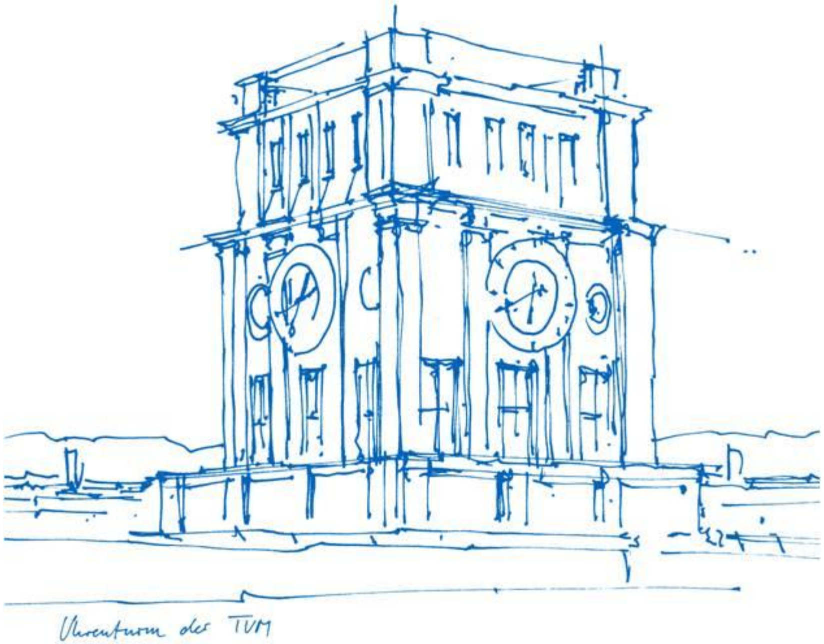
D3VO: Deep Depth, Deep Pose and Deep Uncertainty for Monocular Visual Odometry (2020)

by Nan Yang, Lukas von Stumberg, Rui Wang, Daniel Cremers

The Evolution of Motion Estimation and Real-time 3D Reconstruction Seminar

Supervisor: Lukas Köstler

Başak Melis Öcal
Technical University of Munich
26.01.2021



Introduction

- Monocular sparse direct visual odometry (VO) framework which exploits deep neural networks on **three levels - deep depth, pose and uncertainty.**

- Outperforms SOTA monocular VO methods by a large margin.

- Achieves comparable results to SOTA stereo/LiDAR odometry and visual-inertial odometry (VIO) methods, **while using only a single camera.**

Figure 1. Performance of D3VO on EuRoC MAV Dataset and KITTI Odometry Benchmark [16].

Outline

Overview

Method Description

Experiments & Results

Personal Comments

Summary

Overview

Contributions to Limitations of VO:

Scale drift & low robustness of monocular VO
➔ Deep self-supervised monocular depth estimation network

Limited utilization of deep neural networks
➔ Deep pose estimation

Inconsistent illumination between training image pairs
➔ Brightness alignment of image pairs

Photometric uncertainty
➔ Deep uncertainty estimation

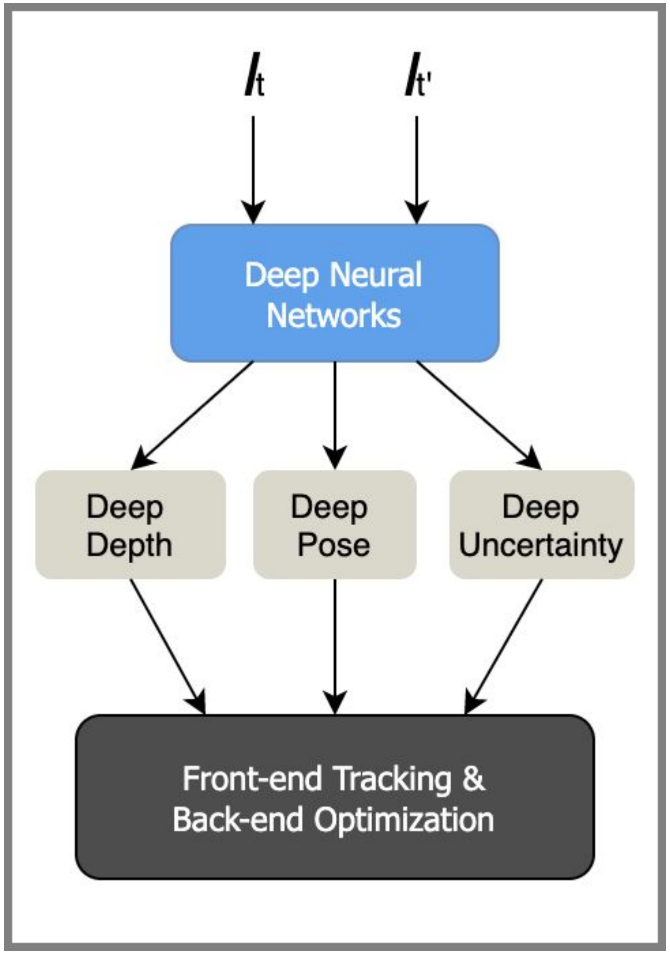


Figure 2. D3VO Framework

Overview

Contributions to Limitations of VO

Integration of predicted depth into VO system

➔ Initialize 3D points with the predicted depth

➔ Virtual stereo term

Integration of predicted pose into VO system

➔ Incorporate into both front-end tracking and back-end optimization

Integration of predicted uncertainty map into VO system

➔ Use the predicted uncertainty map in the weighting function of the VO energy function

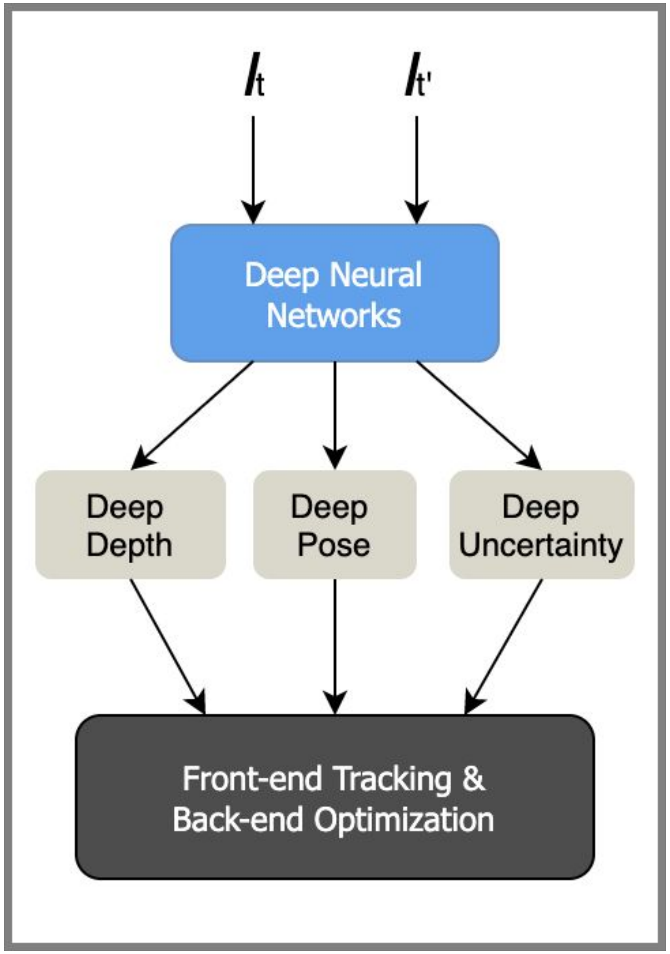


Figure 2. D3VO Framework

Method Description

Self-supervised Network

MonoDepth2 [4]: In the **absence of ground truth depth**, train a depth estimation model using **image reconstruction as the supervisory signal**.

- Learn depth with Depth Net, motion with Pose Net.
- Minimize **photometric reprojection error** based on photometric constancy:

$$L_{self} = \frac{1}{|V|} \sum_{\mathbf{p} \epsilon V} \min_{t'} r(I_t, I_{t' \to t}) \tag{1}$$

$$I_{t' \to t} = I_{t'} \langle proj\,(D_t, T_{t \to t'}, K) \rangle \tag{2}$$

$$I_{t'} \,\epsilon\, \{I_{t-1}, I_{t+1}, I_{t^s}\}$$

t: Index of target frame
t': index of all source frames
V: Set of all pixels on I_t
D_t: Predicted depth
$T_{t \to t'}$: Predicted pose

K: Camera intrinsics
$I_{t' \to t}$: Synthesized I_t
r(): Photometric error
proj(): Projection function
< >: Bilinear sampler

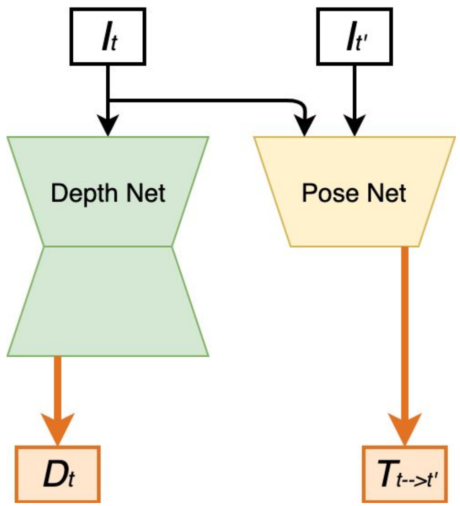


Figure 3. MonoDepth2 architecture

Method Description

Self-supervised Network

Photometric constancy assumption may be violated due to **illumination changes and auto-exposure of the camera** to which both **L1 and SSIM losses are not invariant.**

➔ Align the illumination of *It* to *It'* by predicting affine transformation parameters via pose network.

➔ Minimize **photometric reprojection error** based on **photometric constancy + affine transformation:**

$$L_{self} = \frac{1}{|V|} \sum_{\mathbf{p} \epsilon V} \min_{t'} r(a_{t\to t'} I_t + b_{t\to t'}, I_{t'\to t}) \qquad (3)$$

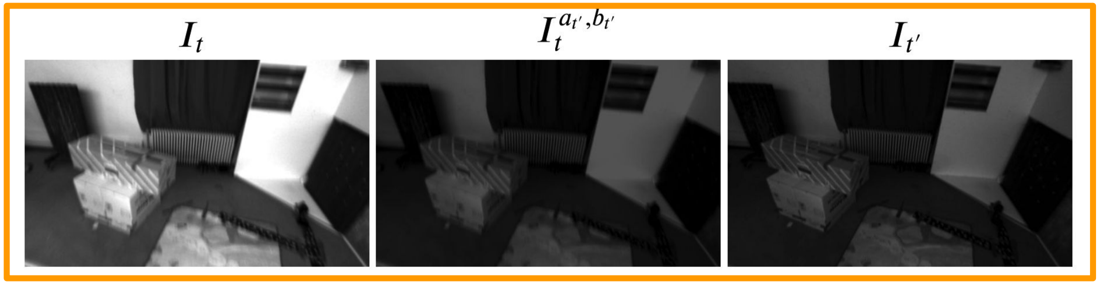


Figure 5. Examples of affine brightness transformation on EuRoC MAV.

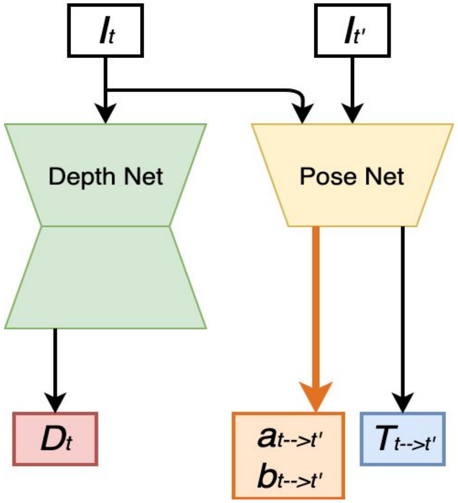


Figure 4. (a) Extended MonoDepth2 architecture. Pose Net predicts additional brightness transformation parameters.

Method Description

Self-supervised Network

Non-Lambertian surfaces, high-frequency areas and moving objects also violate the brightness constancy assumption.

➔ Can be seen as observation noise, leverage the concept of **heteroscedastic aleatoric uncertainty.**

- **Predict a posterior probability distribution for each pixel** parameterized with its mean as well as its variance $p(y|\tilde{y}, \sigma)$. No ground-truth label for σ is needed for training!

$$-\log p(y|\tilde{y},\sigma) = \frac{|y-\tilde{y}|}{\sigma} + \log\sigma + const \qquad (4)$$

- Depth network predicts higher σ for the pixel areas where the assumption may be violated.

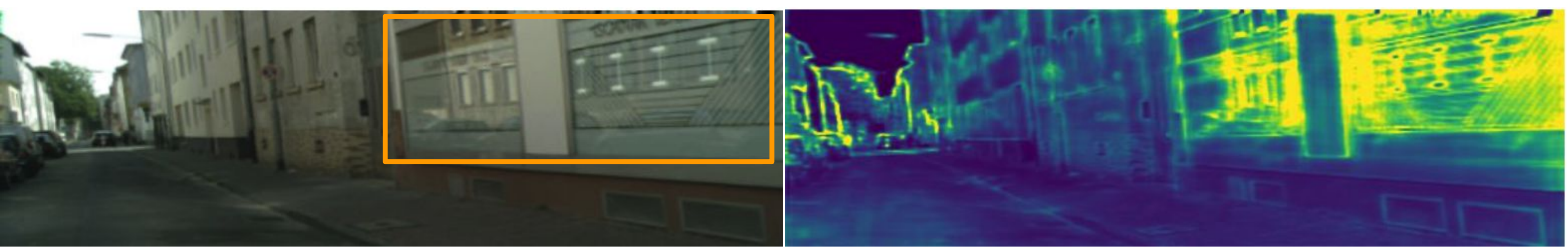


Figure 6. Uncertainty prediction results on Cityscapes with the model trained on KITTI [16].

Method Description

Self-supervised Network

➔ Minimize **photometric reprojection error** based on **photometric constancy + affine transformation + aleatoric uncertainty:**

$$L_{self} = \frac{1}{|V|} \sum_{\mathbf{p} \epsilon V} \frac{\min_{t'} r(a_{t \to t'} I_t + b_{t \to t'}, I_{t' \to t})}{\Sigma_t} + \log \Sigma_t \quad (5)$$

Total loss function is the summation of the self-supervised losses and the regularization losses on multi-scale images:

$$L_{total} = \frac{1}{s} \sum_s (L_s^{self} + \lambda L_s^{reg}) \quad (6)$$

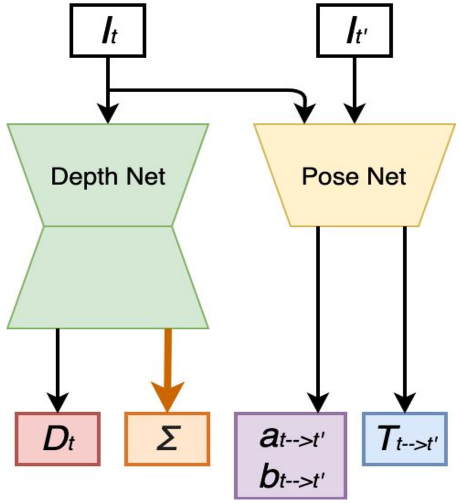


Figure 4. (b) Extended MonoDepth2 architecture. Depth Net predicts an additional uncertainty map.

Method Description

D3VO - Predicted Uncertainty Integration

D3VO aims to minimize a total photometric error E_{photo} defined as:

$$E_{photo} = \sum_{i \epsilon F} \sum_{\mathbf{p} \epsilon Pi} \sum_{j \epsilon obs(\mathbf{p})} E_{\mathbf{p}j} \quad (7)$$

$$E_{\mathbf{p}j} := \sum_{p \epsilon N_{\mathbf{p}}} w_{\mathbf{p}} \left\| (I_j[\mathbf{p}'] - b_j) - \frac{e^{a_j}}{e^{a_i}} (I_i[\mathbf{p}] - b_i) \right\|_{\gamma} \quad (8)$$

$$\mathbf{p}' = \Pi(\mathbf{T}_{\mathbf{i}}^{\mathbf{j}} \Pi^{-1}(\mathbf{p}, d_{\mathbf{p}})) \quad (9)$$

F: Set of all keyframes
P_i: Set of points hosted in keyframe i
obs(p): Set of keyframes in which point p is observable
N: Set of 8 neighboring pixels of p
a, b: affine brightness parameters jointly estimated ·
$||_{\gamma}$: Huber norm
d_p: Depth of point p
Π(·): Projection function

In DSO [1] the residual is down-weighted when the pixels are with high image gradient to compensate small independent geometric noise. In realistic scenarios there are more sources of noise!

→ Incorporate learned uncertainty to the weighting function to make it **dependent to also higher level of noise pattern:**

$$w_{\mathbf{p}} = \frac{a^2}{a^2 + \left\| \tilde{\Sigma}(p) \right\|_2^2} \quad (10)$$

Method Description

D3VO - Predicted Depth Integration

Traditional monocular VO methods [1] initialize d_p randomly.

➔ Incorporate predicted depth into the VO system:

1. Initialize the point with $d_{\mathbf{p}} = \tilde{D}_i[\mathbf{p}]$ which provides metric scale.
2. Introduce a **virtual stereo term** as in **DVSO [15]** to optimize the estimated depth d_p from VO to be consistent with the depth prediction of the Depth Net.

$$E_{photo} = \sum_{i \epsilon F} \sum_{\mathbf{p} \epsilon P_i} (\lambda E_{\mathbf{p}}^{+} + \sum_{j \epsilon obs(\mathbf{p})} E_{\mathbf{p}j}) \qquad (11)$$

$$E_{\mathbf{p}}^{+} = w_{\mathbf{p}} \left\| I_i^{+}[\mathbf{p}^{+}] - I_i[\mathbf{p}] \right\|_{\gamma} \qquad (12)$$

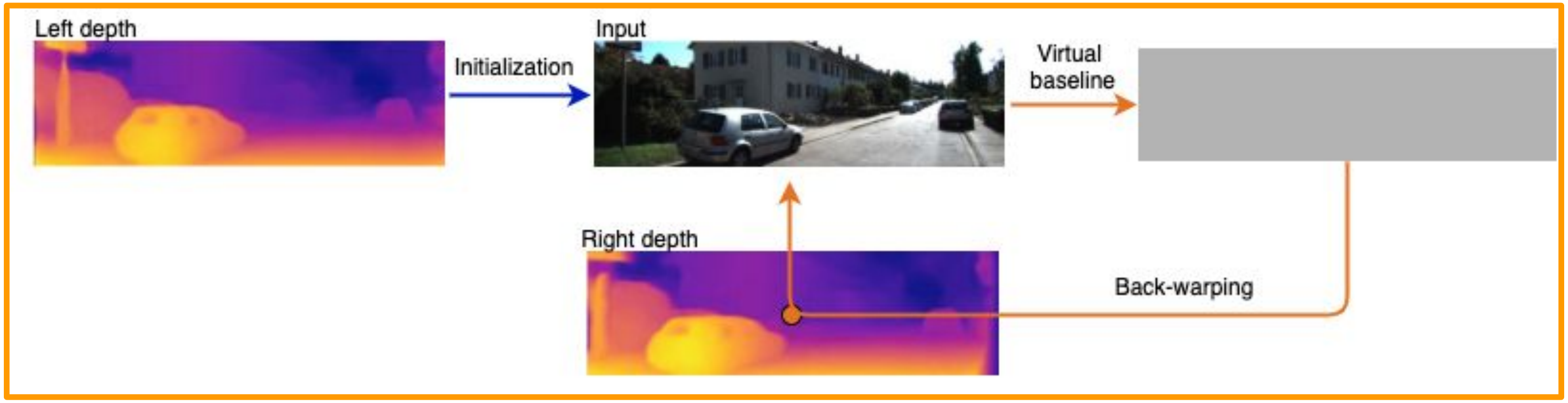


Figure 7. Virtual stereo term descriptive figure.

Method Description

D3VO - Predicted Pose Integration

Traditional direct VO approaches initialize the front-end tracking for each new frame with a constant velocity motion model.

➔ Leverage the predicted poses between consecutive frames to build a non-linear factor graph for direct image alignment.

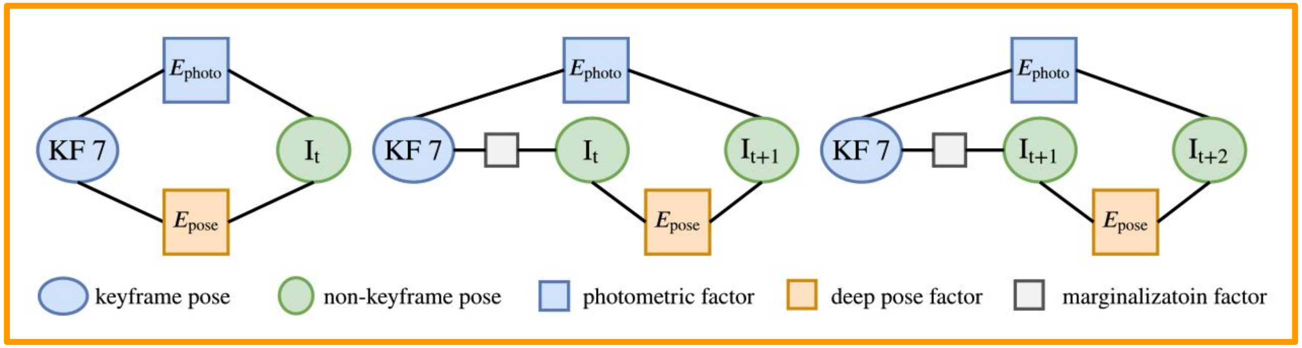


Figure 8. Visualization of the factor graph created for the front-end tracking in D3VO [16]. From left to right are the factor graph when the first, second and the third frame comes after the newest keyframe.

➔ Use the pose estimated from front-end tracking to initialize the photometric bundle adjustment back-end.

Method Description

D3VO - Predicted Pose Integration

➔ Introduce a prior for the relative keyframe pose using the predicted pose:

$$E_{pose} = \sum_{i \epsilon F-\{0\}} \log(\mathbf{\tilde{T}^{i}_{i-1}} \mathbf{T^{i-1}_{i}})^T \Sigma^{-1}_{\tilde{\varepsilon}^{i}_{i-1}} \log(\mathbf{\tilde{T}^{i}_{i-1}} \mathbf{T}^{i-1}_{i}) \quad (13)$$

- **Pose term forces the predicted pose from Pose Net and the estimated pose to be consistent.**

Total energy function:

$$E_{total} = E_{photo} + wE_{pose} \quad (14)$$

Etotal is minimized using the Gauss-Newton method.

Experiments & Results

Monocular Depth Estimation

KITTI Eigen Split

Trained on stereo sequences which gives 9,810 training quadruplets:

- 3 (left) temporal images
- 1 (right) stereo image
- 4,424 for validation

EuRoC MAV Dataset

11 sequences categorized as easy, medium and difficult considering camera motion and illumination both between stereo and temporal images.

- *Experiment 1:* Train models with the monocular setting on *MH* sequences and test on *V2_01* .
- *Experiment 2:* Use 5 sequences *MH_01*, *MH_02*, *MH_04*, *V1_01* and *V1_02* as the training set.
  - Remove static frames for training
  - 11,422 images for training and 1269 images for validation

➢ Ablation study of brightness transformation parameters and photometric uncertainty.

Experiments & Results

Monocular Depth Estimation

Approach	Train	RMSE
MonoDepth2 [4]	MS	4.750
Ours, *uncer*	MS	4.532
Ours, *ab*	MS	4.650
Ours, *full*	MS	**4.485**
[6]	DS	4.621
DVSO [15]	D*S	**4.442**
Ours	MS	4.485

Table 1. Depth evaluation results on the KITTI Eigen split. M: self-supervised monocular supervision; S: self-supervised stereo supervision; D: ground-truth depth supervision; D*: sparse auxiliary depth supervision. Upper part shows the comparison with Monodepth2 [4], lower shows the comparison with the SOTA semi-supervised methods using stereo as well as depth supervision.

Approach	RMSE
MonoDepth2 [4]	0.370
Ours, *ab*	0.339
Ours, *uncer*	0.368
Ours, *full*	**0.337**
[5]	0.971
Ours	**0.943**

Table 2. Upper part shows evaluation results of *V2_01* in EuRoC MAV, lower part shows evaluation results of *V2_01* in EuRoC MAV with the model trained with all *MH* sequences.

Experiments & Results

Monocular Depth Estimation

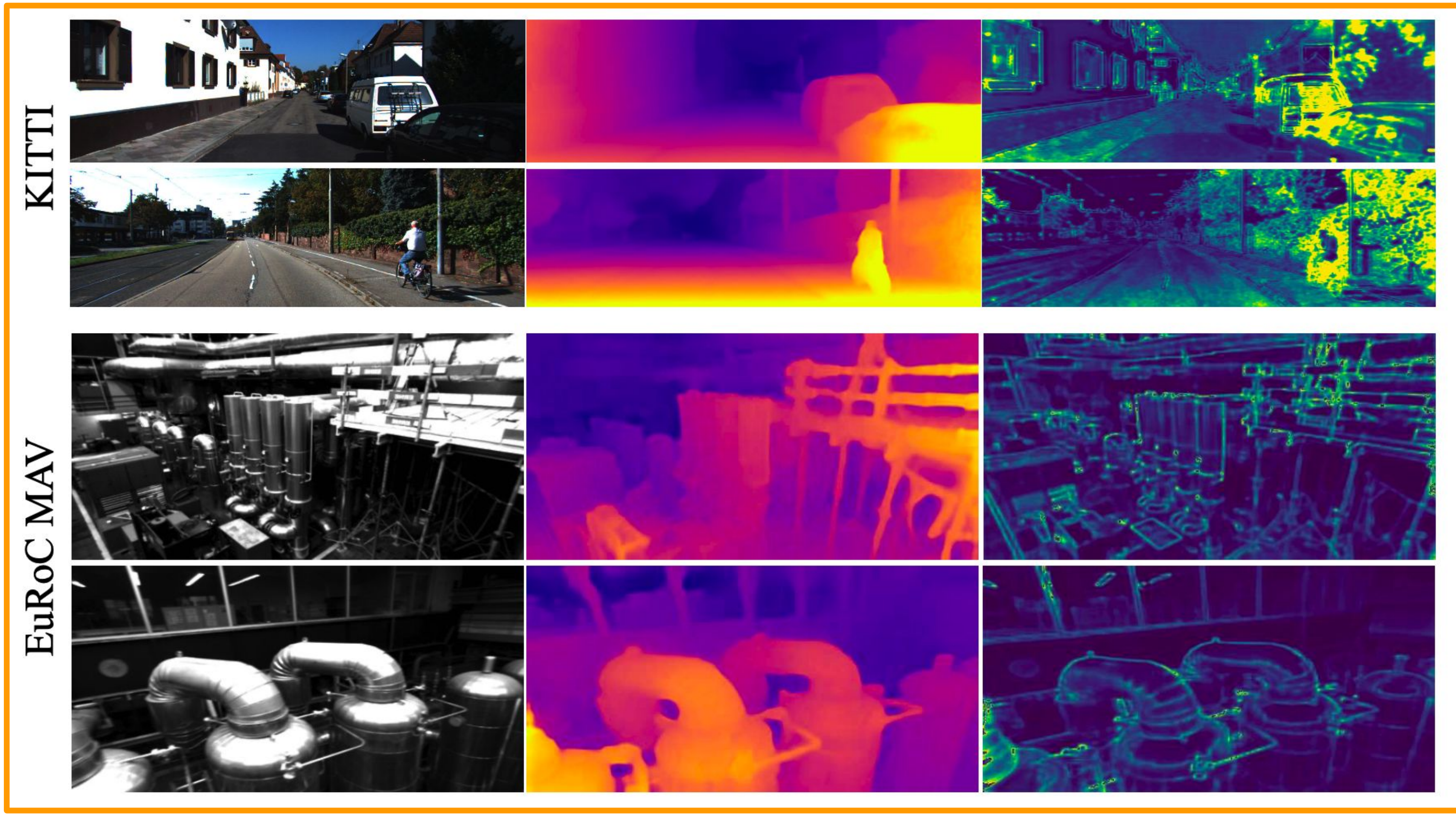


Figure 9. Qualitative results from KITTI and EuRoC MAV. The original image, the predicted depth maps and the uncertainty maps are shown from the left to the right, respectively. In particular, the network is able to predict high uncertainty on object boundaries, moving objects, highly reflecting and high frequency areas [16].

Experiments & Results

Monocular Depth Estimation

Monocular depth estimation performance on Cityscapes Dataset: Network **has the generalization capability** on both depth and uncertainty prediction. Predicts high uncertainties on reflectance, object boundaries, high-frequency areas, and moving objects.

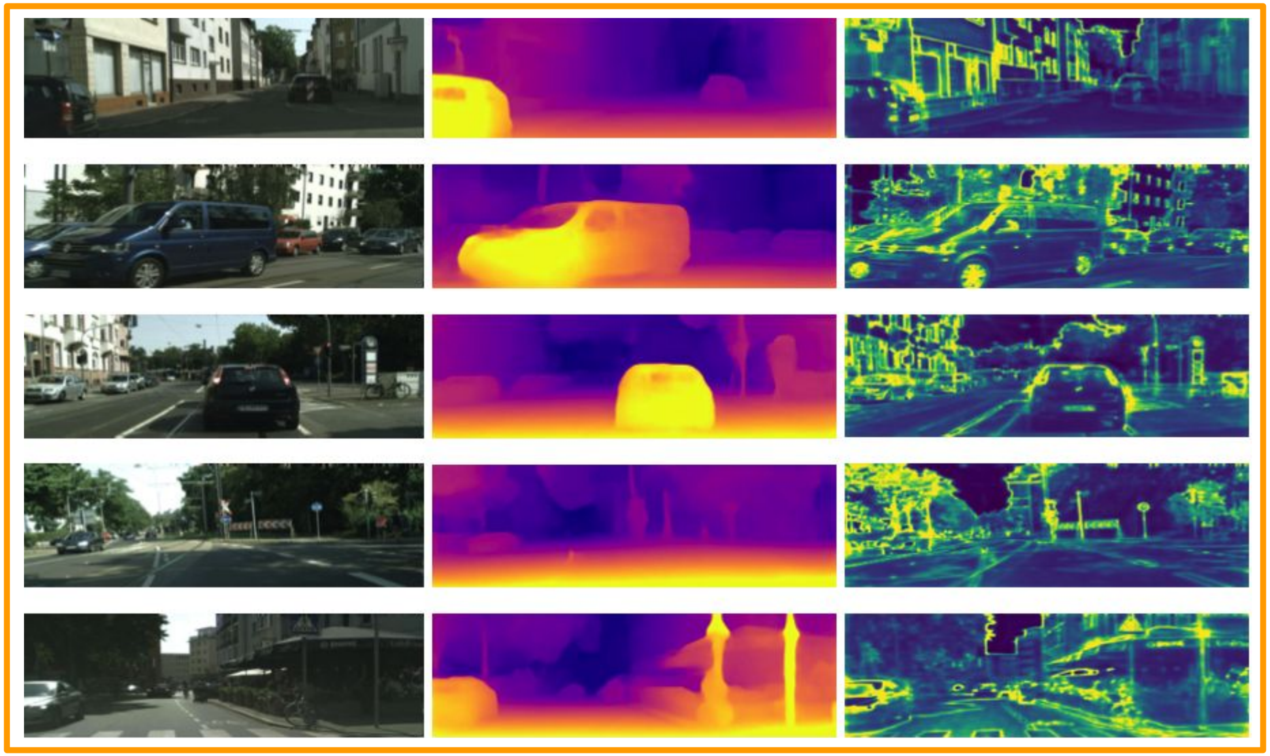


Figure 10. Results on Cityscapes with the model trained on KITTI [16].

Experiments & Results

Monocular Visual Odometry

KITTI Odometry Benchmark

11 sequences with provided ground-truth poses:

- Sequences 00, 03, 04, 05, 07 are in the training set of the Eigen split
- Use the rest of the sequences as the test set

- Evaluation metric: Relative translational error t_{rel}

EuRoC MAV Dataset

MH_03_medium, *MH_05_difficult*, *V1_03_difficult*, *V2_02_medium* and *V2_03_difficult* are used as the test set. All the other sequences are used for training.

- Evaluation metric: Root mean square (RMS) of the absolute trajectory error (ATE) after aligning the estimates with ground truth

➢ Ablation study on the integration of deep depth, pose and uncertainty.

Experiments & Results

Monocular Visual Odometry

Approach		Mean
Mono	**DSO [1]**	65.8
	ORB [9]	37.0
Stereo	**S. LSD [2]**	1.29
	ORB2 [10]	0.91
	S. DSO [14]	0.89
	Dd	0.88
	Dd + Dp	0.87
	Dd + Du	0.84
	D3VO	**0.82**

Table 3. Results of the SOTA monocular methods and SOTA stereo methods on test split of KITTI Odometry. Ablation study for the integration of deep depth (Dd), pose (Dp) as well as uncertainty (Du) is also shown.

Approach		Seq. 09	Seq. 10
End-to-end	**UnDeepVO [7]**	7.01	10.63
	Zhan et al. [18]	11.92	12.45
	SGANVO [3]	4.95	**5.89**
	Gordon et al. [5]	**2.7**	6.8
Hybrid	**CNN-SVO [8]**	10.69	4.84
	Yin et al. [17]	4.14	1.70
	Zhan et al. [19]	2.61	2.29
	DVSO [15]	0.83	0.74
	D3VO	**0.78**	**0.62**

Table 4. Comparison to other hybrid methods as well as end-to-end methods on Seq. 09 and 10 of KITTI Odometry.

Experiments & Results

Monocular Visual Odometry

Approach		M03	M05	V103	V202	V203	Mean
M	**DSO [1]**	0.18	0.11	1.42	0.12	0.56	0.48
	ORB [9]	**0.08**	0.16	1.48	1.72	0.17	0.72
M + I	**VI-ORB [11]**	0.09	**0.08**	X	**0.04**	**0.07**	0.07+X
	VI-DSO [13]	0.12	0.12	**0.10**	0.06	0.17	0.11
	End-end VO	1.80	0.88	1.00	1.24	0.78	1.14
	Dd	0.12	0.11	0.63	0.07	0.52	0.29
	Dd + Dp	0.09	0.09	0.13	0.06	0.19	0.11
	Dd + Du	**0.08**	0.09	0.55	0.08	0.47	0.25
	D3VO	**0.08**	0.09	0.11	0.05	0.19	**0.10**
S + I	**Basalt [12]**	**0.06**	0.12	**0.10**	**0.05**	-	**0.08**
	D3VO	0.08	**0.09**	0.11	0.05	-	**0.08**

Table 5. Evaluation results on EuRoC MAV. Results of DSO and ORB-SLAM as baselines are shown and D3VO is compared with other SOTA monocular VIO (M+I) and stereo VIO (S+I) methods. The best results among the monocular methods are shown as blue bold and the best among the stereo methods are shown as orange bold.

Experiments & Results

Monocular Visual Odometry

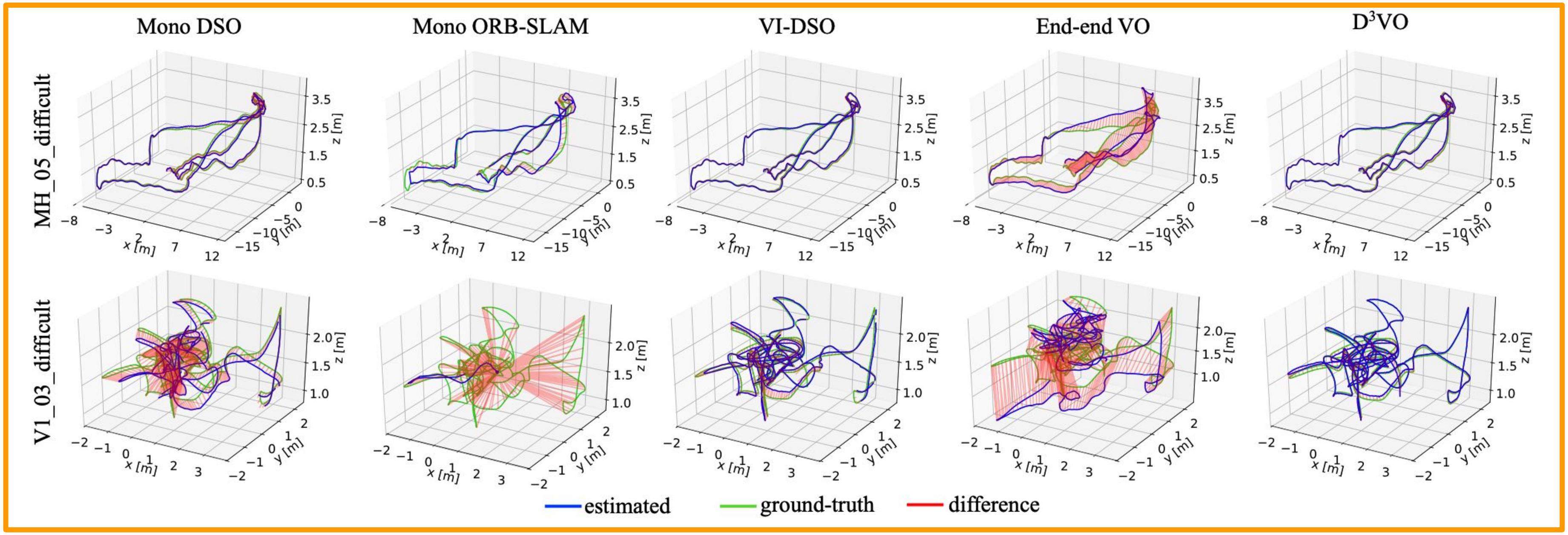
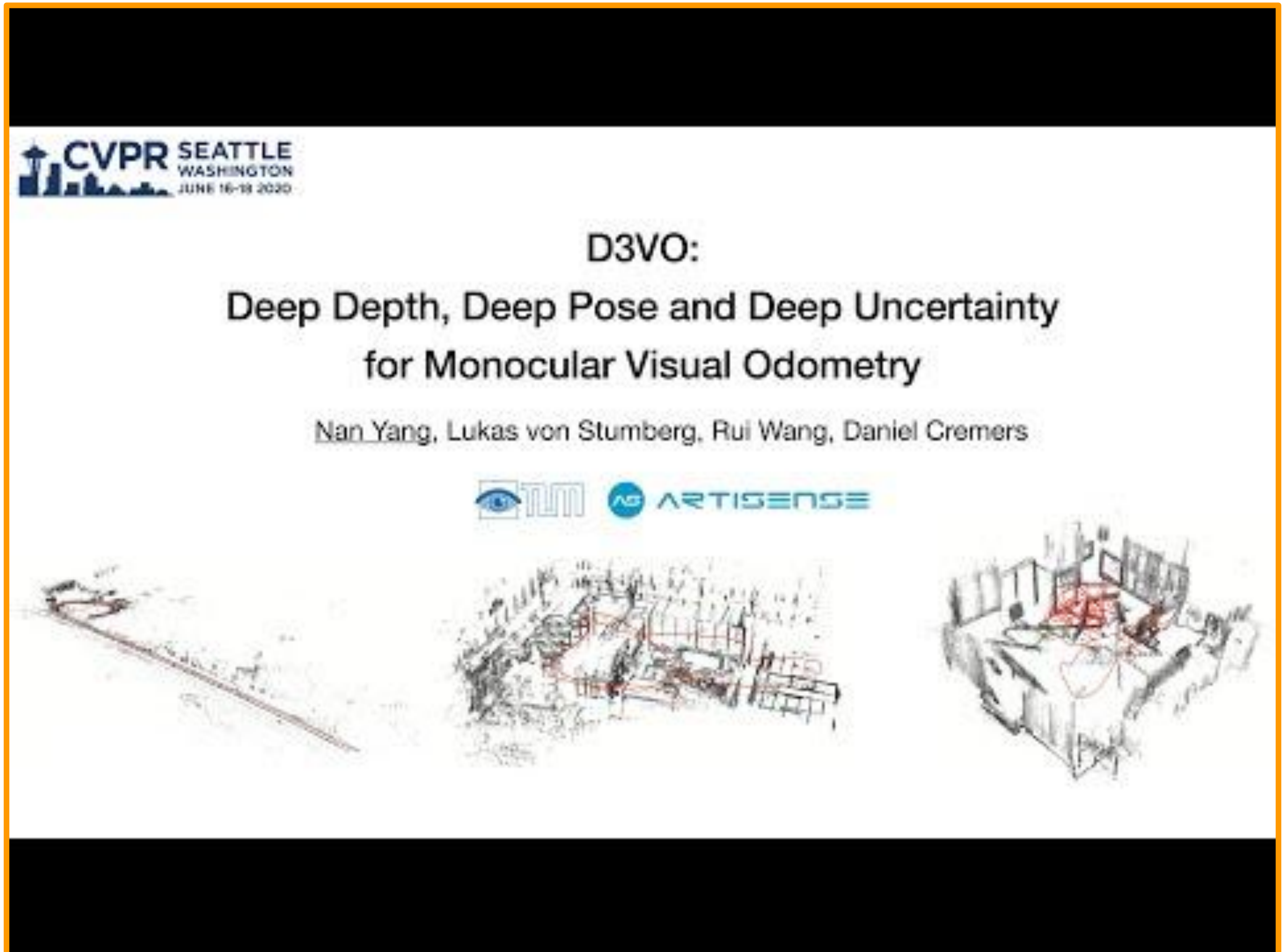


Figure 11. Qualitative comparison of the trajectories on *MH_05_difficult* and *V1_03_difficult* from EuRoC MAV [16].

Experiments & Results

Monocular Visual Odometry



D3VO CVPR presentation video [16].

Personal Comments

- Selected size of training images (512 x 216) might affect the performance of predicted poses and depths.
- ✓ Addressing brightness constancy assumption violation problem also solved most of the failure cases of MonoDepth2 [4].
- Improving the generalization capability of monocular depth estimation among very different scenarios is still a challenge!
- ✓ Comprehensive utilization of deep neural networks and clever integration of predictions.
- ✓ More consistent pose estimations obtained which reflects the lower drift of pose estimations.
- ✓ Achieves the precision of the SOTA stereo/lidar/visual-inertial odometry while using only a single camera.

Summary

- D3VO is a framework for monocular visual odometry that enhances the performance of geometric VO methods by exploiting the deep neural networks on **three levels: monocular depth, photometric uncertainty and relative camera pose.**
- A **self-supervised monocular depth estimation network is introduced which also predicts brightness transformation parameters and uncertainty map** to better address the brightness constancy assumption violation.
- The **predicted depth, uncertainty and pose are incorporated into both the front-end tracking and back-end non-linear optimization** of a direct VO pipeline.
- D3VO sets a new SOTA on KITTI Odometry and also SOTA performance on the challenging EuRoC MAV, **rivaling with leading mono-inertial and stereo-inertial methods while using only a single camera.**

Q & A

References

[1] Jakob Engel, Vladlen Koltun, and Daniel Cremers. Direct sparse odometry. *IEEE transactions on pattern analysis and machine intelligence*, 2017.

[2] Jakob Engel, Jorg Stuckler, and Daniel Cremers. Large-scale direct SLAM with stereo cameras. *In Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*, pages 1935–1942. IEEE, 2015.

[3] Tuo Feng and Dongbing Gu. SGANVO: Unsupervised deep visual odometry and depth estimation with stacked generative adversarial networks. I*EEE Robotics and Automation Letters*, 4(4):4431–4437, 2019.

[4] Clement Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth estimation. *In The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[5] Ariel Gordon, Hanhan Li, Rico Jonschkowski, and Anelia Angelova. Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras. *In The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[6] Yevhen Kuznietsov, Jorg Stuckler, and Bastian Leibe. Semi- supervised deep learning for monocular depth map predic tion. arXiv preprint arXiv:1702.02706, 2017.

[7] Ruihao Li, Sen Wang, Zhiqiang Long, and Dongbing Gu. UnDeepVO: Monocular visual odometry through unsupervised deep learning. arXiv preprint arXiv:1709.06841, 2017.

[8] Shing Yan Loo, Ali Jahani Amiri, Syamsiah Mashohor, Sai Hong Tang, and Hong Zhang. CNN-SVO: Improving the mapping in semi-direct visual odometry using single-image depth prediction. *In 2019 International Conference on Robotics and Automation (ICRA)*, pages 5218–5223. IEEE, 2019.

[9] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. ORB-SLAM: a versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics*, 31(5):1147– 1163, 2015.

[10] Raul Mur-Artal and Juan D Tardo ́s. ORB-SLAM2: An open-source slam system for monocular, stereo, and rgb-d cameras. I*EEE Transactions on Robotics*, 33(5):1255–1262, 2017.

References

[11] Rau´l Mur-Artal and Juan D Tardo´s. Visual-inertial monocular SLAM with map reuse. *IEEE Robotics and Automation Letters*, 2(2):796–803, 2017.

[12] Vladyslav Usenko, Nikolaus Demmel, David Schubert, Jorg Stuckler, and Daniel Cremers. Visual-inertial mapping with non-linear factor recovery. arXiv preprint arXiv:1904.06504, 2019.

[13] Lukas Von Stumberg, Vladyslav Usenko, and Daniel Cremers. Direct sparse visual-inertial odometry using dynamic marginalization. I*n 2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2510–2517. IEEE, 2018.

[14] R. Wang, M. Schworer, and D. Cremers. Stereo DSO: Large-scale direct sparse visual odometry with stereo cameras. *In International Conference on Computer Vision (ICCV)*, Venice, Italy, October 2017.

[15] Nan Yang, Rui Wang, Jorg Stuckler, and Daniel Cremers. Deep virtual stereo odometry: Leveraging deep depth prediction for monocular direct sparse odometry. *In Proceedings of the European Conference on Computer Vision (ECCV)*, pages 817–833, 2018.

[16] Nan Yang, Lukas von Stumberg, Rui Wang and Daniel Cremers. D3VO: Deep Depth, Deep Pose and Deep Uncertainty for Monocular Visual Odometry. *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[17] Xiaochuan Yin, Xiangwei Wang, Xiaoguo Du, and Qijun Chen. Scale recovery for monocular visual odometry using depth estimated with deep convolutional neural fields. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5870–5878, 2017.

[18] H. Zhan, R. Garg, C. S. Weerasekera, K. Li, H. Agarwal, and I. M. Reid. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. *In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 340–349, June 2018.

[19] Huangying Zhan, Chamara Saroj Weerasekera, Jiawang Bian, and Ian Reid. Visual odometry revisited: What should be learnt? arXiv preprint arXiv:1909.09803, 2019.