3D Reconstruction from Single Image Using Occupancy Network with Vision Transformer Architecture

Machine Learning for 3D Geometry SS23

Supervisor: Prof. Dr. Angela Dai

Students: Van Cuong Dam, Volkan Özer, Furkan Yakal, Kaan Özgen

Outline

- Introduction
- Motivation
- Method - Primary Idea
- Method - Final Approach
- Experiment & Evaluation
- Conclusion
- References

Introduction

- 3D reconstruction takes a set of 2D images or collection of 3D point clouds and outputs the shape and structure of the object or scene
- There have been many architectures proposed in recent years, such as Conv. Occupancy Networks [1], DeepSDF [2], and Pix2Vox [3]

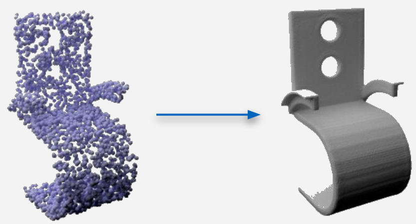


Figure 1. Input and output sample

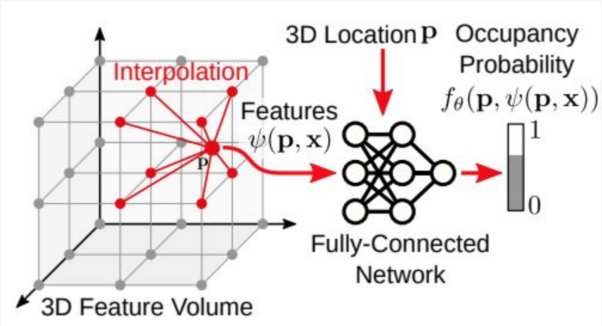


Figure 2. Conv. Occupancy Network[1]

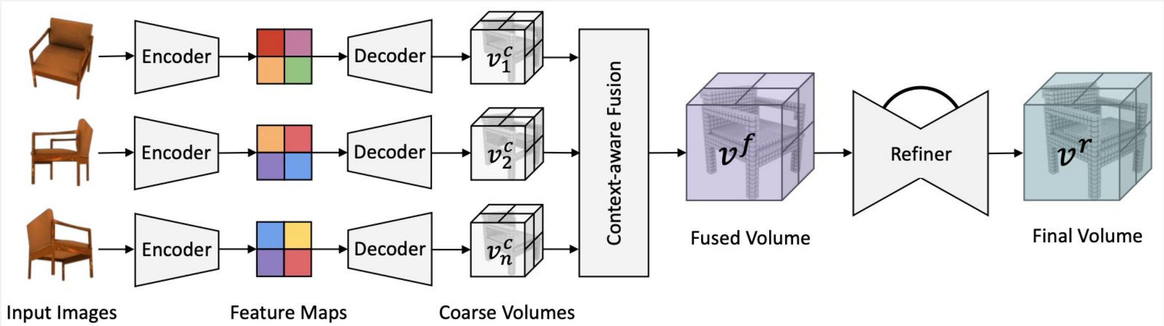


Figure 3. Pix2Vox[3]

Occupancy Networks [4]

- 3D Geometry as the decision boundary of a classifier
- Takes 3D point clouds and 2D as input and outputs their occupancy probability
- 3D reconstruction from point clouds, single images and voxel grids

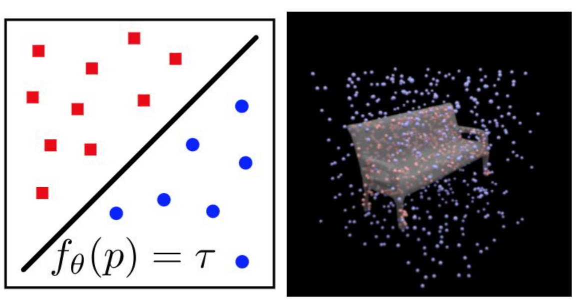


Figure 4. Decision boundary representing the surface of the reconstructed shape

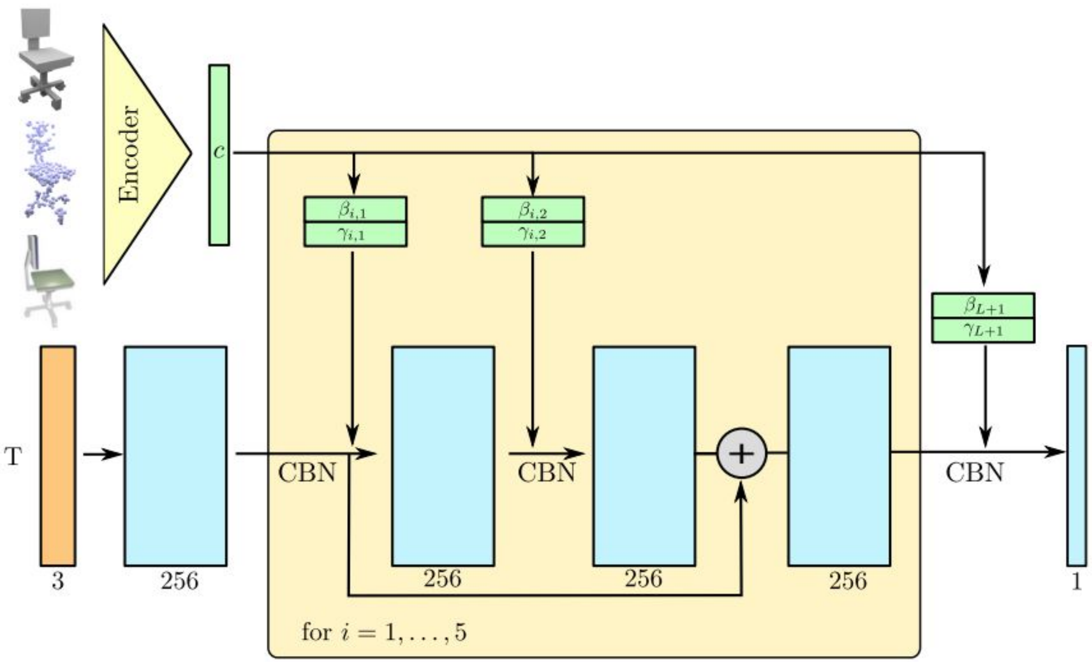


Figure 5. Occupancy Network architecture

Motivation

Vision Transformer (ViT) [5]

- Split an image into a sequence of image patches
- Patch embeddings mixed with positional embeddings
- Transformer Encoder
- Multi-layer perceptron (MLP)

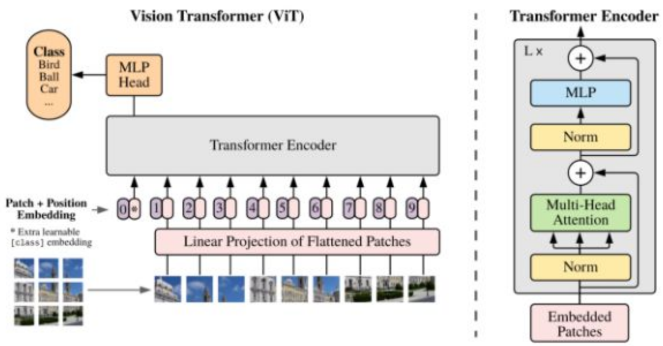


Figure 6. ViT architecture

Detection Transformer (DETR) [6]

- CNN backbone
- Positional encoding
- Transformer Encoder-Decoder
- Feed-forward networks

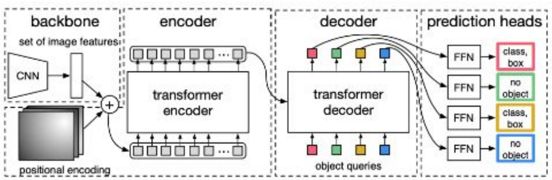


Figure 7. DETR pipeline

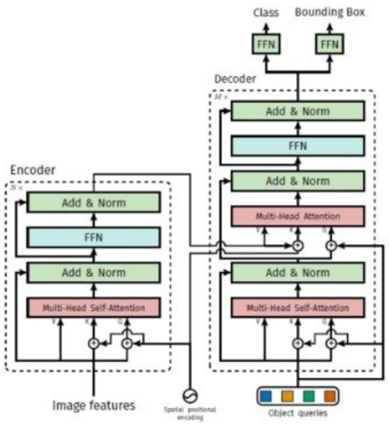
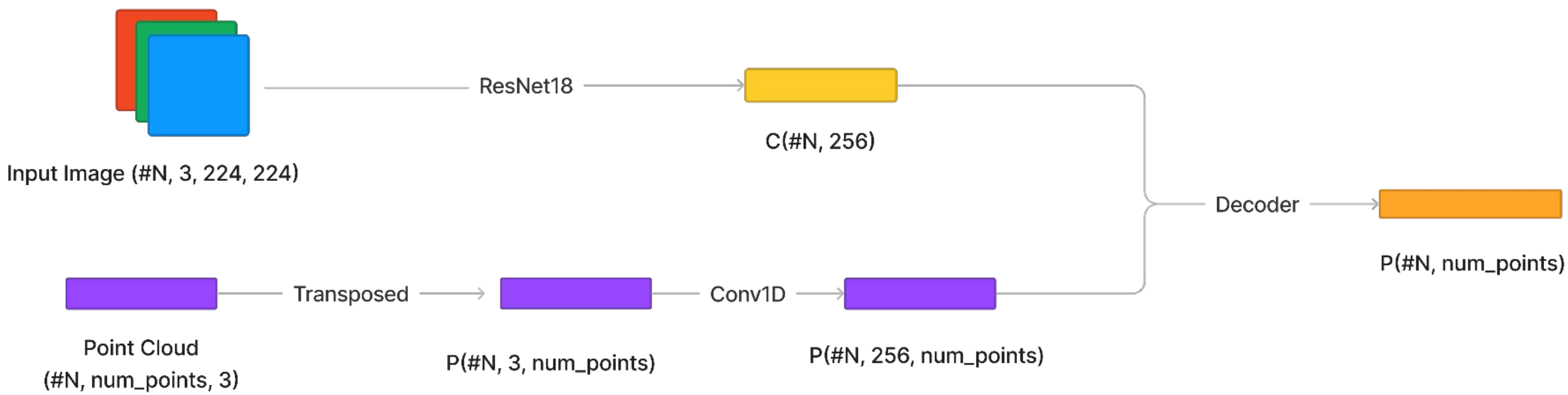


Figure 8. DETR Transformer architecture

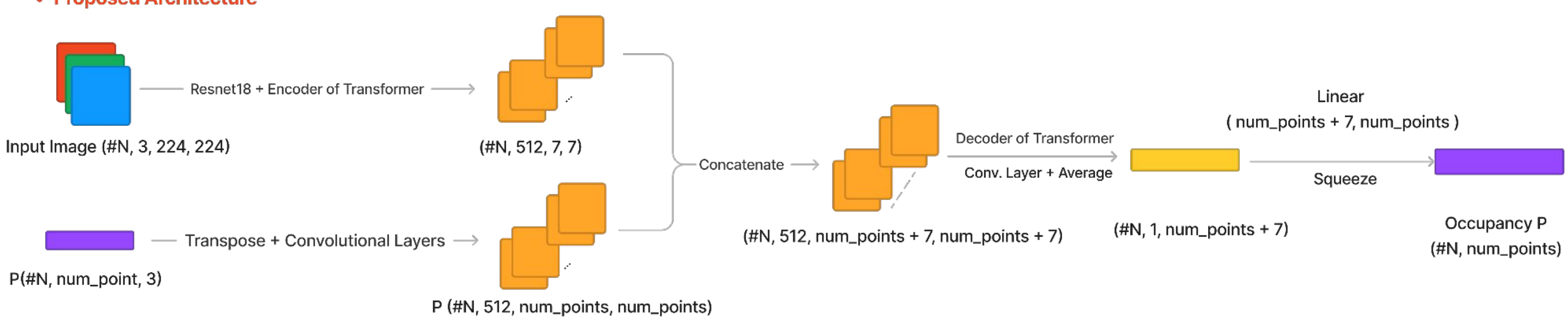
Method - Primary Idea

DeTrOcNet and Problem

- **Occupancy Network Architecture**

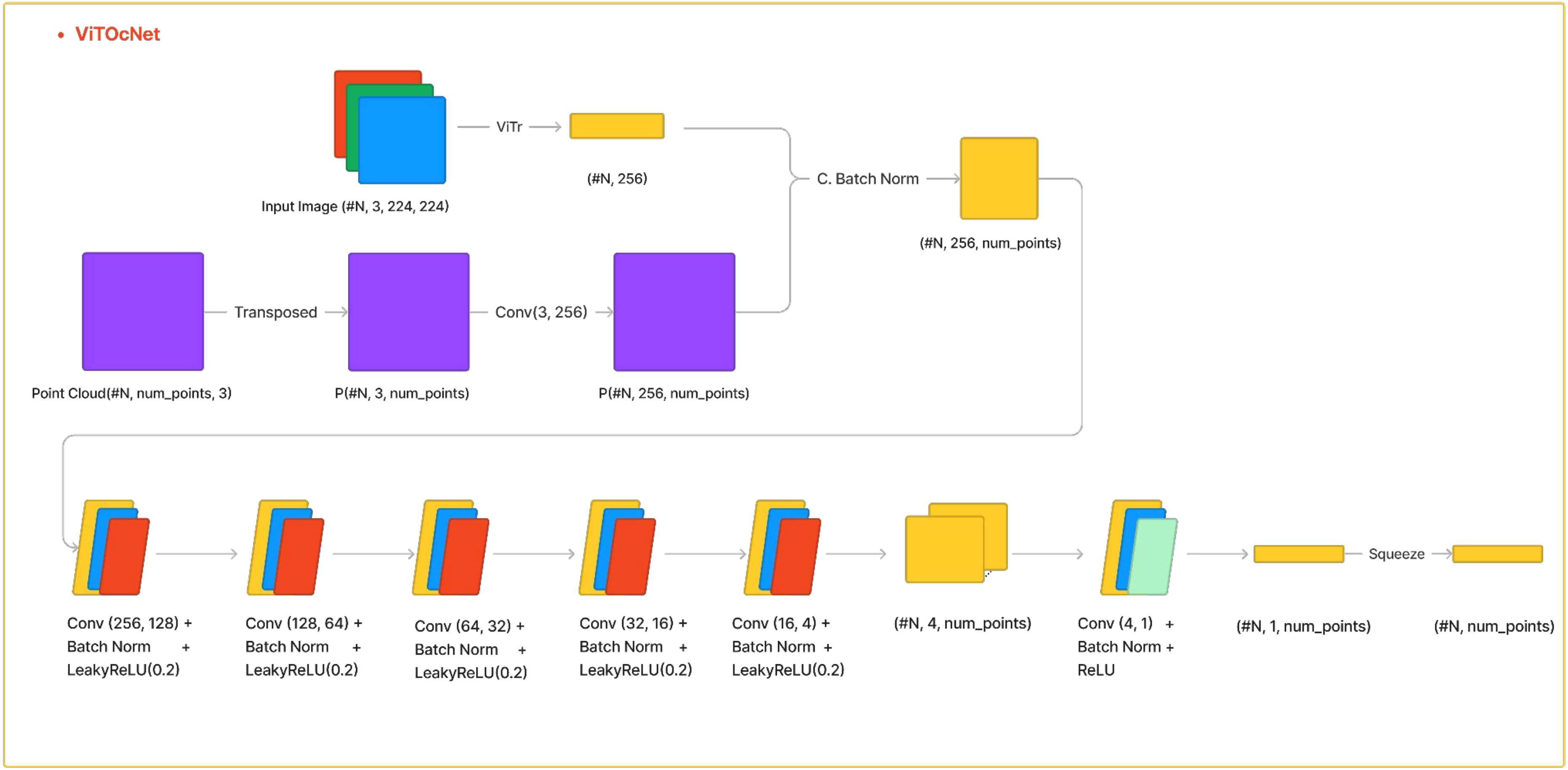


- **Proposed Architecture**



Method
Final Approach

ViTOcNet



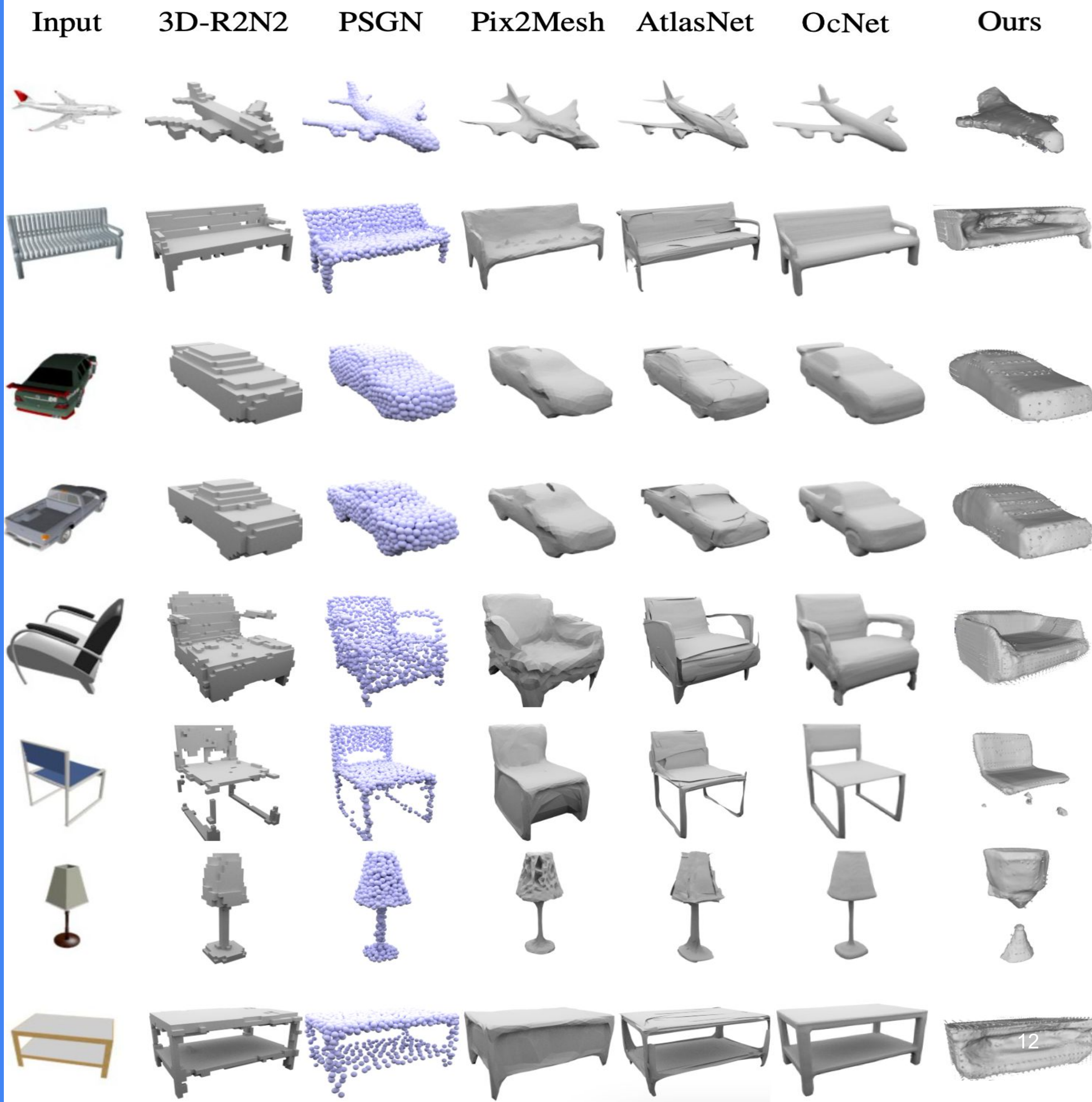
• ViTOcNet
ViTr
(#N, 256)
Input Image (#N, 3, 224, 224)
C. Batch Norm
(#N, 256, num_points)
Transposed
Conv(3, 256)
Point Cloud(#N, num_points, 3)
P(#N, 3, num_points)
P(#N, 256, num_points)
Conv (256, 128) + Batch Norm + LeakyReLU(0.2)
Conv (128, 64) + Batch Norm + LeakyReLU(0.2)
Conv (64, 32) + Batch Norm + LeakyReLU(0.2)
Conv (32, 16) + Batch Norm + LeakyReLU(0.2)
Conv (16, 4) + Batch Norm + LeakyReLU(0.2)
(#N, 4, num_points)
Conv (4, 1) + Batch Norm + ReLU
(#N, 1, num_points)
Squeeze
(#N, num_points)

Experiment & Evaluation

Dataset | ShapeNet [7]

- A repository of shapes represented by 3D models of objects
- Mesh-fusion repository for preprocessing [8]
- shape.obj files to form watertight meshes in pointcloud.npz and points.npz
- OpenGL dependent libraries “glew.h”, “gl.h”, “glu.h”, “glut.h”
- Permission not given to install the libraries

Qualitative Results



Quantitative Results

Table 1: Metrics for Different Objects

Object	IoU	Chamfer-L1	Normal Consistency
airplane	0.310	0.410	0.708
bench	0.117	0.687	0.603
cabinet	0.549	0.344	0.752
car	0.582	0.232	0.766
chair	0.311	0.529	0.706
display	0.296	0.598	0.670
lamp	0.189	0.809	0.520
loudspeaker	0.511	0.504	0.712
rifle	0.255	0.321	0.656
sofa	0.401	0.430	0.666
table	0.222	0.508	0.715
telephone	0.515	0.304	0.828
vessel	0.289	0.394	0.617
mean	0.350	0.467	0.686

category	IoU					Chamfer-L_1					Normal Consistency				
	3D-R2N2	PSGN	Pix2Mesh	AtlasNet	ONet	3D-R2N2	PSGN	Pix2Mesh	AtlasNet	ONet	3D-R2N2	PSGN	Pix2Mesh	AtlasNet	ONet
airplane	0.426	-	0.420	-	**0.571**	0.227	0.137	0.187	**0.104**	0.147	0.629	-	0.759	0.836	**0.840**
bench	0.373	-	0.323	-	**0.485**	0.194	0.181	0.201	**0.138**	0.155	0.678	-	0.732	0.779	**0.813**
cabinet	0.667	-	0.664	-	**0.733**	0.217	0.215	0.196	0.175	**0.167**	0.782	-	0.834	0.850	**0.879**
car	0.661	-	0.552	-	**0.737**	0.213	0.169	0.180	**0.141**	0.159	0.714	-	0.756	0.836	**0.852**
chair	0.439	-	0.396	-	**0.501**	0.270	0.247	0.265	**0.209**	0.228	0.663	-	0.746	0.791	**0.823**
display	0.440	-	**0.490**	-	0.471	0.314	0.284	0.239	**0.198**	0.278	0.720	-	0.830	**0.858**	0.854
lamp	0.281	-	0.323	-	**0.371**	0.778	0.314	0.308	**0.305**	0.479	0.560	-	0.666	0.694	**0.731**
loudspeaker	0.611	-	0.599	-	**0.647**	0.318	0.316	0.285	**0.245**	0.300	0.711	-	0.782	0.825	**0.832**
rifle	0.375	-	0.402	-	**0.474**	0.183	0.134	0.164	**0.115**	0.141	0.670	-	0.718	0.725	**0.766**
sofa	0.626	-	0.613	-	**0.680**	0.229	0.224	0.212	**0.177**	0.194	0.731	-	0.820	0.840	**0.863**
table	0.420	-	0.395	-	**0.506**	0.239	0.222	0.218	0.190	**0.189**	0.732	-	0.784	0.832	**0.858**
telephone	0.611	-	0.661	-	**0.720**	0.195	0.161	0.149	**0.128**	0.140	0.817	-	0.907	0.923	**0.935**
vessel	0.482	-	0.397	-	**0.530**	0.238	0.188	0.212	**0.151**	0.218	0.629	-	0.699	0.756	**0.794**
mean	0.493	-	0.480	-	**0.571**	0.278	0.215	0.216	**0.175**	0.215	0.695	-	0.772	0.811	**0.834**

Conclusion

The original network is indeed a more powerful model for 3D reconstruction

Our approach provides a different view w.r.t preprocessing 2D images and dealing with points clouds.

Due to the limitation of hardware resources, the limit performance of the model could not be tested.

We obtain experience in dealing with 3D points

THANK YOU FOR LISTENING

References

1. Peng, S., Niemeyer, M., Mescheder, L., Pollefeys, M., & Geiger, A. (2020, August 1). *Convolutional Occupancy Networks*. arXiv.org. https://arxiv.org/abs/2003.04618
2. Park, J. J., Florence, P., Straub, J., Newcombe, R., & Lovegrove, S. (2019, January 16). *DEEPSDF: Learning continuous signed distance functions for shape representation*. arXiv.org. https://arxiv.org/abs/1901.05103
3. Xie, H., Yao, H., Sun, X., Zhou, S., & Zhang, S. (2019, July 29). *Pix2Vox: Context-aware 3D reconstruction from single and Multi-view images*. arXiv.org. https://arxiv.org/abs/1901.11153v2
4. Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., & Geiger, A. (2019, April 30). *Occupancy networks: Learning 3D reconstruction in Function Space*. arXiv.org. https://arxiv.org/abs/1812.03828
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021, June 3). *An image is worth 16x16 words: Transformers for image recognition at scale*. arXiv.org. https://arxiv.org/abs/2010.11929
6. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020, May 28). *End-to-end object detection with Transformers*. arXiv.org. https://arxiv.org/abs/2005.12872
7. Chang, A. X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., Xiao, J., Yi, L., & Yu, F. (2015, December 9). *ShapeNet: An information-rich 3D model repository*. arXiv.org. https://arxiv.org/abs/1512.03012
8. D. Stutz and A. Geiger, “Learning 3D Shape Completion under Weak Supervision,” CoRR, vol. abs/1805.07290, 2018, [Online]. Available: http://arxiv.org/abs/1805.07290

Vision Transformer (ViT) [5]

- Split an image into a sequence of image patches
- Patch embeddings mixed with positional embeddings
- Transformer Encoder
- Multi-layer perceptron (MLP)

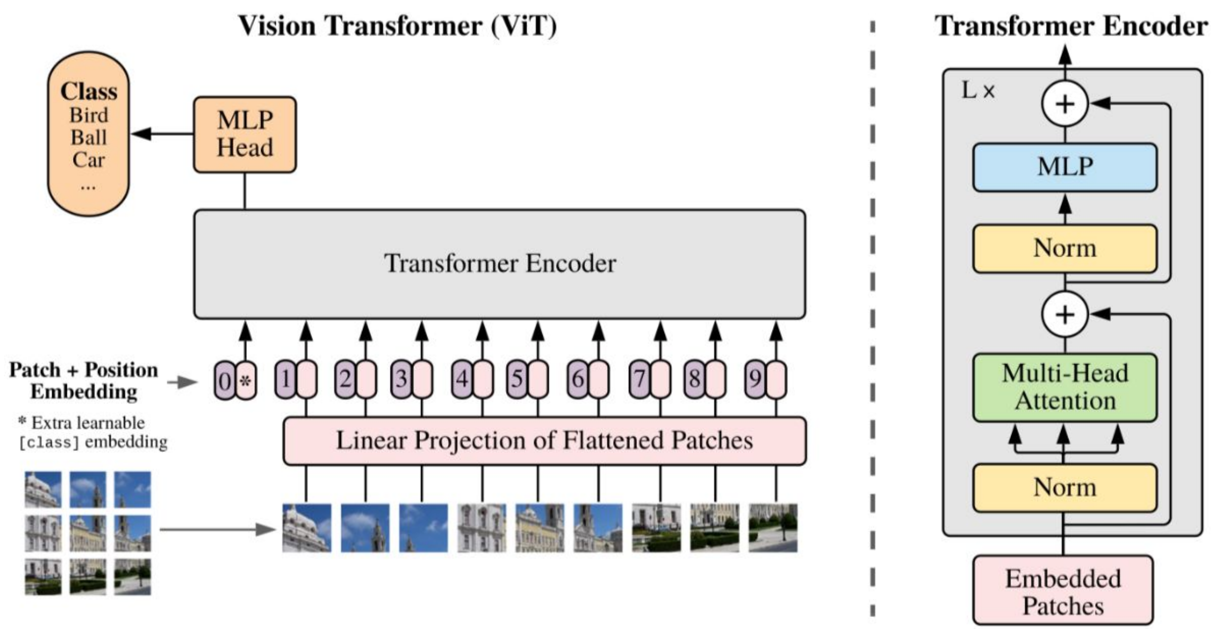


Figure 6. ViT architecture

Detection Transformer (DETR) [6]

- CNN backbone
- Positional encoding
- Transformer Encoder-Decoder
- Feed-forward networks

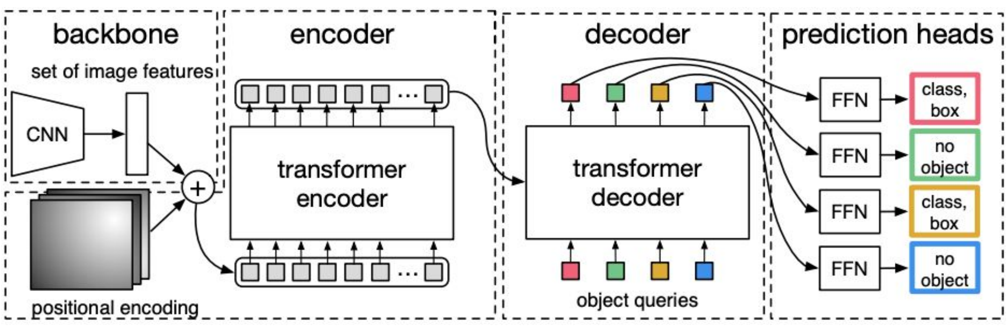


Figure 7. DETR pipeline

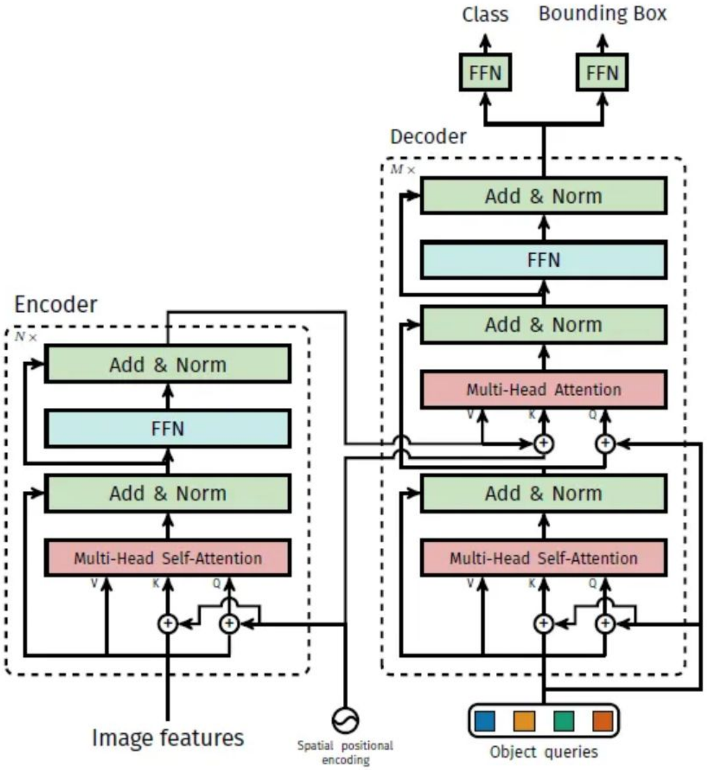


Figure 8. DETR Transformer architecture