Learning Structure from Motion from Motion

Clément Pinard[a,b] Laure Chevalley[a] Antoine Manzanera[b] David Filliat[b]

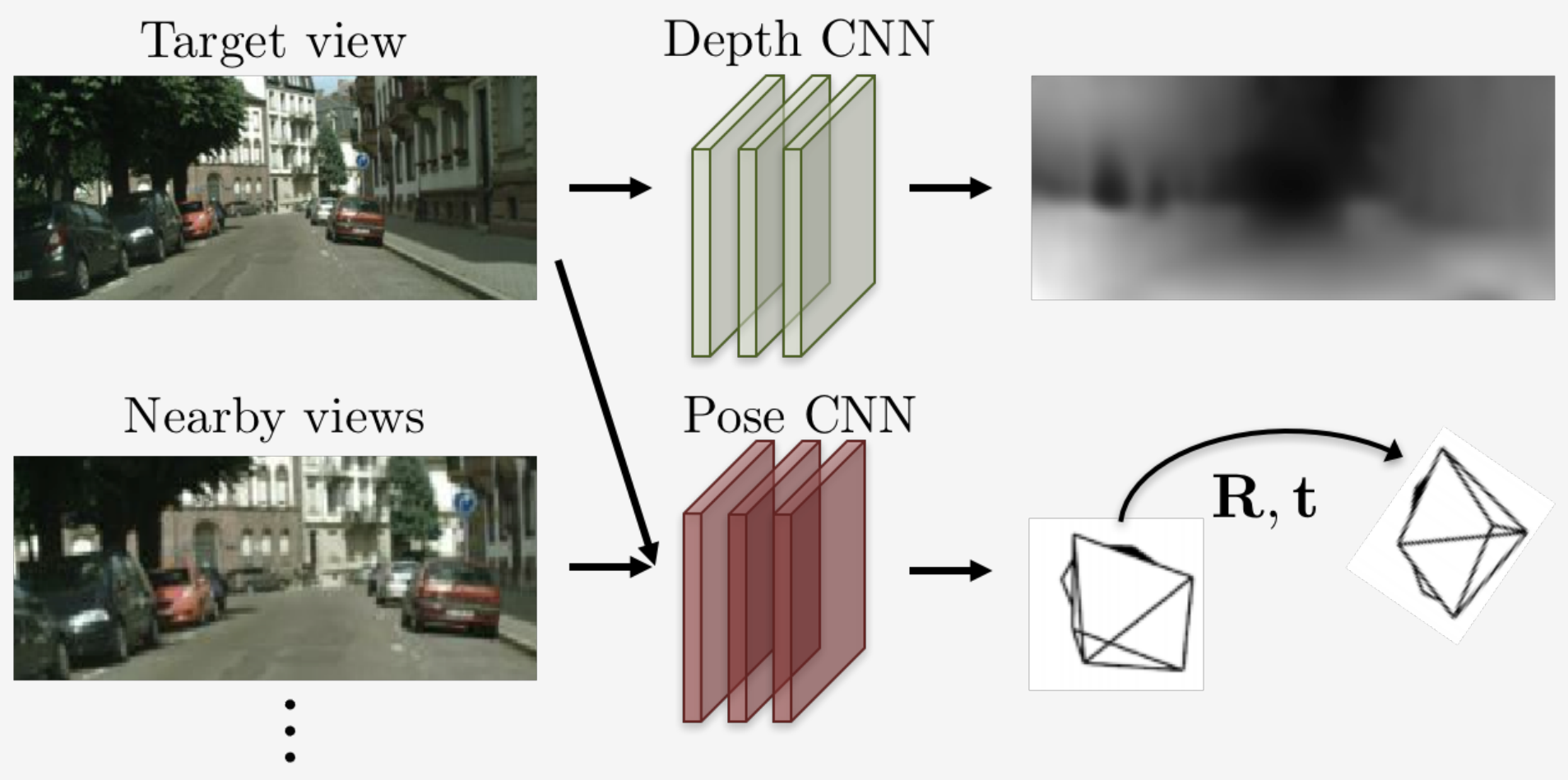
[a]Parrot, Paris, France
(clement.pinard, laure.chevalley)@parrot.com

[b]U2IS, ENSTA ParisTech, Université Paris-Saclay, Palaiseau, France
(clement.pinard, antoine.manzanera,
david.filliat)@ensta-paristech.fr

GMDL@ECCV2018

Context and Motivations

Learning Structure from Motion [4]



Pros :

- Only needs videos with camera intrinsics
- Depth from a **single Image**

Cons :

- Depth from a single Image is **not robust** enough
- Scaling factor is **not known**

A new way of measuring Depth Accuracy

- Current validation from [1] makes **median of groundtruth depth map** available!
- Instead, use **speed estimation** for solving the **scaling factor**.
- Closer to navigation usecase where movement estimation is usually done by other sensors than camera, *e.g.* IMU or GPS.

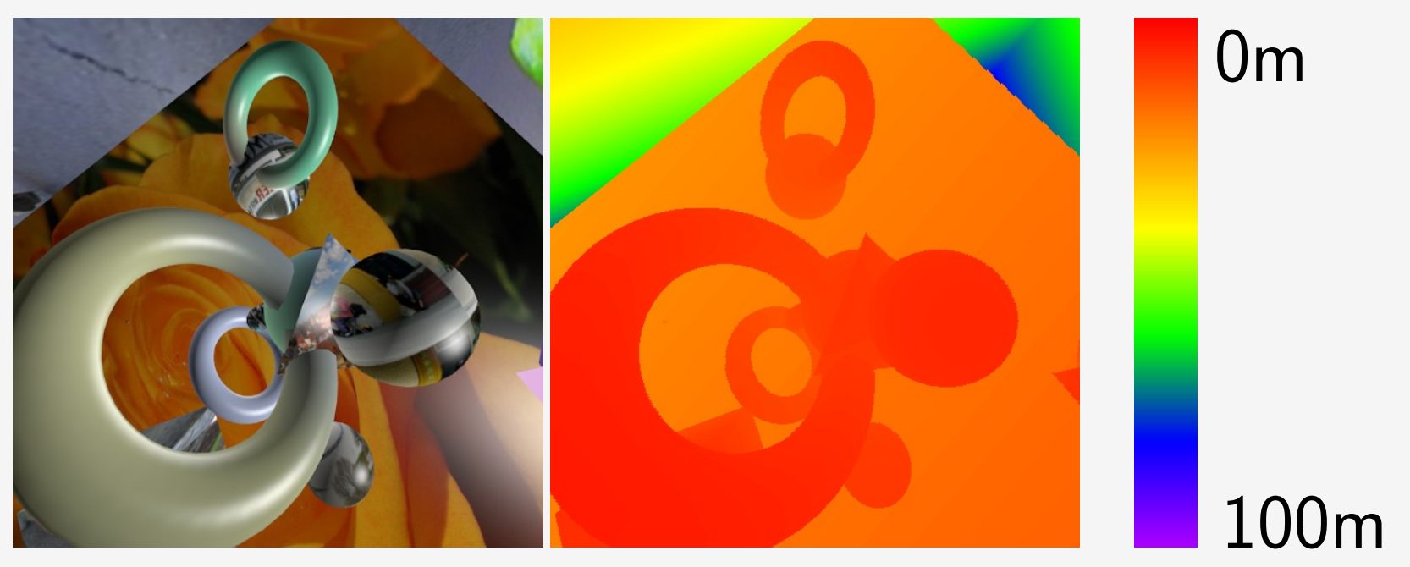
	prior work [1, 4]	Our proposition
Predictions	Depth $\widehat{D}$	Depth $\widehat{D}$, Velocity $\widehat{V}$
Ground Truth	Depth D_{GT}	Depth D_{GT}, Velocity V_{GT}
Measure	$m = \delta\left(D_{GT}, \widehat{D} \times \frac{Me(D_{GT})}{Me(\widehat{D})}\right)$	$m = \delta\left(D_{GT}, \widehat{D} \times \frac{\lvert V_{GT}\rvert}{\lvert\widehat{V}\rvert}\right)$

$Me()$ is the median operator, and δ is a validation measure (e.g. L1 distance)

Training Datasets

Still Box [3]

- Aims at having depth **independent to context**
- **Rigid** scenes
- **Random** orientation and velocity direction
- **Random** textures and shapes



KITTI [2]

- **Realistic**
- **Not Rigid** scenes
- **Always** the same orientation and position w.r.t ground
- Sparse ground truth and not available above horizon

Frame reprojection

$\widehat{I_j}$ is constructed from I_t using the equation:

$$\forall i \in [\![0, N[\![\,, p_t^i = K\widehat{T}_{t\rightarrow i}\left(\widehat{D}(p_t)K^{-1}p_t\right) \tag{1}$$

when only considering rotation and translation :

$$p_t^r = KR_rK^{-1}p_t \tag{2}$$

Loss Functions

$$\mathrm{SSIM}(I_t, I_i) = \frac{(2\mu_{I_t}\mu_{I_i} + C_1) + (2\sigma_{I_tI_i} + C_2)}{(\mu_{I_t}^2 + \mu_{I_i}^2 + C_1)(\sigma_{I_t}^2 + \sigma_{I_i}^2 + C_2)} \tag{3}$$

μ_I is local mean of image I and σ_I is local std of I, obtained with Gaussian and Laplacian 3×3 filters. s is the downsampling factor.

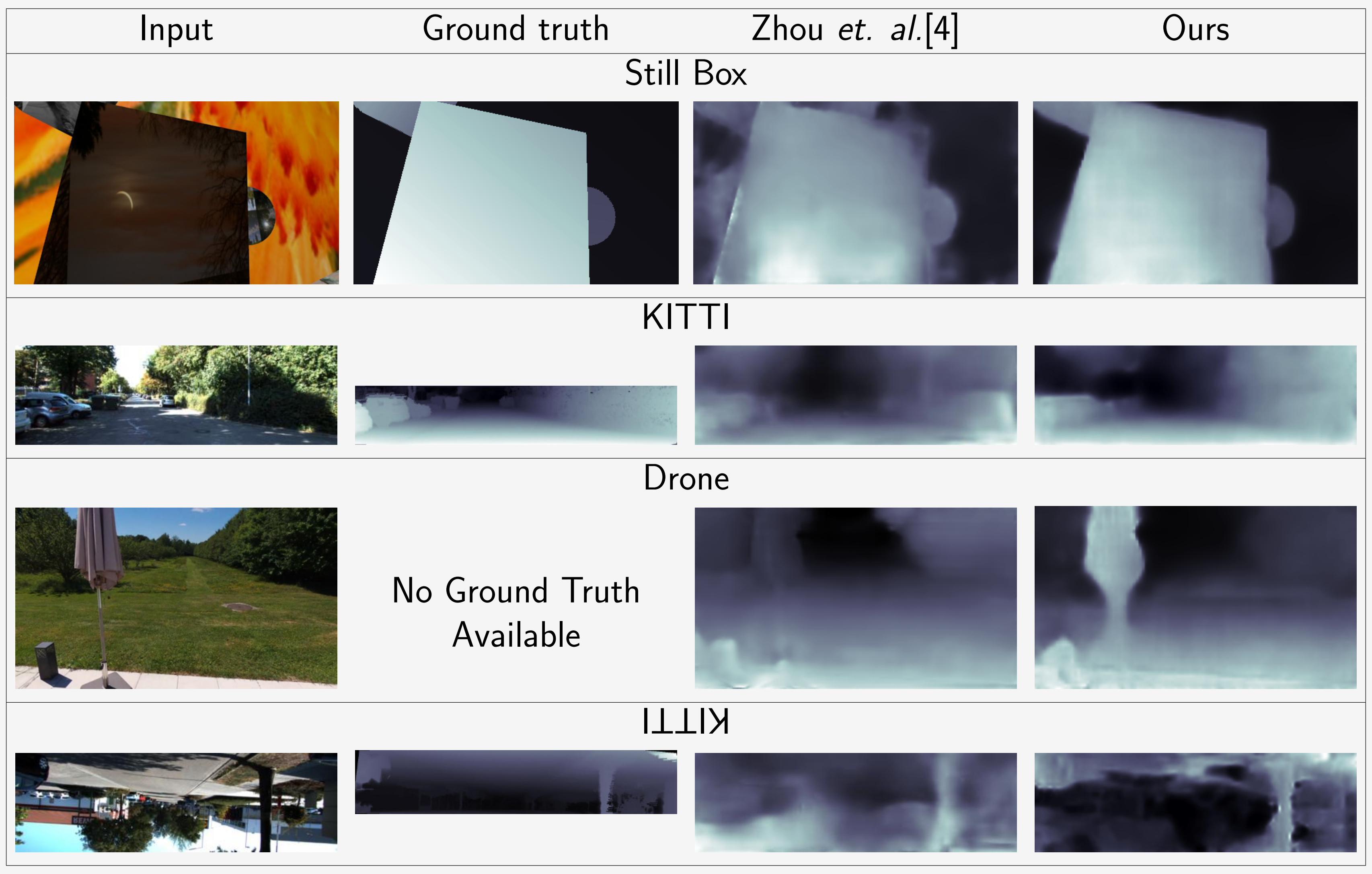
$$\mathcal{L}_p = \sum_i \|\widehat{I_i} - I_t\|_1 - \alpha\mathrm{SSIM}(\widehat{I_i}, I_t) \tag{4}$$

$$\mathcal{L}_g = \left\|\frac{|\Delta\widehat{D}|}{\|\nabla I_t\|}\right\|_1 \times \frac{1}{\|\zeta\|_1} \tag{5}$$

$$\mathcal{L} = \sum_s \frac{1}{2^s}\left(\mathcal{L}_p^s + \lambda\mathcal{L}_g^s\right) \tag{6}$$

For our experiments we used $C_1 = 0.01^2$, $C_2 = 0.03^2$, $\alpha = 0.1$ and $\lambda = 3$

Qualitative results



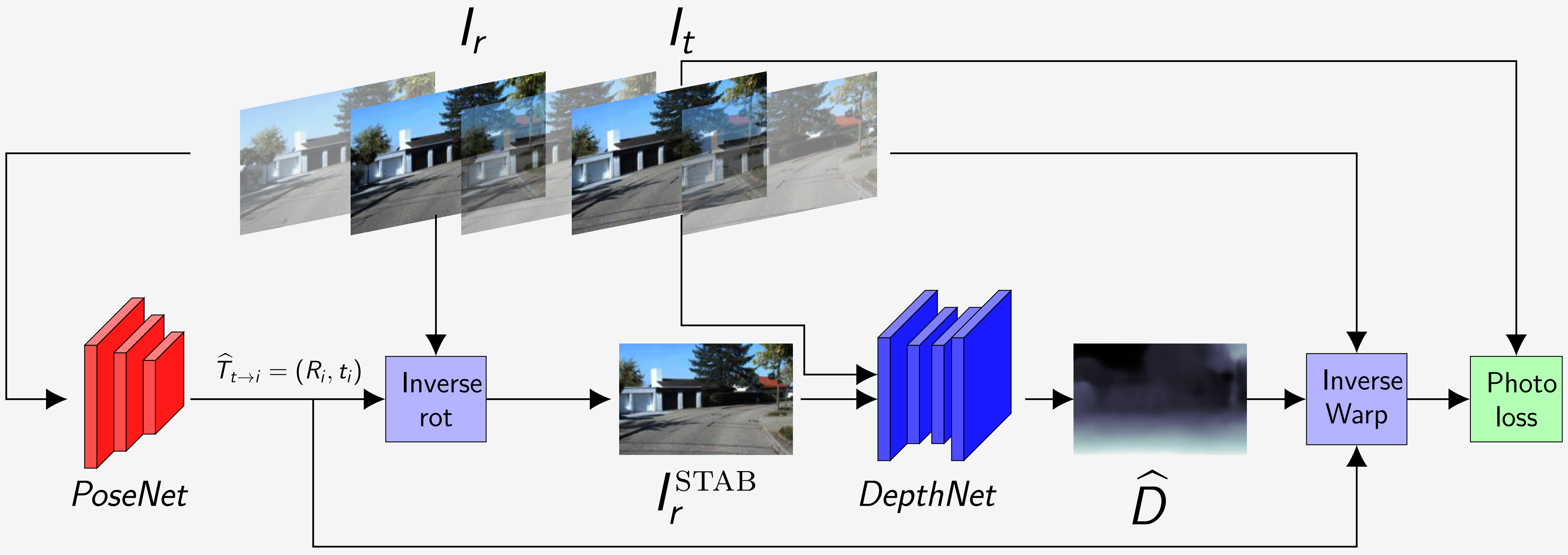
- For Drone results, a finetuning on a 15minutes video is applied.
- For ꓘITTI (KITTI upside down), no finetuning is done.

Quantitative results

Method	training set	scale factor	testing set	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Zhou *et. al.*[4]	K	GT	K	0.183	1.595	6.709	0.270	0.734	0.902	0.959
Zhou *et. al.*	K	V	K	**0.279**	**2.706**	**7.296**	**0.356**	0.582	0.808	0.898
Ours	K	V	K	0.312	5.030	8.498	0.409	0.592	0.796	0.882
Ours	S → K	V	K	0.294	3.992	7.573	0.376	**0.609**	**0.834**	**0.909**
Ours supervised [3]	S	V	S	0.212	2.064	7.067	0.296	0.709	0.881	0.946
Zhou *et. al.*	S	V	S	0.811	11.996	17.274	0.693	0.347	0.573	0.717
Ours	S	V	S	**0.468**	**10.925**	**15.756**	**0.544**	**0.452**	**0.677**	**0.804**
Constant Plane	-	GT	ꓘ	0.457	4.852	12.085	0.600	0.296	0.549	0.752
Zhou *et. al.*	K	GT	ꓘ	0.593	7.541	12.994	0.734	0.222	0.434	0.626
Zhou *et. al.*	K	P	ꓘ	1.588	62.107	21.142	0.958	0.169	0.326	0.474
Ours	S → K	P	ꓘ	**0.648**	**15.391**	**12.432**	**0.624**	**0.382**	**0.617**	**0.761**

Network and Training Specification

- **PoseNet** is the same network as in [3]
- Depth CNN now is feeded **2 images** instead of 1
- second frame is Stabilized beforehand using rotation prediction from **PoseNet**
- translations are normalized so that $\|t_r^{\mathrm{NORM}}\| = T_0$, T_0 is **constant** throughout the whole training.
- Velocity $\widehat{V}$ will then be assumed to be $T_0 \times \mathrm{FPS}$ during testing.



$$\forall i,\, t_i^{\mathrm{NORM}} = t_i\frac{T_0}{\epsilon + \|t_r\|}$$

Conclusion

- Depth from context is suited for KITTI[2], but we showed two datasets on which it performed poorly.
- Current measures don't account for depth scale determination, which makes the problem **too easy compared to a real usecase**.
- Using multiple frames for depth allows much greater **robustness** to unseen scenes or orientations.
- Training code **available on github**!

References

perso.ensta.fr/~pinard/
unsupervised-depthnet

[1] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*, pages 2366–2374, 2014.

[2] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

[3] C. Pinard, L. Chevalley, A. Manzanera, and D. Filliat. End-to-end depth from motion with stabilized monocular videos. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, IV-2/W3:67–74, 2017.

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