Supplementary Material – Kick Back & Relax: Learning to Reconstruct the World by Watching SlowTV

Jaime Spencer
University of Surrey
j.spencermartin@surrey.ac.uk

Chris Russell
Oxford Internet Institute
christopher.m.russell@gmail.com

Simon Hadfield
University of Surrey
s.hadfield@surrey.ac.uk

Richard Bowden
University of Surrey
r.bowden@surrey.ac.uk

A. SlowTV Dataset

Figure 2 shows a frame from each SlowTV video, while Figure 3 shows their map location. Sequences [00-27] are hiking scenes, [28-30] scuba diving and [31-39] driving. As seen, this dataset provides an incredible diversity of environments and locations, enabling us to train models capable of generalizing to previously unseen scene types.

B. Aspect Ratio Augmentation

To make the models invariant to the training image size, we propose to incorporate an aspect ratio augmentation. For more information see Section 4.3 in the main paper. Sample training images obtained using this procedure an be found in Figure 1. The centre crop is uniformly sampled from a set of predetermined aspect ratios:

- Portrait: 6:13, 9:16, 3:5, 2:3, 4:5, 1:1
- Landscape: 5:4, 4:3, 3:2, 14:9, 5:3, 16:9, 2:1, 24:10, 33:10, 18:5

C. Evaluation Datasets

Kitti Eigen-Benchmark [5]. (Test: 652) Subset of the common Kitti Eigen split with corrected LiDAR [15].

Kitti Eigen-Zhou [5]. (Val: 700) Subset of the Kitti Eigen-Zhou val split with corrected LiDAR [15].

Mannequin Challenge [5]. (Test: 1k) Subset of the original test split, using COLMAP [13] depth reconstructions.

SYNS-Patches [1, 14]. (Val: 400, Test: 775) Official val and test splits consisting of dense LiDAR maps.

DDAD [8]. (Test: 1k) Subset of the official val split, featuring LiDAR maps with an increased range up to 250m.

Sintel [5]. (Test: 1064) Official test split, consisting of synthetic image & depth pairs from highly dynamic scenes

Table 1: Learning Camera Intrinsics. Performance when training on a single dataset (Kitti or Mannequin Challenge) and learning camera intrinsics. If the cameras are not perfectly calibrated, learning the intrinsics can improve accuracy.

	Kitti Eigen-Zhou				Mannequin		
	Rel↓	F↑	$\delta_{.25}$↑		Rel↓	F↑	$\delta_{.25}$↑
Baseline	5.69	**60.88**	95.89	Baseline	16.66	14.20	77.18
Learn **K**	**5.68**	60.81	**95.90**	Learn **K**	**16.12**	**14.77**	**78.40**

DIODE Indoors [16]. (Test: 325) Official val split with dense LiDAR depth maps.

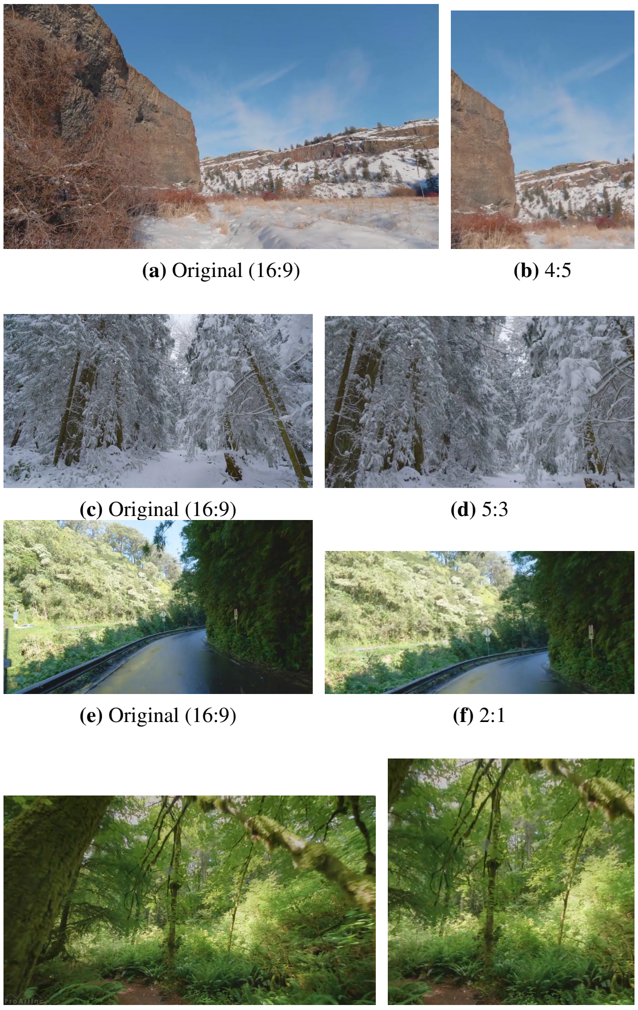
DIODE Outdoors [16]. (Test: 446) Official val split with dense LiDAR depth maps.

NYUD-v2 [10]. (Test: 654) Official test split collected using a Kinect RGB-D camera.

TUM-RGBD [5]. (Test: 2.5k) Subset of dynamic scenes with moving people also collected using a Kinect.

D. Leaning Camera Intrinsics

Estimating the intrinsics parameters is required when training with uncalibrated cameras. However, this procedure can be applied even if the camera parameters are known. Table 1 shows results when training on either Kitti Eigen-Benchmark or Mannequin Challenge. If the dataset provides accurately calibrated cameras (Kitti), self-supervised learning of the intrinsics is on par with using the ground-truth parameters. However, when the ground-truth parameters are estimated using COLMAP [13], learning the intrinsics can slightly improve performance.



(a) Original (16:9) **(b)** 4:5

(c) Original (16:9) **(d)** 5:3

(e) Original (16:9) **(f)** 2:1

(g) Original (16:9) **(h)** 1:1

Figure 1: AR-Aug. Additional augmentations used to diversify the variety of image shapes and object scales seen by the network.

E. Dynamic Objects

MDE models trained exclusively using monocular supervision are prone to artefacts from dynamic objects. For instance, vehicles moving at similar speeds to the camera can produce holes of infinite depth due to their static appearance across images. Meanwhile, other dynamic objects can result in underestimated depth when moving towards the camera, or overestimated depth when moving away from it. This is due to the additional motion causing incorrect correspondences in the warping procedure.

Existing approaches that address these dynamic objects [7, 2, 3] rely on additional labels such as semantic or instance segmentation. We instead opt for the losses proposed by Monodepth2 [6] as a simpler proxy without increased computation or label requirements.

We test the effectiveness of these constraints on a smaller subset of all three training datasets. These results can be found in Table 2 and Figure 4. Despite not explicitly modelling dynamic objects, Monodepth2 drastically increases the accuracy and robustness. This can be seen both in the improved metrics and the reduction in visual artefacts.

F. Median Alignment Results

Table 3 shows results when applying median depth alignment between prediction and ground-truth. As expected, this generally results in worse performance that estimating both scale and shift parameters. This is particularly noticeable for MiDaS, DPT and the SSL baselines.

G. Failure Cases

Whilst representing a significant milestone in SS-MDE, our model still suffers from several failure cases. We show these in Figure 5. For instance, Kitti shows a car estimated as a hole of infinite depth, despite training with the minimum reconstruction loss and automasking [6]. Several visualizations are also characterized by texture-copy artefacts. In some cases, our models estimated incorrect relative object positions (*e.g.* Sintel or DDAD). An interesting failure case for all approaches are highly-reflective surfaces, such as mirrors or TVs. These are challenging due to the fact that they do not violate the photometric error and obtaining LiDAR or Structure-from-Motion (SfM) ground-truth is highly challenging. Finally, due to the strong prior for upright images, our model struggles to adapt to extreme rotations (TUM-RGBD). This could be mitigated with additional augmentations. Finally, it is worth pointing out that, in the vast majority of these cases, our model outperforms the SSL baselines.

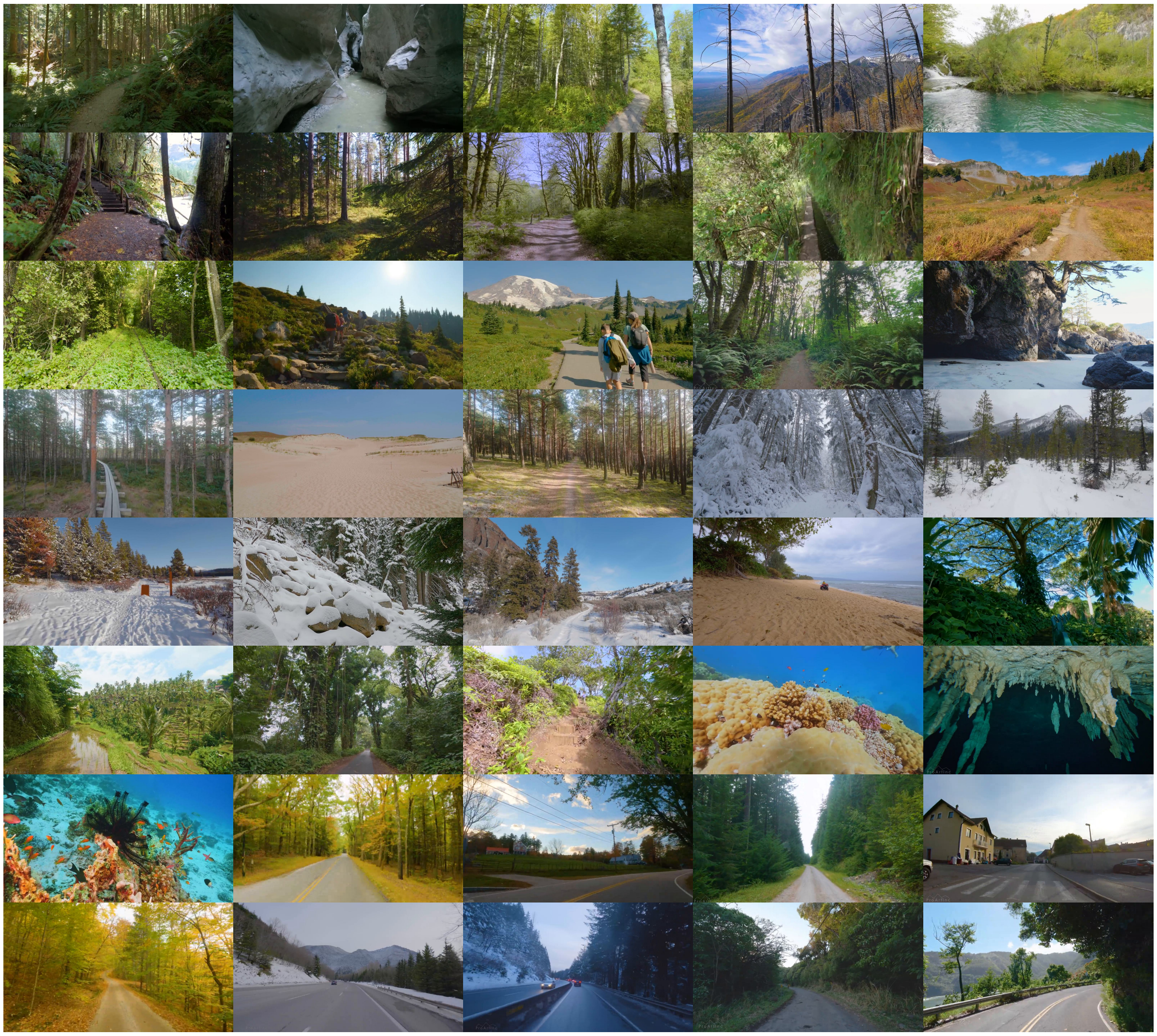


Figure 2: SlowTV Dataset. We show one frame per video from the proposed SlowTV. The dataset contains a diverse set of environments in a range of environmental conditions. The final dataset has a total of 1.7M images, with 1.15M natural, 400k driving and 180k underwater.

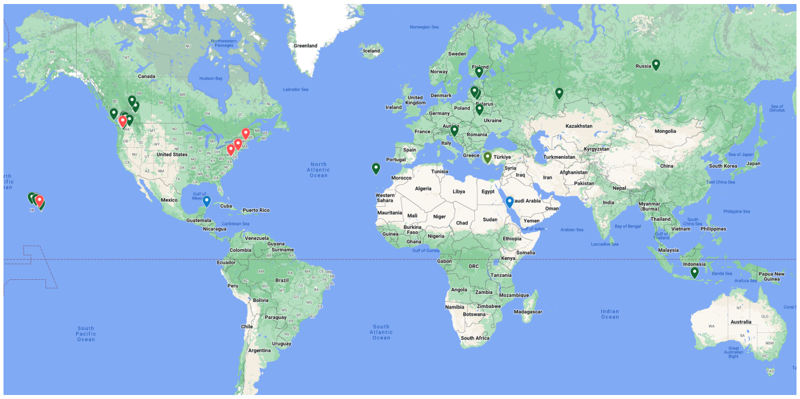


Figure 3: SlowTV Map. Distribution of locations in the proposed dataset. **Green**=Natural, **Red**=Driving, **Blue**=Underwater.

Table 2: Monodepth2 [6] Losses. The minimum reconstruction loss and automasking from Monodepth2 serve as valuable proxies to increase robustness to dynamic objects, while remaining simple and efficient.

	Multi-task		Kitti		Mannequin		DDAD		DIODE		Sintel		SYNS		DIODE		NYUD-v2		TUM	
	Rank↓	Δ↑	Rel↓	F↑	Rel↓	F↑	Rel↓	F↑	Rel↓	$\delta_{.25}$↑	Rel↓	F↑	Rel↓	F↑	Rel↓	$\delta_{.25}$↑	Rel↓	$\delta_{.25}$↑	Rel↓	$\delta_{.25}$↑
Baseline	1.89	0.00	9.00	53.50	**16.89**	**14.66**	23.57	11.13	35.99	52.70	35.33	38.15	25.47	15.73	17.91	75.03	21.68	71.41	17.69	75.67
MinRec+Automask	**1.11**	**7.01**	**6.50**	**55.62**	16.96	14.48	**18.49**	**11.64**	**35.62**	**52.95**	**34.97**	**38.83**	**24.44**	**16.25**	**16.85**	**76.50**	**14.27**	**80.54**	**17.23**	**76.23**

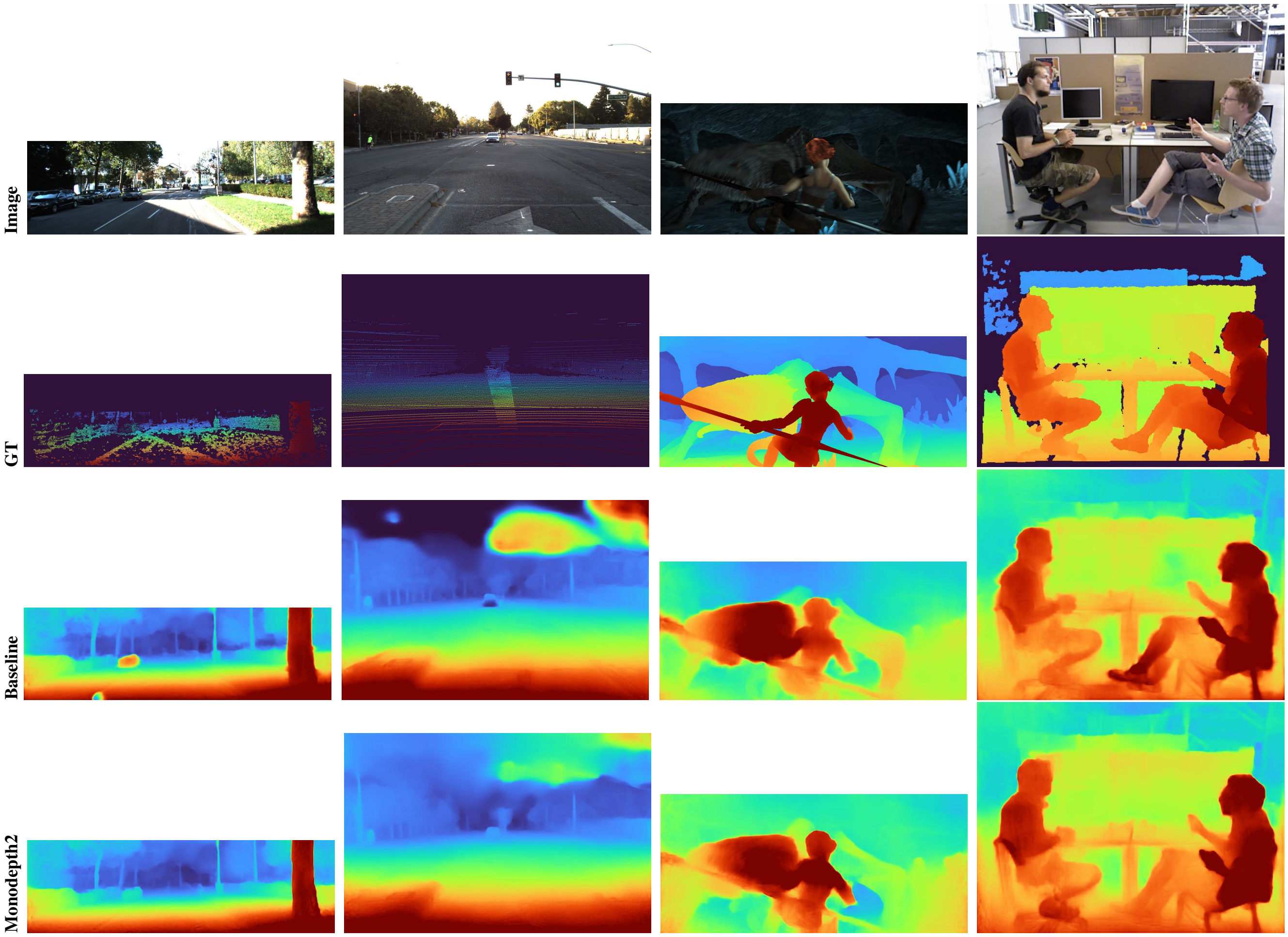


Figure 4: Monodepth2 Losses. Monodepth2 [6] reduces the presence of holes of infinite depth and dynamic object artefacts. The sharpness of object boundaries are also improved due to the refined correspondences from the minimum reconstruction loss.

Table 3: Median-Scaling Results. This represents the common SS-MDE (SS-MDE) evaluation procedure [19]. Removing the shift alignment reduces performance for all approaches. Our method still outperforms all existing SS-MDE models, and NeWCRFs (NeWCRFs) in many cases.

		In-Distribution				**Outdoor**								**Indoor**					
		Kitti		**Mannequin**		**DDAD**		**DIODE**		**Sintel**		**SYNS**		**DIODE**		**NYUD-v2**		**TUM**	
	Train	Rel↓	F↑	Rel↓	F↑	Rel↓	F↑	Rel↓	$\delta_{.25}$↑	Rel↓	F↑	Rel↓	F↑	Rel↓	$\delta_{.25}$↑	Rel↓	$\delta_{.25}$↑	Rel↓	$\delta_{.25}$↑
Garg [4]	S	7.65	53.28	34.55	9.29	26.77	4.77	57.87	42.85	53.16	30.98	31.68	13.58	30.63	51.00	26.78	54.29	27.37	55.26
Monodepth2 [6]	MS	7.90	50.50	35.88	8.18	25.46	4.77	57.61	43.21	54.40	30.11	30.05	13.28	33.51	47.49	29.87	50.08	30.59	49.82
DiffNet [18]	MS	7.98	49.60	35.50	8.15	24.17	4.75	55.68	45.37	55.23	29.44	29.75	13.41	28.67	53.82	26.62	54.69	28.56	53.07
HR-Depth [9]	MS	7.70	51.49	35.89	8.62	24.01	5.08	57.88	43.92	53.91	30.89	29.87	14.03	32.88	47.67	27.32	53.06	29.22	52.31
KBR (Ours)	M	**7.23**	**54.63**	**18.73**	**15.04**	**14.01**	**14.01**	**43.80**	**60.84**	**37.06**	**36.01**	**24.92**	**16.49**	**18.88**	**72.09**	**13.27**	**83.65**	**16.60**	**76.48**
MiDaS [12]	D	18.45	20.13	26.02	10.61	18.38	8.28	**48.63**	**60.15**	**39.09**	**32.72**	35.30	9.18	18.08	74.48	23.11	69.67	17.75	76.99
DPT-ViT [11]	D	14.23	36.25	28.54	11.38	17.83	8.99	72.46	49.09	128.86	29.58	32.69	12.93	36.82	55.15	24.82	67.95	24.33	78.16
DPT-BEiT [11]	D	18.20	37.46	30.79	12.58	15.39	11.78	70.30	50.03	60.20	29.54	31.09	13.76	51.07	53.11	75.32	42.91	25.27	**83.07**
NeWCRFs [17]	D	**5.55**	**56.45**	**22.15**	**13.68**	**11.87**	**13.44**	50.52	51.16	48.42	32.30	**27.79**	**14.50**	**16.15**	**79.52**	**7.00**	**94.44**	**14.93**	80.63

Highlighted cells are **NOT zero-shot** *results.* ***S****=Stereo,* ***M****=Monocular,* ***D****=Ground-truth Depth.*

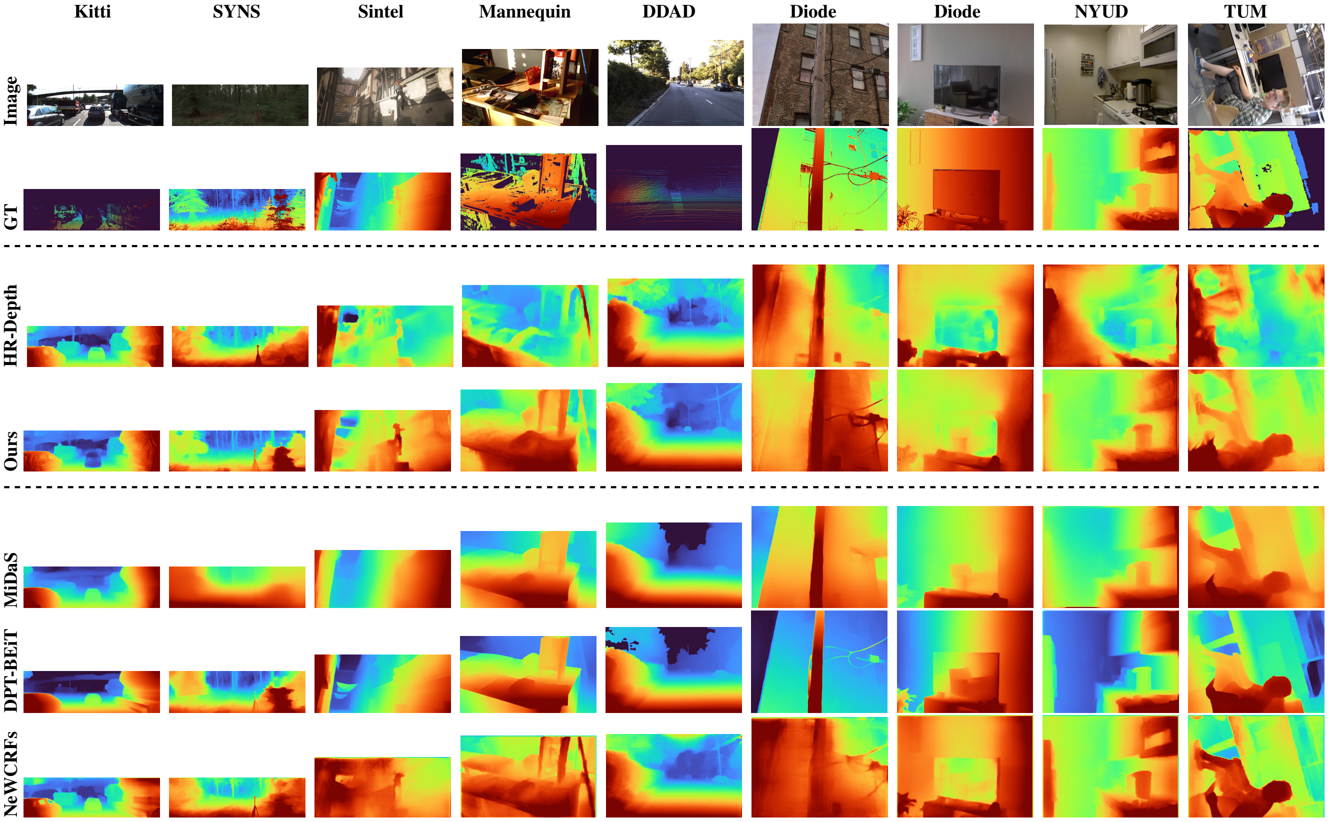


Figure 5: Failure Cases. The proposed model occasionally produces holes of infinite depth or texture-copy artefacts. However, complex regions such as foliage or boundaries tend to be oversmoothed by all approaches. Finally, the upright prior in training data makes the model less robust to strong rotations. *Middle=Self-Supervised – Bottom=Supervised.*

References

[1] Wendy J Adams, James H Elder, Erich W Graf, Julian Leyland, Arthur J Lugtigheid, and Alexander Muryy. The Southampton-York Natural Scenes (SYNS) dataset: Statistics of surface attitude. *Scientific Reports*, 6(1):35805, 2016. 1

[2] Vincent Casser, Soeren Pirk, Reza Mahjourian, and Anelia Angelova. Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8001–8008, 2019. 2

[3] Qi Dai, Vaishakh Patil, Simon Hecker, Dengxin Dai, Luc Van Gool, and Konrad Schindler. Self-supervised object motion and depth estimation from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020. 2

[4] Ravi Garg, Vijay Kumar, Gustavo Carneiro, and Ian Reid. Unsupervised CNN for Single View Depth Estimation: Geometry to the Rescue. In *European Conference on Computer Vision*, pages 740–756, 2016. 5

[5] A Geiger, P Lenz, C Stiller, and R Urtasun. Vision meets robotics: The KITTI dataset. *International Journal of Robotics Research*, 32(11):1231–1237, 2013. 1

[6] Clement Godard, Oisin Mac Aodha, Michael Firman, and Gabriel Brostow. Digging Into Self-Supervised Monocular Depth Estimation. *International Conference on Computer Vision*, 2019-Octob:3827–3837, 2019. 2, 4, 5

[7] Ariel Gordon, Hanhan Li, Rico Jonschkowski, and Anelia Angelova. Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8977–8986, 2019. 2

[8] Vitor Guizilini, Ambrus Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3D packing for self-supervised monocular depth estimation. *Conference on Computer Vision and Pattern Recognition*, pages 2482–2491, 2020. 1

[9] Xiaoyang Lyu, Liang Liu, Mengmeng Wang, Xin Kong, Lina Liu, Yong Liu, Xinxin Chen, and Yi Yuan. HR-Depth: High Resolution Self-Supervised Monocular Depth Estimation. *AAAI Conference on Artificial Intelligence*, 35(3):2294–2301, 2021. 5

[10] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 1

[11] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12179–12188, October 2021. 5

[12] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 5

[13] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-Motion Revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[14] Jaime Spencer, Chris Russell, Simon Hadfield, and Richard Bowden. Deconstructing self-supervised monocular reconstruction: The design decisions that matter. *Transactions on Machine Learning Research*, 2022. Reproducibility Certification. 1

[15] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity Invariant CNNs. *International Conference on 3D Vision*, pages 11–20, 2018. 1

[16] Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z Dai, Andrea F Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R Walter, et al. Diode: A dense indoor and outdoor depth dataset. *arXiv preprint arXiv:1908.00463*, 2019. 1

[17] Weihao Yuan, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, and Ping Tan. Neural window fully-connected crfs for monocular depth estimation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3906–3915, 2022. 5

[18] Hang Zhou, David Greenwood, and Sarah Taylor. Self-Supervised Monocular Depth Estimation with Internal Feature Fusion. In *British Machine Vision Conference*, 2021. 5

[19] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. Unsupervised Learning of Depth and Ego-Motion from Video. *Conference on Computer Vision and Pattern Recognition*, pages 6612–6619, 2017. 5