Supplementary Material for Joint 3D Estimation of Objects and Scene Layout

Andreas Geiger
Karlsruhe Institute of Technology
geiger@kit.edu

Christian Wojek
MPI Saarbrücken
cwojek@mpi-inf.mpg.de

Raquel Urtasun
TTI Chicago
rurtasun@ttic.edu

In this supplementary material, we first show how the 3D trackets can be generated from 2D observations. We then depict additional results for inference when θ is know and when it is unknown.

1 3D Tracklet Generation from 2D Observations

This section describes the mapping

$$\varphi : f, \mathbf{b}, \mathcal{C} \rightarrow \boldsymbol{\pi}, \boldsymbol{\Sigma} \tag{1}$$

which takes a frame index $f \in \mathbb{N}$, an object bounding box $\mathbf{b} \in \mathbb{R}^4$ and the calibration parameters $\mathcal{C}$, and maps it to the object location $\boldsymbol{\pi} \in \mathbb{R}^2$ and uncertainty $\boldsymbol{\Sigma} \in \mathbb{R}^{2\times 2}$ in bird's eye perspective. As cues for this mapping we use the bounding box width and height, as well as the location of the bounding box foot point. Scene depth adaptive error propagation is employed for obtaining $\boldsymbol{\Sigma}$. The unknown parameters of the mapping are the uncertainty in bounding box location σ_u, σ_v, width $\sigma_{\Delta u}$ and height $\sigma_{\Delta v}$ as well as the real-world object dimensions Δ_x, Δ_y along with their uncertainties $\sigma_{\Delta x}, \sigma_{\Delta y}$. Those parameters are learned from a separate stereo image training dataset, including 1020 images with 3634 manually labeled vehicles in total.

Let (u, v) be the image coordinate of the lower middle point of the object's bounding box obtained by a classical 2D object detector. Similarly, let $(\Delta u, \Delta v)$ be the width and height of the bounding box. Further, let (x, y, z) be the 3D coordinates of an object (alignment: right, down, forward) and let $(\Delta x, \Delta y)$ be the object width and height in meters, measured via parallel-projection to the plane $z = 0$, which is coplanar to the image plane.

As $u, v, \Delta u, \Delta v$ are observations from the object detector, we are interested in

$$\begin{aligned} & p(u, v, \Delta u, \Delta v | y, z, \Delta x, \Delta y) \\ = \; & p(u, v | x, z, \Delta x, \Delta y)\; p(\Delta u | x, z, \Delta x, \Delta y)\; p(\Delta v | x, z, \Delta x, \Delta y) \\ = \; & p(u, v | x, z)\; p(\Delta u | z, \Delta x)\; p(\Delta v | z, \Delta y) \\ \propto \; & p(x, z | u, v)\; p(z | \Delta u, \Delta x)\; p(z | \Delta v, \Delta y) \end{aligned} \tag{2}$$

where in the last line we have made the assumption of a uniform prior over x and z. We will now examine each of these terms individually.

1.1 Estimating $p(x, z|u, v)$

We assume $p(x, z|u, v) = \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and the standard pinhole projection model

$$\begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = \mathbf{P}^{3\times 4} \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix}$$

where $\mathbf{P} = \mathbf{KTR}$ is composed of a calibration matrix $\mathbf{K}^{3\times 3}$, a transformation from road plane coordinates to camera coordinates $\mathbf{T}^{3\times 4}$ which is estimated by RANSAC ground plane fitting and

an additional camera pitch error θ, parameterized by the rotation matrix $\mathbf{R}^{4\times4}(\theta)$. Given (u, v) we obtain (x, z) by solving the linear system

$$\mathbf{A}\begin{pmatrix}x\\z\end{pmatrix} = \mathbf{b}$$

with

$$\begin{aligned}
\mathbf{A}(u,v) &= \begin{pmatrix} up_{31}-p_{11} & u(p_{33}\cos\theta - p_{32}\sin\theta) - (p_{13}\cos\theta - p_{12}\sin\theta) \\ vp_{31}-p_{21} & v(p_{33}\cos\theta - p_{32}\sin\theta) - (p_{23}\cos\theta - p_{22}\sin\theta)\end{pmatrix} && (3)\\
&\overset{\theta=0}{=} \begin{pmatrix} up_{31}-p_{11} & up_{33}-p_{13} \\ vp_{31}-p_{21} & vp_{33}-p_{23}\end{pmatrix} && (4)\\
\mathbf{b}(u,v) &= \begin{pmatrix} p_{14}-up_{34} \\ p_{24}-vp_{34}\end{pmatrix} && (5)
\end{aligned}$$

where p_{ij} are elements of $\mathbf{P}$, with $\boldsymbol{\mu}_1 = \mathbf{A}^{-1}\mathbf{b}$. Here, we make use of the fact that the road plane is known and has been adjusted for within $\mathbf{P}$, hence the pitch error θ of $\mathbf{R}$ is zero on average, and only the variance of θ is considered in the following equations, which derive the covariance matrix $\boldsymbol{\Sigma}_1$.

Assuming the covariance of (u, v) is known, the covariance of (x, z) can be obtained by error propagation. Since the transformation is non-linear, we linearize it by means of a first-order Taylor series expansion. Assuming $\sigma_u, \sigma_v, \sigma_\theta$ to be given, the covariance of (x, z) can be computed as

$$\boldsymbol{\Sigma}_1 = \mathbf{J}\begin{pmatrix}\sigma_u^2 & 0 & 0\\ 0 & \sigma_v^2 & 0\\ 0 & 0 & \sigma_\theta^2\end{pmatrix}\mathbf{J}^T$$

with Jacobian

$$\mathbf{J} = \begin{pmatrix}\frac{\partial(\mathbf{A}^{-1}\mathbf{b})}{\partial u} & \frac{\partial(\mathbf{A}^{-1}\mathbf{b})}{\partial v} & \frac{\partial(\mathbf{A}^{-1}\mathbf{b})}{\partial \theta}\end{pmatrix}$$

where

$$\begin{aligned}
\partial(\mathbf{A}^{-1}\mathbf{b}) &= \partial\mathbf{A}^{-1}\mathbf{b} + \mathbf{A}^{-1}\partial\mathbf{b}\\
&= -\mathbf{A}^{-1}\partial\mathbf{A}\mathbf{A}^{-1}\mathbf{b} + \mathbf{A}^{-1}\partial\mathbf{b}\\
&= \mathbf{A}^{-1}\left(\partial\mathbf{b} - \partial\mathbf{A}\mathbf{A}^{-1}\mathbf{b}\right)
\end{aligned}$$

with

$$\begin{aligned}
\frac{\partial\mathbf{A}}{\partial u} &= \begin{pmatrix}p_{31} & p_{33}\\ 0 & 0\end{pmatrix} & \frac{\partial\mathbf{b}}{\partial u} &= \begin{pmatrix}-p_{34}\\ 0\end{pmatrix}\\
\frac{\partial\mathbf{A}}{\partial v} &= \begin{pmatrix}0 & 0\\ p_{31} & p_{33}\end{pmatrix} & \frac{\partial\mathbf{b}}{\partial v} &= \begin{pmatrix}0\\ -p_{34}\end{pmatrix}\\
\frac{\partial\mathbf{A}}{\partial v} &= \begin{pmatrix}0 & p_{12}-up_{32}\\ 0 & p_{22}-vp_{32}\end{pmatrix} & \frac{\partial\mathbf{b}}{\partial v} &= \begin{pmatrix}0\\ 0\end{pmatrix}
\end{aligned}$$

at $\theta = 0$.

1.2 Estimating $p(z|\Delta u, \Delta x)$

We assume $p(z|\Delta u, \Delta x) = \mathcal{N}(\mu_2, \sigma_2^2)$. From the pinhole model we have

$$\Delta u = \frac{f\Delta x}{z}$$

or equivalently

$$\mu_2 = z = \frac{f\Delta x}{\Delta u}$$

which again is a non-linear function, this time in Δu. Using the same reasoning from above, we obtain the variance in z as

$$\sigma_2^2 = \mathbf{J}\begin{pmatrix}\sigma_{\Delta u}^2 & 0\\ 0 & \sigma_{\Delta x}^2\end{pmatrix}\mathbf{J}^T$$

with Jacobian

$$\mathbf{J} = \begin{pmatrix}-\frac{f\Delta x}{(\Delta u)^2} & \frac{f}{\Delta u}\end{pmatrix}$$

1.3 Estimating $p(z|\Delta v, \Delta y)$

Similarly, we can write

$$\mu_3 = z = \frac{f\Delta y}{\Delta v}$$

with variance

$$\sigma_3^2 = \mathbf{J} \begin{pmatrix} \sigma_{\Delta v}^2 & 0 \\ 0 & \sigma_{\Delta y}^2 \end{pmatrix} \mathbf{J}^T$$

and Jacobian

$$\mathbf{J} = \begin{pmatrix} -\frac{f\Delta y}{(\Delta v)^2} & \frac{f}{\Delta v} \end{pmatrix}$$

1.4 Parameter Learning

The unknown parameters for our tracklet model are $\sigma_u, \sigma_v, \sigma_{\Delta u}, \sigma_{\Delta v}$ as well as $\Delta x, \Delta y, \sigma_{\Delta x}, \sigma_{\Delta y}$. A principled way to obtain these values is to estimate them automatically from training data. We do this by building a dataset of 1020 images including 3634 manually annotated vehicles and disparity maps. Our labels include a tightly fitting bounding box as well as the heading of the car, quantized into 8 orientations.

We first estimate the parameters related to detection accuracy $\sigma_u, \sigma_v, \sigma_{\Delta u}, \sigma_{\Delta v}$ by comparing the object detections to manually labeled bounding boxes. Due to the characteristics of sliding-window detectors, we expect the noise to be dependent on the object scale. A good approximation to object scale is the bounding box height Δv, since – in contrast to the bounding box width – it is largely invariant with respect to car orientation. Furthermore, it is readily given by the object detector. Figure 1 depicts $\sigma_u, \sigma_v, \sigma_{\Delta u}, \sigma_{\Delta v}$ as a function of Δv, computed from the differences between object detections and manually labeled bounding boxes. As expected the noise behaviour is approximately linear in Δv, thus we represent this dependency via a linear least-squares fit.

$\Delta x, \Delta y, \sigma_{\Delta x}, \sigma_{\Delta y}$ can be easily computed from our annotated data in conjunction with known stereo disparity via

$$\Delta x = \frac{z\Delta u}{f} = \frac{b\Delta u}{d} \qquad \Delta y = \frac{z\Delta v}{f} = \frac{b\Delta v}{d}$$

where b is the camera baseline and d is the median value of all valid disparities within the bounding box. Here Δu and Δv are given by the manually labeled ground truth. Special care has to be taken for Δx, as it strongly depends on the orientation of the vehicle. We tackle this problem by learning a separate Δx for each of the eight canonical car orientations. Figures 2 and 3 show the results: When viewed from behind or frontally, a typical car appears ~ 2 meters wide, while it appears ~ 4.4 meters wide when viewed from the side.

1.5 Putting it together

Multiplying the terms in the last row of Eq. 2 leads to

$$x, z|u, v, \Delta u, \Delta v \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

with

$$\boldsymbol{\Lambda} = \boldsymbol{\Lambda}_1 + \boldsymbol{\Lambda}_2 + \boldsymbol{\Lambda}_3 \tag{6}$$

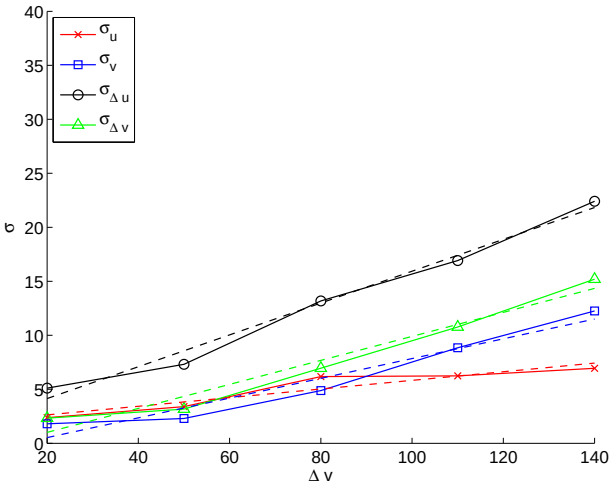
$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}^{-1} \tag{7}$$

$$\boldsymbol{\mu} = \boldsymbol{\Sigma}\boldsymbol{\Lambda}_1\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}\boldsymbol{\Lambda}_2\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}\boldsymbol{\Lambda}_3\boldsymbol{\mu}_3 \tag{8}$$

Here $\boldsymbol{\Lambda}$ is a precision matrix, the index denotes the corresponding term, $\boldsymbol{\Lambda}_1$ has full rank and $\boldsymbol{\Lambda}_2, \boldsymbol{\Lambda}_3$ are singular matrices of the form

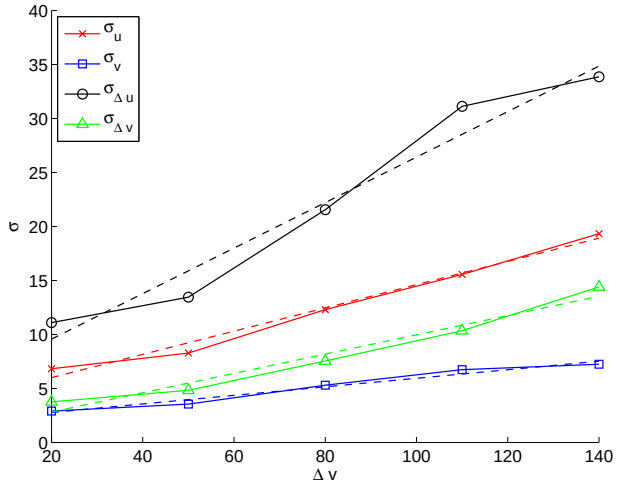
$$\boldsymbol{\Lambda}_2 = \begin{pmatrix} 0 & 0 \\ 0 & \lambda_2 \end{pmatrix} \qquad \boldsymbol{\Lambda}_3 = \begin{pmatrix} 0 & 0 \\ 0 & \lambda_3 \end{pmatrix}$$

Resulting depth probabilities for 3 tracklet detections are depicted in Fig. 4. Here the colored curves are the individual cues discussed in the previous sections and the black curves are the combined results.



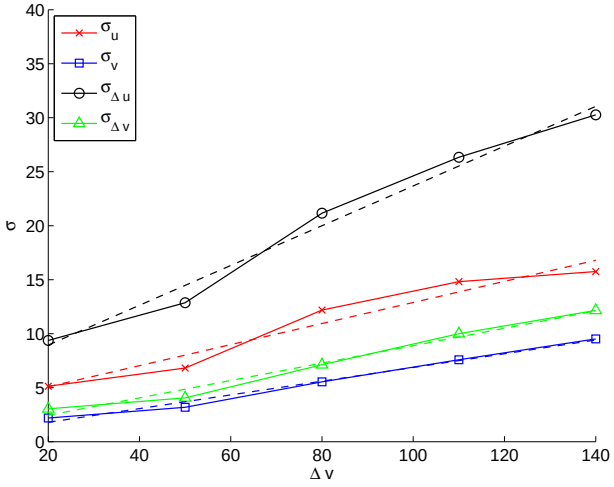
(a) Heading 1,2

	a	b
σ_u	0.0400	1.8239
σ_v	0.0915	-1.3080
$\sigma_{\Delta u}$	0.1474	1.1972
$\sigma_{\Delta v}$	0.1111	-1.2124



(b) Heading 3,4

	a	b
σ_u	0.1076	3.8502
σ_v	0.0396	1.9848
$\sigma_{\Delta u}$	0.2107	5.3625
$\sigma_{\Delta v}$	0.0893	1.0360



(c) Heading 5,..,8

	a	b
σ_u	0.0975	3.1407
σ_v	0.0635	0.5281
$\sigma_{\Delta u}$	0.1842	5.2666
$\sigma_{\Delta v}$	0.0806	0.8323

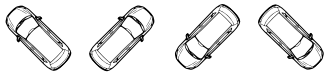


Figure 1: Object detection accuracies and line parameters ($\sigma = a\ \Delta v + b$)

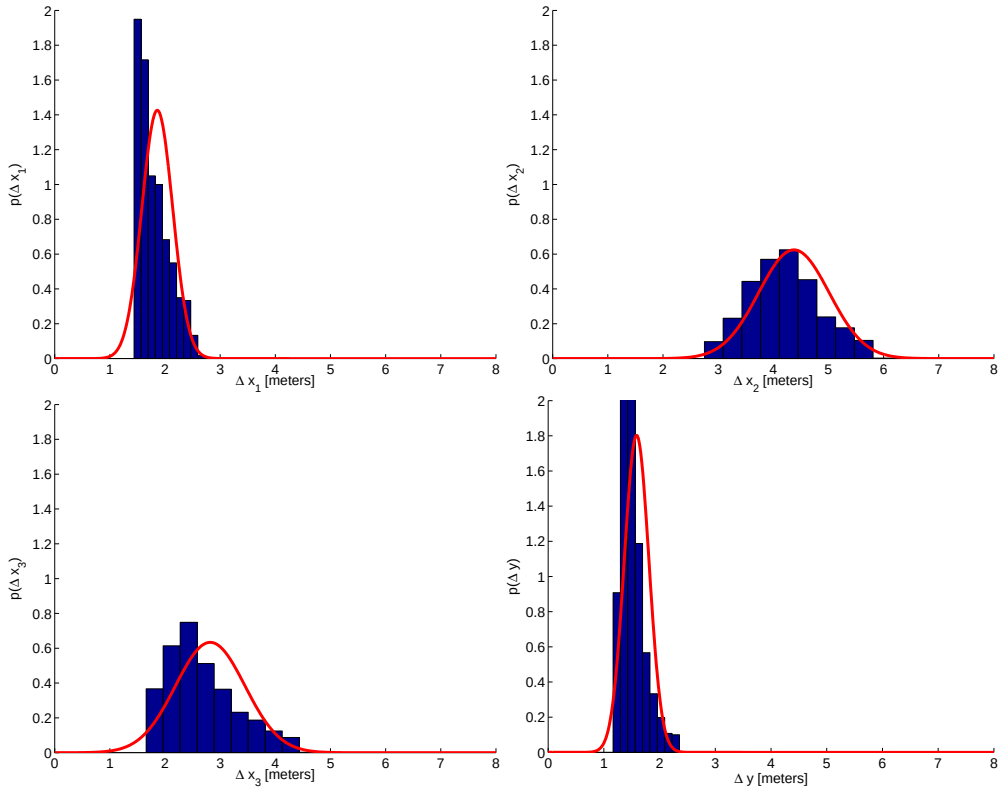
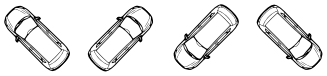


Figure 2: Object size probability density

	 Δx_1	 Δx_2	 Δx_3	Δy
μ	1.86	4.37	2.82	1.59
σ	0.28	0.64	0.63	0.22

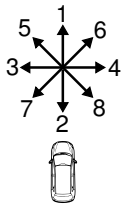


Figure 3: Estimated Object Parameters (Δx = width, Δy = height)

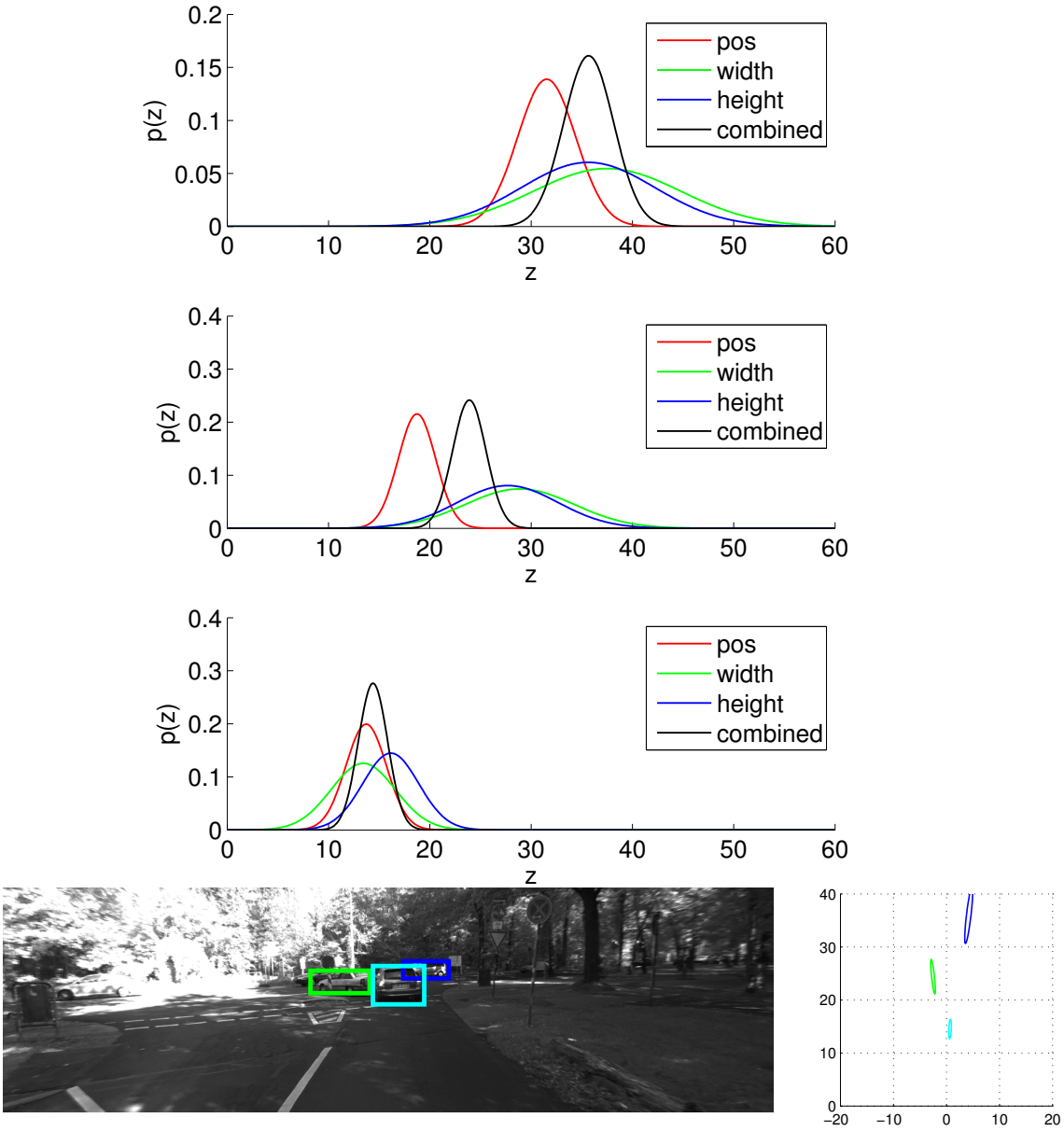


Figure 4: Example of depth probabilities estimated in this case from 3 detections

2 Inference Results

This section shows randomly selected inference results when θ is know and when θ is unknown on 2,3 and 4-armed intersections. Note that in some cases, even for humans, it is extremely difficult to distinguish 3-armed from 4-armed intersections using only monocular cues. For every row, 3D tracklets projected into the image are shown in the left, while results with θ known and θ unknown are shown in the next two images.

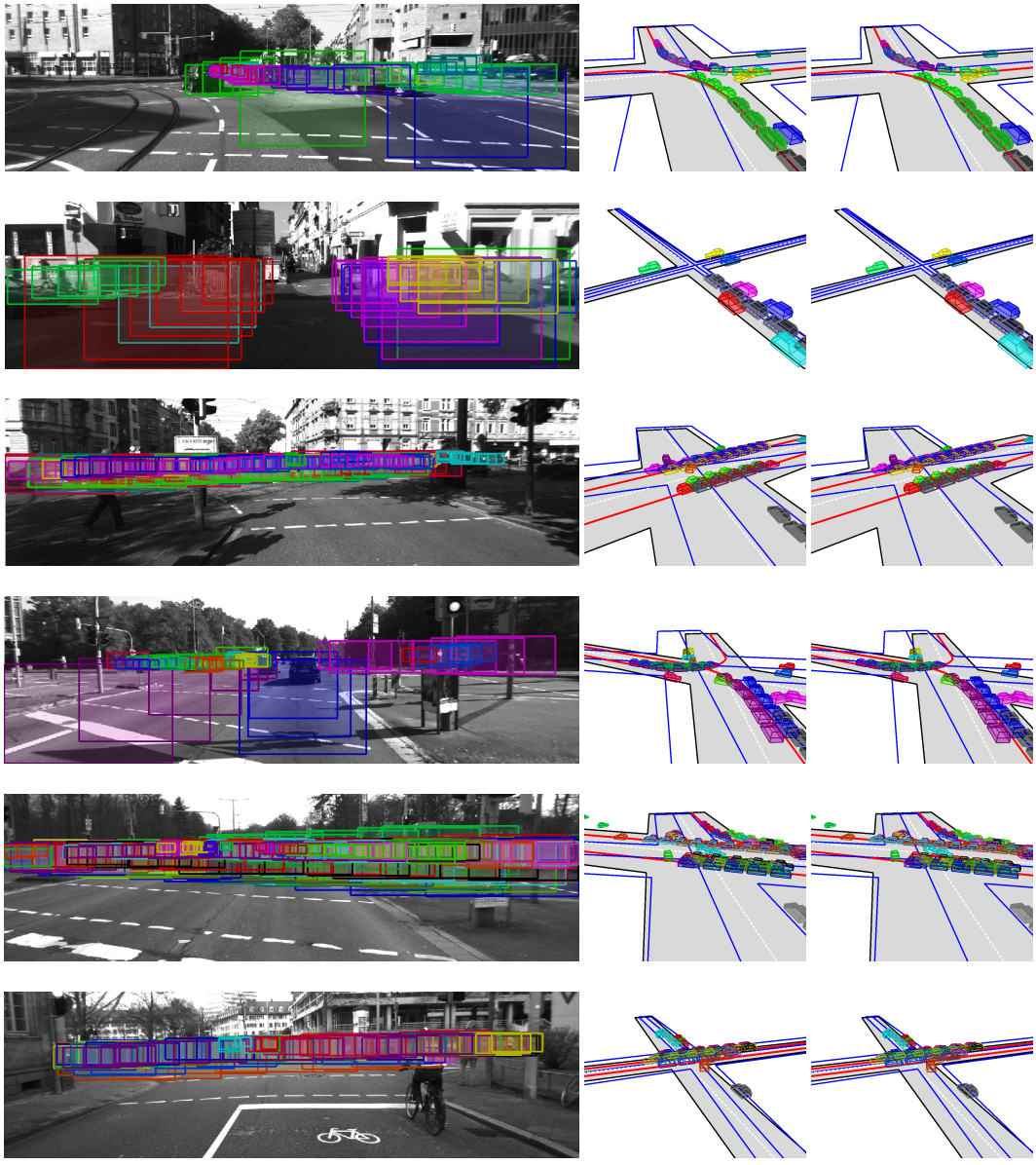


Figure 5: Automatically inferred scene descriptions for 4-armed intersections. (Left) Trackets from all frames are superimposed in the last frame of the sequence. (Middle) Inference result with θ known and (Right) θ unknown. Detections belonging to the same tracklet are grouped by color, the observer is depicted in black. Inferred activities are shown in red. The ground truth labels for the intersection layout are given in blue.

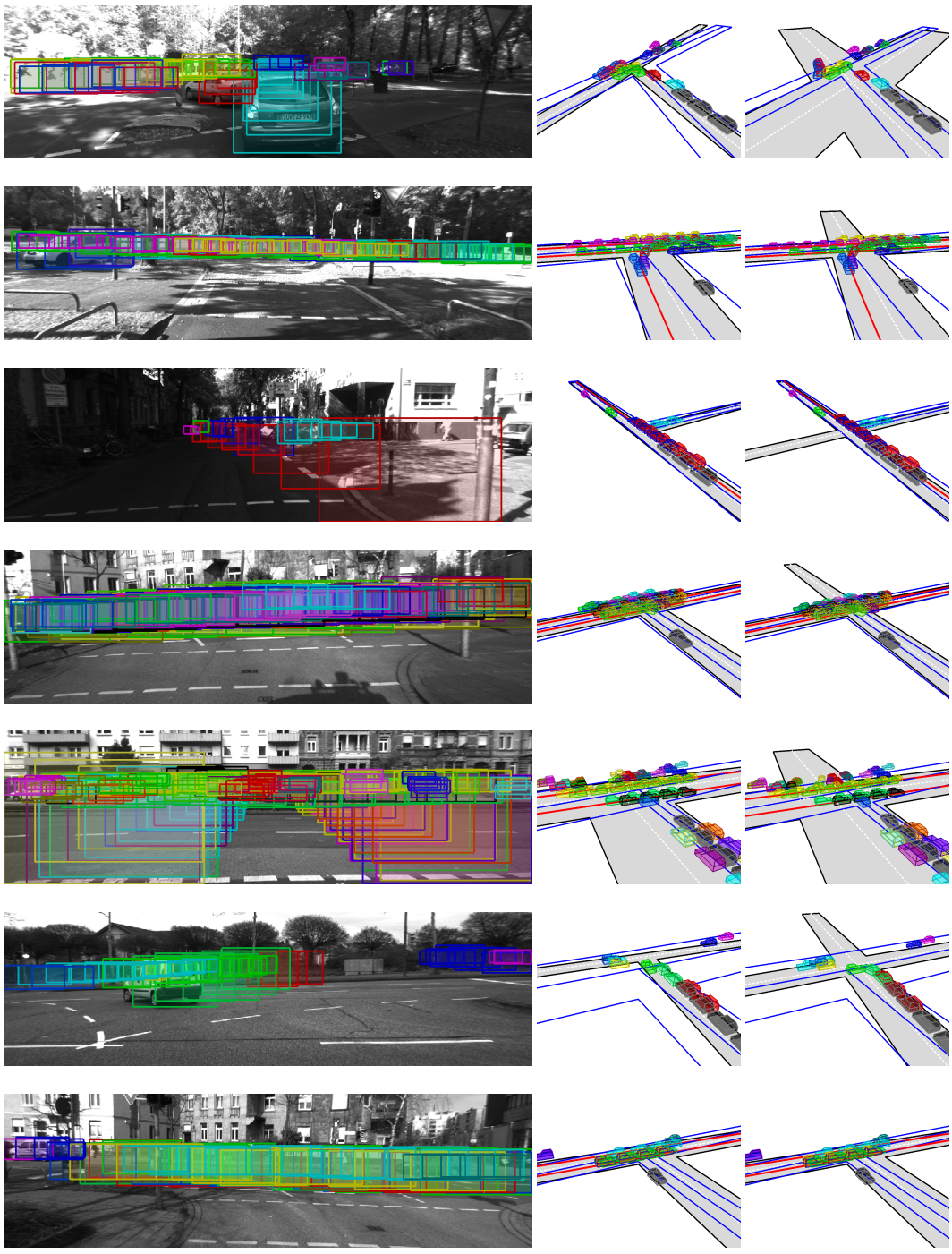


Figure 6: Automatically inferred scene descriptions for 3-armed intersections. (Left) Trackets from all frames are superimposed in the last frame of the sequence. (Middle) Inference result with θ known and (Right) θ unknown. Detections belonging to the same tracklet are grouped by color, the observer is depicted in black. Inferred activities are shown in red. The ground truth labels for the intersection layout are given in blue.

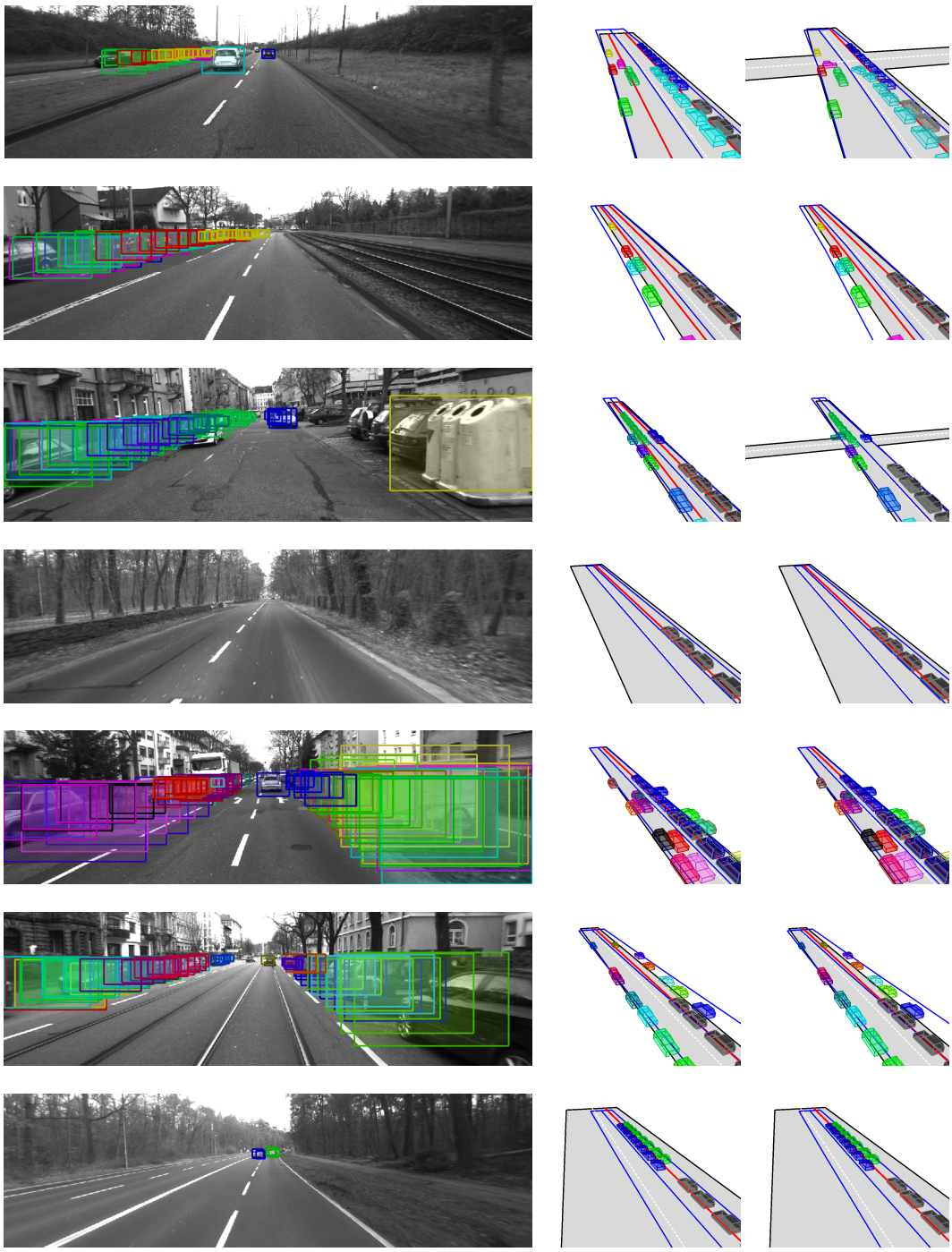


Figure 7: Automatically inferred scene descriptions for 2-armed intersections. (Left) Trackets from all frames are superimposed in the last frame of the sequence. (Middle) Inference result with θ known and (Right) θ unknown. Detections belonging to the same tracklet are grouped by color, the observer is depicted in black. Inferred activities are shown in red. The ground truth labels for the intersection layout are given in blue.