Learning Single-View & Multiple-View 3D Object Reconstruction

Ronak Dedhiya
Indian Institute of Science
Bangalore, India
ronakdedhiya@iisc.ac.in

Abstract

Understanding the 3D world is a fundamental problem in computer vision. However, learning a good representation of 3D objects is still an open problem due to the high dimensionality of the data and many factors of variation involved. In our project, we studied two diverse methods of learning 3d structure, one is unsupervised method which formulates 3D and 2D interaction using novel projection loss defined by the perspective transformation and another supervised method which uses well-designed encoder-decoder to generate a coarse 3D volume from each input image. A context- aware fusion module is used to adaptively select high-quality reconstructions for each part from different coarse 3D volume to obtain a fused 3D Volume. Finally, a refiner further refines the fused 3D volume to generate the final output. We also experiment with different losses i.e, L1, L2 and BCE loss for supervised learning problem.

1. Introduction

3D reconstruction is an important problem in robotics, CAD, virtual reality and augmented reality. Understanding the 3D world is at the heart of successful computer vision applications in robotics, rendering and modeling. It is especially important to solve this problem using the most convenient visual sensory data: 2D images. we looked for lot of work along the lines of constructing 3d structure using 2D Images both supervised and unsupervised method. Unsupervised learning from multiple Images from [2] uses additional albedo, shape and texture loss. Learning Imlicit fields for generative shape modelling from [1] replaces the mlp decoder to learn better shape representation. In Occupancy Networks [4], 3D output can be viewed as continuous boundary instead of voxel or point representation and can be used in supervised setting.

There were generative approached to build the 3D shape, in [6] 3D Gan is used for learning 3D shape, apart from the Images, it used class labels for conditional generation of 3D voxels. Nguyen et.al [5] proposes Holo Gan which is complete unsupervised Learning of 3D representations from natural images. Liu [3] proposes method which learn to infer implicit surfaces without 3D supervision. Another novel unsupervised approach from Yan et al. [8] predict 3D shape from single view without using the groundtruth 3D Volumetric data for training by using a 2D silhoutte loss function based on perspective transformation. We also observe another novel supervised approach proposed by Xie et al [7] which uses a well-designed encoder-decoder, it generates a coarse 3D volume from each input image. Then, a context-aware fusion module is introduced to adaptively select high-quality reconstructions for each part (e.g., table legs) from different coarse 3D volumes to obtain a fused 3D volume. Finally, a refiner further refines the fused 3D volume to generate the final output.

2. Unsupervised Method

2.1. Problem Formulation

From the perspective of a learning agent (e.g., neural network), a natural way to understand one 3D object X is from its 2D views by transformations. By moving around the 3D object, the agent should be able to recognize its unique features and eventually build a 3D mental model of it as illustrated in Figure 1(a). Assume that $I_{(k)}$ is the 2D image from the k-th viewpoint $\alpha_{(k)}$ by projection $I_{(k)} = P(X; \alpha_{(k)})$, or rendering in graphics. An object X in a certain scene is the entanglement of shape, color and texture (its intrinsic properties) and the image $I_{(k)}$ is the further entanglement with viewpoint and illumination (extrinsic parameters). The general goal of understanding 3D objects can be viewed as disentangling intrinsic properties and extrinsic parameters from a single image.

2.2. Methodology

We use the volumetric representation of 3d shape V where each voxel V_i is a binary unit. In other words, the voxel equals to one, i.e., V_i = 1, if the i-th voxel space is occupied by the shape; otherwise V_i = 0. Assuming the 2D silhouette $S_{(k)}$ is obtained from the k-th image $I_{(k)}$, we can

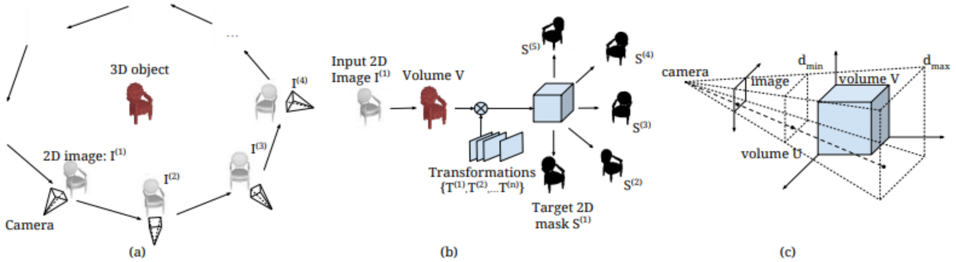


Figure 1. (a) Understanding 3D object from learning agent's perspective; (b) Single-view 3D volume reconstruction with perspective transformation. (c) Illustration of perspective projection. The minimum and maximum disparity in the camera frame are denoted as dmin and dmax.

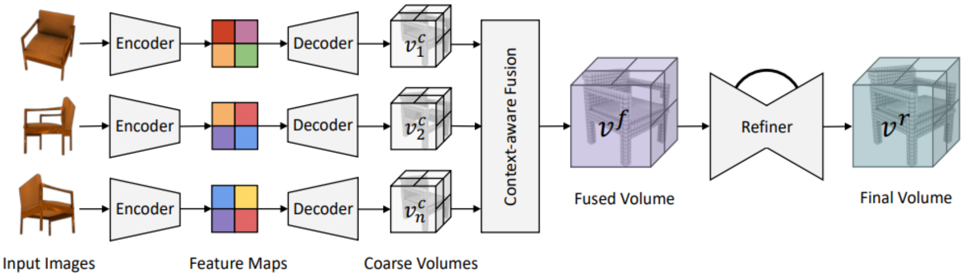


Figure 2. : An overview of the proposed Pix2Vox. The network recovers the shape of 3D objects from arbitrary (uncalibrated) single or multiple images. The reconstruction results can be refined when more input images are available. Note that the weights of the encoder and decoder are shared among all views

specify the 3D-2D projection $S_{(k)} = P(V; \alpha_{(k)})$. A 2D silhouette $S_{(j)}$ projected from the generated volume V under certain camera viewpoint $\alpha_{(j)}$ should match the ground truth 2D silhouette $S_{(j)}$ from image observations. In other words, if all the generated silhouettes $S_{(j)}$ match well with their corresponding ground truth silhouettes $S_{(j)}$ for all j's, then we hypothesize that the generated volume V should be as good as one instance of visual hull equivalent class of the ground truth volume V.

$$L_{proj}(I_{(k)}) = \sum_{j=1}^{n} L^{j}_{proj}(I_{(k)}; S_{(j}, \alpha_{(j)}) \quad (1)$$

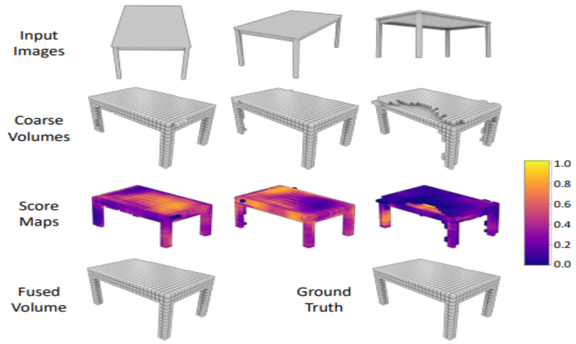
$$L_{proj}(I_{(k)}) = \frac{1}{n}\sum_{j=1}^{n} ||P(f(I_{(k)}); \alpha_j) - S_{(j)}||_2^2 \quad (2)$$

2.3. Perspective Transformer Networks

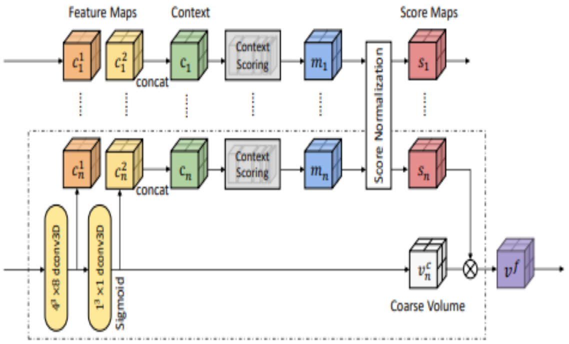
As defined previously, 2D silhouette $S_{(k)}$ is obtained via perspective projection given input 3D volume V and specific camera viewpoint $\alpha_{(k)}$. In this work, we implement the perspective projection 1 with a 4-by-4 transformation matrix θ_{44}, where K is camera calibration matrix and (R, t) is extrinsic parameters.

$$\theta_{4x4} = \begin{bmatrix} K & 0 \\ O^T & 1 \end{bmatrix} \begin{bmatrix} R & t \\ O^T & 1 \end{bmatrix} \quad (3)$$

For each point $p_i^s = (x_i^s, y_i^s, z_i^s, 1)$ in 3D world frame, we compute the corresponding point $p_i^t = (x_i^t, y_i^t, 1, d_i^t)$ in camera frame (plus disparity d_i^t) using the inverse of perspective transformation matrix $p_i^s \theta_{44}^1 p_i^t$. A novel perspective transformation operator that performs dense sampling from input volume (in 3D world frame) to output volume (in camera frame). To obtain the 2D silhouettes from 3D volume, we propose a simple approach using max operator that flattens the 3D spatial output across disparity dimension. This operation can be treated as an approximation to the ray-tracing algorithm. In the experiment, we assume that transformation matrix is always given as input, parametrized by the viewpoint α. Again, the 3D point (x_i^s, y_i^s, z_i^s) in input volume V R^{HWD} and corresponding point (x_i^t, y_i^t, d_i^t) in output volume U R^{H0W0D0} is linked by perspective transformation matrix θ_{44}. Here, (W, H, D) and (W0, H0, D0) are the width, height and depth of input and output volume,respectively.



(a) : Visualization of the score maps in the context aware fusion module. The context-aware fusion module generates higher scores for high-quality reconstructions, which can eliminate the effect of the missing or wrongly recovered parts.



(b) : An overview of the context-aware fusion module. It aims to select high-quality reconstructions for each part to construct the final results. The objects in the bounding box describe the procedure score calculation for a coarse volume v. The other scores are calculated according to the same procedure. Note that the weights of the context scoring network are shared among different views.

3. Supervised Method

3.1. Overview

The proposed Pix2Vox aims to reconstruct the 3D shape of an object from either single or multiple RGB images. The key components of Pix2Vox are shown in Figure 2. First, the encoder produces feature maps from input images. Second, the decoder takes each feature map as input and generates a coarse 3D volume correspondingly. Third, single or multiple 3D volumes are forwarded to the context aware fusion module, which adaptively selects high-quality reconstructions for each part from coarse 3D volumes to obtain a fused 3D volume. Finally, the refiner with skip connections further refines the fused 3D volume to generate the final reconstruction result.

3.2. Network Architecture

3.2.1 Encoder

The encoder is to compute a set of features for the decoder to recover the 3D shape of the object. The first nine convolutional layers, along with the corresponding batch normalization layers and ReLU activations of a VGG16 pretrained on ImageNet, are used to extract a 512×28×28 feature tensor from a 224 × 224 × 3 image. 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.

3.2.2 Decoder

The decoder is responsible for transforming information of 2D feature maps into 3D volumes.There are five 3D transposed convolutional layers. r. Each transposed convolutional layer is followed by a batch normalization layer and a ReLU activation except for the last layer followed by a sigmoid function.

3.2.3 Context-aware Fusion

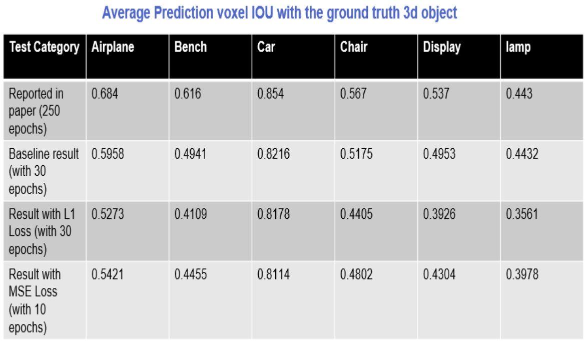
As shown in Figure 3b, given coarse 3D volumes and the corresponding context, the context-aware fusion module generates a score map for each coarse volume and then fuses them into one volume by the weighted summation of all coarse volumes according to their score maps. The spatial information of voxels is preserved in the context-aware fusion module, and thus Pix2Vox can utilize multi-view information to recover the structure of an object better. Specifically, the context-aware fusion module generates the context cr of the r-th coarse volume v c_r by concatenating the output of the last two layers in the decoder. Then, the context scoring network generates a score m_r for the context of the r-th coarse voxel. The learned score m_r for context c_r are normalized across all learnt scores. We choose softmax as the normalization function.

3.2.4 Refiner

The refiner can be seen as a residual network, which aims to correct wrongly recovered parts of a 3D volume. It follows the idea of a 3D encoder-decoder with the U-net connections.

3.3. Loss Function

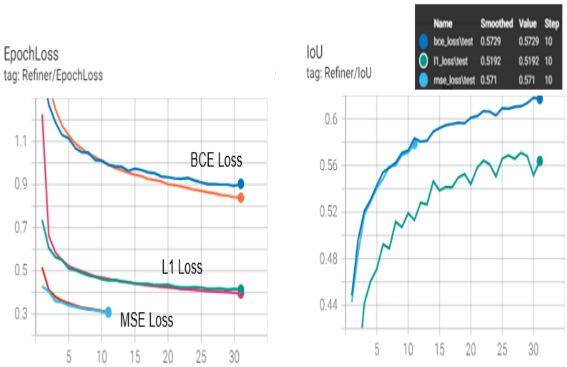
The loss function of the network is defined as the mean value of the voxel-wise binary cross entropies between the reconstructed object and the ground truth.



Average Prediction voxel IOU with the ground truth 3d object

Test Category	Airplane	Bench	Car	Chair	Display	lamp
Reported in paper (250 epochs)	0.684	0.616	0.854	0.567	0.537	0.443
Baseline result (with 30 epochs)	0.5958	0.4941	0.8216	0.5175	0.4953	0.4432
Result with L1 Loss (with 30 epochs)	0.5273	0.4109	0.8178	0.4405	0.3926	0.3561
Result with MSE Loss (with 10 epochs)	0.5421	0.4455	0.8114	0.4802	0.4304	0.3978

(a) : Comparison of baseline results with different types of loss function



(b) : Training loss curve and IOU value on test set while training for each of the losses

4. Experiments

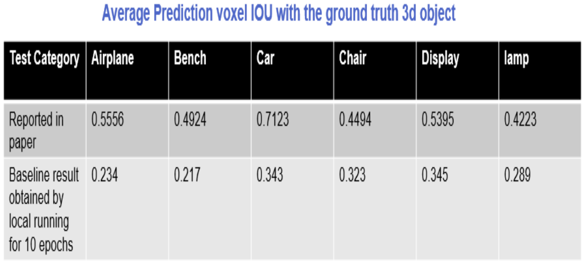
4.1. Dataset and Metrics

Dataset A subset of the ShapeNet dataset consists of 50,000 models and 13 major categories was used in the experimentation. The 13 categories are airplane, bench, cabinet, car, chair, display, lamp, speaker, rifle, sofa, table, telephone, and watercraft. ShapeNet rendered Images are obtained from ftp://cs.stanford.edu/cs/cvgl/ShapeNetRendering.tgz (12.8GB) and ShapeNet voxelized models are obtained from ftp://cs.stanford.edu/cs/cvgl/ShapeNetVox32.tgz (100mb)

Evaluation Metrics To evaluate the quality of the output from the proposed methods, we binarize the probabilities at a fixed threshold of 0.3 and use intersection over union (IoU) as the similarity measure. Higher IoU values indicate better reconstruction results.

4.2. Results with unsupervised approach

The unsupervised implementation of PTN is available in tensorflow 1.x and torch lua which are not compatible to be run directly on the current gpu system. The code implementation uses prespective transformer layer which takes the camera intrinsic and extrinsic parameters and generate the projections from the given voxelized decoder output. These perspective transformer layer implementation is unavailable in latest pytorch tensorflow 2 code. Converting the code is even difficult. Another option is to use the open source rendered, which takes in camera parameters and mesh inputs to generate the projected output. However, this rendered library works with specific cuda version and again difficult to run on google colab or kaggle run. We managed to get it run on the local machine installing tensorflow 1.x and thus ran the experiment for only 10 epochs with the results show in Figure 5. We predict the voxel given the input image to encoder and calculate the iou with the groundtruth voxel, the results is averaged over each category and displayed in the 5



Average Prediction voxel IOU with the ground truth 3d object

Test Category	Airplane	Bench	Car	Chair	Display	lamp
Reported in paper	0.5556	0.4924	0.7123	0.4494	0.5395	0.4223
Baseline result obtained by local running for 10 epochs	0.234	0.217	0.343	0.323	0.345	0.289

Figure 5. : Comparison of baseline results with unsupervised approach

4.3. Results with supervised approach

The pixe2vox is the supervised approach used to construct the 3d objects from single or multi-view images. We experiment with 3 different loss function: a) BCE Loss, b) MSE Loss c) L1 Loss. We ran each of them for 30 epochs and obtained the results shown in figure 4a and 4b. We see that MSE loss matched up with BCE loss in terms of IOU loss obtained while we get less iou value with the L1 loss, which might be bcoz it is turning to make the voxel directly to zero instead of slowing pushing towards 0. Table shows the similar result comparable to the baseline result with just 30 epochs.

5. Conculsion

We studied supervised and unsupervised approach for reconstructing the 3D objects from single or multi view images. We observe that training unsupervised models are hard, requires expertise with the camera, view angle and projection geometry based knowledge and do it correctly. While supervised approach is easy to approach and multiple experiments can be tried easily. We also observe that pix2vox architecture due to context aware fusion, converges faster towards learning the 3D shapes. We will continue our experimentation with both of the approaches.

References

[1] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5939–5948, 2019. 1

[2] Long-Nhat Ho, Anh Tuan Tran, Quynh Phung, and Minh Hoai. Toward realistic single-view 3d object reconstruction with unsupervised learning from multiple images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12600–12610, 2021. 1

[3] Shichen Liu, Shunsuke Saito, Weikai Chen, and Hao Li. Learning to infer implicit surfaces without 3d supervision. *Advances in Neural Information Processing Systems*, 32, 2019. 1

[4] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4460–4470, 2019. 1

[5] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. Hologan: Unsupervised learning of 3d representations from natural images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7588–7597, 2019. 1

[6] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. *Advances in neural information processing systems*, 29, 2016. 1

[7] Haozhe Xie, Hongxun Yao, Xiaoshuai Sun, Shangchen Zhou, and Shengping Zhang. Pix2vox: Context-aware 3d reconstruction from single and multi-view images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2690–2698, 2019. 1

[8] Xinchen Yan, Jimei Yang, Ersin Yumer, Yijie Guo, and Honglak Lee. Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. *Advances in neural information processing systems*, 29, 2016. 1