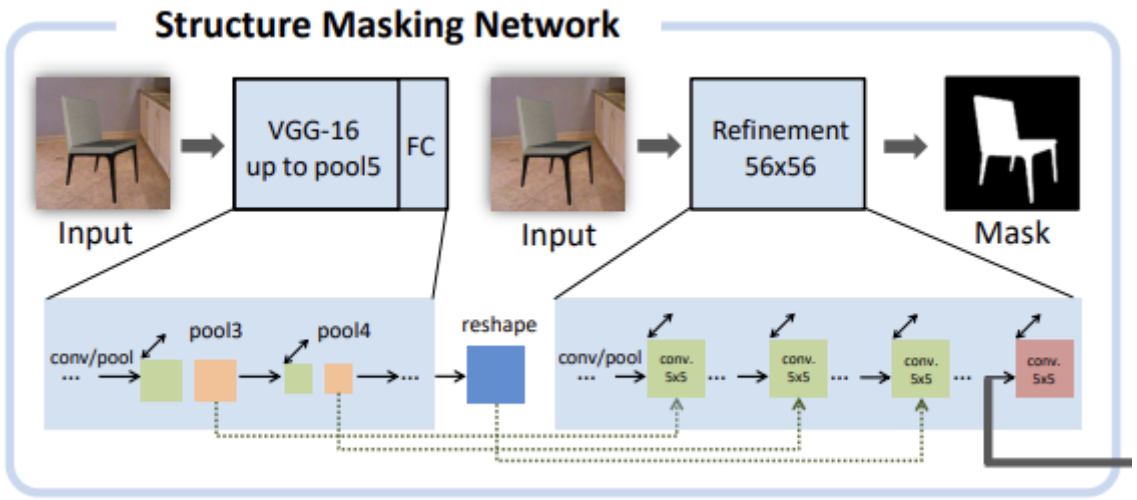


Figure 2: An overview of our network architecture. The structure masking network is a two-scale CNN which is trained to produce a contour mask for the object of interest. The structure recovery network first fuses the feature map of the masking network and the CNN feature of the original image, and decode the fused feature recursively into a box structure. The red arrows in the resultant chair structure (right most) indicate recovered reflectional symmetries between chair legs.

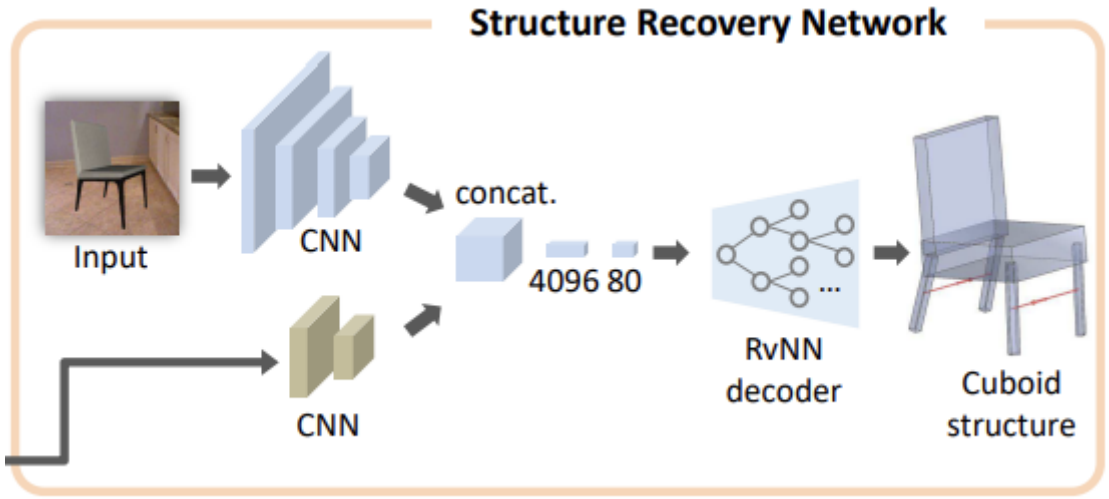
Structure Masking Network

- 상세한 깊이 추정을 위해, 최근 제안된 multi-scale network에서 영감을 받음
- 224 × 224로 리스케일된 입력 RGB 이미지가 주어지면, 입력 해상도의 4분의 1(56 × 56)로 binary contour mask를 출력하는 two-scale structure 네트워크를 설계
- 첫 번째 scale network는 전체 이미지의 정보를 캡처하고 두 번째 scale network는 입력 해상도의 4분의 1로 상세한 마스크 맵을 생성. Prediction target이 binary mask이므로 SoftMax loss을 training loss로 사용.



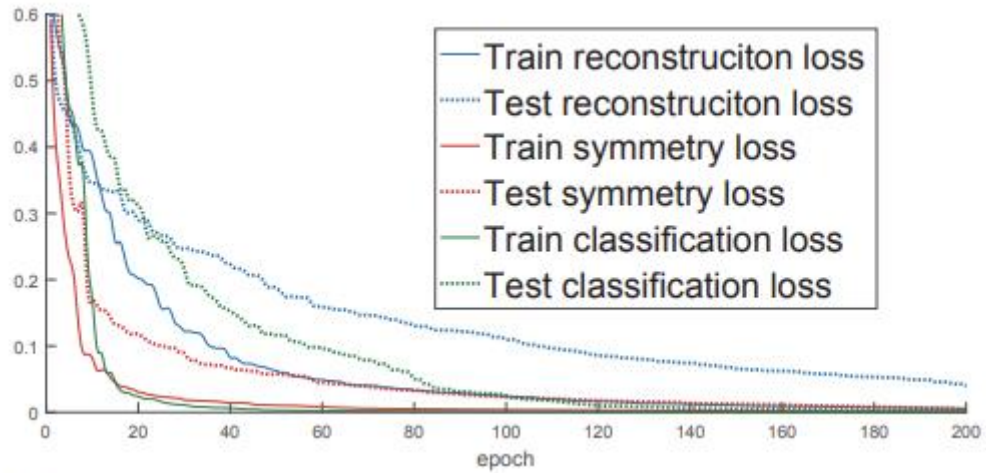
Structure Recovery Network

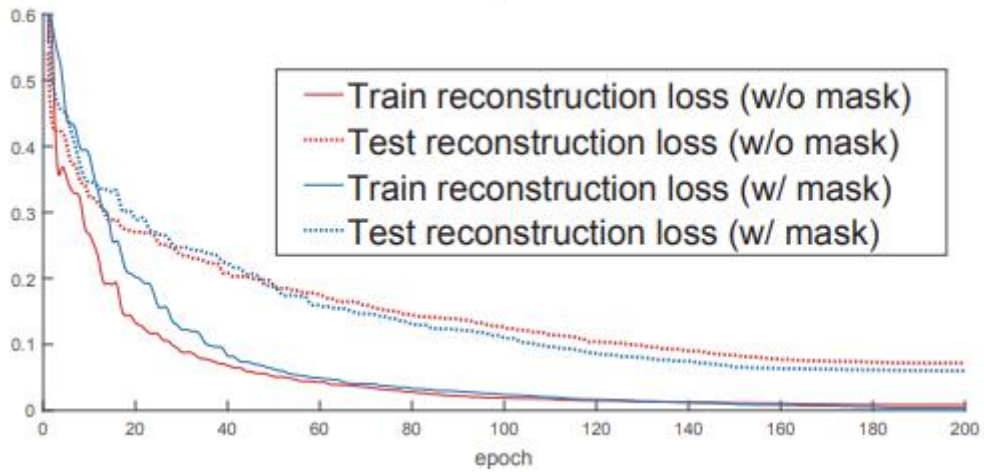
- Structure Masking Network와 input image에 대해 추출된 feature를 bottle feature에 통합하고 이를 재귀적으로 part boxes의 계층으로 디코딩
- Feature fusion
  - 두 컨볼루션 채널에서 특징을 융합
  - 두 채널의 output feature map을 7×7 크기로 concatenate하고 두 개의 fully connected layer를 거쳐 80D 코드로 인코딩하여 input image의 객체 구조 정보를 캡처
- Structure decoding
  - Box structure decoder로 recursive neural network(RvNN)을 채택
  - Root feature code에서 시작하는 RvNN은 재귀적으로 feature의 계층 구조로 디코딩하여 leaf node에 도달할 때까지 디코딩



Training

- Input image에 대한 binary object mask를 추정하기 위해 structure masking network를 학습. Structure masking structure의 첫 번째, 두 번째 scale network는 같이 학습됨.
- 그 다음에는, structure masking network를 refine하고, structure recovery network를 학습한다.
  - structure recovery loss는 box reconstruction error와 노드 분류를 위한 cross entropy loss의 합으로 계산된다..
  - reconstruction error는 각 box와 symmetry node에 대한 input, output parameter 간의 squared differences의 합으로 계산된다.
- 학습 전에 모든 3D shape의 크기를 unit bounding box로 조정하여 서로 다른 shape 간에 reconstruction error를 비교할 수 있도록 함.
- Stochastic Gradient Descent (SGD) 수행.


Train reconstruciton loss
Test reconstruciton loss
Train symmetry loss
Test symmetry loss
Train classification loss
Test classification loss
0.6
0.5
0.4
0.3
0.2
0.1
0
0
20
40
60
80
100
120
140
160
180
200
epoch


Train reconstruction loss (w/o mask)
Test reconstruction loss (w/o mask)
Train reconstruction loss (w/ mask)
Test reconstruction loss (w/ mask)
0.6
0.5
0.4
0.3
0.2
0.1
0
0
20
40
60
80
100
120
140
160
180
200
epoch

Experiment

- dataset containing 800 3D shapes from three categories in ShapeNet: chairs (500), tables (200), aeroplanes (100)
- two subsets for training(70%) and testing (30%)
- Training data generation
  - Image-structure pair generation, Data processing and augmentation
- Qualitative Evaluation
  - Google image challenge for structure recovery
    - Structure recover의 capability, versatility에 대한 정성적 평가를 진행
    - chair, table, airplane 이 키워드를 사용해 google에서 text-based image search를 수행한다.
    - 검색된 이미지 중 상단의 8개 이미지에 대해 3D cuboid structure로 recover를 수행한다..
  - Result
    - Real image에서 상세하고 정확한 방식으로 3D 형상 structure를 recover할 수 있었음.
    - Shape part의 connection과 symmetry relations를 recover할 수 있었으며, 일관되고 그럴듯한 구조의 고품질 결과를 얻을 수 있었다. 그러나 몇몇의 실패 사례도 존재한다.


chair
table
airplane
UNITED

Experiment

- Quantitative evaluation
  - structure masking network의 경우, ground truth mask에 대한 전체 픽셀 정확도와 클래스 별 정확도로 mask accuracy를 평가함

Method	Overall Pixel	Per-Class
single-scale	0.953	0.917
two-scale (w/o jump)	0.982	0.964
two-scale (with jump)	0.988	0.983

Experiment

- Quantitative evaluation
  - shape structure recovery의 경우, 정확도를 평가하기 위해 두 가지 측정 방법을 고안
    - Hausdorff Error, Thresholded Accuracy
    - the structure masking network is simply a vanilla VGG-16 network.
    - Structure masking network는 structure decoding을 촉진함

Method	Hausdorff Error	Thresholded Acc.	
		$\delta < 0.2$	$\delta < 0.1$
Vanilla VGG-16	0.0980	96.8%	67.8%
Structure masking (VGG-16)	0.0894	97.8%	75.3%
Vanilla VGG-19	0.0922	96.4%	72.2%
Structure masking (VGG-19)	0.0846	97.6%	78.5%