3D Shape Reconstruction from Images

David Fang
MIT
fangd@mit.edu

Marie Diane Fadel
MIT
mdfadel@mit.edu

Abstract

The 3D shape reconstruction from 2D images problem is an extremely challenging task for computer vision because of its ill-posed nature. However, it is also a crucial task that has been used to the advantage of many fields including medical imaging, autonomous driving and object detection. Many recent advances in both the quality of datasets and models have enabled rapid progress for this problem. We explore tackling this challenge using encoder-decoder architectures, with single-image and multi-image approaches. We use the ShapeNet Core55 dataset from the 2017 ICCV shape reconstruction challenge to train and evaluate our models. Our best proposed model achieves an average intersection-over-union of 0.41 and a mostly smooth latent space.

1. Introduction

For humans, 3D shape reconstruction seems like a simple task due to the wealth and breadth of knowledge accumulated over life and pattern matching skills. When one sees a side view of a plane or a car, there are very few objects other than a plane or car (basically none) that could have that same view. So while it is impossible to get the exact details correct for anyone, the approximate shape can still be estimated based on our prior knowledge.

However, for a computer to deduce the same thing is very difficult. Such a task would require multiple processing steps: recognizing the important features of the view (such as the wing or wheels), piecing them together to classify it as part of a coherent object, and then remembering what the 3D shape of a object should be. Each of these tasks individually is already difficult, so distilling them into a single model is a challenging problem. Some work has been done in the field but almost all technique either require additional parameters such as camera parameters or depth maps, or use multiple images. Our goal was to try and tackle this problem using the least parameters possible: a single 2D image. This would be revolutionary in fields where data is scarce or costly like medical imaging and space exploration.

We propose models based on encoder-decoder architectures for supervised single image and multi-image to 3D reconstruction without camera parameters. We also demonstrate a somewhat smooth latent space learned in an unsupervised fashion.

2. Related Work

As stated previously, 3D reconstruction given 2D images plays an important role in medical imaging, scene representations, cultural artifact reconstruction, and automation. Previous works have used many techniques to recover the 3D shape of an object, mainly based on the type of data available. The two main techniques related to the work we will present in this paper are the single still image approach and the multi-image 3D reconstruction approach.

2.1. Single Still Image Approach

Using a single 2D image to recover the 3D shape presents more challenges compared to multiple images. However, many approaches have been developed to recover the 3D shape of an object using a single 2D image. One of these approaches includes using a Perspective Transformer network as proposed by Yan et. al [5]. This method ignores color of the image and involves running the image input though an encoder-decoder network that consists of a 2D convolutional encoder, a 3D up-convolutional decoder and a perspective transformer networks.

Another way to use a single image to reconstruct the 3D shape of an object is using its depth map as proposed by Saxena et. al [4]. This was done by predicting the depth map as a function of the image through a supervised machine learning model that uses a hierarchical, multiscale Markov Random Field (MRF).

Neural radiance fields (NeRFs) have also made progress in the domain of single image 3D reconstruction. While NeRFs traditionally require many input views, Yu et. al [6] demonstrates a NeRF-based network that can predict neural representations with few or only one input images. However, such a method requires well-calibrated camera parameters.

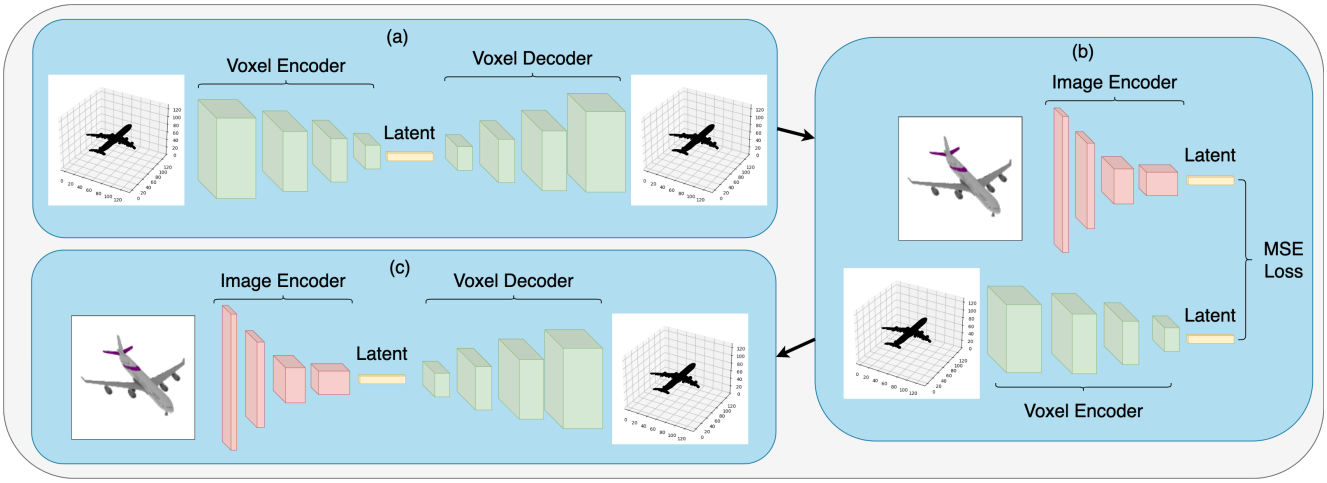

Figure 1. Single-image reconstruction methods for the encoder-decoder model (c) and encoder-encoder-decoder model (a, b, c).

2.2. Multi-image 3D Reconstruction Approach

Structure from Motion (SfM) is one of the classical approaches for 3D reconstruction, which requires capturing the subject from multiple views and then processing them with reconstruction algorithms. However, this method is time consuming and inefficient, which has pushed researchers to find alternative 3D reconstruction methods using deep learning approaches. A lot of such algorithms have been developed but transformer-based encoder-decoders have been found to outperform prior approaches. Peng et al. [3] introduces a transformer-based encoder-decoder for 3D voxel reconstructions that can take sequences of images as input.

3. Proposed Methods

3.1. Encoder-Decoder Model

We propose using an encoder-decoder architecture for this task. The encoder will use single 2D images as input and output embeddings and the decoder will use the embeddings as input and output 3D voxels (Fig. 1c). Intuitively, we believe that the encoder will be able to perform the image-to-embedding classification and the embedding-to-voxel decoder will perform the 3D shape reconstruction of the object from the embeddings. We use a ResNet18 convolutional neural network backbone for the encoder, and 3D up-convolution operators for the decoder. We then pass the voxel output through a sigmoid function to recover probabilities.

3.2. Encoder-Encoder-Decoder Model

The encoder-decoder model may have difficulties classifying the images to objects. Thus, we also propose an encoder-encoder-decoder model, inspired by Girdhar et al. [2]. We retain the same image-to-embedding encoder (with the same ResNet backbone) and embedding-to-voxel decoder, but also add an encoder that takes the voxels as input and outputs embeddings. We can then have a 3-stage training process:

1. Train the voxel-to-embedding encoder with the embedding-to-voxel decoder (Fig. 1a). This is simply an autoencoder.

2. Freeze the parameters of the voxel-to-embedding encoder, and train the image-to-embedding encoder to match the output embeddings using an MSE loss (Fig. 1b).

3. Finetune the image-to-embedding encoder with the embedding-to-voxel decoder (Fig. 1c). This step is the same as the encoder-decoder model.

Thus, in the training process, we force the classification between the images and the voxel representation. We can think of this as ”pretraining” the encoder-decoder model. We include the last stage to further fine-tune the encoder-decoder model. During testing, we can simply remove the voxel-to-embedding encoder and rely just on the image-to-embedding encoder.

3.3. Multi-Image Encoder-Decoder Model

Instead of single images, we can also try recovering the 3D shape from a group or sequence of images, which are all at different viewing angles. We expect that each image can contribute a different smaller part of the 3D structure, so that the combination of all the images will recover most of the overall 3D structure.

We propose an encoder-decoder structure that takes in a sequence of images and outputs a sequence of voxels (Fig.

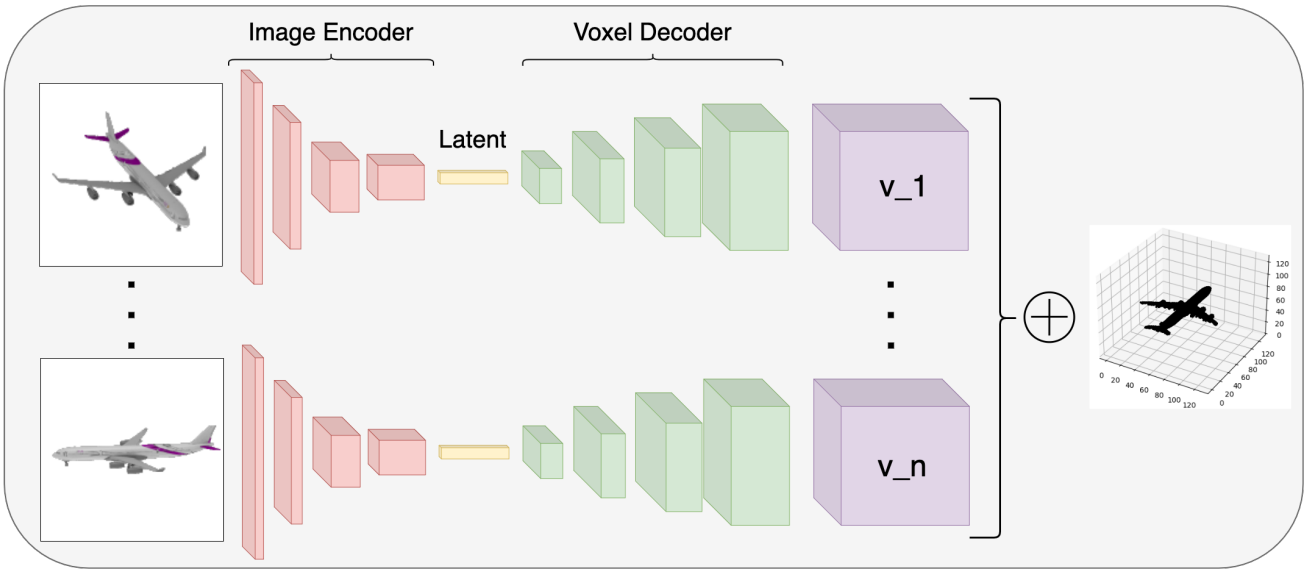

Figure 2. Our multi-image 3D reconstruction method using a variable sequence of images. A sigmoid is applied to the final output voxels to return probabilities.

2). We then take the sum of these voxels and pass them through a sigmoid function to recover probabilities. The architecture of the image encoder remains the same as the other models. The voxel decoder architecture is mostly the same, but without the last sigmoid activation, as that is applied after the voxels are combined.

4. Experiments

4.1. Dataset

We train and evaluate our methods on the ShapeNet Core55 dataset from the 2017 ICCV shape reconstruction challenge [1]. This dataset consists of 48,600 3D models over 55 categories in the ShapeNet dataset. The 3D models are represented by 256^3 voxels as binary values, where 1 indicates occupied space and 0 indicates free space. There are 12 synthesized images at different viewing angles for each model. We use a 70%/10%/20% training/validation/testing split. Due to GPU memory constraints and time limitations, we downsize the voxel representations to 32^3 using trilinear interpolation for training and evaluation.

4.2. Implementation Details

We implemented the encoder-decoder (Sec 3.1), encoder-encoder-decoder (Sec 3.2), and multi-image encoder-decoder (Sec 3.3) models in PyTorch. The decoder was implemented using 4 layers of 3D transposed convolution operators with kernel size 7 and parametric ReLU activations. The ResNet-18 backbone is pretrained on ImageNet-1K and input images are normalized using ImageNet statistics. We use a latent embedding vector size of 64. We train each model separately using binary cross-entropy loss with the Adam optimizer. As the outputs of the model are probabilistic voxels, we convert them to binary voxels using a threshold of 50%. Each model took on the order of 2 days on a Nvidia V100 GPU to train.

4.3. Quantitative Results

We quantitatively evaluate our 3D reconstruction results using the intersection-over-union (IoU) metric on the predicted and ground truth voxels.

$$\text{IoU}(\text{voxel}_{pred}, \text{voxel}_{GT}) = \frac{\text{voxel}_{pred} \cap \text{voxel}_{GT}}{\text{voxel}_{pred} \cup \text{voxel}_{GT}}$$

We report the average IoU over the test split of the ShapeNet dataset for each of our models (Table 1). We consider our encoder-decoder model as the baseline as it is the simplest model to tackle this problem.

Model	Average IoU
Encoder-Decoder (Baseline)	0.33
Voxel Encoder-Decoder	0.75
Encoder-Encoder-Decoder	0.41
Multi-Image Encoder-Decoder	0.24

Table 1. Average IoU for each model (higher is better).

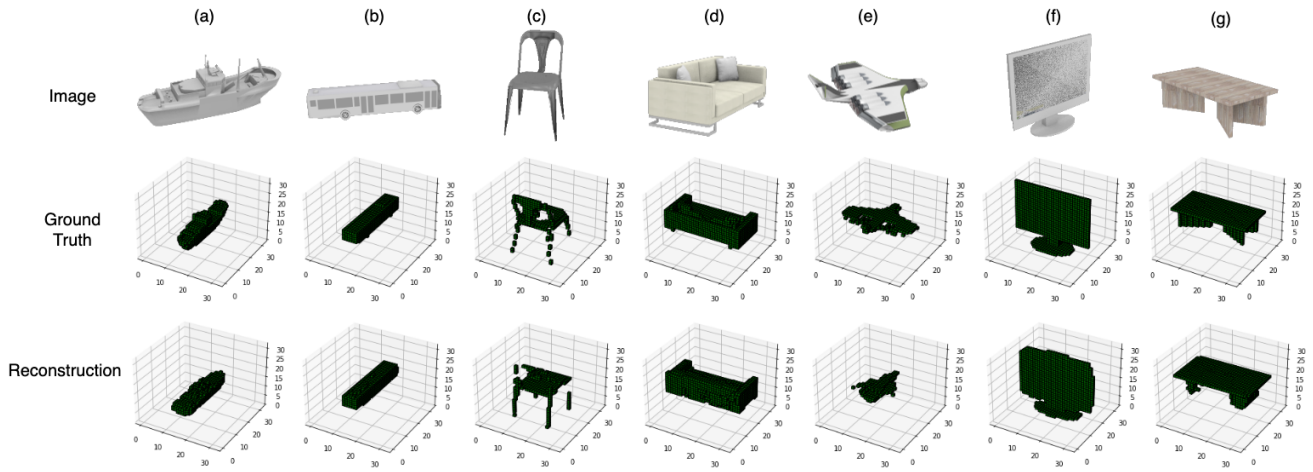

Figure 3. Single-image reconstruction results for assorted objects. Some reconstructions look almost the same as the ground truth (b), while others are significantly different (e).

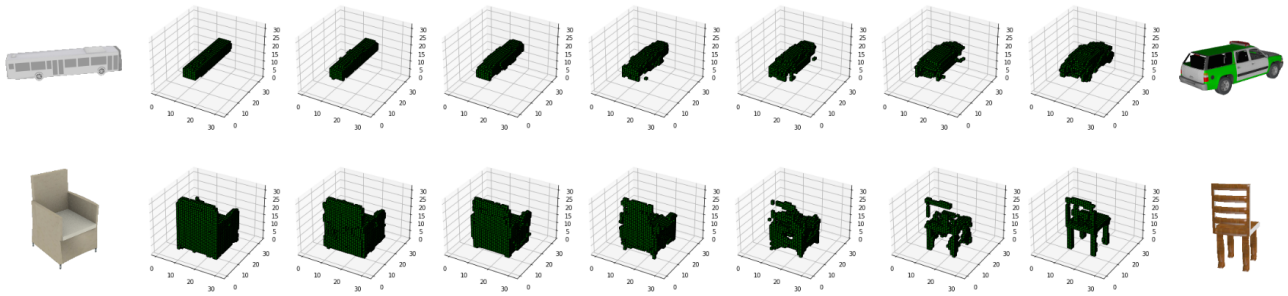

Figure 4. Interpolation between latent embeddings.

4.4. Qualitative Results

We also show qualitative reconstructions of a few objects (Fig 3). The shown reconstructions are predicted with our best model, the single image encoder-encoder-decoder model, in a 32^3 voxel grid with a threshold of 50%. We include both successful and unsuccessful reconstructions. Additionally, we show example interpolations between embeddings for similar objects, such as a bus to a car and a sofa chair to a chair (Fig. 4).

5. Discussion

As expected, the encoder-encoder-decoder model performs better than the baseline encoder-decoder model. This is probably due to the fact that the staged training process gives the model a good initialization to further finetune. We found it surprising that the voxel encoder-decoder only achieved an average IoU of 0.75, when we expected it to be much closer to 1.0.

The biggest surprise was the multi-image encoder-decoder, which performed worse than the baseline. We hypothesize that because the 3D reconstructions have a certain pose orientation, the multi-image approach makes it difficult to capture the orientation because it contains different angles.

In our reconstruction examples, we find that our encoder-encoder-decoder model can often predict the approximate shape, but has difficulty resolving finer details. For example, the bus reconstruction looks quite accurate because the model can capture the simple geometric box shape (Fig 3b), while the plane is more complex and our model only captures the middle body (Fig 3e). The model also struggles with finer details such as with the clean corners and edges of monitor, but can still approximate the overall shape (Fig 3f). Additionally, we find that similar objects such as different types of cars will often look the same because they will be missing some of the details that differentiate the different cars. This matches our intuition that our encoder is classifying the images to an broader category (such as cars) and using that for 3D reconstruction.

As our model is using a single image approach, we find that sometimes the angle of the image does matter. For ex-

ample, if parts of the object partially occlude each other in certain images, then the model has difficulty capturing that part when given those images. For example, the base of the table in Fig 3g is partially occluded by the table surface and also poorly reconstructed.

We also find the our latent space can be somewhat smoothly interpolated. For objects in similar categories, the 3D reconstructions have mostly smooth transitions as the embeddings vector changes (Fig. 4). However, we find that for objects that are too dissimilar, the voxels will fully disappear in the middle, as if it is resetting the drawing board.

5.1. Limitations

The most obvious limitation in our models is the latent embedding size of 64. Ideally, our voxel autoencoder would have an average IoU of close to 1.0, which would mean that it can compress and decode almost all of the 3D information fairly well. However, our autoencoder only achieves an IoU of 0.75. Having an embedding size of only 64 might be restricting the amount of information the models can compress and decode. If we had enough time, we would also ideally experiment with larger latent sizes, such as 128, 256, or even 512.

Another limitation is the depth of our network. Due to time constraints, we only used a ResNet-18 backbone, which performs worse than larger ResNet models on most benchmarks. However, we chose ResNet-18 because it trains faster. Additionally, our decoder only has 4 layers for the same reasons. If we had more time, we could experiment with a larger backbone, such as ResNet-50, and add more layers to our decoder.

6. Conclusion

We propose two models for image to 3D voxel reconstruction. Our single image encoder-encoder-decoder model outperforms the baseline, while our multi-image encoder decoder did worse. We find that the single image model sufficiently captures large scale 3D structure across categories of objects. However, it fails to reconstruct details on a finer scale for each unique object. We also demonstrate that the model learns a mostly smooth interpolatable latent space without supervision as well during the training process.

More complex models could be explored in the future for better performance. The multi-image approach could be extended in the future to include some sort of pose consistency or orientation. Additionally, another model like a recurrent neural network that uses a sequence of images could help maintain the pose while adding on finer details.

Another interesting line of future work could be camera pose estimation combined with neural radiance fields. NeRFs currently require camera poses to successfully model an object. However, only given a set of images, it could be possible to construct a neural radiance field and also predict the camera poses of the original images.

7. Contributions

Marie Diane worked on the abstract, introduction, related works, and conclusion. David worked on the proposed methods, experiments, and discussion. Both of us helped each other with the individual sections as well. Marie Diane presented alone due to David's surgery, so he focused more on the technical experiments.

References

[1] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015. 3

[2] Rohit Girdhar, David F. Fouhey, Mikel Rodriguez, and Abhinav Gupta. Learning a predictable and generative vector representation for objects. In *ECCV*, 2016. 2

[3] Kebin Peng, Rifatul Islam, John Quarles, and Kevin Desai. Tmvnet : Using transformers for multi-view voxel-based 3d reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 221–229, 2022. 2

[4] Ashutosh Saxena, Sung H. Chung, and Andrew Y. Ng. 3-d depth reconstruction from a single still image, 2008. 1

[5] Xinchen Yan, Jimei Yang, Ersin Yumer, Yijie Guo, and Honglak Lee. Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. In *NIPS*, 2016. 1

[6] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *CVPR*, 2021. 1