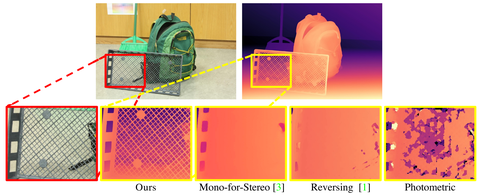
NeRF-Supervised Deep Stereo

Fabio Tosi[1], Alessio Tonioni[2], Daniele De Gregorio[3], Matteo Poggi[1]
[1]University of Bologna, [2]Google Switzerland, [3]Eyecan.ai
{fabio.tosi5, m.poggi}@unibo.it, alessiot@google.com, daniele.degregorio@eyecan.ai

Project Page

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Introduction



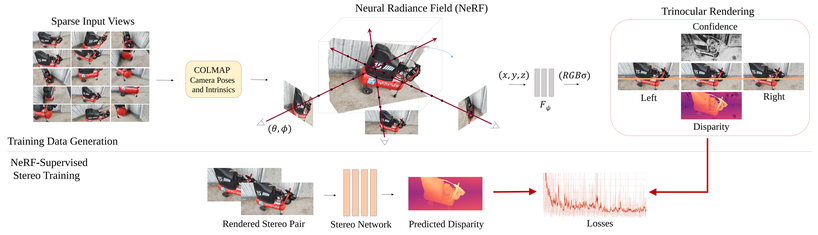
Problems:

- Depth-annotated data for stereo is **hard to source**
- Self-supervised learning paradigms are ineffective at dealing with **challenges for the stereo setting** (e.g. occlusions)
- Training on synthetic images cause **domain-shift** issues when testing on real data.
- Moreover, synthetic assets are **seldom open source** and necessitate additional **human labor**.

Key Contributions:

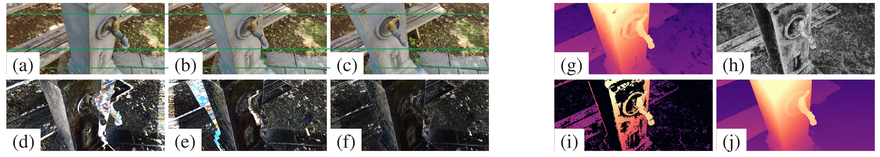
- A new method for collecting stereo training data using **neural rendering** and user-collected image sequences.
- A **NeRF-Supervised** training protocol on rendered image triplets and depth, to deal with occlusions and fine details.
- An **open-source stereo dataset** for training any state-of-the-art model according to such protocol.
- State-of-the art, **zero-shot generalization** results on stereo datasets, without using any ground-truth or real stereo pair.

Training Data Generation



- **Collection and pre-processing.** We acquire image sequences and estimate intrinsic/extrinsic parameters using COLMAP.
- **NeRF-Training.** Train an independent NeRF for each scene, using Instant-NGP as the NeRF engine.
- **Stereo Pairs Rendering.** We render binocular stereo pairs with disparity and uncertainty maps to train deep stereo networks, and also generate a perfectly rectified **stereo triplet** by rendering a third image to the left of the reference frame.

NeRF-Supervised Stereo Regime



- **Triplet Photometric Loss**. We measure the pair-wise photometric difference between warped images – from both left (a) and right (c) – and I_c (b) by means of SSIM and L1 terms. These alone cannot deal with occlusions – (d) and (e) – while the pixel-wise minimum between the two (f) can provide consistent supervision on the entire image.

$$\mathcal{L}_{3\rho}(\hat{I_l}^c, I_c, \hat{I_r}^c) = \min\Big(\mathcal{L}_\rho(\hat{I_l}^c, I_c), \mathcal{L}_\rho(I_c, \hat{I_r}^c)\Big) \quad \text{with} \quad \mathcal{L}_\rho(I_c, \hat{I}_c^x) = \beta \cdot \frac{1 - \text{SSIM}(I_c, \hat{I}_c^x)}{2} + (1-\beta) \cdot |I_c - \hat{I}_c^x| \tag{1}$$

- **Rendered Disparity Loss**. We exploit an additional loss between the predictions and the rendered disparities by NeRF (g). A filtering mechanism based on the rendered uncertainty AO (h) is employed to retain only the most reliable pixels (i).

$$\mathcal{L}_{disp} = |d_c - \hat{d}_c| \qquad \text{AO} = \sum_{i=1}^{N} T_i \alpha_i, \qquad \alpha_i = 1 - \exp(-\sigma_i \hat{\delta}_i) \tag{2}$$

- **NeRF-Supervised Loss.** Our final loss combines the two terms introduced so far, according to a threshold *th* on AO

$$\mathcal{L}_{NS} = \gamma_{disp} \cdot \eta_{disp} \cdot \mathcal{L}_{disp} + \mu \cdot \gamma_{3\rho} \cdot (1 - \eta_{disp}) \cdot \mathcal{L}_{3\rho} \qquad \eta_{disp} = \begin{cases} 0 & \text{if AO} < th \\ \text{AO} & \text{otherwise} \end{cases} \tag{3}$$

On a stereo pair (k) from our dataset, models trained with pair-wise photometric loss suffers at occlusions (m), in contrast to those trained with triplet loss (n). Our full loss yields the best disparity map (o), better than those rendered by NeRF itself (l).

Experimental Results

Comparison with Mono-from-Stereo [3]

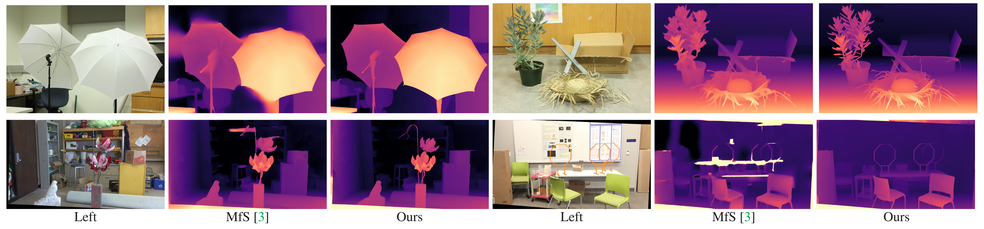
Model	Stereo Network	Dataset	# Images	Pre-Train.	KITTI-12 (> 3px)	Midd-A F (> 2px)	Midd-A H (> 2px)	Midd-A Q (> 2px)	Midd-21 (> 2px)
MfS [3] + MidAs	PSMNet	MfS	535K	~2M	4.70	25.61	17.98	14.78	27.55
MfS [3] + MidAs	PSMNet	Ours	65K	~2M	5.02	28.72	20.66	16.92	26.40
NS (**Ours**)	PSMNet	Ours	65K	0	**4.07**	**19.84**	**13.66**	**9.15**	**19.08**
MfS [3] + MidAs	CFNet	MfS	535K	~2M	**4.47**	22.00	16.69	14.32	23.44
MfS [3] + MidAs	CFNet	Ours	65K	~2M	4.90	24.20	19.11	16.20	23.76
NS (**Ours**)	CFNet	Ours	65K	0	4.64	**17.55**	**12.31**	**11.13**	**19.73**
MfS [3] + MidAs	RAFT-Stereo	MfS	535K	~2M	4.45	19.79	12.67	9.63	22.26
MfS [3] + MidAs	RAFT-Stereo	Ours	65K	~2M	4.67	24.61	17.25	14.05	24.18
NS (**Ours**)	RAFT-Stereo	Ours	65K	0	**4.02**	**13.12**	**6.91**	**7.18**	**12.87**

Comparison with state-of-the-art models trained on SceneFlow [2] – Zero-shot generalization

	Method		KITTI-15 (> 3px) All	KITTI-15 (> 3px) Noc	Midd-T F (> 2px) All	Midd-T F (> 2px) Noc	Midd-T H (> 2px) All	Midd-T H (> 2px) Noc	Midd-T Q (> 2px) All	Midd-T Q (> 2px) Noc	ETH3D (> 1px) All	ETH3D (> 1px) Noc
	Training Set		SceneFlow with GT									
(A)	GANet		10.46	10.15	45.36	40.80	26.75	21.8	15.52	11.49	8.68	7.75
(A)	DSMNet		5.50	5.19	29.95	24.79	16.88	12.03	13.75	9.44	12.52	11.62
(A)	CFNet		6.01	5.94	29.12	24.15	20.11	15.84	13.77	10.32	5.77	5.32
(A)	MS-GCNet	*	6.21	-	-	-	-	18.52	-	-	-	8.84
(A)	RAFT-Stereo	*	5.74	-	18.33	-	12.59	-	9.36	-	3.28	-
(A)	RAFT-Stereo	‡	5.45	5.21	**18.20**	**14.19**	**11.19**	**8.09**	**9.31**	**6.56**	**2.59**	**2.24**
(A)	SGM + NDR		5.41	5.12	27.27	21.09	17.70	13.51	11.75	7.93	**5.20**	**4.78**
(A)	STTR		8.31	6.73	38.10	30.74	26.39	18.17	15.91	8.51	20.49	19.06
(A)	CEST		7.61	6.13	27.44	19.53	19.89	11.82	14.71	7.56	10.99	9.78
(A)	FC-GANet	*	5.3x	-	-	-	-	10.2x	-	7.8x	5.8x	-
(A)	ITSA-GWCNet		5.60	5.39	29.46	25.18	19.38	15.95	14.36	10.76	7.43	7.12
(A)	ITSA-CFNet		**4.96**	**4.76**	26.38	21.41	18.01	14.00	13.32	9.73	5.40	5.14
(A)	CREStereo	‡	5.79	5.40	34.78	30.52	17.57	13.87	12.88	8.85	8.98	8.14
(B)	PSMNet	‡	7.86	7.40	33.69	28.35	21.69	16.92	17.24	12.37	23.19	22.12
(B)	MS-PSMNet	*	7.76	-	-	-	-	19.81	-	-	-	16.84
(B)	Graft-PSMNet		**5.34**	**5.02**	**25.46**	**19.28**	**17.81**	**13.46**	**14.18**	**9.21**	11.42	10.69
(B)	FC-PSMNet	*	5.8x	-	-	-	-	15.1x	-	9.3x	**9.5x**	-
(B)	ITSA-PSMNet		6.00	5.73	32.09	27.46	20.83	17.14	14.68	11.05	10.34	**9.77**
	Training Set		Real-world data without GT									
(C)	Reversing[1]-PSMNet		[4.09]	[3.88]	38.23	30.00	26.45	20.91	20.55	15.08	9.00	8.23
(C)	MfS[3]-PSMNet		**5.18**	**4.91**	26.42	21.38	17.56	13.45	12.07	9.09	**8.17**	**7.44**
(D)	NS-PSMNet (**Ours**)		**5.05**	**4.80**	**20.60**	**15.83**	**12.91**	**9.07**	**11.03**	**7.15**	11.69	11.00
(D)	NS-RAFT-Stereo (**Ours**)		5.41	5.23	**16.45**	**12.08**	**9.67**	**6.42**	**8.05**	**4.82**	**2.94**	**2.23**

(A) Any deep networks not in (B-D) (B) PSMNet variants trained with synthetic ground-truth (C) PSMNet models trained without ground-truth (D) Our best model
* numbers taken from original papers ‡ networks retrained by ourselves

Qualitative Results



References

[1] Aleotti et al. Reversing the cycle: self-supervised deep stereo through enhanced monocular distillation. In *ECCV*, 2020.
[2] Nikolaus Mayer et al. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *CVPR)*, 2016.
[3] Watson et al. Learning stereo from single images. In *ECCV*, 2020.

Links

Paper Supplementary Video