TripleDNet: Exploring Depth Estimation with Self-Supervised Representation Learning

Ufuk Umut Senturk, Arif Akar, Nazli Ikizler-Cinbis

HACETTEPE UNIVERSITY · aselsan

BMVC 2022

Problem Definition

- Models relying on SfM fail disastrously and mask out causing training disrupted.

★ Assumptions (constant illumination, static world) are not met in the real world

(a) Image x_t (b) Projected image $x_{t-1\to t}$ (c) Photometric error (3)

★ Masking out stationary, occluded, or dynamic pixels
- Important signals could be lost

(d) Segmentation m_t (e) Projected segmentation (f) DC object mask μ_t

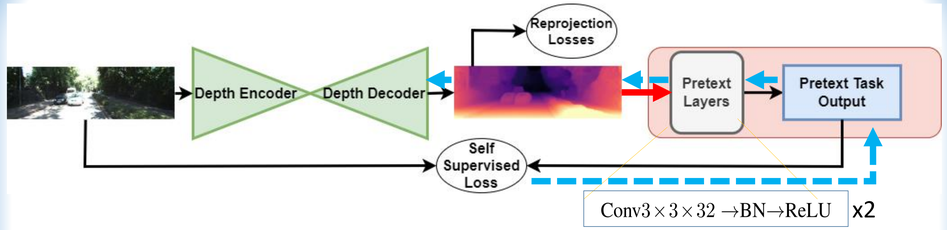
Marvin Klingner, Jan-Aike Termöhlen, Jonas Mikolajczyk, and Tim Fingscheidt. Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance. ECCV, 2020.

- **Our solution:** We propose **TripleDNet** (**Disentangled Distilled Depth Network**), a multi-objective, distillation-based framework for purely SS depth estimation.

Contributions

- Further objectives are added to SfM-based estimation to constrain the solution space and to allow feature space disentanglement.
- **Distillation** and **disentanglement** mechanisms based **on joint learning of novel self-supervised pretext tasks and monocular depth estimation**.
- First work to introduce and **evaluate self-supervised IRL to self-supervised depth estimation**.
- Experimental results on two benchmark datasets show that the proposed approach is able to **achieve state-of-the-art performance** in monocular depth estimation in a **fully self-supervised fashion**.

Pretext Task Distillation



★ Depth-to-Grayscale(**D2G**)

$$\mathcal{L}_{d2g}(x) = \sqrt{(PL(D(x)) - GS(x))^2 + \varepsilon^2}$$

★ Depth and Grayscale-to-Color(**DG2C**)

$$\mathcal{L}_{dg2c}(x) = \sqrt{(PL(D(x) \oplus GS(x)) - AB(x))^2 + \varepsilon^2}$$

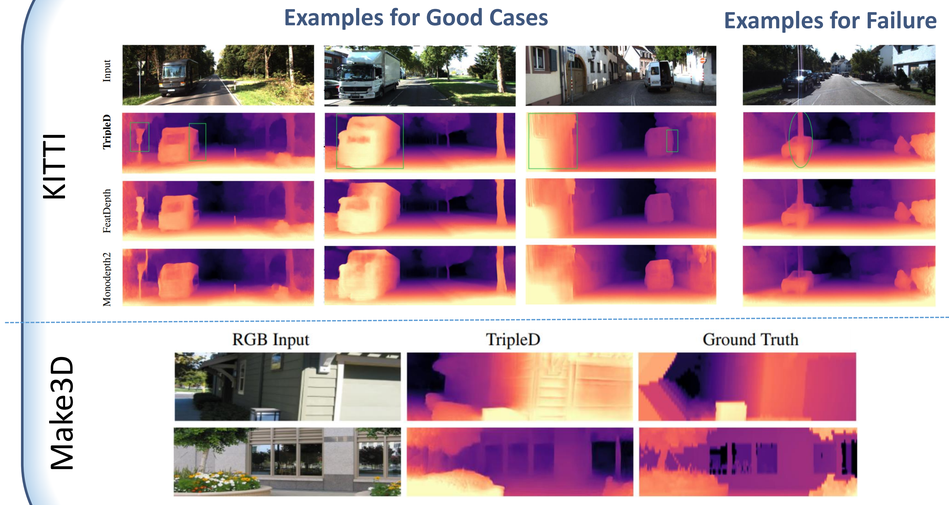
★ Masked D2G(**MD2G**)

$$\mathcal{L}_{md2g}(x) = \hat{M} \odot \sqrt{(PL((1-\hat{M}) \odot D(x)) - GS(x))^2 + \varepsilon^2}$$

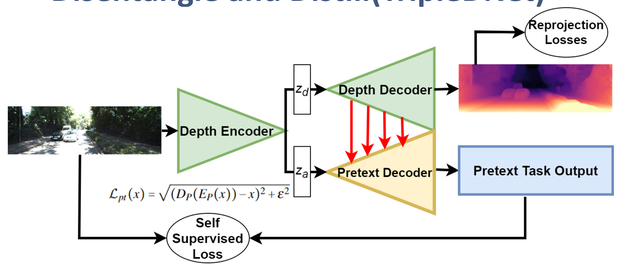
★ Masked DG2C(**MD2C**)

$$\mathcal{L}_{mdg2c}(x) = \hat{M} \odot \sqrt{(PL((1-\hat{M}) \odot (D(x) \oplus GS(x))) - AB(x))^2 + \varepsilon^2}$$

Qualitative Results



Disentangle and Distill(TripleDNet)



Disentanglement for geometry(depth decoder) and appearance(pretext decoder)

$$\mathcal{L}_{rp}(I_t, I_{s\to t}) = \psi * \mathcal{L}_{pw}(I_t, I_{s\to t}) + \lambda * \frac{1 - SSIM(I_t, I_{s\to t})}{2} \qquad \mathcal{L}_{fm}(F_t, F_{s\to t}) = \mathcal{L}_{pw}(F_t, F_{s\to t})$$

$$\mathcal{L}_{pw}(x,y) = \sqrt{(x-y)^2 + \varepsilon^2}$$

$$\mathcal{L}_{total} = \mathcal{L}_{rp} + \alpha * \mathcal{L}_{pt} + \beta * \mathcal{L}_{fm}$$

Quantitative Results

KITTI

Method	Superv.	Encoder	Res.	↓ Abs Rel	↓ Sq Rel	↓ RMSE	↓ RMSElog	↑ δ_1	↑ δ_2	↑ δ_3
				Lower is better				*Higher is better*		
Wang et al.[45]	M	RN18	640x192	0.109	0.779	4.641	0.186	0.883	0.962	0.982
DDV[20]	M	RN101	640x192	0.106	0.861	4.699	0.185	0.889	0.962	0.982
Jung et al. [21]	M+Sem	RN50	640x192	0.102	0.675	4.393	0.178	0.893	0.966	0.984
D2G	M	RN50	640x192	0.108	0.738	4.639	0.185	0.882	0.963	0.983
DG2C	M	RN50	640x192	0.107	0.742	4.607	0.183	0.886	0.964	0.983
TripleD	M	RN50	640x192	0.104	0.714	4.509	0.181	0.890	0.964	0.984
Monodepth2[11]	M	RN50	1024x320	0.110	0.831	4.642	0.187	0.883	0.962	0.982
SGDepth[23]	M+Sem	RN18	1280x384	0.107	0.768	4.468	0.186	0.891	0.963	0.982
PackNet[14]	M	PackNet	1280x380	0.107	0.802	4.538	0.186	0.889	0.962	0.981
HRDepth[31]	M	RN18	1024x320	0.106	0.755	4.472	0.181	0.892	0.966	0.984
FeatDepth[38]	M	RN50	1024x320	0.104	0.729	4.481	0.179	0.893	0.965	**0.987**
CamLessMD[4]	M	RN50	1024x320	0.102	0.723	4.374	0.178	0.898	0.966	0.983
Jung et al. [21]	M+Sem	RN18	1024x320	0.102	0.687	4.366	0.178	0.895	0.967	0.984
X-Distill[2]	M+Sem	RN50	1024x320	0.102	0.698	4.439	0.180	0.895	0.965	0.983
SGRL[15]	M+Sem	PackNet	1024x320	0.100	0.761	**4.270**	0.175	0.902	0.965	0.982
DIFFNet [50]	M	HRNet	1024x320	**0.097**	0.722	4.345	0.174	**0.907**	0.967	0.984
TripleD (*sup.*)	M	RN50	1024x320	0.103	0.726	4.437	0.180	0.896	0.965	0.983
DG2C	M	RN50	1024x320	0.099	0.668	4.448	0.176	0.893	0.966	0.985
D2G	M	RN50	1024x320	0.098	0.676	4.307	0.175	0.903	0.967	0.984
MD2C	M	RN50	1024x320	0.099	0.652	4.338	0.174	0.898	0.968	0.984
MDG2C	M	RN50	1024x320	0.099	0.651	4.336	0.173	0.897	0.967	0.985
TripleD	M	RN50	1024x320	0.099	**0.648**	4.296	**0.173**	0.901	**0.968**	0.985

- Generalizability of our approach Make3D
- Model pretrained on KITTI dataset evaluated on Make3D

Make3D

Method	Superv.	↓ Abs Rel	↓ Sq Rel	↓ RMSE	↓ RMSElog
Monodepth [12]	S	0.544	10.94	11.760	0.193
SfMLearner [52]	M	0.383	5.321	10.470	0.478
DDVO [42]	M	0.387	4.720	8.090	0.204
Monodepth2[11]	M	0.322	3.589	7.417	0.163
X-Distill[2]	M	0.308	3.122	7.015	0.158
TripleD	M	**0.303**	**3.032**	**6.907**	**0.155**

Encoder Initialization

Encoder Init	↓Abs Rel	↓Sq Rel	↓RMSE	↓RMSElog	↑δ_1	↑δ_2	↑δ_3
Supervised	0.103	0.726	4.437	0.180	0.896	0.965	0.984
MoCo	0.103	0.736	4.486	0.177	0.899	0.964	0.984
SimCLR	0.101	0.699	4.443	0.176	0.895	0.967	0.984
SwAV	**0.099**	**0.648**	**4.296**	**0.173**	**0.901**	**0.968**	**0.985**