Three for one and one for three: Flow, Segmentation, and Surface Normals

Hoang-An Le, Anil S. Baslamisli, Thomas Mensink, Theo Gevers
Computer Vision Lab, Informatics Institute, University of Amsterdam

1. Overview

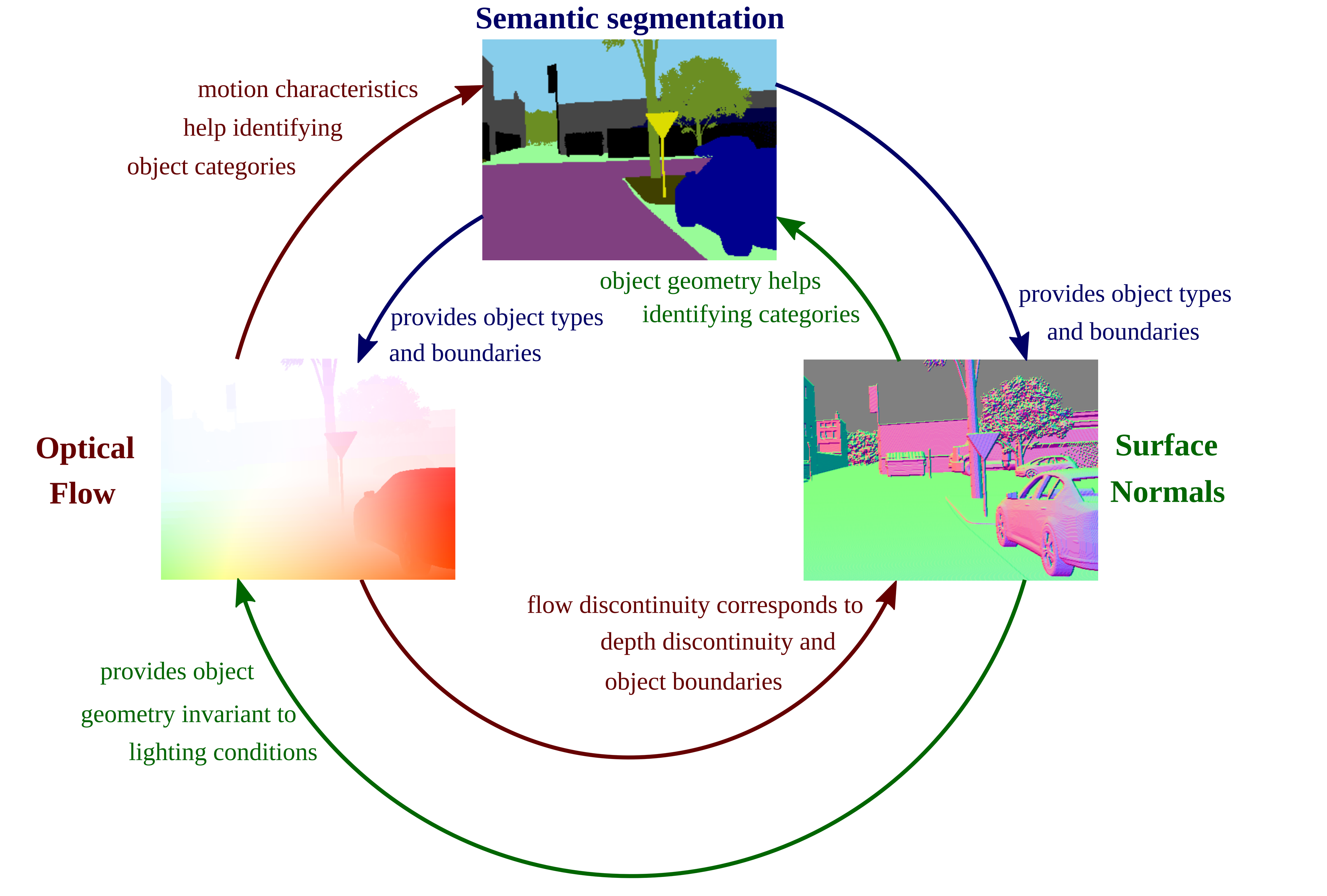
Goal: Study the mutual interaction of different modalities, inspired by human perception which combines different types of information:

Motion : Optical flow
Categories : Semantic segmentation
Geometry : Surface normals

Contributions:

- Analyzing the mutual interaction of the 3 modalities
- Combining modalities to improve the other using CNN
- Large scale synthetic dataset of outdoor nature scenes

2. Motivation



3. Method

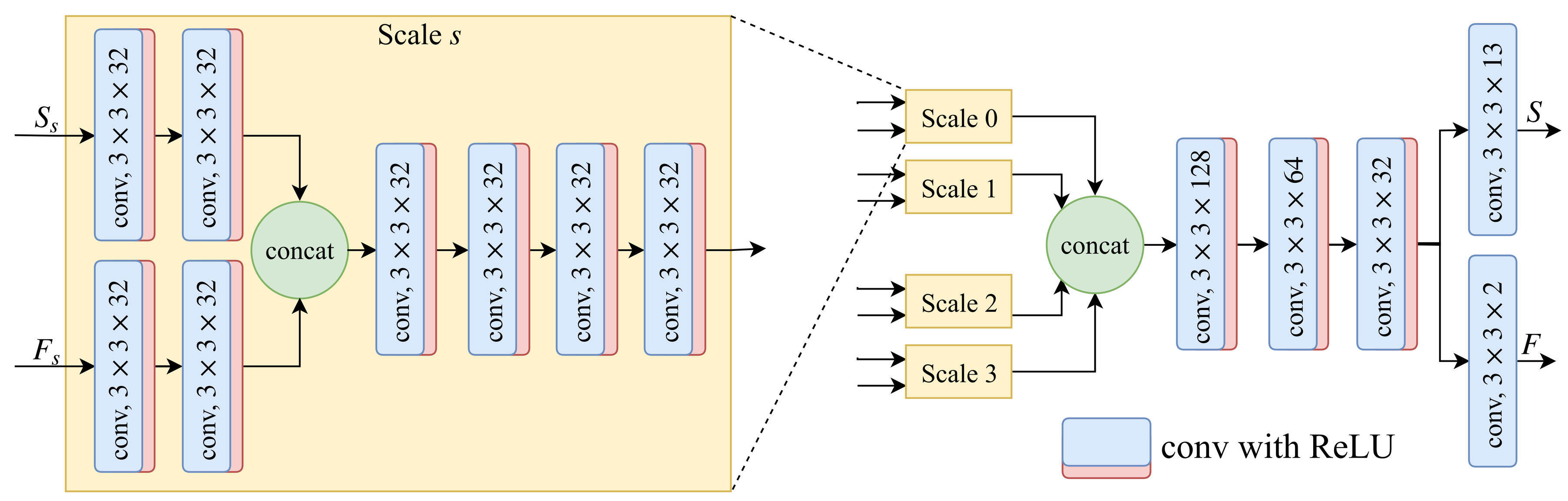
We follow refinement strategy to study the relationship between the 3 modalities:

- Each modality is first learned by a baseline network.
- Predicted modality is refined with other (either ground truth or predicted) using refinement architecture [5].
- The scale box s scales down input size to $\frac{1}{2^s}$, then up-samples the output back to the original size.

We also study cross-modality influence by doing single-task and multi-task refinement, at different coupling levels.

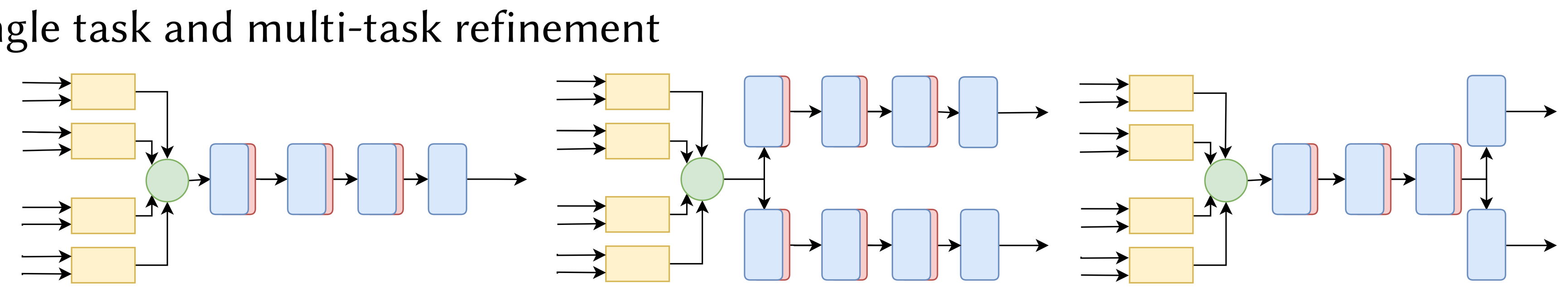
4. Architectures

Refinement architecture, inspired from [5]



(a) Scale box structure (b) refinement architecture

Single task and multi-task refinement



(a) *zero* coupling (b) *loose* coupling (c) *tight* coupling

5. Experiments

We refine each modality from the others (excluding RGB images) using either ground truth (GT) or predicted (PR) results to see how much one modality can be benefited from the others.

Datasets:

- virtual KITTI [4](20k images): synthetic driving city scenes
- UvA-Nature (15k images): synthetic nature scenes x 5 lighting types

Metrics

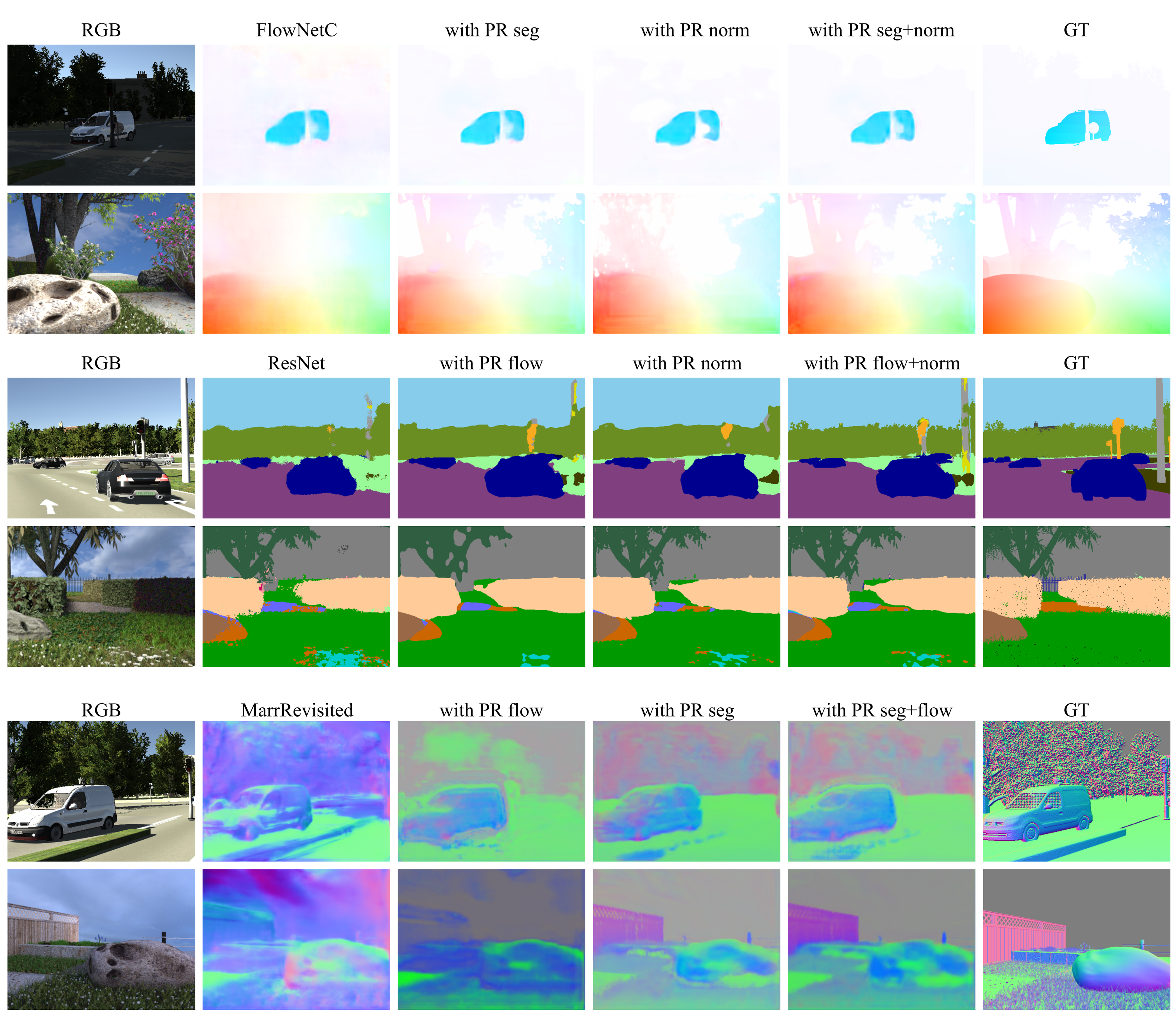
Optical Flow : end-point errors (epe) ↓
Segmentation : mean intersection-over-union (miou) ↑
Surface normals : root mean square errors (rmse) ↓

6. Refinement Couplings

Single task (*zero*) vs. multiple task (*loose*, *tight*) refinement

Target	Baseline	GT	Predicted			
		zero	*zero*	*loose*	*tight*	*tight+*
Segmentation (miou↑)	44.11	46.9	**44.78**	41.2	41.1	43.9
Optical flow (epe↓)	2.68	2.37	**2.40**	2.43	2.41	2.42

7. Qualitative results



8. Quantitative results

Flow

Dataset	FlowNetC [3]	with GT seg	with GT norm	with PR seg	with PR norm	with PR seg+norm
VKITTI	2.68	2.37	**2.36**	2.40	2.50	**2.39**
Nature	16.19	14.09	**13.92**	**14.16**	16.62	14.21

Segmentation

Dataset	ResNet [2]	with GT flow	with GT norm	with PR flow	with PR norm	with PR flow+norm
VKITTI	44.11	46.90	**50.0**	44.78	45.36	**47.55**
Nature	37.88	38.4	**41.6**	37.57	**38.83**	38.00

Surface normals

Dataset	MarrR [1]	with GT flow	with GT seg	with PR flow	with PR seg	with PR flow+seg
VKITTI	57.44	17.29	**16.78**	18.02	**17.24**	17.28
Nature	50.25	**13.20**	13.48	14.38	**12.56**	12.71

Acknowledgement

This project was funded by the EU Horizon 2020 program No. 688007 (TrimBot2020).

References

[1] A. Bansal, B. Russell, and A. Gupta. Marr Revisited: 2D-3D Alignment via Surface Normal Prediction. In *CVPR*, 2016.
[2] J. Cheng, Y.-h. Tsai, S. Wang, M.-H. Yang, and S. W. M.-h. Yang. SegFlow : Joint Learning for Video Object Segmentation and Optical Flow. In *ICCV*, 2017.
[3] A. Dosovitskiy, P. Fischery, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. V. D. Smagt, D. Cremers, and T. Brox. FlowNet: Learning optical flow with convolutional networks. In *ICCV*, 2016.
[4] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig. Virtual Worlds as Proxy for Multi-Object Tracking Analysis. In *CVPR*, 2016.
[5] O. H. Jafari, O. Groth, A. Kirillov, M. Y. Yang, and C. Rother. Analyzing modular CNN architectures for joint depth prediction and semantic segmentation. In *ICRA*, 2017.