Final Report

Title: 3D Model Reconstruction from 2D Images
Authors: Shubh Patel, Tejas Patel

Abstract

This deep learning project focuses on reconstructing 3D objects from 2D images captured from different angles i.e. 5 views. The goal is to take multiple 2D images of an object and reconstruct its 3D shape using advanced deep learning techniques alongside predicting its labels. Bridging the gap between 2D visual data and 3D object reconstruction has significant applications in fields like Augmented Reality (AR), Virtual Reality (VR), gaming, robotics, and 3D printing. Utilizing convolutional neural networks (CNNs) and advanced reconstruction algorithms such as 3DR2N2 and Nerf based techniques, this study achieves accurate and efficient transformations of 2D images into 3D voxel grids.

As input, we provided some images with different angled photos as well as suitable masks for each photo and then feed it to the network and as a result, we got the output as a 3D object constructed by very small dots using voxel grids of the form of 0 and 1, alongside the label of the input image. We have achieved a satisfactory result by using this model with pix3D dataset and handpicked input, validation and testing sets.

For the 3d object reconstruction model we were successfully able to construct the voxel grid for the input image and for the CNN classifier model we were able to achieve 0.702 accuracy, 0.728 Precision, and Recall of around 0. 70253, we were also able to visualize the confusion matrix for the labels to better understand the perks and perils for this algorithm.

---

1. Introduction

Reconstructing 3D models from 2D images is a critical task in computer vision with numerous applications, including AR/VR, robotics, gaming, and 3D printing. Most visual data, such as images and videos, are captured in 2D, while objects naturally exist in 3D. Bridging this gap enhances the realism and functionality of virtual environments and enables efficient 3D model generation for practical applications. This project explores deep learning techniques to create an end-to-end pipeline for transforming multiple 2D images into 3D models. Our approach emphasizes efficiency and accuracy in generating high-quality 3D reconstructions. For this project the whole idea was to create and train the 3d construction models from scratch rather than using any pretrained model for that.

---

2. Related Work

Significant research has been conducted in the domain of 3D reconstruction.

1. Choy et al. proposed 3D-R2N2, which uses recurrent neural networks to predict volumetric shapes from 2D views

2. Wang et al. developed Pixel2Mesh, leveraging graph convolution networks for mesh-based 3D reconstruction.

3. Mescheder et al. introduced Occupancy Networks, which represent shapes as continuous fields, enabling high-resolution reconstruction.

Our work integrates aspects of these approaches, enhancing scalability and accuracy through preprocessing and tailored CNN architectures.

---

3. Overview

Here, we have mainly 3 components in our Nerf Based Architecture,

1. Feature extractor – we are using CNN based design to extract features from the images.
2. Feature Aggregator – Here, we are combining all collected features into a latent.
3. Decoder – It transforms the latent files into 3D voxel grid of the size (50,50,50) which sums up to 250000 voxels point which can be active/unactive during visualization.

We have also trained our model on the near to replica of 3DR2N2 architecture proposed by one of the papers of the StandFord University. For that architecture we have used different stages of CNNS along side combing a Recurrent based architecture such as LSTM or else normal RNNs for experiment purposes.

---

4. Problem Statement

Most visual data captured in daily life, such as images or videos, are of the type 2D, while objects exist in 3D. Bridging this gap enables the creation of realistic models for AR/VR applications, gaming, and robotics. This technology can also be applied to 3D printing, where a 2D image generates a 3D model ready for printing, eliminating the need for manual creation in tools like Blender.

We are trying to make our lives and technology we use easy to use by integrating it into our life, 3D technology is very useful in that. Our project is makes 3D model from same object but in different scenarios, As we remove background with the help of masks it loses the focus on background and can know where the object is situated in the image alongside with the background removing techniques we have also used some of the image preprocessing techniques such as Data Augmentation which included flipping and transformation of the images alongside we have also use the Image normalization concept.

There are so many specific challenges here, like too few images to train the model for one object and we use augmentation techniques like rotation and flipping to overcome that.

Another thing was we had a very big dataset of 13 GB, and we downsized it to only 3 GB by manually filtering all the proper images, their masks as well as the voxel grid to train the model.

Our voxel grid size was also 80*80*80, which would take very long to be trained as there would be around 512000 active/unactive voxels. So, we down sampled it to 50*50*50.

One of the challenges for this model was to predict perfect threshold to visualize an image. For example, as we know a voxel grid has a value of 0,1 so it is kind of binary classification. But the architecture we discussed above which are used for this problem statement give output in the range of the continuous variable although the end activation function is sigmoid, so we need to find a perfect threshold for visualization, we have also used one of the scikit-learn available functions for that alongside we have also done some hyperparameters tuning for that.

---

5. Dataset

The Project employs Pix3d Dataset which have four thing under it first of all the images related to the labels, voxels grid for all of the images, 3d object of the extension ".obj", and at last there were mask available for each of the object which were further used for the purpose of the background removing of an image. This dataset spans nine object categories, including beds, bookcases, and chairs, among others. Each object has multiple 2D images captured from different angles, paired with masked images and ground-truth 3D models for reconstruction. Extensive preprocessing, including manual cleaning, alignment, manually created annotation files as well as handpicked images, was performed to create a structured and standardized dataset for training, validation and testing. We also have to create a image matching file which we have used in our custom dataset and data loader class. For the training dataset we have used the batch size of 7 for the training loader class and for the validation loader we have used the batch size of around 5 with shuffling = "On" and number of workers=2.

We have 70% training, 20% Validation and 10% of training data in dataset and all the data was fed into batches of 5 images along with their masks after which the images in one row were concatenate to input the data in the model. So there were multiple rows of images with each row corresponding to there 5 images.

---

6. Methods

Our methodology includes 3 type of algorithms, Nerf based architecture, 3DR2N2 based architecture and Normal CNN based classifier:

1. **Feature Extraction:** CNNs extract spatial features from 2D images [5].
2. **Voxel Grid Generation:** Intermediate 3D representations are generated, allowing for structured learning.

3. **Model Architecture:** Advanced algorithms such as 3D-R2N2 and NERF based architecture are integrated to refine the reconstruction process.
4. **Data Handling:** Batch processing using a custom Data loader and Dataset framework ensures efficient training and scalability.

---

7. Experiments

Initial experiments validated the feasibility of the pipeline:

- **Data Preprocessing:** Successful standardization of the Pix3D dataset for model compatibility with manual filtering and augmentation techniques such as image flipping and filtering. For Data preprocessing we have use one of the widely used techniques available in computer vision, which is background removal techniques, for this we have used for this we have used one of the inbuild functions available in cv2 such as cv2.threshold, and cv2.bitwise_and.
- **Training:** Iterative refinement of the model architecture improved early performance.

  We have used Adam and L2 regularization loss to make the model performance better. We have trained our algorithm on 90 epochs which approximately took 24 hours to get the perfect parameters.
- **Testing:** For testing purpose, we have used unseen dataset for testing purpose and we have used the parameters of our saved models to predict the labels alongside the 3d object of the different images alongside prediction of different labels.

---

8. Conclusion

Our project has achieved significant milestones in dataset preparation, model design, initial training, and testing. Future work includes further optimization, testing, and comparative analysis with existing approaches to measure the system's effectiveness. This research aims to contribute to advancements in 3D reconstruction technologies, driving applications in multiple fields.

For future improvements we will refine the model, so it could handle more complex models, and we are trying to enhance the resolution of the output 3D object. We will also take some more computation power to train the model of larger epoch such as 1000 or 2000 as it may require ton amount of memory.

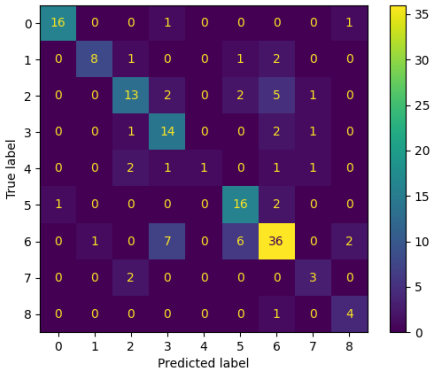
Supplementary Material

- Libraries Used- Matplotlib, NumPy, Pytorch, Pandas, OS, torch vision, torch audio, pillow, cv2, torch, python-dateutil, OpenCV-python, contourpy, pillow, scikit-learn, piglet.
- **Source Code:** It is available on our GitHub Repository.

  GITHUB REPOSITORY LINK-LINK
- Model Testing Results:

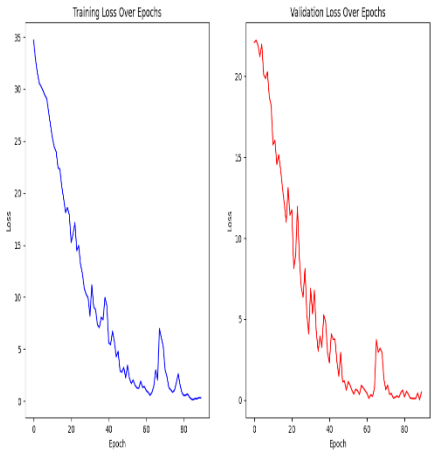
```
Using CPU
Testing Results:
Accuracy: 0.7025316455696202
Precision: 0.7287715805842658
Recall: 0.7025316455696202

Process finished with exit code 0
```

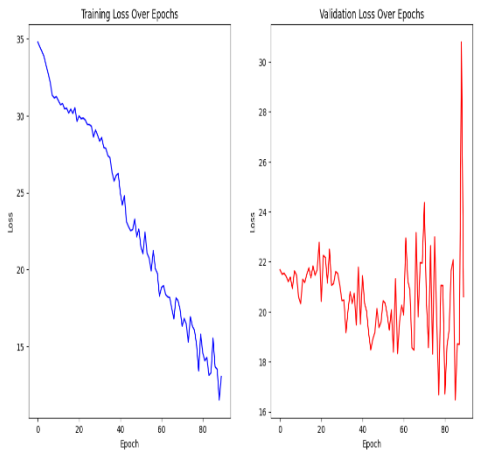
Testing Accuracy, Precision, Recall



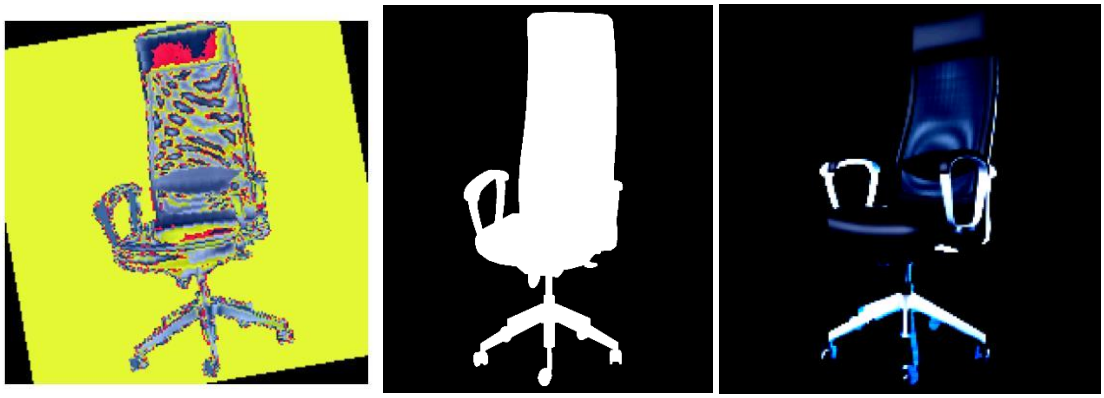
Testing Confusion Matrix

- Loss On Validation and Training Set.





- **Pre-processed Images:**

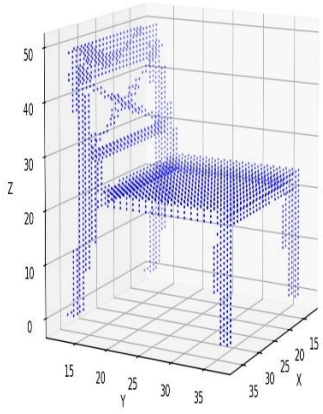


NORMALIZED IMAGE MASK OF THE IMAGE IMAGE WITH BG REMOVAL

- **Dataset Samples Along Side Predicted 3D Voxel Grid:**

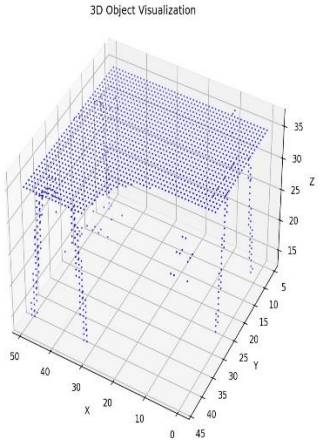
  Input images –

  1. Chair



INPUT IMAGE OUTPUT IMAGE i.e. Voxel Grid

  2. DESK



INPUT IMAGE OUTPUT IMAGE i.e. Voxel Grid

References

1. C. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese, "3D-R2N2: A Unified Approach for Single and Multi-view 3D Object Reconstruction," *European Conference on Computer Vision (ECCV)*, 2016.
2. Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. Bronstein, and J. M. Solomon, "Pixel2Mesh: Generating 3D Mesh Models from Single RGB Images," *European Conference on Computer Vision (ECCV)*, 2018.
3. L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger, "Occupancy Networks: Learning 3D Reconstruction in Function Space," *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
4. X. Sun, X. Wu, Z. Wang, X. Zhou, and S. Deng, "Pix3D: Dataset and Methods for Single-Image 3D Shape Modelling," *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
5. K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.