Click Here: Human-Localized Keypoints as Guidance for Viewpoint Estimation

Ryan Szeto and Jason J. Corso
Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI
{szetor,jjcorso}@umich.edu

UNIVERSITY OF MICHIGAN

ICCV 17

Project URL: **ryanszeto.com/projects/ch-cnn**

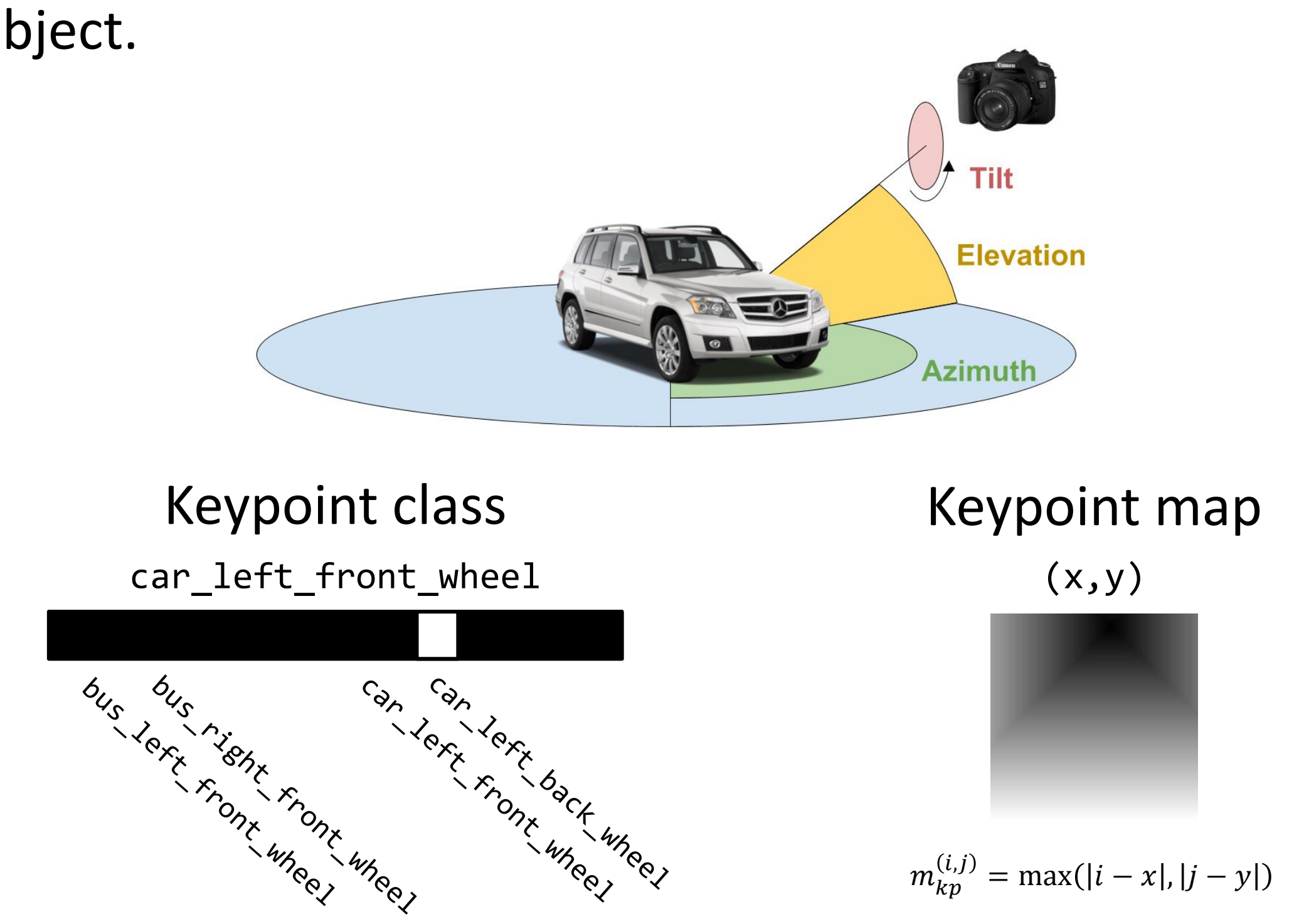
Objective and Contributions

We leverage human guidance at inference time to improve monocular viewpoint estimation performance over image-only approaches.

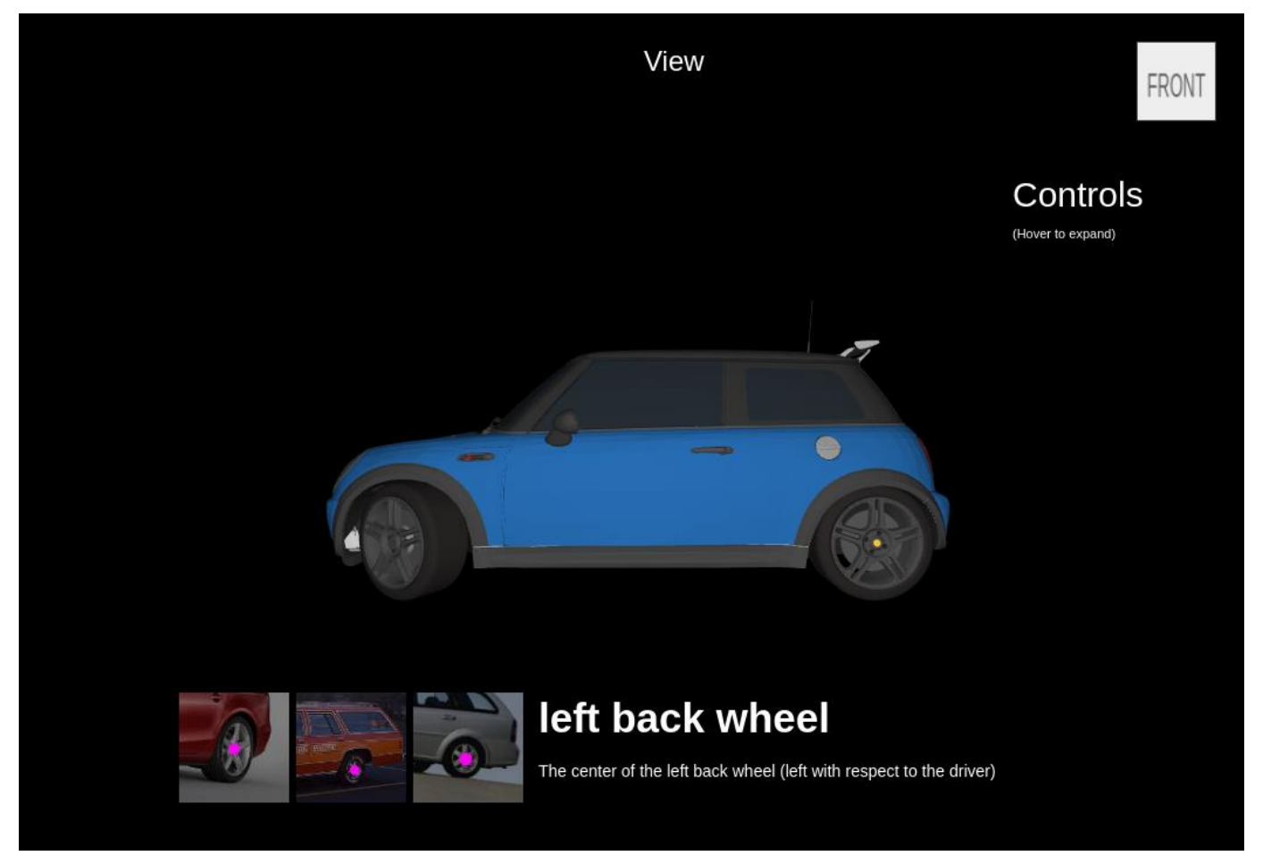
- Motivations:
  - Human guidance can help overcome challenges due to occlusion, truncation, and symmetry
  - High benefit-to-human-effort ratio: Humans can quickly locate keypoints, and the information can help disambiguate viewpoint candidates
- Contributions:
  - Click-Here CNN (CH-CNN), a model that estimates the viewpoint from an image and information about a single keypoint
  - A publicly-available dataset of keypoint locations on over 8,500 CAD models from ShapeNet [2]
  - Better viewpoint estimates: CH-CNN achieves 90.7% accuracy on PASCAL 3D+ [3], whereas the state-of-the-art image-only model [1] obtains 85.7%

Problem Statement

Given an image, information about one keypoint (2D location and class), and the object class, predict the azimuth, elevation, and tilt of the camera w.r.t. the object.

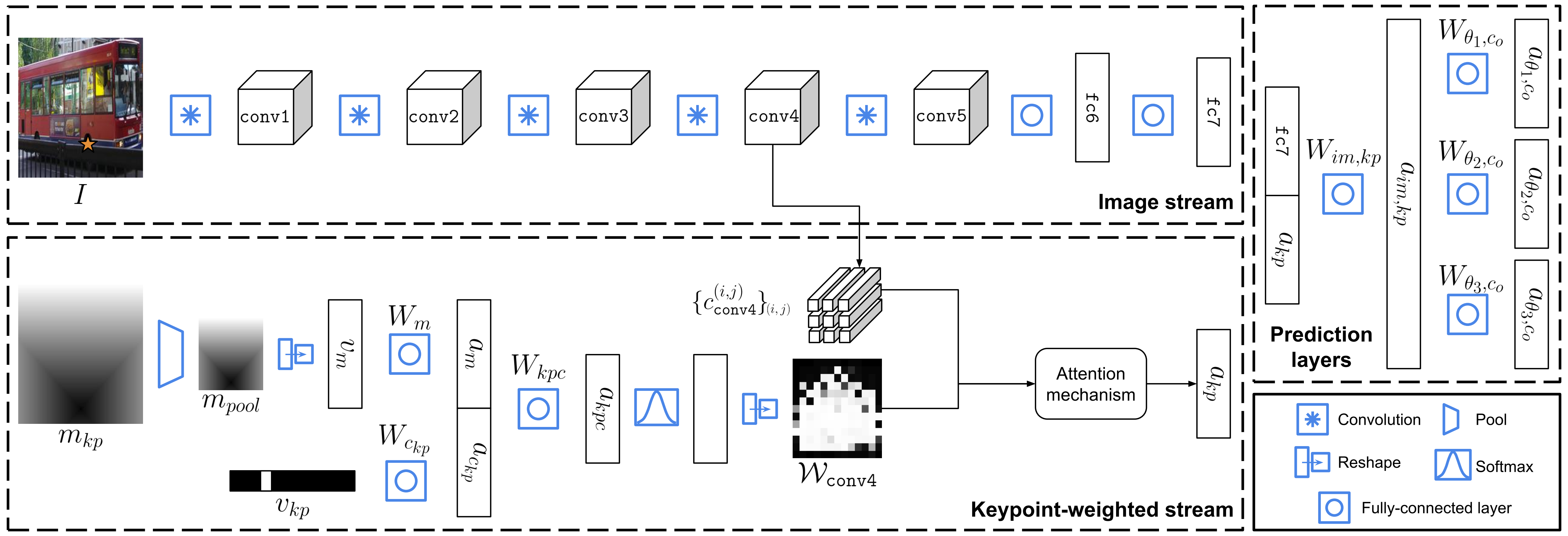


Keypoint Collection



Our dataset, with annotations for 918 bus, 7,377 car, and 320 motorcycle models, includes over ten times more models than the next-largest ShapeNet keypoint dataset [4]. It is available on our project website.

Click-Here CNN (CH-CNN)



Training Data

Synthetic Data

- Extend Render for CNN pipeline [1] to generate synthetic images with keypoint data
- Sample viewpoint and cropping parameters, add background image
- Extract keypoint location and visibility from rendering engine

PASCAL 3D+ Data

- Object bounding boxes, keypoint locations, and viewpoint labels included
- Crop the object and create one training instance for each visible keypoint

Training Procedure

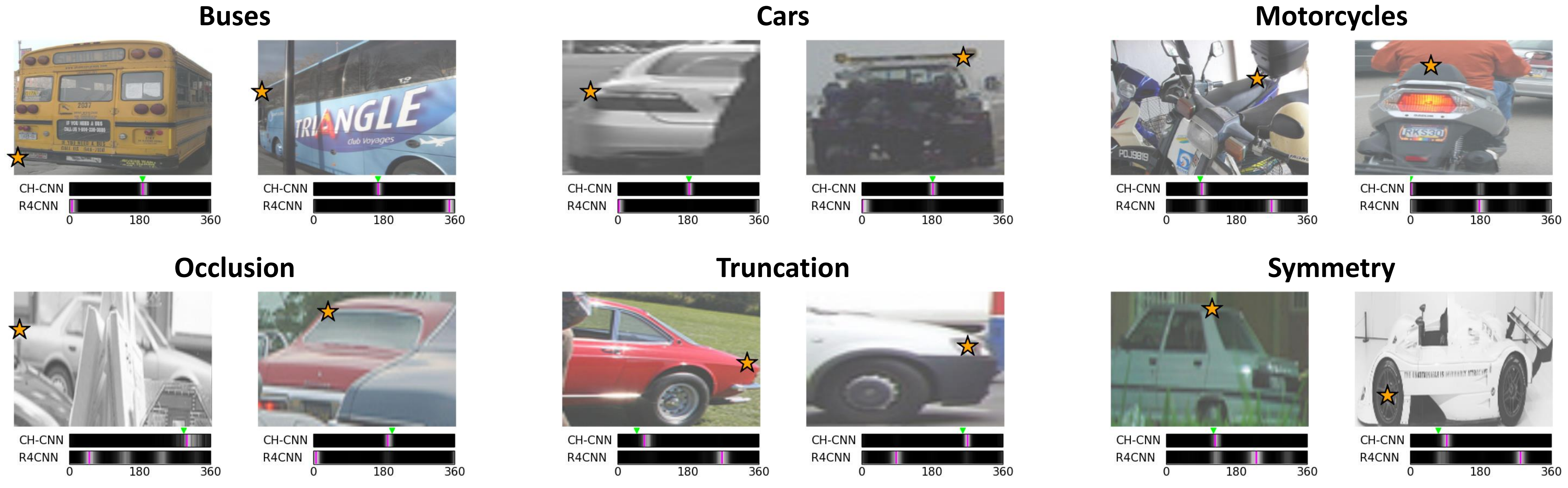
1. Initialize image feature layers with R4CNN [1] weights
2. Fine-tune on synthetic data until convergence
3. Fine-tune on PASCAL 3D+ data until convergence

Geometric structure-aware loss [1]:

$$L_{\theta_i}(\mathcal{S}) = -\sum_{s \in \mathcal{S}} \sum_{\theta \in \Theta} e^{-d(\theta, \theta_{gt})/t} \log(P(\theta|s))$$

Qualitative Results

- Gray bars indicate confidence across 360 azimuth angles
- Green triangles mark the ground truth, purple lines indicate each model's most confident prediction



- R4CNN: Multiple peaks, wide bands, or high confidence for the angle opposite the ground truth
- CH-CNN: Higher confidence within smaller intervals around the ground truth compared to R4CNN
- CH-CNN: More robust to occlusion, truncation, and symmetry

Experiments: Accuracy and Median Error

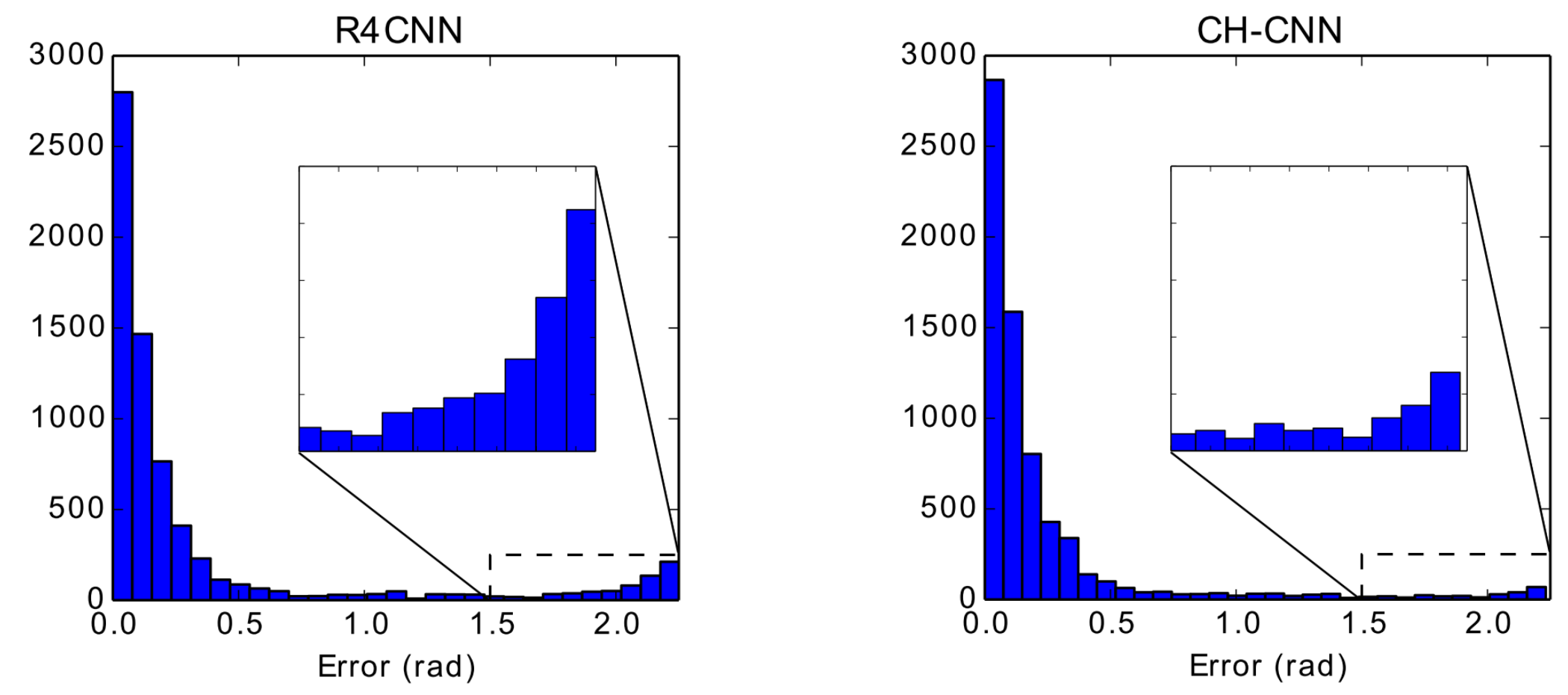
- $Acc_{\pi/6}$: Geodesic distance between predicted rotation matrix and GT is less than $\pi/6$ in radians [1, 5]
- $MedErr$: The median error (in degrees) for estimates in the object class [1, 5]
- Fixed weights: `conv4` columns are weighted by 2D Gaussian or uniform map

$Acc_{\pi/6}$				
	bus	**car**	**motor**	**mean**
R4CNN [1]	92.4	78.5	81.4	84.1
R4CNN [1], fine-tuned	90.6	82.4	84.1	85.7
Fixed weights, Gaussian	88.9	81.3	82.8	84.4
Fixed weights, uniform	90.6	82.0	83.7	85.4
CH-CNN (KPM only)	90.6	82.0	84.2	85.6
CH-CNN (KPC only)	90.9	86.3	83.1	86.8
CH-CNN (full model)	**96.8**	**90.2**	**85.2**	**90.7**

$MedErr$				
	bus	**car**	**motor**	**mean**
R4CNN [1]	5.04	7.86	14.5	9.14
R4CNN [1], fine-tuned	2.93	5.63	11.7	6.74
Fixed weights, Gaussian	3.00	5.88	11.4	6.76
Fixed weights, uniform	3.01	5.72	12.1	6.93
CH-CNN (KPM only)	3.04	5.73	11.3	6.68
CH-CNN (KPC only)	2.92	5.29	**11.0**	6.41
CH-CNN (full model)	**2.64**	**4.98**	11.4	**6.35**

- CH-CNN surpasses state-of-the-art image-only model
- Weighing `conv4` columns dynamically from keypoint data works better than hand-crafted maps
- Best results from using both keypoint location and class

Experiments: Error Histograms



- CH-CNN: Large errors are less frequent
- CH-CNN takes advantage of keypoint features when image features are insufficient

References

[1] Su et al. Render for CNN: *Viewpoint Estimation in Images Using CNNs Trained with Rendered 3D Model Views.* ICCV 2015.
[2] Chang et al. *ShapeNet: An Information-Rich 3D Model Repository.* ArXiv 2015.
[3] Xiang et al. *Beyond PASCAL: A Benchmark for 3D Object Detection in the Wild.* WACV 2014.
[4] Li et al. *Deep Supervision with Shape Concepts for Occlusion-Aware 3D Object Parsing.* CVPR 2017.
[5] Tulsiani and Malik. *Viewpoints and Keypoints*. CVPR 2015.

Acknowledgements

This work was partially supported by the Denso Corporation, NSF CNS 1463102, and DARPA W31P4Q-16-C-0091.