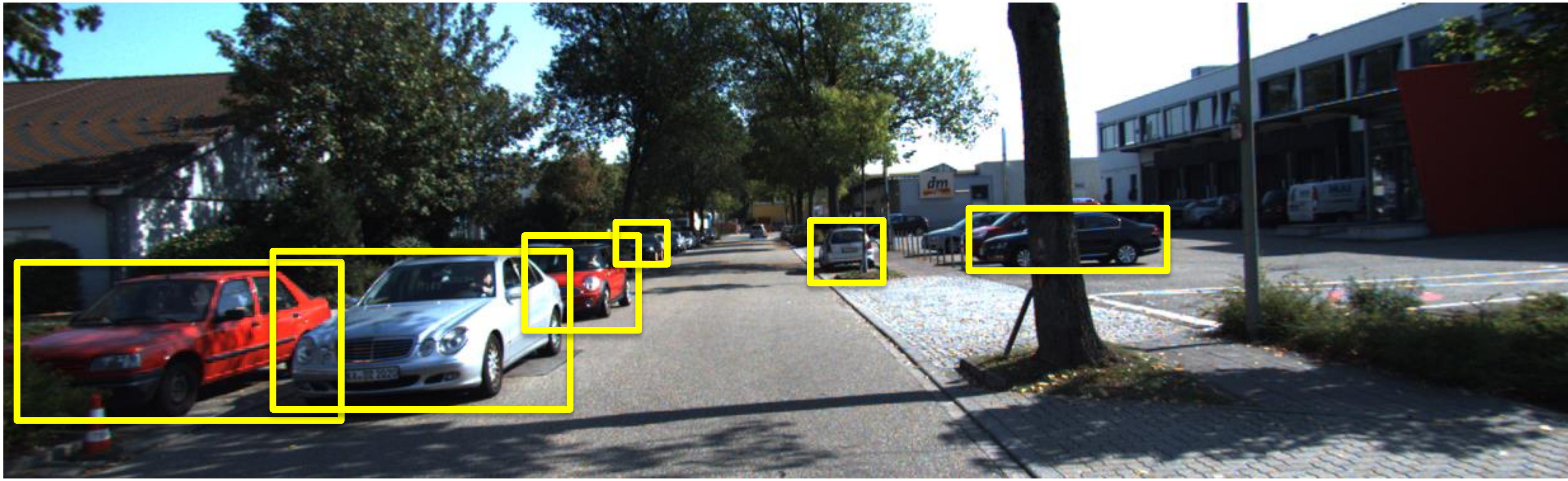
Continuous Models for Scene and Traffic Participant Interactions in Road Scene Understanding

Vikas Dhiman
SUNY at Buffalo

Mentor : Manmohan Chandraker

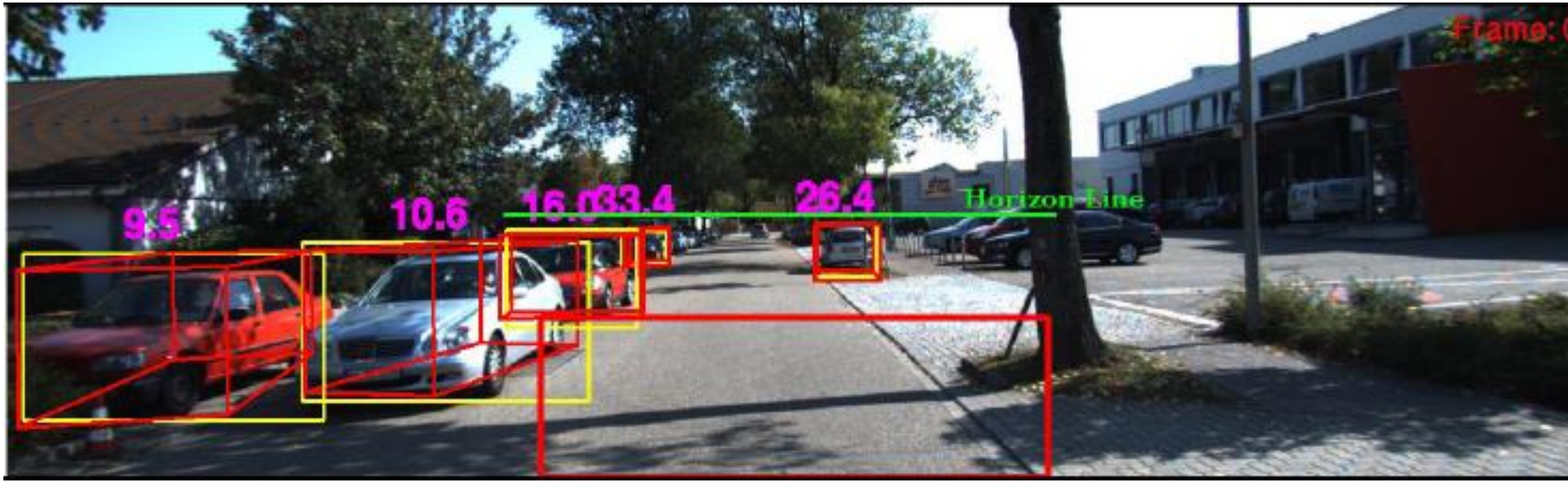
Monocular Road Scene Understanding

Monocular Road Scene Understanding



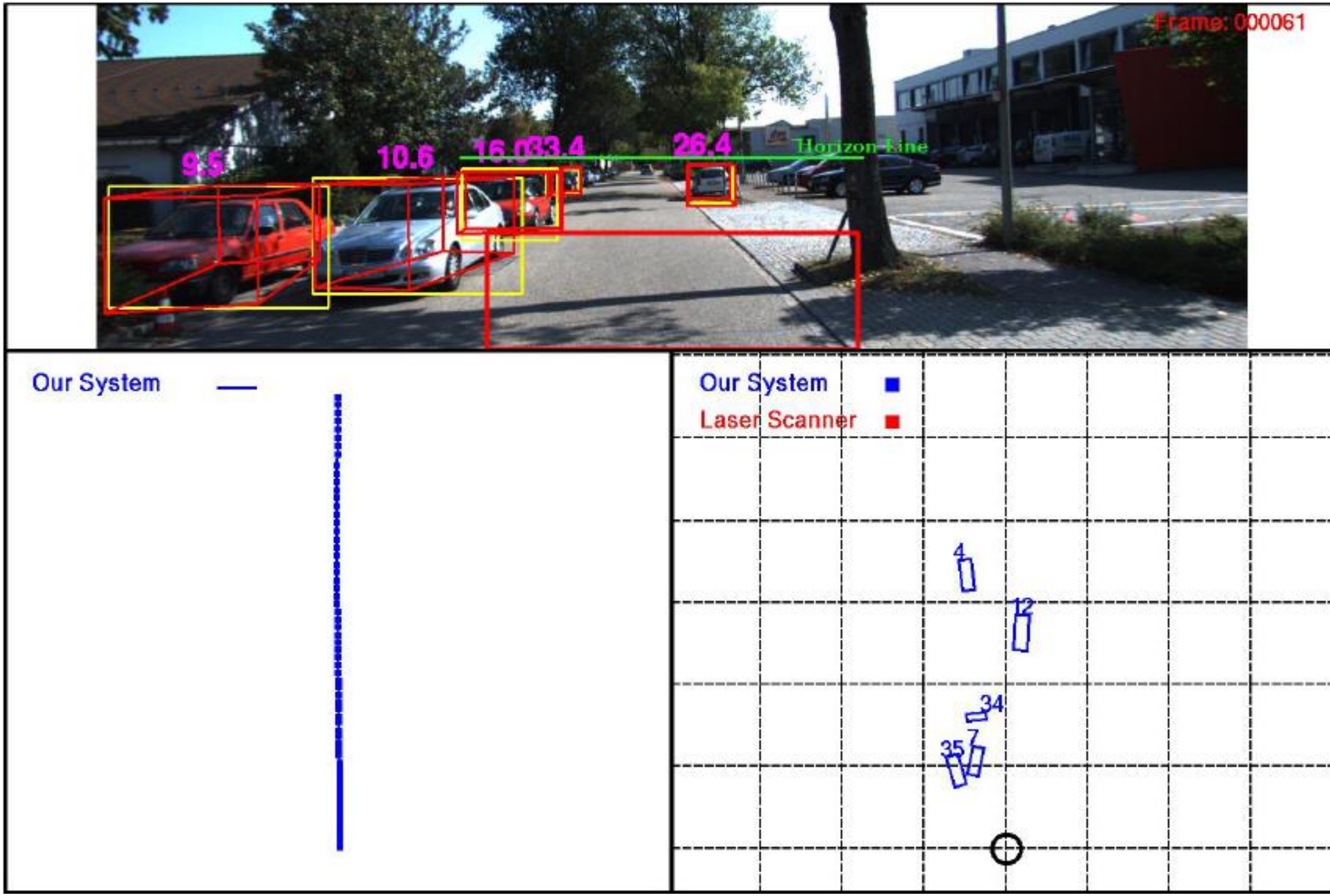
- Object detection: Detect various traffic participants (TP)

Monocular Road Scene Understanding



- Object detection: Detect various traffic participants (TP)
- Object localization: position and orientation of TPs in 3D

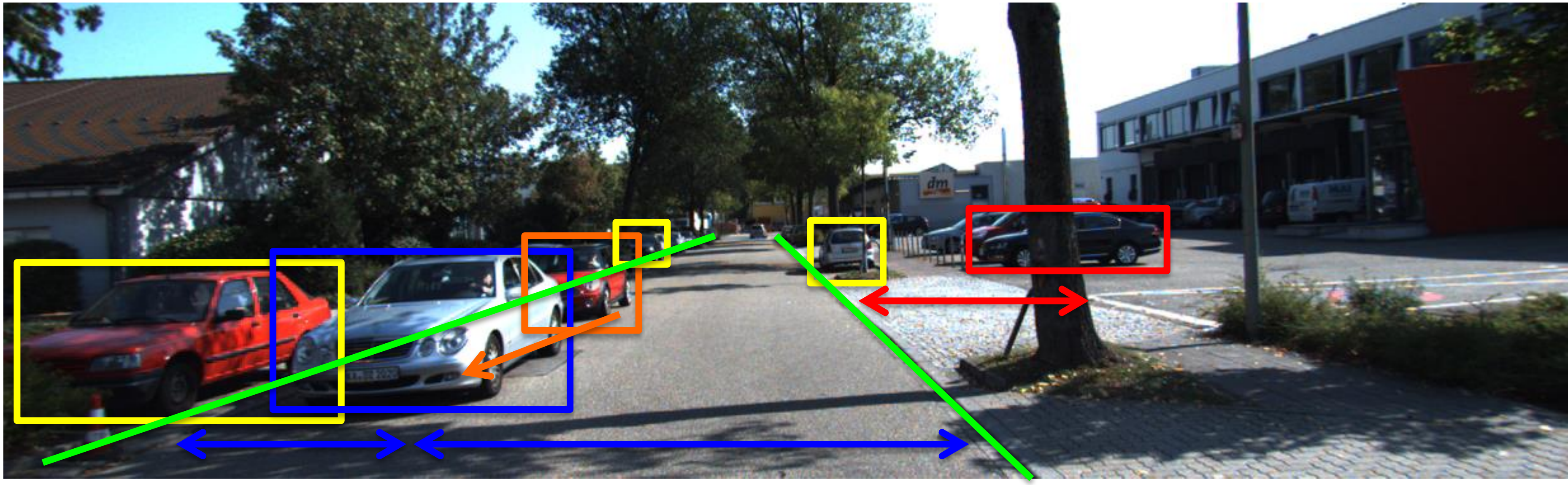
Monocular Road Scene Understanding



Monocular Road Scene Understanding

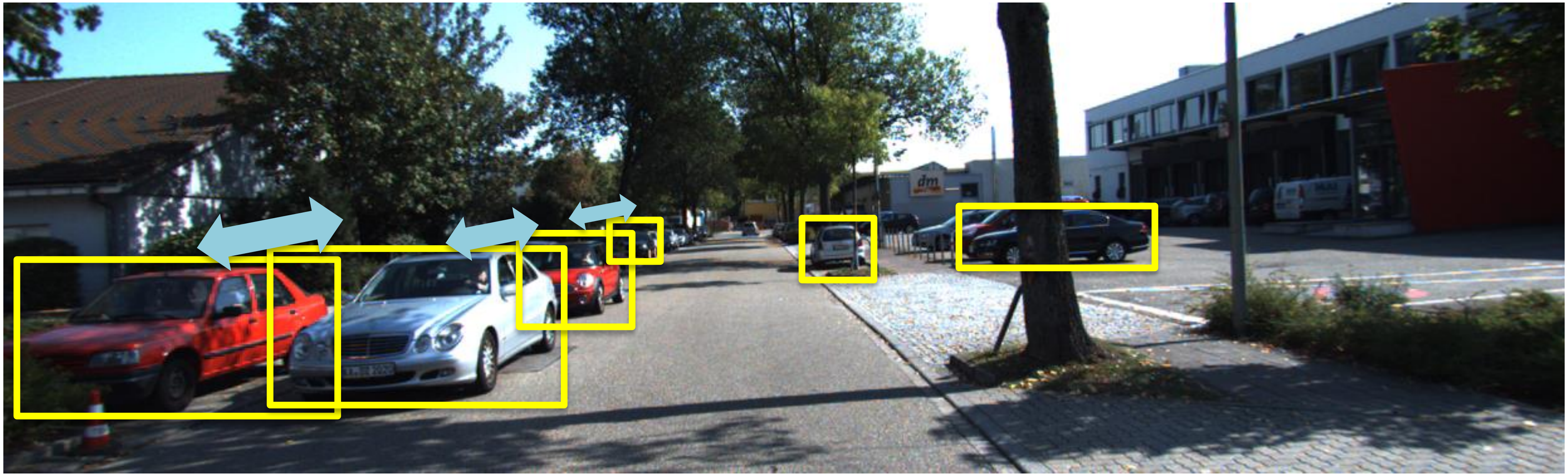
- Object detection: Detect various traffic participants (TP)
- Object localization: position and orientation of TPs in 3D
- Detect various scene elements (SE)

Monocular Road Scene Understanding



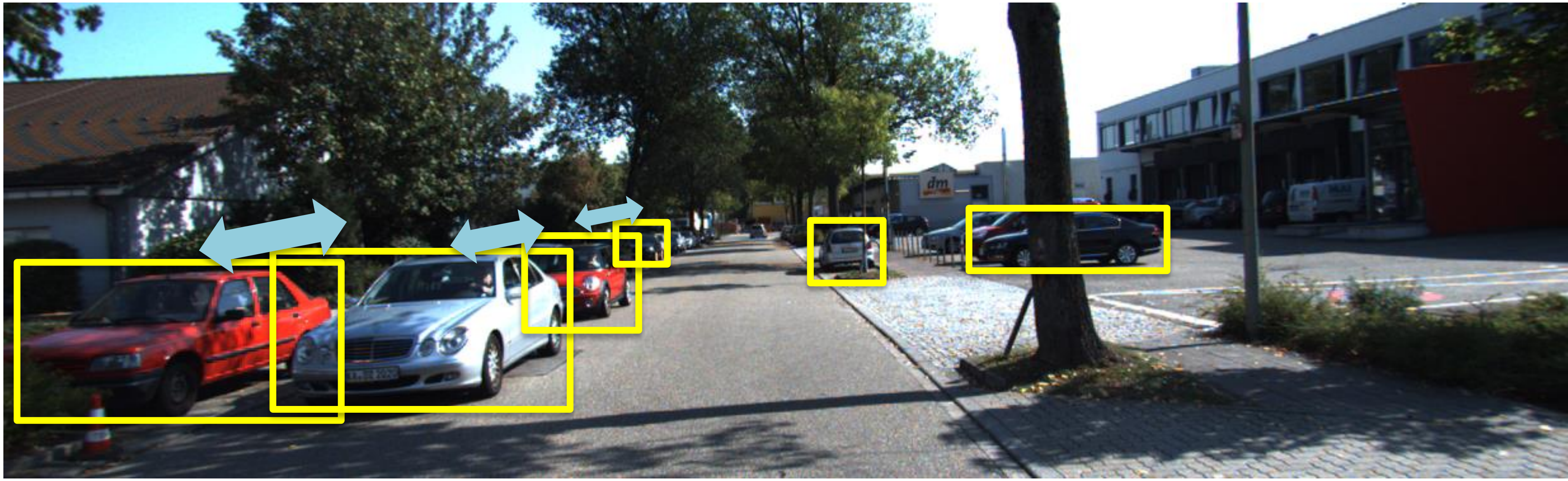
- Object detection: Detect various traffic participants (TP)
- Object localization: position and orientation of TPs in 3D
- Detect various scene elements (SE)
- Enforce relations between TPs and SEs

Monocular Road Scene Understanding



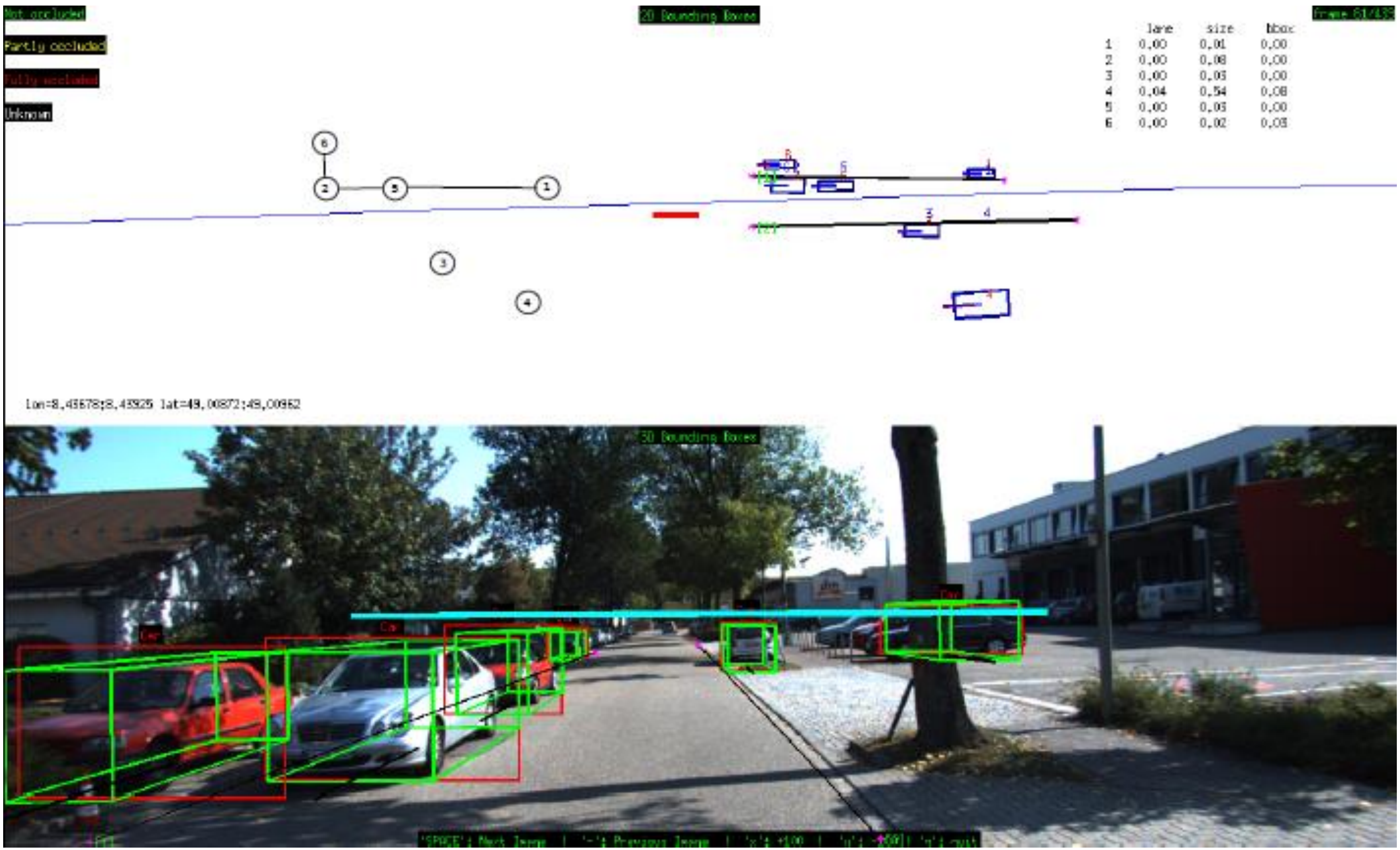
- Object detection: Detect various traffic participants (TP)
- Object localization: position and orientation of TPs in 3D
- Detect various scene elements (SE)
- Enforce relations between TPs and SEs
- Enforce relations between TPs

Monocular Road Scene Understanding

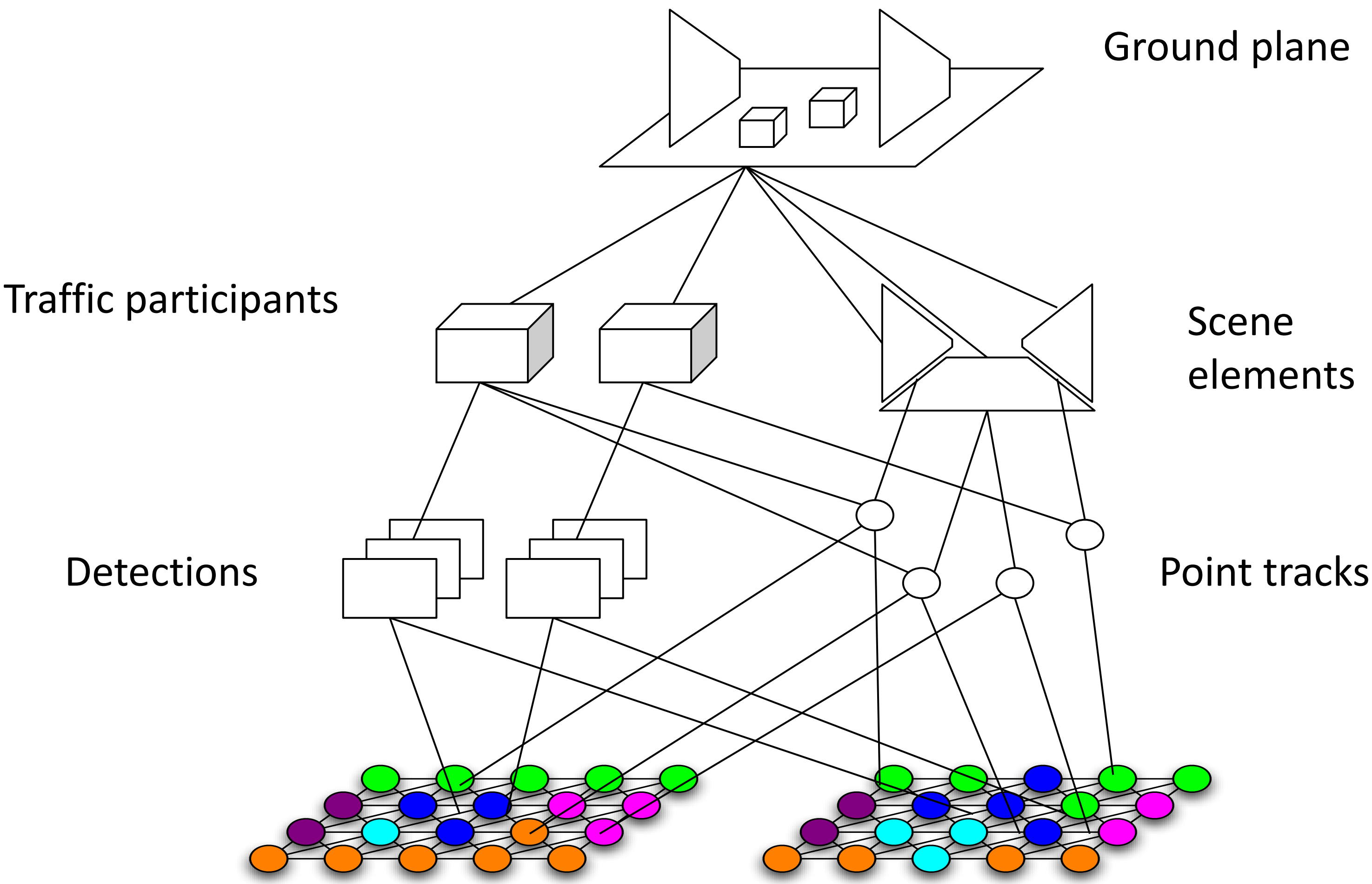


- Object detection: Detect various traffic participants (TP)
- Object localization: position and orientation of TPs in 3D
- Detect various scene elements (SE)
- Enforce relations between TPs and SEs
- Enforce relations between TPs
- Spatially and temporally consistent relationships.

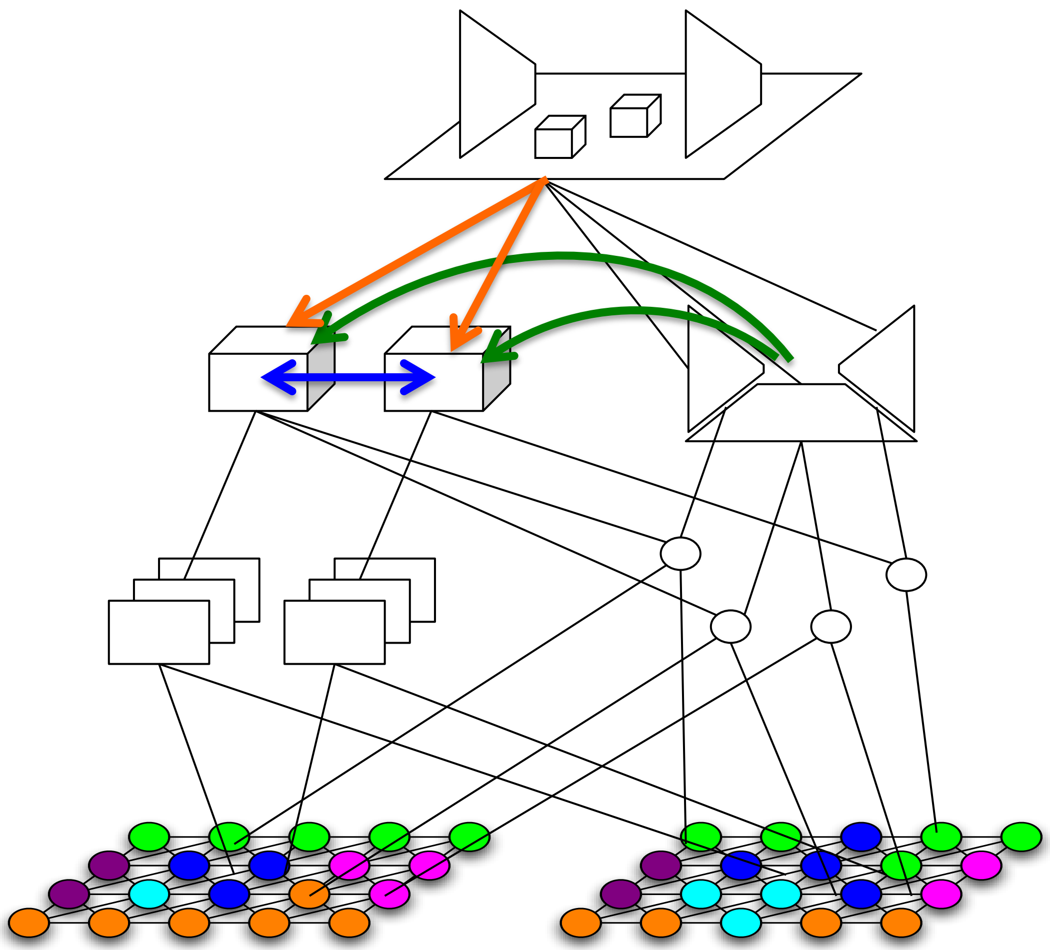
Monocular Road Scene Understanding



Relation to Overall Framework



Relation to Overall Framework



Prior Works

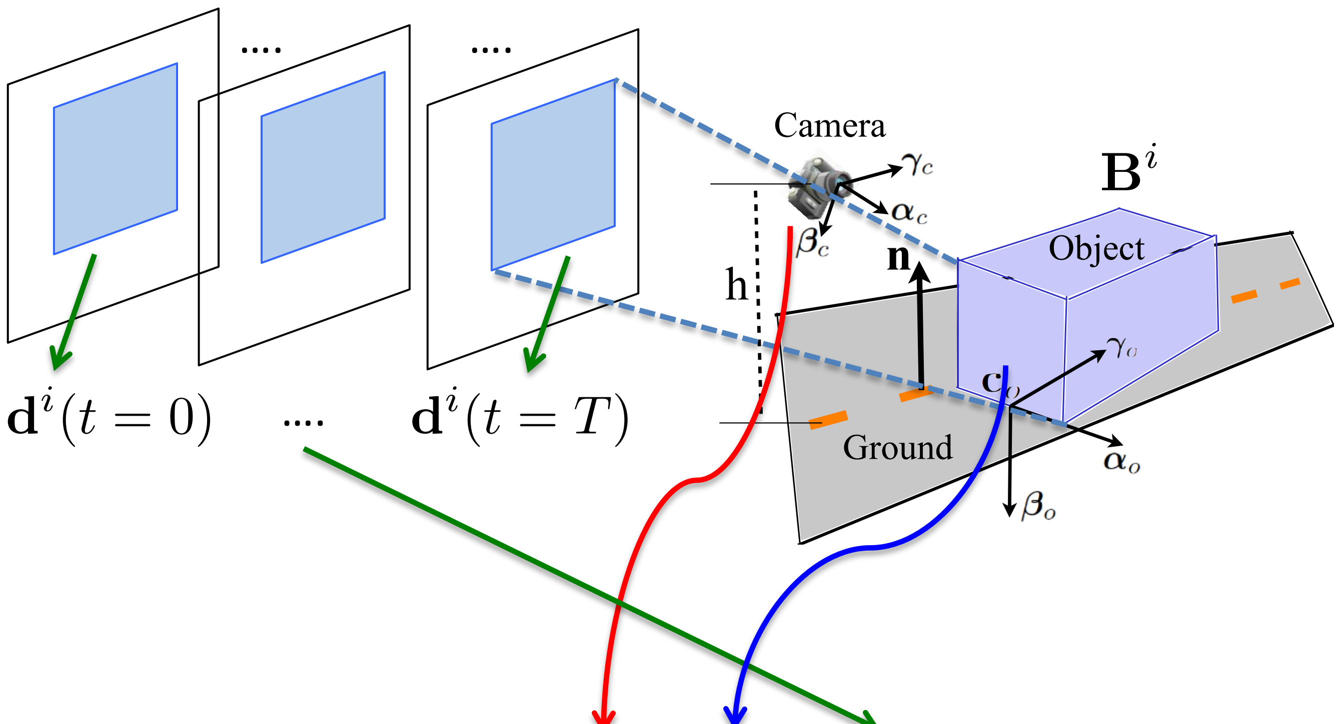
- Localize individual objects
  - [Wojek et al. 2013, Song and Chandraker 2014]
  - Cannot capture interactions
  - We model TP-Scene and TP-TP relationships
- Use stereo
  - [Ess et al. 2011, Geiger et al. 2013]
  - Dense depth information available from stereo
  - We use a single camera (monocular)
- Discontinuous occlusion modeling
  - [Zia et al. 2014]
  - Harder optimization, unpredictable output
  - We develop continuous occlusion models, which yields probabilistically meaningful interactions.

Input-Output

- Inputs:
  - Camera poses and ground plane from SFM
  - 2D object detection
  - Feature tracks on objects
  - GPS
- Outputs:
  - 3D object bounding boxes
  - Consistent TP-Scene relations
    - How objects relate to lane geometry
  - Consistent TP-TP relations
    - Occlusion relationships between objects
    - Which point belongs to which object.

Bounding Box Energy

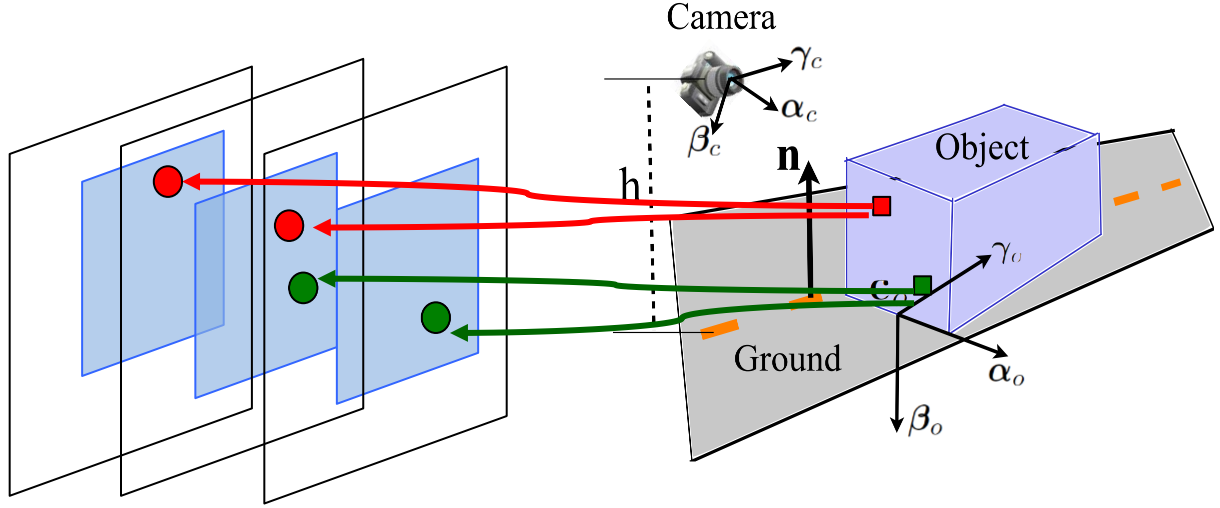
- Simpler version without occlusion
  - Uses prior size, contact of 2D bounding box with ground.



$$E_{bbox} = \|\pi_{\Omega^i(t)}(\mathbf{B}^i) - \mathbf{d}^i(t)\|^2$$

3D Points Energy

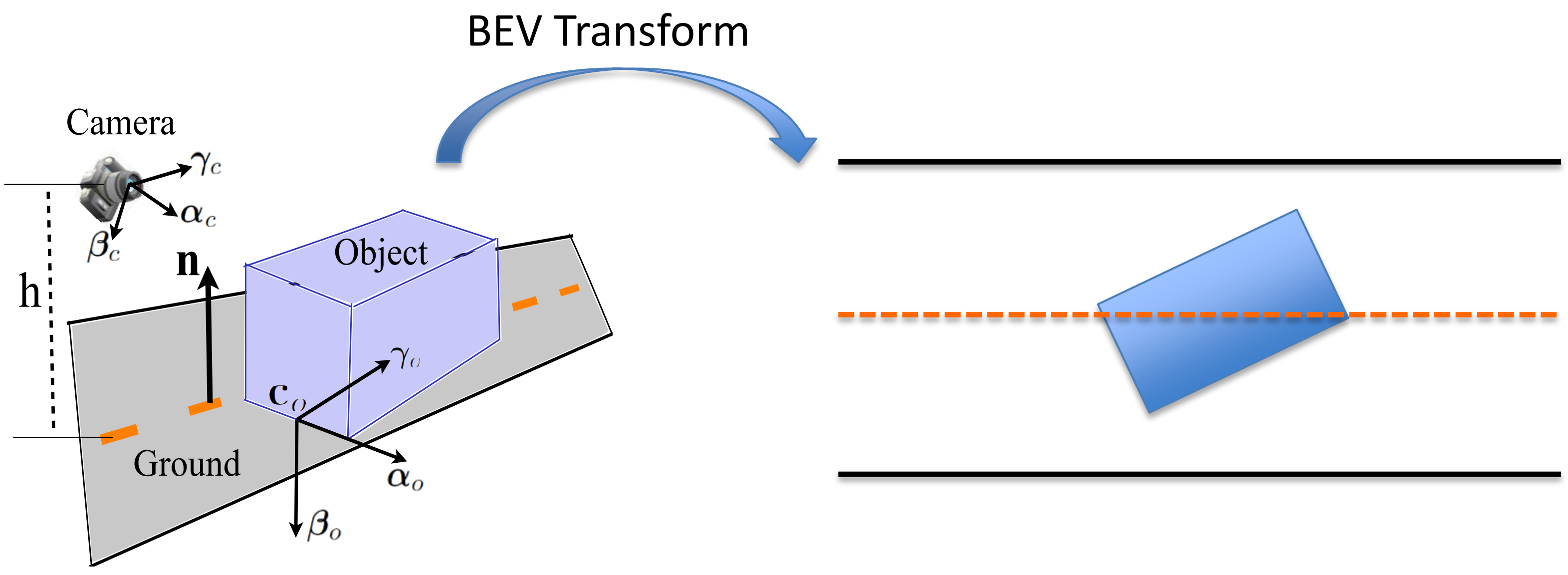
- Simpler version without occlusion
  - Backproject a point at time *t-1* to 3D bounding box
  - Compute reprojection error with observation at time *t*.



$$E_{track} = \sum_{j \in \text{tracks}} \| \mathbf{u}^j(t) - \pi_{\Omega^i(t)}(\pi^{-1}_{\Omega^i(t-1)}(\mathbf{u}_j(t-1)))\|^2$$

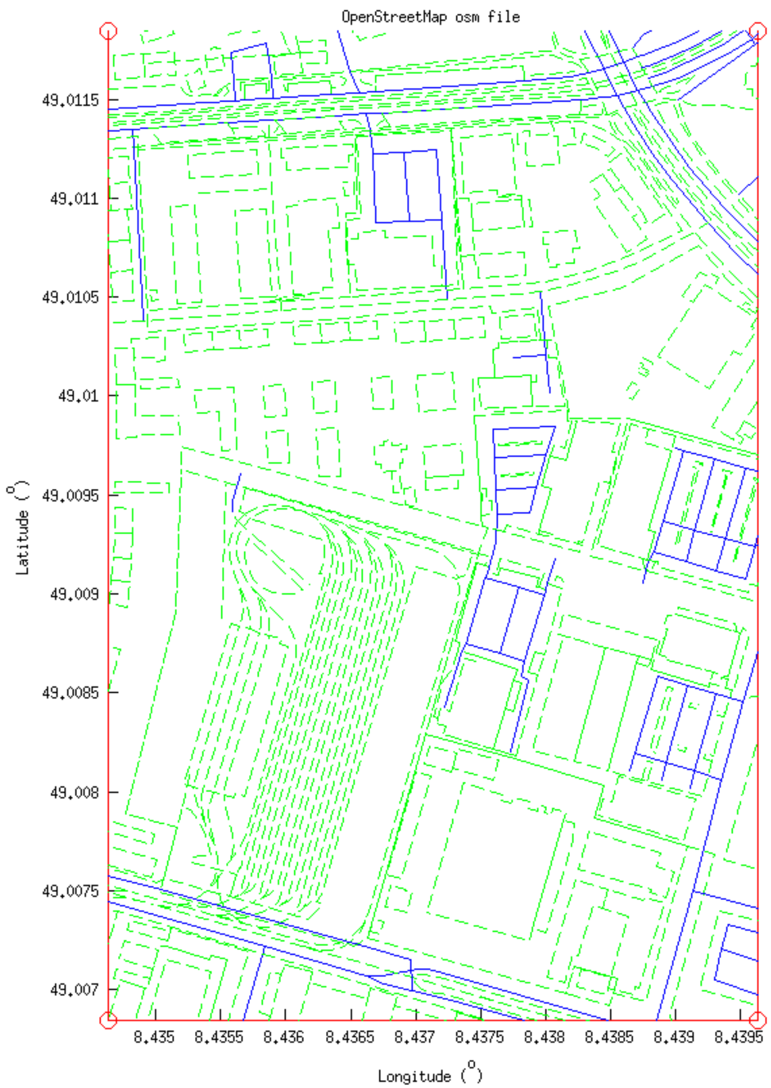
Bird-Eye View

- Use SFM camera pose and ground plane to represent each TP in BEV.



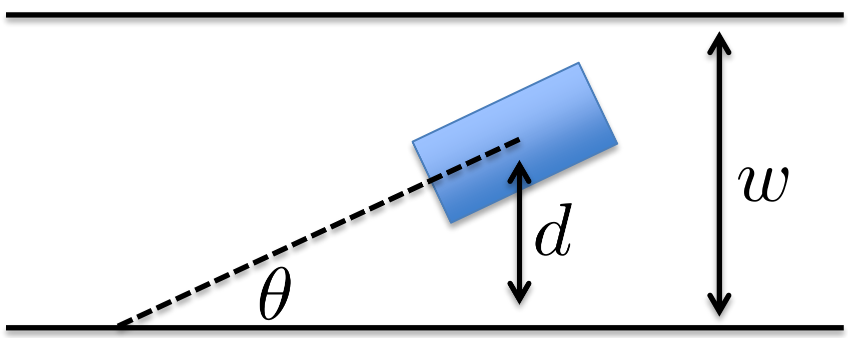
Extracting Scene Elements

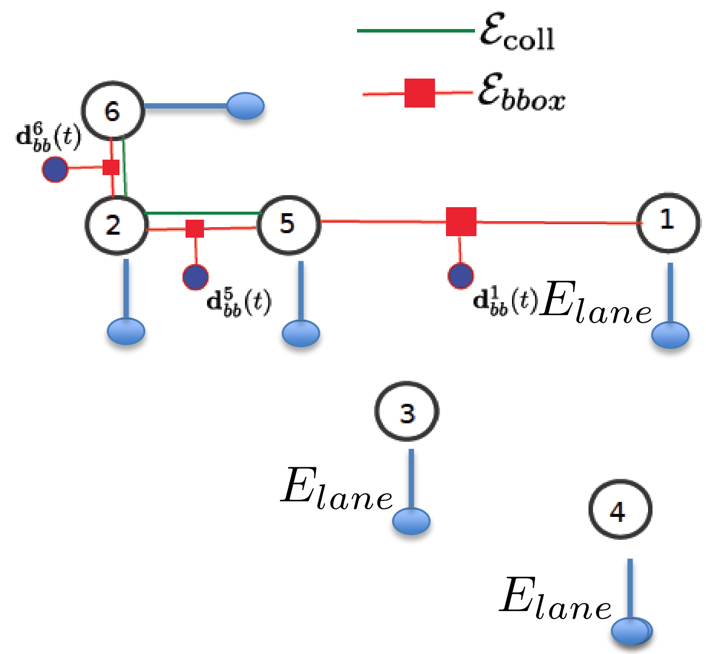
- Use OpenStreetMaps to extract lane geometry
  - Use GPS coordinates
  - Automatically filter out small lanes and side streets
- Annotated lanes (to be replaced by lane detector)
- Align SFM poses with lane geometry.



TP-Scene Constraints

- Lane position and orientation
  - filter away far objects
  - align objects with closest lane directions.

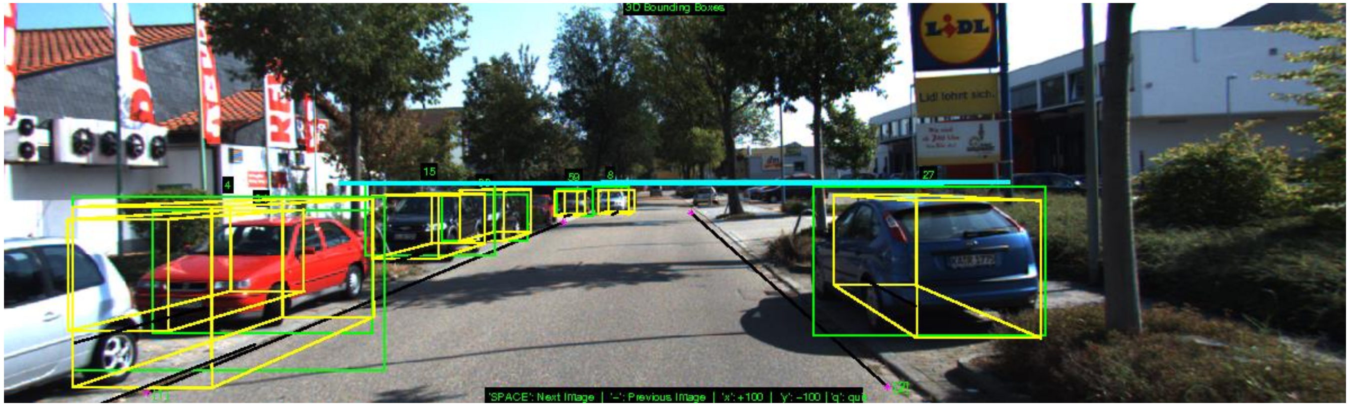


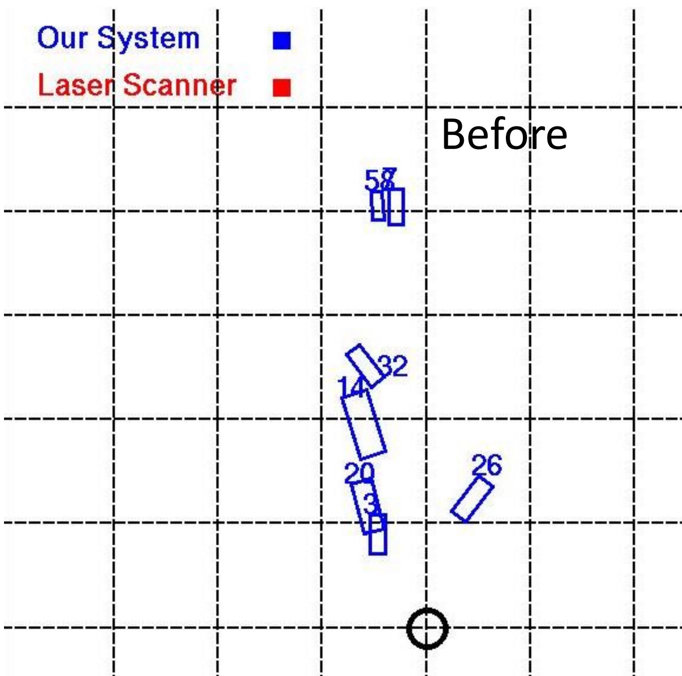


Energy for 1 lane: $\dfrac{1-\cos^2\theta}{1+\exp(-\lambda|w-d|)}$

Soft energy for closest lanes: $E_{lane} = \sum_{k:d_k<\tau} \dfrac{1-\cos^2\theta_k}{1+\exp(-\lambda|w-d_k|)}$

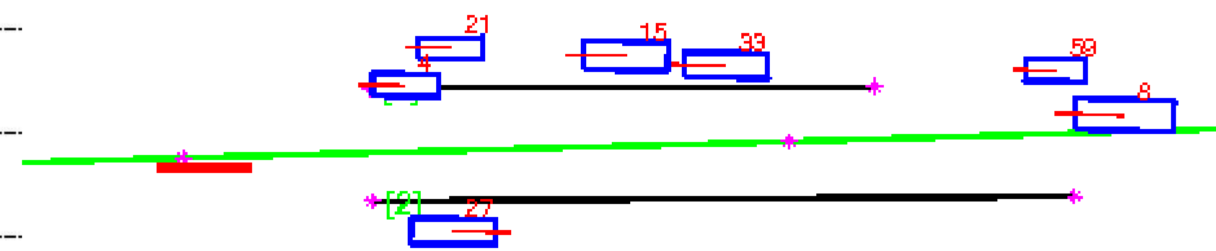
Effect of Lane Energy



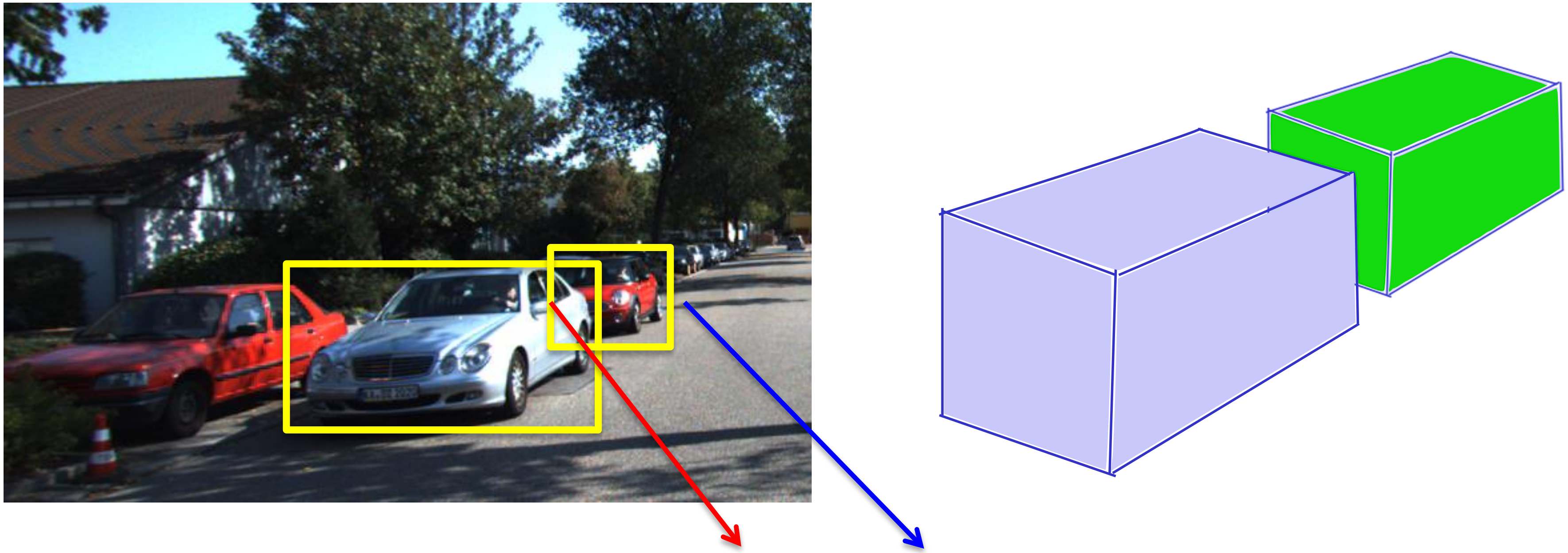


Before

After

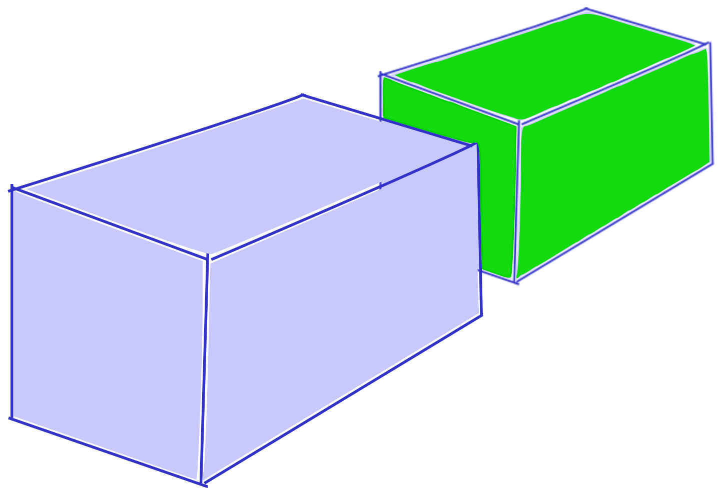


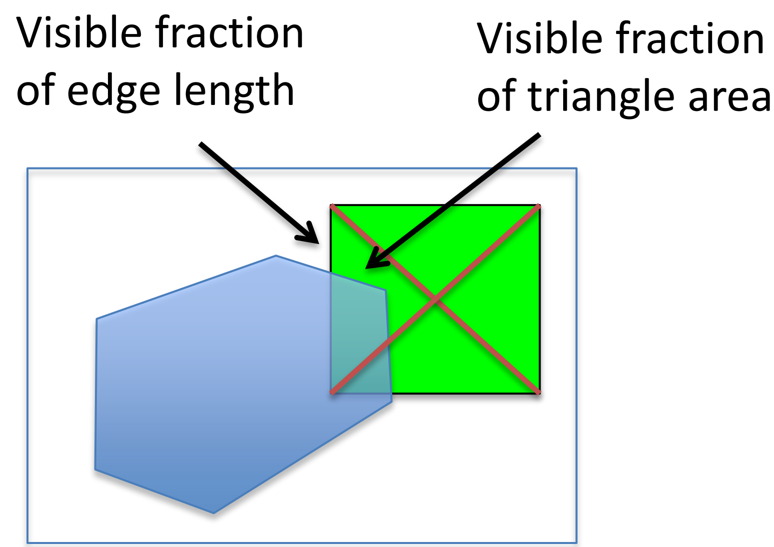
TP-TP Relation: Bounding Box Visibility



- Determine 3D bounding boxes aware of occlusions due to objects in front
- Encourage alignment for unoccluded edges
- Relax alignment for occluded edges.

TP-TP Relation: Bounding Box Visibility

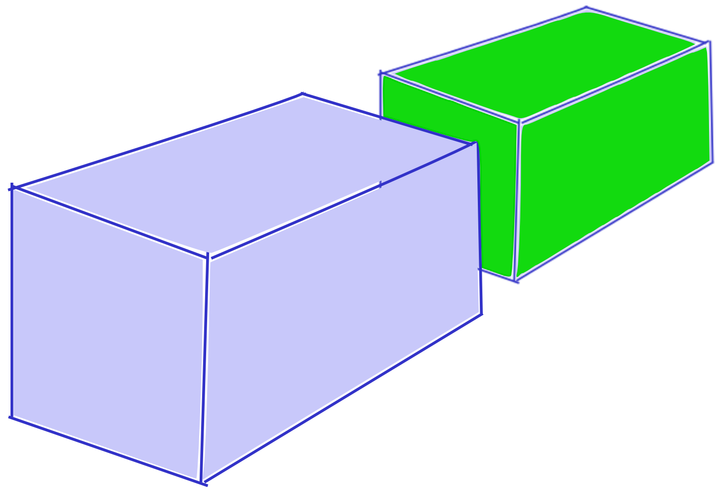


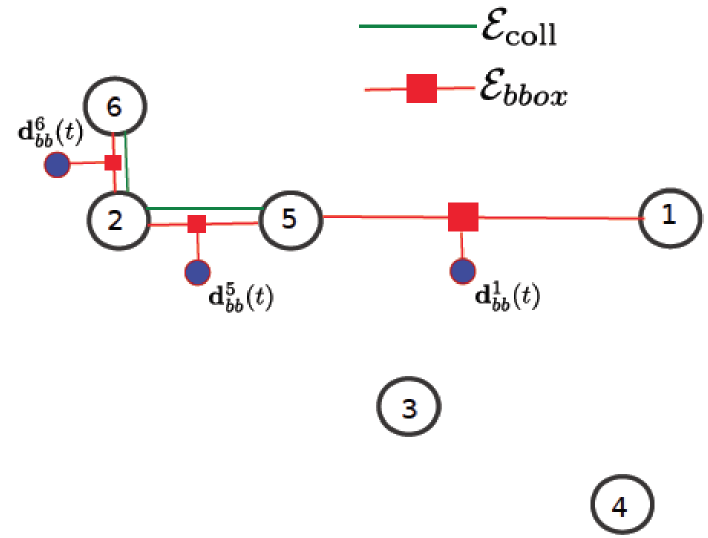


Visibility fraction for a hypothesized bounding box edge: $v^{ij} = \dfrac{\text{Visible area of triangle}}{\text{Area of triangle}}$

Bounding box energy with occlusion: $E_{bboxOcc} = \sum_{k \in \text{edges}} v_k^{ij} |\pi_{\Omega^j}(\mathbf{B}^j) - \mathbf{d}^j|_k$

TP-TP Relation: Bounding Box Visibility

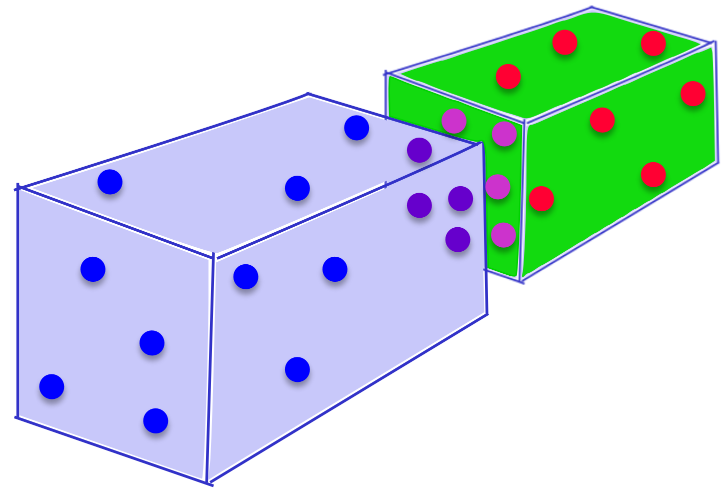




Visibility fraction for a hypothesized bounding box edge: $v^{ij} = \frac{\text{Visible area of triangle}}{\text{Area of triangle}}$

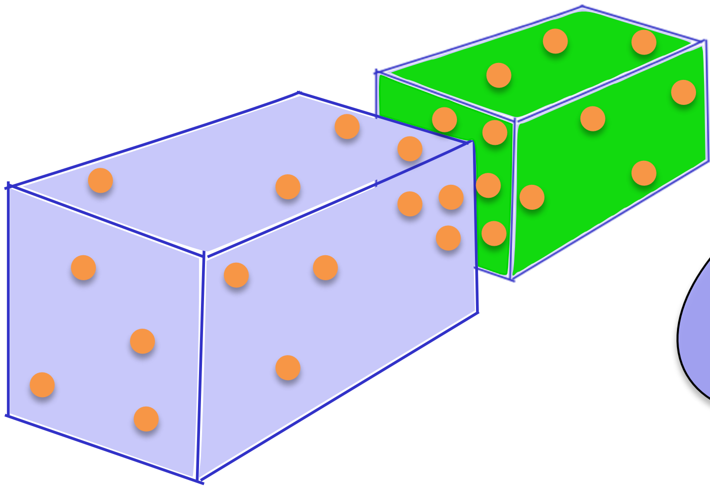
Bounding box energy with occlusion: $E_{bboxOcc} = \sum_{k \in \text{edges}} v_k^{ij} |\pi_{\Omega^j}(\mathbf{B}^j) - \mathbf{d}^j|_k$

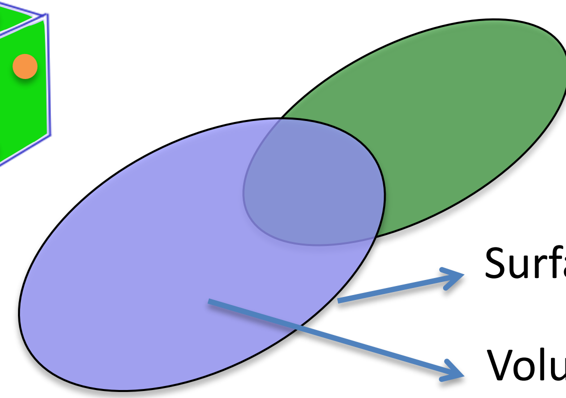
TP-TP: Probabilistic Occlusion



- Determine soft assignment of 2D point tracks to each 3D bounding box
- Probabilistic visibility for each point track.

TP-TP: Probabilistic Occlusion





Represent TPs as translucent 3D ellipsoids

Surface = Reflection (3D Point)

Volume = Transmission
= 1 - Occlusion

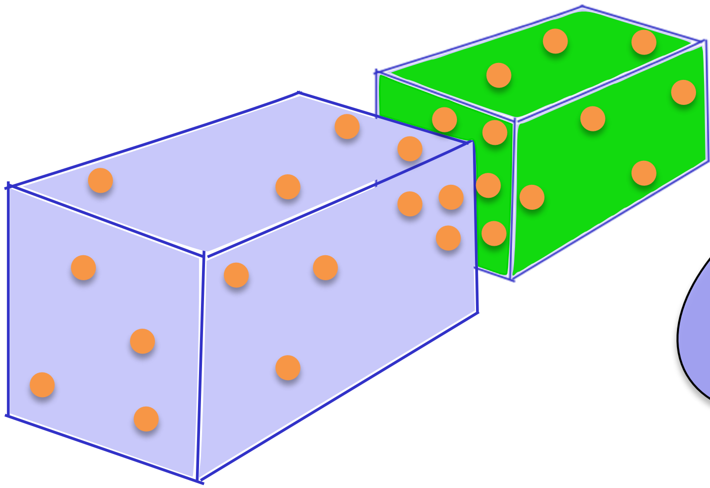
Projection of bounding box in image: $[u_l^i, v_t^i, u_r^i, v_b^i] = \pi_{\Omega^i(t)}(\mathbf{B}^i)$

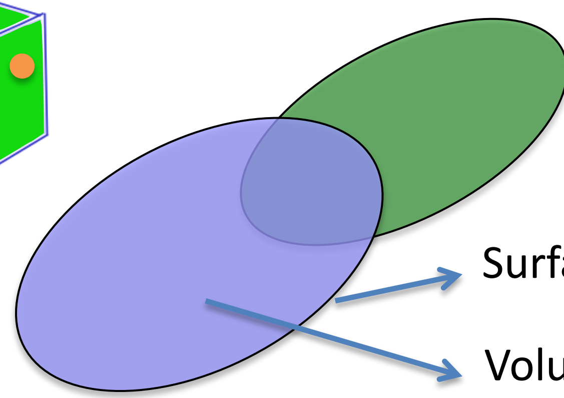
Mean and covariance of ellipsoid: $\mu_i = \frac{1}{2}\begin{bmatrix} u_l^i + u_r^i \\ v_t^i + v_b^i \end{bmatrix} \quad \Sigma_i = \begin{bmatrix} \frac{2}{(u_l^i - u_r^i)^2} & 0 \\ 0 & \frac{2}{(v_t^i - v_b^i)^2} \end{bmatrix}$

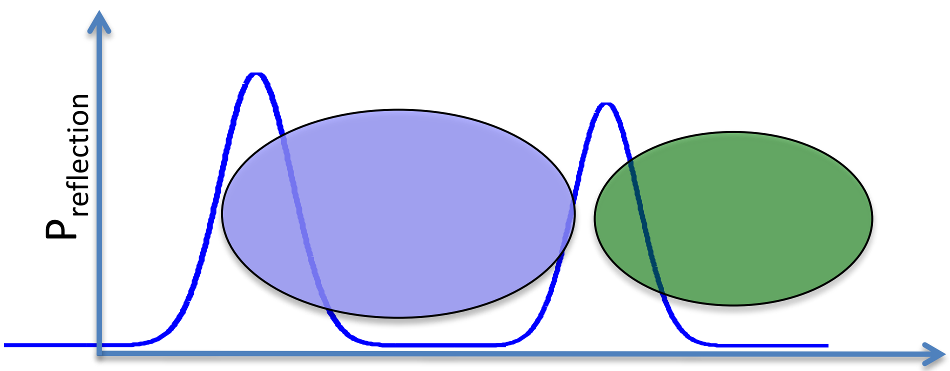
Model occlusion as a continuous soft probability:

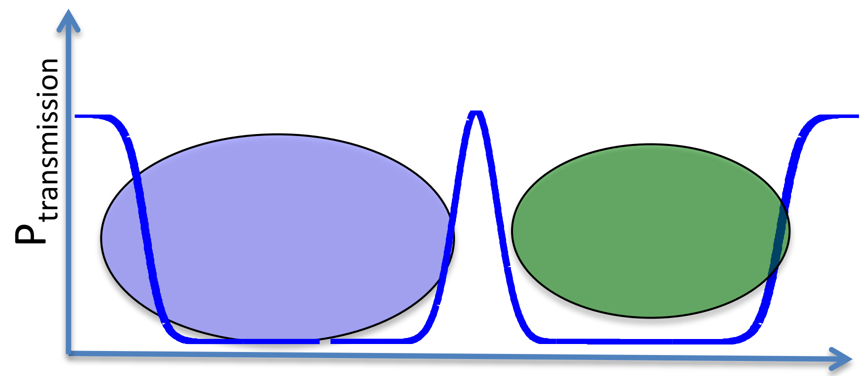
$$f_{\text{occ}}^i(u, v, \lambda) = \frac{N(u, v; \mu_i, \Sigma_i)}{1 + e^{-\frac{\lambda - \mu_i^{(d)}}{\beta}}} \text{where } \mu_d = \Omega^i(t)_z$$

TP-TP: Probabilistic Occlusion



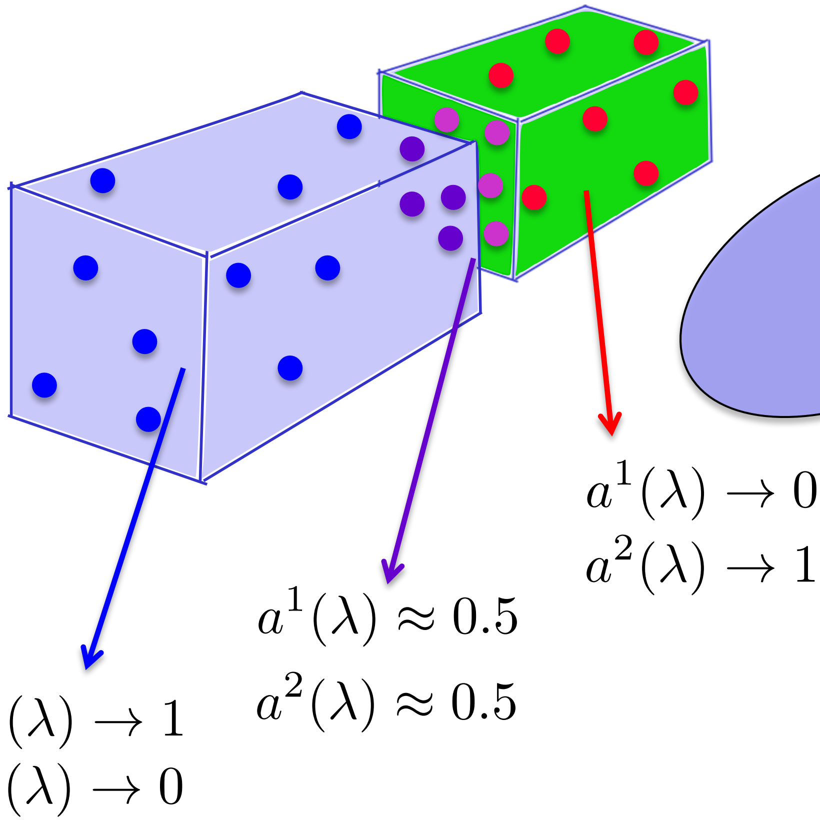


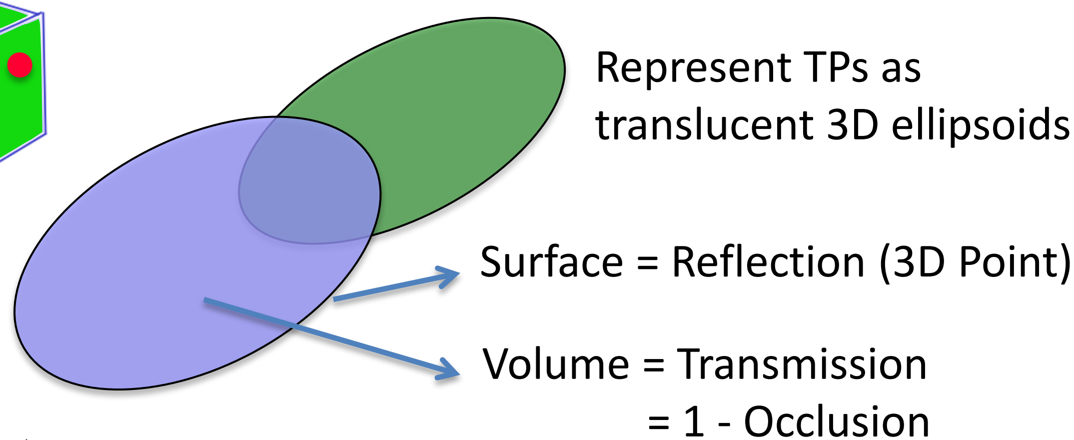


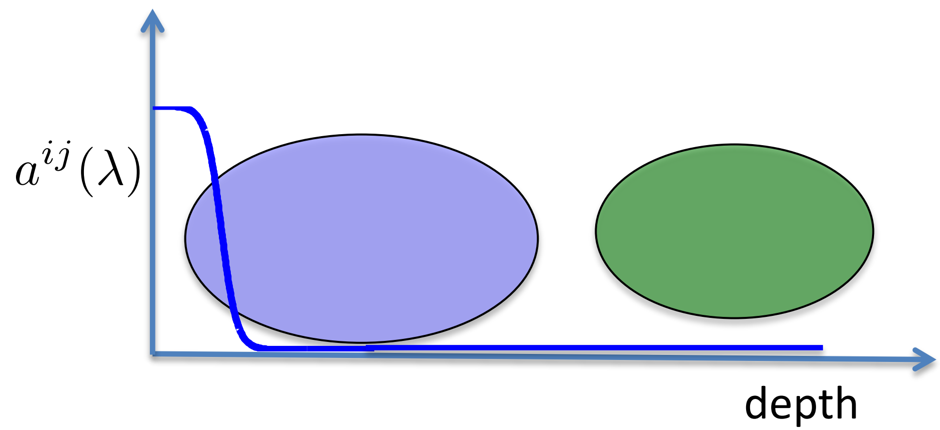


Association probability for point j with object i : $a^{ij}(\lambda) = P^{i}_{refl} \prod_{0}^{\lambda} P^{d\lambda}_{trans}$

TP-TP: Probabilistic Occlusion

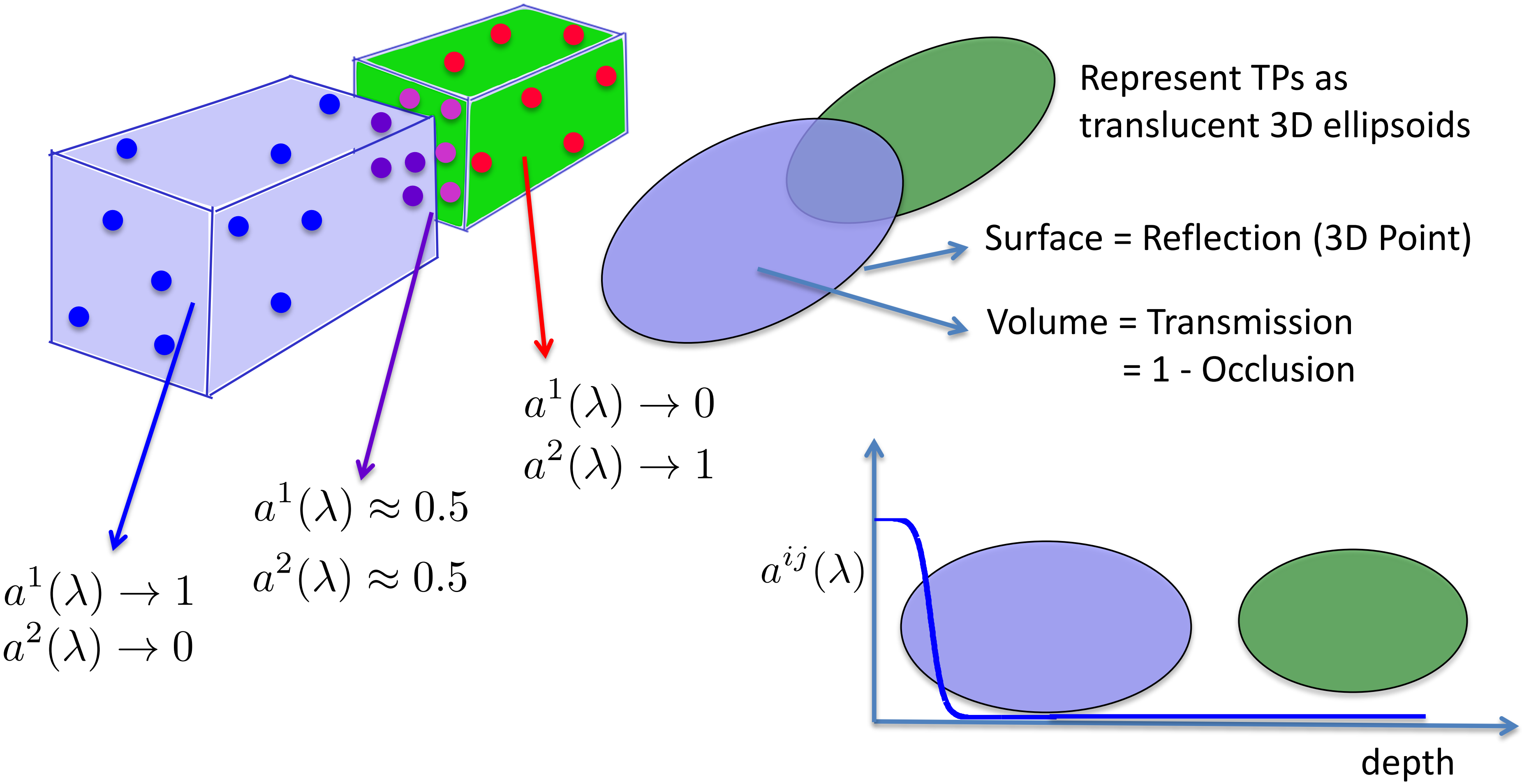






Association probability for point j with object i : $a^{ij}(\lambda) = P^i_{refl} \prod_0^{\lambda} P^{d\lambda}_{trans}$

TP-TP: Probabilistic Occlusion



Point track energy considering TP-TP occlusions:

$$E_{trackOcc} = \sum_{i \in \text{objects}} \sum_{j \in \text{tracks}} a^{ij} \| \mathbf{u}^j(t) - \pi_{\Omega^i(t)}(\pi^{-1}_{\Omega^i(t-1)}(\mathbf{u}_j(t-1))) \|^2$$

Probabilistic Occlusion Levels

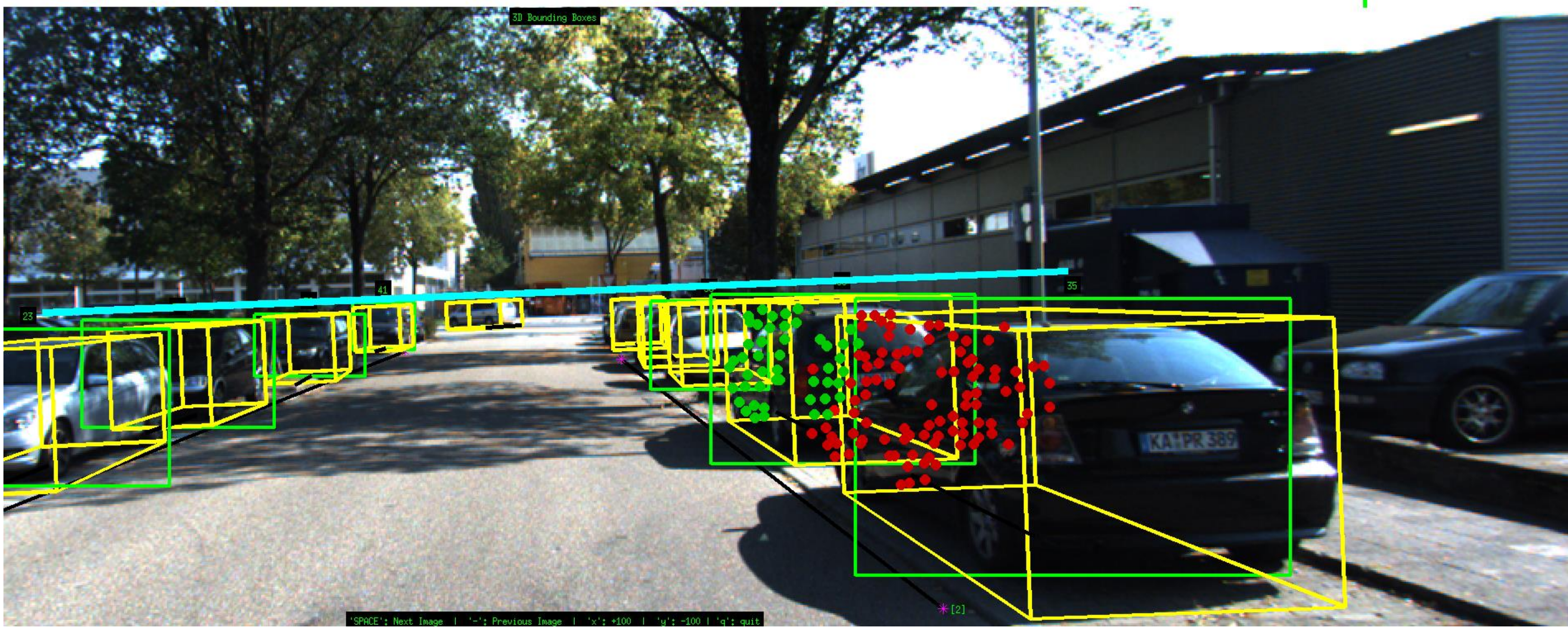
Not occluded

Partly occluded

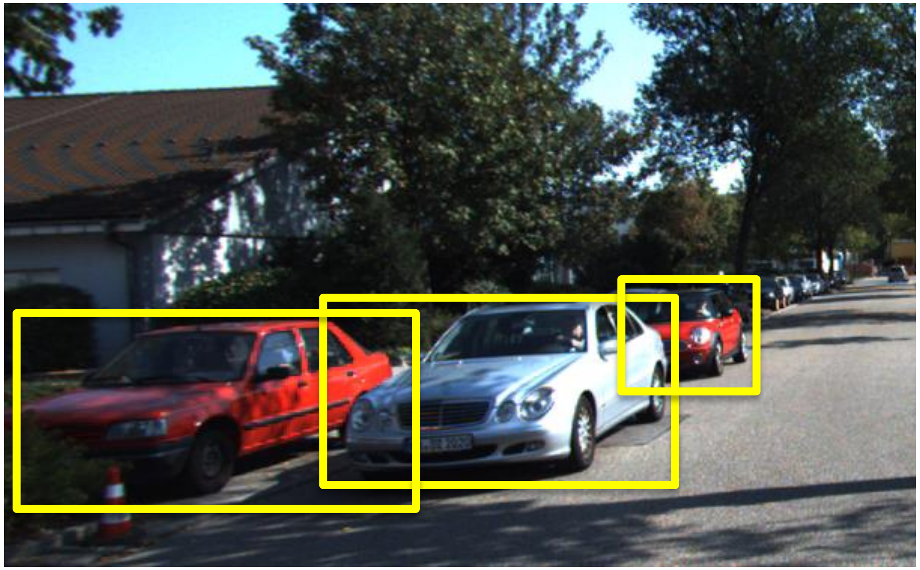
Mostly occluded

- Probabilistic specification of occlusion level for each object

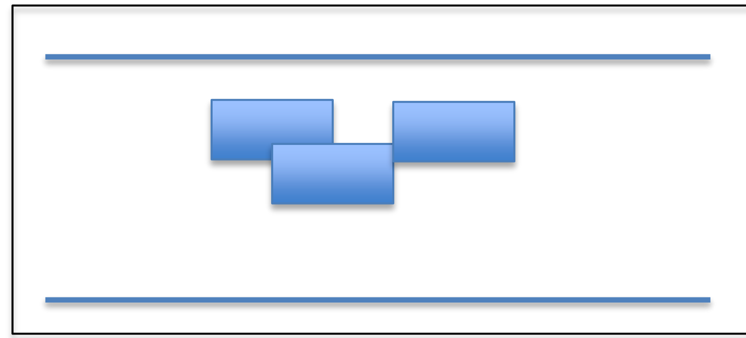
Effect of Occlusion Energy



TP-TP Relationships: Collision

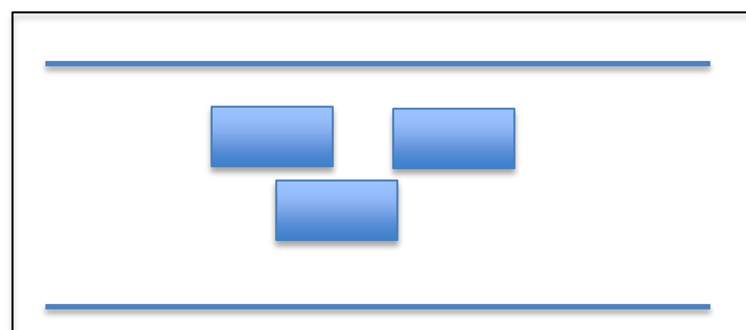


BEV Localization

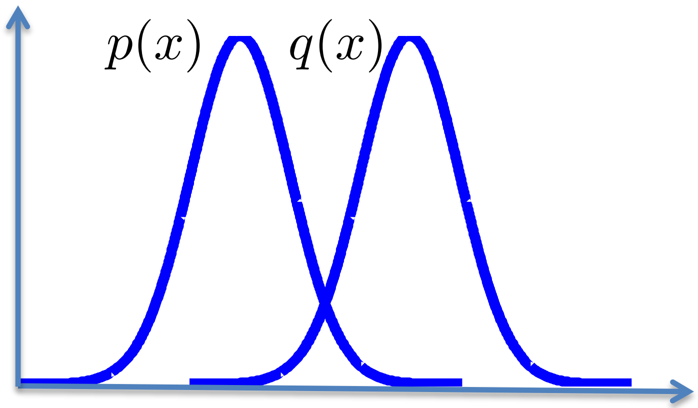


Objects cannot physically occupy the same 3D space.

Collision Resolution



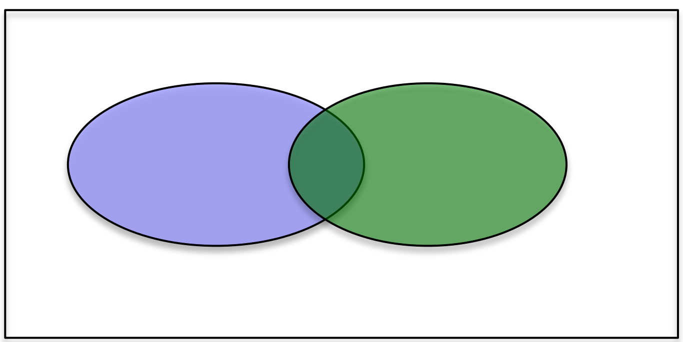
TP-TP Relationships: Collision

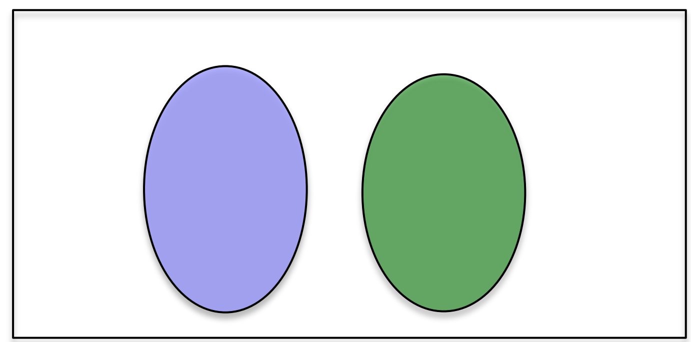


Bhattacharya coefficient for distance:

$$BC(p,q) = \int_0^\infty \sqrt{p(x)q(x)}\; dx$$

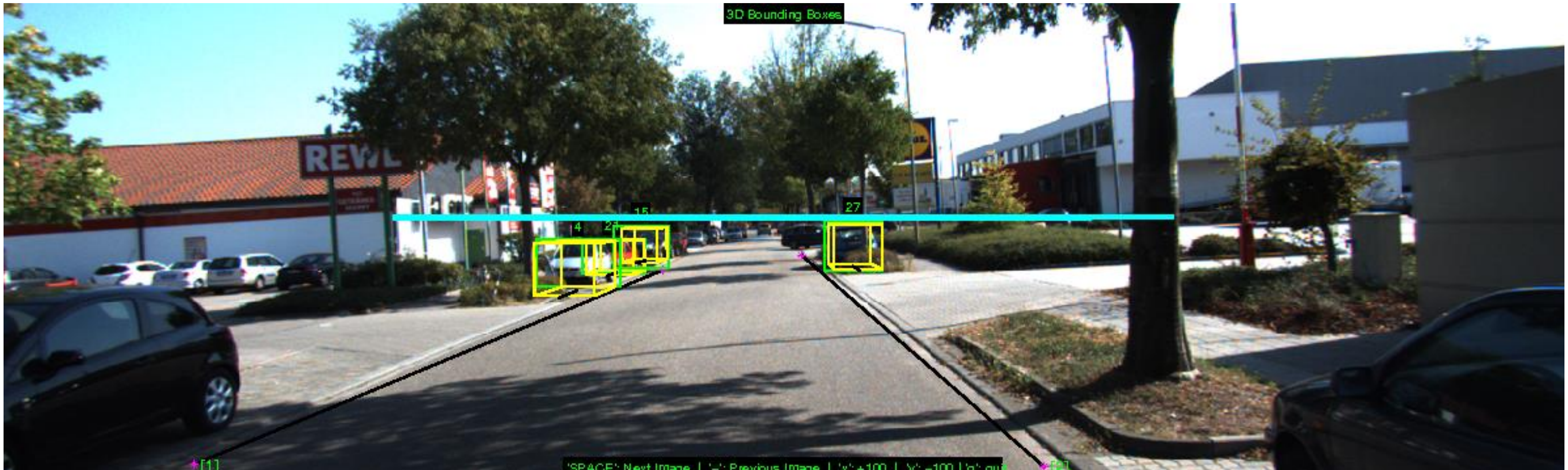
Has analytic form for Gaussian distributions.

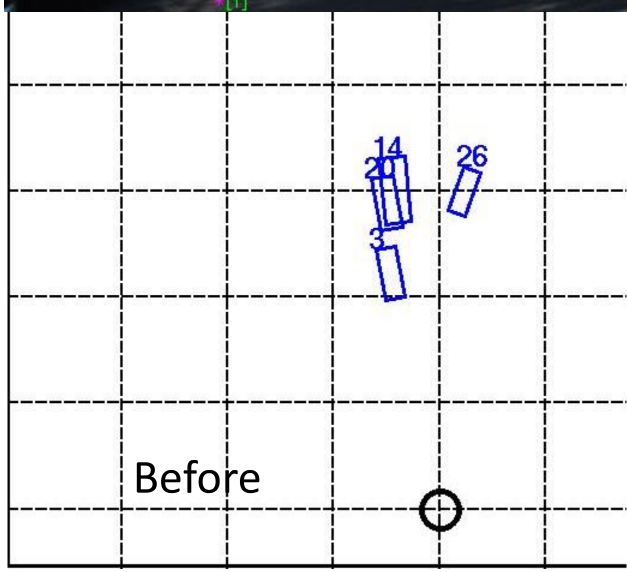




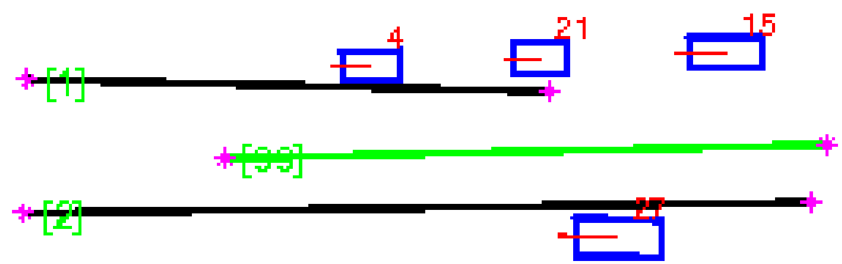
$$\mathcal{E}_{\mathrm{col}}^{ijt} = \frac{|\Sigma_i|^{\frac{1}{4}}|\Sigma_j|^{\frac{1}{4}}}{|\frac{1}{2}\Sigma_i + \frac{1}{2}\Sigma_j|^{\frac{1}{2}}} e^{-\frac{1}{8}\left(\mathbf{p}^{(i)}(t)-\mathbf{p}^{(i)}(t)\right)^\top \left(\frac{1}{2}\Sigma_i+\frac{1}{2}\Sigma_j\right)^{-1}\left(\mathbf{p}^{(i)}(t)-\mathbf{p}^{(i)}(t)\right)}$$

Effect of Collision Energy



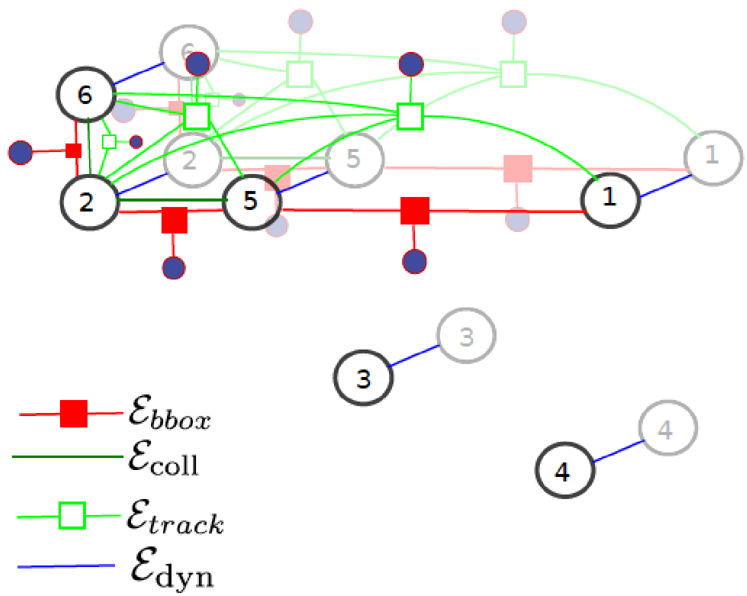


Before



After

Temporal Consistency



- Dynamic terms
  - holonomic, orientation and velocity constraints.

$$\mathcal{E}^{it}_{\text{dyn-hol}} = 1 - \omega^{(i)}(t-1) \cdot (\mathbf{p}^{(i)}(t) - \mathbf{p}^{(i)}(t-1))$$

Car moves only in forward direction

$$\mathcal{E}^{it}_{\text{dyn-ori}} = \|\omega^{(i)}(t) - \omega^{(i)}(t-1)\|^2$$

Smoothness for orientation

$$\mathcal{E}^{it}_{\text{dyn-vel}} = \|(\mathbf{p}^{(i)}(t) - 2\mathbf{p}^{(i)}(t-1)) + \mathbf{p}^{(i)}(t-2)\|^2$$

Constant velocity

Inference

- Just use unconstrained minimization for now
- Alternatingly minimize for a few iterations:
  - Lane + Dynamic energies
  - Bounding box + Size energies
  - Occlusion + Collision energies

- Future work:
  - Message passing to exploit graph structure.

Results

- Dataset : sequences from KITTI
- Metrics :
  - Translation error
  - Orientation error (yaw angle along ground plane)
  - Size error (averaged over length, width and height)
  - Position error in Z (depth)
  - Position error in X (lateral)

Results

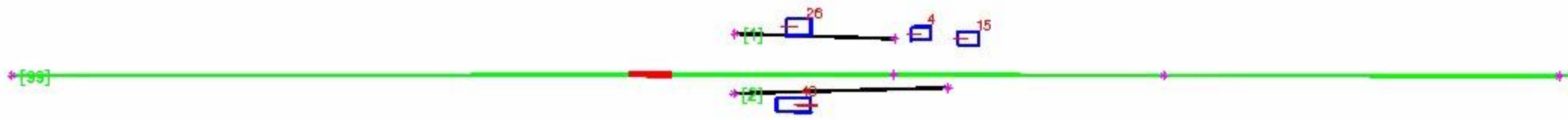
	Translation (%)	Yaw (Degrees)	Size (%)	Z (%)	X (%)
Independent Localization	8.210	19.824	**1.209**	7.922	**1.469**
Bounding Box + Lane	7.761	4.787	1.283	7.358	1.689
Bounding Box + Lane + Dynamic	7.704	**4.635**	1.264	7.294	1.660
Occlusion + Lane + Dynamic	**7.697**	4.764	1.264	**7.285**	1.661
Occlusion + Lane + Dynamic + Collision	7.802	4.655	1.259	7.362	1.727

- Errors decrease using scene and TP constraints
  - Scene elements constrain object orientation (yaw)
  - Also better translation and depth errors

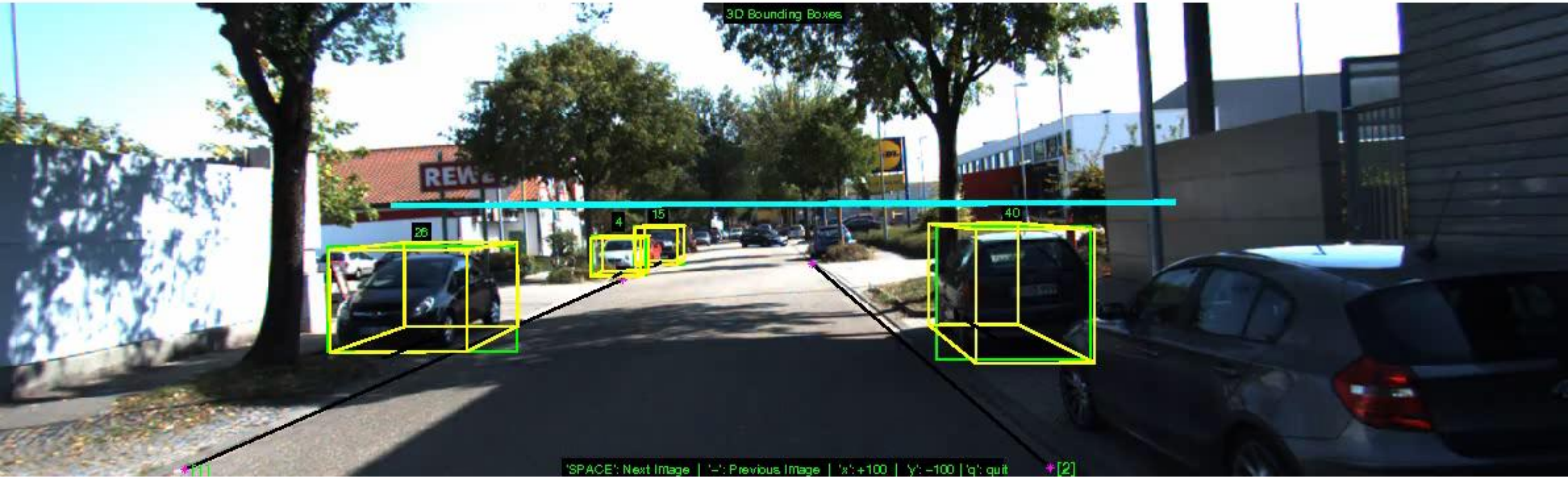
Videos

frame 7/11

	bboxO	size	lane	posT	yawT	col
40	16.41	0.06	0.43	0.00	0.00	0.38
26	3.38	0.66	0.43	0.00	0.00	0.45
15	0.67	0.11	0.18	0.00	0.00	1.02
4	2.41	0.18	0.88	0.00	0.00	1.11



X=-82.80000:82.80000 Y=-25.00000:25.00000



Conclusions

- TP-Scene interactions lead to better localization
  - Significant improvement in orientation accuracy
- Modeling TP-TP interactions lends consistency
  - Probabilistically reason about occlusions
  - 3D object localization incorporating visibility
  - Soft point track associations to handle occlusions
  - Resolve collisions
- Better accuracy than independent localization
  - For “important” metrics (depth and orientation)
- Probabilistic notion of TP-Scene and TP-TP interactions
  - Forms input to scene recognition applications.

Future Work

- Learning the weights
- Better optimization
- More extensive evaluation.