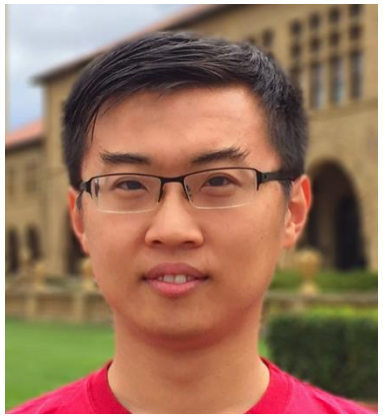
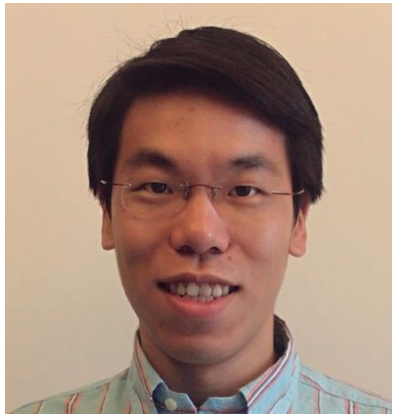
# Render for CNN:

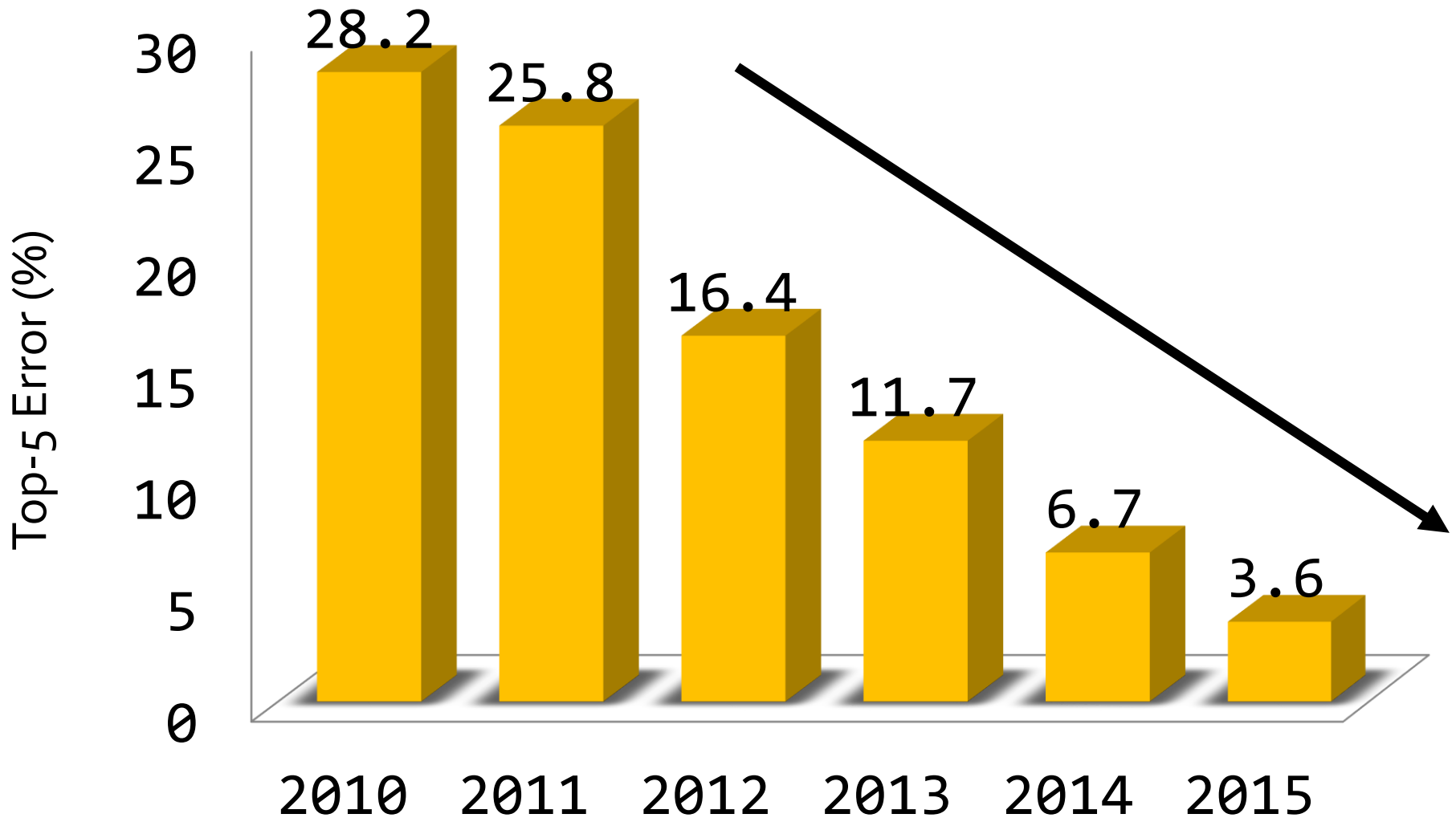
## Viewpoint Estimation in Images Using CNNs Trained with Rendered 3D Model Views

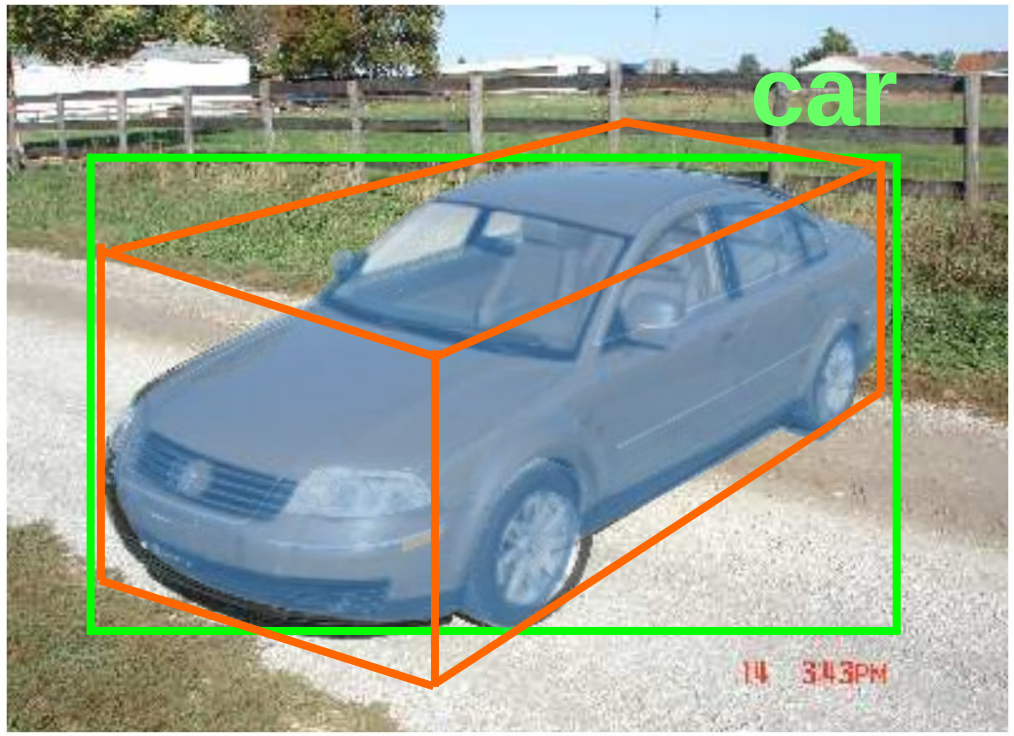

Hao Su*
Charles R. Qi*
Yangyan Li
Leonidas J. Guibas

Stanford University

# ILSVRC Image Classification Top-5 Error (%)

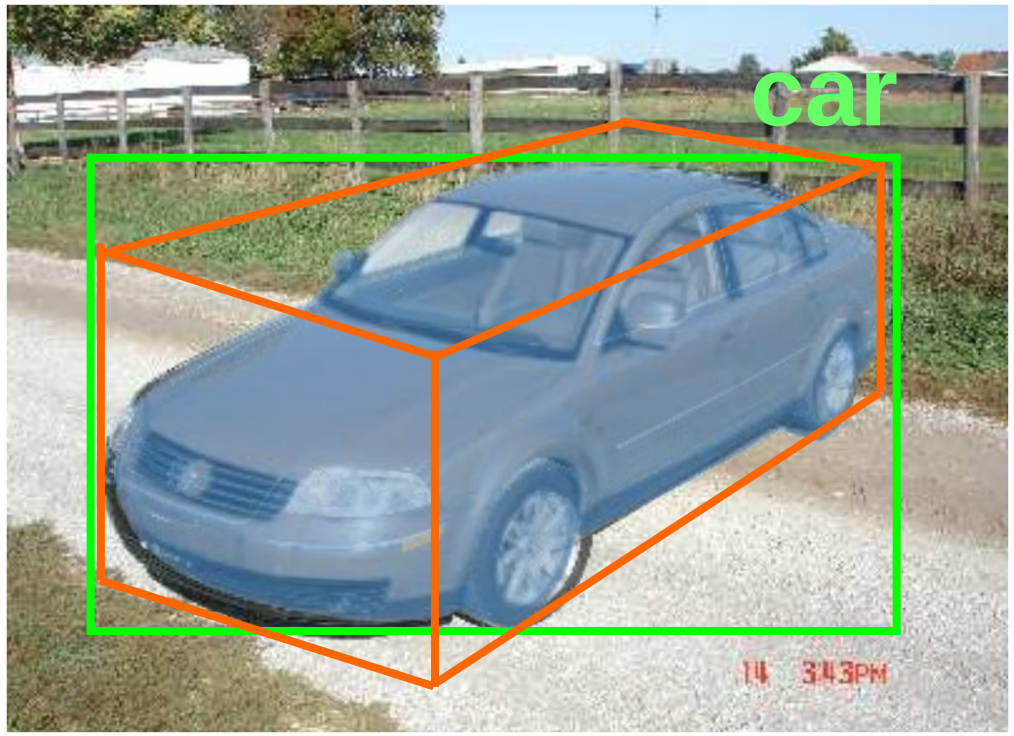


# Go beyond 2D Image Classification

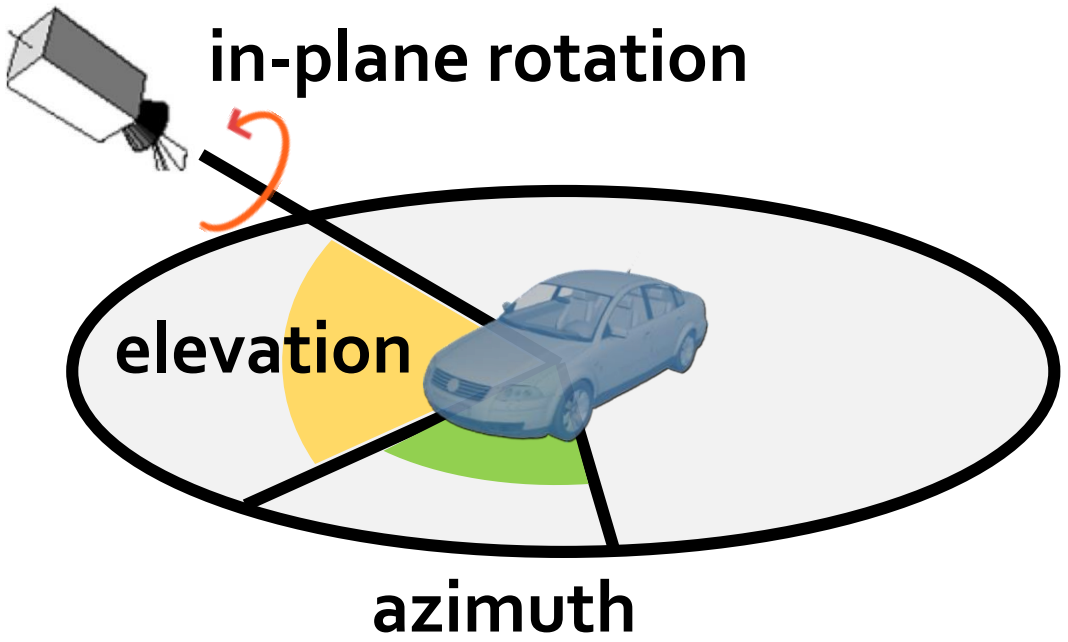



- 3D bounding box
- 3D alignment
- 3D model retrieval

# Go beyond 2D Image Classification




## 3D Viewpoint Estimation



# 3D Viewpoint Estimation in the Wild

Images in the Wild

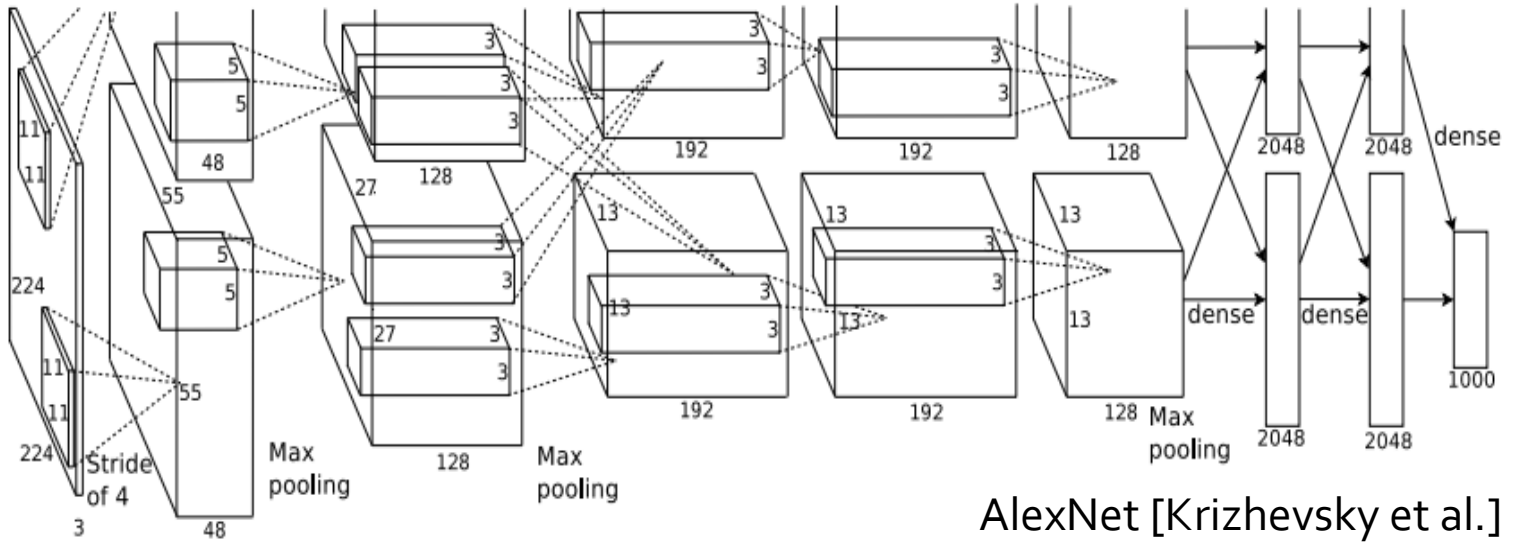
Models unknown


# 3D Perception in the Wild

Images in the Wild

Models unknown

## Learn from Data



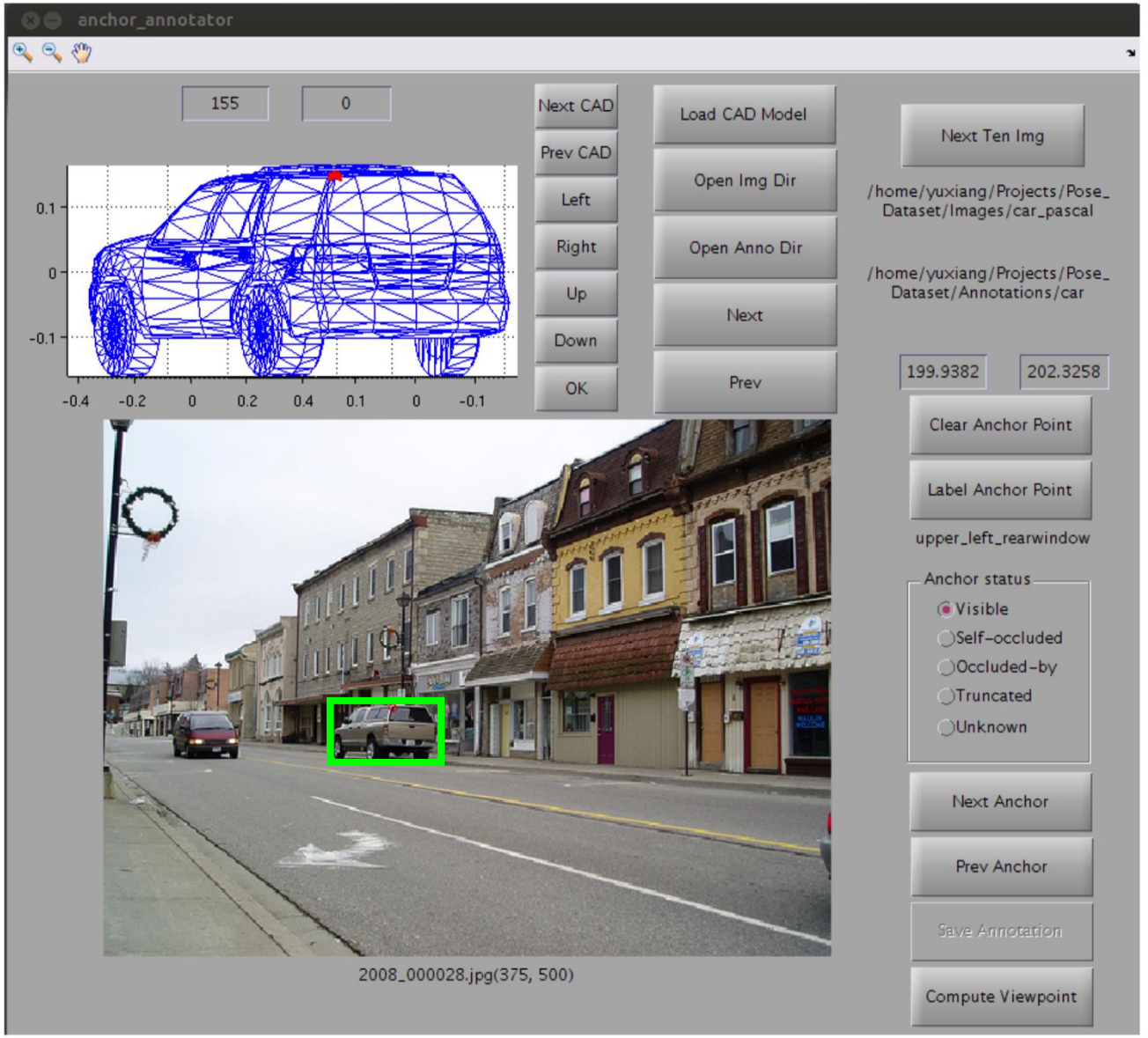

AlexNet [Krizhevsky et al.]

# However.. Accurate Label Acquisition is Expensive

**What's the camera viewpoint angles to the SUV in the image?**

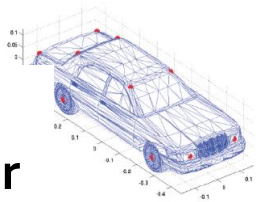
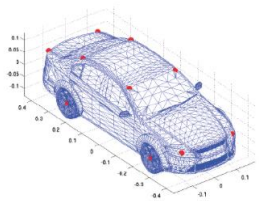
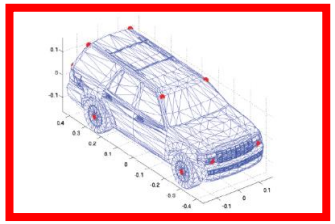
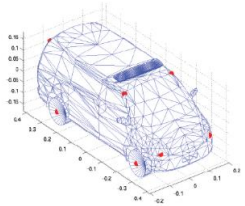
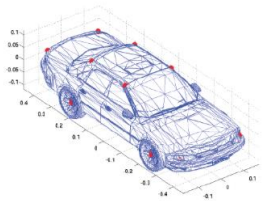
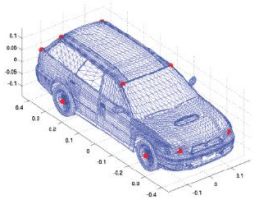
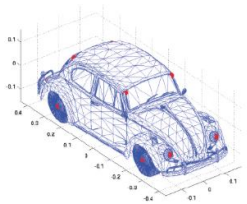
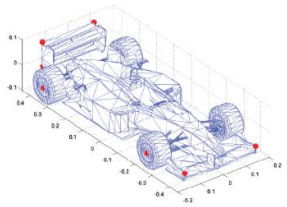
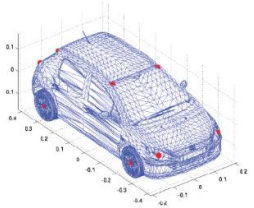
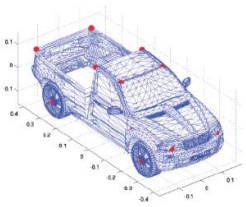
# However.. Accurate Label Acquisition is Expensive

PASCAL3D+ dataset [Xiang et al.]



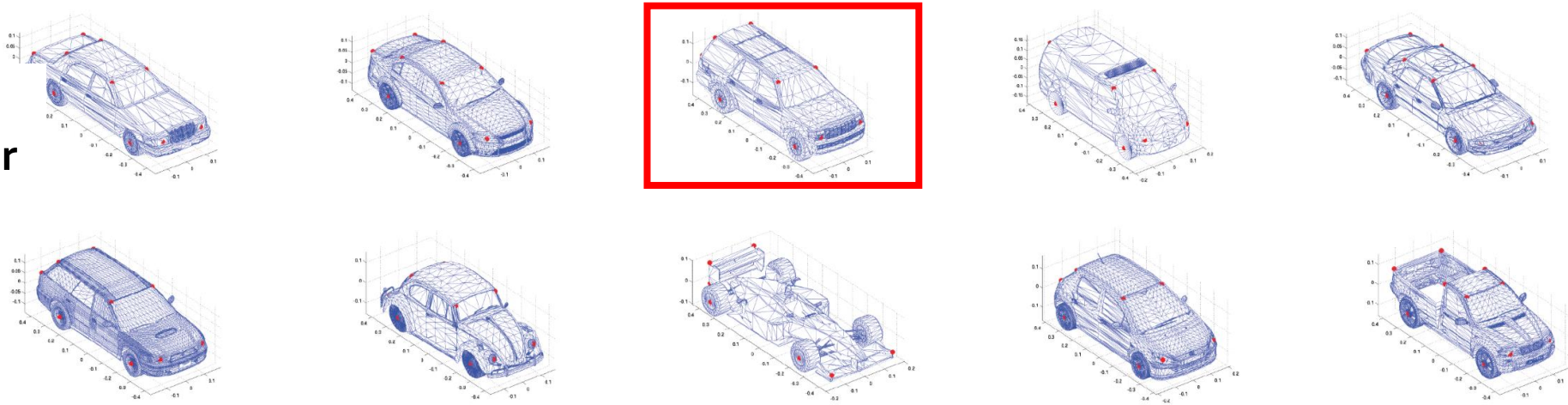
# However.. Accurate Label Acquisition is Expensive

**Step1:**
**Choose similar model**

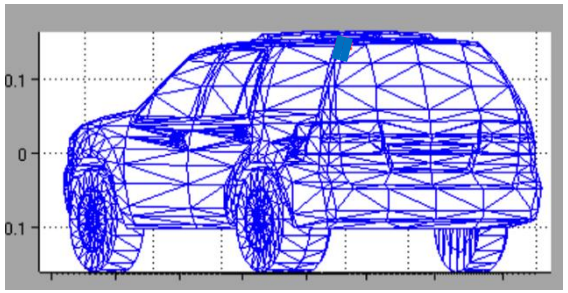












# However.. Accurate Label Acquisition is Expensive

**Step1:**
**Choose similar model**

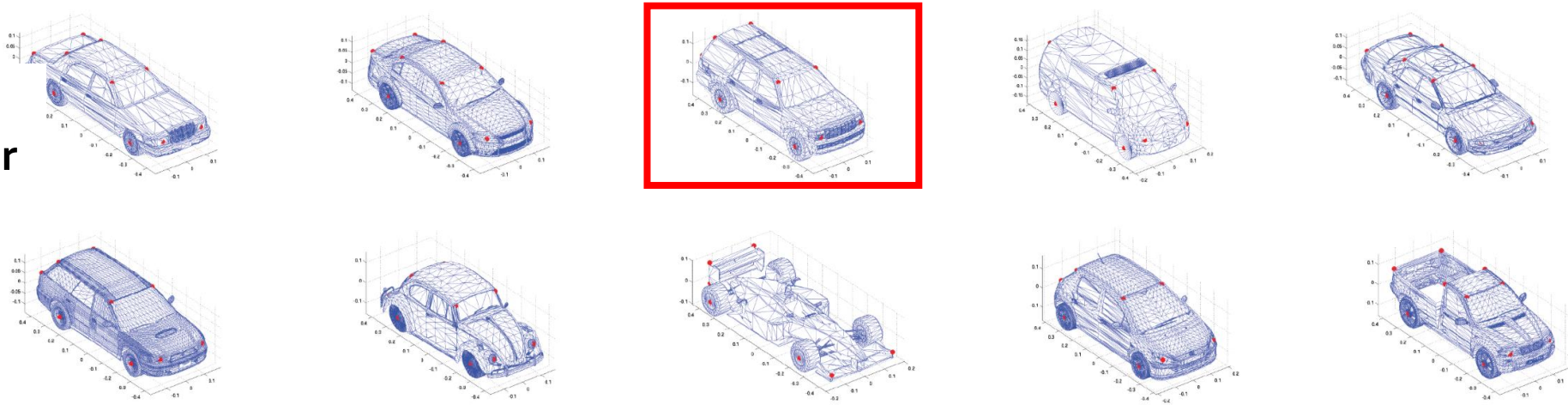


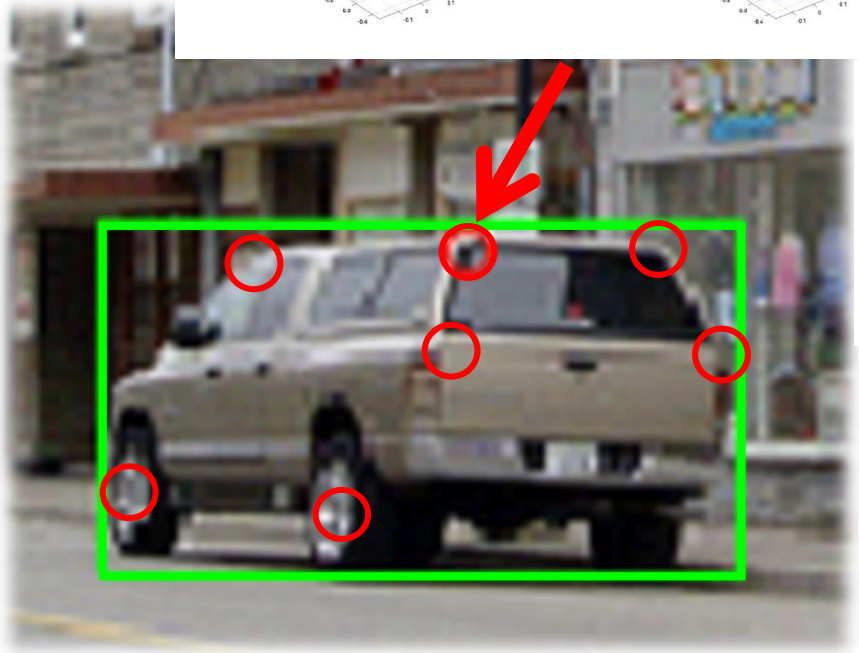
**Step2:**
**Coarse Viewpoint Labeling**



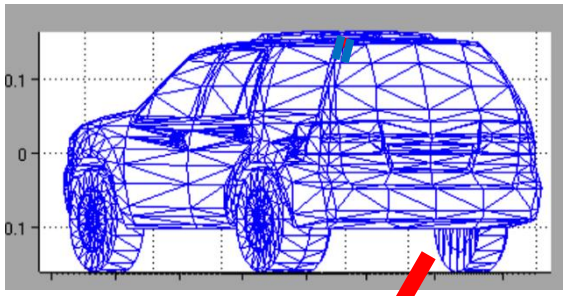
# However.. Accurate Label Acquisition is Expensive

**Step1:**
**Choose similar model**

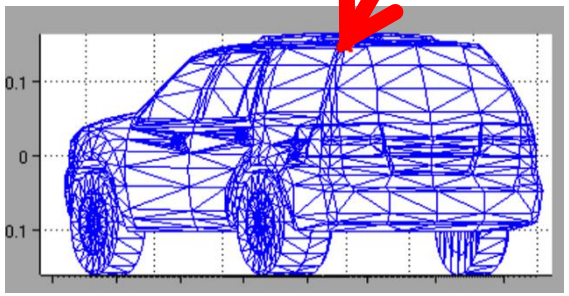




**Step2:**
**Coarse Viewpoint Labeling**

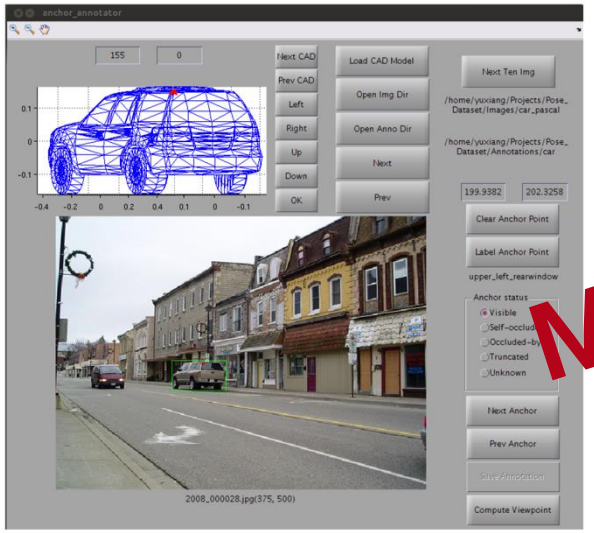


**Step3:**
**Label keypoints For alignment**



**Annotation takes ~1 min per object**

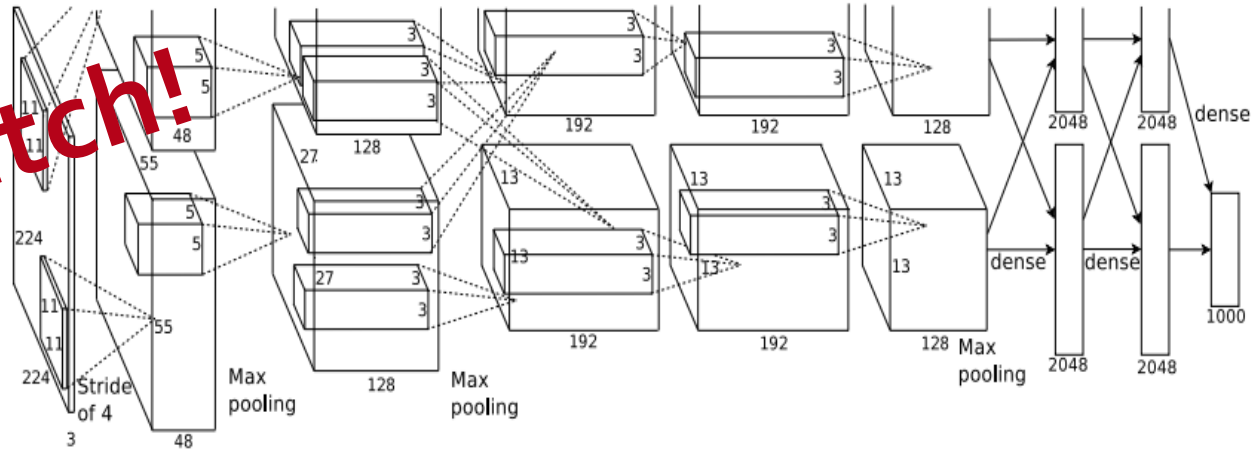
# High-cost Label Acquisition




**30K** images with viewpoint labels in PASCAL3D+ dataset [Xiang et al.]

Mismatch!

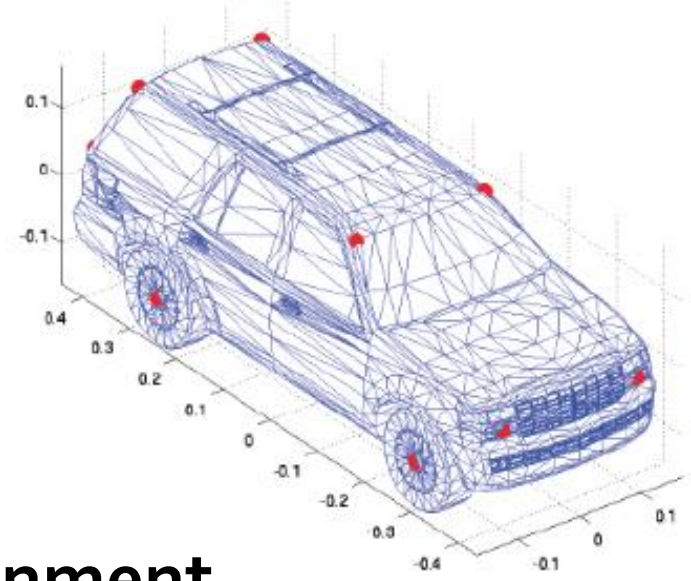
# High-capacity Model




**60M** parameters. AlexNet [Krizhevsky et al.]

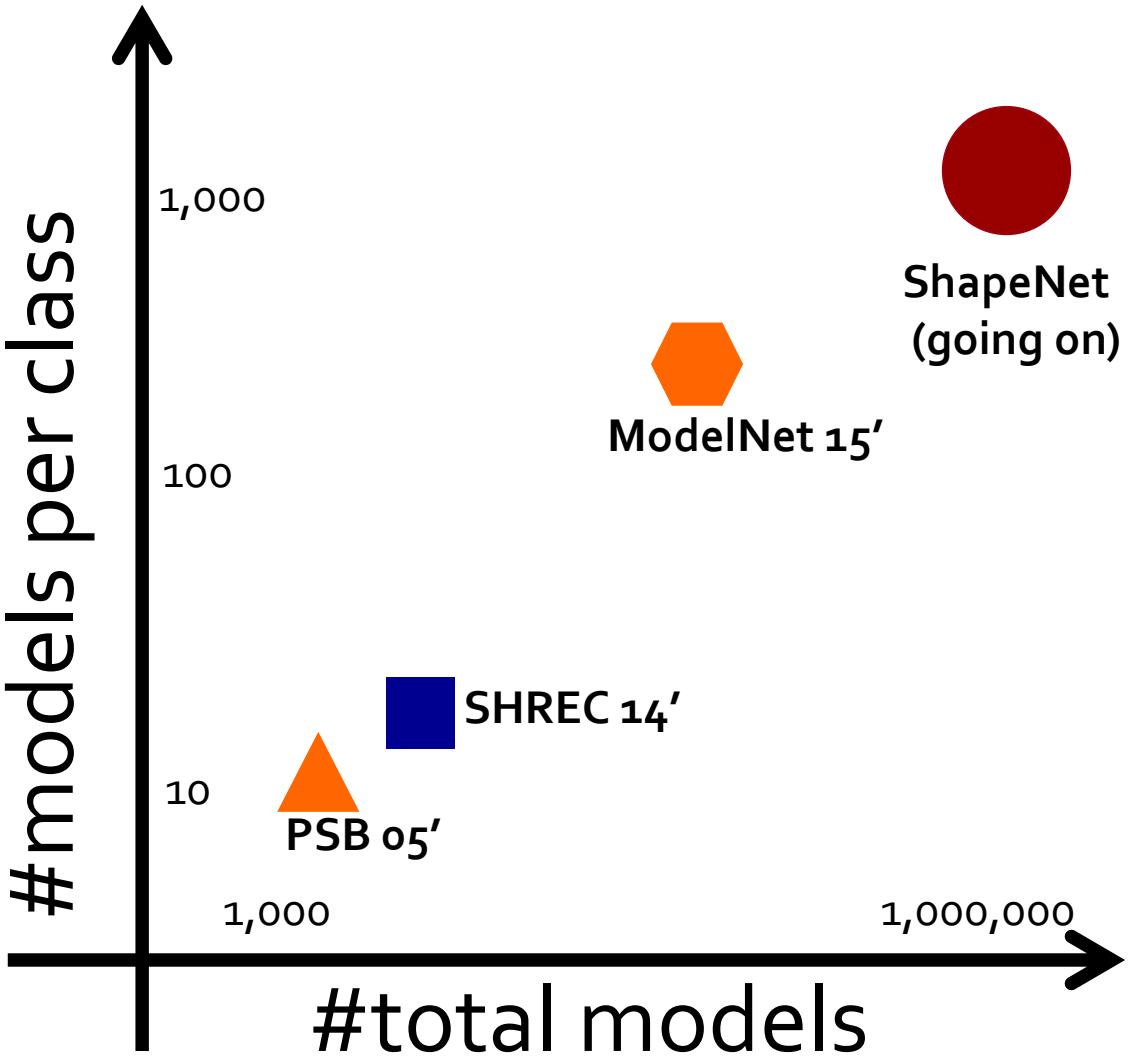
**How to get MORE images with ACCURATE viewpoint labels?**

**Manual** alignment
by annotators




**Auto** alignment
through rendering

# Good News: ShapeNet

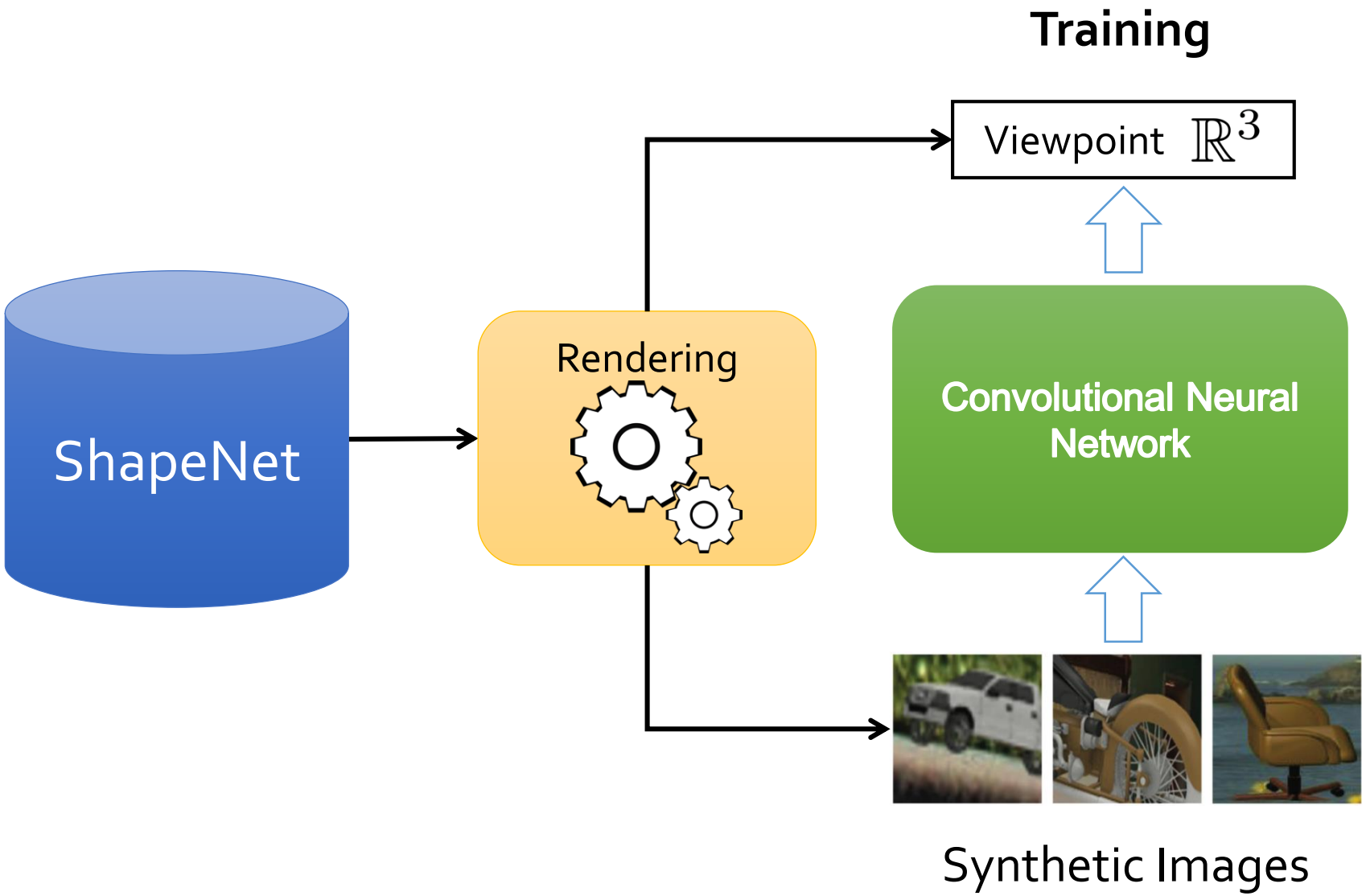



**3M** models in total
**330K** from **4K categories** annotated

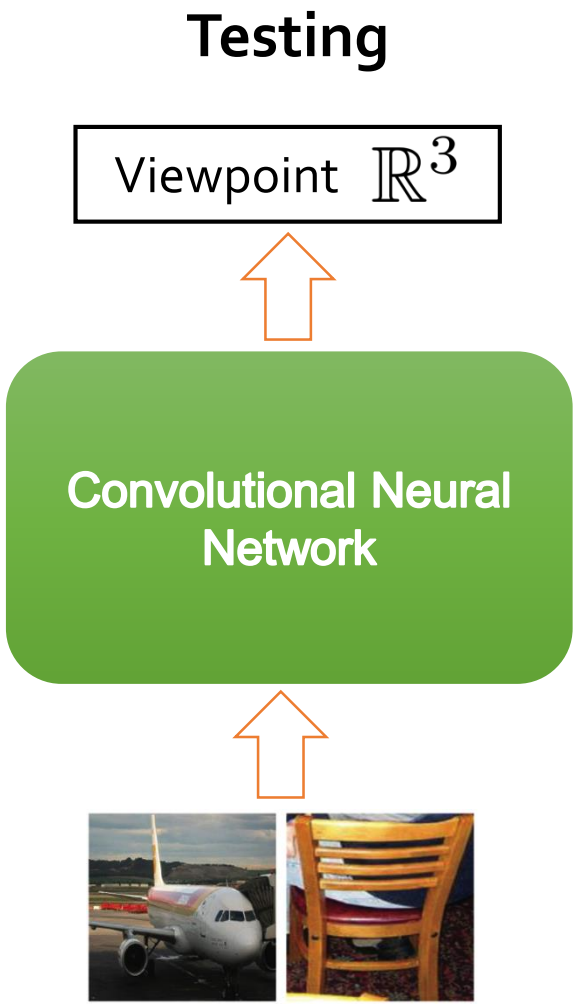
SHAPENET

**http://shapenet.cs.stanford.edu**

# Key Idea: Render for CNN



# Key Idea: Render for CNN



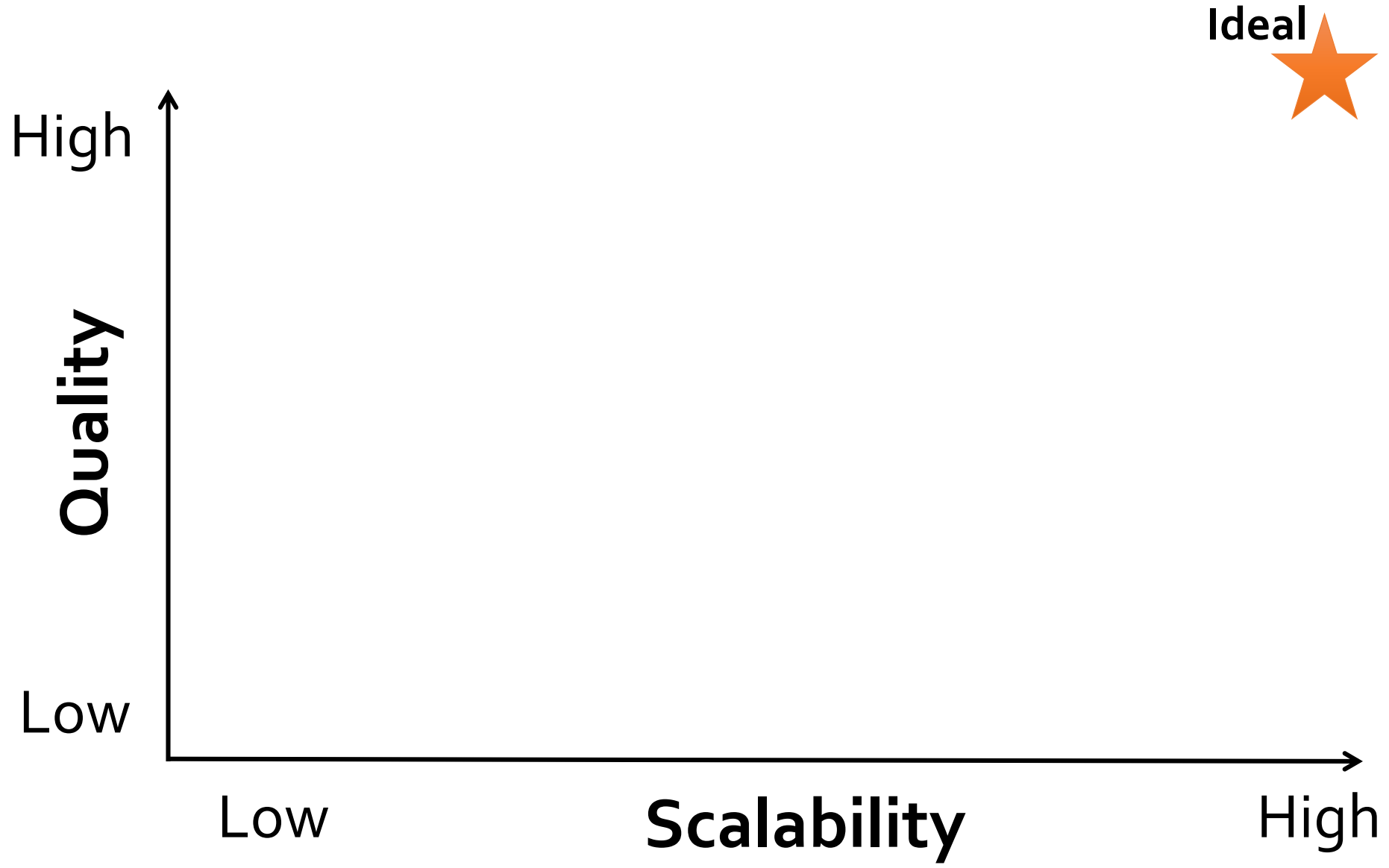
Convolutional Neural Network

I want data!

How to render data with both quantity and quality?

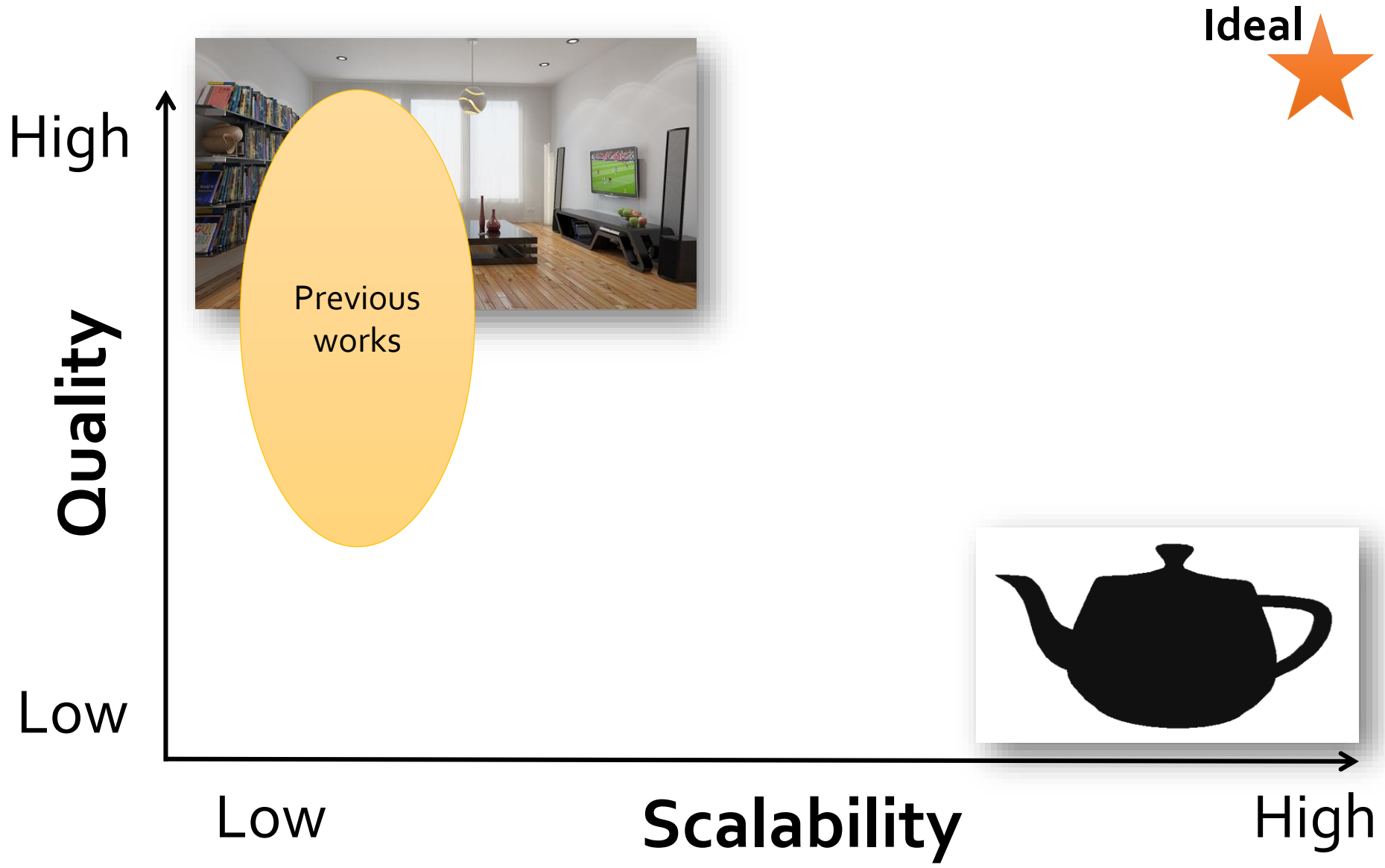
Rendering

# Synthesize: Scalability vs Quality

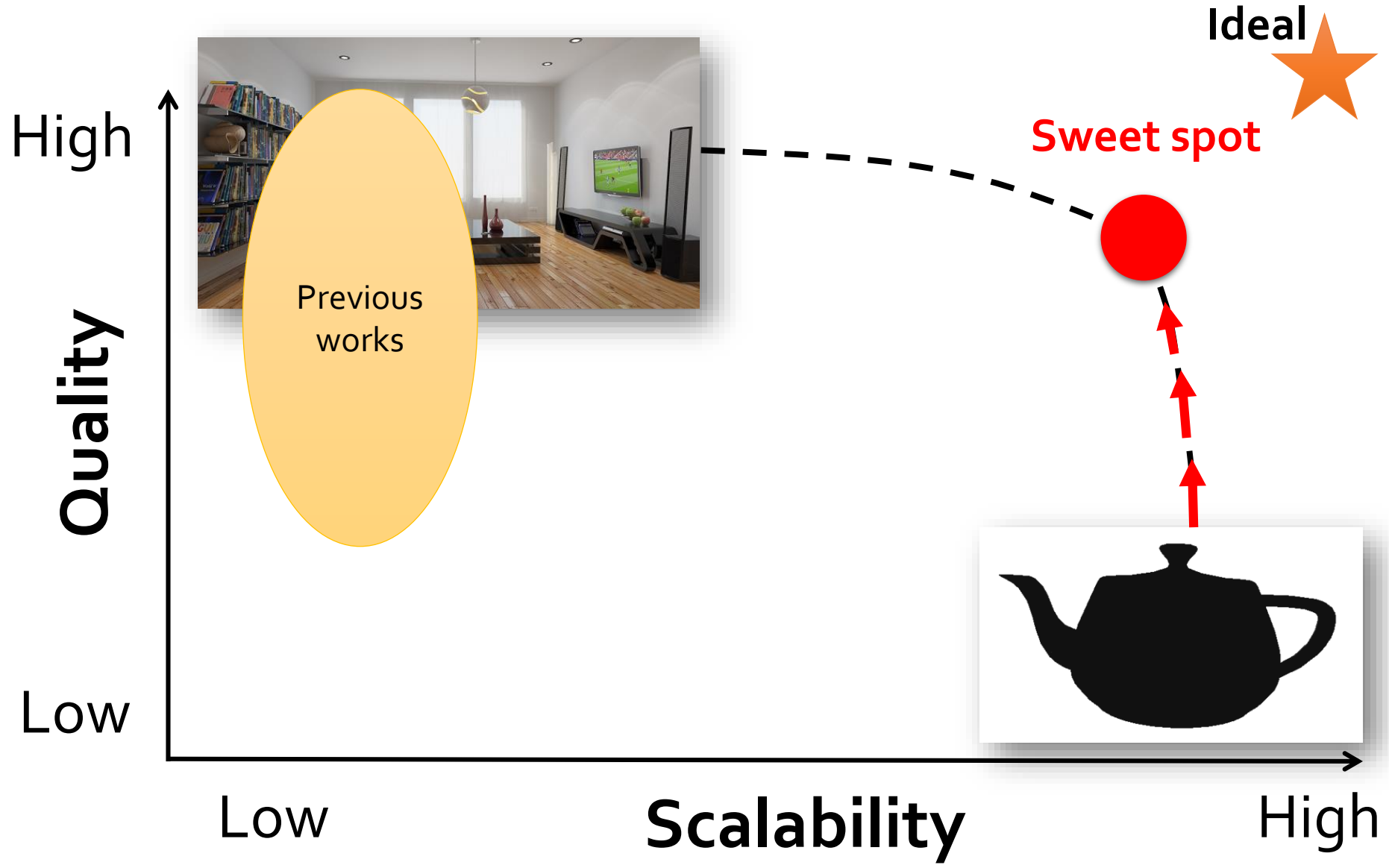


# Synthesize: Scalability vs Quality

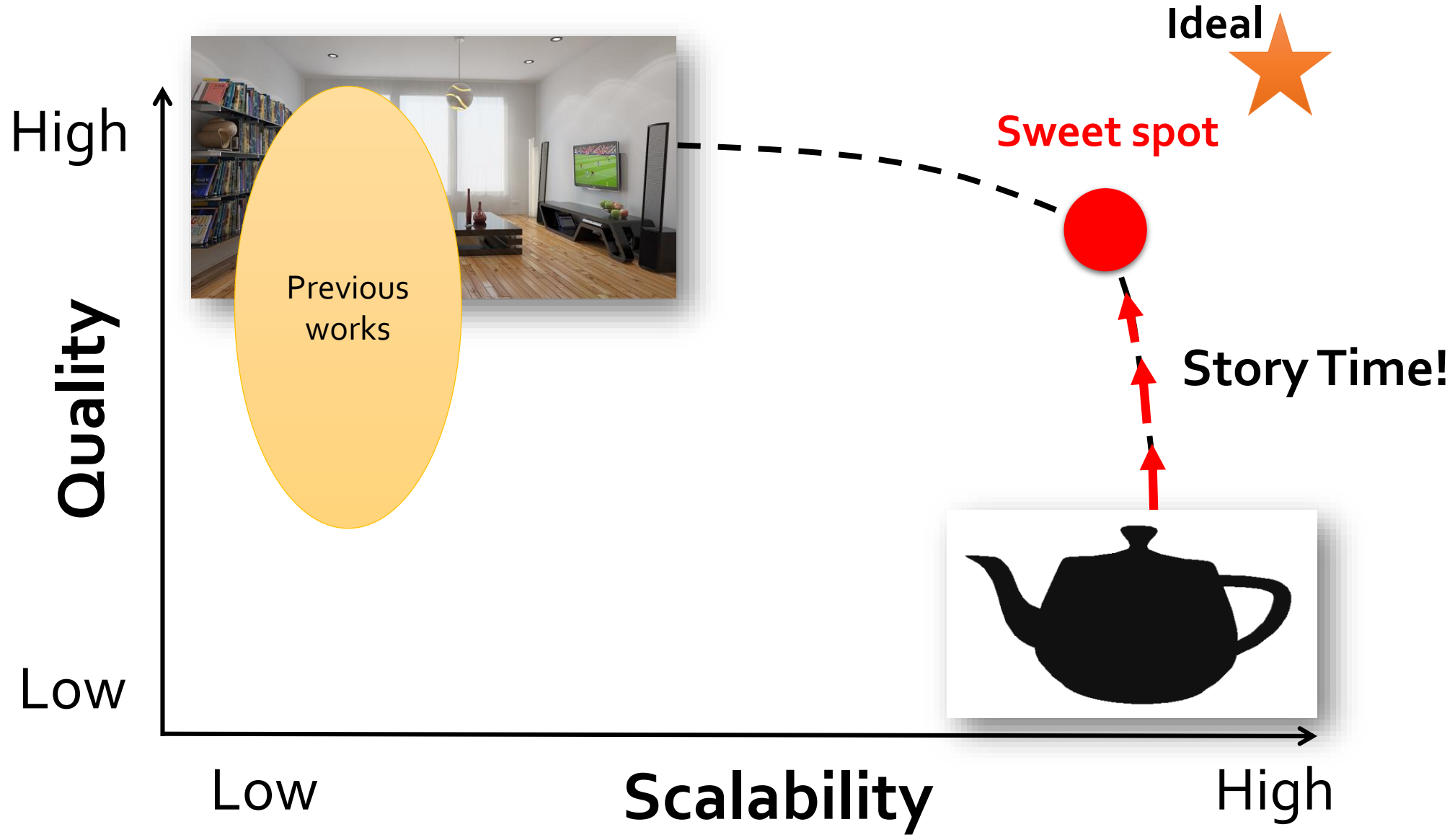
# Synthesize: Scalability vs Quality



# Synthesize: Scalability vs Quality



# Synthesize: Scalability vs Quality



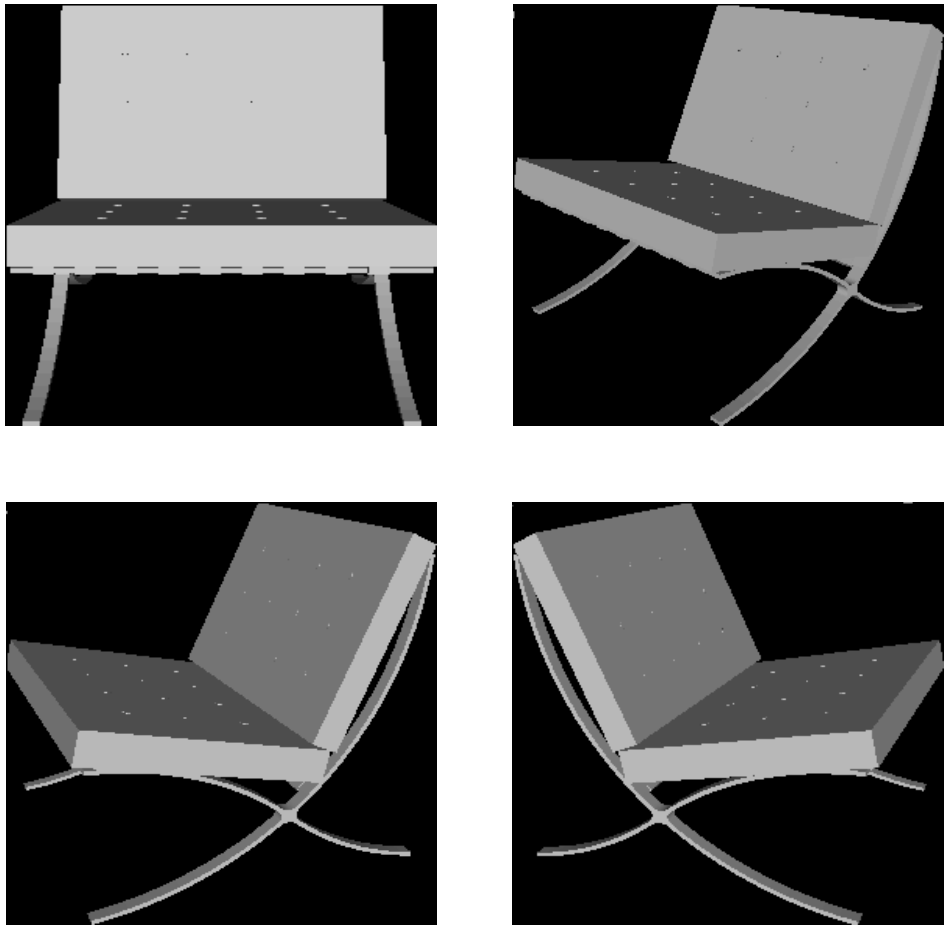
# A “Data Engineering” Journey

- 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



**95% on synthetic val set**

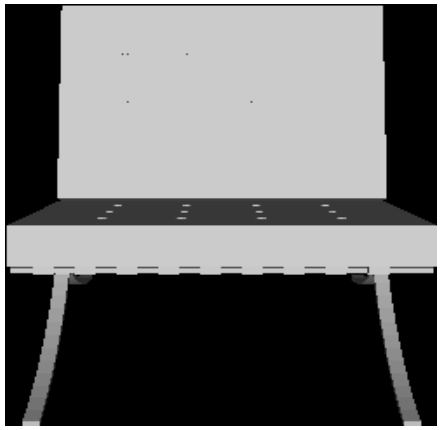
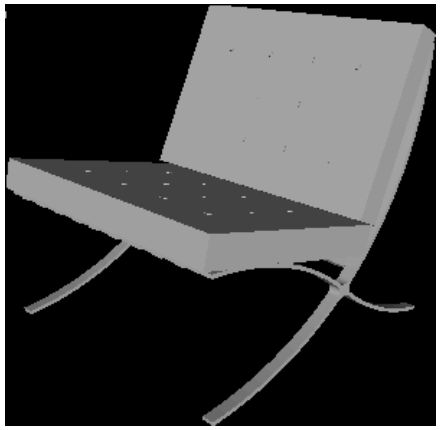
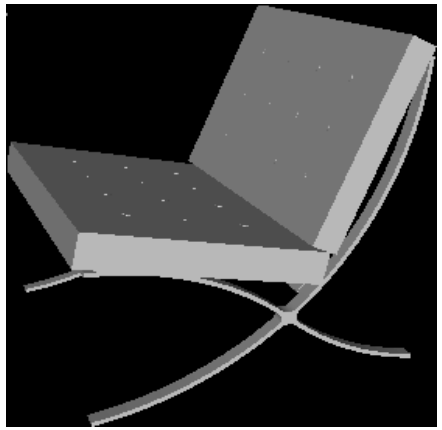
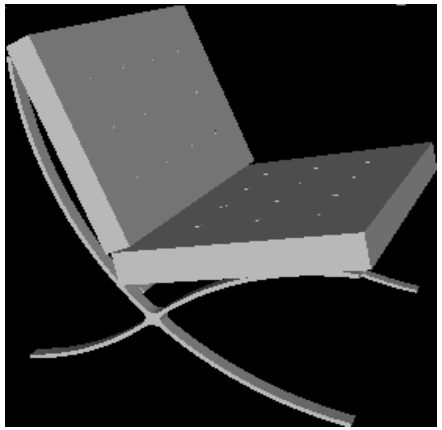
**47% on real test set ☹**

**ConvNet:**

Ah ha, I know!

Viewpoint is just the brightness pattern!

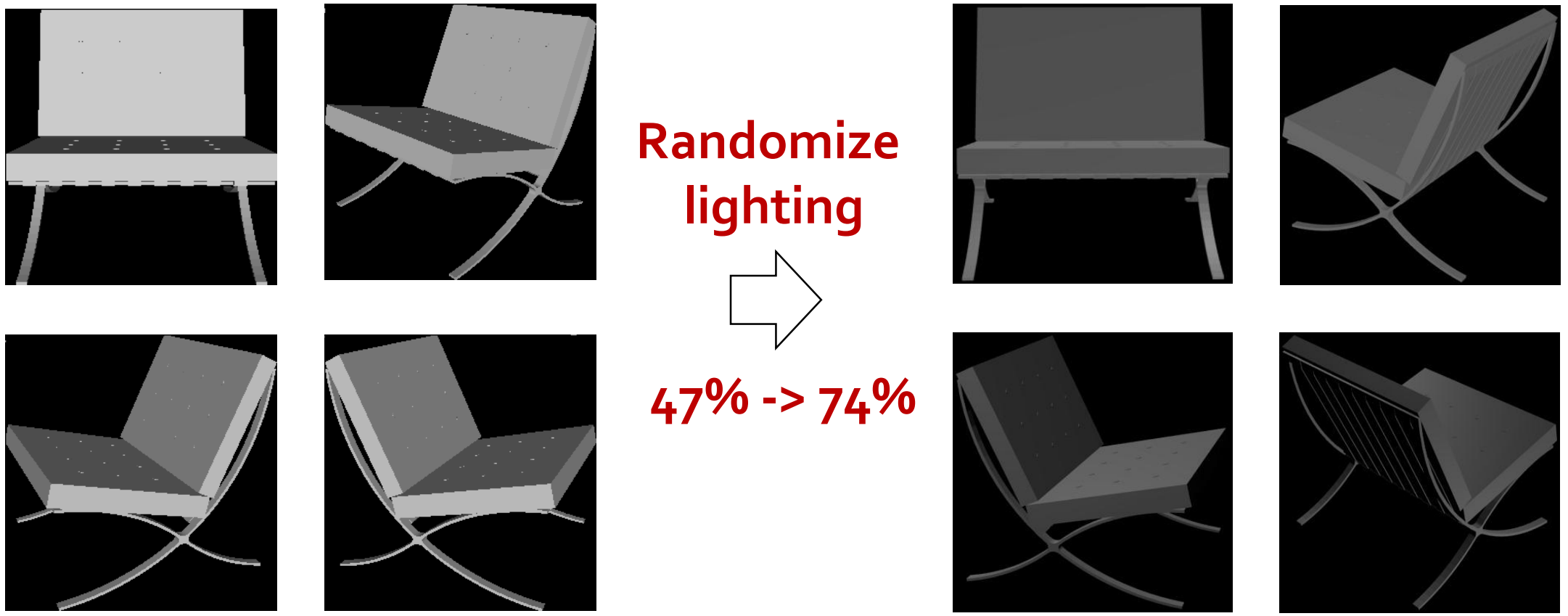
# A “Data Engineering” Journey

**95% on synthetic val set**

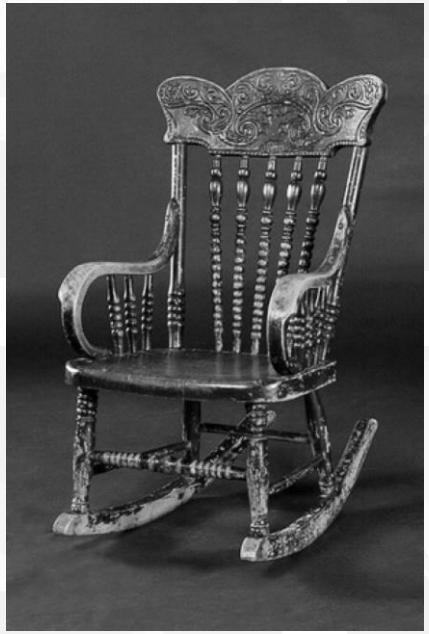
**47% on real test set ☹**

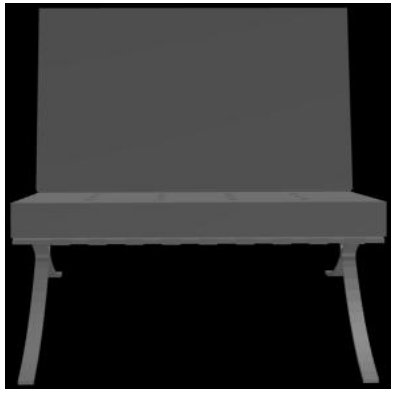
# A “Data Engineering” Journey

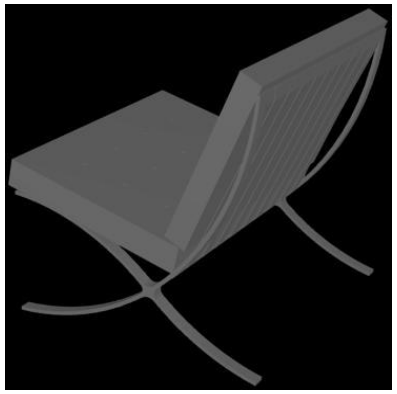


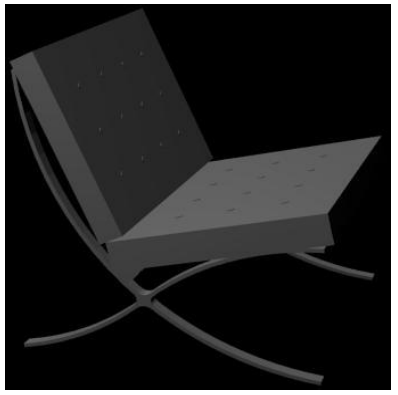

**ConvNet:** hmm.. viewpoint is not the brightness pattern. Maybe it’s the contour?

# A “Data Engineering” Journey



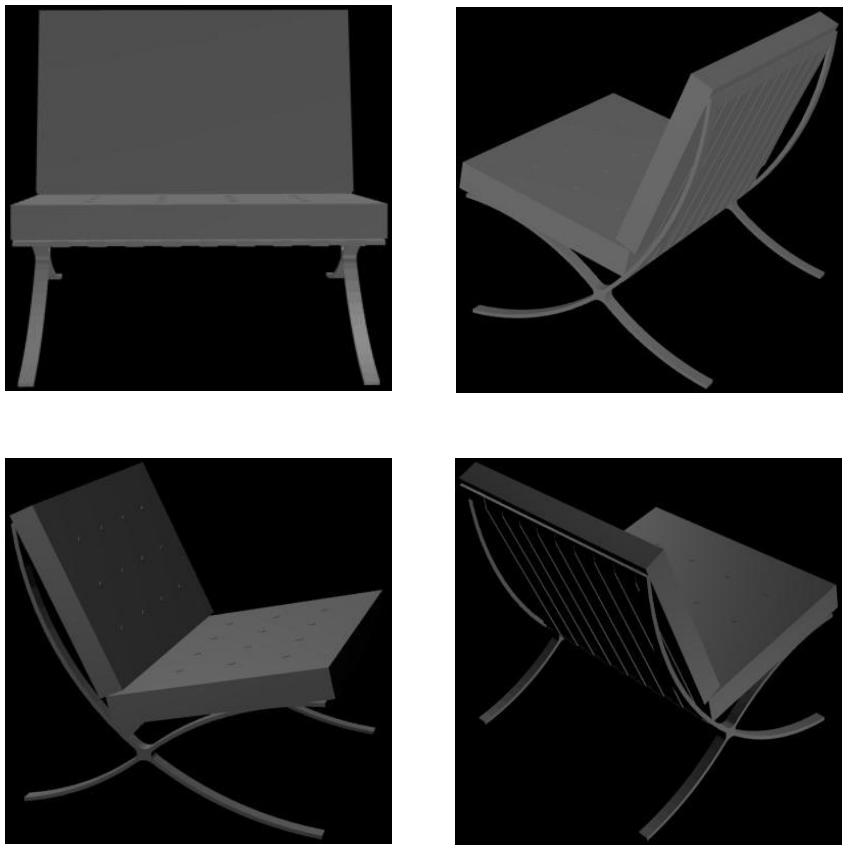






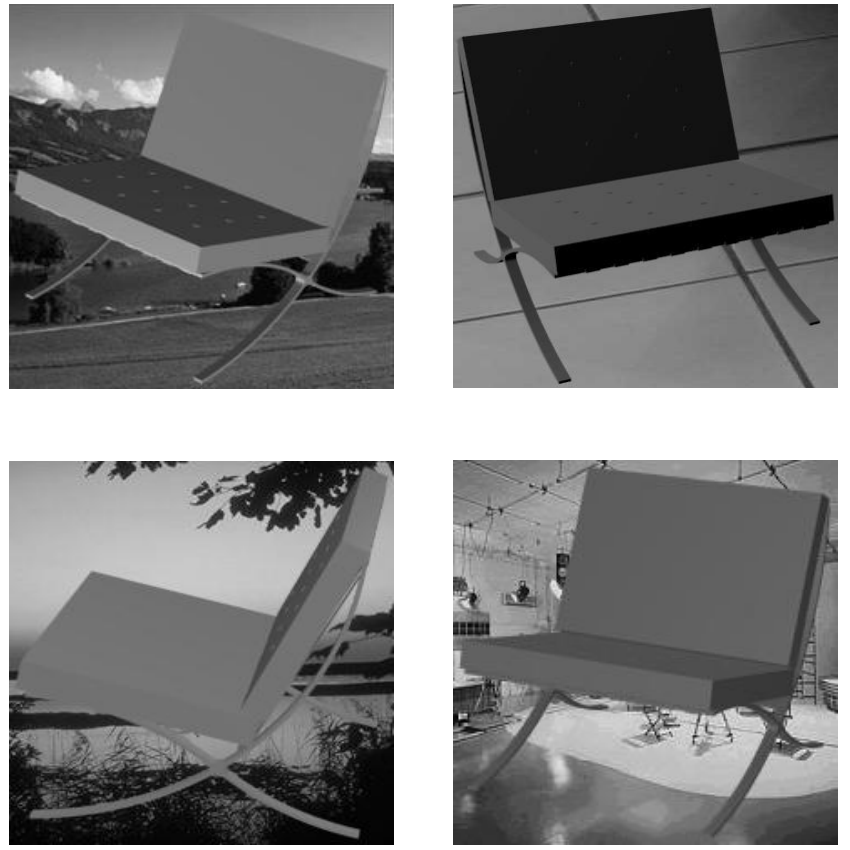
**ConvNet:** hmm.. viewpoint is not the brightness pattern. Maybe it’s the contour?

# A “Data Engineering” Journey



Add backgrounds

74% -> 86%



**ConvNet:** It becomes really hard! Let me look more into the picture.

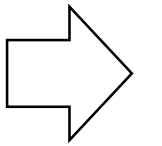
# A “Data Engineering” Journey

**bbox crop**
**texture**

**86% -> 93%**

# A “Data Engineering” Journey

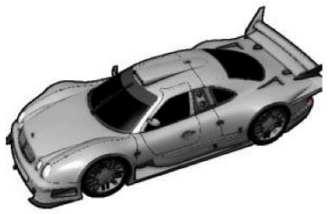
**bbox crop**
**texture**



**86% -> 93%**

**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**

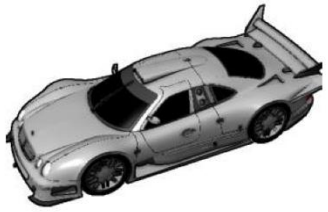
Sample bkg. Image Alpha-blending

**Crop**

Sample cropping params

**Hyper-parameters estimation from real images**

# Render for CNN Image Synthesis Pipeline



**3D model**

**Rendering**

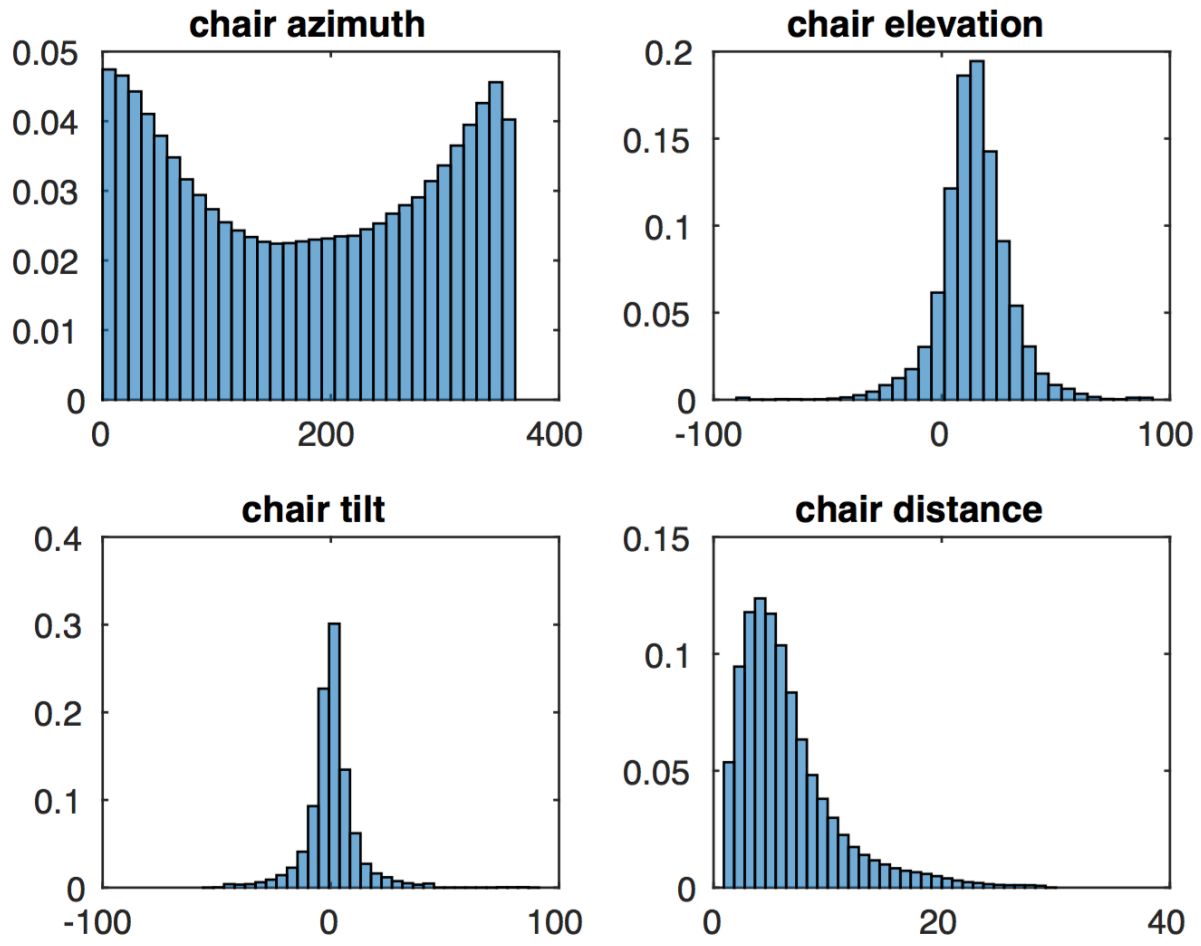
Sample lighting and camera params

# Rendering

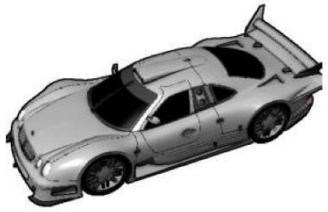
Lighting params
Randomly sampled

- 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**

Sample lighting and camera params

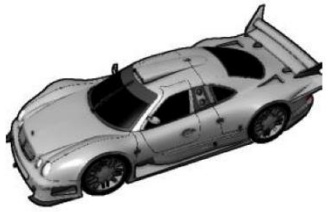
**Add bkg**

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

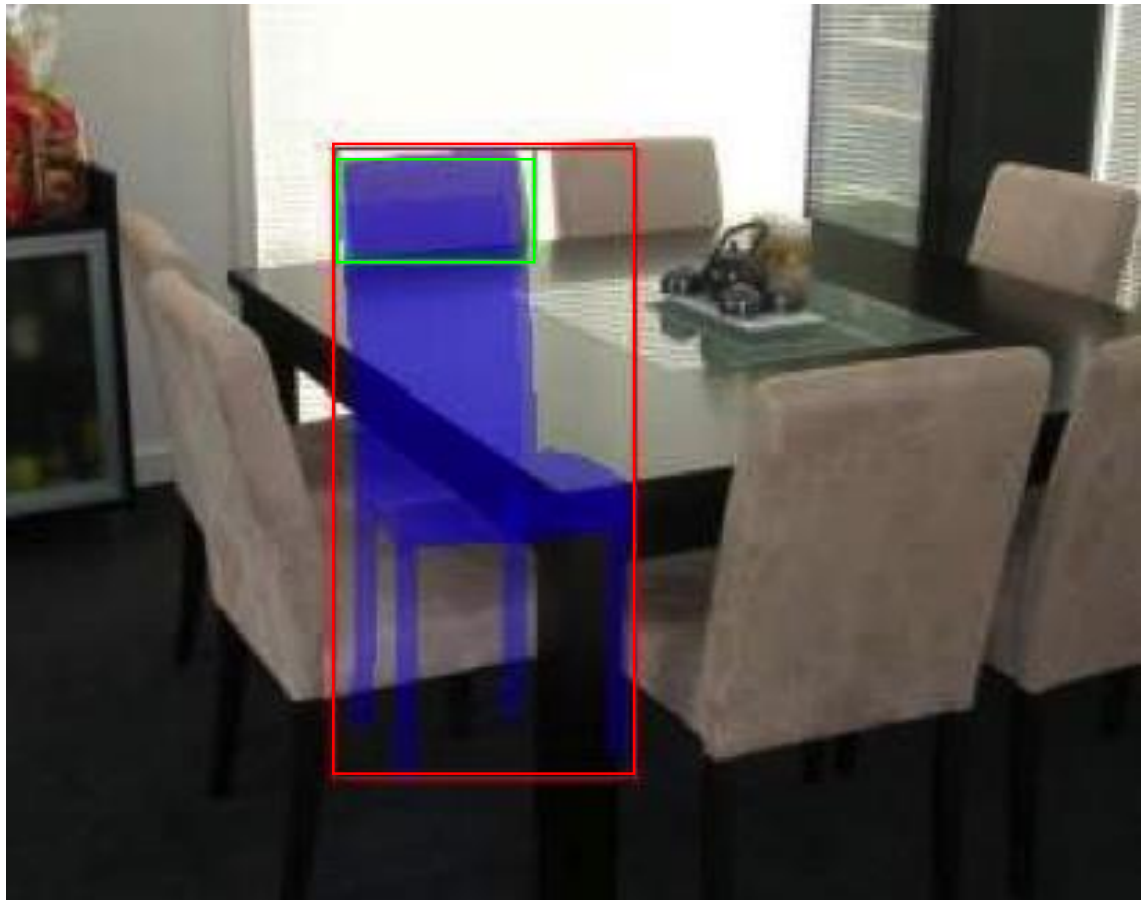
Sample bkg. Image Alpha-blending

## Crop

Sample cropping params

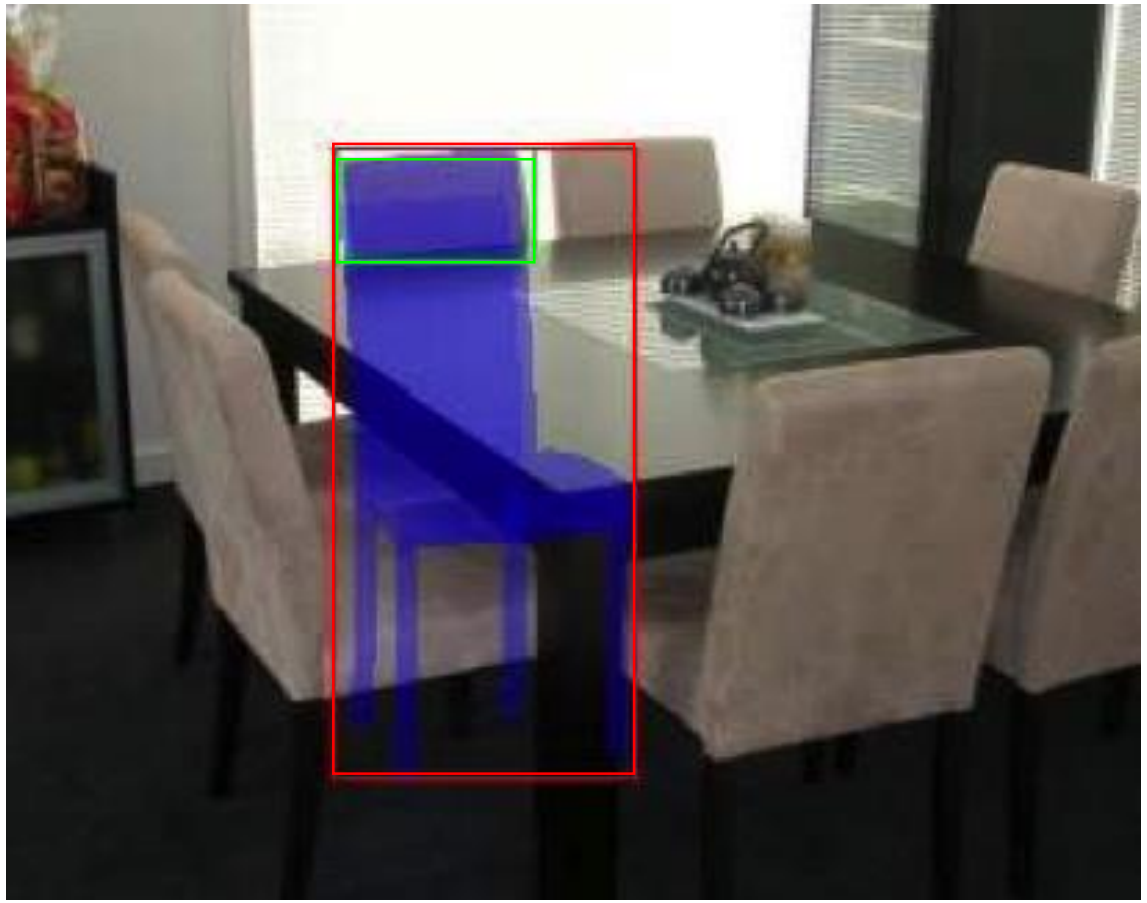
# Image Cropping

Cropping patterns KDE from PASCAL3D+ train set



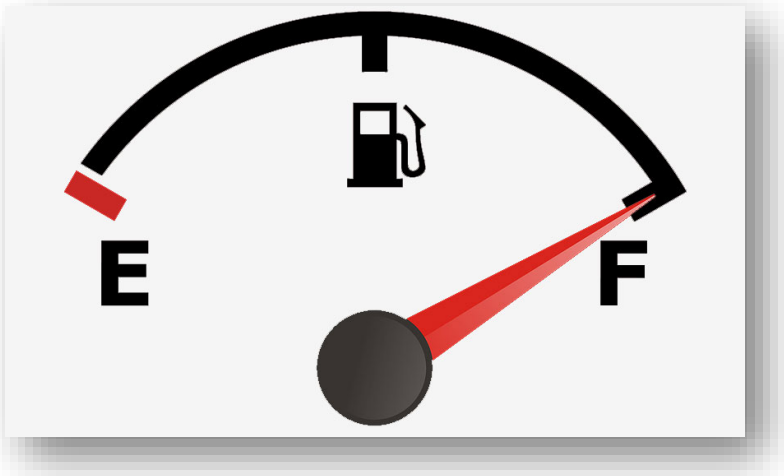
# Image Cropping

Cropping patterns KDE from PASCAL3D+ train set



# 2.4M Synthesized Images for 12 Categories

- High scalability
- High quality
  - Overfit-resistant
  - Accurate labels




# Results

# 3D Viewpoint Estimation Evaluation

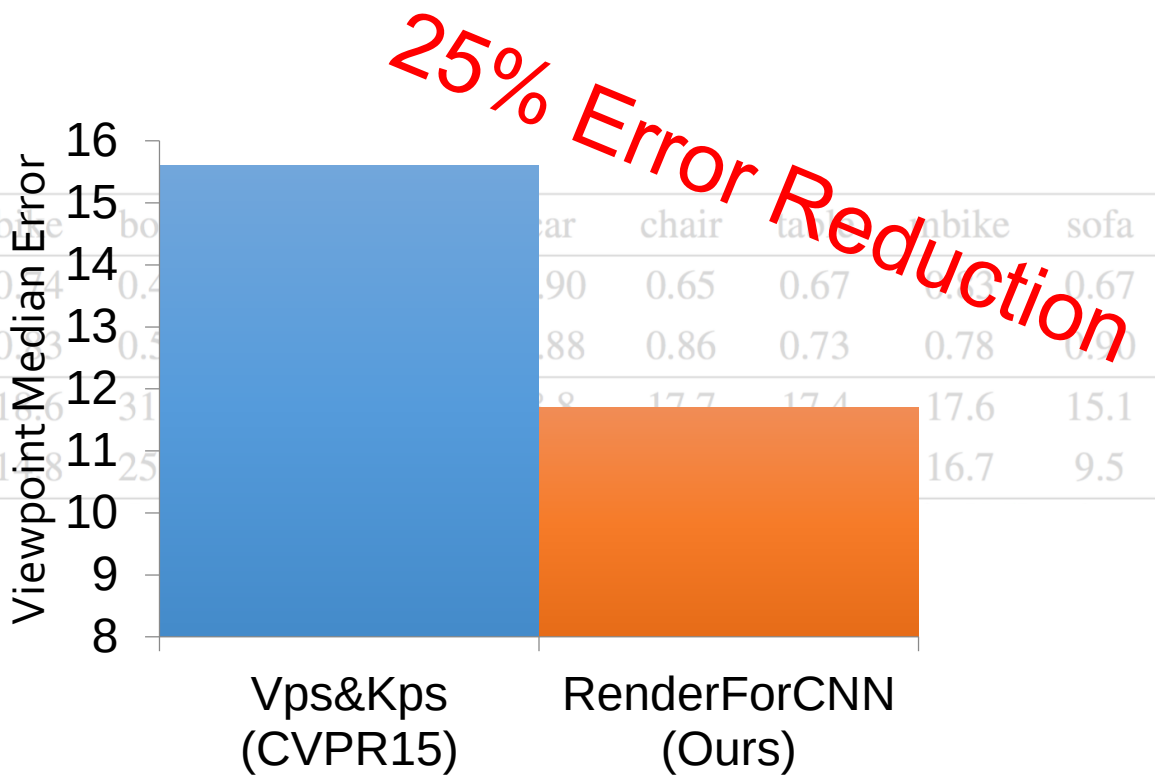
Metric: **median angle error** (lower the better)

**Real test images** from PASCAL3D+ dataset

| | aero | bike | boat | bottle | bus | car | chair | table | mbike | sofa | train | tv | mean |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $Acc_{\frac{\pi}{6}}$ (Tulsiani, Malik) | 0.78 | 0.74 | 0.49 | 0.93 | 0.94 | 0.90 | 0.65 | 0.67 | 0.83 | 0.67 | 0.79 | 0.76 | 0.76 |
| $Acc_{\frac{\pi}{6}}$ (Ours-Render) | 0.74 | 0.83 | 0.52 | 0.91 | 0.91 | 0.88 | 0.86 | 0.73 | 0.78 | 0.90 | 0.86 | 0.92 | 0.82 |
| $MedErr$ (Tulsiani, Malik) | 14.7 | 18.6 | 31.2 | 13.5 | 6.3 | 8.8 | 17.7 | 17.4 | 17.6 | 15.1 | 8.9 | 17.8 | 15.6 |
| $MedErr$ (Ours-Render) | 15.4 | 14.8 | 25.6 | 9.3 | 3.6 | 6.0 | 9.7 | 10.8 | 16.7 | 9.5 | 6.1 | 12.6 | 11.7 |

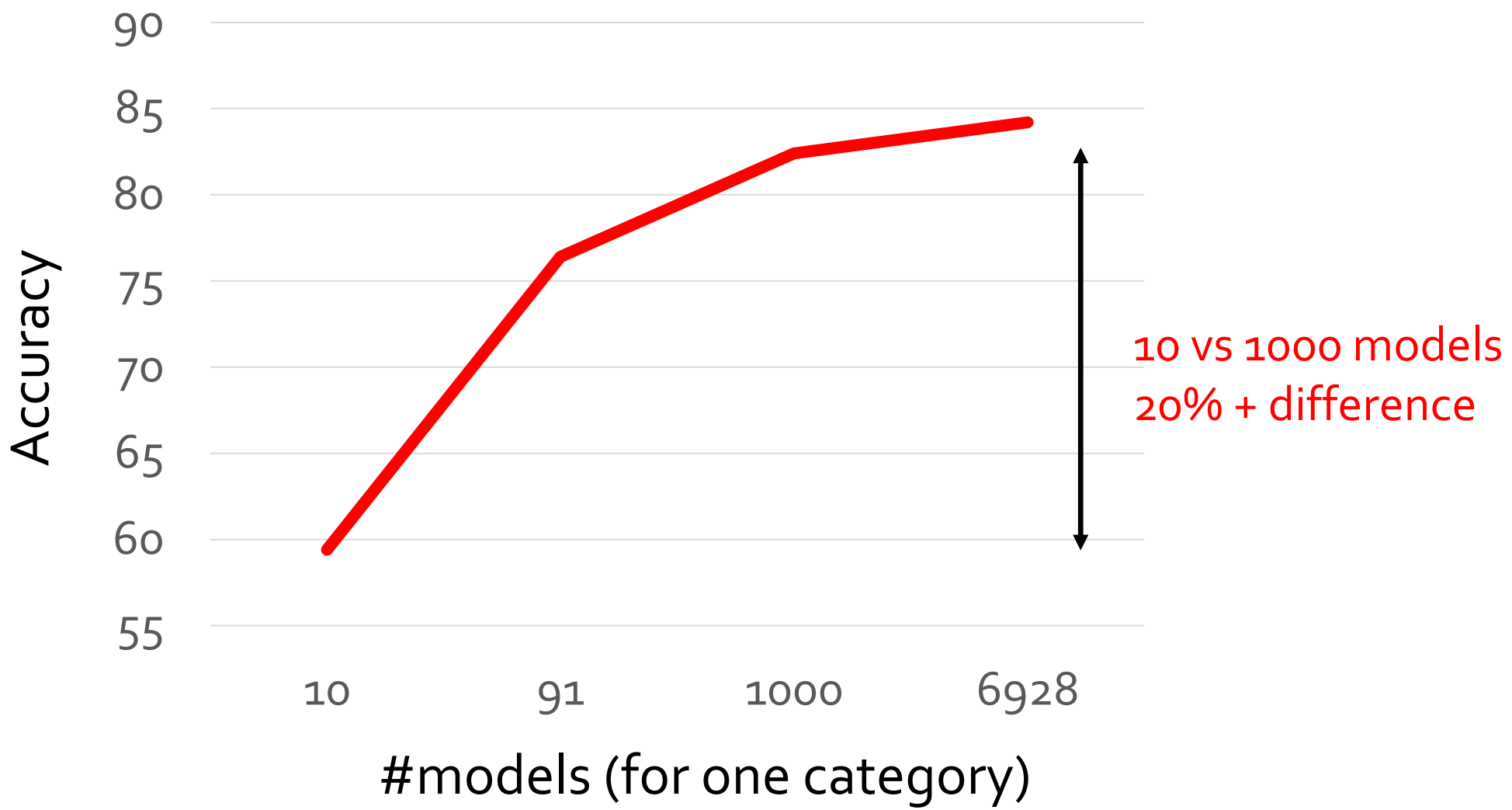
# 3D Viewpoint Estimation Evaluation

Our model **trained on rendered images** outperforms state-of-the-art model **trained on real images** in PASCAL3D+.




| | aero | bike | bo | | chair | tab | mbike | sofa | train | tv | mean |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $Acc_{\frac{\pi}{6}}$ (Tulsiani, Malik) | 0.78 | 0.4 | 0.4 | .90 | 0.65 | 0.67 | 0.7 | 0.67 | 0.79 | 0.76 | 0.76 |
| $Acc_{\frac{\pi}{6}}$ (Ours-Render) | 0.74 | 0.3 | 0.5 | .88 | 0.86 | 0.73 | 0.78 | 0.90 | 0.86 | 0.92 | 0.82 |
| $MedErr$ (Tulsiani, Malik) | 14.7 | 15.6 | 31 | 8 | 17.7 | 17.4 | 17.6 | 15.1 | 8.9 | 17.8 | 15.6 |
| $MedErr$ (Ours-Render) | 15.4 | 15.8 | 25 | | | | 16.7 | 9.5 | 6.1 | 12.6 | 11.7 |

# How many 3D models are necessary?



# 3D Viewpoint Estimation

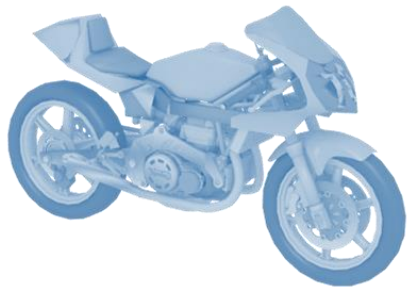




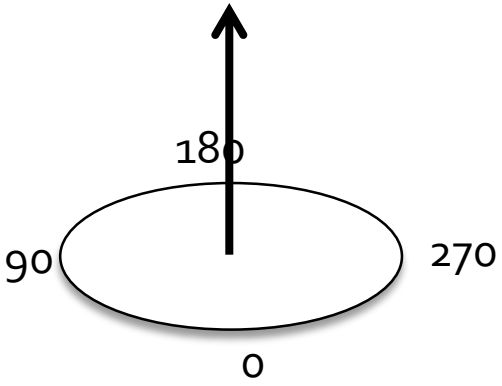






# Azimuth Viewpoint Estimation

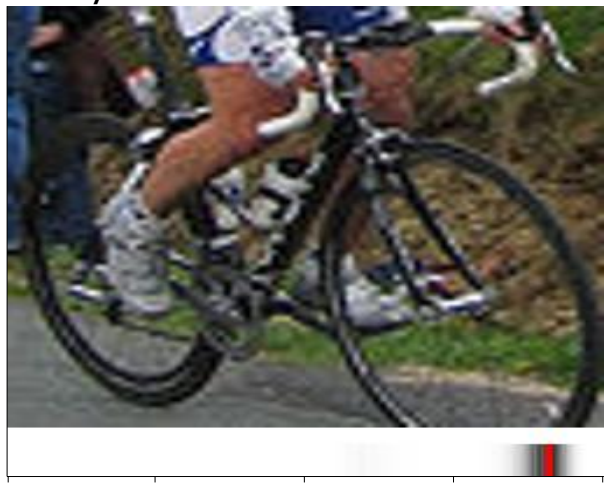



Ground truth view

Estimated view confidence

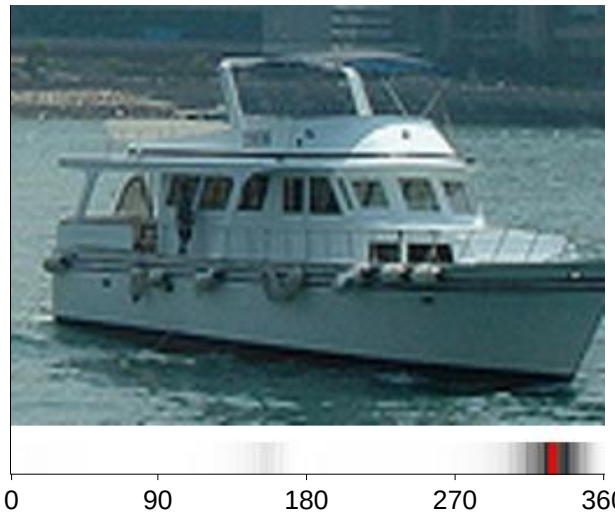
airplane


bicycle




bicycle

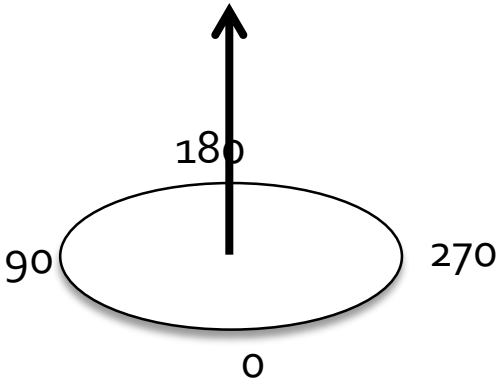

boat




car


motorbike

# Azimuth Viewpoint Estimation




Ground truth view

Estimated view confidence

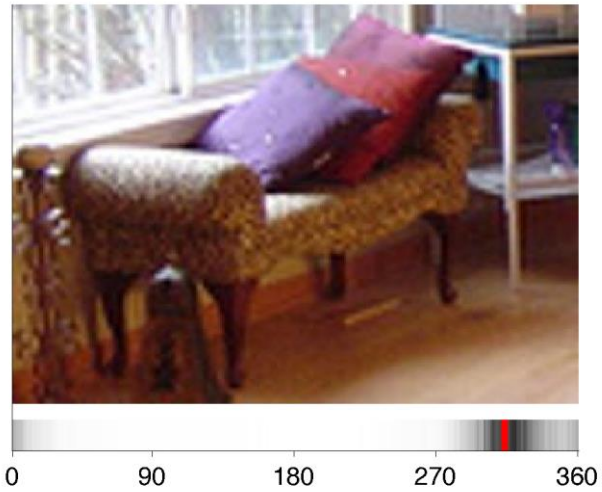
table


chair


chair


monitor

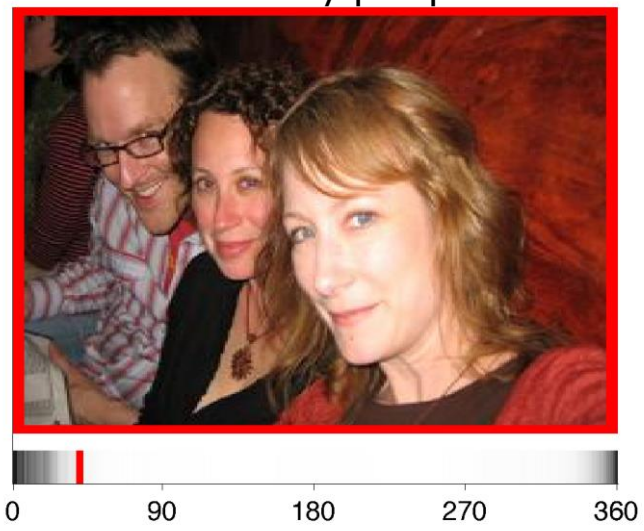

sofa



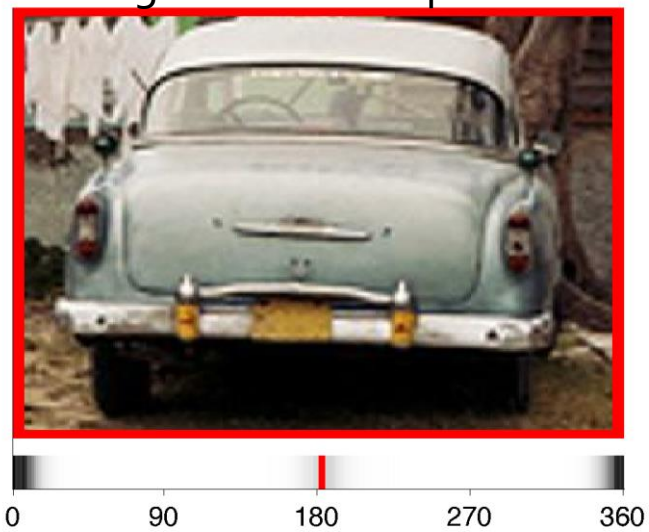

sofa

# Failure Cases

sofa occluded by people



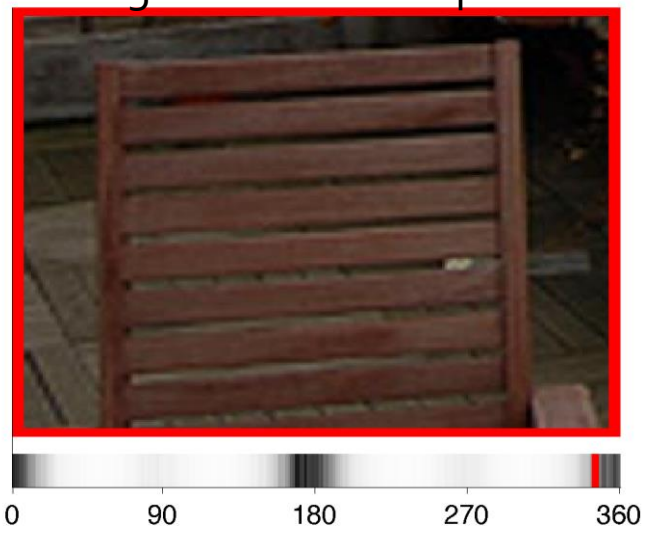

ambiguous car viewpoint




multiple cars


car occluded by motorbike


ambiguous chair viewpoint




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

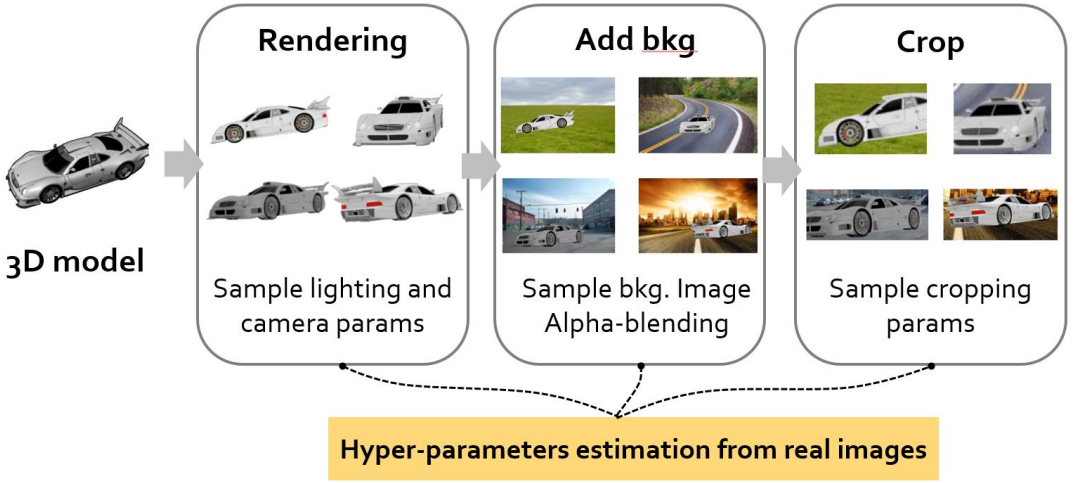
**Images rendered from 3D models can be effectively used to train CNNs, especially for 3D tasks. State-of-the-art result has been achieved.**

**Keys to success**

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

SHAPENET

http://shapenet.cs.stanford.edu



# THE END

# THANK YOU!