3D Pose-by-Detection of Vehicles via Discriminatively Reduced Ensembles of Correlation Filters

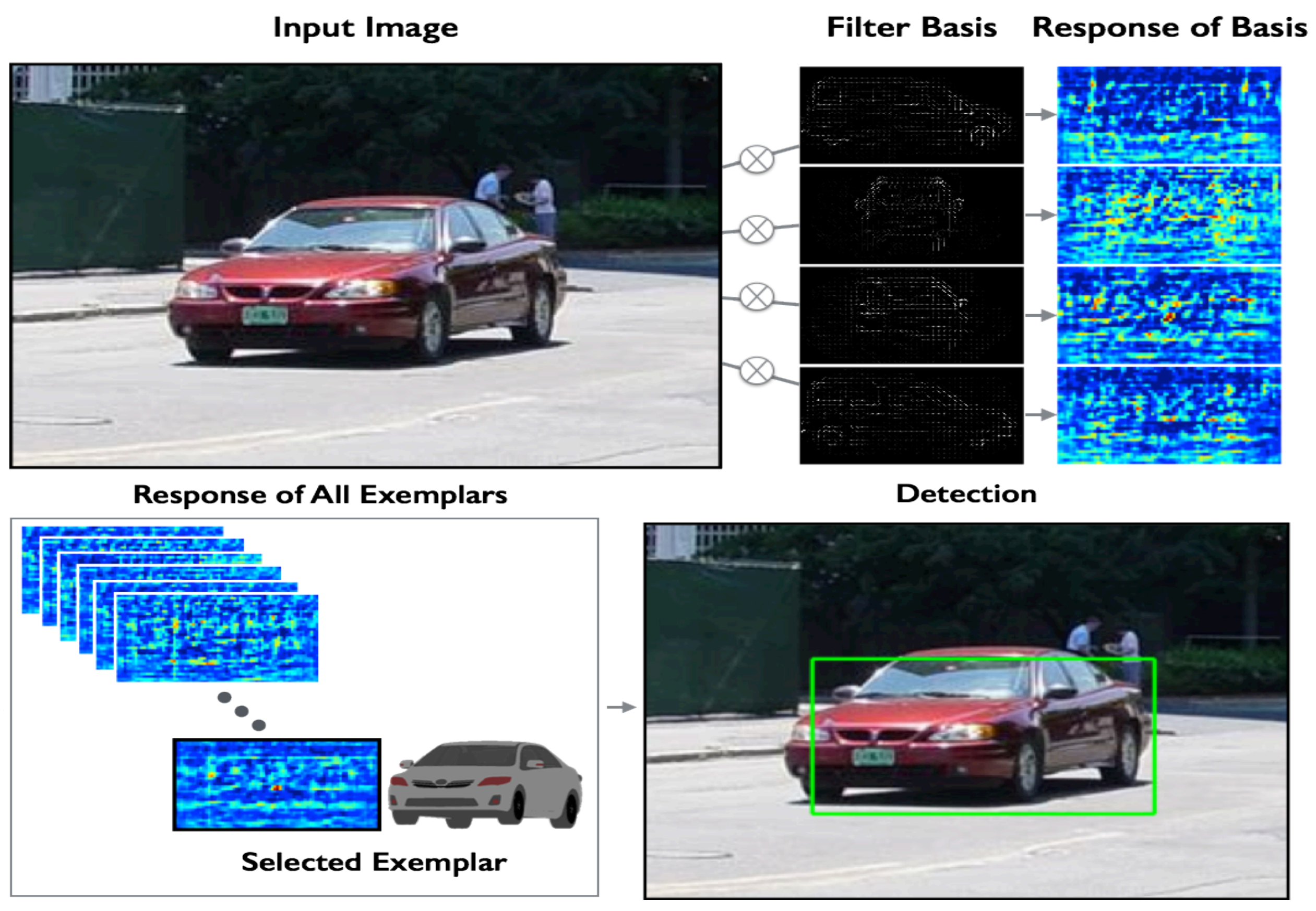
Yair Movshovitz-Attias[1], Vishnu Naresh Boddeti[2], Zijun Wei[2], Yaser Sheikh[2]
1. Computer Science Department, Carnegie Mellon University
2. Robotics Institute, Carnegie Mellon University
Contact: yair@cs.cmu.edu, Website: www.cs.cmu.edu/~ymovshov

Carnegie Mellon University
School of Computer Science

1. Introduction

Estimating the precise pose of a 3D model in an image is challenging; explicitly identifying correspondences is difficult, particularly at smaller scales and in the presence of occlusion. Exemplar classifiers have demonstrated the potential of detection-based approaches to problems where precision is required. In particular, correlation filters explicitly suppress classifier response caused by slight shifts in the bounding box. This property makes them ideal exemplar classifiers for viewpoint discrimination, as small translational shifts can often be confounded with small rotational shifts. However, exemplar based pose-by-detection is not scalable because, as the desired precision of viewpoint estimation increases, the number of exemplars needed increases as well. We present a training framework to reduce an ensemble of exemplar correlation filters for viewpoint estimation by directly optimizing a discriminative objective. We show that the discriminatively reduced ensemble outperforms the state-of-the-art on three publicly available datasets and we introduce a new dataset for continuous car pose estimation in street scene images.

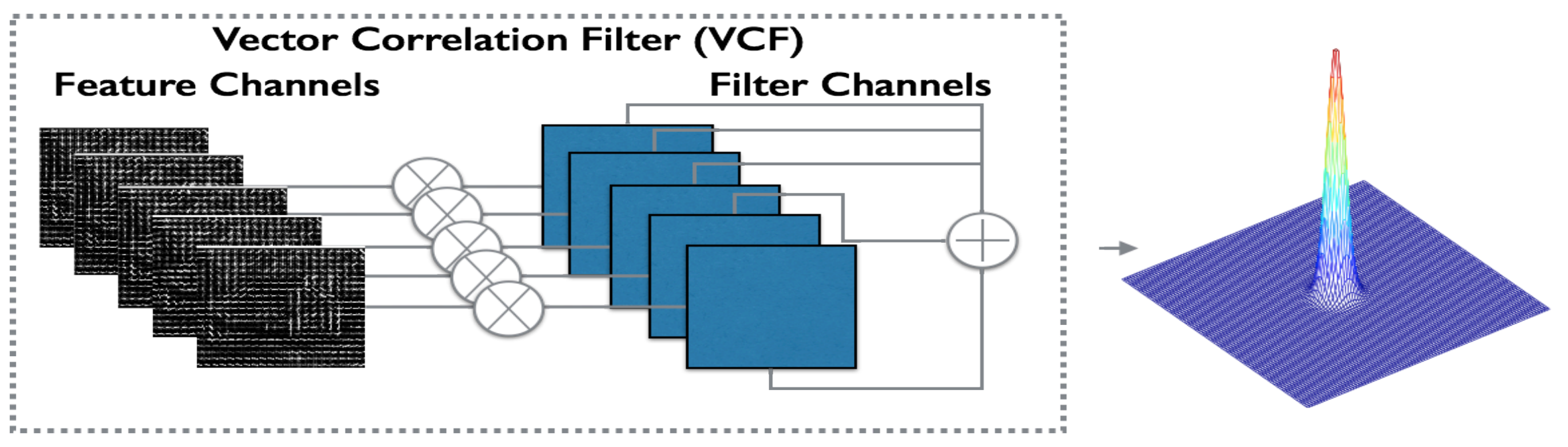
Method Overview



2. Background: Exemplar Correlation Filter

Correlation filters are described by the equation below. They are a type of classifier that explicitly controls the shape of the entire cross correlation output between an image, x_i, and the filter, g_v, and are designed to give a sharp peak response, r_i, at the location of the object, and no such peak anywhere else. Here an image is represented by a HOG feature with C channels.

$$\min_{\mathbf{g}_v^1,\cdots,\mathbf{g}_v^C} \sum_{i=1}^{N} \left\| \sum_{c=1}^{C} \mathbf{x}_i^c \otimes \mathbf{g}_v^c - \mathbf{r}_i \right\|_2^2 + \lambda \sum_{c=1}^{C} \left\| \mathbf{g}_v^c \right\|_2^2$$



3. Learning Ensembles of Correlation Filters

To use exemplar correlation filters to predict view points, we can learn one filter, g_v, per desired view. However, this results in a large set of filters. Applying many filters during test time is computationally expensive.

Therefore, our method jointly learns a set of K correlation filters, F, and a set of V sparse correlation filters, A, such that a detector g_v for any viewpoint is defined by: $g_v = F\alpha_v$

The joint learning objective is:

$$\begin{aligned} \arg\min_{\mathbf{F},\mathbf{A}} \sum_{i:\mathbf{x}_i \in V} \left\| \sum_{k=1}^{K} \alpha_k^i \left(\sum_{c=1}^{C} \mathbf{f}_k^c \otimes \mathbf{x}_i^c \right) - \mathbf{r}^{\text{pos}} \right\|_2^2 + & \quad \text{Controls response on positive instances} \\ \sum_{j:\mathbf{x}_j \in B} \sum_{i:\mathbf{x}_i \in V} \left\| \sum_{k=1}^{K} \alpha_k^i \left(\sum_{c=1}^{C} \mathbf{f}_k^c \otimes \mathbf{x}_j^c \right) - \mathbf{r}^{\text{neg}} \right\|_2^2 + & \quad \text{Controls response on negative instances} \\ \lambda_1 \|\mathbf{F}\|_2^2 + \lambda_2 \|\mathbf{A}\|_1 & \quad \text{Controls sparsity} \end{aligned}$$

4. **CMU-car**: A dataset of street scenes with car 3D viewpoint annotations

The MIT street scene data set was augmented by Boddeti et al. with landmark annotations. To allow for evaluation of precise viewpoint estimation we further augment this data set by providing camera matrices for **3,240 cars**. To get the camera viewpoint matrices, we manually annotated a 3D CAD car model with the same landmark locations as the images and used the POSIT algorithm to align the model to the images.

5. Experiments

We tested our method for object detection and view point estimation on 4 dataset: (1) Weizmann Car View Point, (2) EPFL Multi-View Cars, (3) VOC2007 Car Viewpoint, (4) CMU-car.

Displayed are the median angular errors when the model of the car in the image is known/unknown. We compare the results when using an ensemble of size K, with a full set of examples correlation filters, and with previous work.

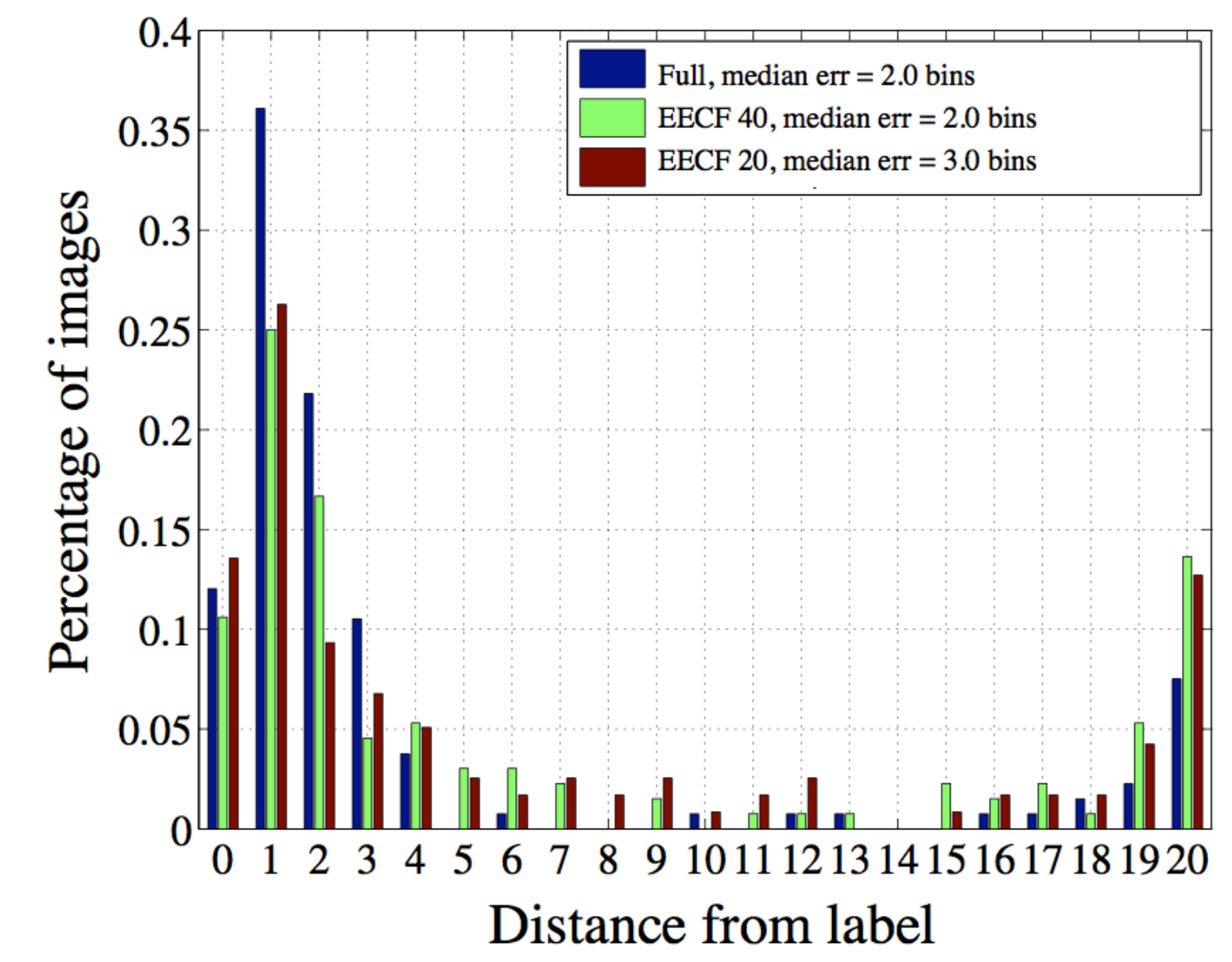
	Median Angular Error	
Method	*KNOWN*	*UNKNOWN*
EECF, $K = 20$	10.2°	9.4
EECF, $K = 40$	7.6°	8.4°
ECF full (360)	6.9°	7.5
Glasner et al. [14]	-	12.25°

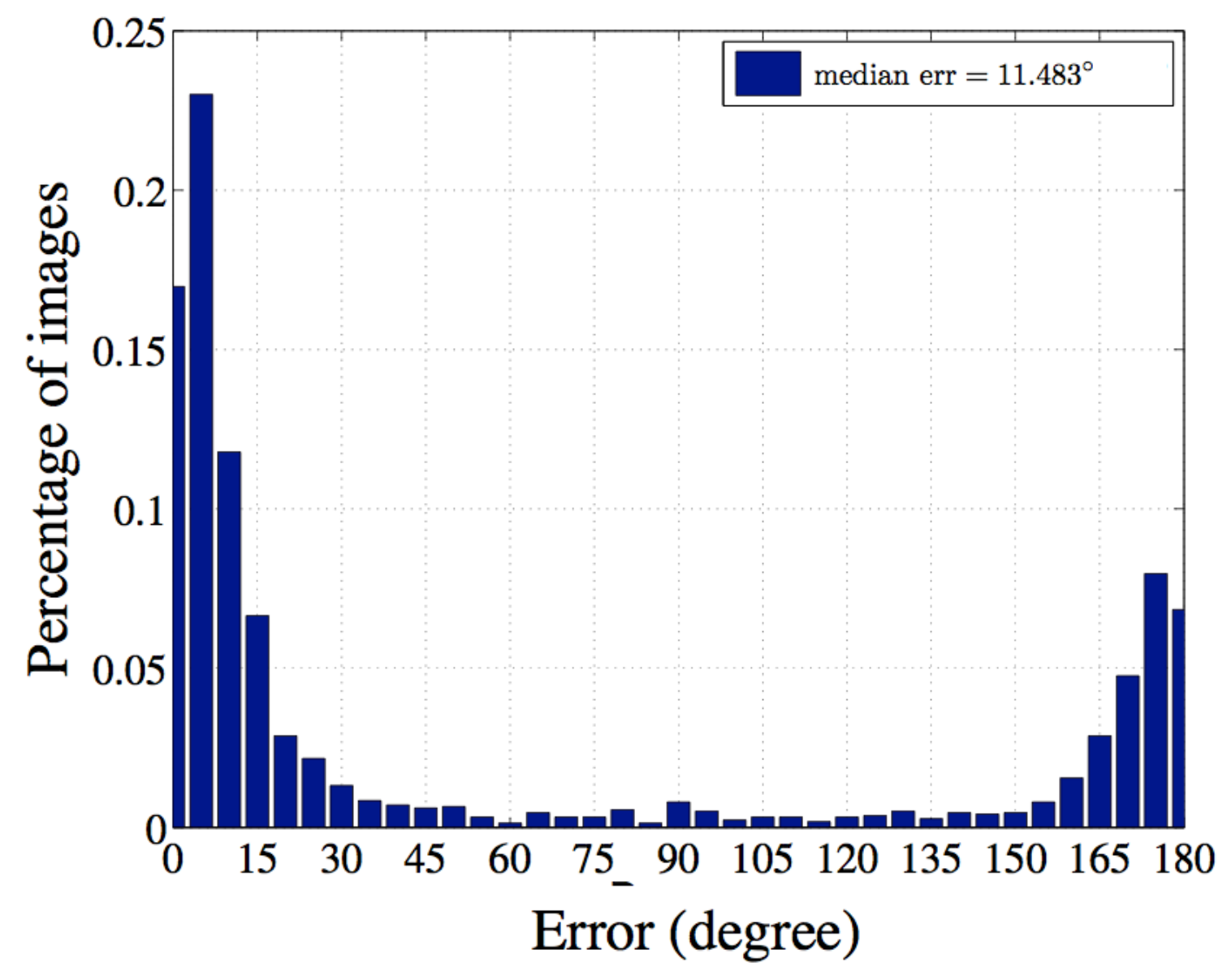
(a) Azimuth Estimation: WCVP

	Median Angular Error	
Method	Azimuth	Elevation
EECF, $K = 20$	26.0°	3.8°
EECF, $K = 40$	11.48°	3.6°
ECF Full (360)	3.2°	3.0°
Glasner et al. [14]	-	-

(b) Pose Estmation: CMU Car

Below is the distribution of angular errors on VOC2007 Car Viewpoint (left) and CMU-car (right). We achieve a median distance of 2 bins, which is an improvement over previous state-of-the-art of 3 bins.

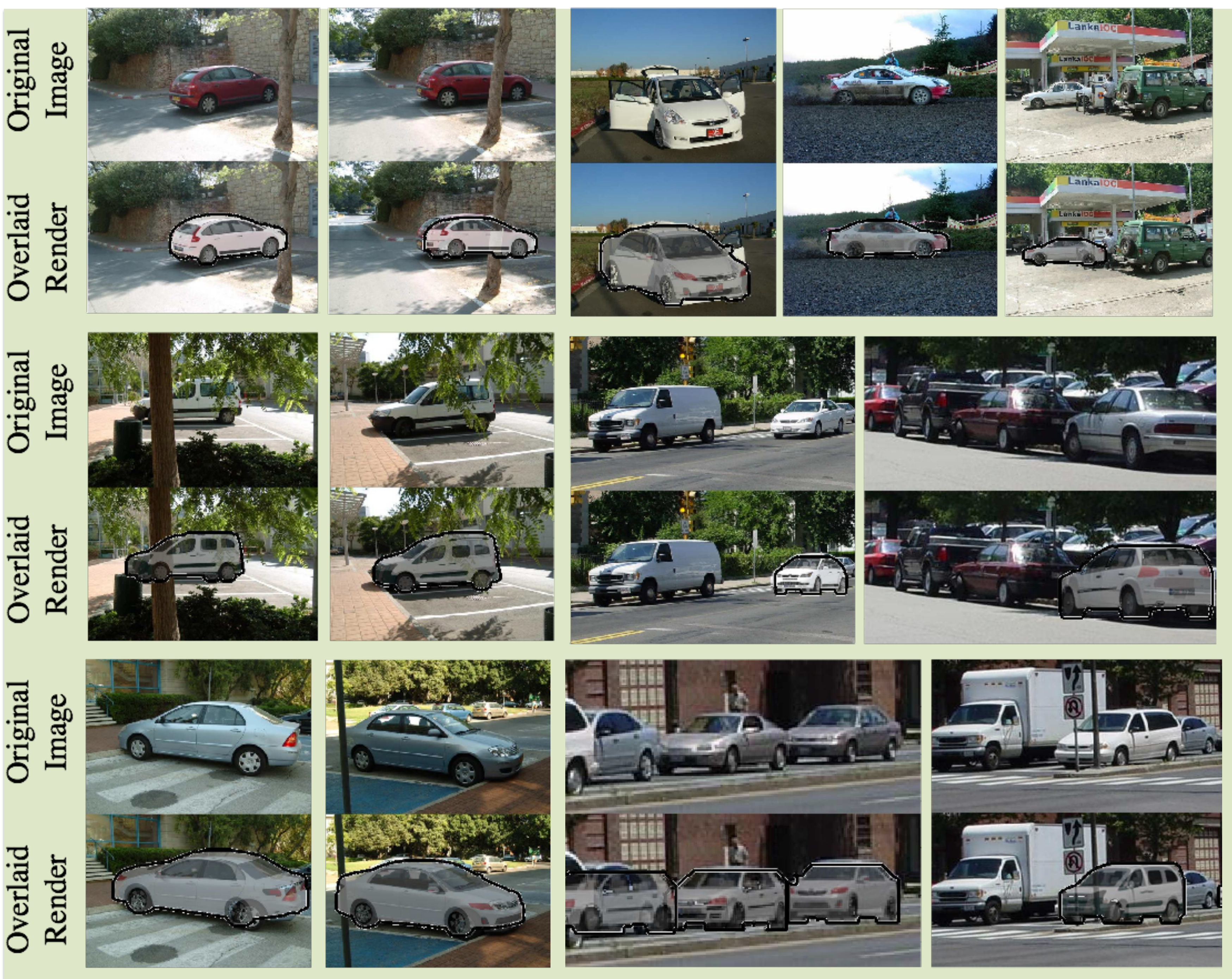




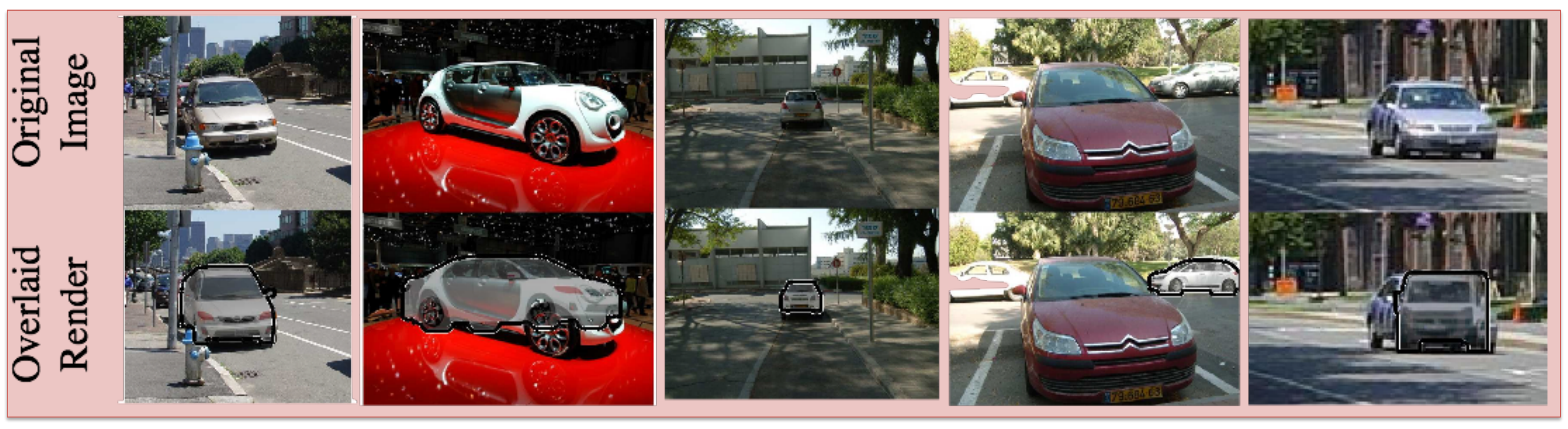
6. Overlaid Results

Each row shows input images (top) and overlaid pose estimation results (bottom).

Successful estimation:

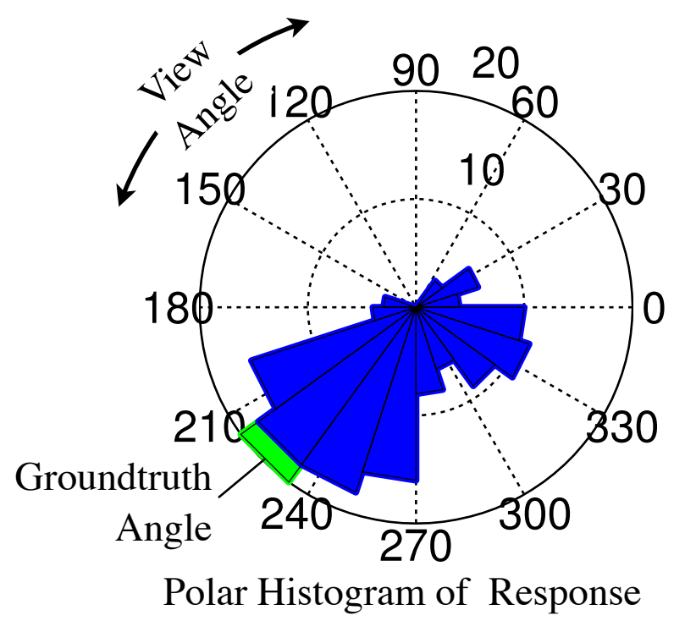


Failure cases:



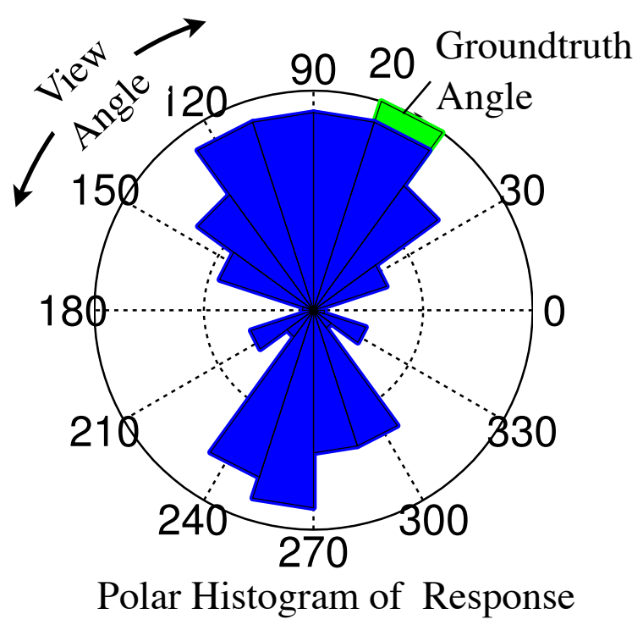
7. Polar histograms

Polar histogram of scores. The left example shows a van at an oblique angle, with little ambiguity in the distribution of responses. The right example shows a side view with a distinctive symmetric ambiguity.



Polar Histogram of Response

Input Image



Polar Histogram of Response

Input Image