CSC2457 3D & Geometric Deep Learning

Multi-view 3D Reconstruction for Foot Models with Pix2Vox-LSTM

Date: April. 12nd, 2021

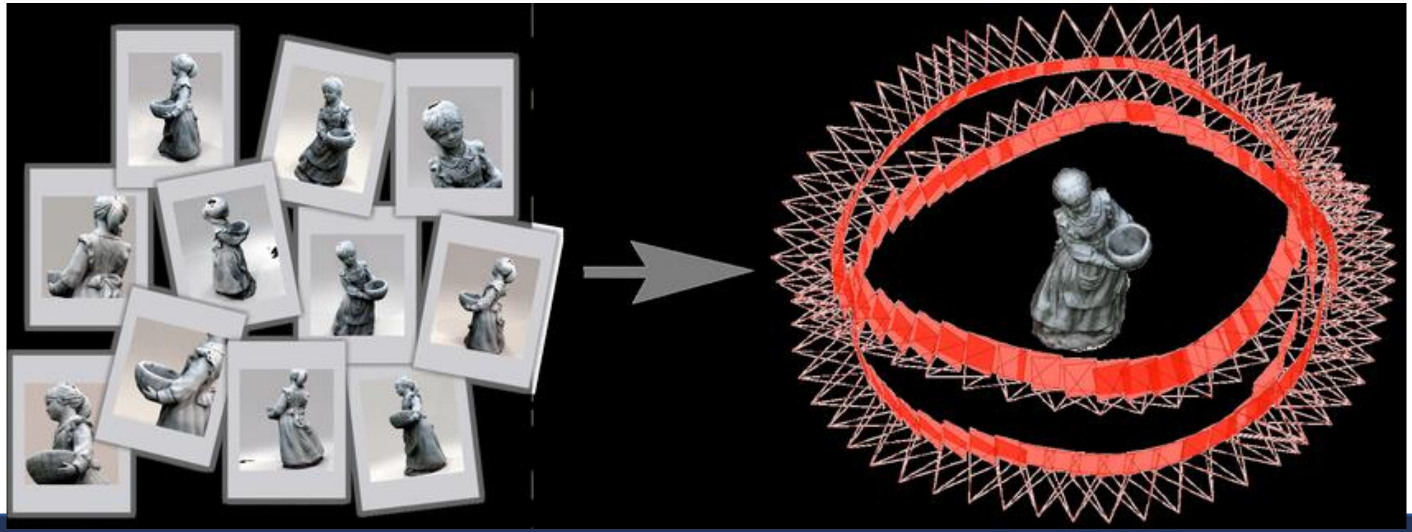
Presenter: Xiang Cao

Instructor: Animesh Garg

UNIVERSITY OF TORONTO

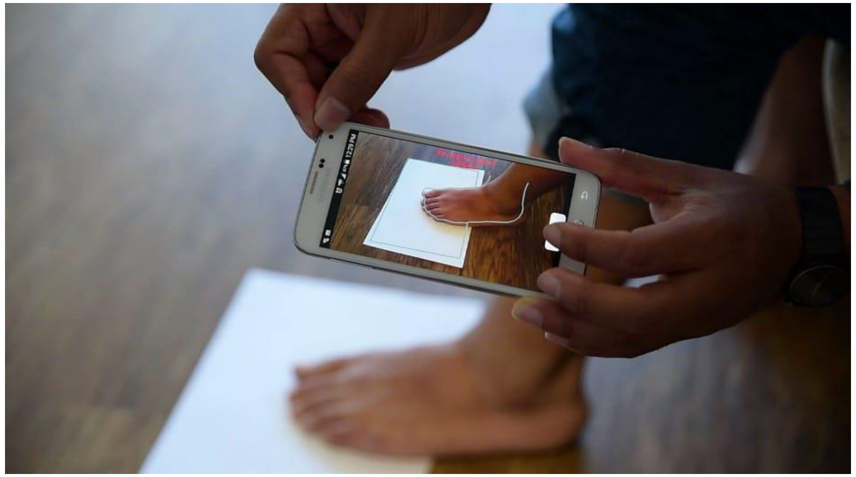
General Background : 3D Reconstruction

- From 2D images to 3D models
- Classical CV methods: structure from motion, multi-view stereo
- Deep learning methods:
    - Multi-view images
    - scene understanding
    - Differentiable rendering



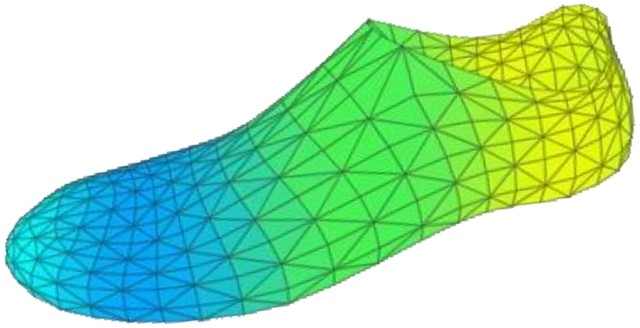
General Background : Foot 3D Scan

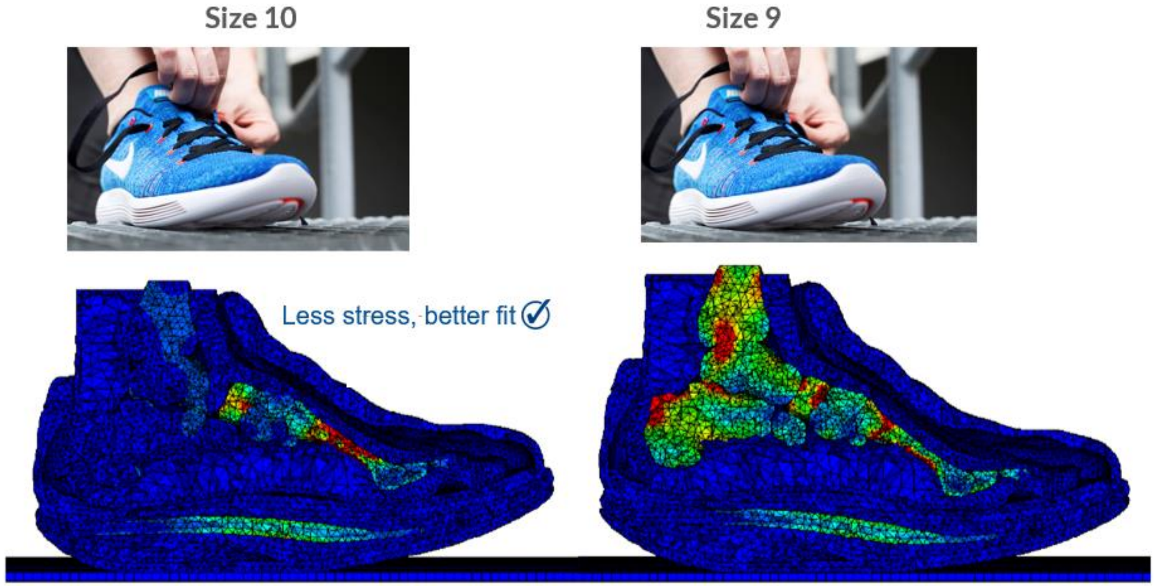
- Precise 3D model of foot are necessary for athletes
- Creating such 3D model of foot require specialized hardware
- Looking for convenient solution to create foot 3D model



Motivation

- 3D Reconstruction: AR, VR, Medical imaging
- Create a PERFIT 3D solution for everyone’s foot
- Improve athletes’ performance and consumers’ comfortness
- 35 billion dollar shoe e-commerce market





Prior Works and Their Limits

- SfM: Structure from Motion, match image feature across views
- 3D- R2N2: 3D recurrent reconstruction neural network

  3D IoU = 0.56

  Use RNN to fuse feature maps from images sequentially

  Recurrent unit is permutation variant
- PSGN: Point Set Generation Network

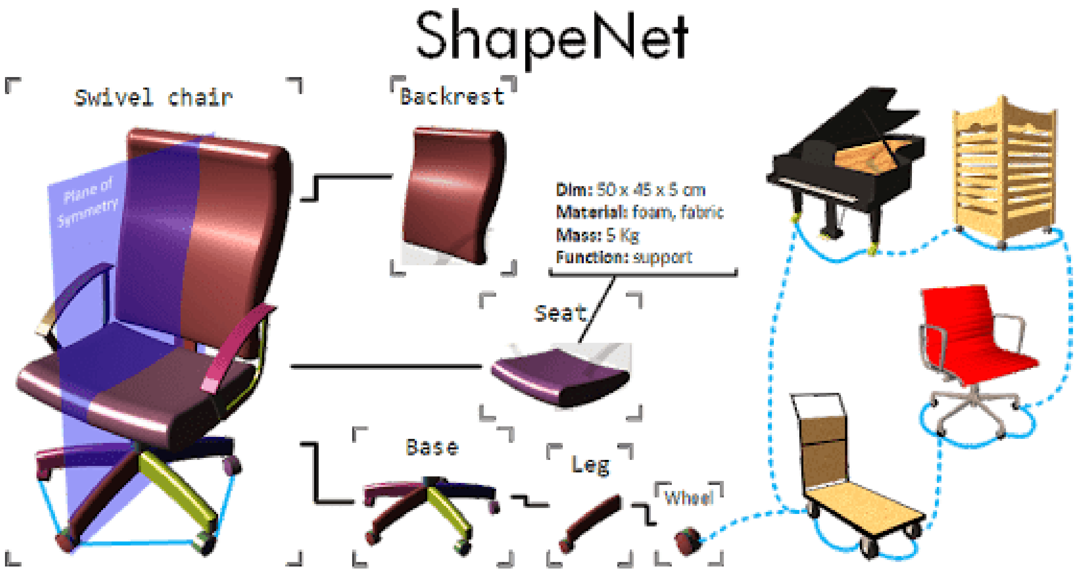
  3D IoU =0.64

  Use conditional shape sampler to predict multiple plausible 3D point clouds

  Model size large and slow on inference

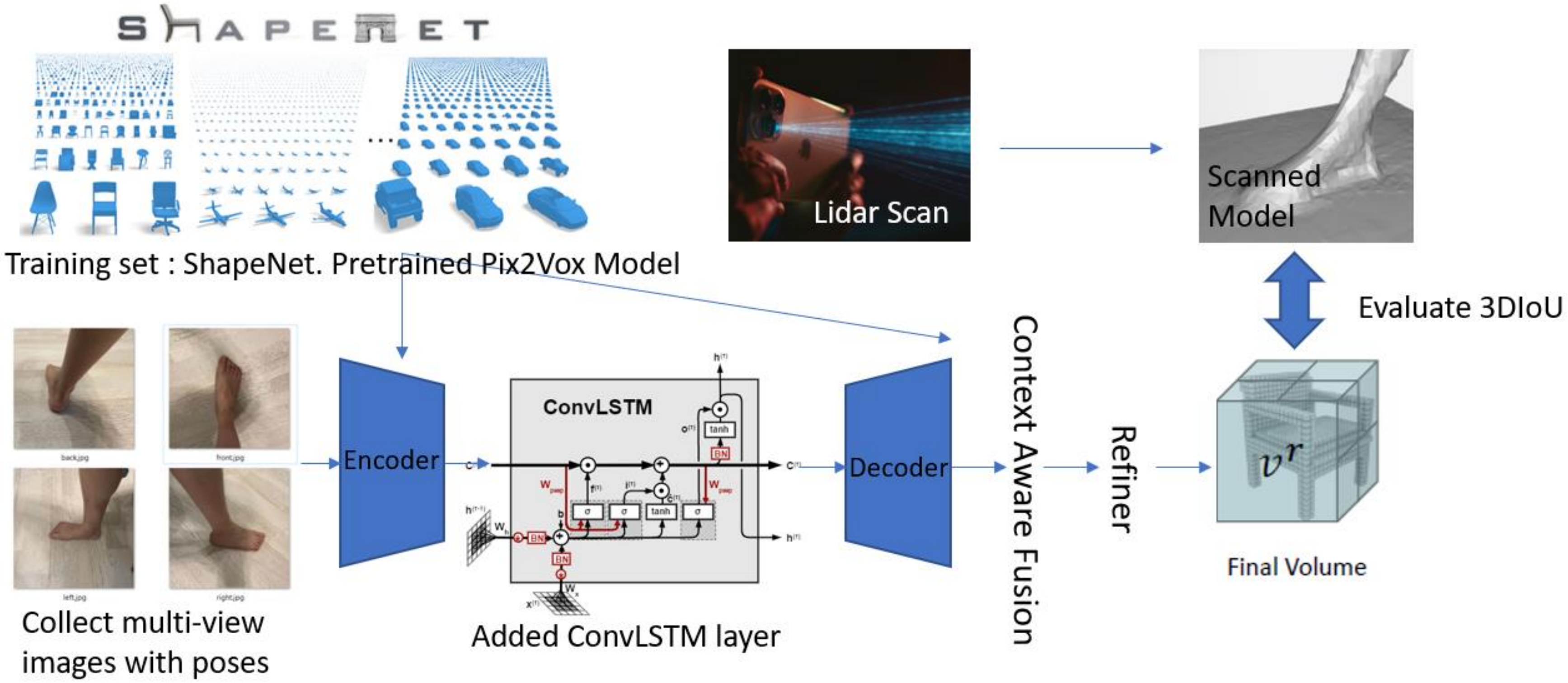
Problem with current 3D Reconstruction

- Many 3D dataset are CAD based with synthetic photos
- Perform poorly on real life photos with challenging lighting and background
- ShapeNet: large-scale richly annotated 3D dataset consist of 3D CAD models from a multitude of semantic categories



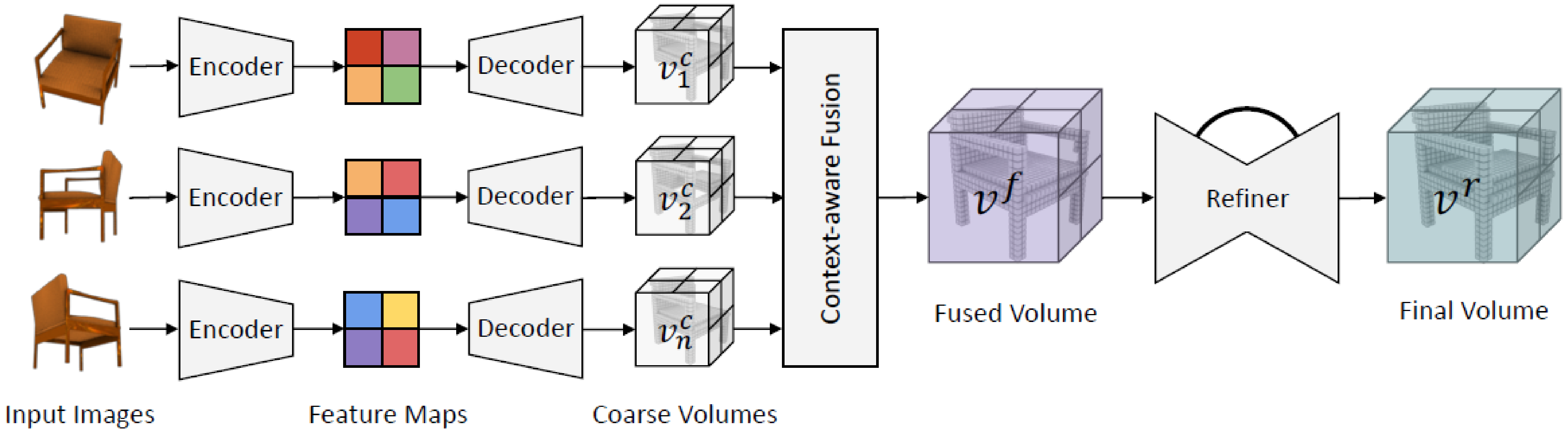
Overall Workflow

- Investigate Multi-view CNN based3D reconstruction method – Pix2Vox
- Create pipeline to hand craft 3D foot dataset with mobile phone
- Evaluate and improve 3D reconstruction method



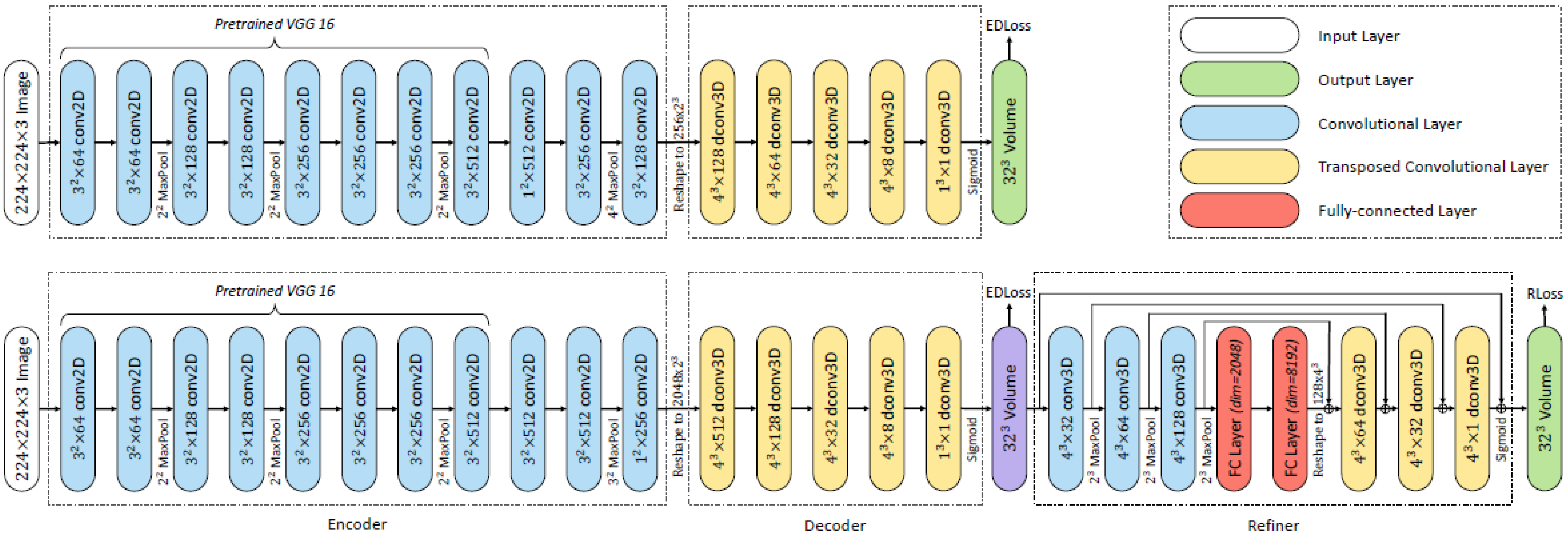
Approach: Pix2Vox

- STOA on ShapeNet dataset
- 3D IoU = 0.66
- Encoder – Decoder – Context aware fusion - Refiner

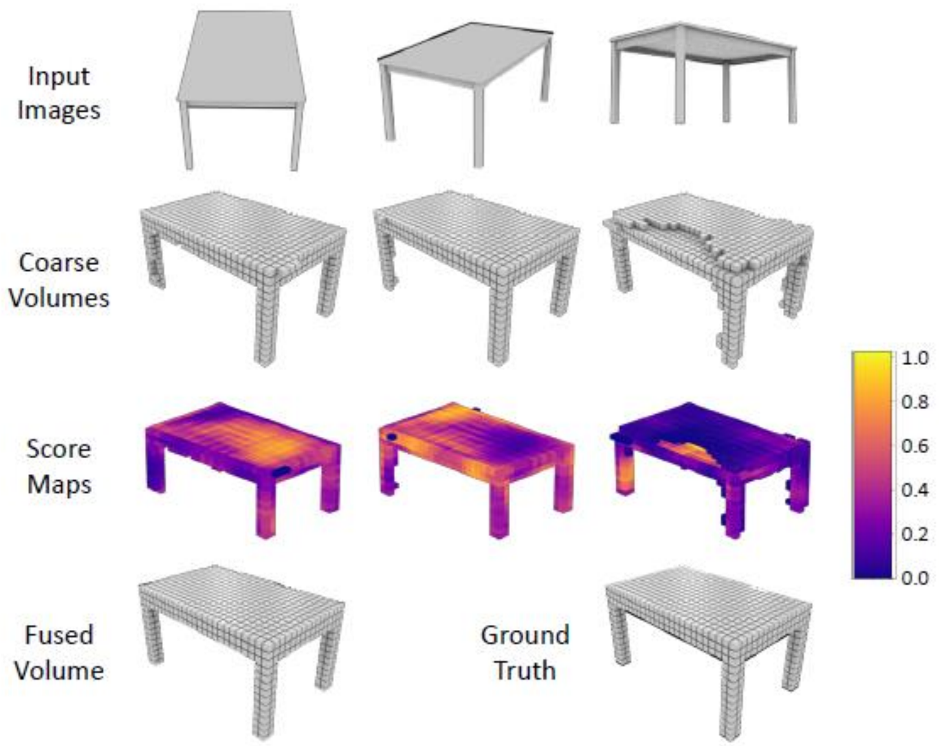


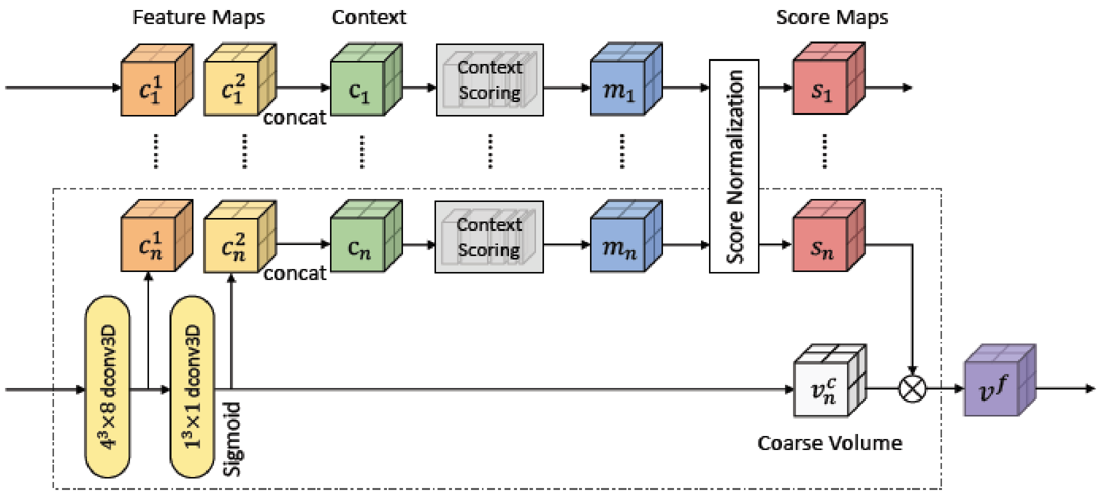
Network Architecture

- Pix2Vox – Fast
- Pix2Vox-Accurate



Context-aware Fusion





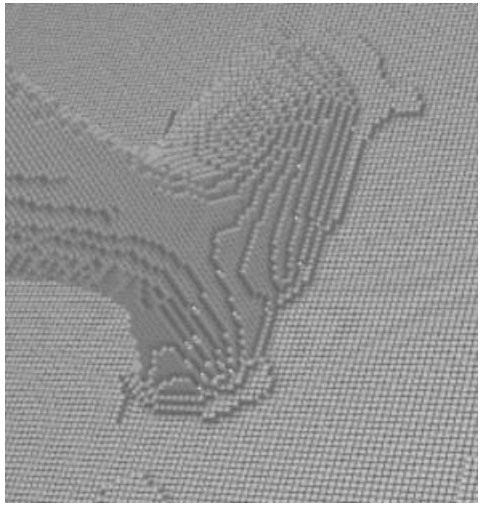
Refiner and loss Function

- Refiner: A residual network
- correct the wrongly recovered parts of 3D Volume
- U-net connections
- Loss function: mean of voxel-wise binary cross entropies

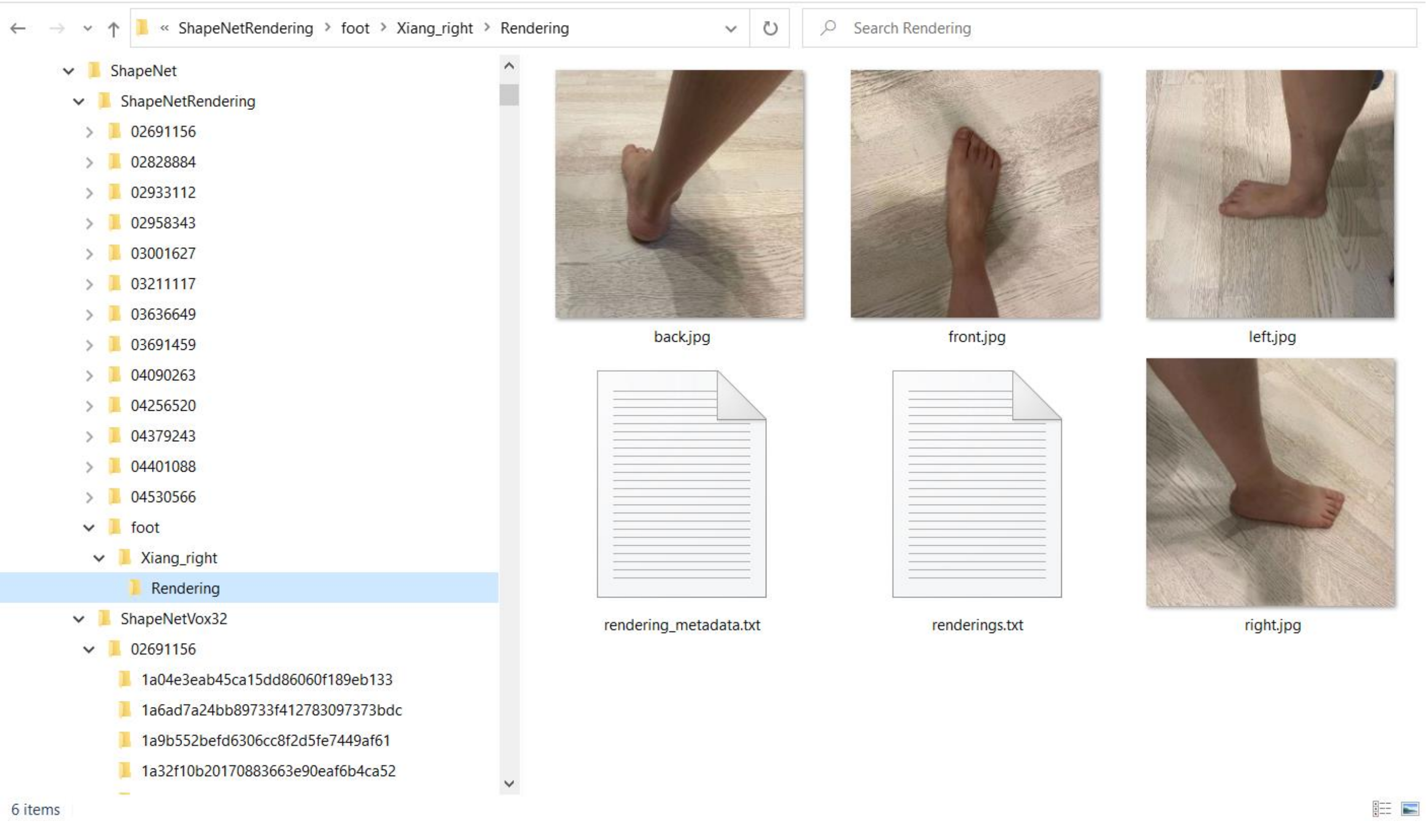
$$\ell = \frac{1}{N} \sum_{i=1}^{N} [gt_i \log(p_i) + (1 - gt_i) \log(1 - p_i)]$$

Dataset collection for foot

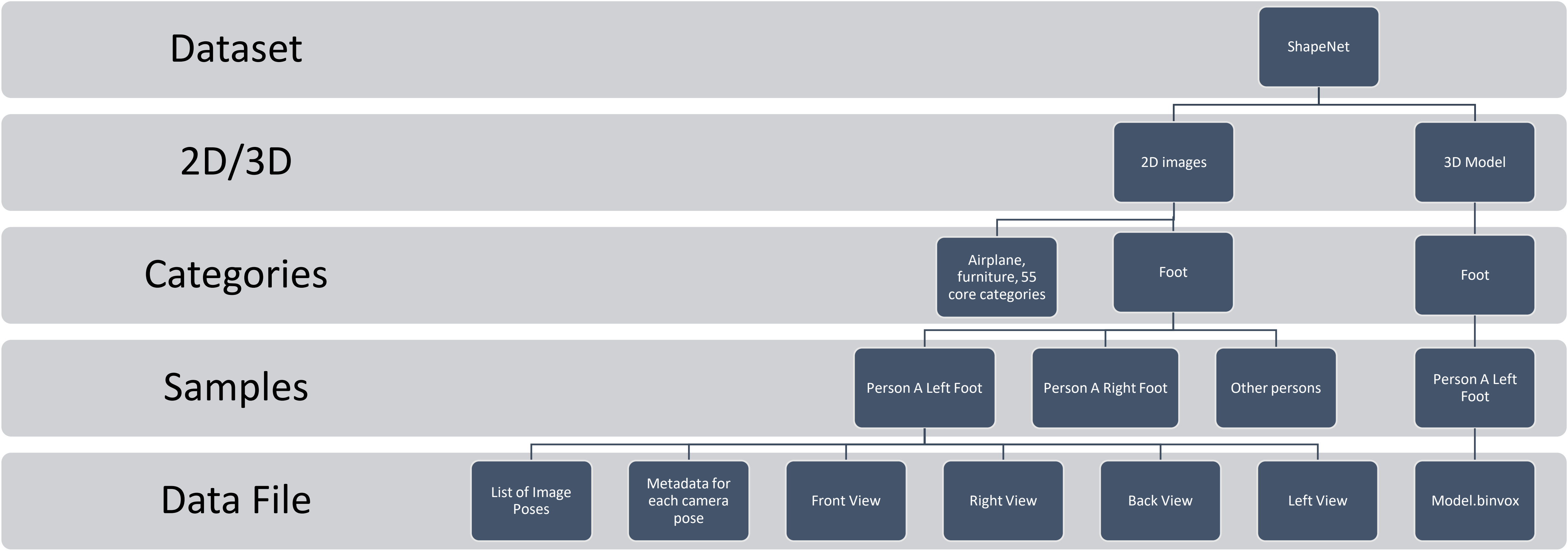
- iOS Apps that use LiDAR scanning:
  - **Polycam**, Qlone, 3d Scanner APP, Scanierverse.
- Convert from textured mesh.obj to .binvox
- Take multi-view images, and record metadata of camera poses:
  - Azimuth
  - elevation
  - in-plane rotation
  - Distance
  - Field of view



Data Collection Process



Augment data with ShapeNet’s Taxonomy



Pix2Vox Results -3D IoU

- Significantly decreased from CAD data
- Similar to result on Pix3D dataset

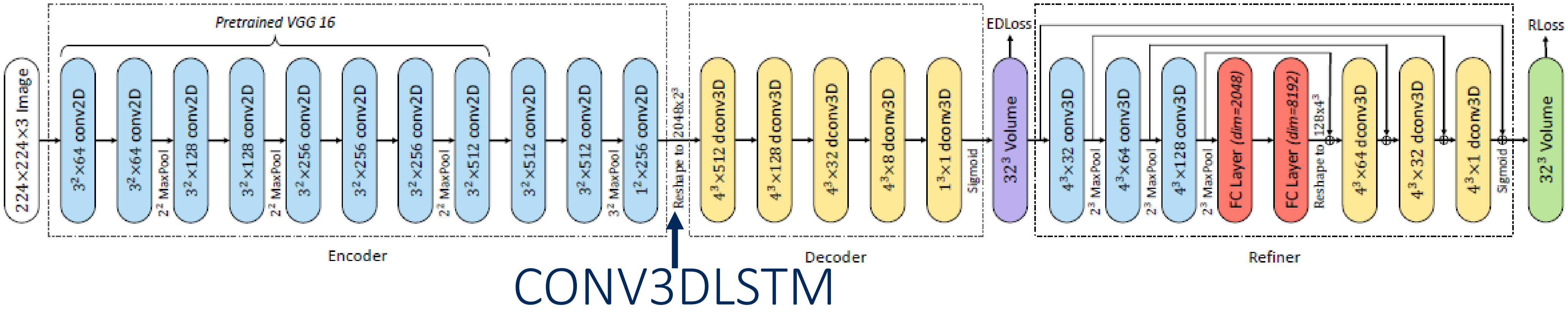
```
================================ TEST RESULTS ================================
Taxonomy                #Sample Baseline        t=0.20  t=0.30  t=0.40  t=0.50
aeroplane               810     0.5130          0.6665  0.6842  0.6903  0.6889
bench                   364     0.4210          0.6024  0.6157  0.6203  0.6151
cabinet                 315     0.7160          0.7895  0.7924  0.7919  0.7847
car                     1501    0.7980          0.8476  0.8548  0.8568  0.8533
chair                   1357    0.4660          0.5638  0.5666  0.5631  0.5493
display                 220     0.4680          0.5347  0.5373  0.5336  0.5188
lamp                    465     0.3810          0.4481  0.4430  0.4340  0.4170
speaker                 325     0.6620          0.7166  0.7144  0.7090  0.6939
rifle                   475     0.5440          0.6042  0.6148  0.6148  0.6050
sofa                    635     0.6280          0.7061  0.7092  0.7080  0.6984
table                   1703    0.5130          0.5977  0.6006  0.5983  0.5857
telephone               211     0.6610          0.7696  0.7764  0.7792  0.7783
watercraft              389     0.5130          0.5898  0.5946  0.5902  0.5779
foot                    20                      0.2556  0.2824  0.2923  0.2765
Overall                                         0.6552  0.6607  0.6608  0.6503
```

Pix3D Testing

Method	IoU
Training on ShapeNet-Chairs	
Pix2Vox++/F	0.179
Pix2Vox++/A	0.204
Training on Things3D-Chairs	
Pix2Vox++/F	0.256
Pix2Vox++/A	0.269

Improving Pix2Vox for real-life photos

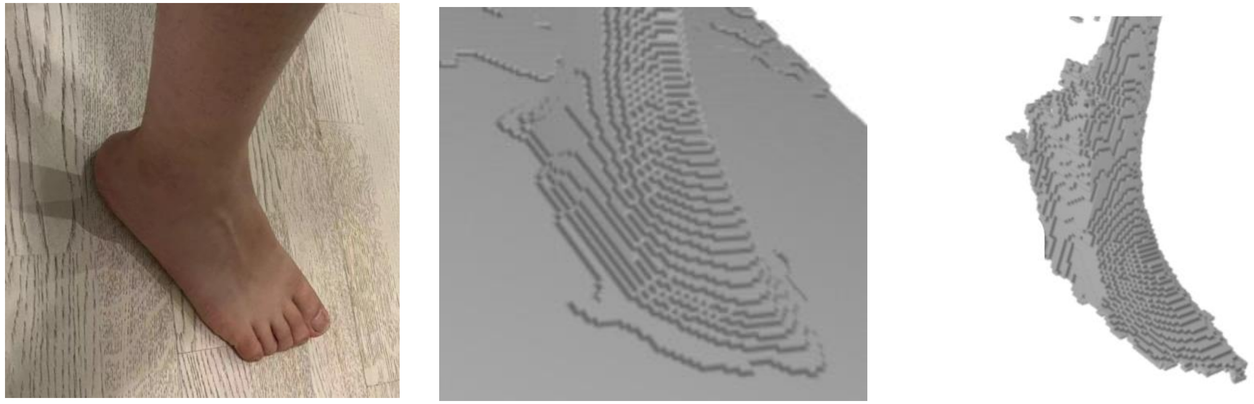
- Reorder 2D images in clockwise order
- Add CONV3DLSTM after encoder
- Extract sequential relation between images
- Change of lighting condition and background has sequential relation



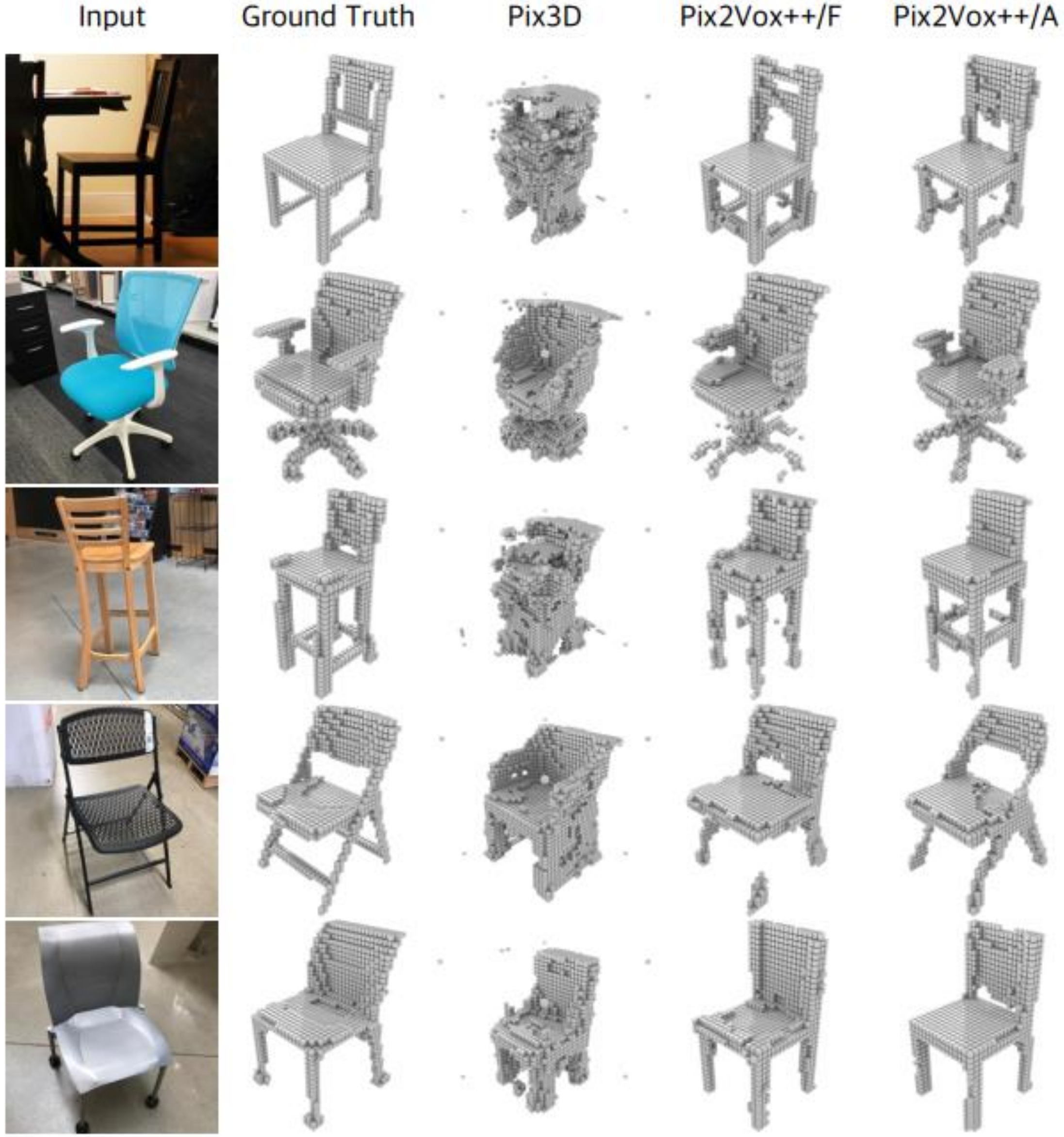
foot	20	0.2897	0.3013	0.3225	0.2963
Overall		0.5852	0.6013	0.6134	0.5987

Qualitative result: Pix2Vox-LSTM

Input	Ground Truth	Pix2Vox-LSTM



Pix3D



Limitations

- 3D Models captured with iPhone Lidar were interfered with flooring and suffered from detail loss.
- Camera pose acquisition not streamlined for crowd-sourcing
- Pix2VOX-LSTM performance is still inadequate for commercial application

Future work

- Find/Develop an iOS APP for taking images with camera poses recorded.
- Explicit model the 3D geometry of camera rays in multi-views to learn better representations
  - Idea from “Atlas: End-to-End 3D Scene Reconstruction from Posed Images”
- Experiment attentional aggregation method called AttSets.

Recap

- Create pipeline to handcraft 3D dataset of foot model
  - Scan 3D model with iPhone12 Pro LiDAR
  - Take multi-view photos with camera poses recorded
- Used Pix2Vox to implement accurate and fast 3D reconstruction
- Evaluate and Improve Pix2Vox with added CONV3DLSTM module, named as Pix2Vox-LSTM