Single Image 3D Interpreter Network

Jiajun Wu*[1], Tianfan Xue*[1], Joseph J. Lim[2], Yuandong Tian[3], Joshua B. Tenenbaum[1], Antonio Torralba[1], William T. Freeman[1,4]
1 MIT 2 Stanford University 3 Facebook AI Research 4 Google Research (* equal contributions)

Overview

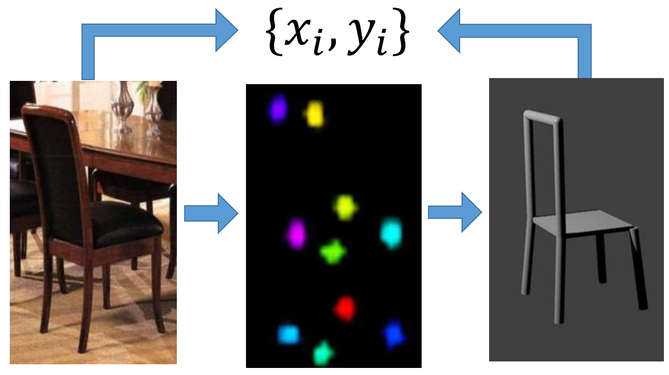
Problem: 3D structure and pose estimation from a single RGB image

Challenge: 3D annotations are hard to obtain
Solution: Use synthetic 3D object models for training

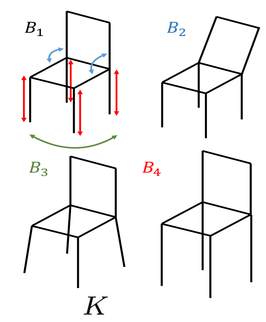
Challenge: Hard to render realistic images with synthetic 3D data
Solution: Use heatmaps of 2D keypoints as intermediate representations

Challenge: Errors propagate in a two-stage model
Solution: Add a 3D-to-2D projection layer for end-to-end finetuning

Pipeline



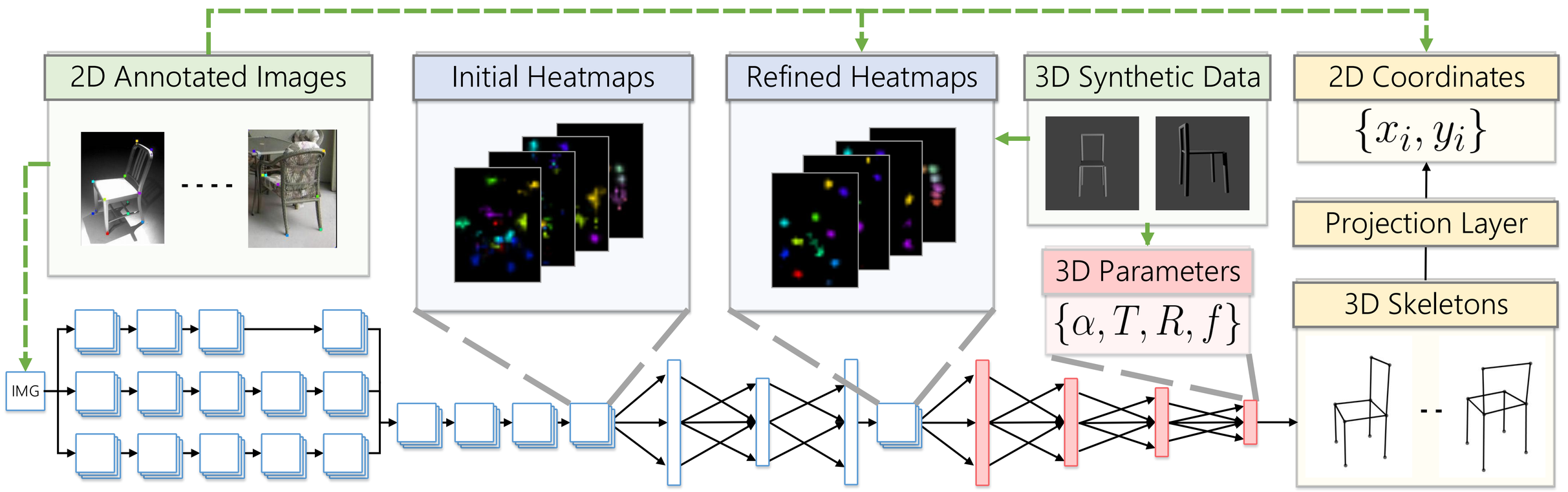
3D Skeletons



$$P(R\sum_{k=1}^{K}\alpha_k B_k + T)$$

weighted sum of base shapes

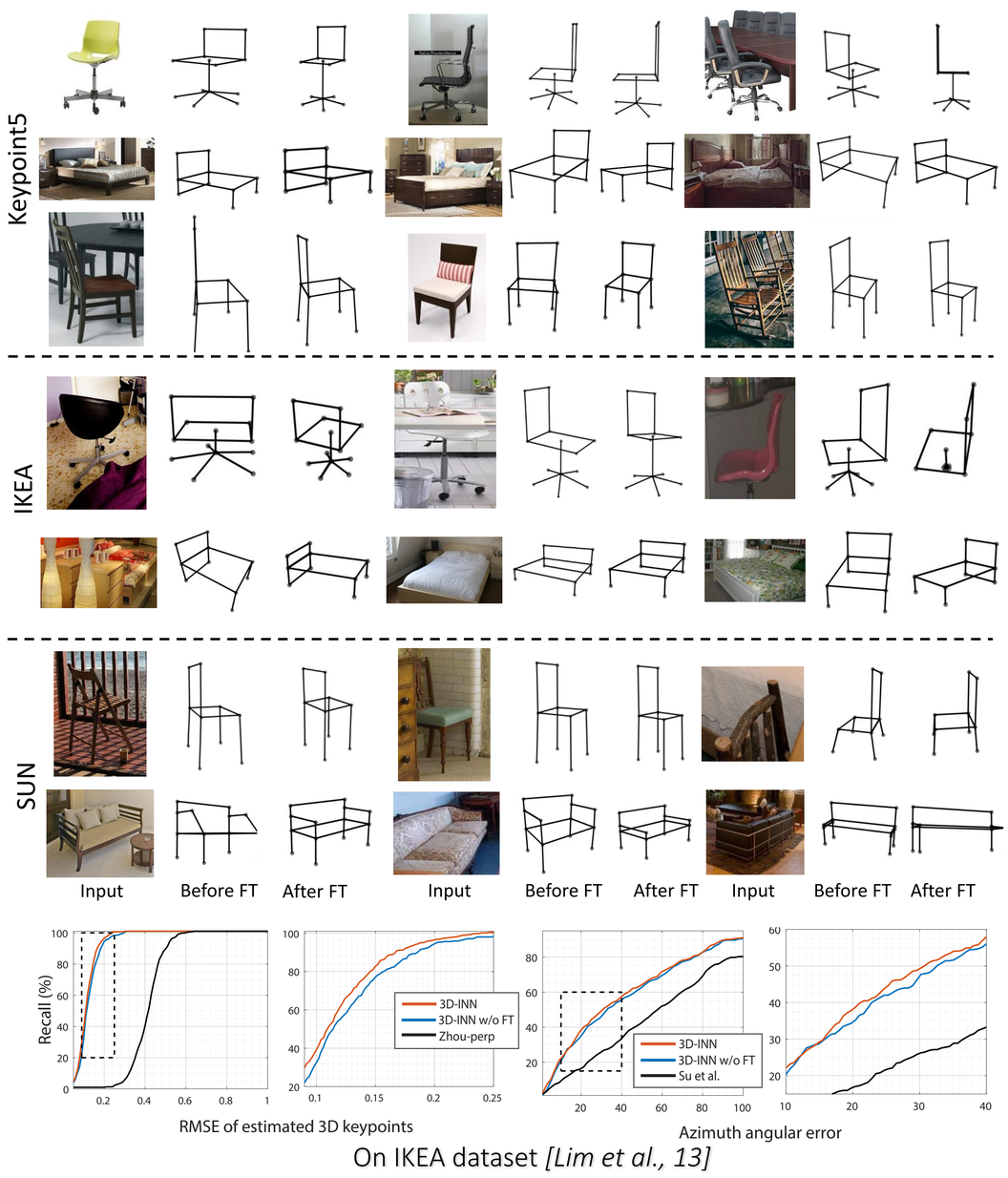
3D INterpreter Network (3D-INN)



Three-stage training: 1) image to 2D keypoint 2) 2D keypoint to 3D skeleton 3) end-to-end fine-tuning

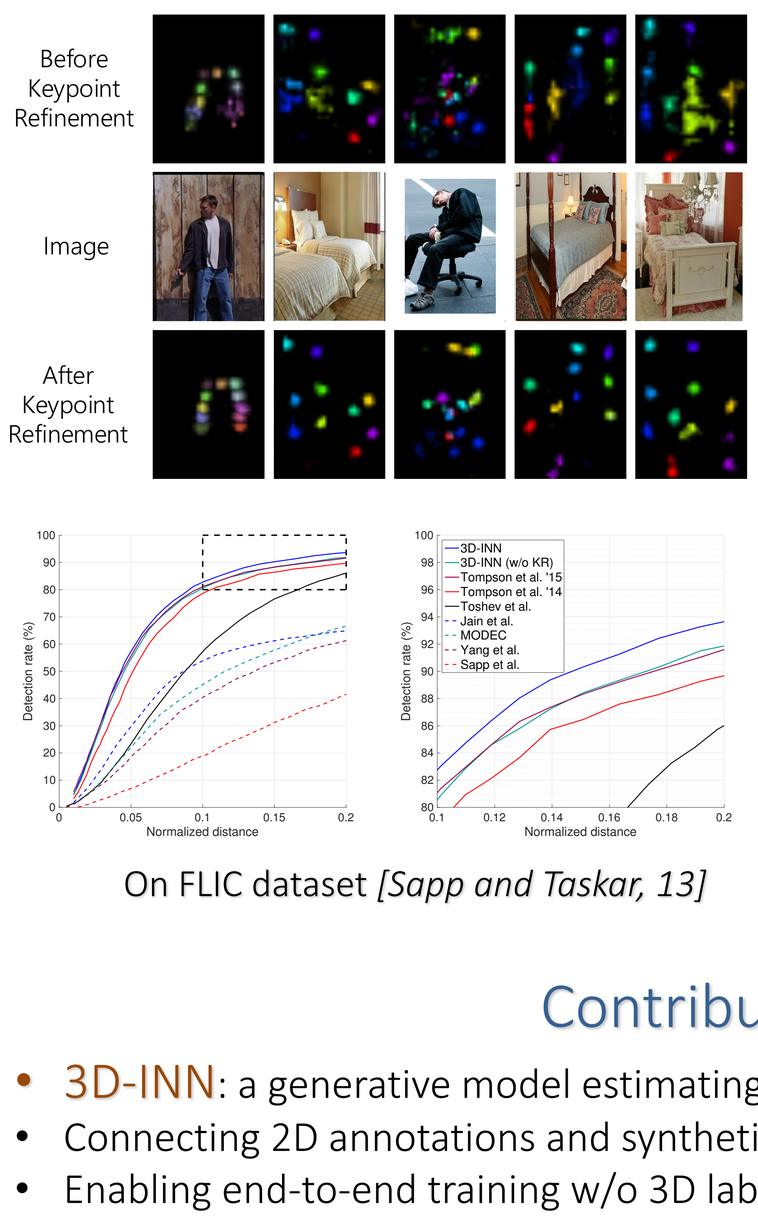
Evaluation

3D Structure and Pose Estimation



On IKEA dataset *[Lim et al., 13]*

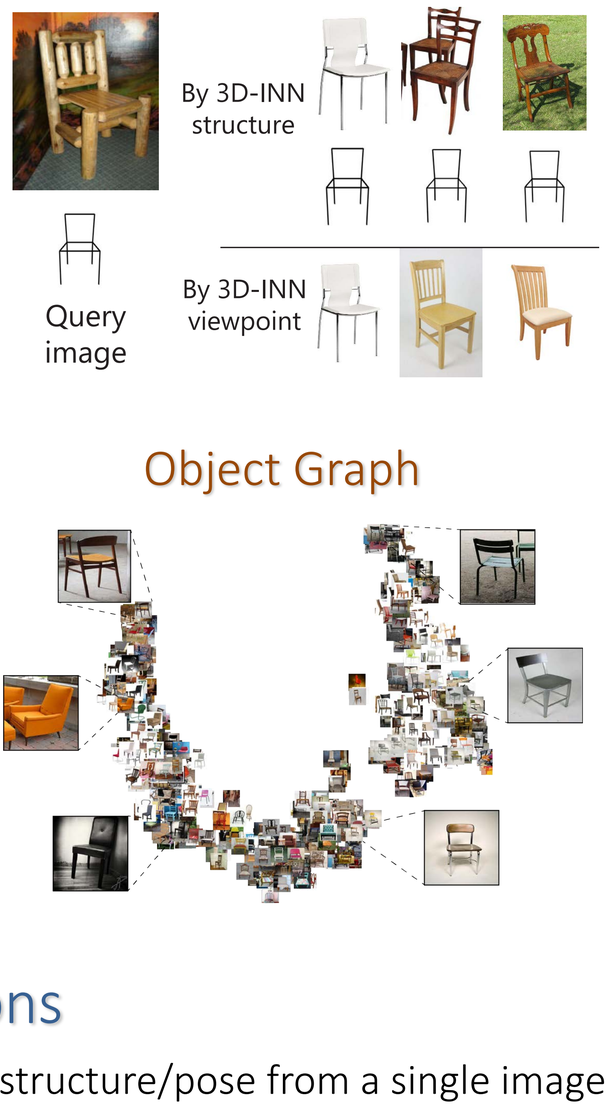
2D Keypoint Estimation



On FLIC dataset *[Sapp and Taskar, 13]*

Visualization

Retrieval



Object Graph

Contributions

- 3D-INN: a generative model estimating 3D structure/pose from a single image
- Connecting 2D annotations and synthetic 3D objects via heatmaps of keypoints
- Enabling end-to-end training w/o 3D labels through a 3D-to-2D projection layer