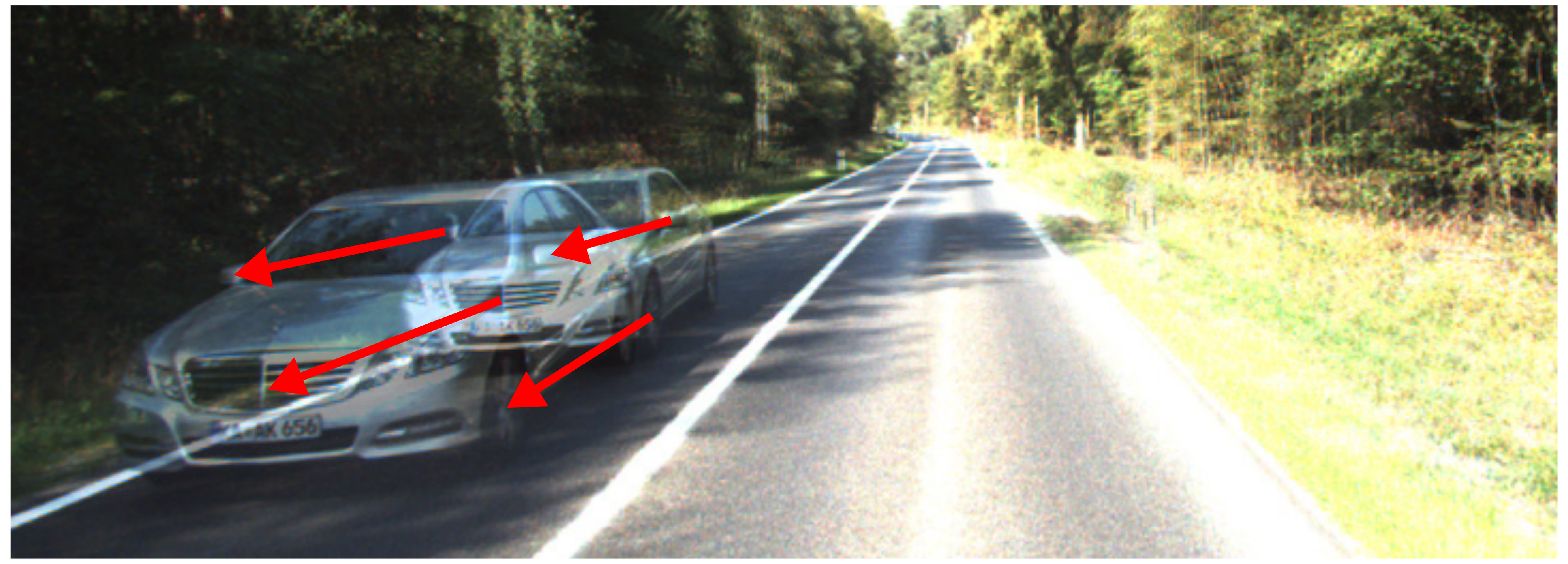
Bounding Boxes, Segmentations and Object Coordinates: How Important is Recognition for 3D Scene Flow Estimation in Autonomous Driving Scenarios?

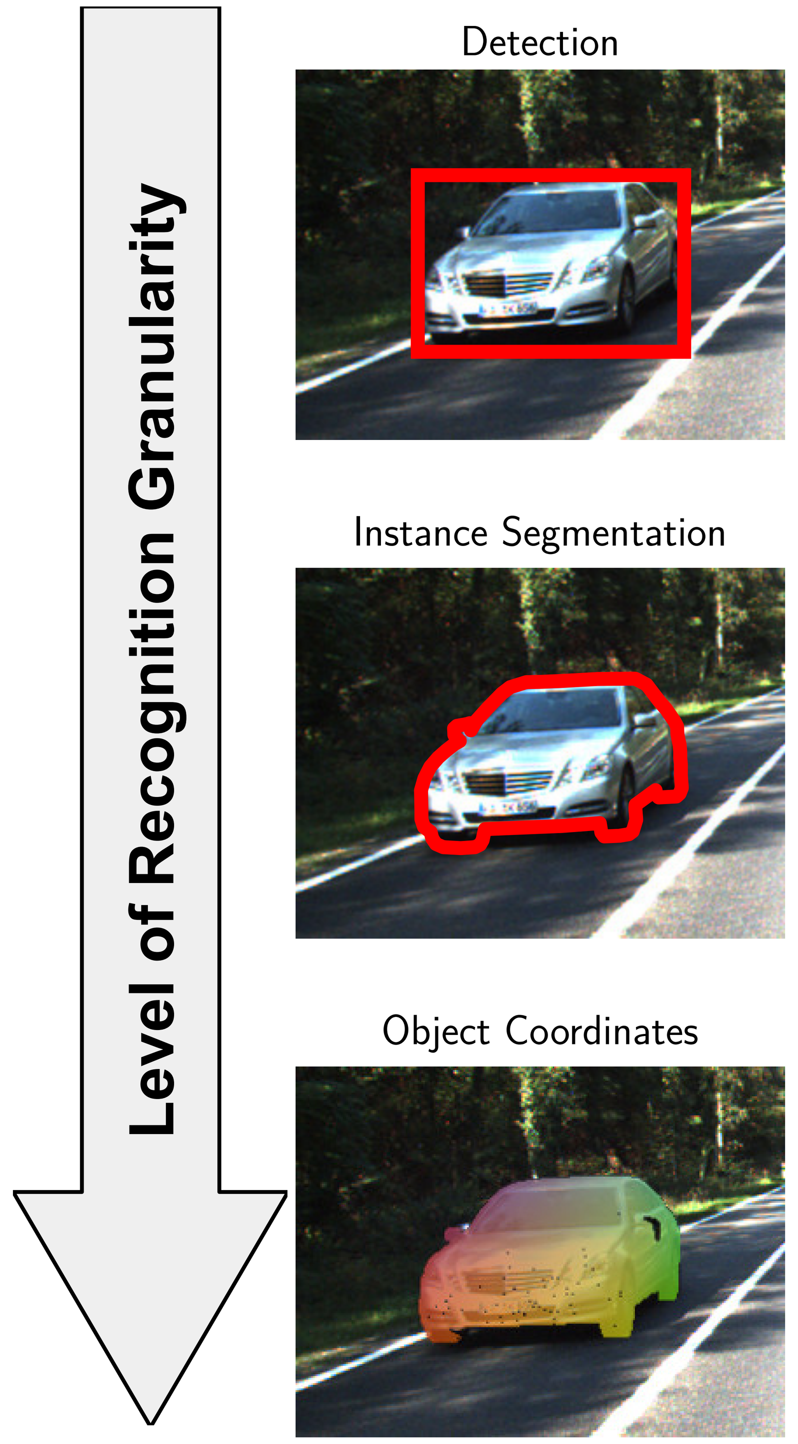
Aseem Behl[2,*] Omid Hosseini Jafari[1,*] Siva Karthik Mustikovela[1,*] Hassan Abu Alhaija[1] Carsten Rother[1] Andreas Geiger[2,3]

[1]Computer Vision Lab, TU Dresden, [2]Autonomous Vision Group, MPI for Intelligent Systems Tübingen, [3]Computer Vision and Geometry Group, ETH Zürich (* indicates equal contributions)

TECHNISCHE UNIVERSITÄT DRESDEN

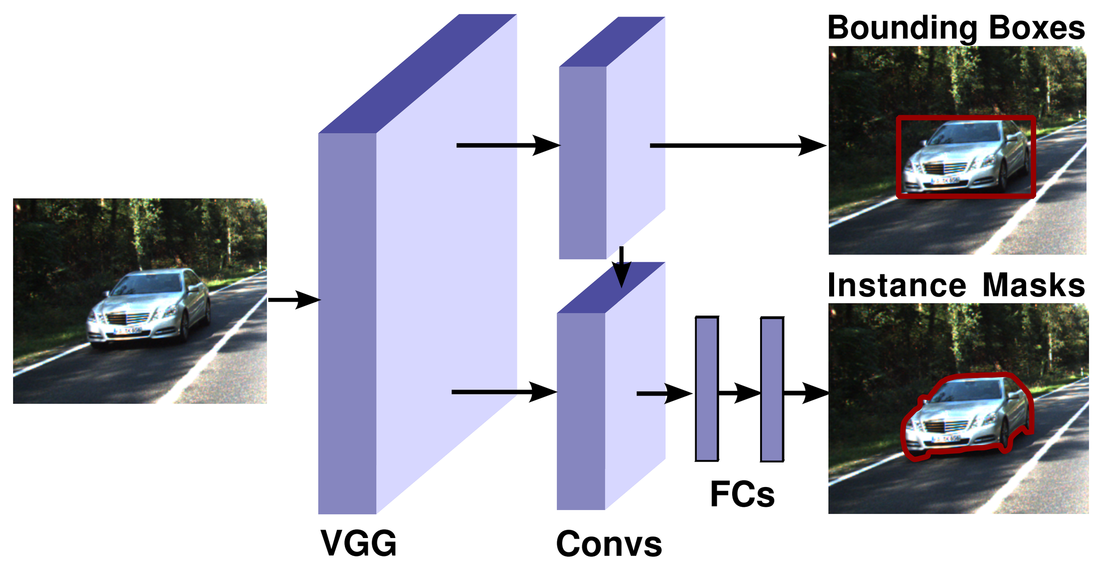


- **Motivation:** Scene flow estimation often fails in the presence of large displacements or reflective surfaces, e.g. the front wheel in the first frame appears similar to the back wheel in the second frame.
- **Goal:** To study the impact of three levels of recognition granularity on scene flow.



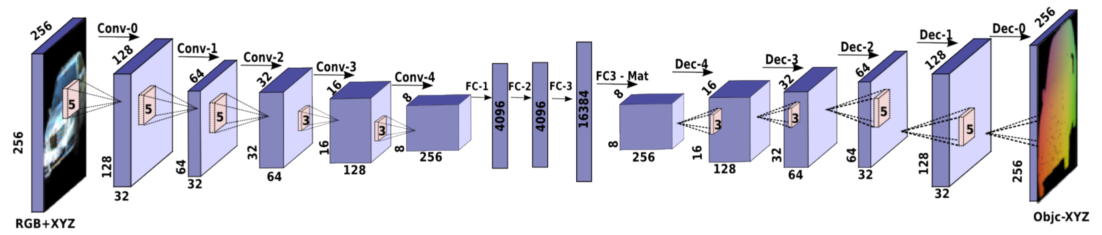
Semantic grouping

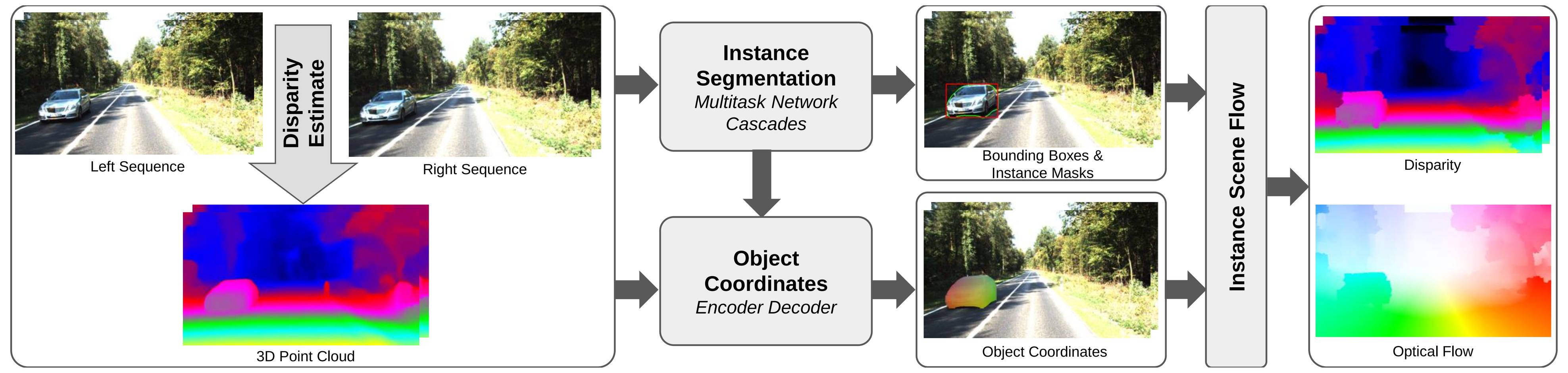
- Detection and instance segmentation can provide powerful cues to identify potential dynamic objects.
- Pixels grouped together are likely to move as a single rigid object in the case of vehicles.



Fine-grained geometric recognition

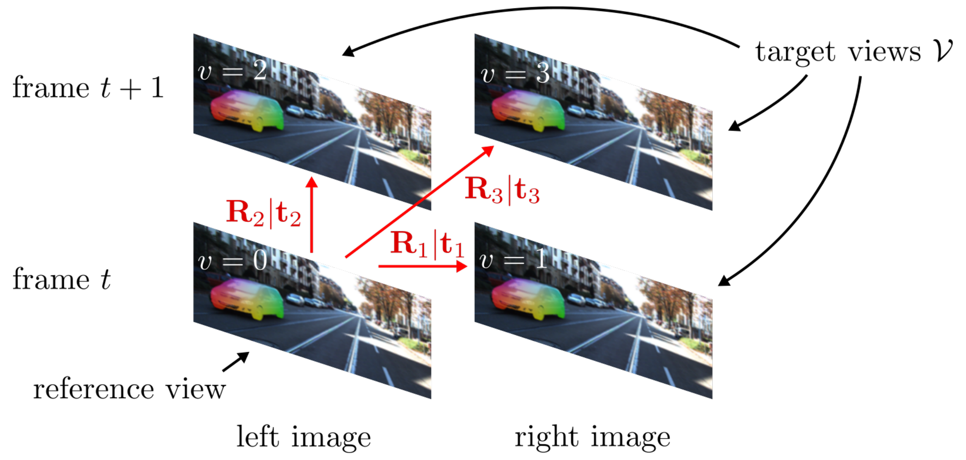
- Object coordinates are unique geometric labels of points on the object's surface in its local coordinate space.





Representation

Input



For each view, we compute the following information:

- Instance label maps $\mathbf{M}_v$, 2D bounding box segmentations or 2D instance segmentations.
- 3D object coordinates predicted by the network $\mathbf{C}_v$.

Parameters

- Each superpixel $i \in \mathcal{S}$ is parameterized by a plane $\mathbf{n}$ and an object index k_i
$$\mathbf{s}_i = (\mathbf{n}_i, k_i)^\mathsf{T}$$
- Each object $j \in \mathcal{O}$ is parameterized by its rigid motion parameters $\mathbf{o}_j \in SE(3)$.
- Superpixels inherit their rigid motion parameters from the associated object.

Model

$$\hat{\mathbf{s}}, \hat{\mathbf{o}} = \underset{\mathbf{s},\mathbf{o}}{\operatorname{argmin}} \underbrace{\varphi(\mathbf{s},\mathbf{o})}_{\text{data}} + \underbrace{\psi(\mathbf{s})}_{\text{smooth.}} + \underbrace{\chi(\mathbf{s},\mathbf{o})}_{\text{instance}}$$

Data term

- Enforces compatibility of appearance for corresponding points across views.

$$\varphi(\mathbf{s},\mathbf{o}) = \sum_{i \in \mathcal{S}} \sum_{\mathbf{p} \in \mathcal{R}_i} \sum_{v \in \mathcal{V}} \varphi_v^{\mathsf{D}}(\mathbf{p},\mathbf{q})$$

- $\mathbf{q}$ is the location of pixel $\mathbf{p}$ in reference view mapped to target view.

Smoothness term

- Encourages coherence of adjacent superpixels in terms of depth, orientation and motion. It decomposes as:

$$\psi(\mathbf{s}) = \sum_{i \sim j} \underbrace{\psi_{ij}^{\mathsf{G}}(\mathbf{n}_i,\mathbf{n}_j)}_{\text{geometry}} + \underbrace{\psi_{ij}^{\mathsf{M}}(\mathbf{s}_i,\mathbf{s}_j)}_{\text{motion}}$$

Instance term

- Enforces compatibility of appearance and part labeling (object coordinates) of detected instances across views.

$$\chi(\mathbf{s},\mathbf{o}) = \sum_{i \in \mathcal{S}} \sum_{\mathbf{p} \in \mathcal{R}_i} \sum_{v \in \mathcal{V}} \chi_v^{\mathsf{I}}(\mathbf{p},\mathbf{q})$$

Inference

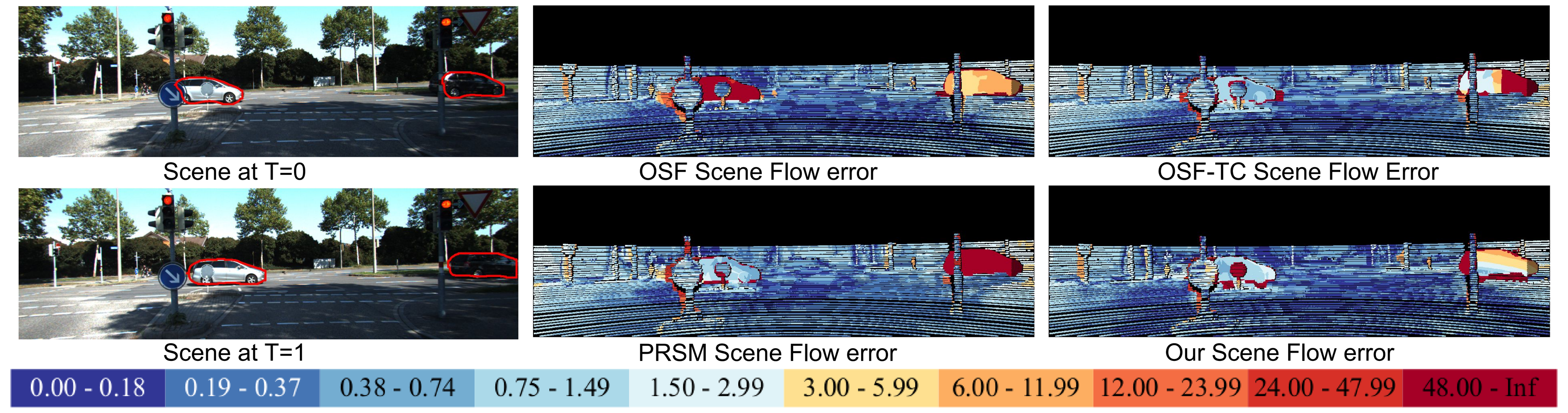
We use max-product particle BP to jointly infer shape and motion parameters with TRW-S for the inner loop.

At each iteration, particles are sampled for:

- Geometry variables: from a Gaussian distribution around the previous MAP estimate.
- Rigid motion variables: based on photoconsistency for each object individually after warping.

Experimental Results

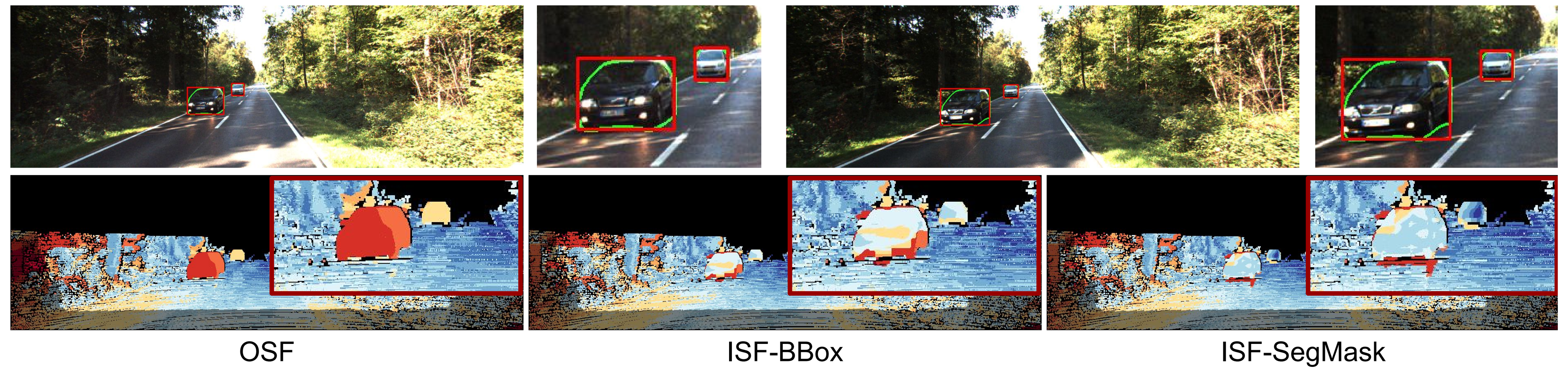
Comparison with state-of-the-art



	D1			D2			Fl			SF		
	bg	*fg*	*bg+fg*	*bg*	*fg*	*bg+fg*	*bg*	*fg*	*bg+fg*	*bg*	*fg*	*bg+fg*
PRSM [1]	3.02	10.52	**4.27**	5.13	15.11	6.79	**5.33**	13.40	6.68	6.61	20.79	8.97
OSF+TC [2]	4.11	9.64	5.03	5.18	15.12	6.84	5.76	13.31	7.02	7.08	20.03	9.23
OSF [3]	4.54	12.03	5.79	5.45	19.41	7.77	5.62	18.92	7.83	7.01	26.34	10.23
ISF-SegMask-CNNDisp	**4.12**	**6.17**	4.46	**4.88**	**11.34**	**5.95**	5.40	**10.29**	**6.22**	**6.58**	**15.63**	**8.08**

#1 KITTI optical flow benchmark & #1 KITTI scene flow benchmark

Ablation study



	D1			D2			Fl			SF		
	bg	*fg*	*bg+fg*	*bg*	*fg*	*bg+fg*	*bg*	*fg*	*bg+fg*	*bg*	*fg*	*bg+fg*
OSF	4.00	8.86	4.74	5.16	17.11	6.99	6.38	20.56	8.55	7.38	24.12	9.94
ISF-BBox	3.94	8.81	4.69	5.10	10.77	5.97	6.46	12.90	7.44	7.42	17.11	8.90
ISF-SegMask	4.06	7.97	4.66	5.26	9.20	5.86	6.72	10.78	7.34	7.74	14.60	8.79
ISF-SegMask-ObjCoord	4.08	7.98	4.68	5.27	9.20	5.87	6.72	10.84	7.35	7.75	14.66	8.80
ISF-SegMask-CNNDisp	3.55	3.94	3.61	4.86	4.72	4.84	6.36	7.31	6.50	7.23	8.72	7.46

References

[1] C. Vogel, K. Schindler, and S. Roth. Piecewise rigid scene flow. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2013.

[2] Michal Neoral and Jan Åăochman. Object scene flow with temporal consistency. In *Proc. of the Computer Vision Winter Workshop (CVWW)*, 2017.

[3] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015.