Pix2Vox: Context-aware 3D Reconstruction from Single and Multi-view Images

Haozhe Xie[1] Hongxun Yao[1] Xiaoshuai Sun[1] Shangchen Zhou[2] Shengping Zhang[1,3]

[1] Harbin Institute of Technology [2] SenseTime Research [3] Peng Cheng Laboratory

ICCV 2019 Seoul, Korea

Project Page
http://bit.ly/Pix2Vox

Introduction

Goal

- Recover the 3D shape of an object from single or multiple viewpoints.
- Unknown shape priors (e.g., lighting condition) and camera intrinsic and extrinsic parameters.

The Problems of 3D-R2N2

- RNN is not permutation invariant
- Long-short memory loss of RNN
- RNN-based methods are time-consuming

How to Solve it?

- Propose a new module, called *Context-aware Fusion*.

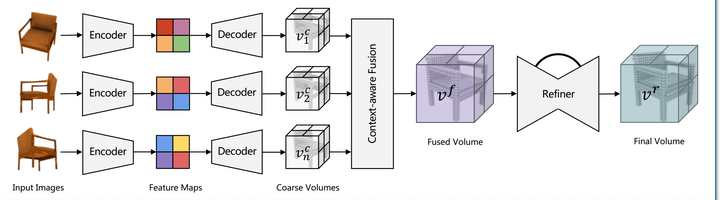
The Advantages of Context-aware Fusion

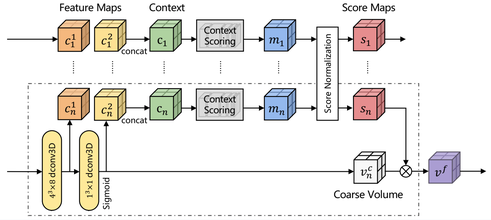
- Selects high-quality reconstruction parts from different coarse 3D volumes in parallel.
- Better preserves multi-view spatial constraints when selecting reconstruction parts.

Contributions

- A unified framework for both single-view and multi-view 3D object reconstruction, which consists of a well-designed encoder, decoder and refiner.
- A context-aware fusion module that adaptively selects high-quality reconstruction parts from different coarse 3D volumes to produce a fused reconstruction of the whole object.
- Outperform all state-of-the-arts methods on the ShapeNet and Pix3D datasets in terms of both accuracy and efficiency.
- Additional experiments show strong generalization abilities in reconstructing unseen 3D objects.

Proposed Method





Score Normalization

$$s_r^{(i,j,k)} = \frac{\exp(m_r^{(i,j,k)})}{\sum_{p=1}^{n}\exp(m_p^{(i,j,k)})}$$

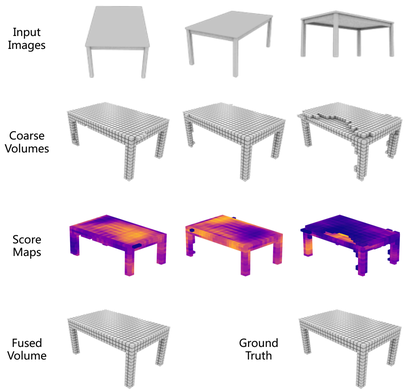
Weighting Coarse Volumes

$$v^f = \sum_{i=1}^{n} s_r v_r^c$$

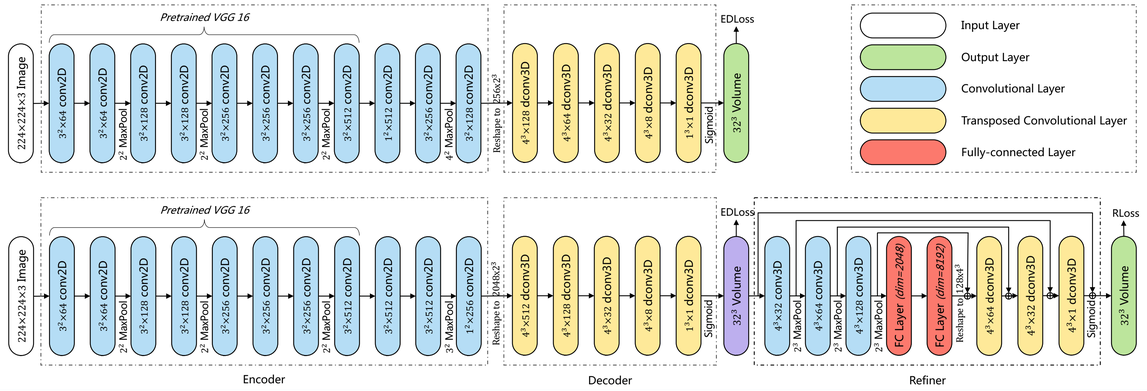
Loss Function

$$\ell = \frac{1}{N}\sum_{i=1}^{n}[g_{t_i}\log p_i + (1-g_{t_i})\log(1-p_i)]$$

Context-aware Fusion



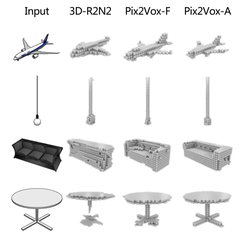
Network Architecture

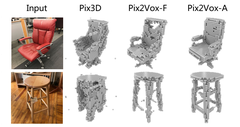


Experimental Results

Single-view Reconstruction on ShapeNet

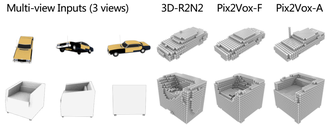
Category	3D-R2N2	OGN	DRC	PSGN	Pix2Vox-F	Pix2Vox-A
airplane	0.513	0.587	0.571	0.601	0.600	0.684
bench	0.421	0.481	0.453	0.550	0.538	0.616
cabinet	0.716	0.729	0.635	0.771	0.765	**0.792**
car	0.798	0.828	0.755	0.831	0.837	**0.854**
chair	0.466	0.483	0.469	0.544	0.535	**0.567**
display	0.468	0.503	0.419	**0.552**	0.511	0.537
lamp	0.381	0.398	0.415	**0.462**	0.435	0.443
speaker	0.662	0.637	0.609	**0.737**	0.707	0.714
rifle	0.544	0.593	0.608	0.604	0.598	**0.615**
sofa	0.628	0.646	0.606	0.708	0.687	**0.709**
table	0.513	0.536	0.424	**0.606**	0.587	0.601
telephone	0.661	0.702	0.413	0.749	0.770	**0.776**
watercraft	0.513	**0.632**	0.556	0.611	0.582	0.594
IoU	0.560	0.596	0.545	0.640	0.634	**0.661**

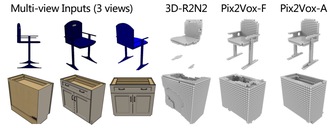




Multi-view Reconstruction on ShapeNet

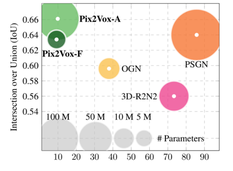
# views	1	2	3	4	5	8	12	16	20
3D-R2N2	0.560	0.603	0.617	0.625	0.634	0.635	0.636	0.636	0.636
Pix2Vox-F	0.634	0.660	0.668	0.673	0.676	0.680	0.682	0.684	0.684
Pix2Vox-A	**0.661**	**0.686**	**0.693**	**0.697**	**0.699**	**0.702**	**0.704**	**0.705**	**0.706**





Memory usage and Running Time

Methods	3D-R2N2	OGN	PSGN	Pix2Vox-F	Pix2Vox-A
# Parameters (M)	35.97	12.46	210.46	7.41	114.24
Memory (MB)	1407	793	2733	673	2729
Training Time (hours)	169	192	50	12	25
Backward Time (ms)	312.50	312.25	97.90	12.93	72.01
Forward Time, 1 view (ms)	73.35	37.90	85.73	9.25	9.90
Forward Time, 2 views (ms)	108.11	N/A	N/A	12.05	13.69
Forward Time, 4 views (ms)	112.36	N/A	N/A	23.26	26.31



References

- [3D-R2N2] 3D-R2N2: A Unified Approach for Single and Multi-view 3D Object Reconstruction. *ECCV 2016.*
- [OGN] Octree Generating Networks: Efficient Convolutional Architectures for High-resolution 3D Outputs. *ICCV 2017.*
- [DRC] Multi-view Supervision for Single-View Reconstruction via Differentiable Ray Consistency. CVPR 2017.
- [PSGN] A Point Set Generation Network for 3D Object Reconstruction from a Single Image. CVPR 2017.