Render for CNN:
Viewpoint Estimation in Images Using CNNs Trained with Rendered 3D Model Views

Hao Su*  
Charles R. Qi*  
Yangyan Li  
Leonidas J. Guibas

Stanford University
ILSVRC Image Classification Top-5 Error (%)
Go beyond 2D Image Classification

- 3D bounding box
- 3D alignment
- 3D model retrieval
Go beyond 2D Image Classification

3D Viewpoint Estimation

- in-plane rotation
- elevation
- azimuth
3D Viewpoint Estimation in the Wild

Images in the Wild
Models unknown
3D Perception in the Wild

Learn from Data

Images in the Wild
Models unknown

AlexNet [Krizhevsky et al.]
However, Accurate Label Acquisition is Expensive

What’s the camera viewpoint angles to the SUV in the image?
However, accurate label acquisition is expensive.

PASCAL3D+ dataset [Xiang et al.]
However, Accurate Label Acquisition is Expensive

Step 1: Choose similar model
However.. Accurate Label Acquisition is Expensive

Step 1:
Choose similar model

Step 2:
Coarse Viewpoint Labeling
However, Accurate Label Acquisition is Expensive

**Step 1:** Choose similar model

**Step 2:** Coarse Viewpoint Labeling

**Step 3:** Label keypoints For alignment

Annotation takes ~1 min per object
30K images with viewpoint labels in PASCAL3D+ dataset [Xiang et al.]

60M parameters. AlexNet [Krizhevsky et al.]

How to get MORE images with ACCURATE viewpoint labels?
Manual alignment by annotators

Auto alignment through rendering
Good News: ShapeNet

- 3M models in total
- 330K from 4K categories annotated

http://shapenet.cs.stanford.edu
Key Idea: Render for CNN

ShapeNet → Rendering → Synthetic Images

Training

Viewpoint $\mathbb{R}^3$ → Convolutional Neural Network → Synthetic Images
Key Idea: Render for CNN

Testing

Viewpoint $\mathbb{R}^3$

Convolutional Neural Network

Real Images
I want data!

How to render data with both quantity and quality?
Synthesize: Scalability vs Quality

<table>
<thead>
<tr>
<th>Scalability</th>
<th>Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>High</td>
<td>Low</td>
</tr>
</tbody>
</table>

Ideal
Synthesize: Scalability vs Quality

Scalability vs Quality

Ideal

Low Quality

High Quality

Low Scalability

High Scalability
Synthesize: Scalability vs Quality

Scalability vs Quality:
- Low Quality
- High Quality
- Low Scalability
- High Scalability

Ideal: Previous works

Ideal point on the graph.
Synthesize: Scalability vs Quality

Scalability vs Quality:
- High Quality
- Low Quality
- High Scalability
- Low Scalability

Ideal Sweet spot
Previous works
Synthesize: Scalability vs Quality

- Scalability
- Quality

- Low
- High

Previous works

Ideal

Sweet spot

Story Time!
A “Data Engineering” Journey

• 80K rendered chair images
• Metric: 16-view classification accuracy tested on real images

At beginning..
• Lighting: 4 fixed point light sources on the sphere
• Background: clean
A “Data Engineering” Journey

ConvNet: Ah ha, I know!
Viewpoint is just the brightness pattern!

95% on synthetic val set
47% on real test set 😞
A “Data Engineering” Journey

95% on synthetic val set
47% on real test set 😞
A “Data Engineering” Journey

Randomize lighting

47% -> 74%

**ConvNet:** hmm.. viewpoint is not the brightness pattern. Maybe it’s the contour?
A “Data Engineering” Journey

ConvNet: hmm.. viewpoint is not the brightness pattern. Maybe it’s the contour?
A “Data Engineering” Journey

Add backgrounds

74% -> 86%

**ConvNet:** It becomes really hard! Let me look more into the picture.
A “Data Engineering” Journey

bbox crop

texture

86% -> 93%
A “Data Engineering” Journey

**Key Lesson:** Don’t give CNN a chance to “cheat” - it’s very good at it. When there is no way to cheat, true learning starts.
Render for CNN Image Synthesis Pipeline

3D model

Rendering
Sample lighting and camera params

Add bkg
Sample bkg. Image Alpha-blending

Crop
Sample cropping params

Hyper-parameters estimation from real images
Render for CNN Image Synthesis Pipeline

3D model

Rendering

Sample lighting and camera params
Lighting params
Randomly sampled

- Number of light sources
- Light distances
- Light energies
- Light positions
- Light types

Camera params KDE from PASCAL3D+ train set
Render for CNN Image Synthesis Pipeline

3D model → Rendering → Add bkg

- Sample lighting and camera params
- Sample bkg. Image Alpha-blending
Background Composition

• Simple but effective!

• Backgrounds randomly sampled from SUN397 dataset [Xiao et al.]

• Alpha blending composition for natural boundaries
Render for CNN Image Synthesis Pipeline

3D model

Rendering
- Sample lighting and camera params

Add bkg
- Sample bkg. Image Alpha-blending

Crop
- Sample cropping params
Image Cropping

Cropping patterns KDE from PASCAL3D+ train set
Image Cropping

Cropping patterns KDE from PASCAL3D+ train set
2.4M Synthesized Images for 12 Categories

• High scalability
• High quality
  • Overfit-resistant
  • Accurate labels
Results
Metric: **median angle error** (lower the better)

**Real test images** from PASCAL3D+ dataset

<table>
<thead>
<tr>
<th></th>
<th>aero</th>
<th>bike</th>
<th>boat</th>
<th>bottle</th>
<th>bus</th>
<th>car</th>
<th>chair</th>
<th>table</th>
<th>mbike</th>
<th>sofa</th>
<th>train</th>
<th>tv</th>
<th>mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Acc_{\theta}$ (Tulsiani, Malik)</td>
<td>0.78</td>
<td>0.74</td>
<td>0.49</td>
<td>0.93</td>
<td>0.94</td>
<td>0.90</td>
<td>0.65</td>
<td>0.67</td>
<td>0.83</td>
<td>0.67</td>
<td>0.79</td>
<td>0.76</td>
<td>0.76</td>
</tr>
<tr>
<td>$Acc_{\theta}$ (Ours-Render)</td>
<td>0.74</td>
<td>0.83</td>
<td>0.52</td>
<td>0.91</td>
<td>0.91</td>
<td>0.88</td>
<td>0.86</td>
<td>0.73</td>
<td>0.78</td>
<td>0.90</td>
<td>0.86</td>
<td>0.92</td>
<td>0.82</td>
</tr>
<tr>
<td>$MedErr$ (Tulsiani, Malik)</td>
<td>14.7</td>
<td>18.6</td>
<td>31.2</td>
<td>13.5</td>
<td>6.3</td>
<td>8.8</td>
<td>17.7</td>
<td>17.4</td>
<td>17.6</td>
<td>15.1</td>
<td>8.9</td>
<td>17.8</td>
<td>15.6</td>
</tr>
<tr>
<td>$MedErr$ (Ours-Render)</td>
<td>15.4</td>
<td>14.8</td>
<td>25.6</td>
<td>9.3</td>
<td>3.6</td>
<td>6.0</td>
<td>9.7</td>
<td>10.8</td>
<td>16.7</td>
<td>9.5</td>
<td>6.1</td>
<td>12.6</td>
<td>11.7</td>
</tr>
</tbody>
</table>
Our model trained on rendered images outperforms state-of-the-art model trained on real images in PASCAL3D+.
How many 3D models are necessary?

Accuracy

#models (for one category)

10 vs 1000 models
20% + difference
3D Viewpoint Estimation
Azimuth Viewpoint Estimation

Ground truth view

Estimated view confidence
Azimuth Viewpoint Estimation

Ground truth view

Estimated view confidence
Failure Cases

- sofa occluded by people
- car occluded by motorbike
- ambiguous car viewpoint
- ambiguous chair viewpoint
- multiple cars
- multiple chairs
Limitations of Current Synthesis Pipeline

- Modeling Occlusions?
- Modeling Background Context?
- Shape database augmentation by interpolation?
Render for CNN – Beyond Viewpoint

• 3D model retrieval
  • Joint Embedding [Li et al sigasia15]

• Object detection
• Segmentation
• Intrinsic image decomposition

• Controlled experiments for DL
• Vision algorithm verification
Conclusion

Images rendered from 3D models can be effectively used to train CNNs, especially for 3D tasks. State-of-the-art result has been achieved.

Keys to success

• Quantity: Large scale 3D model collection (ShapeNet)
• Quality: Overfit-resistant, scalable image synthesis pipeline

http://shapenet.cs.stanford.edu
THE END

THANK YOU!