Learning Depth Estimation for Transparent and Mirror Surfaces

Alex Costanzino, Pierluigi Zama Ramirez, Matteo Poggi, Fabio Tosi, Stefano Mattoccia, Luigi Di Stefano

University of Bologna - Department of Computer Science and Engineering

{alex.costanzino, pierluigi.zama, m.poggi, fabio.tosi5}@unibo.it

ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

Computer Vision Laboratory

Project Page

Paper

ICCV23 PARIS

Abstract

Inferring the depth of transparent or mirror (ToM) surfaces represents a hard challenge for either sensors, algorithms, or deep networks. We propose a simple pipeline for learning to estimate depth properly for such surfaces with neural networks, without requiring any ground-truth annotation. We unveil how to obtain reliable pseudo labels by in-painting ToM objects in images and processing them with a monocular depth estimation model. These labels can be used to fine-tune existing monocular or stereo networks, to let them learn how to deal with ToM surfaces. Experimental results on the Booster dataset show the dramatic improvements enabled by our remarkably simple proposal.

Overview & Contributions

Overview: We propose a simple yet very effective strategy that consists in tricking monocular depth estimation networks by replacing ToM objects with virtually textured ones, inducing them to hallucinate their depths, to produce annotations. Annotations are then fed to a fine-tuning pipeline to enhance state-of-art models to deal also with ToM surfaces.

Key Contributions:

- We propose a robust virtual depth generation strategy;
- We introduce a pipeline for fine-tuning monocular depth estimation networks to deal with ToM objects;
- We present an extension for stereo depth estimation settings, i.e. stereo matching.

Virtual Depth Results – GT vs Proxy

Training on only the test set of MSD and Trans10K with proxy masks, results on the Booster train set at quarter resolution. All models start from the official weights.

		MiDaS							DPT						
Category	Method	$\delta < 1.25$ ↑ (%)	$\delta < 1.20$ ↑ (%)	$\delta < 1.15$ ↑ (%)	$\delta < 1.10$ ↑ (%)	$\delta < 1.05$ ↑ (%)	MAE ↓ (mm)	RMSE ↓ (mm)	$\delta < 1.25$ ↑ (%)	$\delta < 1.20$ ↑ (%)	$\delta < 1.15$ ↑ (%)	$\delta < 1.10$ ↑ (%)	$\delta < 1.05$ ↑ (%)	MAE ↓ (mm)	RMSE ↓ (mm)
All	Base	94.56	91.72	85.68	74.00	50.12	90.82	120.51	96.79	94.45	89.71	79.00	56.26	75.35	100.68
All	Virtual Depth (Proxy)	91.78	87.52	79.57	66.00	42.21	105.67	140.00	93.23	89.43	81.98	68.06	43.62	98.09	128.36
All	Ft. Virtual Depth (GT)	94.99	92.12	86.32	74.99	49.91	88.63	117.40	98.09	**96.85**	**93.91**	**83.50**	**58.74**	**65.52**	**86.41**
All	Ft. Virtual Depth (Proxy)	**95.00**	**92.17**	**86.48**	**75.46**	**50.24**	**88.10**	**117.10**	**98.11**	96.68	93.48	83.00	58.13	66.43	87.18
ToM	Base	87.44	83.40	72.71	59.63	36.28	122.33	140.31	92.77	88.77	80.98	62.46	37.70	113.14	136.28
ToM	Virtual Depth (Proxy)	86.37	79.35	69.91	54.84	32.31	112.92	124.10	89.75	85.26	75.17	56.57	31.75	110.74	123.28
ToM	Ft. Virtual Depth (GT)	**91.85**	**89.93**	**83.25**	**70.63**	**47.55**	**92.05**	**103.54**	**96.95**	**96.26**	**93.27**	**81.84**	**55.49**	**67.95**	**80.88**
ToM	Ft. Virtual Depth (Proxy)	91.12	88.36	81.18	69.15	45.02	96.33	107.59	96.82	96.05	92.57	80.62	53.74	70.67	83.44
Other	Base	94.57	91.81	85.99	74.01	50.28	91.08	119.86	97.10	94.84	90.08	79.76	57.31	73.19	95.63
Other	Virtual Depth (Proxy)	91.80	87.47	79.41	65.86	41.95	108.09	143.64	93.22	89.39	82.07	68.30	43.88	98.64	129.48
Other	Ft. Virtual Depth (GT)	**95.27**	92.30	86.60	75.43	49.95	88.81	118.48	98.19	**96.85**	**93.85**	**83.70**	**58.93**	**65.65**	**86.30**
Other	Ft. Virtual Depth (Proxy)	**95.27**	**92.42**	**86.84**	**76.02**	**50.48**	**88.07**	**117.99**	**98.20**	96.67	93.35	83.20	58.40	66.51	87.06

Merged Virtual Depth Quantitatives – GT vs Proxy

Our pipeline improves the performance of both RAFT-Stereo and CREStereo on ToM objects, even in the case of extremely noisy proxy semantic annotations.

		RAFT-Stereo						CREStereo					
Category	Method	bad-2 ↓ (%)	bad-4 ↓ (%)	bad-6 ↓ (%)	bad-8 ↓ (%)	MAE ↓ (px)	RMSE ↓ (px)	bad-2 ↓ (%)	bad-4 ↓ (%)	bad-6 ↓ (%)	bad-8 ↓ (%)	MAE ↓ (px)	RMSE ↓ (px)
All	Base	17.42	13.49	11.59	10.11	4.07	8.63	15.13	10.70	8.91	7.57	3.15	7.40
All	Ft. Merged (Proxy)	19.56	13.53	10.53	8.28	2.51	5.37	12.51	7.33	5.16	3.56	**1.38**	**3.30**
All	Ft. Merged (GT)	**14.68**	**9.63**	**7.36**	**5.58**	**1.95**	**4.58**	**10.85**	**6.11**	**4.39**	**3.12**	1.51	3.62
ToM	Base	56.77	44.38	38.43	33.31	13.45	16.56	51.83	37.88	32.86	28.19	12.42	15.60
ToM	Ft. Merged (Proxy)	51.85	35.32	27.17	20.88	6.56	8.04	40.63	23.38	16.55	10.81	**4.11**	**5.55**
ToM	Ft. Merged (GT)	**47.54**	**30.55**	**22.81**	**16.62**	**5.83**	**7.43**	**36.90**	**20.75**	**15.38**	**10.65**	5.02	6.69
Other	Base	8.48	5.74	4.52	3.83	1.58	3.79	8.11	4.83	3.50	2.78	1.14	2.77
Other	Ft. Merged (Proxy)	11.95	7.36	5.31	4.15	1.41	3.41	6.51	3.65	2.52	1.88	0.82	2.00
Other	Ft. Merged (GT)	**7.14**	**4.09**	**2.88**	**2.23**	**0.96**	**2.64**	**5.27**	**2.72**	**1.81**	**1.34**	**0.68**	**1.69**

Mono Distillation Pipeline

Given an RGB and a segmentation mask, we in-paint pixels belonging to transparent and mirror surfaces with a random uniform color and process these augmented images with a pre-trained monocular network. The obtained virtual depths are aggregated to obtain a pseudo-labeled dataset for fine-tuning the network itself.

Mono Out-of-domain Results

Examples of predictions by MiDaS and DPT trained on all MSD and Trans10K and tested on the Booster train set at quarter resolution. All models start from the official weights. For each model, we show results achieved by the original model and by fine-tuned instances using proxy or GT segmentation masks.

		MiDaS							DPT						
Category	Method	$\delta < 1.25$ ↑ (%)	$\delta < 1.20$ ↑ (%)	$\delta < 1.15$ ↑ (%)	$\delta < 1.10$ ↑ (%)	$\delta < 1.05$ ↑ (%)	MAE ↓ (mm)	RMSE ↓ (mm)	$\delta < 1.25$ ↑ (%)	$\delta < 1.20$ ↑ (%)	$\delta < 1.15$ ↑ (%)	$\delta < 1.10$ ↑ (%)	$\delta < 1.05$ ↑ (%)	MAE ↓ (mm)	RMSE ↓ (mm)
All	Base	94.56	91.72	85.68	74.00	50.12	90.82	120.51	96.79	94.45	89.71	79.00	56.26	75.35	100.68
All	Ft. Base	94.09	91.28	85.29	73.55	49.14	93.34	124.54	96.83	94.60	90.14	79.46	56.89	76.00	100.98
All	Ft. Virtual Depth	**95.07**	**92.31**	**86.39**	**75.20**	**50.57**	**88.83**	**118.23**	**97.99**	**96.65**	**93.55**	**83.94**	**60.46**	**64.93**	**85.93**
ToM	Base	87.44	83.40	72.71	59.63	36.28	122.33	140.31	92.77	88.77	80.98	62.46	37.70	113.14	136.28
ToM	Ft. Base	86.30	82.96	72.84	61.04	37.91	126.71	145.69	92.69	89.49	81.12	63.62	37.95	118.84	141.27
ToM	Ft. Virtual Depth	**91.81**	**89.12**	**81.68**	**70.75**	**47.85**	**93.66**	**104.80**	**96.68**	**95.96**	**92.23**	**79.96**	**54.67**	**70.68**	**83.06**
Other	Base	94.57	91.81	85.99	74.01	50.28	91.08	119.86	97.10	94.84	90.08	79.76	57.31	73.19	95.63
Other	Ft. Base	94.27	91.49	85.87	73.77	49.38	92.48	122.46	97.15	95.02	90.97	80.77	58.05	72.45	94.07
Other	Ft. Virtual Depth	**95.33**	**92.53**	**86.80**	**75.43**	**50.46**	**89.17**	**119.58**	**98.07**	**96.64**	**93.52**	**84.30**	**61.19**	**64.70**	**85.57**

Point Clouds

Examples of point cloud predictions by MiDaS and DPT, RAFT-Stereo and CREStereo. For each model, we show resulting point clouds achieved by the original model and by fine-tuned instances using proxy or GT segmentation masks.

Stereo Distillation Pipeline

Given a stereo pair and a segmentation mask for the left image, we merge predictions from a pre-trained stereo network with virtual depths obtained by a monocular network with our strategy. We merge the monocular or stereo maps, by taking values belonging to either ToM or other surfaces from mono or stereo, respectively. These final merged depth labels are used to fine-tune the original stereo network.

Stereo Out-of-domain Results

Examples of predictions by RAFT-Stereo and CREStereo trained on all MSD and Trans10K and tested on the Booster train set at quarter resolution. All models start from the official weights. For each model, we show results achieved by the original model and by fine-tuned instances using proxy or GT segmentation masks.

		RAFT-Stereo						CREStereo					
Category	Method	bad-2 ↓ (%)	bad-4 ↓ (%)	bad-6 ↓ (%)	bad-8 ↓ (%)	MAE ↓ (px)	RMSE ↓ (px)	bad-2 ↓ (%)	bad-4 ↓ (%)	bad-6 ↓ (%)	bad-8 ↓ (%)	MAE ↓ (px)	RMSE ↓ (px)
All	Base	17.42	13.49	11.59	10.11	4.07	8.63	15.13	10.70	8.91	7.57	3.15	7.40
All	Ft. Base	18.42	14.02	12.00	10.47	4.32	8.91	14.66	10.16	8.41	6.99	2.88	6.72
All	Ft. Virtual Depth	16.80	12.35	9.99	8.09	2.60	6.04	18.83	14.28	12.33	10.80	5.09	9.78
All	Ft. Merged	**14.68**	**9.63**	**7.36**	**5.58**	**1.95**	**4.58**	**10.85**	**6.11**	**4.39**	**3.12**	**1.51**	**3.62**
ToM	Base	56.77	44.38	38.43	33.31	13.45	16.56	51.83	37.88	32.86	28.19	12.42	15.60
ToM	Ft. Base	58.57	44.42	38.00	32.99	13.84	16.47	49.89	34.89	29.40	24.63	10.60	13.43
ToM	Ft. Virtual Depth	57.08	42.89	34.74	27.62	8.48	10.75	49.51	35.14	29.60	24.57	11.94	13.85
ToM	Ft. Merged	**47.54**	**30.55**	**22.81**	**16.62**	**5.83**	**7.43**	**36.90**	**20.75**	**15.38**	**10.65**	**5.02**	**6.69**
Other	Base	8.48	5.74	4.52	3.83	1.58	3.79	8.11	4.83	3.50	2.78	1.14	2.77
Other	Ft. Base	9.20	6.20	4.89	4.14	1.70	3.98	7.01	4.20	3.15	2.53	1.00	2.56
Other	Ft. Virtual Depth	7.97	5.04	3.82	3.16	1.19	3.21	11.44	8.17	6.96	6.25	2.72	5.49
Other	Ft. Merged	**7.14**	**4.09**	**2.88**	**2.23**	**0.96**	**2.64**	**5.27**	**2.72**	**1.81**	**1.34**	**0.68**	**1.69**