Gurram, Akhil; Urfalioglu, Onay; Halfaoui, Ibrahim; [et al.]. «Semantic monocular depth estimation based on artificial intelligence». IEEE Intelligent Transportation Systems Magazine, Vol. 13, issue 4 (2021), p. 99-103. 5 pàg. DOI 10.1109/MITS.2019.2926263

Semantic Monocular Depth Estimation based on Artificial Intelligence

Akhil Gurram   Onay Urfalioglu   Ibrahim Halfaoui   Fahd Bouzaraa   Antonio M. López

***Abstract*—Depth estimation provides essential information to perform autonomous driving and driver assistance. A promising line of work consists of introducing additional semantic information about the traffic scene when training CNNs for depth estimation. In practice, this means that the depth data used for CNN training is complemented with images having pixel-wise semantic labels where the same raw training data is associated with both types of ground truth, i.e., depth and semantic labels. The main contribution of this paper is to show that this hard constraint can be circumvented, i.e., that we can train CNNs for depth estimation by leveraging the depth and semantic information coming from heterogeneous datasets. In order to illustrate the benefits of our approach, we combine KITTI depth and Cityscapes semantic segmentation datasets, outperforming state-of-the-art results on monocular depth estimation.**

***Index Terms*—Monocular Depth Estimation, Semantic Segmentation, Multi-task learning.**

I. Introduction

In contrast to stereo vision, monocular depth estimation is a relatively young topic, which has become affordable thanks to convolutional neural networks (CNNs).

Godard *et al.* [6] propose an unsupervised method to learn a monocular depth estimator from stereo data; a photometric loss function with terms accounting for left-right consistency is used during CNN training. Kuznietsov *et al.* [8] propose a semi-supervised method to estimate inverse depth maps by combining an appearance matching loss similar to [6] and a supervised objective function using sparse depth ground truth (GT) from LIDAR.

Supervised methods, *i.e.* fully relying on depth GT, are proposed by several authors too. Xu *et al.* [12] fuse complementary information derived from multiple CNNs by means of Conditional Random Fields (CRFs). Similarly, Liu *et al.* [9] present a CNN with a CRF-based loss layer. In Cao *et al.* [1] the depth GT is discretized into several distance ranges for training a FCN-residual network that predicts these ranges pixel-wise; which is followed by a CRF post-processing enforcing local depth coherence. Xu *et al.* [13] propose a structured attention model to automatically regulate the amount of information transferred between CNN features at different scales. Luo *et al.* [14] reformulate monocular depth estimation as a view synthesis procedure followed by stereo matching; obtaining competitive results by fine-tuning based on additional 200 high-quality disparity labels.

II. Monocular Depth Estimation

In this paper, we propose to leverage heterogeneous datasets to train a CNN for depth estimation; *i.e.* training can rely on one dataset having *only* depth GT, along with a different dataset with *only* pixel-wise semantic GT. We divide the training process into two phases.

In the first phase, we use multi-task learning [7] for pixel-wise depth and semantic CNN-based classification (Fig. 1). This means that at this stage depth is discretized, a task that has been shown to be useful for supporting instance segmentation [11]. We use a CNN architecture consisting of a common feature extractor followed by two task-specific branches. We denote the layers in the common sub-net as DSC (depth-semantic classification) layers, the depth specific sub-net as DC layers, and the semantic segmentation specific sub-net as SC layers. At training time, we apply a conditional calculation of gradients during back-propagation, which we call *conditional flow*. More specifically, the common sub-net is always active, but the origin of each data sample determines which specific sub-net branch is also active during back-propagation (Fig. 1). We alternate batches of depth and semantic GT samples.

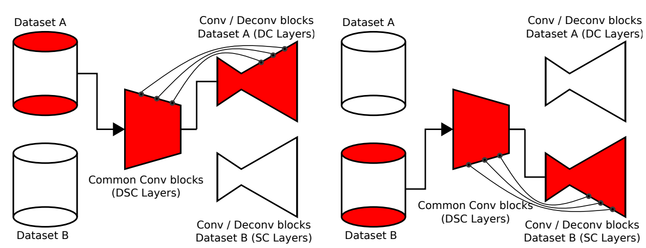


Fig. 1: Phase one: conditional backward passes (see main text). We also use skip connections linking convolutional and deconvolutional layers with equal spatial sizes.

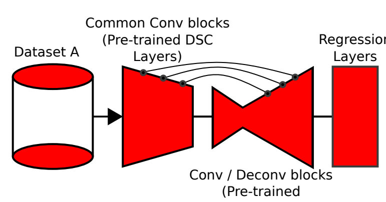


Fig. 2: Phase two: the pre-trained (DSC+DC) CNN is augmented by regression layers for fine-tuning, resulting in the (DSC-DRN) network for depth estimation.

Approaches \ metrics	cap (m)	Lower the better					Higher the better		
		rel	sq-rel	rms	rms-log	$\log_{10}$	δ<1.25	δ<1.25^2	δ<1.25^3
Liu fine-tune [9]	80	0.217	1.841	6.986	0.289	-	0.647	0.882	0.961
Godard – K [6]	80	0.155	1.667	5.581	0.265	0.066	0.798	0.920	0.964
Godard – K + CS [6]	80	0.124	1.240	5.393	0.230	0.052	0.855	0.946	0.975
Cao [1]	80	0.115	-	4.712	0.198	-	0.887	0.963	0.982
kuznietsov [8]	80	0.113	0.741	4.621	0.189	-	0.862	0.960	0.986
Xu [12]	80	0.125	0.899	4.685	**0.154**	-	0.816	0.951	0.983
Xu [13]	80	0.122	0.897	4.677	-	-	0.818	0.954	0.985
Luo [14](same dataset)	80	0.102	0.700	4.681	0.200	-	0.872	0.954	0.978
Luo [14](fine-tuned*)	80	**0.094**	0.626	**4.252**	0.177	-	**0.891**	0.965	0.984
Ours (DRN)	80	0.112	0.701	4.424	0.188	0.0492	0.848	0.958	0.986
Ours (DC-DRN)	80	0.110	0.698	4.529	0.187	0.0487	0.844	0.954	0.984
Ours	80	0.100	**0.601**	4.298	0.174	**0.0440**	0.874	**0.966**	**0.989**
Garg [4]	50	0.169	1.512	5.763	0.236	-	0.836	0.935	0.968
Godard – K [6]	50	0.149	1.235	4.823	0.259	0.065	0.800	0.923	0.966
Godard – K + CS [6]	50	0.117	0.866	4.063	0.221	0.052	0.855	0.946	0.975
Cao [1]	50	0.107	-	3.605	0.187	-	0.898	0.966	0.984
kuznietsov [8]	50	0.108	0.595	3.518	0.179	-	0.875	0.964	0.988
Luo [14] (same dataset)	50	0.097	0.539	3.503	0.187	-	0.885	0.960	0.981
Luo [14] (fine-tuned*)	50	**0.090**	0.499	**3.266**	0.167	-	**0.902**	0.968	0.986
Ours (DRN)	50	0.109	0.618	3.702	0.182	0.0477	0.862	0.963	0.987
Ours (DC-DRN)	50	0.107	0.602	3.727	0.181	0.0470	0.865	0.963	0.988
Ours	50	0.096	**0.482**	3.338	**0.166**	**0.0420**	0.886	**0.980**	**0.995**

TABLE I: Results on Eigen *et al.*'s KITTI split [3]. DRN - Depth regression network, DC-DRN - Depth regression model with pretrained classification network, DSC-DRN - Depth regression network trained with our conditional flow approach. Evaluation metrics as follows, rel: avg. relative error, sq-rel: square avg. relative error, rms: root mean square error, rms-log: root mean square log error, log_{10}: avg. log_{10} error, $\delta < \tau$: % of pixels with relative error $< \tau$ ($\delta \geq 1$; $\delta = 1$ no error). Godard − K means using KITTI for training, and "+ CS " adding Cityscapes too. Bold stands for **best**, underline for second best. Luo *et al.* [14] (fine-tuned*) approach uses additional 200 HQ disparity labels in training.

In the second phase, we focus on depth regression. In particular, we add layers that perform regression taking the depth classification layers as input (Fig. 2). We use standard losses for classification and regression tasks, *i.e.* cross-entropy and L1 losses, respectively.

III. Experimental Results

A. Datasets

We evaluate our approach on KITTI dataset [5], following the commonly used Eigen *et al.* [3] split for depth estimation. It consists of 22,600 training images and 697 testing images, *i.e.* RGB images with associated LIDAR data. To generate dense depth ground truth for each RGB image we follow Premebida *et al.* [10]. We use half down-sampled images, *i.e.* 188×620 pixels, for training and testing. Moreover, we use 2,975 images from Cityscapes dataset [2] with per-pixel semantic labels.

B. Results

We compare our approach to supervised methods such as Liu *et al.* [9] and Cao *et al.* [1], Xu *et al.* [12] [13], Luo *et al.* [14] and unsupervised methods such as Garg *et al.* [4] and Godard *et al.* [6], and semi-supervised method Kuznietsov *et al.* [8]. Quantitative results are shown in Table I for two different distance ranges (cap), namely [1,50]m and [1,80]m. As for the mentioned works, we follow the metrics proposed by Eigen *et al.* [3]. Note how our method outperforms the state-of-the-art models in all metrics but one (being second best). Fig. 3 shows qualitative results on KITTI comparing with Godard *et al.* [6]. Fig. 4 shows similar results for Cityscapes; *i.e.* illustrating generalization by the model trained on KITTI.

IV. Conclusion

The underlying assumption in the presented work is that object contours are shared between depth and semantic segmentation GT up to a large extend. Accordingly, we have presented a method to train a CNN for monocular depth estimation using datasets with depth GT, while improving its accuracy by leveraging semantic GT from other datasets as main novelty. The presented qualitative and quantitative experiments confirm our assumption by a multi-task training using KITTI RGB images with their depth GT, as well as Cityscapes RGB images with their semantic segmentation GT. In particular, we obtain state-of-the-art results on the depth-from-mono task of the

About the Authors

KITTI dataset. As future work we plan to incorporate temporal coherence in line with Zhou *et al.* [15].

Acknowledgement. Antonio M. López acknowledges the financial support by the Spanish TIN2017-88709-R (MINECO/AEI/FEDER, UE), and by ICREA under the ICREA Academia Program. As CVC/UAB researcher,

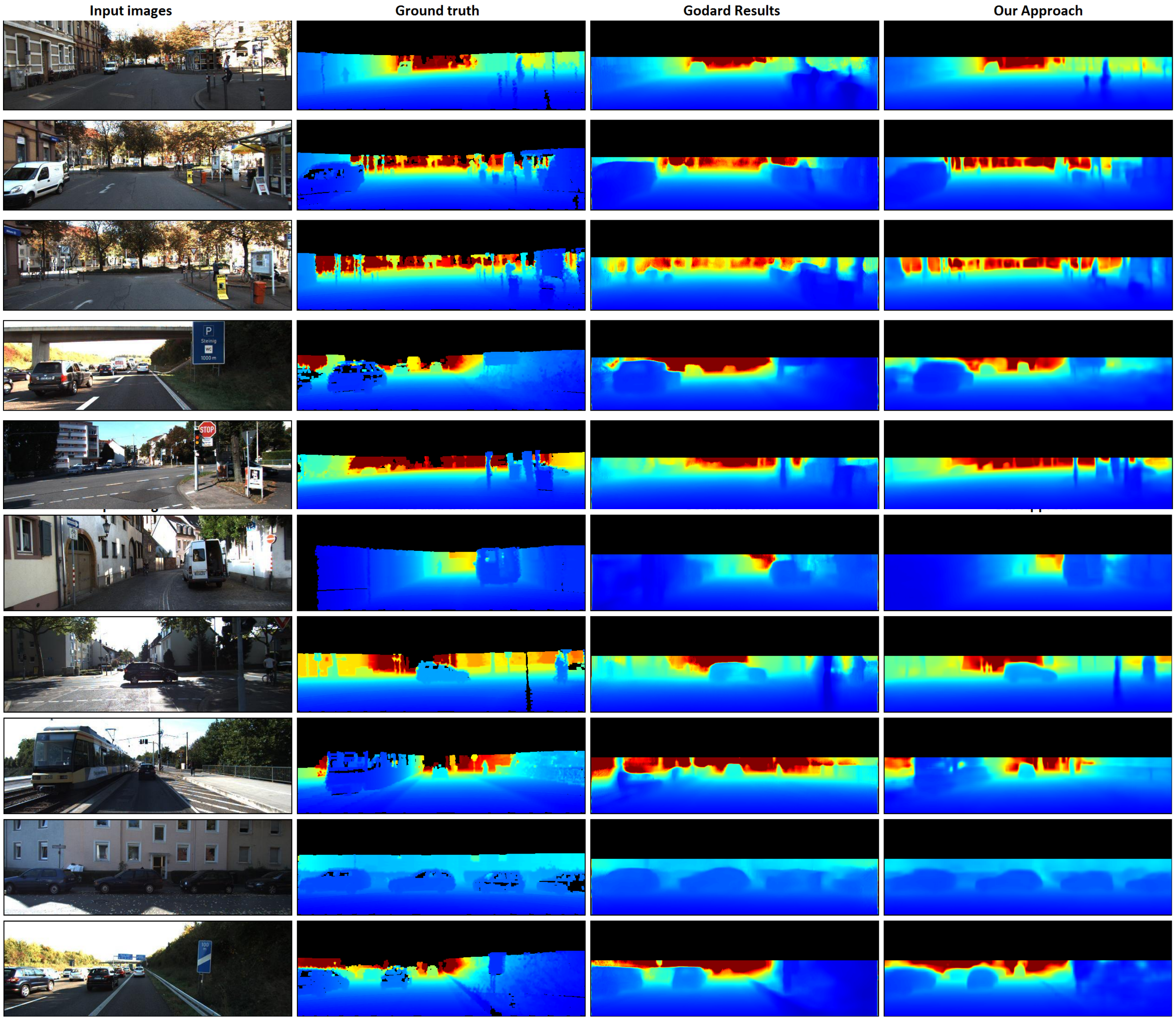


Fig. 3: Left to right: RGB image (KITTI), depth ground truth, Godard *et al.* [6] and our depth estimation results. In this figure, we show on the right side of the image that Godard *et al.* [6] results yield poor detection quality along with inaccurate depth estimation for specific relevant objects such as cars, tram or poles. On the other hand, our method provides a more accurate depth estimation which can be seen on the right most column.

Antonio also acknowledges the Generalitat de Catalunya CERCA Program and its ACCIO agency.

References

[1] Y. Cao, Z. Wu, and C. Shen, "Estimating depth from monocular images as classification using deep fully convolutional residual networks," *IEEE T-CSVT*, 2017.

[2] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The Cityscapes dataset for semantic urban scene understanding," in *CVPR*, 2016.

[3] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *NIPS*, 2014.

[4] R. Garg, V. Kumar, G. Carneiro, and I. Reid, "Unsupervised cnn for single view depth estimation: Geometry to the rescue," in *ECCV*, 2016.

[5] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *IJRR*, vol. 32, no. 11, pp. 1231–1237, 2013.

[6] C. Godard, O. Aodha, and G. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *CVPR*, 2017.

[7] I. Kokkinos, "Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6129–6138.

[8] Y. Kuznietsov, J. Stückler, and B. Leibe, "Semi-supervised deep learning for monocular depth map prediction," in *CVPR*, 2017.

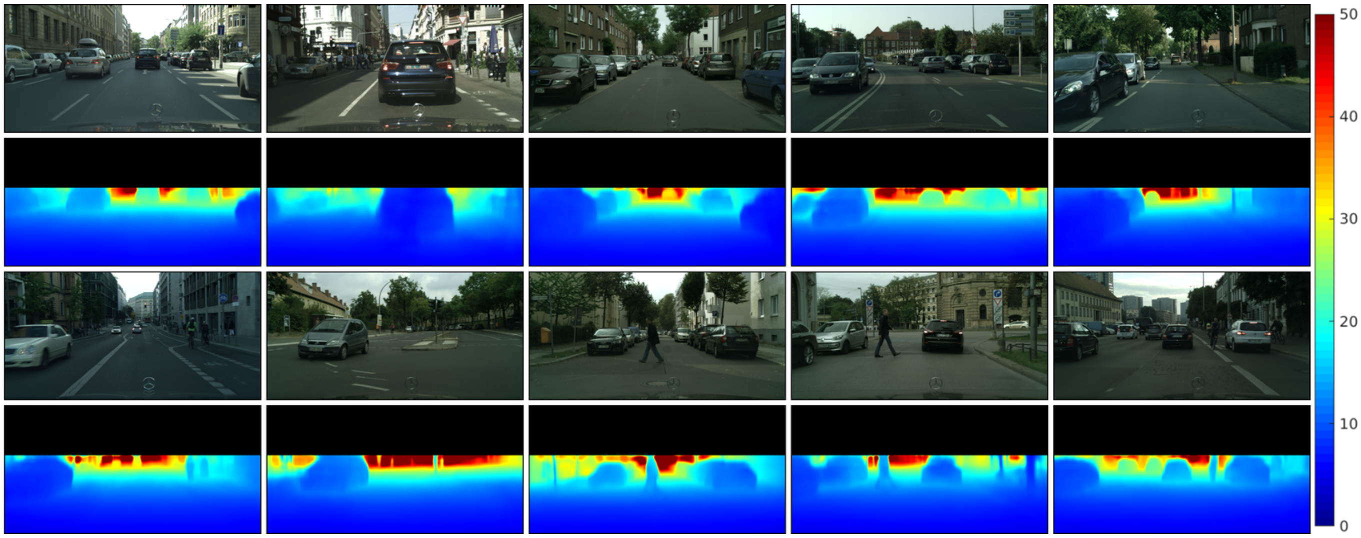


Fig. 4: Depth estimation results on Cityscapes validation and testing set images. This Cityscapes dataset is used for the task of semantic segmentation and we couldn't provide quantitative results as it doesn't have depth ground truth. **Note**: The validation set images are not used for training the network.

[9] F. Liu, C. Shen, G. Lin, and I. Reid, "Learning depth from single monocular images using deep convolutional neural fields," *IEEE T-PAMI*, vol. 38, no. 10, pp. 2024–2039, 2016.

[10] C. Premebida, J. Carreira, J. Batista, and U. Nunes, "Pedestrian detection combining RGB and dense LIDAR data," in *IROS*, 2014.

[11] J. Uhrig, M. Cordts, U. Franke, and T. Brox, "Pixel-level encoding and depth layering for instance-level semantic labeling," in *GCPR*, 2016.

[12] D. Xu, E. Ricci, W. Ouyang, X. Wang, and N. Sebe, "Monocular depth estimation using multi-scale continuous crfs as sequential deep networks," *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[13] D. Xu, W. Wang, H. Tang, H. Liu, N. Sebe, and E. Ricci, "Structured attention guided convolutional neural fields for monocular depth estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3917–3925.

[14] M. L. J. P. W. S. H. L. L. L. Yue Luo, Jimmy Ren, "Single view stereo matching," in *CVPR*, 2018.

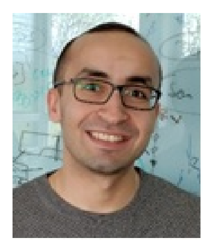
[15] T. Zhou, M. Brown, N. Snavely, and D. Lowe, "Unsupervised learning of depth and ego-motion from video," in *CVPR*, 2017.

About the Authors

Akhil Gurram received his Master's degree with focus on computer vision and deep learning at the Universitat Autonoma de Barcelona in September 2016 and currently working as a PhD-student at Huawei GRC in Munich. His area of interests are in Multi-task Learning for Depth estimation, Learnable localization and End-to-End driving modules using computer vision and deep learning.

Onay Urfalioglu is currently team leader at Huawei Technologies. He obtained a Ph.D. degree at the Leibniz University of Hannover in 2006. From 2006-2011, he was a Post-Doctoral researcher at the ISTI/CNR Pisa and at the Bilkent University, Ankara. His recent activities are using Deep Learning for Mapping & Localization applications.

Ibrahim Halfaoui studied electrical engineer at the Technical University of Munich and received his Master degree in June 2014. He is currently working as a PhD student at Huawei GRC in Munich. His research interests lie in the areas of deep learning and computer vision.



Fahd Bouzaraa studied Electrical Engineering and Information Technology at the Technical University of Munich. His areas of expertise revolves around developing machine learning and specifically deep learning-based solutions for computer vision tasks such as HDR rendering for dynamics scenes and scene understanding.

Antonio M. López received his PhD degree from the Univ. Autonoma de Barcelona (UAB) in 2000. Since 1992, he has been giving lectures at the UAB, where he has a tenure position as associate professor. In 1996, he participated in the foundation of the UAB's Computer Vision Center (CVC). In 2003, he started the CVC's ADAS group; since then, Antonio is the Principal Investigator on ADAS and Autonomous Driving at CVC. He has also been awarded with an ICREA Academia grant.