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

Xiang Cao
University of Toronto
Institute for Aerospace Study
tonyxiang.cao@mail.utoronto.ca

Abstract

3D reconstruction has been a popular field in 3D computer vision, which can be used to determine object's 3D profile and 3D coordinates of any point on the profile. In the last 5 years, the performance of 3D reconstruction has been greatly improved by the advancement of deep learning . Meanwhile, an accessible solution to create foot models is in great demand to manufacture well-fit shoes that improve performance for athletes and comfort for consumers. We attempt to reconstruct 3D Models of foot using a multi-view based deep learning approach, built upon Pix2Vox framework. Pix2Vox is one of the state of the art 3D reconstruction method trained on the ShapeNet dataset. In this project we also explored methods to handcraft 3D dataset of foot by 3D scanning foot with the LiDAR sensor equipped on the iPhone12 Pro. In experiments, the Pix2Vox model doesn't perform well on our handcrafted small test set of foot, which reveals a common challenge in 3D reconstruction. Models trained on syntehtic CAD datasetoften failed to generalize well on real world data with more challenging backgrounds and lighting. To address this issue, we rearranged our multi-view images into a sequential orders (clockwise), and added a Convolution-LSTM module to the Pix2Vox model to extract the sequential relation between the images at different angle of the object. The modified model, named as Pix2Vox-LSTM, increased performance on the foot dataset although decreased for ShapeNet test set. Overall, this project implemented and modified the Pix2Vox 3D reconstruction, evaluated on our handcraft realworld foot 3D dataset, and explored options for future application.

1 Introduction

Multi-view 3D Reconstruction aims to create 3D models by using multiple 2D images. 2D images are the most common data format in computer vision, and was originally created by projecting from a 3D scene onto a 2D plane with digital camera. However, the depth information is lost during this projection process. The goal of 3D reconstruction is to project 2D features back to 3D spaces while minimizing the reprojection errors [1]. 3D reconstruction has been actively researched over the last few decades as it has wide range of industry application, from CAD modelling, virtual reality, augment reality, medical imaging, etc. Classical computer vision methods, such as multi-view stereo and Structure from Motion (SfM) were able to construct 3D models from several images taken at multiple views. Classical methods is implemented by feature extraction and matching, which would become very difficult when viewpoints are separated by large margin [2].

Deep learning methods for 3D reconstruction have been rapidly developed over the last 5 years. Various deep learning methods, including RGB-D, volumetric, point cloud processing,and multi-view CNN, have been proposed for different application of 3D reconstruction [3]. The Multi-view CNN methods focuses on reconstructing 3D objects using only 2D images, the most common format of computer vision data. Multi-view CNN methods are built upon deep learning research outcome from

2D image feature extraction using Convolutional Neural Networks(CNN). Many multi-view methods are trained and benchmarked on ShapeNet dataset, a large-scale richly annotated 3D dataset consist of 3D CAD models from a multitude of semantic categories [4]. In this project we thoroughly study and improve a Multi-view based deep learning framework called Pix2Vox [5], and applied it to a real-world application of 3D reconstruction of foot model.

The real world problem to be resolved is constructing 3D models of foot which will be used to customize shoes making or find best fitting shoes. Athletes across multiple sports including runners and soccer players require customized shoes to maximize their performance, while everyday consumers can also benefit from 3D models of their foot to find matching shoes that will improve comfort. Traditionally, the creation of foot 3D model involve additional equipment and cumbersome process such as casting or laser scanners. A more accessible and convenient solution such as using cell phone to create a 3D scan is needed. To resolve this real world problem, we investigate the effectiveness of using Multi-view CNN 3D reconstruction method to create 3D model of foot using multiple photos of foot taken from camera. To evaluate the 3D reconstruction on actual foot data, we also create a workflow to generate 3D dataset from scratch using the Lidar sensor equipped on iPhone12 Pro. The matching 2D images in multiple views are also captured with iPhone12 Pro camera, with Camera poses recorded. The dataset is organized in the same structure as ShapeNet taxonomy. A small test set containing scanned 3D models of foot are collected to evaluate the Pix2Vox performance.

The evaluation results shows a common challenge for Multi-view based deep learning methods: they are trained on synthetic dataset such as ShapeNet and failed to generalize well on real world images. To address this problem, we added a Convolution-LSTM layer after the encoder of the Pix2Vox model to extract the sequential relation between the images at different angle of the object. The improved Pix2Vox model shows improved performance on the foot testset but decreased performance on ShapeNet test set.

In summary, this project works on the following 3 aspect to create a 3D reconstruction work flow of using Multi-view deep learning model to create 3d model of foot.

1. Create pipeline to hand craft 3D foot dataset with mobile phone
2. Implement and evaluate multi-view CNN based 3D reconstruction method: Pix2Vox
3. Add CONV-LSTM module and create variation model named Pix2Vox-LSTM

The overall workflow is shown in Figure 1, where we feed hand-crafted multi-view images to a Pix2Vox model pre-trained on ShapeNet, and evaluate its performance based on 3DIoU against Lidar scanned foot model.

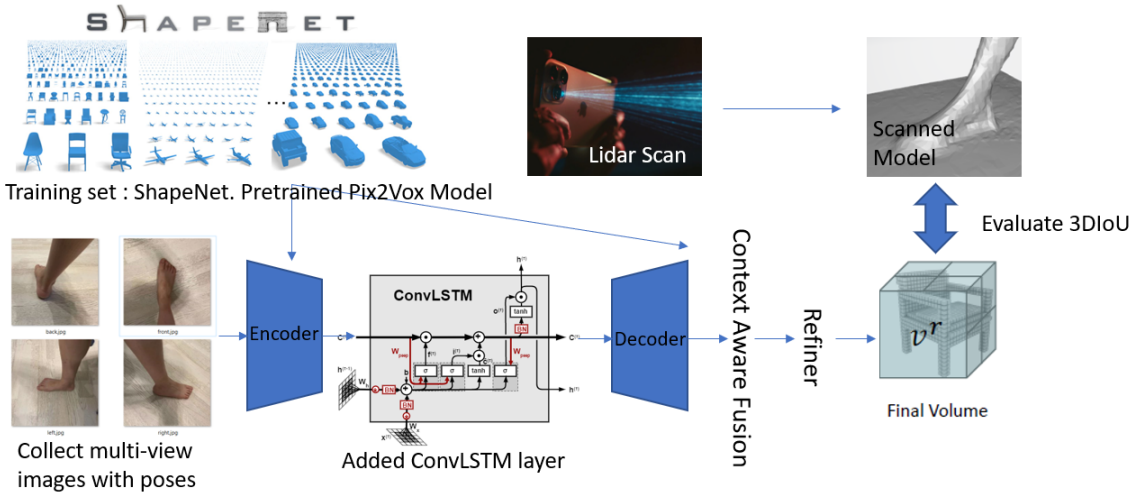

Figure 1: Project's overall work flow to reconstruct 3D model of foot

2 Related Work

2.1 Classical 3D reconstruction method

Classical computer vision methods for 3D reconstruction had been developed and applied in industry for several decades. Prior to the prevailing of deep learning approaches, there are 4 major methods for constructing 3D volume: depth based method, model based methods, structure from motion, multi view stereo. However, those traditional methods would fails in situation when viewpoint changes drastically, lighting condition is challenging, object texture isn't distinctive or baselines are too wide. For example, Structure from Motion is a technique which utilizes a series of 2-dimensional images to reconstruct the 3-dimensional structure of a scene or object. It determines the spatial and geometric relationship of the target through the movement of the camera. However it requires taking photos at fine intervals and does not handle occlusion challenge well [6].

2.2 Deep learning 3D reconstruction method

With the advancement of deep learning, researchers developed multiple approaches to address the 3D reconstruction problem, including RGB-D, volumetric approaches, point cloud processing, differentiable rendering, and Multi-view CNNs approach. RGB-D methods require depth information and has many application in the indoor scene. Point cloud processing methods often use Lidar for autonomous driving application. Some of the recent researches showed promising result with differentiable rendering method. Niemeyer et al. [7] proposed Differentiable Volumetric Rendering method , which can learn implicit shape without 3D supervision. They achieved this by analytically derived depth gradients using implicit differentiation. Liu et al. [8] proposed Soft Rasterizer, which views rendering as an aggregation function that fuses the probabilistic contributions of all mesh triangles with respect to the rendered pixels. This project focuses on 3D reconstruction of individual objects, where the multi-view CNN approach excels at. It utilize Convolutional Neural Network to extract 2D features from 2D images, and later fused to 3D volumes.

2.3 Dataset for 3D Reconstruction

One of the contributor of Multi-view CNN methods is that it can take advantage of abudnance of large 2D image training dataset such as ImageNet and CoCo. This distinctive advantage of large size real world dataset is not availble in 3D datasets. As a context, ImageNet has more than 14 million images while Pix3D only has 418 scanned 3D models. Due to the difficulty of acquiring real-world 3D models to build a dataset, synthetic 3D dataset are most commonly used in 3D reconstruction model training. ShapeNet is one of the most widely benchmarked synthetic 3D dataset with more than 3 million 3D CAD models. This large scale dataset is a richly annotated from a multitude of semantic categories and focused heavily on manufactured items including furniture and transportation tools.

2.4 Comparable multi-view method

Pioneer work in multi-view deep learning method such as 3D-R2N2 [9] and LSM [10] used Recurrent Neural Network to fuse multiple feature maps extracted from input images sequentially. In 3D-R2N2 , several images are taken into the networks that can be from arbitrary viewpoints and outputs the 3D reconstruction in 3D occupancy grid. This model were benchmarked on the ShapeNet dataset and achieved 4-view 3D IoU of 0.625. However such recurrent network is permutation variant and would introduce problem at change of viewpoint. Fan et al. [11] proposed a Point Set Generation Network(PSGN) from a single image, and achieved a 3D IoU of 0.64, by using a conditional shape sampler to predicting multiple plausible 3D point clouds from an input image. The limit of PSGN is it has a large size model and is slow on inference. A novel multi-view approach, Pix2Vox, published in 2019 IICV by Xie et al., outperformed the 2 aforementioned methods on the ShapeNet dataset benchmark with 4-view 3D IoU of 0.697 which was the highest at the time of publication. With its state of art performance, the Pix2Vox model is used as the base model in this project for 3D reconstructing foot model.

3 Methods

3.1 Pix2Vox multi-view reconstruction method

The project starts with evaluating the performance of the Pix2Vox model, which will be explained in this section. Compared with previous methods that fuse deep features generated by a shared encoder, Pix2Vox uses context-aware fusion on multiple coarse volumes produced by a decoder and preserves multi-view spatial constraints better. The overall end-to-end deep learning framework of Vox2Ppix is shown in Figure 3. This workflow recovers the shape of 3D objects from arbitrary single or multiple images. The reconstruction results can be refined when more input images are available.

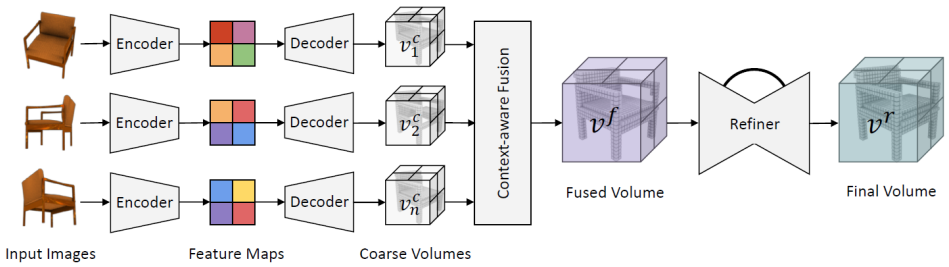

Figure 2: Vox2Pix Overall Framework

The encoder of the Pix2Vox framework is based on CNN and the first 8 layers uses transfer learning from VGG16 to compute features in 2D images.This feature extraction is followed by three sets of 2D convolutional layers, batch normalization layers and ELU layers to embed semantic information into feature vectors. The decoder take the feature map output from the encoder, and transform them into 3D volumes. The decoder consists of 5 Convolution 3D ayers, and outputs a 32^3 voxelized shape in object's canonical view. The context-aware fusion model use the coarse 3D volume from the decoder and corresponding context, and then generates a score map for each coarse volume. It then fuses them into one volume by the weighted sum of all coarse volume according to the score maps. This fusion helps to preserve the spatial information of voxel and score map can adaptively select high quality construction for each part. The refiner is similar to a residual network which can correct 3D volumes that were recovered incorrectly, by using a U-net connections to preserve local structure in the fused volume. Finally, the loss function of Pix2vox is calculating the mean of voxel-wise binary cross entropies between the reconstructed object and the ground truth, as shown in the formula :

$$l = \frac{1}{N}\sum_{i=1}^{N}(gt_i\, log(p_i) \;+\; (1-gt_i)\, log(1-p_i))$$

Xie et al. proposed 2 models of Pix2Vox framework in their paper, Pix2Vox-F and Pix2Vox-A, where the -F suffix stands for fast version and -A suffix stand for accurate version. As shown in Figure 3, the Accurate version of Pix2Vox used an extra refiner with skip connections on top of encoder and decoder shared with the fast version.

3.2 Pipeline for hand craft 3D dataset

Since there are currently no available dataset that contains foot 3D model with their multi-view 2D images, we thus collects additional multi-view foot images and 3D scan models to evaluate the model for this specific application. We want to create a work flow that is convenient and accurate for crowd sourcing dataset using mobile phone equipped with Lidar. This will allow future scaling up the size of dataset as well as generalize this work flow for other categories of data. Several Lidar scanning app on iOS platform were evaluated and Polycam was selected as the final choice with best scan quality. The 3D scans in polycam can be output as a textured mesh format such as .OBJ, .GLB, .USDZ or point cloud format such as .DXF, .PLY, .XYZ, .PTS. We choose to output it to .USDZ format, the native 3D format developed by Pixar and Apple, and use open source software Blender to convert it

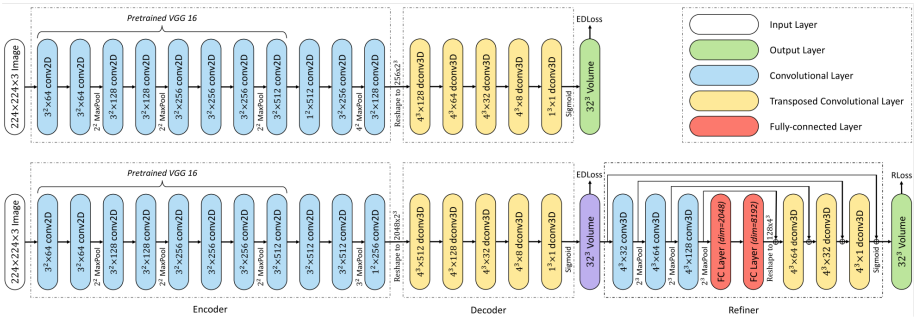

Figure 3: Vox2Pix's 2 Network Structure

to 3D Voxel format .binvox, which is the output format of Pix2Vox model. We convert it to voxel format for the consistency of comparison with Pix2Vox output and ShapeNet dataset, essential for evaluating 3D IoU metric. A sample of converted 3D voxel model is shown in Figure 4.

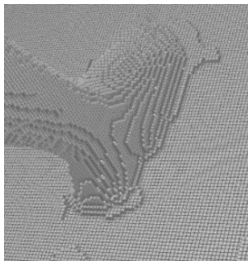

Figure 4: 3D scanned foot from Polycam APP in .binvox format. The toe area suffered from detail loss after the voxel conversion. However the key geometry for shoe making such as foot length, ball and grith circumference, are well preserved.

In addition to 3D Scans, we also create workflow to capture multi-view 2D images in the same format and taxonomy with ShapeNet dataset. For each sample of in ShapeNet dataset, it contains multiple 2D images synthetically rendered from different views, as well as the camera poses information of each view. In order to match the format of camera poses, we also created a metadata.txt file under each folder of 2D images, recording the following metadata for each of the view. As shown in Figure 5, we captured images of foot in 4 different views respectively :front, right, back, left. These 4 views correspondent to Azimuth angle of 0, 90, 180, 270 degree. The elevation angle of back view is smaller than the rest 3 views, as it is located at a position difficult to reach.

Azimuth	Elevation	in-plane rotation	Distance	Field of view
0	60	0	0.5	40
90	60	0	0.5	40
180	45	0	0.5	40
270	60	0	0.5	40

Figure 5: Camera poses metadata

The hand-crafted dataset is organized in the same taxonomy structure as in ShapeNet for ease of evaluation and comparison, as shown in Figure 6. "Foot" is created as a new category, in parallel with other 55 categories such as airplane. Each foot is a unique sample that has both 2D images and 3D models. Under the image folder, we create a list of name of images used, as well as a text file containing the camera poses as in Figure 5. 4 different views of images file in .jpg format and 3D model in .binvox format is saved in different folder respectively.

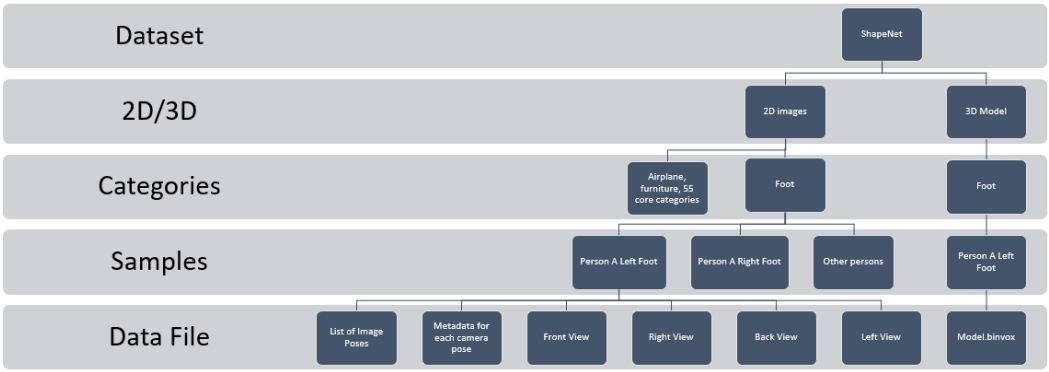

Figure 6: Hand-crafted dataset taxonomy following ShapeNet template

3.3 Model improvement

As pointed out by author of Pix2Vox and Occupancy Network [12], a common challenge in Multi-view models trained on ShapeNet is that they perform poorly on real-world photo when facing with challenge of various lighting condition and backgrounds of objects.In this project, we also explore methods to increase performance of the model on real photos. We propose a method commonly used in video analytics which extract sequential correlations between different views of the images. Initially proposed by Shi et al., a Convolutional LSTM network passes previous hidden state to the next step of sequence [13]. It can thus holding information on previous data the network has seen before to make prediction. We added a Conv3DLSTM Layer after the encoder in the hope of capturing such sequential relations, and the result will be discussed in detail in Section 4.3. One of the pre-requisite of using Convolution LSTM layer is the input images needs to be in sequential order, which is not the case for ShapeNet dataset where images are rendered in random camera view. To accomodate this pre-requisite, we re-ordered images of our foot model testset in the clockwise order of (front, right, back and left views).

4 Experiments

4.1 Experiments on hand-craft 3D dataset

The experiments started with evaluating the effectiveness of 3D scanning on iOS using the latest Lidar equipped on iPhone 12 Pro. Three 3D scanning Apps with high positive ratings on App Store are tested: 3d Scanner App, Scaniverse and Polycam. The 3 Apps have similar functionalities where they use Lidar to scan the environment with distance up to 5 meters, and can output as either Point cloud or textured mesh representation. We choose the textured mesh representation for qualitative evaluation. They also support in-app measurement of scanned 3D models. We use measure tape to record the ground truth of author's actual foot length, and then compare it with the result from the 3D scanning Apps for quantitative evaluation of the scanning accuracy. The evaluation result is shown in Figure 7. The 3D scan generated by Polycam has the best quality with less distortion few occlusion losses and got the correct measurement of 26cm for foot length. After evaluation we select Polycam as our 3D scanning App for dataset collection, and choose .usdz output for further conversion into .binvox using Blender software.

Next we experimented the 2D images collection methods, where the main challenge is recording the metadata of camera poses.We controlled several variable in this experiment: Azimuth, elevation, in-plane rotation and distance. Firstly, We used measuring tape to fix the distance from camera to foot is 0.5m. The we use the perpendicular gap on floor and tiles to confirm the 4 views of image taken are 90 degrees apart from each other. Lastly, we used gimbal for mobile phone to fix the elevation angle at 60/45 degree with zero in-plane rotation angle. The image taking process is cumbersome which would be a challenge for future large scale crowdsource dataset. Due to physical restriction of Covid, only 20 samples were collected from volunteers and their household members in this stage of the project.

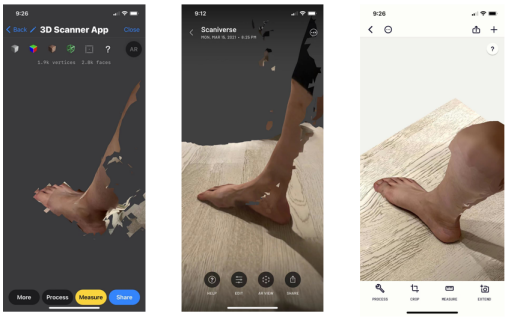

Figure 7: Lidar scan App evaluation result. Ground truth of foot length is 26cm.

4.2 Experiment on Pix2Vox

The Pix2Vox model is implemented in Pytorch, and we ran a benchmark test of ShapeNet Testset on our local computer to evaluate the inference time. As shown in Figure 8, Pix2Vox model perform well on ShapeNet with overall 3D IoU over 0.66 when the voxelization probabilities threshold is selected at 0.2 or 0.3. The total run time for testing Pix2Vox-Accurate-ShapeNet over 8770 3D models only takes 191 seconds, which including inferencing from multi-view images and calculating 3D IoU difference. This 0.02 second/sample inference speed is achieved on our laptop with Nvidia RTX 2060 graphic card, which suggest it can also achieve a fast real-time processing speed on latest mobile processor optimized for machine learning, such as Apple's A14 with 16-core neural engine.

```
============================ 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  0.6147  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
Overall                                 0.6555  0.6610  0.6600  0.6506

[INFO/MainProcess] process shutting down
```

Figure 8: Pix2Vox Testing Result

Next, we modify the ShapeNet dataset .json index file, to add our handcrafted testset of foot 3D Models. We ran the testing again on the 20 test sample and received a 3D IoU score of 0.292 with 0.4 voxelization threshold. The result is significantly lower than the result on ShapeNet however the decreased performance is expected on real world photos, as suggested by Xie et. al's experiment on Pix3D dataset [5]. The 3D IoU scoring of Pix2Vox-A on the Pix3D testset is only 0.204.

4.3 Experiment on Model Improvement

Seeing the challenge of real world photos in the previous experiment, we implemented experiment on model improvement methods as described in section 3.3 of this paper. We rearranged the multi-view 2D images in a sequential clockwise order, so that the image features, lighting condition and background also changes sequentially. We added a CONV3DLSTM module between the encoder and decoder, implemented with Pytorch, retrained on ShapeNet training set, and obtained a slightly improved scoring on foot dataset(real photo) yet decreased scoring on overall ShapeNet testset(rendered photo randomly ordered), as shown in Figure 9. The highest 3D IoU on foot testset is 0.323. We name this modified Pix2Vox model as Pix2Vox-LSTM for the added LSTM module. The implementation code is available on Github :`https://github.com/Tony-Xiang-Cao/3D-Reconstruction-Pix2Vox-LSTM` .

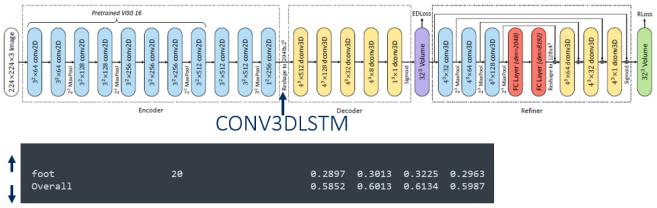

Figure 9: Testing result with modified Pix2Vox-LSTM model

The qualitative result of 3D reconstruction form Pix2Vox-LSTM is shown in Figure 10. The bottom of the foot wasn't reconstructed due to occlusion since all photos were taken at a high elevation angle. This suggest in future a foot bottom image/scan should be also incorporated.

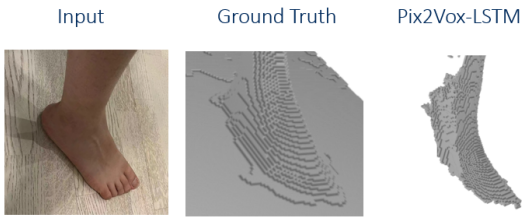

Figure 10: Qualitative 3D Reconstruction result from modified Pix2Vox-LSTM

5 Conclusion

5.1 Limitation and future work

Throughout the experiment, several limitations are found, which suggest directions for future work improvement. The 3D scan generated from the iPhone12 Pro Lidar and Polycam App captures the overall dimension well but loss some detail. This could be improved with further development on 3D scanning Apps which was launched less than 1 year. The 3D scan and 2D images should also include angle aiming at the bottom of the foot, which will create a true 360 degree scan and avoid interference from the floor. The camera pose acquisition process is too cumbersome for large scale crowdsourcing as it still requires measure tape and gimbal. A more streamlined process could involve developing/finding an App that will record all the camera poses information digitally. Another limitation at current stage of the project is that the sample size 20 of the foot dataset is too small to show the statistical significance of the result.

More importantly, the performance of 3D reconstruction model is still inadequate for commercial application. Our initial goal for the Pix2Vox based method is to achieve a 3D IoU of 0.5 on foot testset. However, the performance only increased from 0.292 of original Pix2Vox-A to 0.323 of Pix2Vox-LSTM. Further study and experiment on the model is required to achieve a better result on real world dataset. We will also investigate modifying other state of art Multi-view 3D reconstruction method such as OccNet [12], Atlas [14], and AttSets [15].

5.2 Summary

To sum up, in this project 3 areas were explored for creating 3D reconstruction method for foot model. Firstly, we created a pipeline to handcraft 3D dataset of foot models. We used Lidar sensor and Polycam App to create ground truth 3D model of foot and take multi-view images with camera poses, and organized in ShapeNet's taxonomy. Secondly, we implemented Pix2Vox model for 3D reconstruction, evaluated its accuracy and speed on ShapeNet and foot testset. We found the performance of 3D reconstruction is compromised on real world dataset. Lastly, we added CONV-LSTM module to the network and name it as Pix2Vox-LSTM. Evaluation found slightly improved performance on foot dataset, yet decreased 3D IoU on ShapeNet, which requires further research.

References

[1] O. Ozyesil, V. Voroninski, R. Basri, and A. Singer, "A survey of structure from motion," 2017.

[2] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[3] D. Griffiths and J. Boehm, "A review on deep learning techniques for 3d sensed data classification," *Remote. Sens.*, vol. 11, no. 12, p. 1499, 2019.

[4] A. X. Chang, T. A. Funkhouser, L. J. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu, "Shapenet: An information-rich 3d model repository," *CoRR*, vol. abs/1512.03012, 2015.

[5] H. Xie, H. Yao, X. Sun, S. Zhou, and S. Zhang, "Pix2vox: Context-aware 3d reconstruction from single and multi-view images," in *ICCV*, 2019.

[6] K. Cordes, B. Scheuermann, B. Rosenhahn, and J. Ostermann, "Foreground segmentation from occlusions using structure and motion recovery," in *Computer Vision, Imaging and Computer Graphics. Theory and Application* (G. Csurka, M. Kraus, R. S. Laramee, P. Richard, and J. Braz, eds.), (Berlin, Heidelberg), pp. 340–353, Springer Berlin Heidelberg, 2013.

[7] M. Niemeyer, L. Mescheder, M. Oechsle, and A. Geiger, "Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision," 2020.

[8] S. Liu, T. Li, W. Chen, and H. Li, "Soft rasterizer: A differentiable renderer for image-based 3d reasoning," 2019.

[9] C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese, "3d-r2n2: A unified approach for single and multi-view 3d object reconstruction," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[10] A. Kar, C. Häne, and J. Malik, "Learning a multi-view stereo machine," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA* (I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, eds.), pp. 365–376, 2017.

[11] H. Fan, H. Su, and L. J. Guibas, "A point set generation network for 3d object reconstruction from a single image," *CoRR*, vol. abs/1612.00603, 2016.

[12] L. M. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger, "Occupancy networks: Learning 3d reconstruction in function space," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 4460–4470, Computer Vision Foundation / IEEE, 2019.

[13] X. Shi, Z. Chen, H. Wang, D. Yeung, W. Wong, and W. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," *CoRR*, vol. abs/1506.04214, 2015.

[14] Z. Murez, T. van As, J. Bartolozzi, A. Sinha, V. Badrinarayanan, and A. Rabinovich, "Atlas: End-to-end 3d scene reconstruction from posed images," in *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part VII* (A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, eds.), vol. 12352 of *Lecture Notes in Computer Science*, pp. 414–431, Springer, 2020.

[15] B. Yang, S. Wang, A. Markham, and N. Trigoni, "Attentional aggregation of deep feature sets for multi-view 3d reconstruction," *CoRR*, vol. abs/1808.00758, 2018.