Supplementary Material for - SeasonDepth: Cross-Season Monocular Depth Prediction Dataset and Benchmark under Multiple Environments

Hanjiang Hu[1,2] **Baoquan Yang**[1] **Zhijian Qiao**[1] **Ding Zhao**[2] **Hesheng Wang**[1*]
[1]Shanghai Jiao Tong University [2]Carnegie Mellon University
{hanjianghu,dingzhao}@cmu.edu {yangbaoquan,qiaozhijian,wanghesheng}@sjtu.edu.cn

1 Building SeasonDepth Dataset

In this section, we present more details about the process of building *SeasonDepth* dataset and statistical analysis of depth maps in each environment.

1.1 Details in Building Dataset

We adopt the categorized slices of Urban part according to [1] as original images after rectification though camera intrinsic file. Specifically, we use `slice2, slice3, slice7, slice8` as the split test slices for evaluation and benchmark, and the other slices `slice4, slice5, slice6` are intended to treat as training set. Note that since not all images from original dataset are appropriate for depth prediction due to huge noise, *e.g.*, a moving truck covering almost all the pixels, we remove such images in the final version. The released dataset split can be found on `https://seasondepth.github.io/`. The numbers of images under all the environments for all slices in training set and test set are shown in Tab. 1. The abbreviations of environments are `S` for `Sunny`, `C` for `Cloudy`, `O` for `Overcast`, `LS` for `Low Sun`, `Sn` for `Snow`, `F` for `Foliage`, `NF` for `No Foliage`, and `MF` for `Mixed Foliage`. It could be seen that the total number of test set is a little bit larger than that of training set with more different slices, which helps to make the benchmark results more accurate and reliable. Also, the training set can be used for fine-tuning for pretrained models, which do not need too many images. Images from left and right cameras are merged together in the same slice for calculation.

We adopt COLMAP's MVS pipeline [2, 3] to find the 3D structure and depth map. We follow the instruction on `https://colmap.github.io/` with sequential SIFT matching with RANSAC, sparse reconstruction and dense reconstruction. Some important detailed hyperparameters can be found in Tab. 2, while others are with default configuration.To make full use of the image sequences, we adjust the sequential matching overlap to be 15 instead of the whole sequence, improving the local optimization with less noise. During each iteration of RANSAC algorithm in triangulation, the minimum inlier ratio for SIFT matching is set to be 0.65 for the consideration that most pixels of a single image are static in most cases. The maximum SIFT matching distance is 0.55 to adapt the distance of dynamic objects and improve the efficiency. The image samples after SfM can be found in Fig. 1-(b)

The valid pixels of the original depth map are between lower threshold and upper threshold to filter most noise pixels. For one thing, since the fields, forests and cloud in the far distance away from the camera matter little to the depth prediction applications for the autonomous driving, we truncate the depth values over 92% (80% in some cases) of the whole image to focus more on the near roads, vehicles, buildings, vegetation, *etc.* For another, due to the camera placement on the both sides of the car, the very near descriptors of the road cannot be correctly matched during SfM and reconstructed

*Corresponding author

Submitted to the 35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks. Do not distribute.

Table 1: Numbers of Images under All the Environments for All Slices

Environments	Training Set				Test Set				
	slice4	slice5	slice6	All Slices	slice2	slice3	slice7	slice8	All Slices
S+NF Apr. 4th	446	197	544	1187	382	450	190	449	1471
S+F Sept. 1st	496	243	496	1235	385	464	249	490	1588
S+F Sept. 15th	404	222	526	1152	335	329	462	457	1583
C+F Oct. 1st	406	216	628	1250	347	438	350	244	1379
S+F Oct. 19th	382	201	558	1141	301	439	412	230	1382
O+MF Oct. 28th	394	204	536	1134	333	418	362	442	1555
LS+MF Nov. 3rd	448	212	518	1178	335	447	203	416	1401
LS+MF Nov. 12th	484	221	558	1263	352	500	357	501	1710
C+MF Nov. 22nd	394	200	578	1172	298	436	380	423	1537
LS+NF+Sn Dec. 21st	482	213	592	1287	284	512	56	147	999
LS+F Mar. 4th	380	214	576	1170	354	222	0	512	1088
O+F Jul. 28th	458	223	560	1241	256	425	384	467	1532
All Environments	5174	2566	6670	14410	3962	5080	3405	4778	17225

for dense depth map, which should be removed by filtering the pixel values less than 5% of the whole depth map. Besides, in the special cases that all the near-road noises appear on the bottom of the images, we directly filter the pixels with depth values greater than a threshold in that rectangular bottom area of the images. The samples after depth range truncation can be seen in Fig. 1-(c).

Although depth range truncation removes some pixels with too large depth values, there are still misrecontructed pixels of sky, cloud or shadow with normal depth values. We use `PowerToys` from `https://github.com/microsoft/PowerToys` to pick up typical HSV values for further refinement and denoising. As Tab. 3 shows, the minimal and maximal HSV values are given for some typical noises, including sky, cloud, reflections and shadows. For the clear or cloudy sky, Value tends to be high around 200 and Hue is usually blue or white. However, for those areas in the shadow of low sun, Saturation and Value are extremely low to be about 10% so that the depth map pixels are too hard to be correctly reconstructed, which need to be filtered. The samples after HSV refinement are shown in Fig. 1-(d).

Though RANSAC algorithm inside the SfM and MVS pipeline largely removes pixels of the dynamic objects to ensure the accuracy of overall depth values, the dynamic pixels cannot been fully eliminated and the contours of objects are not clear as well. Therefore, we employ MaskRCNN [4] with pretrained models from Detectron2 on `https://github.com/facebookresearch/detectron2`. We adopt the pretrained model with config-

Table 2: Some Important Hyperparameters for COLMAP

Process	Hyperparameter	Value
Sequential SIFT Matching	min_inlier_ratio	0.65
	max_distance	0.55
	min_num_inliers	50
	overlap_num	15
RANSAC	dyn_num_trials_multiplier	3.0
	confidence	0.99
	min_inlier_ratio	0.1
Sparse Reconstucion	abs_pose_min_inlier_ratio	0.25
	filter_max_reproj_error	4.0
	filter_min_tri_angle	1.5
Dense Reconstucion	geom_consistency_max_cost	3.0
	geom_consistency_regularizer	0.3

Table 3: Some Typical Noises and HSV Thresholds

Noise Source and Type	minimal threshold (H, S, V)	maximal threshold (H, S, V)
Blue Sky	(172, 5%, 40%)	(240, 90%, 100%)
White Cloud and Bright Reflections from Windows	(0, 0%, 100%)	(360, 100%, 100%)
Dark and Black Shadows	(0,0%,0%)	(0,0%,0)%
Dusk Cloud and Refections from Roads and Cars	(0,0%,70%)	(90,20%,100%)
Dusk Sky	(140, 11%, 40%)	(160, 50%, 100%)

uration file of `COCO-InstanceSegmentation/mask_rcnn_R50_FPN_3x.yaml` and modify the `MODEL.ROI_HEADS.SCORE_THRESH_TEST` to be `0.5` to find the instance segmentation with the class of `car`, `person` and `bus`. To process the image directly, we modify the visualization part in the official colab notebook, omitting boxes, keypoints and labels and letting $\alpha = 1$ in `draw_polygon` function to set the pixels of the target objects to be black. But semantic or instance segmentation cannot distinguish dynamic objects that need to be removed, we use human annotation to check whether segmented vehicles or pedestrians are moving or not, relabeling the missing dynamic objects and correcting the mislabeled objects. The depth map samples after all the post-processing can be found in Fig. 1-(e).

1.2 Statistics and Analysis of Depth Map for Each Environment

Here we give the statistical analysis of the proposed *SeasonDepth* dataset for each environment. Since all the depth values are scale-free and not absolute for distance, it is not applicable to directly find the pixel value distribution for the dataset as [5, 6] do. However, the depth values of sequential frames in similar urban scenes under the same environment are similarly distributed, *i.e.* the depth of images along the similar streets and blocks are consistent. Then the key point is to align the distribution of each environment to the mean of all environments, obtaining the normalized whole distribution map and dismissing the scale discrepancy.

Therefore, we first find the original depth value distribution $p_{D_i}(x)$ for all the slices under each environment i. Then lower quartile Q_1 (25%), median Q_2 (50%) and upper quartile Q_3 (75%) are calculated for the original distribution of every environment and the mean value of quartiles can be found as reference quartiles$Q_{1_{ref}}, Q_{2_{ref}}, Q_{3_{ref}}$ for all n environments,

$$Q_{1_{ref}} = \frac{1}{n}\sum_{i=1}^{n} Q_{1_i}, Q_{2_{ref}} = \frac{1}{n}\sum_{i=1}^{n} Q_{2_i}, Q_{3_{ref}} = \frac{1}{n}\sum_{i=1}^{n} Q_{3_i}$$

Figure 1: The processing samples given RGB image followed by normalized depth maps for clear visualization of (a) dense reconstruction, (b) range filtering, (c) HSV-based refinement and (d) manual post-processing.

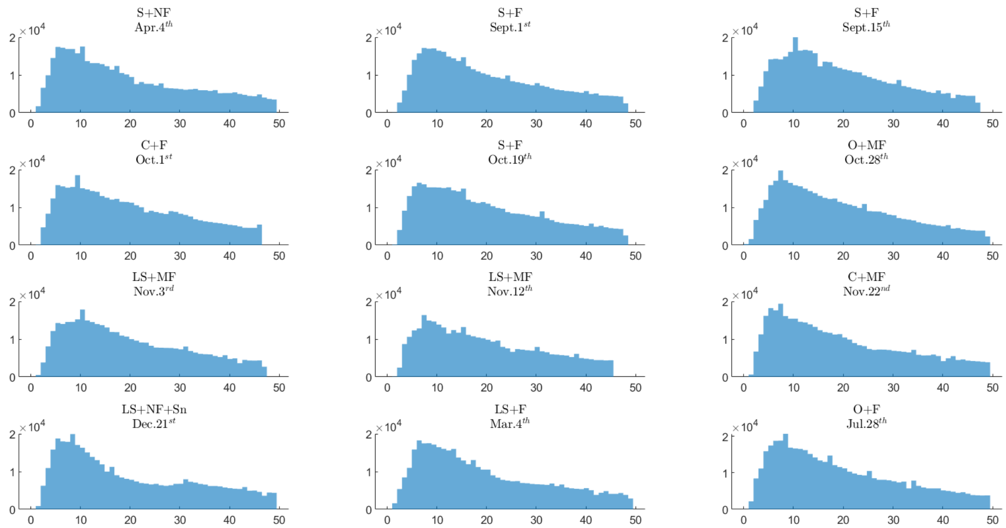


Figure 2: The normalized depth map distribution under all environments. The values of y-axes are number of pixels with the value of abscissa on each image on average.

To find the scale normalization ratio r_i, we use arithmetic mean to measure the ratio of reference quartiles $Q_{1_{ref}}, Q_{2_{ref}}, Q_{3_{ref}}$ and other quartiles $Q_{1_i}, Q_{2_i}, Q_{3_i}$,

$$r_i = \frac{1}{3}\left(\frac{Q_{1_{ref}}}{Q_{1_i}} + \frac{Q_{2_{ref}}}{Q_{2_i}} + \frac{Q_{3_{ref}}}{Q_{3_i}}\right) \tag{1}$$

Then the distribution $p_{D_i}(x)$ can be normalized to mean reference environment to obtain $p_{D\_norm_i}(x)$,

$$p_{D\_norm_i}(x) = r_i p_{D_i}(x) \tag{2}$$

After that, the normalized distribution of all the environments can be added directly to get the whole distribution. The distribution map of each environment can be found in Fig. 2. It can be seen that all the pixels follow similar long-tail distribution and the average y-axis numbers of per-image pixels overcome the bias caused by unbalanced image quantities across different environments. The normalization makes each distribution aligned on the x-axis, which can be directly added to obtain the total distribution map as Fig.3 in the main body paper shows.

2 SeasonDepth Benchmark

2.1 Details about Evaluated Models

For the fairness to evaluate the performance of off-the-shelf depth prediction algorithms under changing environments, we investigate a large amount of depth prediction methods and choose to benchmark the representative and recent state-of-the-art supervised, self-supervised and domain adaptation models from well-known *KITTI* leaderboard [7], which are with open-source codes and pretrained models for fair comparison. Here give important details for all the evaluated baselines. Our experiments are conducted on two NVIDIA 2080Ti cards with 64G RAM on Ubuntu 18.04 system. The evaluation metrics are modified based on `development kit` [7] on `http://www.cvlibs.net/datasets/kitti/eval_depth.php?benchmark=depth_prediction`.

For the supervised methods, we evaluate four representative methods, Eigen *et al.* [8], *BTS* [9], *MegaDepth* [10] *VNL* [11]. Eigen *et al.* propose the first CNNs-based depth prediction method and introduce the famous Eigen split of *KITTI* dataset for depth prediction benchmark. We hence evaluate this representative method through `https://github.com/DhruvJawalkar/Depth-Map-Prediction-from-a-Single-Image-using-a-Multi-Scale-Deep-Network` with the improved image gradient component in the newer loss to see the performance across multiple environments. Recent supervised work *BTS* ranks $4th$ on *KITTI* benchmark and we test it on `https://github.com/cogaplex-bts/bts` using the pretrained model `DenseNet161` on Eigen split. We further fine-tune this pretrained model of *BTS* on our training set of `slice6` using default configuration file for more 65 epochs to get the best performance of $Average$ metric. Note that `focal value` does not influence the experimental results due to the relative scale of the depth metrics. We test *MegaDepth* method according to `https://www.cs.cornell.edu/projects/megadepth/` with the `MegaDepth` pretrained models as described in the paper and all the hyperparameters are set as default. *VNL* are evaluated using `https://github.com/YvanYin/VNL_Monocular_Depth_Prediction` with the pretrained model of `ResNext101_32x4d` backbone and trained on *KITTI* dataset.

For self-supervised methods, we further categorize them and choose baselines respectively, *i.e. Monodepth* [12], *adareg* [13] and *monoResMatch* [14] for stereo geometry based methods, *SfMLearner* [15], *Monodepth2* [16] and *PackNet* [5] for monocular video SfM based methods, and *CC* [17] and *SGDepth* [18] for multi-task learning with monocular SfM unsupervised pipeline.

For stereo geometry based unsupervised methods, *Monodepth* method is evaluated using `https://github.com/OniroAI/MonoDepth-PyTorch`, which is able to reproduce similar results to those in the paper on Eigen split. We test the model of *adareg* from `https://github.com/alexklwong/adareg-monodispnet` pretrained with Eigen split. *monoResMatch* is tested through `https://github.com/fabiotosi92/monoResMatch-Tensorflow` with *KITTI* pretrined model with default hyperparameters.

For sequence SfM based unsupervised methods, we adopt `https://github.com/ClementPinard/SfmLearner-Pytorch` to benchmark *SfMLearner* for better performance than original repo with slight modification. We further fine-tune this pretrained model of *SfMLearner* on our training set of `slice6` using default configuration file for more 35 epochs to get the best performance of $Average$ metric. We use the model of `ResNet18` pretrained on *ImageNet* and fine-tuned on *KITTI* with the resolution of 640×192 to test *PackNet* on `https://github.com/TRI-ML/packnet-sfm`. Similarly, in order to incorporate stereo geometric constraint into the monocular SfM framework, we use the model of `mono+stereo` pretrained on *ImageNet* and *KITTI* with the resolution of 640×192 to evaluate the performance of *Monodepth2* on `https://github.com/nianticlabs/monodepth2`. For the multi-task SfM unsupervised learning methods, *CC* is evaluated with `DispNet, PoseNet, MaskNet and FlowNet` pretrained model on *KITTI* through `https://github.com/anuragranj/cc`. We also test another recent work *SGDepth* on `https://github.com/ifnspaml/SGDepth` with the `full` model of semantic segmentation and depth prediction with the resolution of 640×192.

Since synthetic datasets like *V-KITTI* include multiple environments in spite of existing domain gap, we additionally evaluate the performance of three domain adaptation methods from *KITTI* benchmark, Atapour *et al.* [19], *T2Net* [20] and *GASDA* [21]. We follow the instruction on `https://github.com/atapour/monocularDepth-Inference` to evaluate the method proposed by Atapour *et al.* with the model pretrained on `KITTI` and `DeepGTAV` [22]. *T2Net* is tested on `https://github.com/lyndonzheng/Synthetic2Realistic` with the weakly-supervised pretrained model for outdoor scenes of *KITTI* and *V-KITTI*. We then evaluate the performance of *GASDA* on `https://github.com/sshan-zhao/GASDA` with the model pretrained on *V-KITTI* and *KITTI* using self-supervised stereo geometric information.

2.2 Qualitative Evaluation Results across Environments

In this section, the results of all the evaluated baselines are visualized after adjustment under typical challenging environments, including dark illumination, snowy scene and complex vegetation. See Fig. 3 for more details. From the results of supervised methods, it can be seen that the patterns of predicted depth maps are similar especially for *BTS* [9] and *VNL* [11], where the top and bottom areas are dark while the middle areas are bright due to overfitting, see buildings as examples. But *VNL* [11] shows advantage on depth details (*e.g.* telephone poles and vegetation) in the middle areas which accounts for the best average performance.

Stereo training involved self-supervised methods (including *Monodepth2* [16] and *GASDA* [21]) perform best continuous depth results for the same entity under all environments, *e.g.* depth values of buildings. Monocular video based self-supervised methods do better in distinguishing relative depth from far and near areas, *e.g.* depth values for objects along different directions of roads, especially for multi-task learning ones *CC* [17] and *SGDepth* [18]. Besides, domain adaptation methods still suffer from domain gaps, which shows that synthetic multi-environment image help little to the performance under the real-world changing environments.

References

[1] Torsten Sattler, Will Maddern, Carl Toft, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, et al. Benchmarking 6dof outdoor visual localization in changing conditions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8601–8610, 2018.

[2] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4104–4113, 2016.

[3] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision*, pages 501–518. Springer, 2016.

[4] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[5] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2485–2494, 2020.

[6] 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.

[7] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. In *International Conference on 3D Vision (3DV)*, 2017.

[8] David Eigen, Christian Puhrsch, and Rob 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.

[9] Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326*, 2019.

[10] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2041–2050, 2018.

[11] Wei Yin, Yifan Liu, Chunhua Shen, and Youliang Yan. Enforcing geometric constraints of virtual normal for depth prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5684–5693, 2019.

[12] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 270–279, 2017.

[13] Alex Wong and Stefano Soatto. Bilateral cyclic constraint and adaptive regularization for unsupervised monocular depth prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5644–5653, 2019.

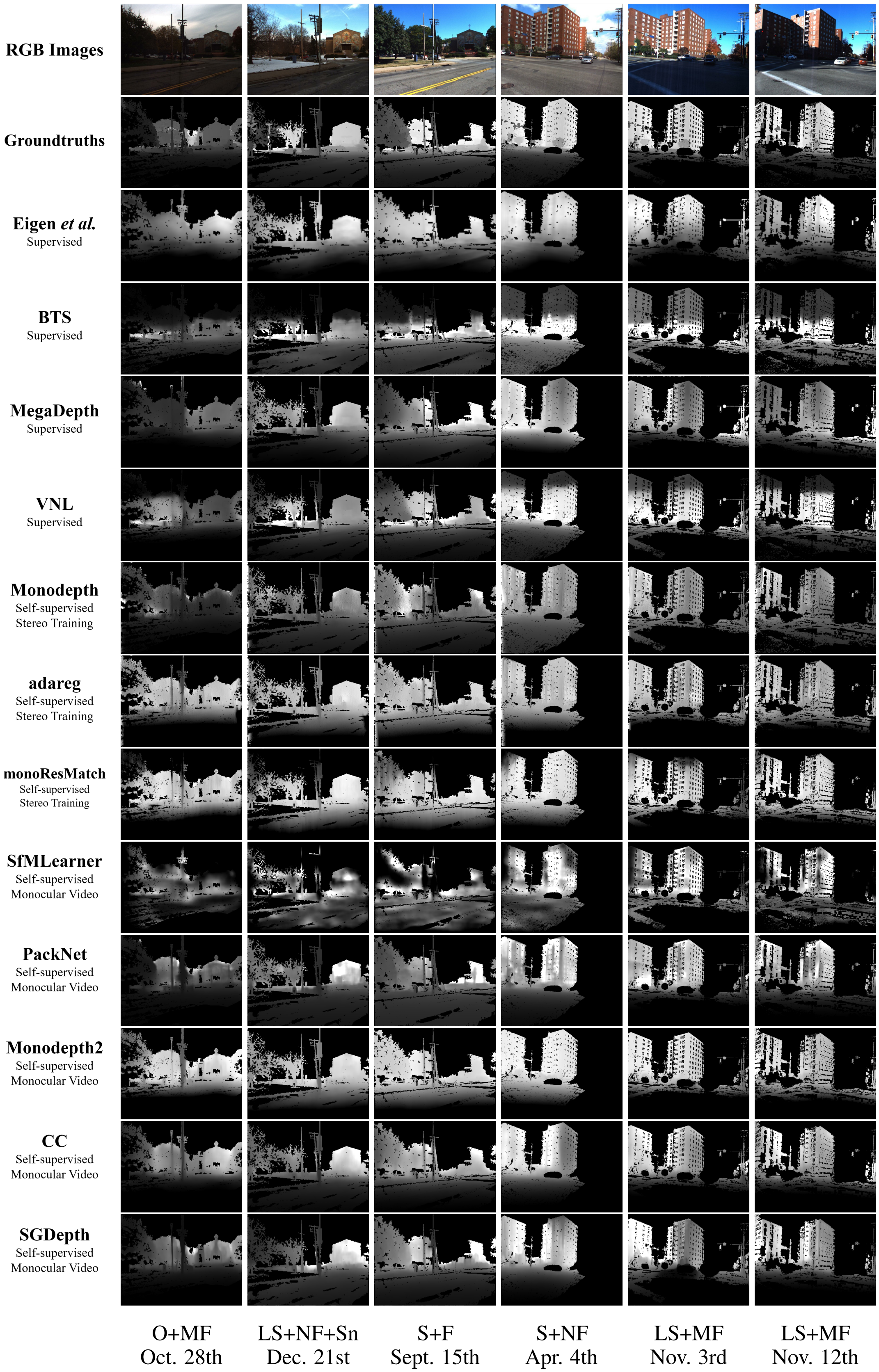
[14] Fabio Tosi, Filippo Aleotti, Matteo Poggi, and Stefano Mattoccia. Learning monocular depth estimation infusing traditional stereo knowledge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9799–9809, 2019.

[15] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1851–1858, 2017.

[16] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3828–3838, 2019.

[17] Anurag Ranjan, Varun Jampani, Lukas Balles, Kihwan Kim, Deqing Sun, Jonas Wulff, and Michael J Black. Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12240–12249, 2019.

[18] Marvin Klingner, Jan-Aike Termöhlen, Jonas Mikolajczyk, and Tim Fingscheidt. Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance. In *European Conference on Computer Vision*, pages 582–600. Springer, 2020.



To be continued

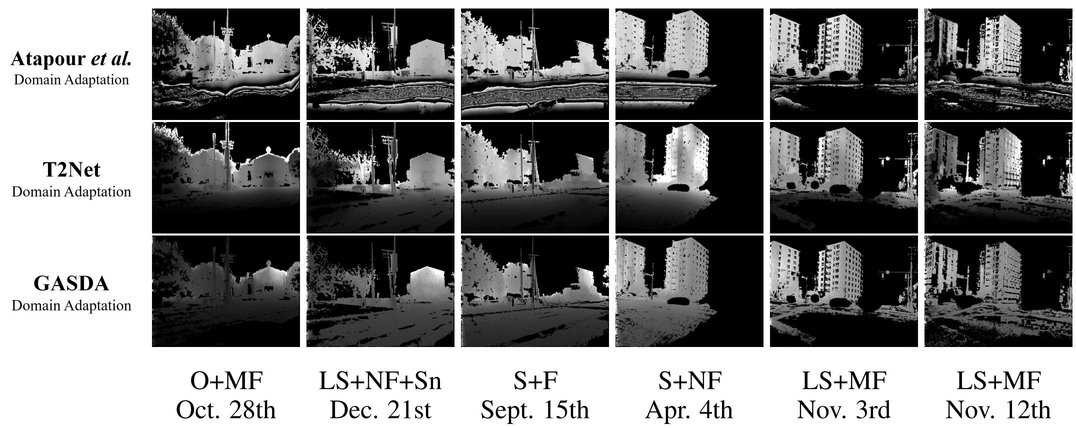


Figure 3: Qualitative results for all the baselines with multiple illumination, vegetation and weather conditions.

[19] Amir Atapour-Abarghouei and Toby P Breckon. Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2800–2810, 2018.

[20] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. T2net: Synthetic-to-realistic translation for solving single-image depth estimation tasks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 767–783, 2018.

[21] Shanshan Zhao, Huan Fu, Mingming Gong, and Dacheng Tao. Geometry-aware symmetric domain adaptation for monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9788–9798, 2019.

[22] Aitor Ruano Miralles. An open-source development environment for self-driving vehicles. `http://openaccess.uoc.edu/webapps/o2/bitstream/10609/63765/6/aruanomTFM0617memory.pdf`, 2017.

[23] Hernán Badino, Daniel Huber, and Takeo Kanade. Visual topometric localization. In *2011 IEEE Intelligent Vehicles Symposium (IV)*, pages 794–799. IEEE, 2011.

[24] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*, 2018.

[25] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[26] Mans Larsson, Erik Stenborg, Lars Hammarstrand, Marc Pollefeys, Torsten Sattler, and Fredrik Kahl. A cross-season correspondence dataset for robust semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9532–9542, 2019.

A Appendix

A.1 Discussion on Societal Impacts

To our best knowledge, we are the first work focusing on the influence of changing environments on monocular depth prediction task, which has great significance to the long-term or lifelong autonomous driving and outdoor mobile robotics. The robustness of depth prediction algorithm is important to the safety of vehicles and pedestrians from the long-run perspective.

However, there are also some potential negative societal impacts. First, our dataset is not that general because the original dataset CMU Visual Localization dataset is only collected in one city, which may mislead the algorithm to overfit on the similar scenes, leading to unstability and risks when used in the complex scenes for applications. Second, privacy is another problem. Although the dataset is secondarily derived and there are many licenses on it, malicious and unintended uses may still happen, *e.g.* collect the human faces or properties of the locals, which may violate the privacy right and cause other problems.

Dismissing such concerns needs the efforts from communities of research, industry and other social organization. Researchers and engineers should fully evaluate the performance and robustness of algorithms with environmental changes despite using our dataset, to make sure the safety of autonomous driving. Social organization should also keep an eye on the use of such open-source real-world dataset to avoid the illegal use.

A.2 Information about SeasonDepth Dataset and Benchmark Toolkit

Our dataset and benchmark can be accessed through `https://seasondepth.github.io/` and the toolkit for evaluation is available on `https://github.com/SeasonDepth/SeasonDepth`. The dataset is placed in long-term preserved `figshare`. The persistent identifier of Digital Object Identifier is `10.6084/m9.figshare.14731323`, which can be accessed on `https://doi.org/10.6084/m9.figshare.14731323` or `https://figshare.com/articles/dataset/SeasonDepth_Cross-Season_Monocular_Depth_Prediction_Dataset/14731323`. The dataset has also adhered to Schema.org and DCAT metadata standards on `https://seasondepth.github.io/` through the tool on `https://www.google.com/webmasters/markup-helper/u/0/`. Due to the fact that SeasonDepth is repackaged from existing dataset, according to the original license, the license of our dataset is BY-NC-SA-4.0 `https://creativecommons.org/licenses/by-nc-sa/4.0/` considering the official announcement to use the latest version `https://creativecommons.org/faq/#why-should-i-use-the-latest-version-of-the-creative-commons-licenses`. Since the metrics are based on *KITTI* benchmark [7], the toolkit and benchmark are under BY-NC-SA-4.0 license `https://creativecommons.org/licenses/by-nc-sa/4.0/` according to `http://www.cvlibs.net/datasets/kitti/` and the official announcement to use the latest version `https://creativecommons.org/faq/#why-should-i-use-the-latest-version-of-the-creative-commons-licenses`.

A.3 Licenses of All Used Assets

The license of original CMU Visual Localization Dataset [23] and CMU-Seasons Dataset [1] have the license of `BY-NC-SA 3.0`. The license of *KITTI* benchmark is `BY-NC-SA 3.0`. Codes and pretrained models of Eigen *et al.*, *BTS*, *MegaDepth*, *VNL*, *Monodepth*, *adareg*, *monoResMatch*, *SfMLearner*, *PackNet*, *Monodepth2*, *CC*, *SGDepth*, Atapour *et al.*, *T2Net* and *GASDA* are under the licenses of `MIT`, `GPL 3.0`, `MIT`, `2-clause BSD`, `UCLB ACP-A`, `MIT`, `MIT`, `MIT`, `MIT`, `Monodepth2 license` on `https://github.com/nianticlabs/monodepth2/blob/master/LICENSE`, `MIT`, `MIT`, `MIT`, `MIT` and `MIT`.

The copyright and terms of service from `https://www.visuallocalization.net/` can be found on `https://www.visuallocalization.net/privacypolicy/` and the license is `Creative Commons Attribution-NonCommercial-ShareAlike 3.0 License`, see `https://creativecommons.org/licenses/by-nc-sa/3.0/`.

A.4 Dataset Documentation and Intended Uses

We adopt datasheet for dataset [24] template from `https://www.overleaf.com/latex/templates/datasheet-for-dataset-template/jgqyyzyprxth`, keeping the original format of the documentation in blue color but with single column for clarity. We authors state that we will bear all responsibility in case of violation of rights and data is under the license of `BY-NC-SA 4.0`. We host our dataset and benchmark through GitHub and GitHub Pages with license and dedicated people are responsible for the long-term maintenance.

Datasheet for SeasonDepth Dataset

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The dataset is created for the purpose of monocular depth prediction under changing environments, which is really essential to robust outdoor long-term visual perception and localization for autonomous driving and mobile robots but has not been studied yet to our best knowledge.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The dataset is created by Hanjiang Hu, Baoquan Yang, Zhijian Qiao, Ding Zhao and Hesheng Wang, where Hanjiang Hu, Baoquan Yang, Zhijian Qiao and Hesheng Wang are with Intelligent Robotics and Machine Vision Lab, Department of Automation, Shanghai Jiao Tong University. Hanjiang Hu and Ding Zhao are with Safe AI Lab at Carnegie Mellon University.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

This work was supported in part by the Natural Science Foundation of China under Grant 62073222 and U1913204, in part by grants from NVIDIA Corporation.

Any other comments? No.

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

SeasonDepth is derived from CMU Visual Localization dataset through structure from motion and the instances include RGB images together with corresponding depth maps as groundtruths, saved in `jpg` and `png` files respectively. Besides, some documents are also involved, including README.md and intrinsic paraeters file.

How many instances are there in total (of each type, if appropriate)?

We totally provide 17725 pairs of RGB image and depth map over 12 different environments for 7 slices from left and right cameras. See detailed numbers of instances in Tab. 1.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset contains RGB images and depth maps with 7 different lefet-right-side urban road sequences along a route of 8.5km in Pittsburgh under 12 different environmental conditions. It is a sample of instances from the larger set which contains RGB images and depth maps of all the outdoor urban road views with all the camera perspectives under all the possible environments from all over the world. It is not fully representative for the whole larger dataset since rainy or night scenes are not involved due to the lack of the original dataset and unavailable. But it covers commonly-seen environments for current outdoor visual perception so we believe it can be representative for the study of depth prediction under changing environments.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each instance consists of one undistorted RGB image saved in `jpg` format.

Is there a label or target associated with each instance? If so, please provide a description.

The label is the normalized depth map saved in `uint16 png` format obtained from SfM for each RGB image instance.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

The information of distortion from original image is missing due to undistortion using intrinsics parameters of the cameras. But we think what we have done is meaningful and the distortion of original image is not what we want and even can bring about noise and error.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

Yes, the relationships between individual instances are make explicit. We provide the timestamp for each image instance which contains information of sequential index, camera index and environmental information so that every two instances can be compared due to the relationship.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

We adopt the categorized slices of Urban part according to [1] with rectification through camera intrinsic file. Specifically, we use slice2, slice3, slice7, slice8 as the split test slices for evaluation and benchmark, and the other slices slice4, slice5, slice6 are intended to treat as training set. The reason lies in that the dataset is priory to benchmark the performance of existing depth prediction methods, so the number of test set is large. For the sake of fine-tuning through small subset of multi-environment images in the future, we also provide the training set in a limited quantity.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

We have filtered most of the noise using RANSAC algorithm, range filtering, HSV refinement and segmentation-based manual post-processing. Although errors and noise are inevitable, we try our best to avoid the influence from them.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

Our dataset is built based on CMU Visual Localization dataset [23] and CMU-Seasons dataset [1]. a) Guarantees and maintenance can be found in their papers. b) Our SeasonDepth dataset has been archived on `https://figshare.com/articles/dataset/SeasonDepth_Cross-Season_Monocular_Depth_Prediction_Dataset/14731323` with DOI of `10.6084/m9.figshare.14731323`. External resources are arhived on `https://www.visuallocalization.net/`. c) The license of original CMU Visual Localization Dataset [23] and CMU-Seasons Dataset [1] have the license of `BY-NC-SA 3.0` with detailed restrictions on `https://creativecommons.org/licenses/by-nc-sa/3.0/`.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.

No.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

No.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Yes, many of the images contain people.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

No, although we adopt instance segmentation for pedestrians, we do not identify the subpopulations and tell how old or what gender they are. What we are only interested in is whether they are moving or not, *i.e.* dynamic objects.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

No, it is not possible to identify individuals from our dataset without any semantic information but only human body identification is possible. For those pedestrians that are static (*e.g.* waiting for traffic light), the corresponding depth values are kept and can be used to get the person body through clustering algorithm of depth values.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

Yes, the dataset contains some information about the cars, people and locations in Pittsburgh. But the detailed locations are not available in our dataset but may be found in the original datasets [23, 1] with GPS or map information.

Any other comments? No

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

The data associated with each instance is actually the depth map and the RGB image. Depth maps are obtained by SfM, and we filtered its noise using RANSAC, range filtering, HSV filtering and mannual post-processing. However, we can not verify it because the outdoor RGB images from the originally collected dataset [23] do not have accurate depth values. But RGB image data was validated and released on `https://www.visuallocalization.net/` by [1].

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

We use off-the-shelf algorithms and softwares like COLMAP [2, 3], OpenCV [25], PowerToys from `https://github.com/microsoft/PowerToys` and Painting App on Windows 10. These procedures are all commonly-used and validated by previous works like [1, 26].

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

Since we use slice2, slice3, slice7, slice8 as the split test slices for evaluation and benchmark, and the other slices slice4, slice5, slice6 are intended to treat as training set as our dataset, we sample the data deterministically from larger set given specific sequences and environments.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

For RGB images, we direct use CMU Visual Localization dataset after undistortion. For depth maps, we hired several undergraduate students for manual post-processing and double check and they were compensated by their honored projects and scores.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The original data was collected in 2010-2011 and the detailed date can be found in Tab. 1. The timeframe does not match the creation timeframe of the data associated with the instances, *i.e.* the creation of depth map and the undistroted RGB images. The timeframe of the depth map is from 2019 to 2021, with many rounds to polish the quality of the dataset.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

No, at least the derived dataset not. But the original datasets [23, 1] may be reviewed by such ethical review borad.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Yes, some images contain people.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

We use the data from original datasets [23, 1], where the images were collected through a car with two cameras in Pittsburgh under different environments.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

We have no idea whether people were notified when the original dataset was collected in 2010-2011, but we believe so and please refer to [23] for more details.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

We have no idea whether people consent to the collection and use of data when the original dataset was collected in 2010-2011, but we believe so and please refer to [23] for more details.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

We have no idea whether people were provided a mechanism to revoke their consent when the original dataset was collected in 2010-2011, but we believe so and please refer to [23] for more details.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

Yes, the potential impact of the dataset has been analyzed and it can befound in Appendix A.1. To our best knowledge, we are the first work focusing on the influence of changing environments on depth prediction tasks, which has great significance to the long-term or lifelong autonomous driving and outdoor mobile robotics. The robustness of depth prediction algorithm is important to the safety of vehicles and pedestrians from long-run perspective. However, there are also some potential negative societal impacts. First, our dataset is not that general because the original dataset CMU Visual Localization dataset is only collected to one city, which may mislead the algorithm to overfit on the single scene, leading to unstable when used in the real complex scenes for applications and causing danger then. Second, privacy is another problem. Although the dataset is secondarily derived and there are many licenses on it, malicious and unintended uses may still happen, *e.g.* collect the human faces or properties of the locals, which may violate the privacy right and cause other problems. Dismissing such concerns need the communities from research, industry and other social organization. Researcher and engineers should clearly evaluate the performance of the robustness across environmental changes despite using our dataset, to make sure the safety of autonomous driving. Social organization should also keep an eye on the use of such open-source real-world dataset to avoid the illegal use.

Any other comments?

No.

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

Yes, we use COLMAP for SfM and dense depth resonctruction, followed by range filtering, HSV refinement and manual post-processing to label the dataset. More details can be found in Sec. 1.1.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

Yes, the "raw" data is from original datasets [23, 1] which can be download through `https://www.visuallocalization.net/datasets/`..

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

Yes, the softwares we used in the data processing are available, including `https://colmap.github.io/`, `https://opencv.org/`, `https://github.com/microsoft/PowerToys` and `https://github.com/facebookresearch/detectron2`.

Any other comments?

No.

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

Yes, we have conducted extensive evaluation using representative baseline methods together with recent open-source state-of-the-art pretrained models on *KITTI* leaderboard [7] to benchmark the perfromance of monocular depth prediction under changing environments. More details can be found in Sec. 2.1.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

Yes, we have listed the evaluated algorithms in Sec. 2.1 with detailed links and model parameters.

What (other) tasks could the dataset be used for?

We can not think of other tasks that the dataset can be used for.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

We have make all the terms of use clear in the REAME.md file in the downloaded zip dataset, users must follow the instructions in it and the toolkit readme file on `https://github.com/SeasonDepth/SeasonDepth`. The harms are totally avoidable if users follow the instructions strictly and there is risk of unfair treatment of individuals or groups.

Are there tasks for which the dataset should not be used? If so, please provide a description.

Although the dataset is secondarily derived and there are many licenses on it, malicious and unintended uses may still happen, *e.g.* collect the human faces or properties of the locals, which may violate the privacy right and cause other problems. Dismissing such concerns need the communities from research, industry and other social organization. Researcher and engineers should clearly evaluate the performance of the robustness across environmental changes despite using our dataset, to make sure the safety of autonomous driving. Social organization should also keep an eye on the use of such open-source real-world dataset to avoid the illegal use.

Any other comments? No.

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

Yes, the dataset is open-source and can be accessed on figshare https://figshare.com/articles/dataset/SeasonDepth_Cross-Season_Monocular_Depth_Prediction_Dataset/14731323. More descriptions can be found on https://seasondepth.github.io/.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub) Does the dataset have a digital object identifier (DOI)?

The dataset is distributed on figshare https://figshare.com/articles/dataset/SeasonDepth_Cross-Season_Monocular_Depth_Prediction_Dataset/14731323 and the toolkit for the benchmark can be found on https://github.com/SeasonDepth/SeasonDepth. The DOI of the dataset is 10.6084/m9.figshare.14731323 which can be accessed through https://doi.org/10.6084/m9.figshare.14731323.

When will the dataset be distributed?

Now our dataset is available on https://doi.org/10.6084/m9.figshare.14731323.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

Due to SeasonDepth is repackaged from existing dataset, according to the original license, the license of our dataset is BY-NC-SA-4.0 https://creativecommons.org/licenses/by-nc-sa/4.0/ considering the official announcement to use the latest version https://creativecommons.org/faq/. Since the metrics are based on *KITTI* benchmark [7], the toolkit and the benchmark is under BY-NC-SA-4.0 license https://creativecommons.org/licenses/by-nc-sa/4.0/ according to http://www.cvlibs.net/datasets/kitti/ and the official announcement to use the latest version on https://creativecommons.org/faq/.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No.

Any other comments?

No.

Maintenance

Who will be supporting/hosting/maintaining the dataset?

Hanjiang Hu and Baoquan Yang will mainly maintain the dataset and other members from Intelligent Robotics and Machine Vision Lab, Department of Automation, Shanghai Jiao Tong University may also be involved in the future for support if necessary.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

Email address: seasondepth@outlook.com . Main maintainer: Hanjiang Hu, Baoquan Yang, Hesheng Wang.

Is there an erratum? If so, please provide a link or other access point.

New updates outside the work will be released on https://seasondepth.github.io/.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

Yes, we will release the training set on `https://seasondepth.github.io/` and toolkit intructions on `https://github.com/SeasonDepth/SeasonDepth` before July by Hanjiang Hu and Baoquan Yang and anyone who watches our work will be noticed through github email.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

For the consideration that we adopt the long-retained dataset from [23, 1], we think we will not set any limits on retention as long as the original dataset on `https://www.visuallocalization.net/datasets/` is still available.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

We will continue support the release version even if we release newer dataset in the future and all the mantenance can be found on `https://seasondepth.github.io/`.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

Yes, we welcome all contribution to our work from others. We plan to use collaboration mechanism on GitHub to do so and their contribution will be shown and validated on our official website`https://seasondepth.github.io/` and any user can view and cite their contribution.

Any other comments?

No.