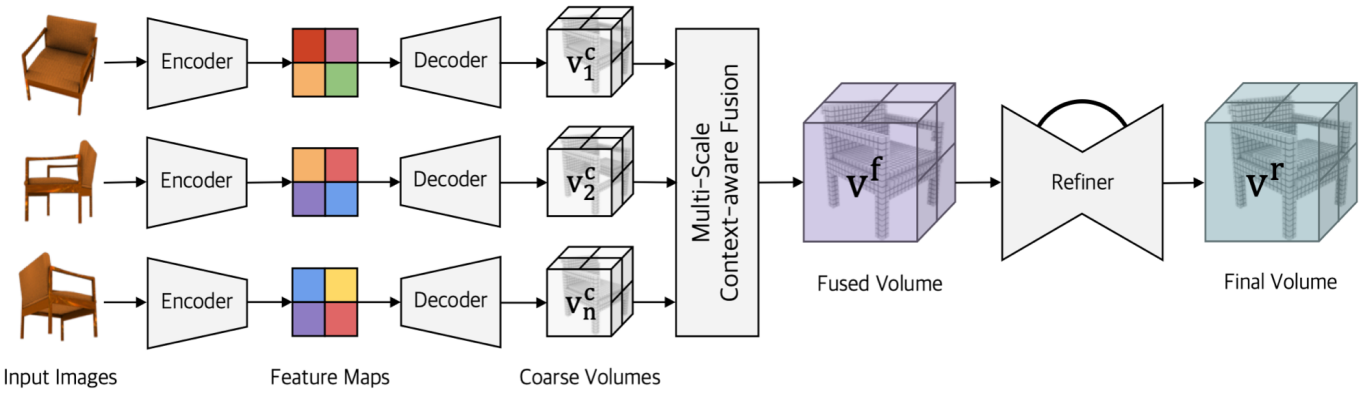
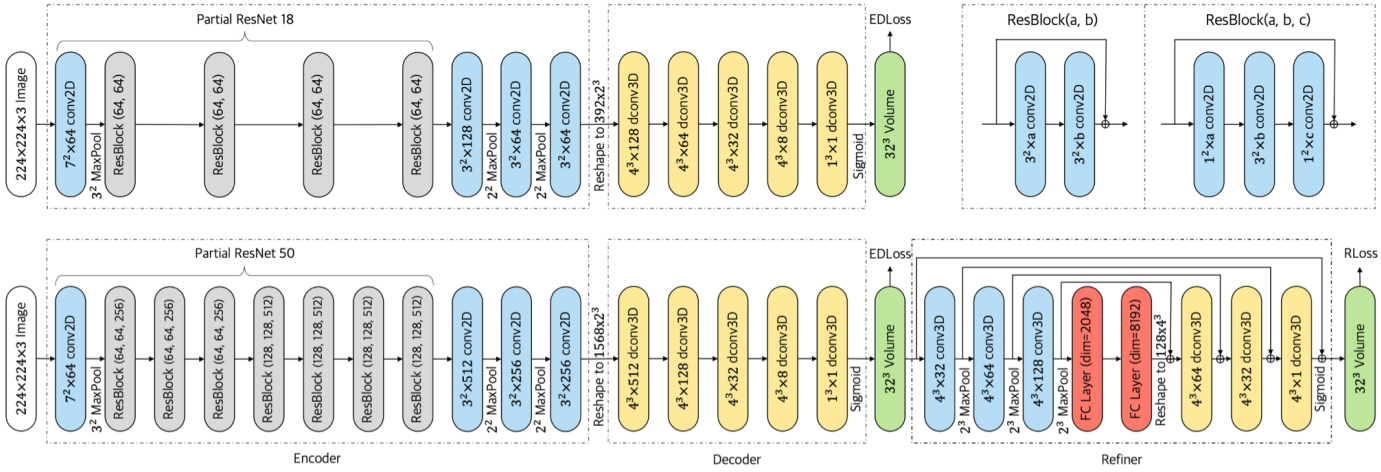
Pix2Vox++: Multi-scale Context-aware 3D Object Reconstruction from Single and Multiple Images

Report



Dataset used: Shapenet, Pix3D, Things3D
Pytorch Code: https://gitlab.com/hzxie/Pix2Vox
Paper: https://arxiv.org/pdf/2006.12250.pdf



Key Points to note:

Encoder:

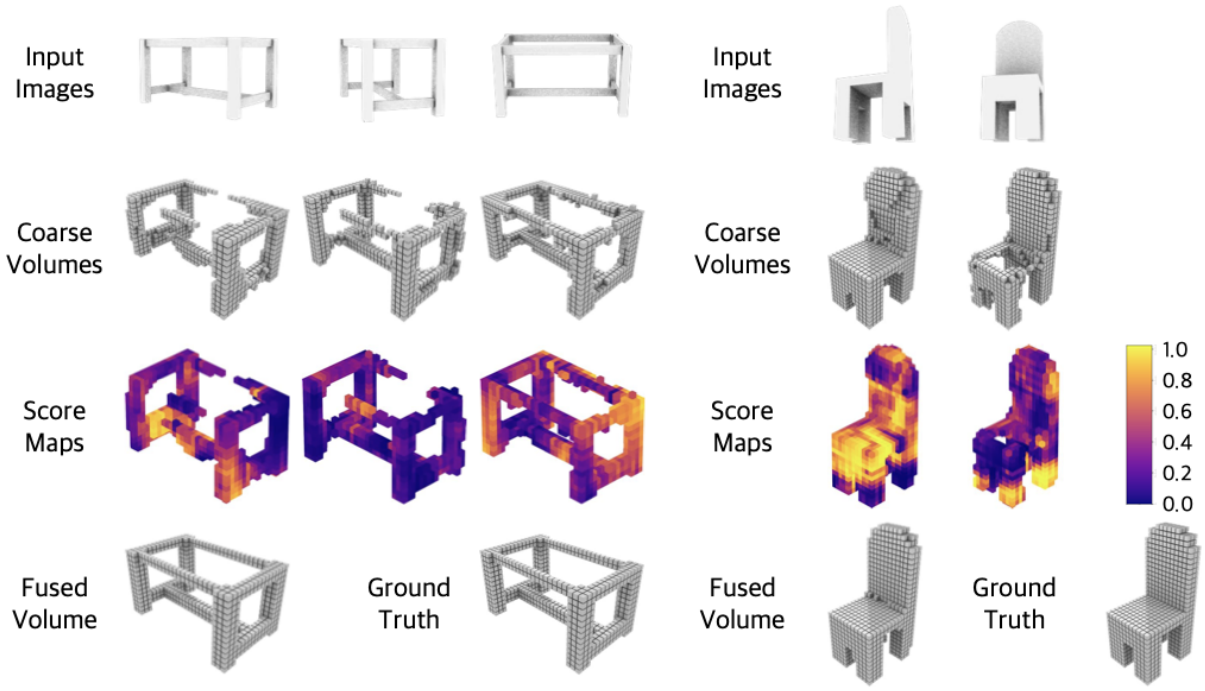
- The part of the encoder is made up of Resnet 18 or 50 depending on the model Pix2Vox++/F or Pix2Vox++/A

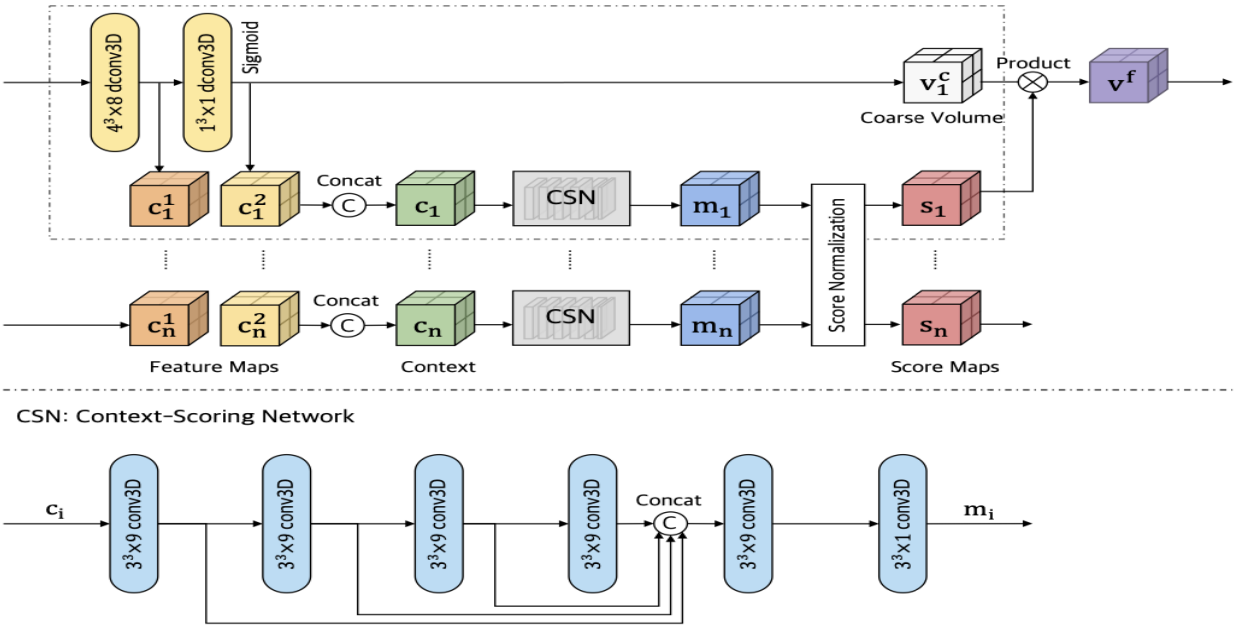
- At the end of the encoder, we attach 3 Convolution layers so that the output of the encoder can be made available for the decoder

Decoder:

- The decoder part is completely made up of 3D convs so that we can get the desired 3D model from the 2D image
- If we want to output high-resolution models then we just need to increase the number of convnets and also the number of channels used in the 3D convnets

Multi-scale Context-aware Fusion





- This part is responsible for Multi-view 3D reconstruction. Basically, it selects the high-quality portion from the coarse 3D volumes and then fuses the 3D model in such a way that only the best portion is selected.
- So basically it tries to give a score to each voxel of the 3D model based on its priority and quality. Then we multiply the score map with course volume and then add all of them to get the final 3D model.
- We also try to concatenate multiple feature map so that we don't lose the important previous features while calculating the score map.

Refiner

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

Training

- We first train the model without the multi-scale fusion using only single view images and then train the network with multi-scale context-aware fusion for multiple view images.