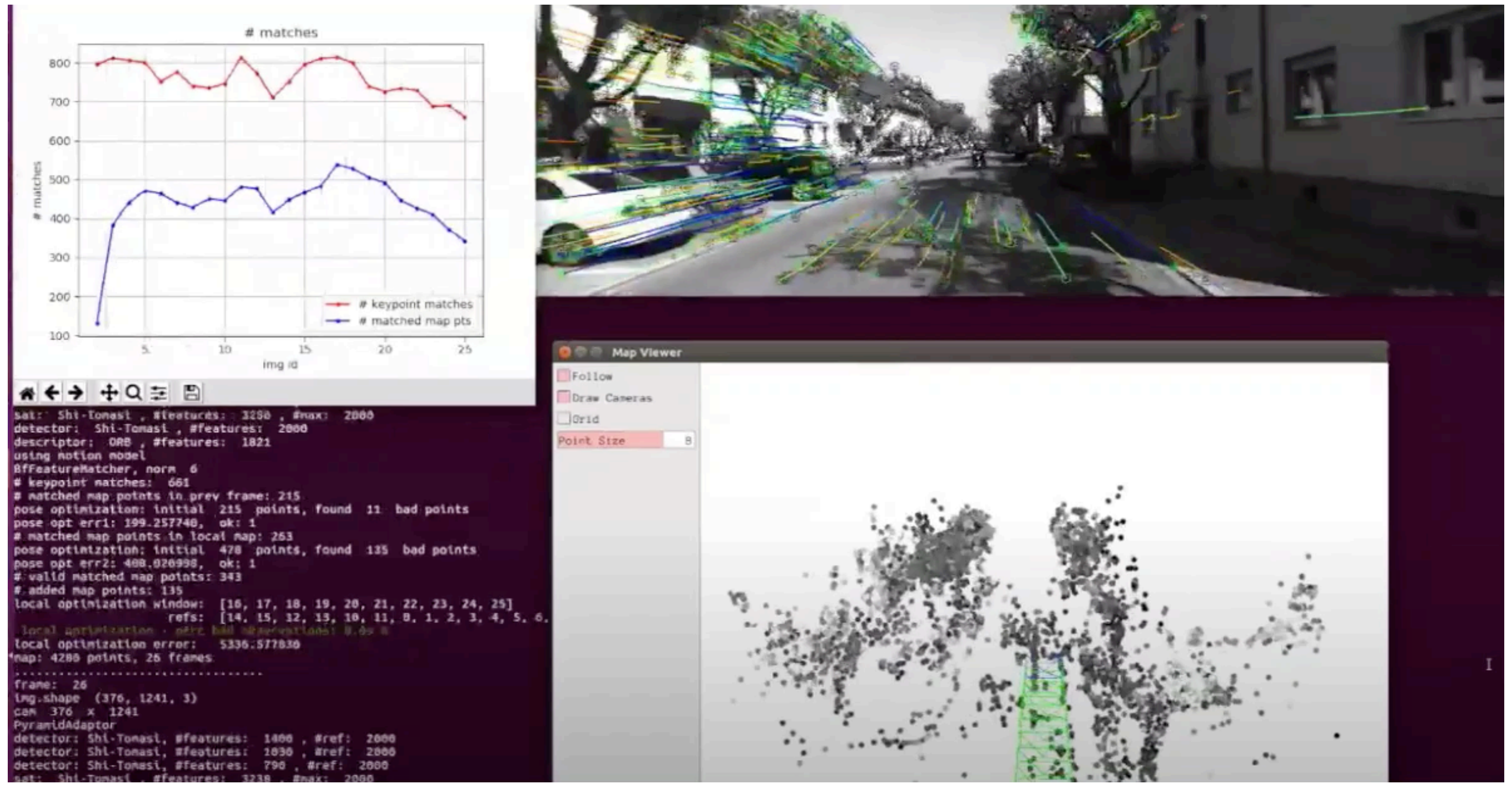
OrcVIO: Object residual constrained Visual-Inertial Odometry

Mo Shan Qiaojun Feng Nikolay Atanasov

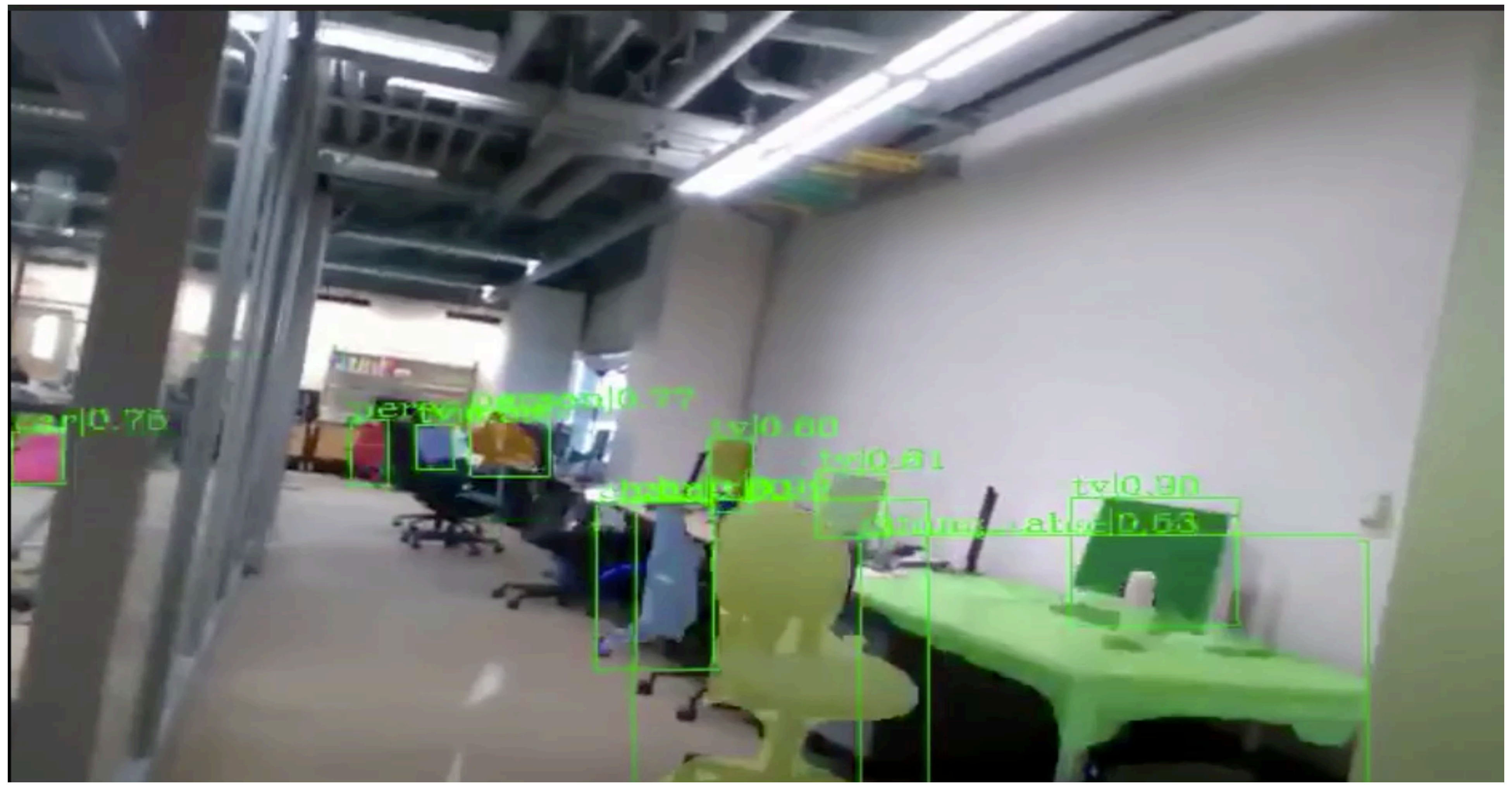
Existential Robotics Laboratory
Department of Electrical and Computer Engineering
University of California, San Diego

Motivation

- Most SLAM/VIO methods produce geometric environment representations



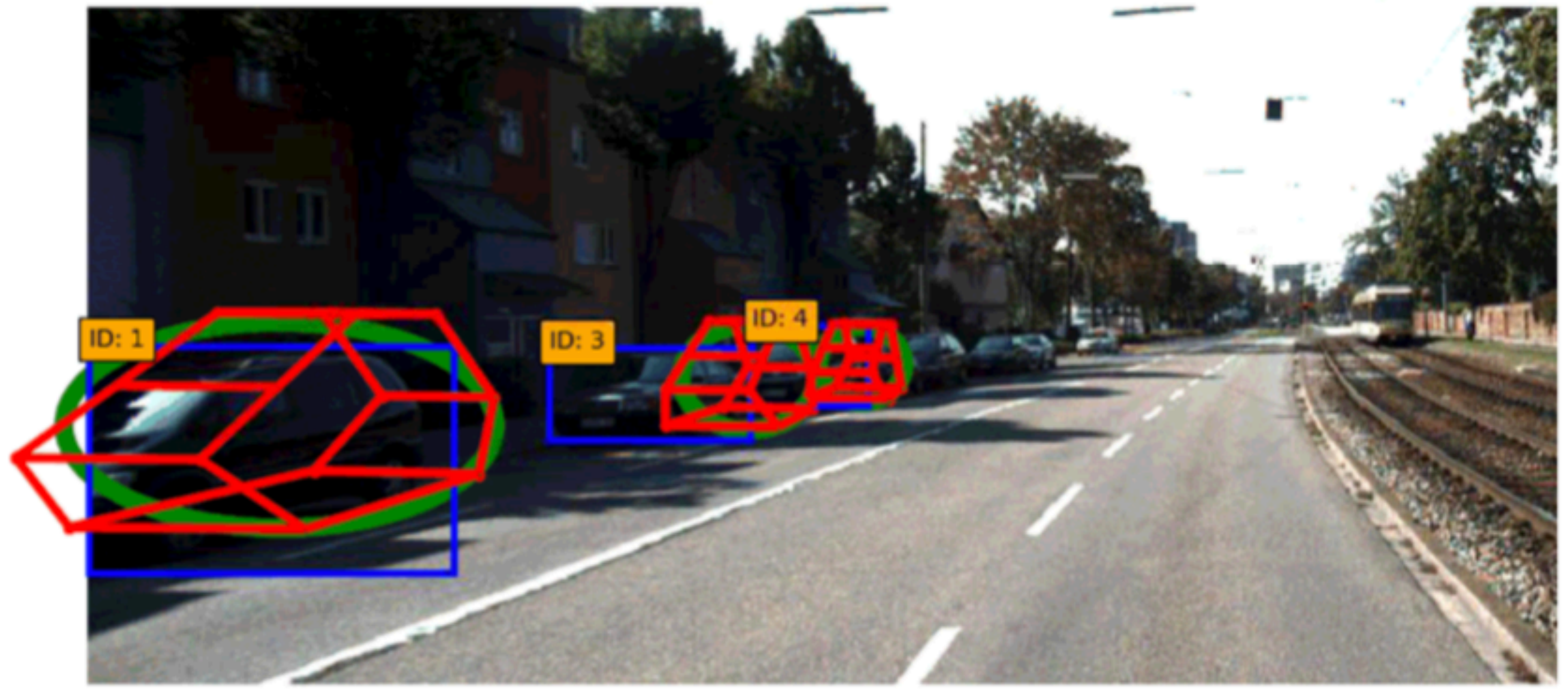
- Object recognition using deep neural networks have impressive results

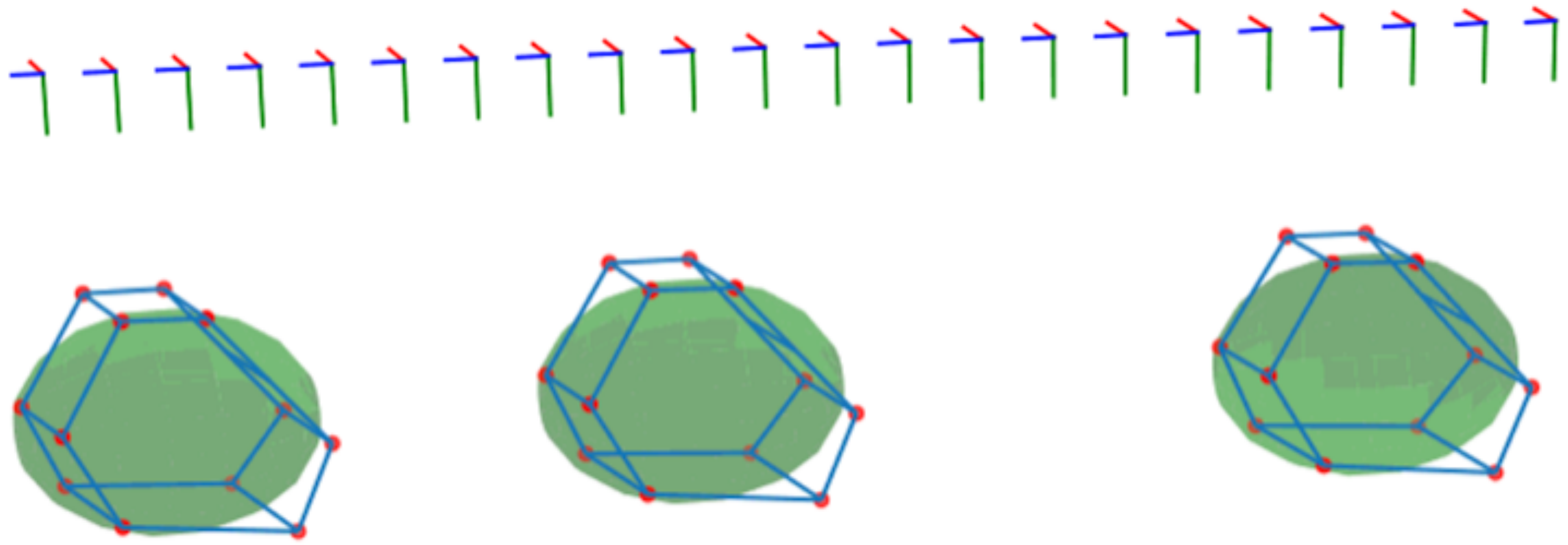


Motivation

- This work harnesses the strength of both VIO and deep neural networks
- We propose Object residual constrained Visual-Inertial Odometry (OrcVIO)
- OrcVIO outputs geometrically consistent, semantically meaningful maps

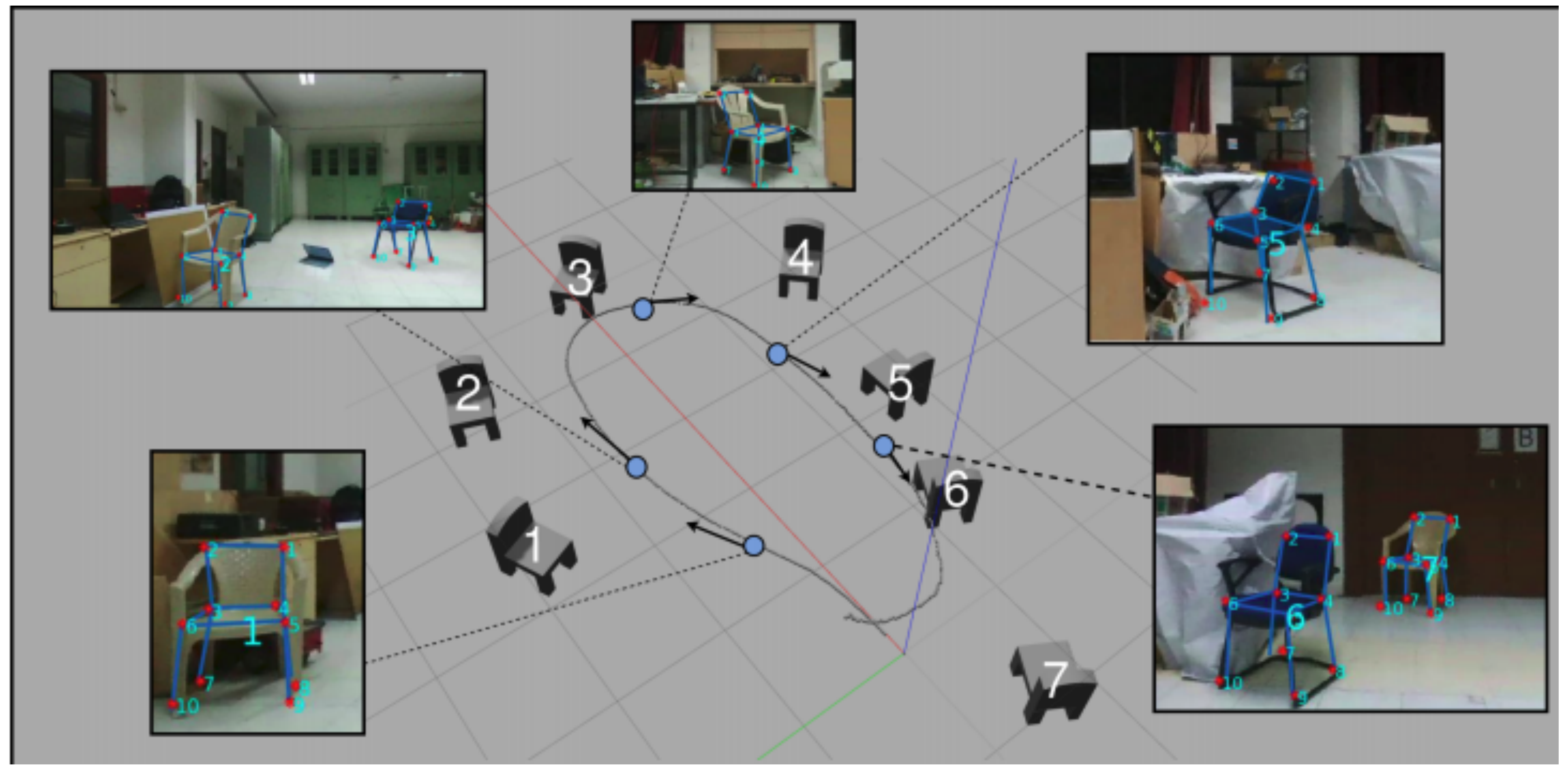
OrcVIO

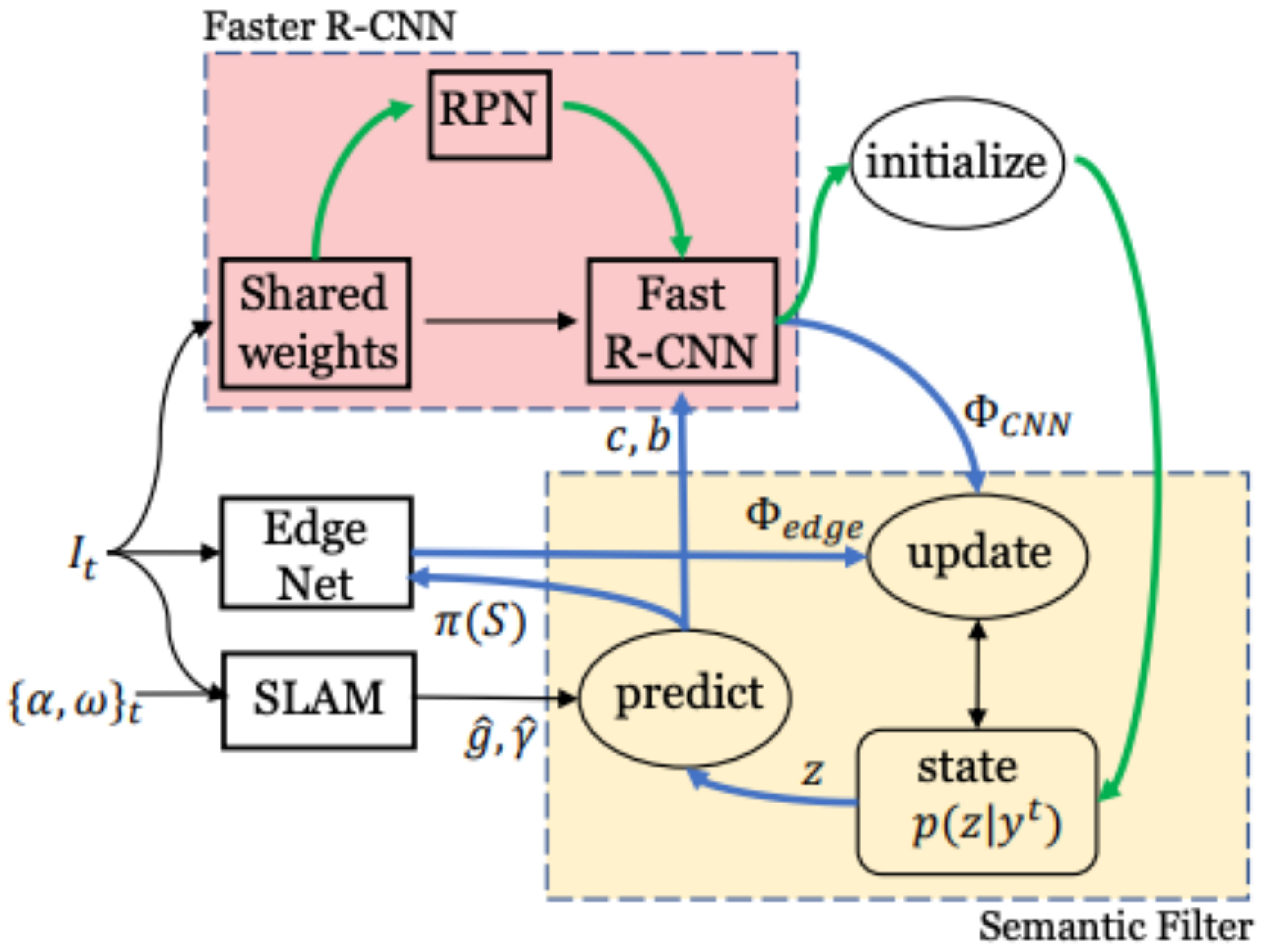




Related Work

- Category-specific approaches optimize the pose and shape of object instances using 3D shape models/semantic keypoints





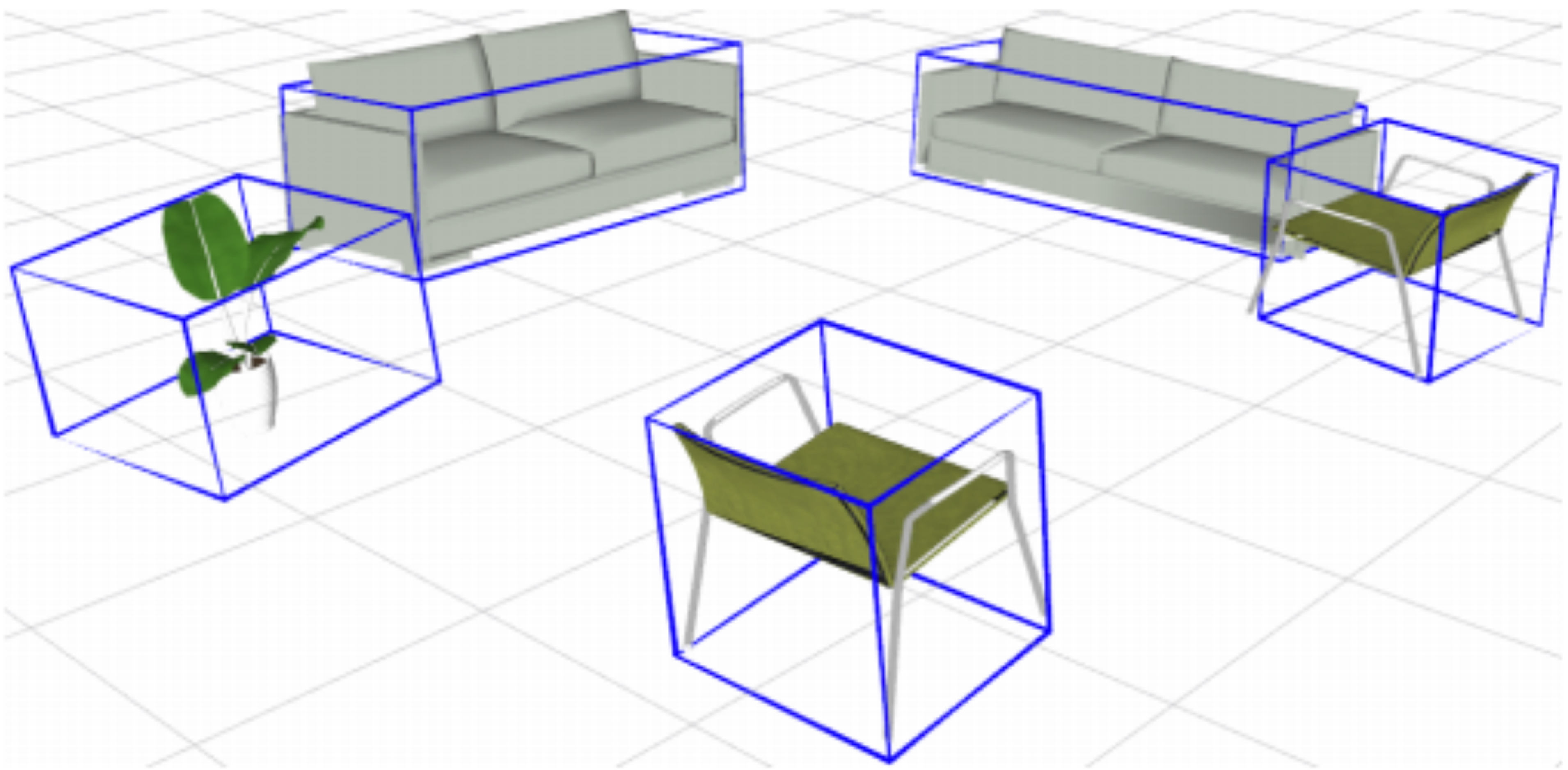
Parkhiya et al., 2018, ICRA

Fei, X., & Soatto, S., 2018, ECCV

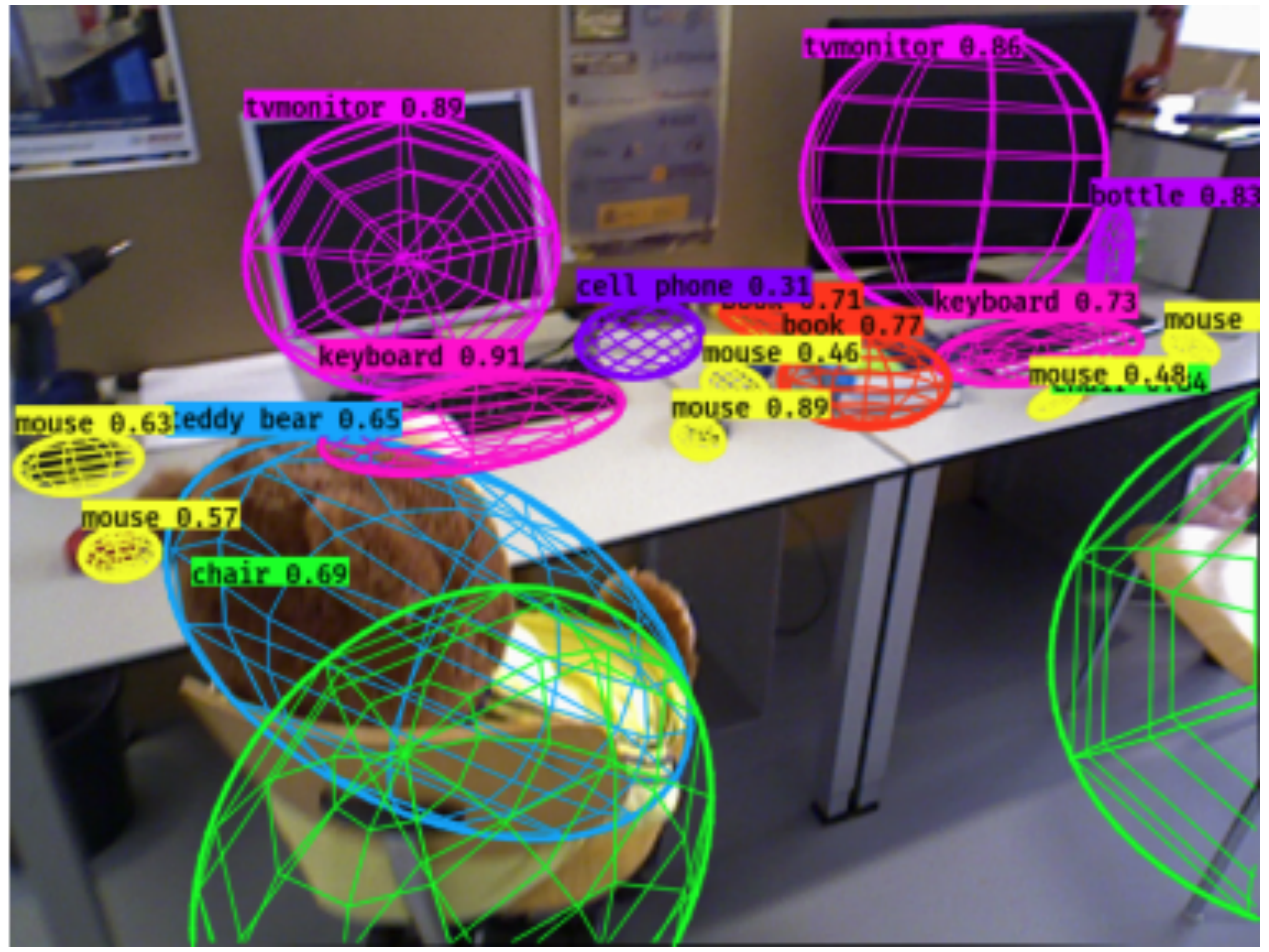
- Parkhiya, P., Khawad, R., Murthy, J.K., Bhowmick, B. and Krishna, K.M., 2018, May. Constructing category-specific models for monocular object-SLAM. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*
- Fei, X. and Soatto, S., 2018. Visual-inertial object detection and mapping. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 301-317).

Related Work

- Category-agnostic approaches use geometric shapes such as ellipsoids or cuboids to represent objects



CubeSLAM, Yang, S. and Scherer, S., 2019, TRO

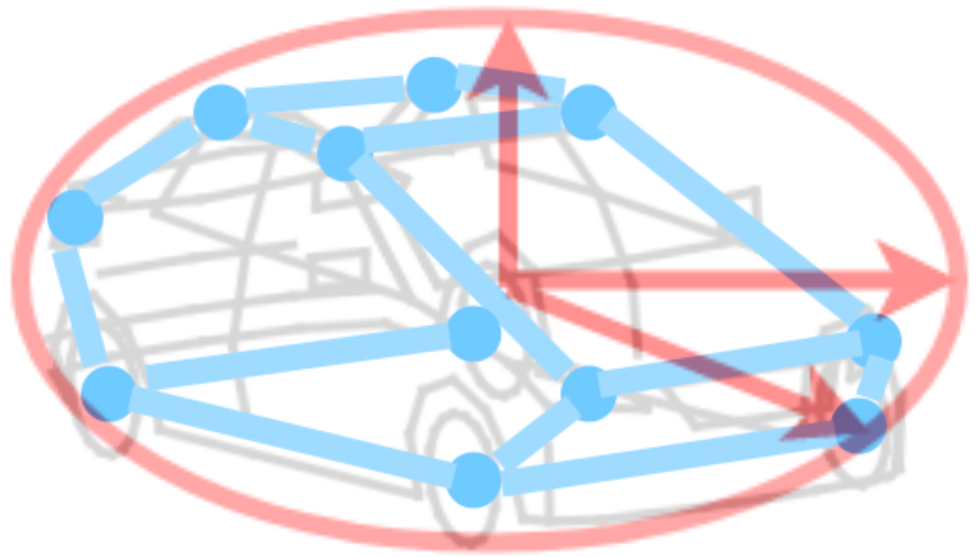


QuadricSLAM, Nicholson et al., 2018, RAL

- Yang, S. and Scherer, S., 2019. Cubeslam: Monocular 3-d object slam. IEEE Transactions on Robotics, 35(4), pp.925-938.
- Nicholson, L., Milford, M. and Sünderhauf, N., 2018. Quadricslam: Dual quadrics from object detections as landmarks in object-oriented slam. IEEE Robotics and Automation Letters, 4(1), pp.1-8.

Object Class

- Coarse level: ellipsoid (red)
- Fine level: keypoints (blue)

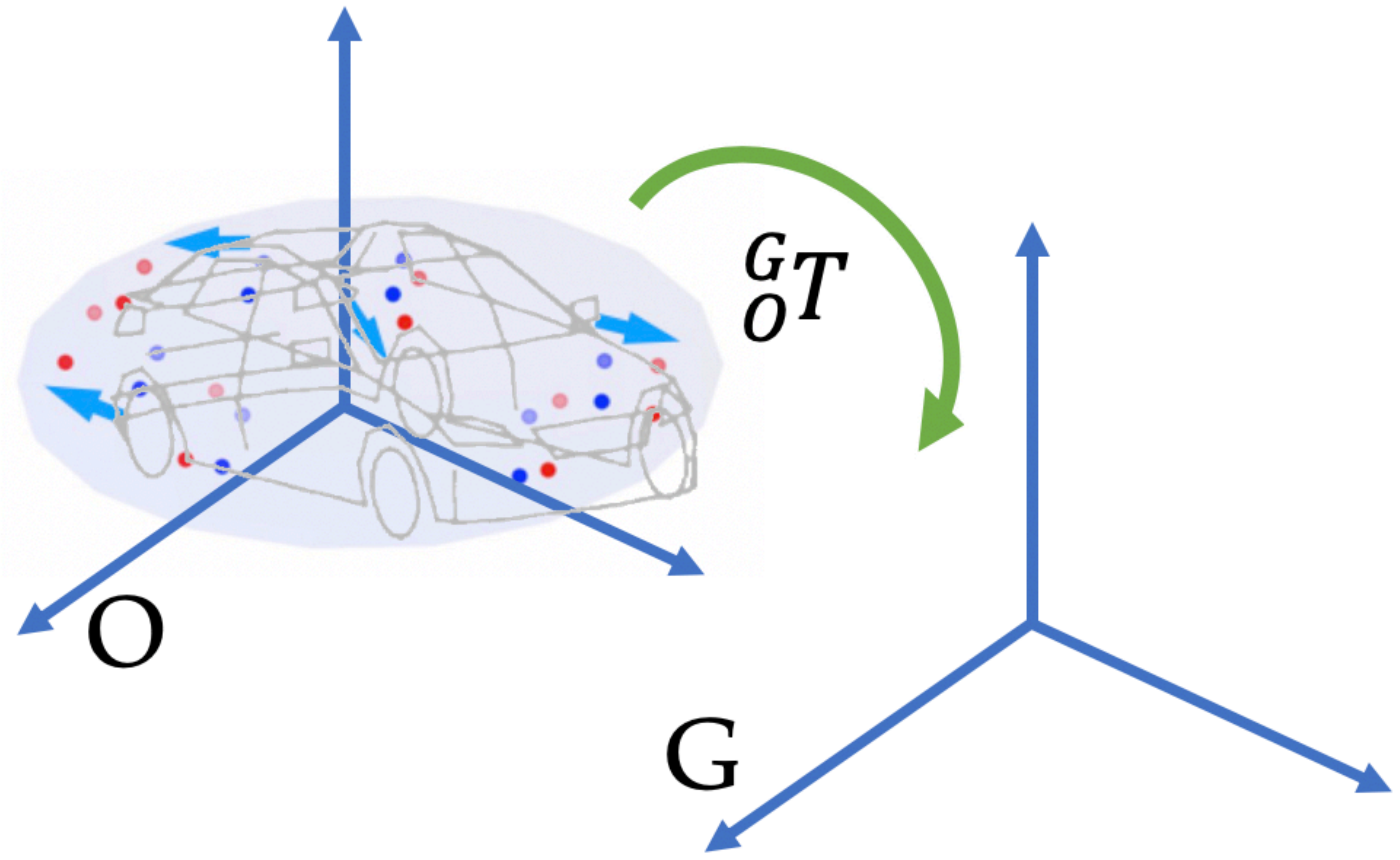


"Treat nature by means of the cylinder, the sphere, the cone, everything brought into proper perspective"

Paul Cezanne

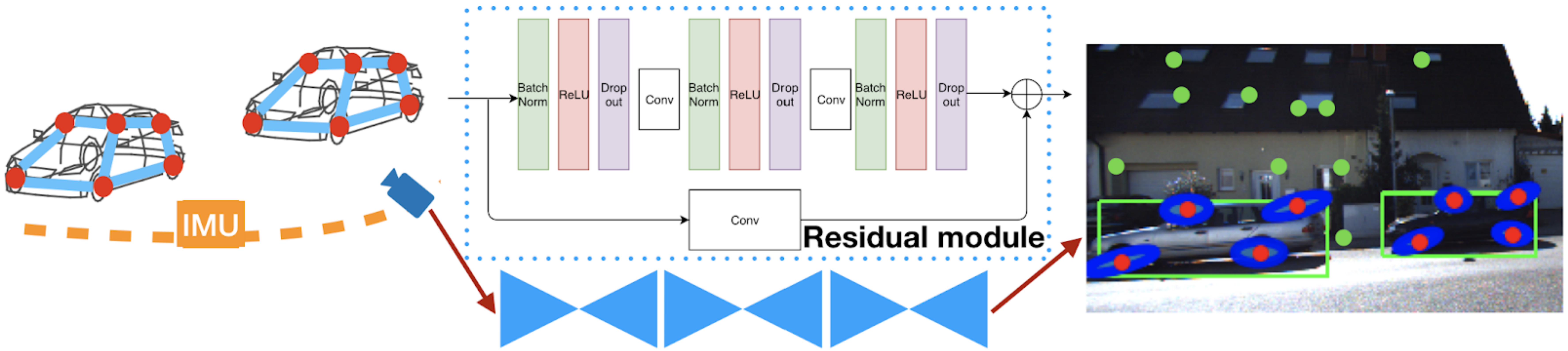
Object Instance

- Deformation (blue arrows)
- Pose (green arrow)



Problem Formulation

- Determine the sensor trajectory, geometric landmarks, and object states using inertial, geometric, semantic, and bounding-box measurements



(a) 3D scene (b) Stacked hourglass network (c) Visual observation

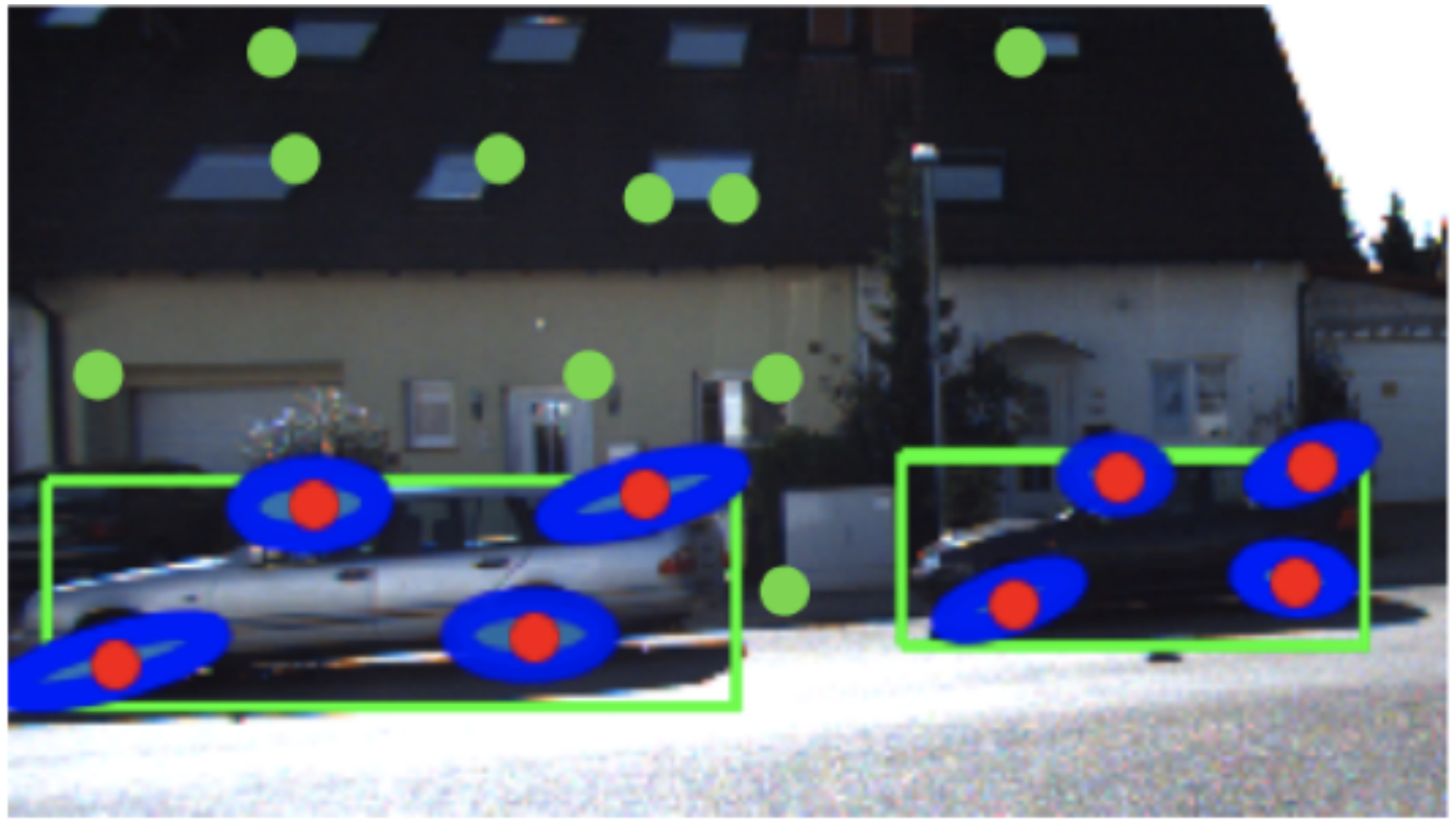
min TrajectoryCost + GeometricReprojectionCost + SemanticReprojectionCost + BoundingBoxCost + ShapeRegularization

Objective Function

Problem. Determine the sensor trajectory $\mathcal{X}^*$, geometric landmarks $\mathcal{L}^*$, and object states $\mathcal{O}^*$ that minimize the weighted sum of squared errors:

$$\min_{\mathcal{X},\mathcal{L},\mathcal{O}} {}^{i}w \sum_{t} \|{}^{i}\mathbf{e}_{t,t+1}\|^2_{{}^{i}\mathbf{V}} + {}^{g}w \sum_{t,m,n} \mathbb{1}_{t,m,n} \|{}^{g}\mathbf{e}_{t,m,n}\|^2_{{}^{g}\mathbf{V}}$$
$$+ {}^{s}w \sum_{t,i,j,k} \mathbb{1}_{t,i,k} \|{}^{s}\mathbf{e}_{t,i,j,k}\|^2_{{}^{s}\mathbf{V}} + {}^{b}w \sum_{t,i,j,k} \mathbb{1}_{t,i,k} \|{}^{b}\mathbf{e}_{t,i,j,k}\|^2_{{}^{b}\mathbf{V}}$$
$$+ {}^{r}w \sum_{i} \|{}^{r}\mathbf{e}\left(\mathbf{o}_i\right)\|^2$$

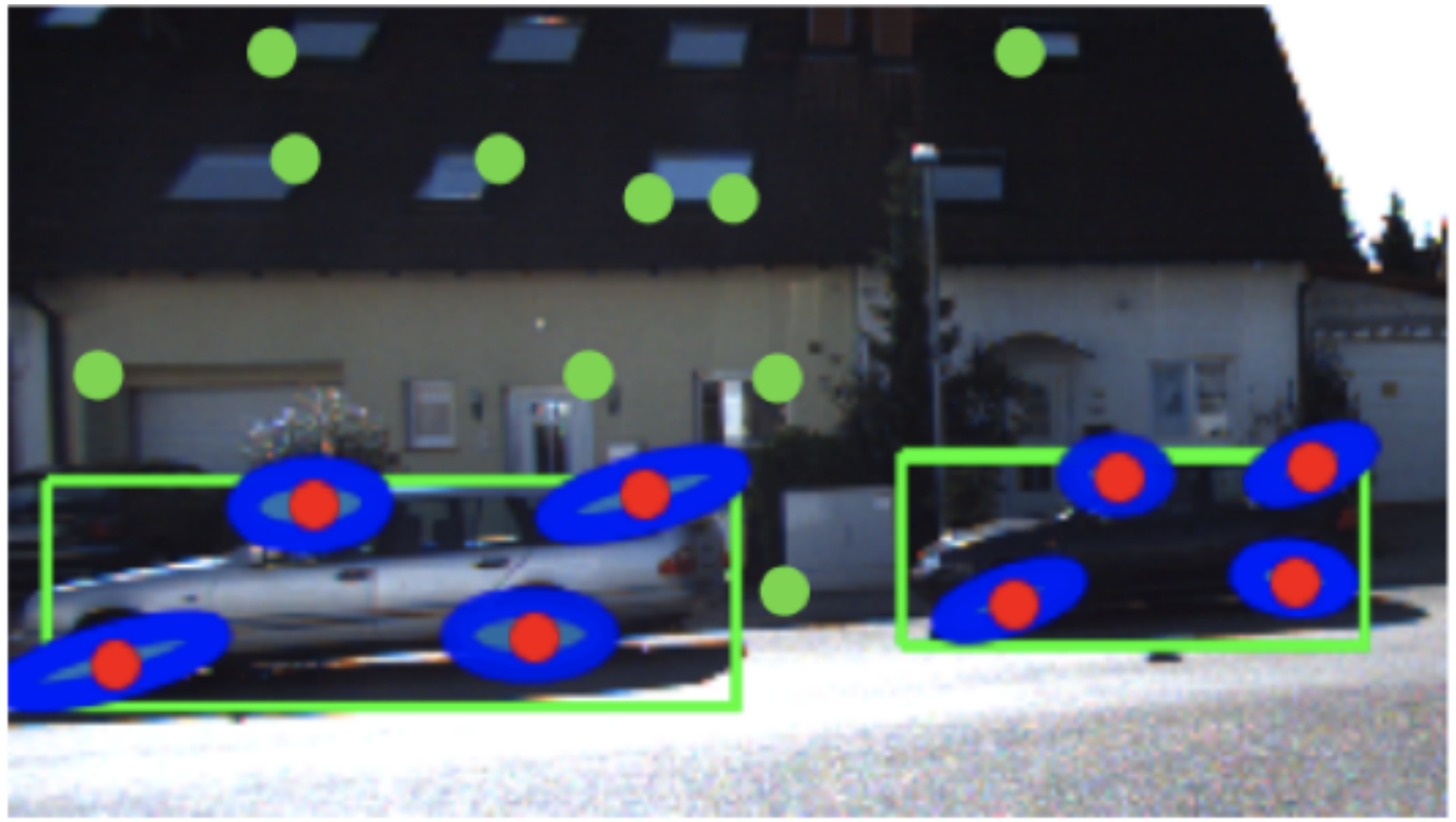
Geometric Keypoints



Define the geometric keypoint error as the difference between the image projection of a geometric landmark $\boldsymbol{\ell}$ using camera pose ${}_{C}\mathbf{T}$ and its associated keypoint observation ${}^{g}\mathbf{z}$:

$$
{}^{g}\mathbf{e}\left(\mathbf{x}, \boldsymbol{\ell}, {}^{g}\mathbf{z}\right) \triangleq \mathbf{P}\pi\left({}_{C}\mathbf{T}^{-1}\underline{\boldsymbol{\ell}}\right) - {}^{g}\mathbf{z},
$$

Semantic Keypoints

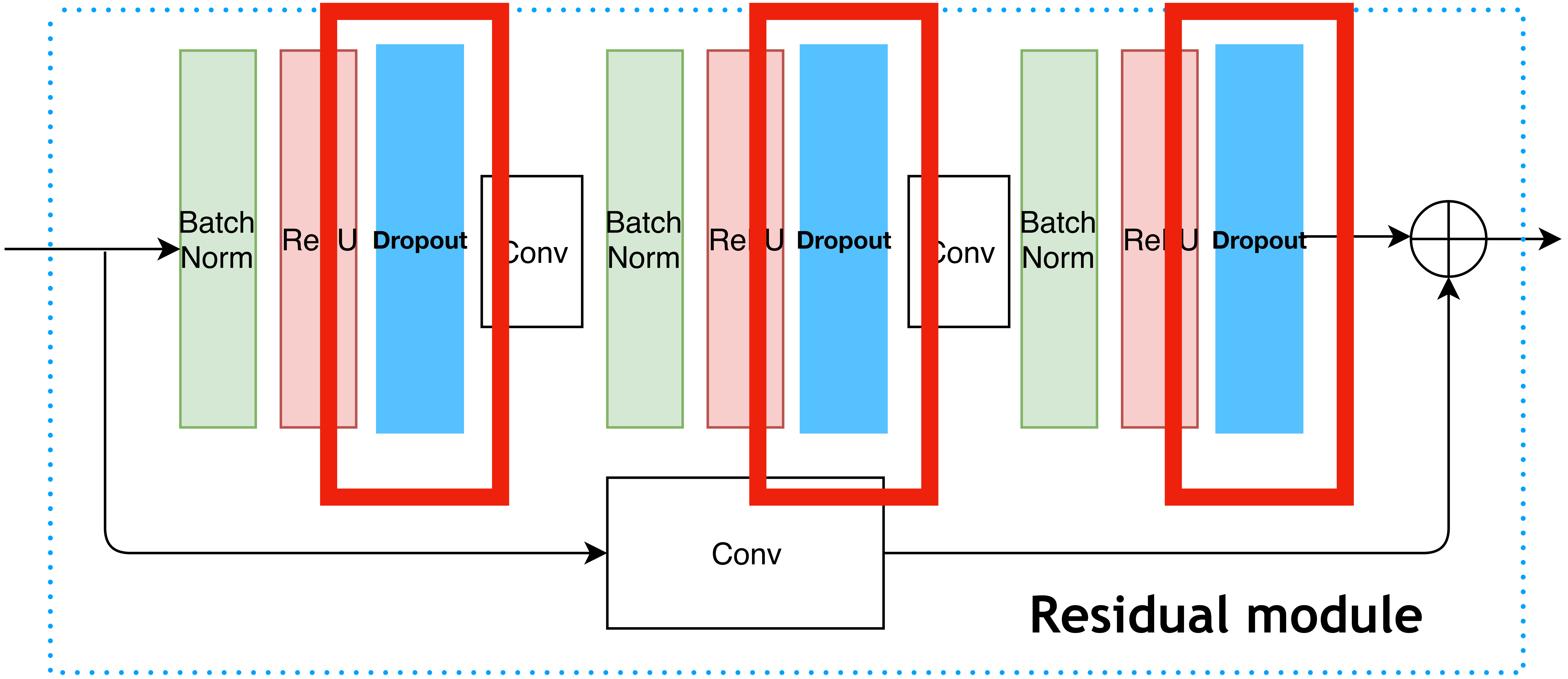


The semantic-keypoint error is defined as the difference between a semantic landmark $\mathbf{s}_j + \delta\mathbf{s}_j$, projected to the image plane using instance pose ${}_O\mathbf{T}$ and camera pose ${}_C\mathbf{T}_t$, and its corresponding semantic keypoint observation ${}^s\mathbf{z}_{t,j,k}$:

$$^s\mathbf{e}(\mathbf{x}_t, \mathbf{o}, {}^s\mathbf{z}_{t,j,k}) \triangleq \mathbf{P}\pi\left({}_C\mathbf{T}_t^{-1}\,{}_O\mathbf{T}\left(\underline{\mathbf{s}}_j + \delta\underline{\mathbf{s}}_j\right)\right) - {}^s\mathbf{z}_{t,j,k}.$$

Semantic Keypoints

- StarMap is used to detect semantic keypoints
- We add drop out layers in original network to obtain covariance



- Zhou, X., Karpur, A., Luo, L. and Huang, Q., 2018. Starmap for category-agnostic keypoint and viewpoint estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 318-334).

Semantic Keypoints

- We use Kalman Filter to track the semantic keypoints on an object level

Object Initialization

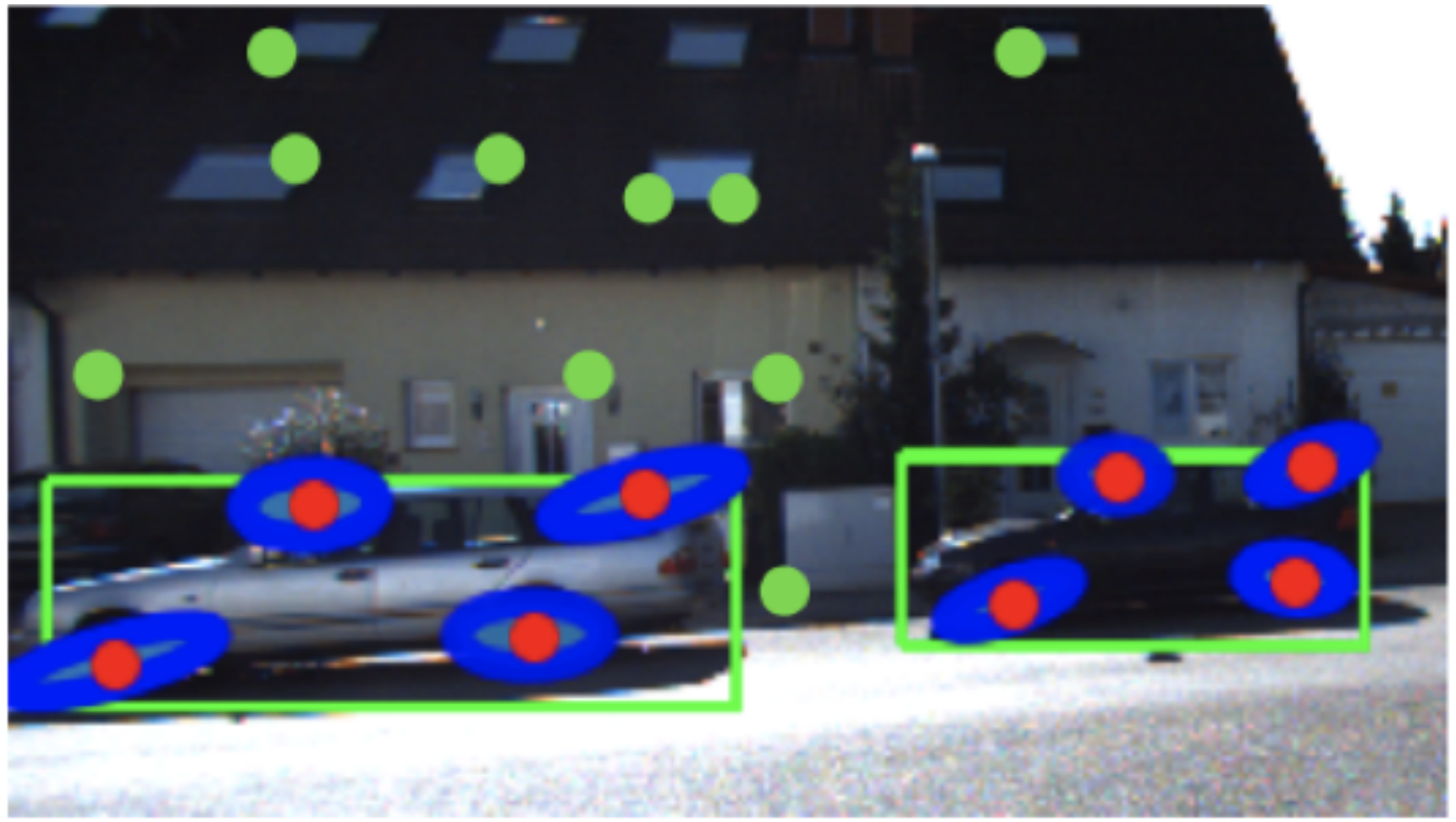
$$\mathbf{0} = \mathbf{P}_C\hat{\mathbf{T}}_t^{-1}{}_O\hat{\mathbf{T}}\underline{\mathbf{s}}_j - \lambda_{t,j,k}{}^s\mathbf{z}_{t,j,k}$$

Rearranging that leads to

$$_C\hat{\mathbf{R}}_t^\top\left(\boldsymbol{\xi}_j - {}_C\hat{\mathbf{p}}_t\right) = \lambda_{t,j,k}{}^s\mathbf{z}_{t,j,k}$$
$$_C\hat{\mathbf{R}}_t^\top\boldsymbol{\xi}_j - {}^s\mathbf{z}_{t,j,k}\lambda_{t,j,k} = {}_C\hat{\mathbf{R}}_t^\top{}_C\hat{\mathbf{p}}_t$$
$$\boldsymbol{\xi}_j - {}_C\hat{\mathbf{R}}_t{}^s\mathbf{z}_{t,j,k}\lambda_{t,j,k} = {}_C\hat{\mathbf{p}}_t$$

Tracked Targets

Bounding-box Measurements



To define a bounding-box error, we observe that if the dual ellipsoid $\mathbf{Q}^*_{(\mathbf{u}+\delta\mathbf{u})}$ of instance $\mathbf{i}$ is estimated accurately, then the lines ${}^b\underline{\mathbf{z}}_{t,j,k}$ of the k-th bounding-box at time t should be tangent to the image plane conic projection of $\mathbf{Q}^*_{(\mathbf{u}+\delta\mathbf{u})}$:

$$^b\mathbf{e}(\mathbf{x}, \mathbf{o}, {}^b\underline{\mathbf{z}}) \triangleq {}^b\underline{\mathbf{z}}^\top \mathbf{P}_C\mathbf{T}^{-1}{}_O\mathbf{T}\mathbf{Q}^*_{(\mathbf{u}+\delta\mathbf{u})}{}_O\mathbf{T}^\top{}_C\mathbf{T}^{-\top}\mathbf{P}^\top{}^b\underline{\mathbf{z}}.$$

Jacobians

$$\frac{\partial^s \mathbf{e}}{\partial_O \boldsymbol{\xi}} = \mathbf{P}\frac{d\pi}{d\underline{\mathbf{s}}}\left({}_C\hat{\mathbf{T}}_t^{-1}{}_O\hat{\mathbf{T}}\left(\underline{\mathbf{s}}_j + \underline{\delta\hat{\mathbf{s}}}_j\right)\right){}_C\hat{\mathbf{T}}_t^{-1}\left[{}_O\hat{\mathbf{T}}\left(\underline{\mathbf{s}}_j + \underline{\delta\hat{\mathbf{s}}}_j\right)\right]^{\odot}$$

$$\frac{\partial^s \mathbf{e}}{\partial \delta\tilde{\mathbf{s}}_j} = \mathbf{P}\frac{d\pi}{d\underline{\mathbf{s}}}\left({}_C\hat{\mathbf{T}}_t^{-1}{}_O\hat{\mathbf{T}}\left(\underline{\mathbf{s}}_j + \underline{\delta\hat{\mathbf{s}}}_j\right)\right){}_C\hat{\mathbf{T}}_t^{-1}{}_O\hat{\mathbf{T}}\begin{bmatrix}\mathbf{I}_3\\ \mathbf{0}^\top\end{bmatrix} \in \mathbb{R}^{2\times 3}.$$

$$\frac{\partial^b \mathbf{e}}{\partial_O \boldsymbol{\xi}} = 2^b\underline{\mathbf{z}}^\top\mathbf{P}{}_C\hat{\mathbf{T}}_t^{-1}{}_O\hat{\mathbf{T}}\hat{\mathbf{Q}}^*_{(\mathbf{u}+\delta\hat{\mathbf{u}})}{}_O\hat{\mathbf{T}}^\top\left[{}_C\hat{\mathbf{T}}_t^{-\top}\mathbf{P}^{\top b}\underline{\mathbf{z}}\right]^{\circledcirc}$$

$$\frac{\partial^b \mathbf{e}}{\partial \delta\tilde{\mathbf{u}}} = (2(\mathbf{u}+\delta\hat{\mathbf{u}})\odot\mathbf{y}\odot\mathbf{y})^\top \in \mathbb{R}^{1\times 3}$$

$$\mathbf{y} \triangleq \begin{bmatrix}\mathbf{I}_3 & \mathbf{0}\end{bmatrix}{}_O\hat{\mathbf{T}}^\top{}_C\hat{\mathbf{T}}_t^{-\top}\mathbf{P}^{\top b}\underline{\mathbf{z}}.$$

Visual-Inertial Odometry

- We propose a framework similar to MSCKF for fusing the visual and inertial observations to estimate the robot states
- Instead of using quaternion, we use rotation matrix to parameterize the robot state

$$ {}_I\mathbf{x}_t \triangleq ({}_I\mathbf{R}_t,\ {}_I\mathbf{p}_t,\ {}_I\mathbf{v}_t,\ \mathbf{b}_g,\ \mathbf{b}_a) $$

- Moreover, we have derived a closed-form integration to propagate the robot state

$$ {}_I\hat{\mathbf{p}}^p_{t+1} = {}_I\hat{\mathbf{p}}_t + {}_I\hat{\mathbf{v}}_t\tau + \mathbf{g}\frac{\tau^2}{2} + {}_I\hat{\mathbf{R}}_t\mathbf{H}_L\left(\tau\left({}^i\boldsymbol{\omega}_t - \hat{\mathbf{b}}_{g,t}\right)\right)\left({}^i\mathbf{a}_t - \hat{\mathbf{b}}_{a,t}\right)\tau^2 $$

$$ {}_I\hat{\mathbf{v}}^p_{t+1} = {}_I\hat{\mathbf{v}}_t + \mathbf{g}\tau + {}_I\hat{\mathbf{R}}_t\mathbf{J}_L\left(\tau\left({}^i\boldsymbol{\omega}_t - \hat{\mathbf{b}}_{g,t}\right)\right)\left({}^i\mathbf{a}_t - \hat{\mathbf{b}}_{a,t}\right)\tau $$

$$ \mathbf{J}_L(\boldsymbol{\omega}) = \mathbf{I}_3 + \frac{1-\cos\|\boldsymbol{\omega}\|}{\|\boldsymbol{\omega}\|^2}\boldsymbol{\omega}_\times + \frac{\|\boldsymbol{\omega}\| - \sin\|\boldsymbol{\omega}\|}{\|\boldsymbol{\omega}\|^3}\boldsymbol{\omega}^2_\times $$

$$ \mathbf{H}_L(\boldsymbol{\omega}) = \frac{1}{2}\mathbf{I}_3 + \frac{\|\boldsymbol{\omega}\| - \sin\|\boldsymbol{\omega}\|}{\|\boldsymbol{\omega}\|^3}\boldsymbol{\omega}_\times + \frac{2(\cos\|\boldsymbol{\omega}\| - 1) + \|\boldsymbol{\omega}\|^2}{2\|\boldsymbol{\omega}\|^4}\boldsymbol{\omega}^2_\times . $$

Qualitative Results

- Backprojection of estimated keypoints and ellipsoid

Quantitative Results

TABLE II

PRECISION-RECALL EVALUATION ON KITTI OBJECT SEQUENCES

	Translation error →	≤ 0.5 m		≤ 1.0 m		≤ 1.5 m	
Rotation error	Method	Precision	Recall	Precision	Recall	Precision	Recall
≤ 30°	SubCNN [36]	0.10	0.07	0.26	0.17	0.38	0.26
	VIS-FNL [14]	**0.14**	0.10	**0.34**	**0.24**	**0.49**	**0.35**
	OrcVIO	0.10	**0.12**	0.18	0.21	0.22	0.25
≤ 45°	SubCNN [36]	0.10	0.07	0.26	0.17	0.38	0.26
	VIS-FNL [14]	**0.15**	0.11	**0.35**	0.25	**0.50**	**0.36**
	OrcVIO	**0.15**	**0.17**	0.25	**0.28**	0.31	0.35
–	SubCNN [36]	0.10	0.07	0.27	0.18	0.41	0.28
	VIS-FNL [14]	0.16	0.11	0.40	0.29	0.58	0.42
	OrcVIO	**0.29**	**0.33**	**0.50**	**0.56**	**0.62**	**0.69**

Thank you!

http://me-llamo-sean.cf/orcvio_githubpage/

Mo Shan
moshan@ucsd.edu

ERL

CONTEXTUAL ROBOTICS INSTITUTE

UC San Diego
JACOBS SCHOOL OF ENGINEERING
Electrical and Computer Engineering