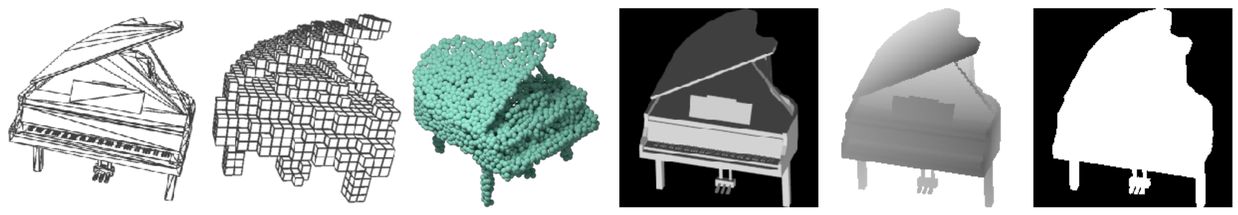
A Deeper Look at 3D Shape Classifiers

Jong-Chyi Su Matheus Gadelha Rui Wang Subhransu Maji
University of Massachusetts, Amherst

ECCV 2018 European Conference on Computer Vision
8 – 14 September 2018 | Munich, Germany

Motivation

We compare different representations and architectures for classifying 3D shapes in terms of 1) computational efficiency, 2) generalization, and 3) robustness to adversarial examples, on ModelNet40 dataset.



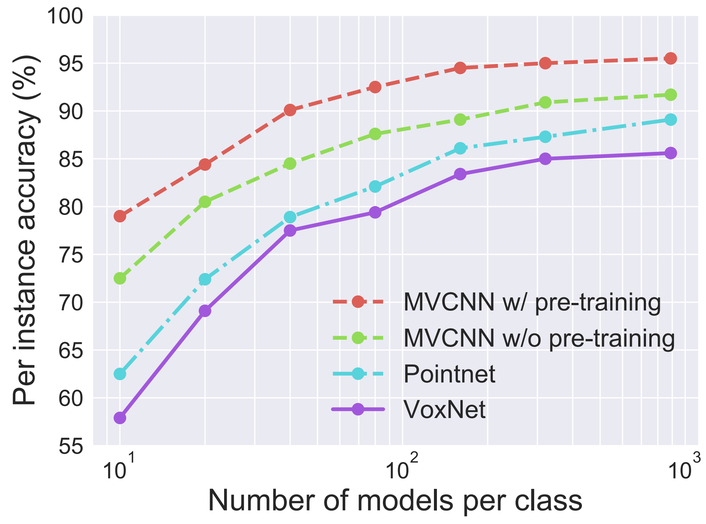
Meshes | Voxels | Point Cloud | Shaded | Depth | Silhouette

Prior Work

- **Multiview Network (MVCNN [1])**
  - Use rendered images from different views as input.
- **Point-based Network (PointNet [2])**
  - Performs operations on each input point and performs orderless aggregations using max-pooling operations.
- **Voxel-based Network (VoxNet [3])**
  - Use convolutional and pooling layers defined on 3D voxel grids.

Learning from Few Examples

MVCNN outperforms VoxNet and PointNet, and generalizes better.



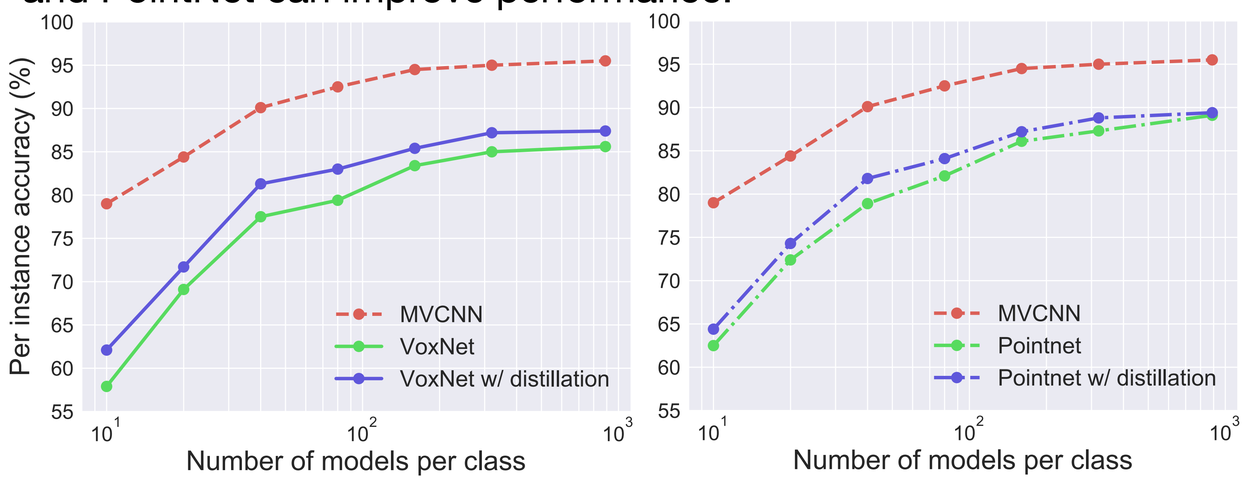
Effect of Shape Rendering

With better renderings, MVCNN can obtain 95.0% accuracy.
Without ImageNet pre-training, MVCNN can still achieve 91.3%.

Model	Rendering	Full Training/Test		80/20 Training/Test
		Per Class Acc. (%)	Per Inst. Acc. (%)	Per Class / Per Inst. Acc. (%)
VGG-M	Shaded from [1]	-	-	89.9
	Shaded from [1] (80x)	-	-	90.1
VGG-11	Shaded from [1]	-	-	89.1
	Shaded	**92.4**	**95.0**	**92.4**
	Shaded, w/o ImageNet	88.7	91.3	
	Depth	89.8	91.6	-
	Shaded + Depth	94.7	96.2	-
	Silhouettes	90.7	93.6	-

Cross Modal Distillation

Using representations from MVCNN to guide learning of VoxNet and PointNet can improve performance.



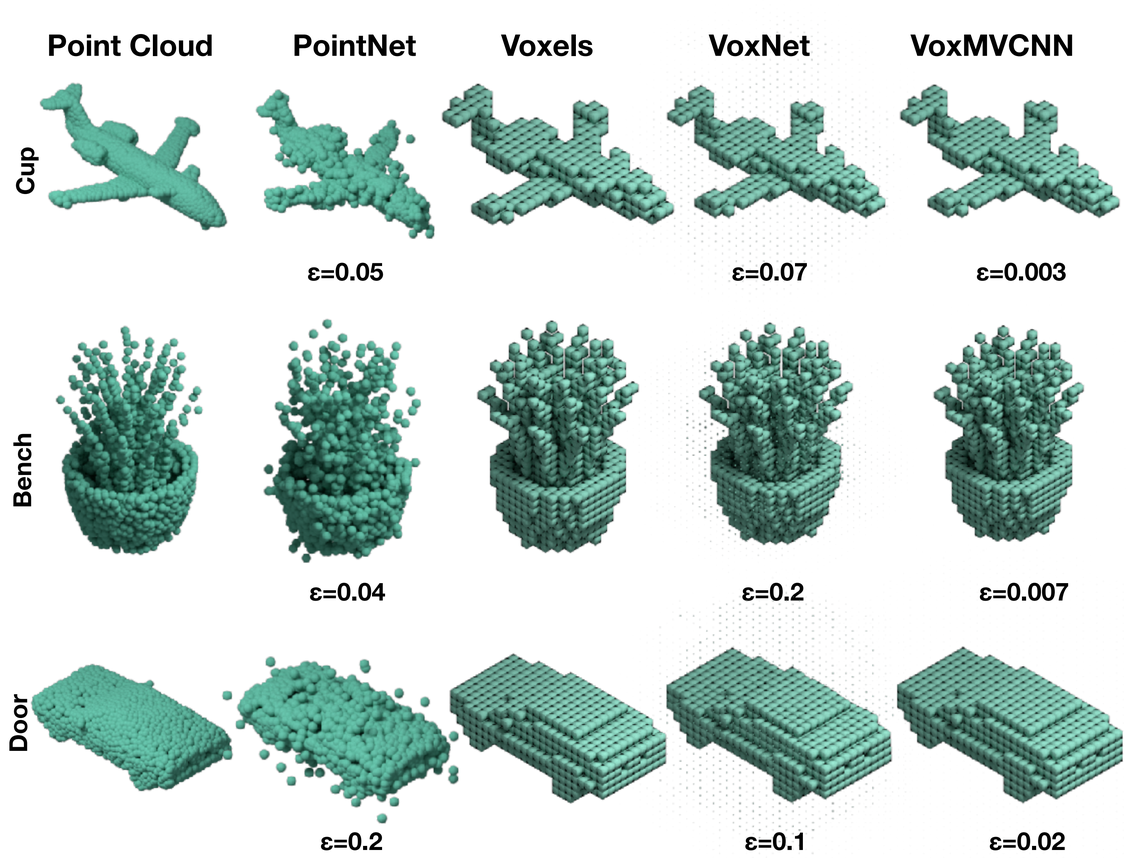
Tradeoffs between Representations

	Forward Time (ms)	#Params	Memory (GB)	Per Class Acc. (%)	Per Inst. Acc. (%)
MVCNN	25.8	128.9M	10.0	92.4	95.0
VoxNet	1.3	1.4M	2.0	81.4	85.6
PointNet	3.1	3.5M	4.4	86.1	89.1

Robustness to Adversarial Examples

- Adding perturbations to voxel occupancy values or coordinates of points to generate adversarial examples.
- **VoxMVCNN**: For multiview methods, we use voxels as input and generate images from 6 different views.
- Point-based network is more robust, while voxel-based and multiview networks are easily fooled by adding imperceptible noise to the input.
- ImageNet pre-training reduces the robustness of VoxMVCNN.
- ε: minimum threshold of the perturbation where we can generate adversarial examples.

	PointNet	VoxNet	VoxMVCNN w/o pre-training	VoxMVCNN w/ pre-training
# Adversarial Examples (out of 400)	379	370	290	399
ε	0.045	0.061	0.041	0.006
Per Inst. Acc. (%)	86.8	85.6	84.4	88.2



Comparison to Prior Work

	Input	Per Class Acc. (%)	Per Inst. Acc. (%)
Our MVCNN	Images	**92.4**	**95.0**
Rotation Net		-	94.8
Dominant Set Clustering		-	93.8
MVCNN-MultiRes		91.4	93.8
MVCNN [1]		90.1	90.1
DynamicGraph	Point Clouds	90.2	92.2
Kd-Networks		-	91.8
PointNet++		-	90.7
PointNet [2]		86.2	89.2
VRN Single	Voxels	-	91.3
O-CNN		-	90.6
VoxNet [3]		-	83.0
3DShapeNets		77.3	84.7
PointNet++	PC+Normal	-	91.9
FusionNet	Voxels+Images	-	90.8

References

[1] Su et al., Multi-view convolutional neural networks for 3d shape recognition, *ICCV* 2015
[2] Su et al., Pointnet: Deep learning on point sets for 3d classification and segmentation, *CVPR* 2017
[3] Maturana and Scherer, Voxnet: A 3d convolutional neural network for real-time object recognition, *IROS* 2015