Distilled Semantics for Comprehensive Scene Understanding from Videos

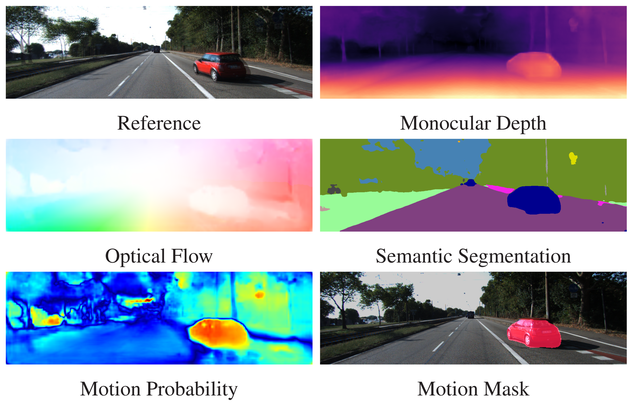
Fabio Tosi*, Filippo Aleotti*, Pierluigi Zama Ramirez*, Matteo Poggi, Samuele Salti, Luigi Di Stefano, Stefano Mattoccia

University of Bologna - Department of Computer Science and Engineering

{fabio.tosi5, filippo.aleotti2, pierluigi.zama, m.poggi, samuele.salti, luigi.distefano, stefano.mattoccia}@unibo.it, *joint first authorship

CVPR SEATTLE WASHINGTON JUNE 16-18 2020

Problem Definition and Contributions



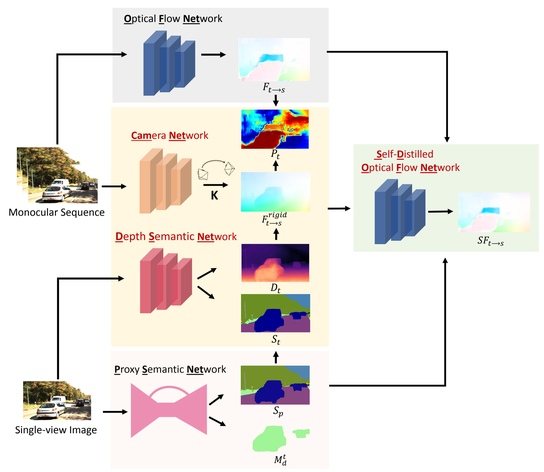
Purpose:

Obtain comprehensive information about the scene from monocular videos.

Key Contributions:

- The first real-time network for joint prediction of depth, optical flow, semantics and motion segmentation from monocular videos leveraging only on self-supervision and cheap proxy labels.
- A novel training protocol relying on proxy semantics and self-distillation to effectively address the self-supervised multi-task learning problem.
- State-of-the-art in monocular depth estimation, optical flow estimation among monocular multi-task frameworks and motion segmentation.

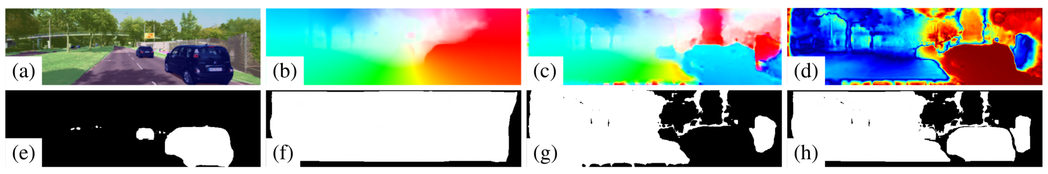
Architecture & Method



ΩNet Framework

1. Train the Depth and Semantic Network (DSNet) and the Camera Network. We train the single-image **depth** and **pose** estimation branches using a *self-supervised* approach from sequences of N images. Moreover, we *distill* cheap proxy labels by a pre-trained s.o.t.a. semantic segmentation network and we use them to train the **semantic** branch.
2. Train the **teacher** Optical Flow Network (OFNet) using a self-supervised strategy on 3 images.
3. Train a **self-distilled** OFNet with our semantic aware self-distillation paradigm to address common problems as moving objects and occlusions.
4. Compute the **motion segmentation** by joint reasoning about optical flow and semantics.

Semantic-Aware Self-Distillation Training Protocol



Based on two intuitions: **1)** Optical flow from OFNet is prone to errors in occluded regions due to the lack of photometric information, but it can provide good estimates for dynamic objects in the scene **2)** On the contrary, the rigid flow (derived from DSNet's depth) is more robust in such regions, but it cannot explain moving objects.

To soften these issues, we leverage **semantic segmentation** (a) together with **rigid flow** (b), **flow** from OFNet (c), and **motion probabilities** (d), the warmer the higher. From (a) we obtain **semantic priors** (e), that we combine with **boundary mask** (f) and **consistency mask** (g), derived from (d), to obtain the **final mask** M (h). We train a robust self-distilled OFNet sourcing supervision from rigid flow where M is black, from OFNet otherwise.

Experiments & Results

Monocular Depth Estimation - Eigen split of KITTI

Method	Abs Rel	Sq Rel	RMSE	RMSE log
Godard *et al.*[5] (640 × 192)	0.132	1.044	5.142	0.210
Godard *et al.*[5] (1024 × 320)	**0.115**	0.882	4.701	0.190
Chen *et al.*[3]	0.135	1.070	5.230	0.210
Luo *et al.*[2]	0.141	1.029	5.350	0.216
Ranjan *et al.*[1]	0.139	1.032	5.199	0.213
Gordon *et al.*[4]	0.128	0.959	5.230	-
ΩNet(640 × 192)	0.120	0.792	4.750	0.191
ΩNet(1024 × 320)	0.118	**0.748**	**4.608**	**0.186**

State-of-the-art for monocular depth estimation

Semantic segmentation - Cityscapes and KITTI

Method	Train	Test	mIoU Class	mIoU Cat.	Pix.Acc.
Chao et al. [7]	CS(S)	CS	**76.37**	**89.22**	**95.35**
ΩNet	CS(P)	CS	54.80	82.92	92.50
Chao et al [7]	CS(S)	K	**44.74**	68.20	72.07
ΩNet	CS(P)	K	43.80	**74.31**	**88.31**
ΩNet	CS(P) + K(P)	K	46.68	75.84	88.12

Generalization results from Cityscapes (CS) to KITTI (K)

Optical Flow - KITTI 2015

	train			test
Method	Noc	All	F1	F1
Chen *et al.*[3] †	5.40	8.95	-	-
Ranjan *et al.*[1]	-	6.21	26.41%	-
Luo *et al.*[2]	-	5.84	-	21.56%
ΩNet	**3.29**	**5.39**	**20.0%**	**19.47%**

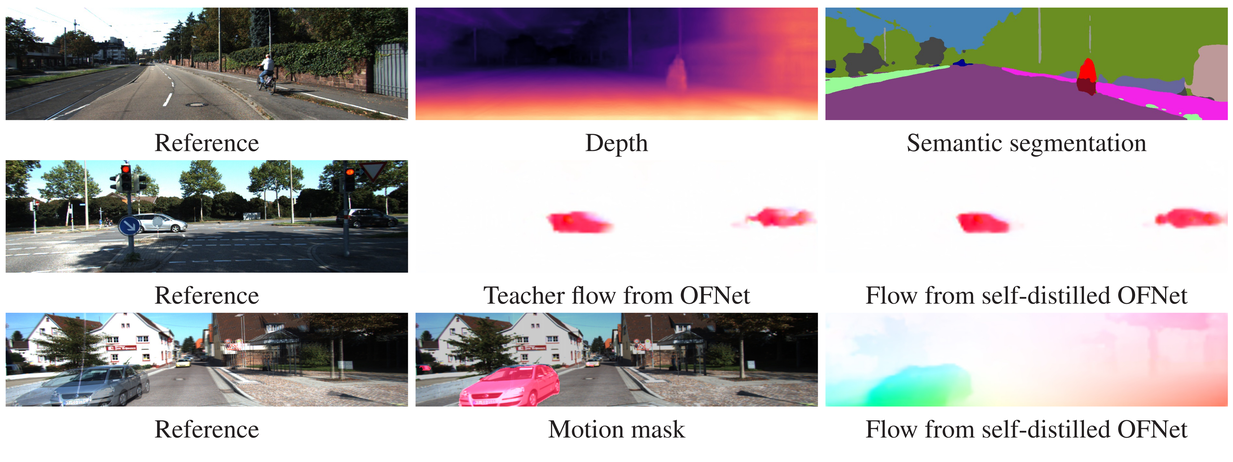
State-of-the-art for optical flow estimation among multi-task frameworks.

Motion Segmentation - KITTI 2015

Method	Pixel Acc.	Mean Acc.	Mean IoU	f.w. IoU
Luo *et al.*[2]	0.88	0.63	0.50	0.86
Ranjan *et al.*[1]	0.87	0.79	0.53	0.85
ΩNet	**0.98**	**0.86**	**0.75**	**0.97**
ΩNet (Proxy [6])	0.98	0.87	0.77	0.97

State-of-the-art for motion-segmentation.

Qualitative results



References

[1] Ranjan et al., Competitive collaboration: Joint unsupervised learning of depth, camera motion,optical flow and motion segmentation.

[2] Luo et al., Every pixel counts++:Joint learning of geometry and motion with 3d holistic understanding.

[3] Chen et al., Self-supervised learning with geometric constraints in monocular video: Connecting flow, depth, and camera.

[4] Gordon et al., Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras.

[5] Godard et al., Unsupervised monocular depth estimation with left-right consistency.

[6] Chen et al., Searching for efficient multi-scale architectures for dense image prediction.

[7] Chao et al., Hardnet: A low memory traffic network.

Links

Paper

GitHub Code

Acknowledgement

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.