Deep Supervision with Shape Concepts for Occlusion-Aware 3D Object Parsing

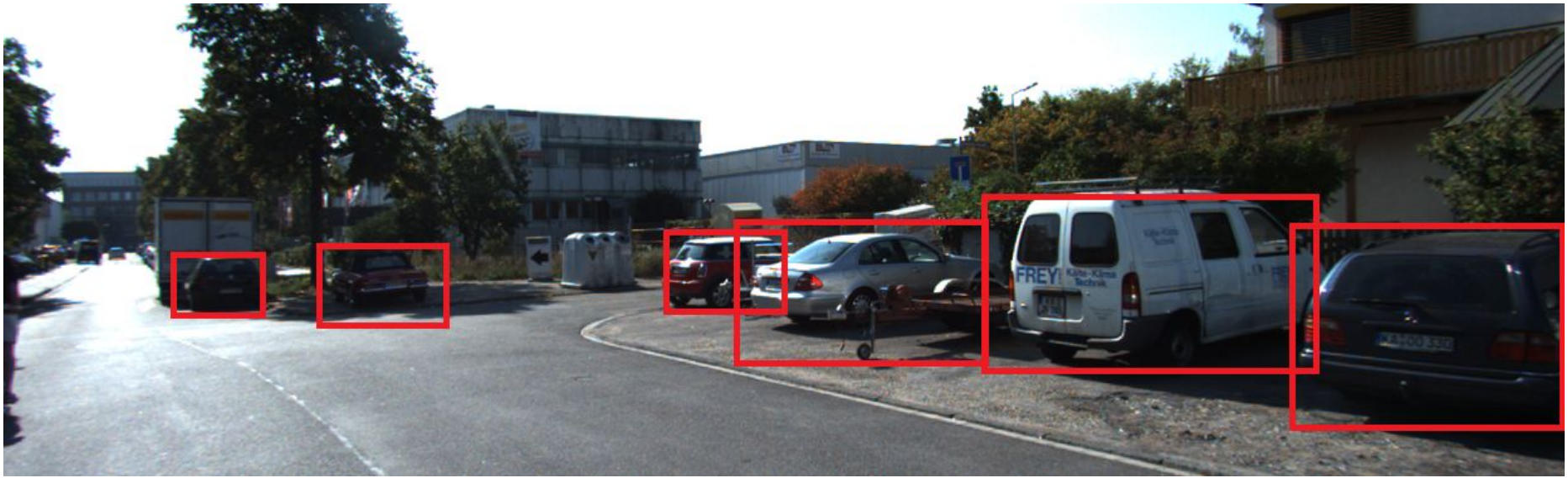
Chi Li, **Zeeshan Zia**, Quoc-Huy Tran, Xiang Yu, Gregory D. Hager, Manmohan Chandraker

zeeshan.zia@outlook.com
www.ZeeshanZia.com

CVPR 2017 work presented at Deep Learning Study Group
Hacker Dojo, Santa Clara
27 February 2017

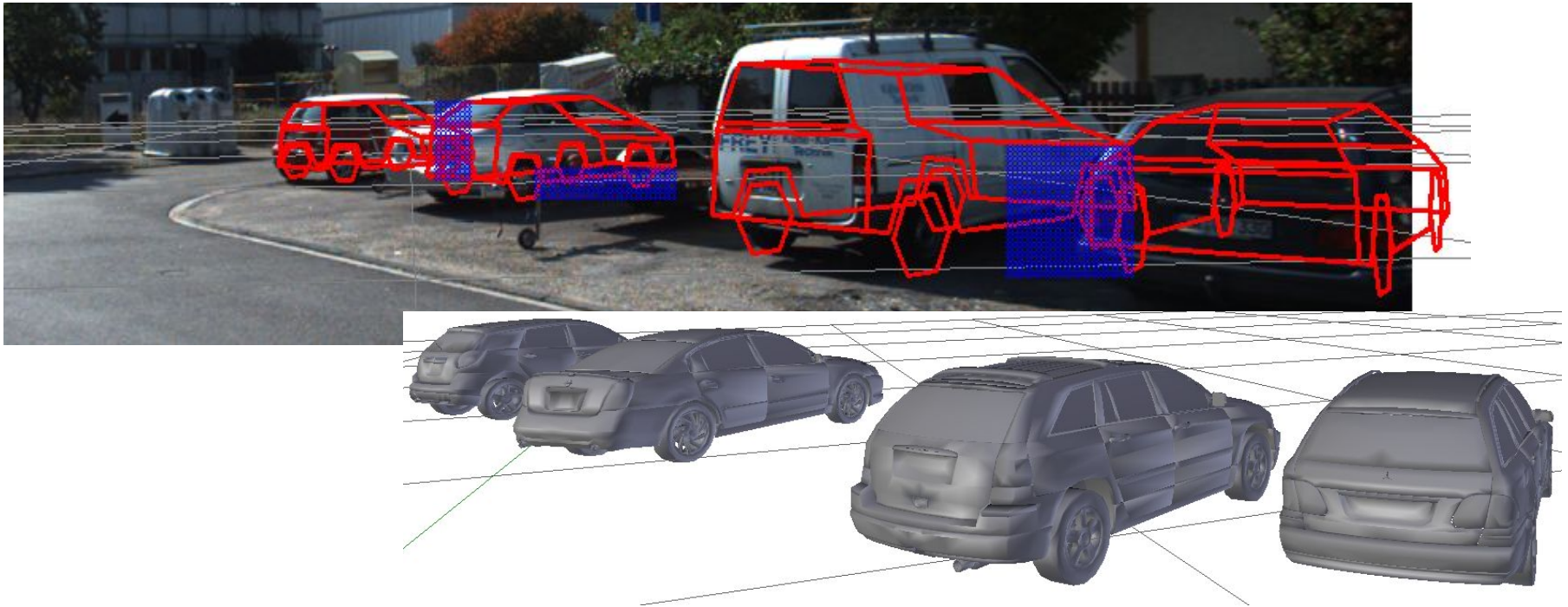
Objective

- Current object models: coarse estimates



Objective

- Finer models to aid scene-level reasoning



Challenges

- Real-world images

- Training data
- Occlusion, truncation
- Processing time

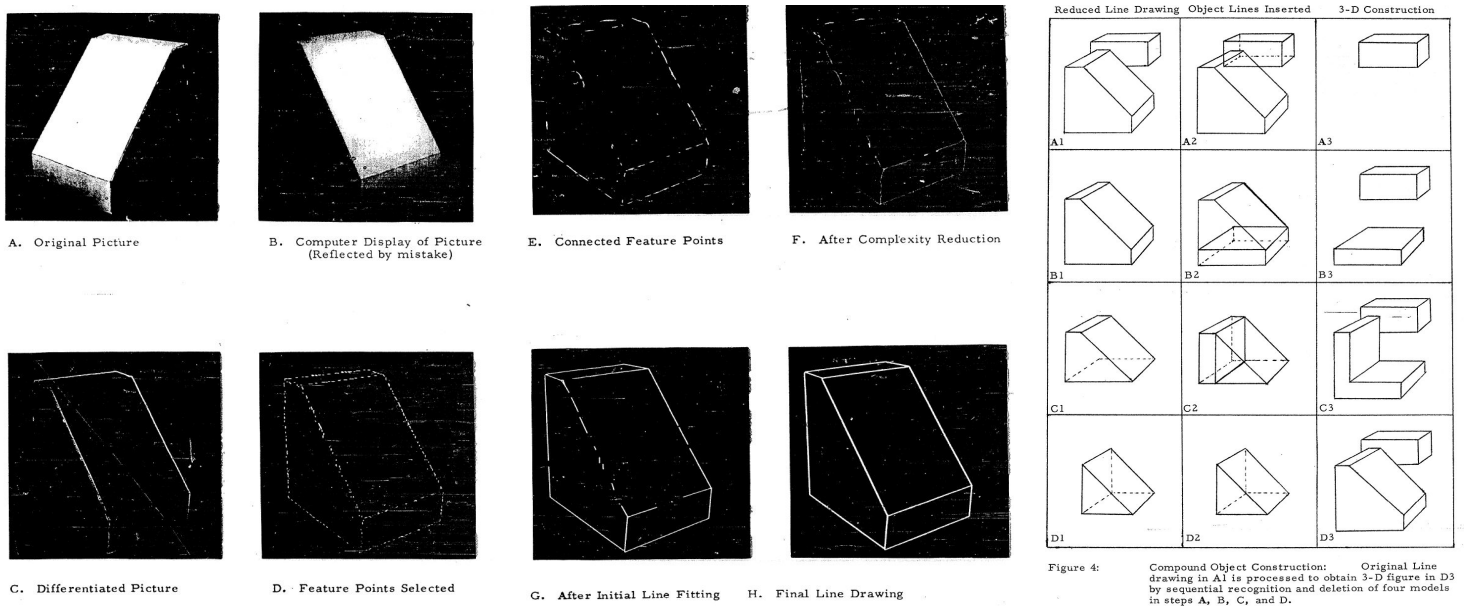
Literature

Three major components in past approaches:

- Geometric model
- Appearance model
- Inference

Geometric model

Machine Perception of Three-Dimensional Solids, L.G. Roberts, PhD Thesis, MIT, 1963



Geometric model

Symbolic Reasoning Among 3D Models and 2D Images, Rodney Brooks, PhD Thesis, MIT, 1981

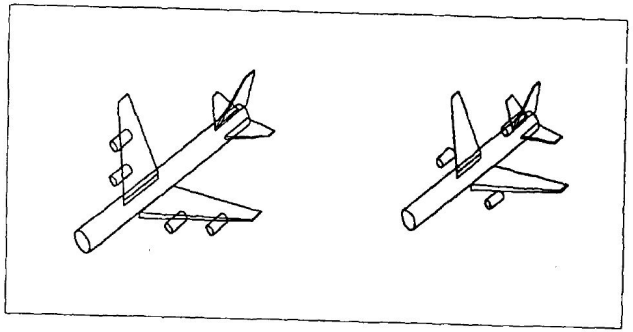
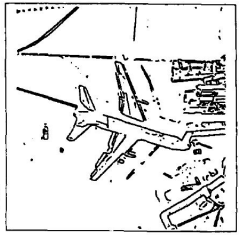
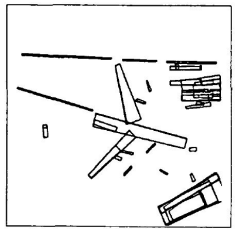


Fig. 9.1: Instances of class models of Boeing-747s and L-1011s.





```
ENG-DISP-GAP ε [6, 10]
ENG-DISP ε [0, 4]
ENG-GAP ε [7, 10]
STAB-ATTACH ε [3, 5]
R-ENG-ATTACHMENT ε [3, 6]
ENG-OUT ε [5, 12]
WING-ATTACHMENT ε [20, 40]
        WING-ATTACHMENT ≥ 0.4*FUSELAGE-LENGTH
        WING-ATTACHMENT ≤ 0.6*FUSELAGE-LENGTH
STAB-RATIO ε [0.2, 0.55]
STAB-SWEEP-BACK ε [3, 7]
STAB-LENGTH ε [7.6, 13]
STAB-THICK ε [0.7, 1.1]
STAB-WIDTH ε [5, 11]
RUDDER-RATIO ε [0.3, 0.4]
RUDDER-SWEEP-BACK ε [3, 9]
RUDDER-LENGTH ε [8.5, 14.2]
RUDDER-X-HEIGHT ε [7, 13]
RUDDER-X-WIDTH ε [0.7, 1.1]
WING-RATIO ε [0.35, 0.45]
WING-THICK ε [1.5, 2.5]
WING-WIDTH ε [7, 12]
        WING-WIDTH ≤ 0.5*WING-LENGTH
WING-LIFT ε [1, 2]
WING-SWEEP-BACK ε [13, 18]
WING-LENGTH ε [22, 33.5]
        WING-LENGTH ≥ 2*WING-WIDTH
        WING-LENGTH ≥ 0.43*FUSELAGE-LENGTH
        WING-LENGTH ≤ 0.65*FUSELAGE-LENGTH
REAR-ENGINE-LENGTH ε [6, 10]
ENGINE-LENGTH ε [4, 7]
ENGINE-RADIUS ε [1, 1.8]
FUSELAGE-RADIUS ε [2.5, 4]
FUSELAGE-LENGTH ε [40, 70]
        FUSELAGE-LENGTH ≥ 1.66666666*WING-ATTACHMENT
        FUSELAGE-LENGTH ≥ 1.53846154*WING-LENGTH
        FUSELAGE-LENGTH ≤ 2.5*WING-ATTACHMENT
        FUSELAGE-LENGTH ≤ 2.3255814*WING-LENGTH
R-ENG-QUANT ε [0, 1]
        R-ENG-QUANT ≤ 2 + -1*F-ENG-QUANT
F-ENG-QUANT ε [1, 2]
        F-ENG-QUANT ≤ 2 + -1*R-ENG-QUANT
```

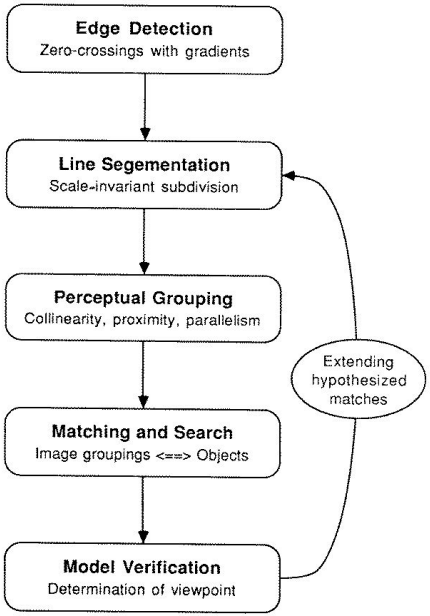
Fig. 9.2: Constraints on quantifiers in the generic model of wide-bodied passenger jet aircraft.

```
STARBOARD-WING.CAMZ ε [-3323.668, -2201.568]
        STARBOARD-WING.CAMZ ≥ -584.08894*WING-WIDTH
        STARBOARD-WING.CAMZ ≥ -99.213672*WING-LENGTH
        STARBOARD-WING.CAMZ ≥ -2.666668 + -1*HEIGHT
        STARBOARD-WING.CAMZ ≤ -314.50943*WING-WIDTH
        STARBOARD-WING.CAMZ ≤ -53.4227467*WING-LENGTH
        STARBOARD-WING.CAMZ ≤ -1.6666675 + -1*HEIGHT
STARBOARD-WING.CAMY ε [-∞, ∞]
        STARBOARD-WING.CAMY ≥ -44 + AIRCRAFT-Y
        STARBOARD-WING.CAMY ≤ 44 + AIRCRAFT-Y
STARBOARD-WING.CAMX ε [-∞, ∞]
        STARBOARD-WING.CAMX ≥ -44 + AIRCRAFT-X
        STARBOARD-WING.CAMX ≤ 44 + AIRCRAFT-X
AIRCRAFT-Y ε [-∞, ∞]
        AIRCRAFT-Y ≥ -44 + STARBOARD-WING.CAMY
        AIRCRAFT-Y ≥ -44 + PORT-WING.CAMY
        AIRCRAFT-Y ≤ 44 + STARBOARD-WING.CAMY
        AIRCRAFT-Y ≤ 44 + PORT-WING.CAMY
AIRCRAFT-X ε [-∞, ∞]
        AIRCRAFT-X ≥ -44 + STARBOARD-WING.CAMX
        AIRCRAFT-X ≥ -44 + PORT-WING.CAMX
        AIRCRAFT-X ≤ 44 + STARBOARD-WING.CAMX
        AIRCRAFT-X ≤ 44 + PORT-WING.CAMX
F-ENG-QUANT ε [1, 2]
        F-ENG-QUANT ≤ 2 + -1*R-ENG-QUANT
R-ENG-QUANT ε [0, 1]
        R-ENG-QUANT ≤ 2 + -1*F-ENG-QUANT
FUSELAGE-LENGTH ε [40.7801786, 70]
        FUSELAGE-LENGTH ≥ 0.0162526593*HEIGHT
        FUSELAGE-LENGTH ≥ 1.53846154*WING-LENGTH
        FUSELAGE-LENGTH ≥ 2.03900883*WING-ATTACHMENT
        FUSELAGE-LENGTH ≤ 0.0301835102*HEIGHT
        FUSELAGE-LENGTH ≤ 2.3255814*WING-LENGTH
        FUSELAGE-LENGTH ≤ 2.5*WING-ATTACHMENT
FUSELAGE-RADIUS ε [2.5, 4]
        FUSELAGE-RADIUS ≥ -1.05197865E-3*PORT-WING.CAMZ
        FUSELAGE-RADIUS ≥ 1.11117898E-3*HEIGHT
        FUSELAGE-RADIUS ≤ -1.75329772E-3*PORT-WING.CAMZ
        FUSELAGE-RADIUS ≤ 2.0636181E-3*HEIGHT
```

Fig 9.3: continued overleaf ...

Geometric model

Three-Dimensional Object Recognition from Single Two-Dimensional Images, David Lowe, Artificial Intelligence, 1987



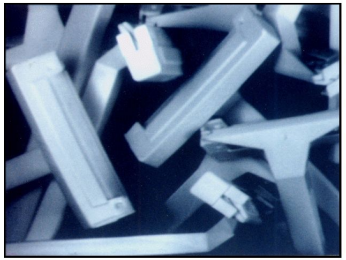


Figure 9: The original image of a bin of disposable razors, taken at a resolution of 512×512 pixels.

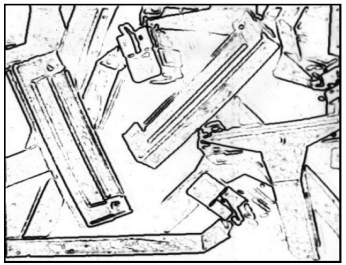


Figure 10: The zero-crossings of a $\nabla^2 G$ convolution. Grey levels are proportional to gradient magnitude at the zero-crossing.

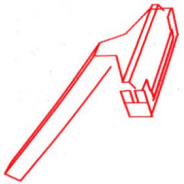
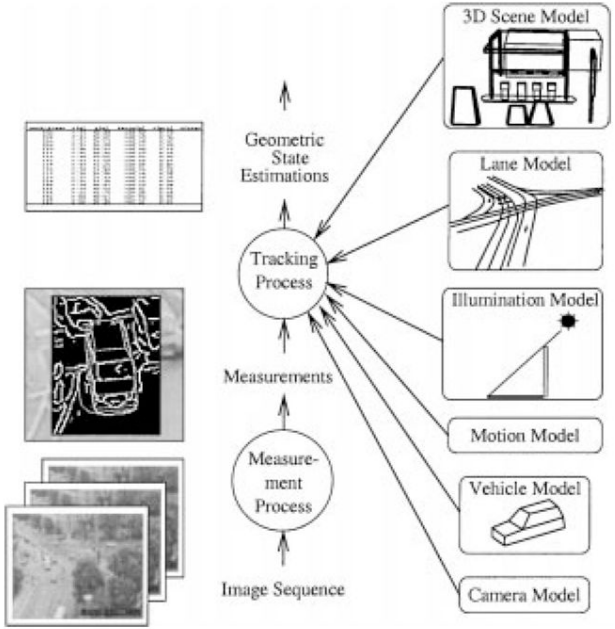


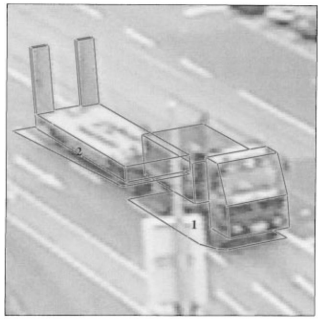
Figure 13: The three-dimensional wire-frame model of the razor shown from a single viewpoint.

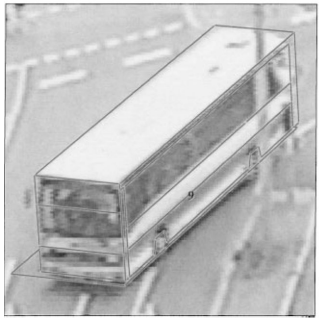
Figure 14: Successful matches between sets of image segments and particular viewpoints of the model.

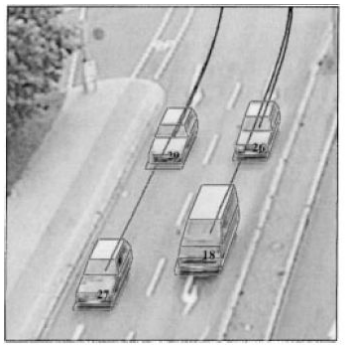
Geometric model

Combination of Edge Element and Optical Flow Estimates for 3D-Model-Based Vehicle Tracking in Traffic Image Sequences, Haag and Nagel, IJCV, 1999









Geometric model

Acquisition of a Dense 3D Model Database for Robotic Vision, Zia et al., ICAR, 2009

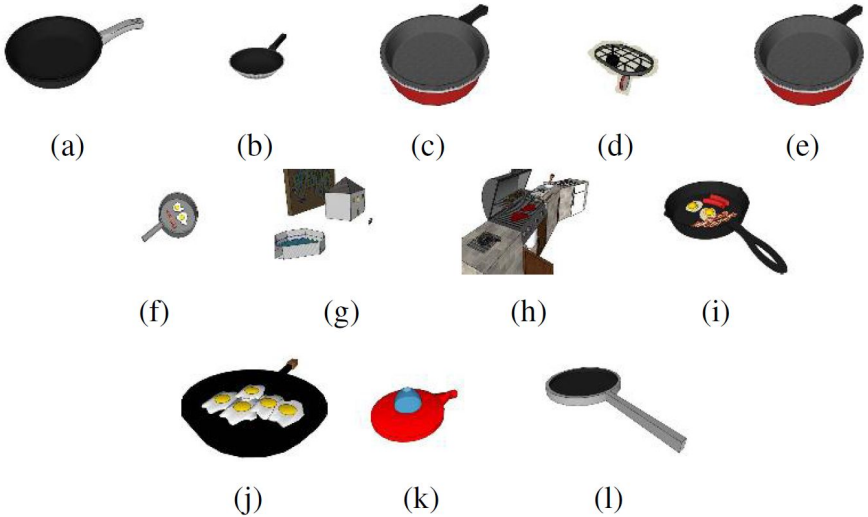
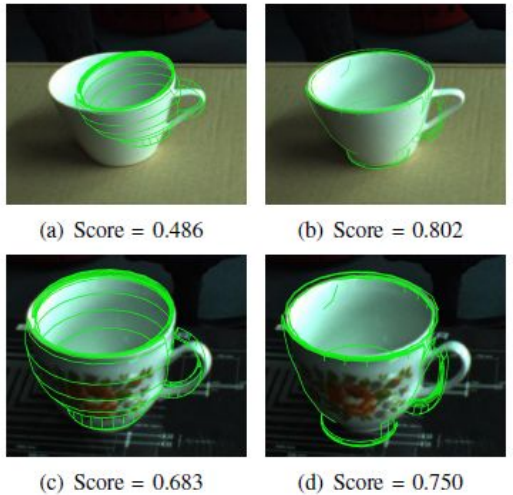
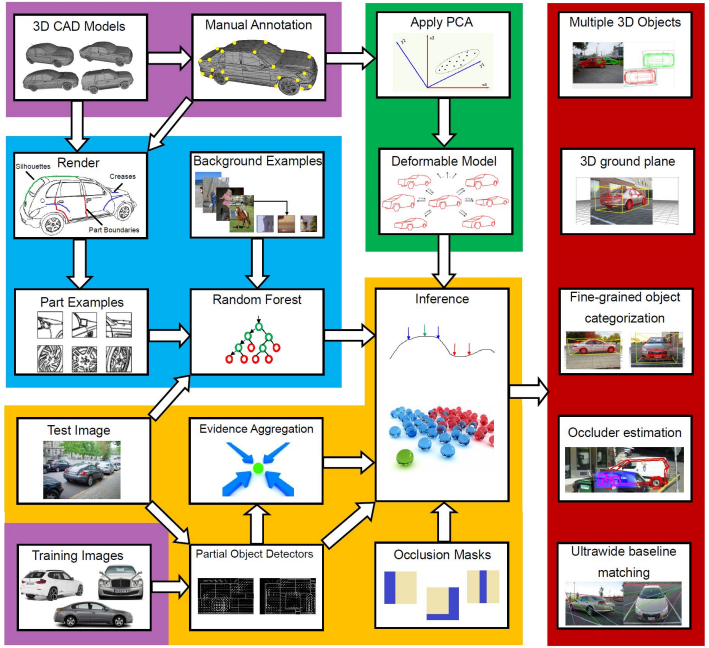


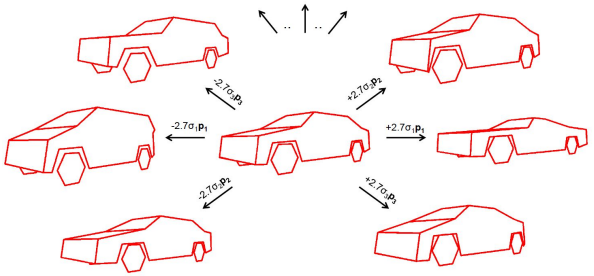
Fig. 6. The resulting four clusters for the query for "pan", a - e are inliers, all the others are classified as outliers (clusters are f-i, j-k and l).

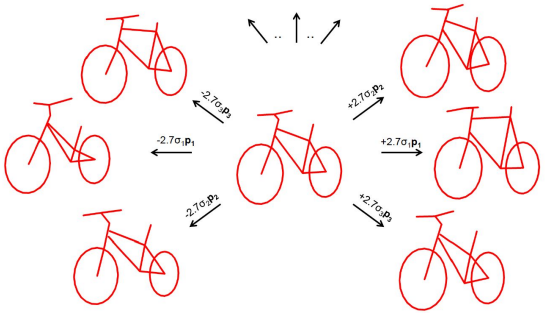


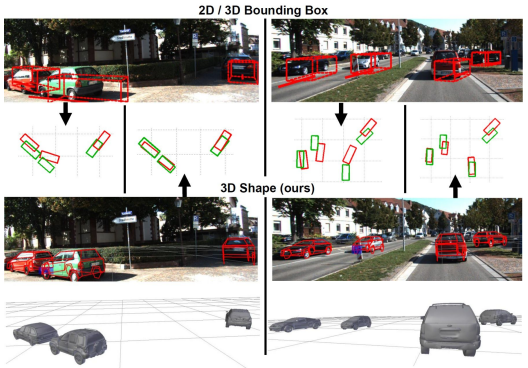
Geometric model

Revisiting 3D geometric models for accurate object shape and pose, Zia et al., 2009-2013, (ICCV 2011, TPAMI 2013, CVPR 2014/2015, IJCV 2015)



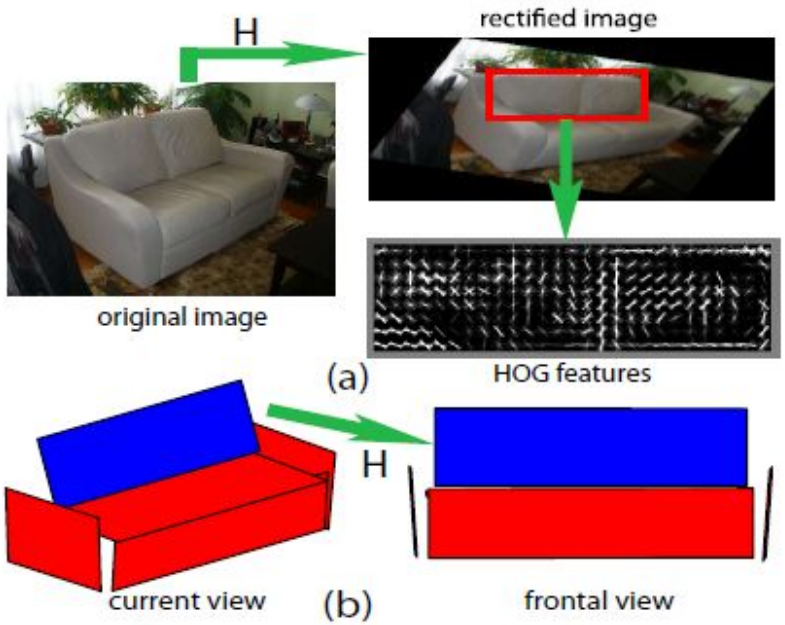


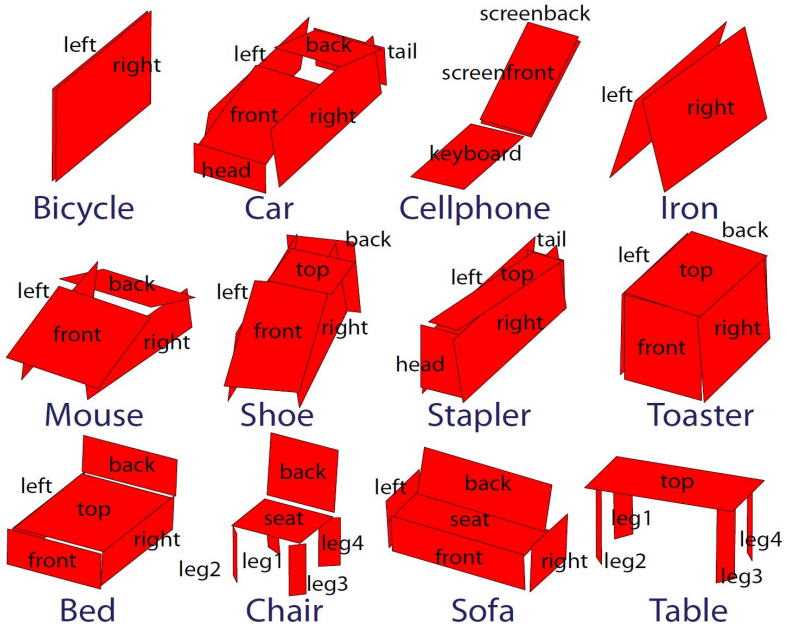




Geometric model

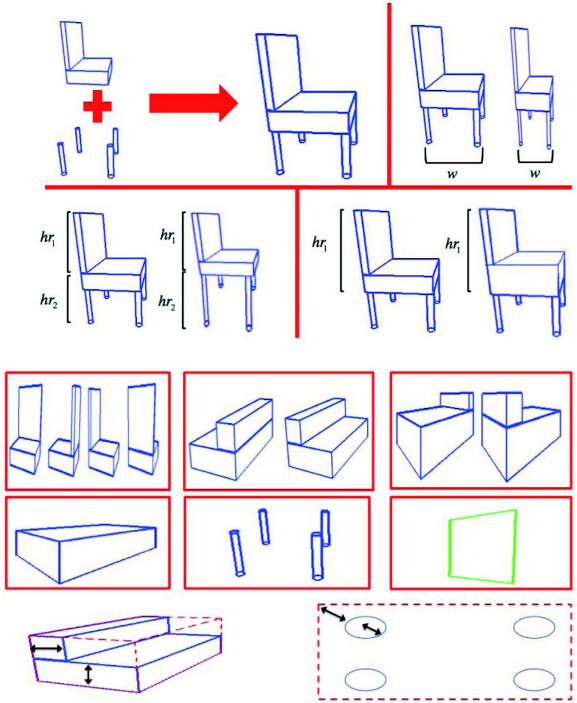
Estimating the Aspect Layout of Object Categories, Xiang and Savarese, CVPR, 2012





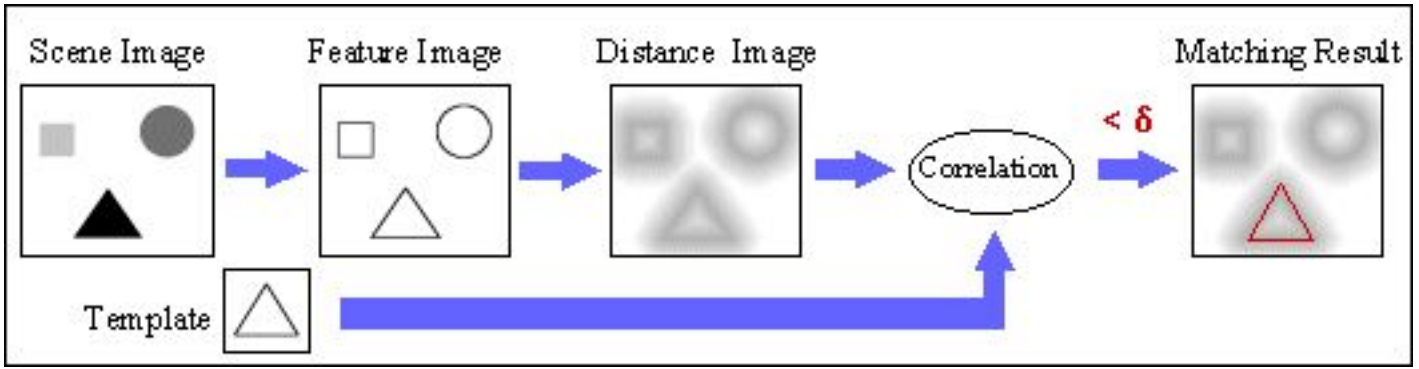
Geometric model

Understanding Bayesian Rooms using composite 3D object models, Del Pero et al., CVPR, 2013



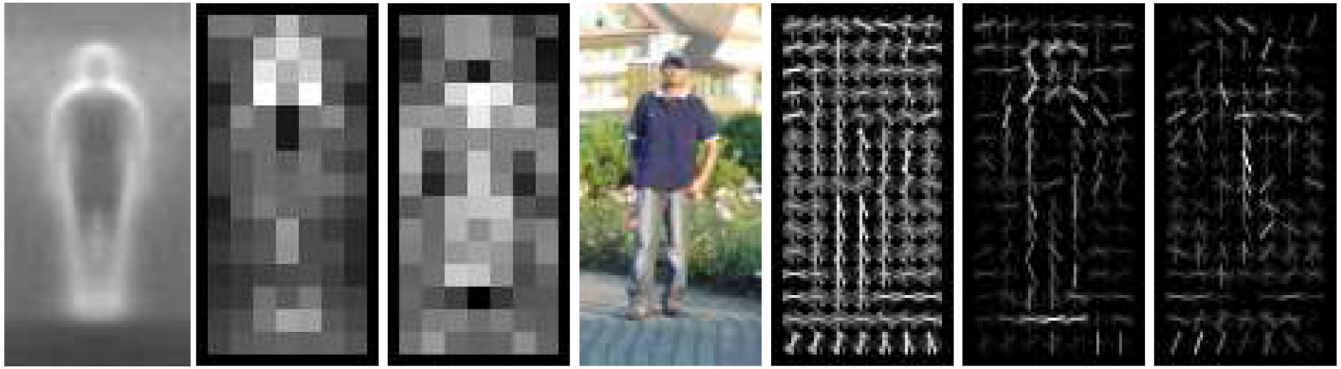
Appearance model

Image intensity and edge matching (1960-~2005)



Appearance model

Handcrafted features (HOG, Shape Context, SIFT) + discriminative classifiers (SVM, RFs), CVPR 2005 onwards

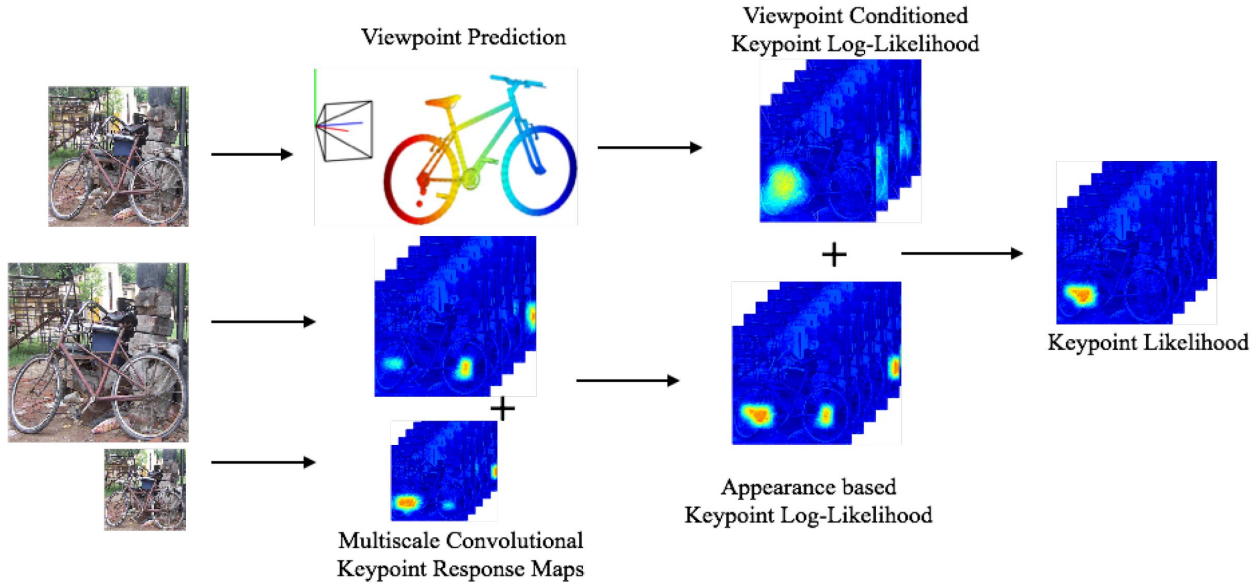


(a) (b) (c) (d) (e) (f) (g)

Figure 6. Our HOG detectors cue mainly on silhouette contours (especially the head, shoulders and feet). The most active blocks are centred on the image background just *outside* the contour. (a) The average gradient image over the training examples. (b) Each "pixel" shows the maximum positive SVM weight in the block centred on the pixel. (c) Likewise for the negative SVM weights. (d) A test image. (e) It's computed R-HOG descriptor. (f,g) The R-HOG descriptor weighted by respectively the positive and the negative SVM weights.

Appearance model

Viewpoints and Keypoints, Tulsiani and Malik, CVPR 2015

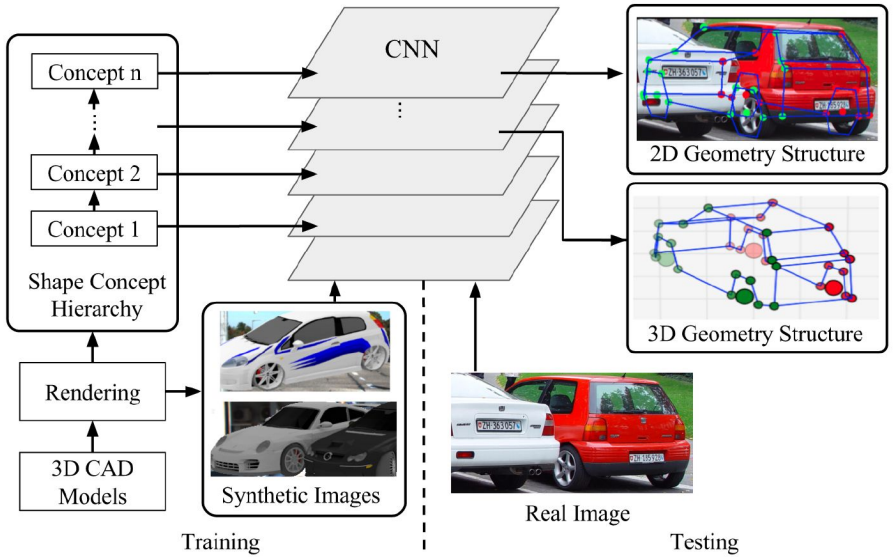


Inference

- Belief Propagation
- Markov Chain Monto Carlo
- Gradient descent

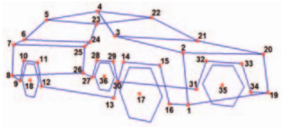
Overview of our approach

- Combine geometric + appearance model into one CNN.
- Inference in single forward pass
- Training using synthetic data, occlusion modeling

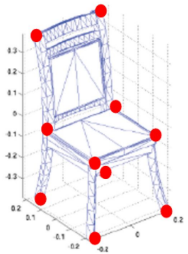


Training data

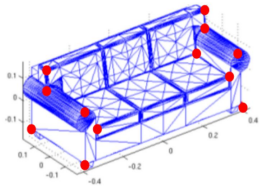
- Hand defined keypoints
- 3D CAD models labeled
- Renderings against real backgrounds
- Multiple object occlusions represented in training data



(a) Car Keypoints



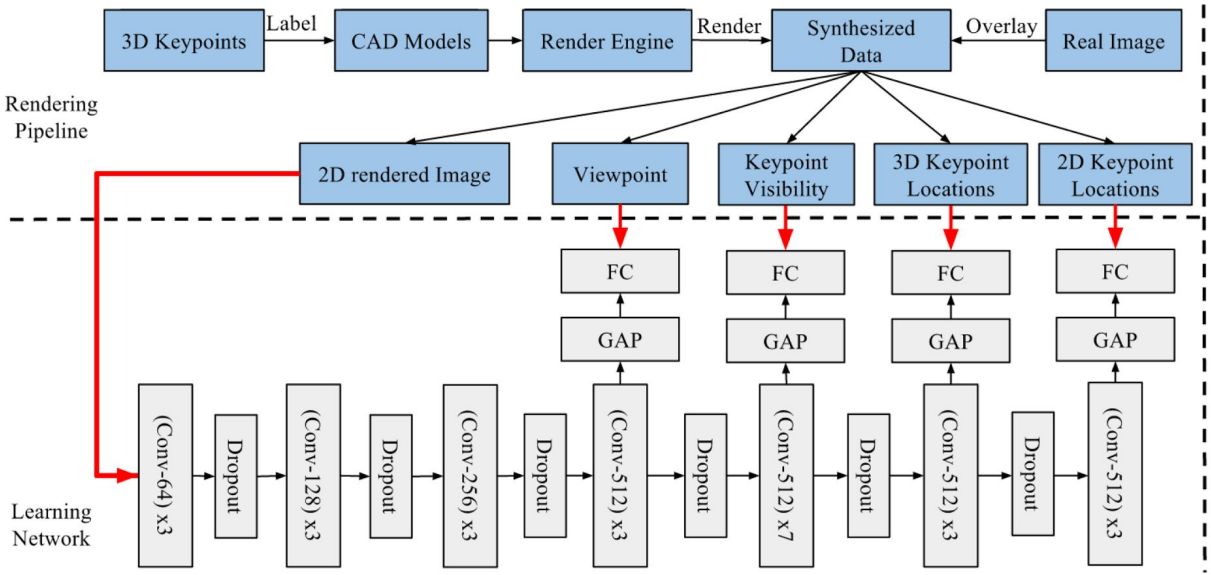
(b) Chair Keypoints



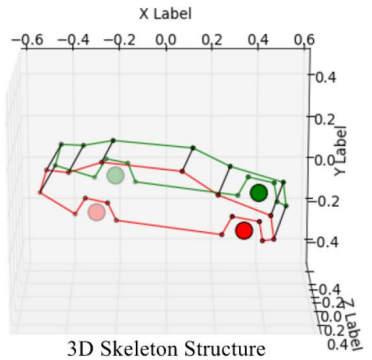
(c) Sofa Keypoints

Deep supervision with shape concepts (DISCO)

- Prevent over fitting to synthetic renderings

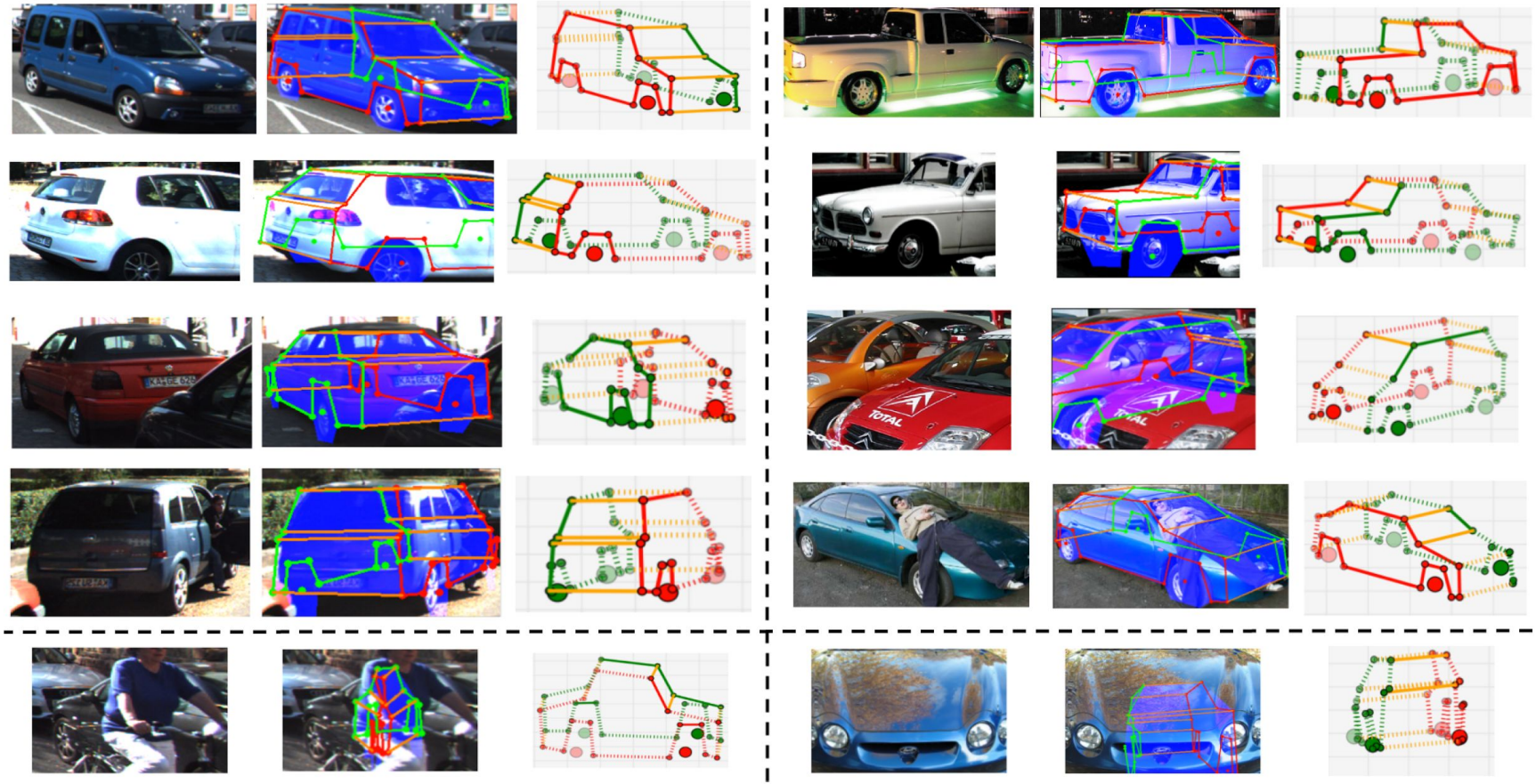


Rendered Image & 2D Keypoints

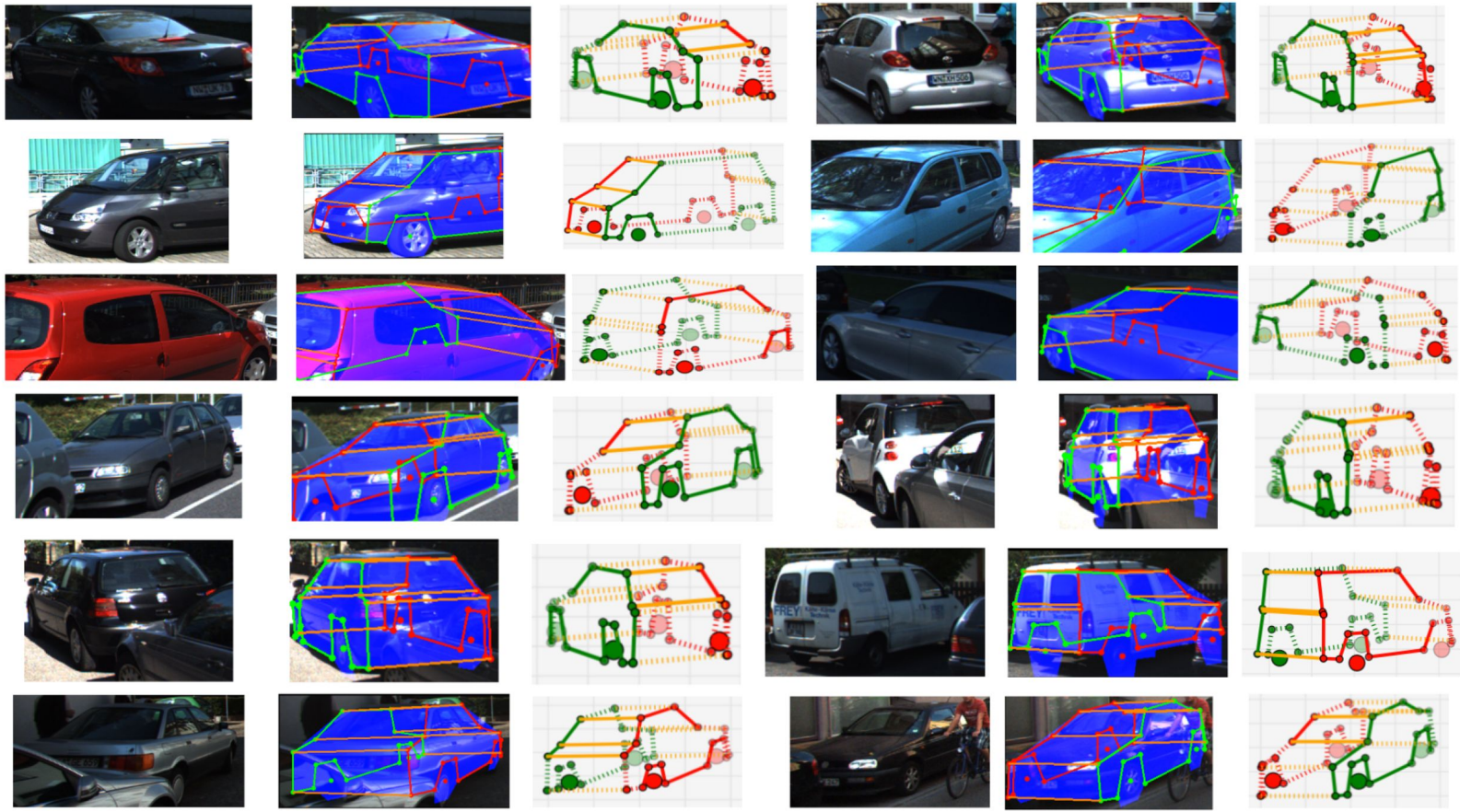


3D Skeleton Structure

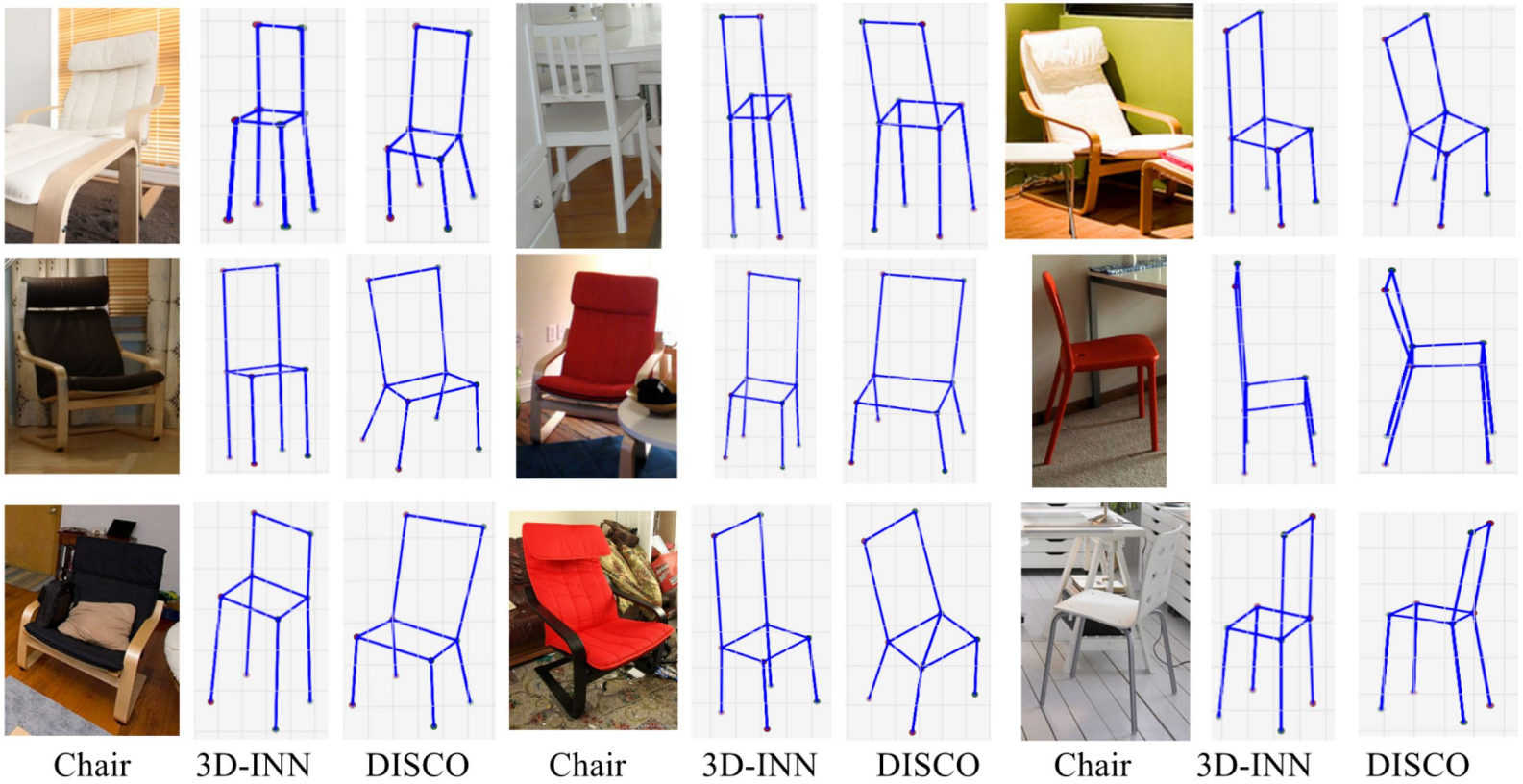
Qualitative results



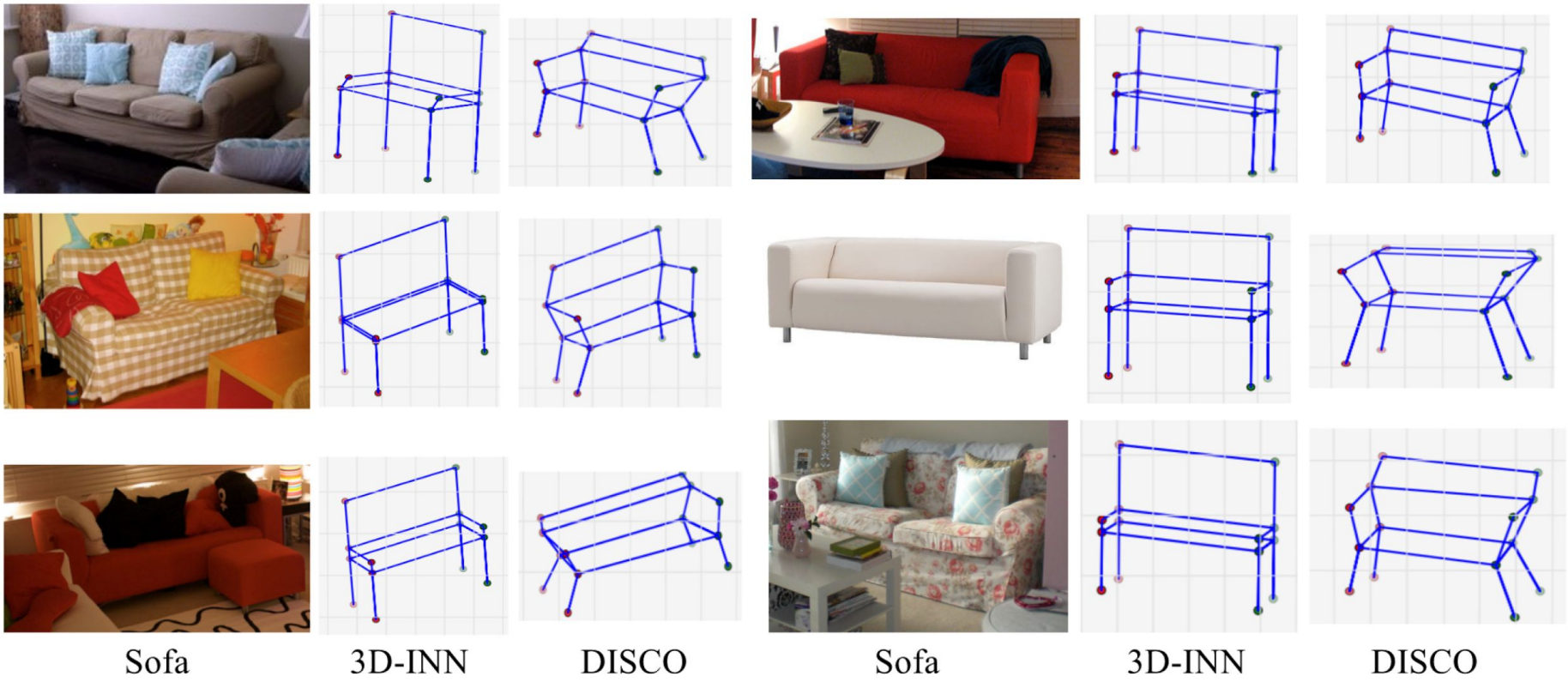
Qualitative results



Qualitative results



Qualitative results



Percentage of Correct Keypoints (PCK)

- Correct Keypoint => it falls within a distance of 10% (of image extent) from ground truth location.
- KITTI-3D dataset, cars

Method	2D					3D	3D-yaw
	Full	Truncation	Multi-Car Occ	Other Occ	All	Full	Full
DDN [44]	67.6	27.2	40.7	45.0	45.1	NA	
WN-gt-yaw* [13]	88.0	76.0	81.0	82.7	82.0	NA	
Zia et al. [47]	73.6	NA				73.5	7.3
plain-2D	88.4	62.6	72.4	71.3	73.7	NA	
plain-3D	NA					90.6	6.5
plain-all	90.8	72.6	78.9	80.2	80.6	92.9	3.9
DISCO-3D-2D	90.1	71.3	79.4	82.0	80.7	94.3	3.1
DISCO-vis-3D-2D	92.3	75.7	81.0	83.4	83.4	95.2	2.3
DISCO-Vgg	83.5	59.4	70.1	63.1	69.0	89.7	6.8
DISCO	93.1	78.5	82.9	85.3	85.0	95.3	2.2

Percentage of Correct Keypoints (PCK)

- Pascal VOC dataset
- 2D pixel locations
- Long [21], “Do ConvNets learn correspondence?”, NIPS 2014

PCK[$\alpha = 0.1$]	Long[21]	VKps[38]	DISCO
Full	55.7	81.3	81.8
Full[$\alpha = 0.2$]	NA	88.3	93.4
Occluded	NA	62.8	59.0
Big Image	NA	90.0	87.7
Small Image	NA	67.4	74.3
All [APK $\alpha = 0.1$]	NA	40.3	45.4

Object segmentation accuracy

- Percentage of pixels correctly segmented

Method	CAD alignment GT	Manual GT
VDPM-16 [42]	NA	51.9
Xiang et al. [28]	64.4	64.3
Random CAD [42]	NA	61.8
GT CAD [42]	NA	67.3
DISCO	71.2	67.6

Conclusion

- Joint object geometry and appearance modeling
- Exploiting the ability of CNNs to model context: occlusion reasoning
- Inference in a single forward pass

- Deep supervision approach to better exploit synthetic training data