

**Abstract**— Different environments pose a great challenge to the outdoor robust visual perception for long-term autonomous driving, and the generalization of learning-based algorithms on different environments is still an open problem. Although monocular depth prediction has been well studied recently, few works focus on the robustness of learning-based depth prediction across different environments, e.g., changing illumination and seasons, owing to the lack of such a multi-environment real-world dataset and benchmark. To this end, the cross-season monocular depth prediction dataset and benchmark, SeasonDepth, is introduced to benchmark the depth estimation performance under different environments. We investigate several state-of-the-art representative open-source supervised and self-supervised depth prediction methods using newly-formulated metrics. Through extensive experimental evaluation on the proposed dataset and cross-dataset evaluation with current autonomous driving datasets, the performance and robustness against the influence of multiple environments are analyzed qualitatively and quantitatively. We show that long-term monocular depth prediction is still challenging and believe our work can boost further research on the long-term robustness and generalization for outdoor visual perception. The dataset is available on https://seasondepth.github.io.

I. INTRODUCTION

Perception and localization for autonomous driving and mobile robotics have made significant progress due to the boost of deep neural networks [1] in recent years. However, since the outdoor environmental conditions are changing because of different seasons, weather and daytime [2], [3], the pixel-level appearance is drastically affected, which casts a big challenge for robust long-term visual perception and localization. Monocular depth prediction plays a critical role in long-term visual perception and localization [4], [5] and is also significant to safe applications such as self-driving cars under different environmental conditions. Although some depth prediction datasets [6], [7], [8], [9] include different environments for diversity, it is still not clear what kind of algorithm is more robust to adverse conditions and how they influence depth prediction performance. Besides, the generalization of learning-based depth prediction methods on different weather and illumination effects is still an open problem. Therefore, building a new dataset and benchmark under multiple environments is needed to study this problem systematically. To the best of our knowledge, we are the first to study the generalization of learning-based depth prediction under changing environments, which is essential and significant to both robust machine learning algorithms and practical applications like autonomous driving.

The outdoor high-quality dense depth maps are not easy to obtain using LiDAR or laser scanner projection [10], [11], [8], [12], or stereo matching [6], let alone collections under multiple environments. We adopt Structure from Motion (SfM) and Multi-View Stereo (MVS) pipeline with RANSAC followed by careful manual post-processing to build a scaleless dense depth prediction dataset SeasonDepth with multi-environment traverses based on the urban part of CMU Visual Localization dataset [2], [13]. Some examples in the dataset are shown in Fig. 1.

For the benchmark on the proposed dataset, several statistical metrics are proposed for the experimental evaluation of the representative and state-of-the-art open-source methods from KITTI leaderboard [10], [14]. The typical baselines we choose include supervised [1], [15], [16], [17], [18], [19], stereo training based self-supervised [20], [21], [22], monocular video based self-supervised [23], [24], [25], [26], [27], [28], [29], [30], [31], [32] and domain adaptation [33], [34], [35] algorithms. Through thoroughly analyzing benchmark results, we find that most well-tuned methods cannot present satisfactory performance in terms of both mean and variance under multiple environments. Besides, through cross-dataset evaluations, current KITTI pretrained models cannot generalize well on our dataset while the models tuned on our dataset perform better on KITTI [10] compared to models tuned on Cityscapes [6]. Furthermore, the performance under each adverse environment is investigated both qualitatively and quantitatively to show hints to address robust perception against challenging environments.

For the open problem of generalizability of learning-based depth prediction methods on different environmental conditions, our dataset is the first one that contains real-world RGB images with multiple environments under the same routes so that fair cross-environment evaluation and comparison can be conducted, giving hints to the future research on robust perception in changing environments. In summary, our contributions in this work are listed as follows.

- A new monocular depth prediction dataset SeasonDepth with the same multi-traverse routes under changing environments is introduced through SfM and MVS.
pipeline and is publicly available to the community.

- We benchmark best and representative open-sourced supervised and self-supervised prediction methods on SeasonDepth using several new statistical metrics.
- From the extensive cross-environment and cross-dataset evaluation, we find that long-term robust depth prediction is still challenging and our dataset and benchmark can give future research direction by pointing out how adversary environments affect the performance with some promising hints to enhance robustness.

II. SEASONDEPTH DATASET

Our proposed dataset SeasonDepth is derived from CMU Visual Localization dataset [13] through SfM algorithm. The original CMU Visual Localization dataset covers over one year in Pittsburgh, USA, including 12 different environmental conditions. Images were collected from two identical cameras on the left and right of the vehicle along a route of 8.5 kilometers. And this dataset is also derived for long-term visual localization [2] by calculating the 6-DoF camera pose of images with more appropriate categories about the weather, vegetation and area. To be consistent with the content of driving scenes in other datasets like KITTI, we adopt images from Urban areas categorized in [2] to build our dataset.

A. Dense Reconstruction and Post-processing

We reconstruct the dense model for each traversal under every environmental condition through SfM and MVS pipeline [36], which is commonly used for depth reconstruction [24], [17] and most suitable for multi-environment dense reconstruction for 3D mapping [37], [2] and show advantage on the aspects of high dense quality despite of huge computational efforts compared to active sensing from LiDAR. Specifically, similar to MegaDepth [17], COLMAP [36] is used to obtain the depth maps through photometric and geometric consistency from sequential images. Furthermore, RANSAC algorithm is used in the SfM to remove the inaccurate values of dynamic objects in the images, where dynamic objects with additional motion besides relative camera motion do not obey the multi-view geometry constraint and should be removed as noise via RANSAC in bundle adjustment optimization.

Besides, from our justification experiments in Sec. IV-C, it is validated that using relative depth values and removing dynamic noise does not significantly influence the training and the performance of depth prediction models. Because the MVS algorithm generates the depth maps with error pixel values that are out of range or too close, like the cloud in the sky or noisy points on a very near road, we manually filter those outliers of the depth map.

After the reconstruction, based on the observation of noise distribution in the HSV color space, e.g. blue pixels always appear in the sky and dark pixels always appear in the shade of the low sun, which tend to be noise in most cases, we remove the noisy values in the HSV color space given some specific thresholds. Though outliers are set to be empty in RANSAC, instance segmentation is adopted through MaskRCNN [38] to fully remove the noise of dynamic objects. However, since it is difficult to generate accurate segmentation maps only for dynamic objects under drastically changing environments, we leverage human annotation as the last step to finally check the depth map. Since we are rigorous and serious to the quality of valid depth pixels which are used for benchmark, we set most noise to be invalid (which causes some “holes” on the boundary from appearance) to avoid any possible pollution to the following benchmark, ensuring the reliable evaluation and benchmark results.

B. Comparison with Other Datasets

The distinctive feature of the proposed dataset is that SeasonDepth contains comprehensive outdoor real-world multi-environment sequences with repeated scenes, just like virtual synthetic datasets [39], [40], [41], [42] but they are rendered from computer graphics and suffer from the huge domain gap. Though real-word datasets [8], [7], [6] include different environments, they lack the same-route traverses under different conditions so they are not able to fairly evaluate the performance across changing environments. Similar to outdoor datasets [43], [17], the depth maps of ours are scaleless with relative depth values, where the metrics should be designed for evaluation as the following section shows. The depth map ground truth from SfM and MVS is semi-dense compared to LiDAR-based sparse depth maps.
III. BENCHMARK SETUP

A. Evaluation Metrics

The challenge for the design of evaluation metrics lies in two folds. One is to cope with scaleless and partially-valid dense depth map ground truth, and the other is to fully measure the depth prediction average performance and the stability or robustness across different environments. Due to the scaleless ground truth of relative depth value, some common metrics [14] cannot be used for evaluation directly. Since the focal lengths of two cameras are close enough to generate similarly-distributed depth values, unlike [23], [17], we align the distribution of depth prediction to depth ground truth via mean value and variance for a fair evaluation. The other key point for multi-environment evaluation lies in the reflection of robustness to changing environments for same-route sequences, which has not been studied in the previous work to the best of our knowledge. We formulate our metrics below.

First, for each pair of predicted and ground truth depth maps, the valid pixels $D_{valid\_predicted}^{ij}$ of the predicted depth map $D_{valid\_predicted}$ are determined by non-empty valid pixels $D_{valid\_GT}^{ij}$ of the depth map ground truth. And then the valid mean and variance of both $D_{valid\_GT}$ and $D_{valid\_predicted}$ are calculated as $Avg_{GT}, Avg_{pred}$ and $Var_{GT}, Var_{pred}$. Then we adjust the predicted depth map $D_{adj}$ to get the same distribution with $D_{valid\_GT}$:

$$D_{adj} = (D_{pred} - Avg_{pred}) \times \sqrt{Var_{GT}/Var_{pred}} + Avg_{GT}$$

Denote the adjusted valid depth prediction $D_{adj}$ as $D_{P}$ in the following formulation. To measure the depth prediction performance, we choose the most distinguishing metrics under multiple environments from commonly-used metrics in [14], $AbsRel$ and $\delta < 1.25$ ($a_1$). For environment $k$, we have,

$$AbsRel^{k} = \frac{1}{n} \sum_{i,j} |D_{P}^{k}_{i,j} - D_{GT}^{k}_{i,j}| / D_{GT}^{k}_{i,j}$$

$$a_1^{k} = \frac{1}{n} \sum_{i,j} \mathbb{1}(\max\{\frac{D_{P}^{k}_{i,j}}{D_{GT}^{k}_{i,j}}, \frac{D_{GT}^{k}_{i,j}}{D_{P}^{k}_{i,j}}\} < 1.25)$$

For the evaluation under different environments, 6 secondary metrics are derived based on original metrics,

$$AbsRel^{avg} = \frac{1}{m} \sum_{k} AbsRel^{k}, a_1^{avg} = \frac{1}{m} \sum_{k} a_1^{k}$$

$$AbsRel^{var} = \frac{1}{m} \sum_{k} \left| AbsRel^{k} - \frac{1}{m} \sum_{k} AbsRel^{k} \right|^2$$

$$a_1^{var} = \frac{1}{m} \sum_{k} \left| a_1^{k} - \frac{1}{m} \sum_{k} a_1^{k} \right|^2$$

where $avg$ terms $AbsRel^{avg}$, $a_1^{avg}$ and $var$ terms $AbsRel^{var}$, $a_1^{var}$ come from Mean and Variance in statistics, indicating the average performance and the fluctuation around the mean value across multiple environments.

Considering the depth prediction applications, it should be more rigorous to prevent better results fluctuation than worse results under changing conditions. Therefore, we use the Relative Range terms $AbsRel^{relRng}$, $a1^{relRng}$ to calculate the relative difference of maximum and minimum for all the environments.

$$AbsRel^{relRng} = \frac{\max\{AbsRel^{k}\} - \min\{AbsRel^{k}\}}{\frac{1}{m} \sum_{k} AbsRel^{k}}$$

$$a1^{relRng} = \frac{\max\{1 - a_1^{k}\} - \min\{1 - a_1^{k}\}}{\frac{1}{m} \sum_{k} (1 - a_1^{k})}$$

Relative Range terms for $AbsRel$ and $1 - a_1$ are more strict than the Variance terms $AbsRel^{var}$, $a1^{var}$ and note that $1 - a_1$ instead of $a_1$ is used to calculate $a1^{relRng}$ to make relative range fluctuation more distinguishable for better methods.

B. Benchmark Design and Algorithms

In the experiment, we aim to first benchmark the well-tuned performance on SeasonDepth using state-of-the-art algorithms and then present the cross-dataset performance with other datasets using representative baselines of each category.

We first split the split training set, validation set, and test set with 11407, 17225 and 3944 images respectively. Note that the detailed analysis for each environment is based on the validation set which requires more images. For the benchmark on SeasonDepth, though there is no limit to other datasets or pre-trained models to obtain the best performance, since SeasonDepth only has monocular images as the training set, we categorize the state-of-the-art evaluated algorithms as supervised methods and self-supervised methods with monocular video training. Specifically, DepthFormer [18], BTS [15] are DPT [19] are supervised baselines, while SUB-Depth [28], VADepth [30], Monodepth2 [25], SFMLearner [23] and ManyDepth [29] are self-supervised baselines.

For the cross-dataset performance with other datasets, we choose the other two popular autonomous driving datasets KITTI and Cityscapes together with SeasonDepth. To analyze the performance under each environment, we report the results on the validation set of SeasonDepth. We first present generalization performance from KITTI to SeasonDepth. Some representative baseline models on KITTI leaderboard [14] are chosen to evaluate the performance on the SeasonDepth dataset without fine-tuning. These methods include supervised methods (Eigen et al. [1], BTS [15], MegaDepth [17] and VNL [16]), self-supervised methods with stereo training (Monodepth [20], aadereg [21], monoResMatch [22]), self-supervised methods with monocular video training (SFMLearner [23], Monodepth2 [25], PackNet [24], CC [26], SGDepth [27], FSRE-Depth [31] CADepth-Net [32] VADepth [30]), and domain adaptation methods (Atapour et al. [33], T2Net [34], GASDA [35]) trained on the virtual dataset with multiple environments.

We then introduce cross-dataset comparison evaluation to justify that the depth accuracy and the ground truth are eligible for the dataset usage of autonomous driving for model
training in spite of the lack of dynamic objects. Specifically, inspired by cross-dataset transfer degradation evaluation in [7], we compare our dataset with the stereo depth dataset Cityscapes [6] in terms of the degraded performance on KITTI dataset after cross-dataset fine-tuning. Based on the pre-trained models on KITTI, we fine-tune BTS [15] and SfMLearner [23] models on SeasonDepth and Cityscapes dataset with the same amount of images for 50 epochs, and evaluate the depth prediction on KITTI validation set using the metrics of MAE, absErrorRel, iMAE, iRMSE, sqErrorRel from [14] and report the mean and standard deviation from the last 10 training epochs.

IV. EXPERIMENTAL EVALUATION RESULTS

We present the main results of our experimental evaluation. We refer readers to the pre-print version [44] for more details.

A. SeasonDepth Benchmark Results

In this section, we present the evaluation results on the test set of SeasonDepth in Tab. I. The models are well tuned on SeasonDepth training set and have impressive performance on the test set, especially for Average performance. We can see that self-supervised methods do not perform worse than supervised ones after well-tuning. It can be found that DepthFormer [18] and SUB-Depth[28] perform the best on Average but not that well on Variance or RelativeRange, showing that even the well-tuned models cannot perform well consistently across different environments. Therefore, there is still a long way to go even for the state-of-the-art methods towards long-term robust depth estimation.

B. Cross-dataset Generalization Results

In this section, we show the generalization performance from KITTI to SeasonDepth in Tab. II. First we can see that in the zero-shot cross-dataset generalization setting, self-supervised methods show more robustness to different environments than supervised ones, which suffer from large values of Variance and RelativeRange and more sensitive. Also, the gap between KITTI results and SeasonDepth Average results is clear, showing that the generalization without fine-tuning is challenging especially in different environments. Interestingly, supervised methods with good Variance performance are not consistent with those with good Average performance, which indicates that algorithms tend to work well in specific environments instead of being robust to all conditions, validating the significance of the cross-environment study with SeasonDepth dataset.

### TABLE I

**SeasonDepth Benchmark** (↓: LOWER BETTER, ↑: HIGHER BETTER, BEST FOR EACH CATEGORY)

| Category       | Method       | Average Variance (10−2) | Relative Range |
|----------------|--------------|-------------------------|----------------|
|                |              | AbsRel ↓ | a1 ↑ | AbsRel ↓ | a1 ↑ | AbsRel ↓ | a1 ↓ | 1 − a1 ↓ |
| Supervised     | DepthFormer [18] | 0.135      | 0.835 | 0.0210 | 0.120 | 0.294 | 0.576 |
|                | BTS [15]     | 0.242      | 0.587 | 0.0222 | 0.0632 | 0.220 | 0.220 |
|                | DPT [19]     | 0.152      | 0.790 | 0.0286 | 0.1574 | 0.364 | 0.637 |
| Self-supervised| SUB-Depth [28] | 0.095      | 0.920 | 0.008  | 0.015  | 0.398 | 0.668 |
| Monocular      | VADepth [30] | 0.131      | 0.852 | 0.006  | 0.024  | 0.247 | 0.397 |
| Video Training | Monodepth2 [25] | 0.144      | 0.824 | 0.011  | 0.046  | 0.305 | 0.502 |
|                | SfMLearner [23] | 0.325      | 0.482 | 0.107  | 0.155  | 0.298 | 0.236 |
|                | ManyDepth [29] | 0.227      | 0.649 | 0.080  | 0.262  | 0.486 | 0.549 |

### TABLE II

**Cross-dataset Generalization from KITTI to SeasonDepth** (↓: LOWER BETTER, ↑: HIGHER BETTER, BEST FOR EACH CATEGORY)

| Category       | Method          | KITTI Eigen Split Average Variance (10−2) | Relative Range |
|----------------|-----------------|------------------------------------------|----------------|
|                |                 | AbsRel ↓ | a1 ↑ | AbsRel ↓ | a1 ↑ | AbsRel ↓ | a1 ↓ | 1 − a1 ↓ |
| Supervised     | Eigen et al. [1] | 0.203 | 0.702 | 1.093 | 0.340 | 0.346 | 0.0170 | 0.206 | 0.0746 |
|                | BTS [15]        | 0.060 | 0.955 | 0.676 | 0.209 | 0.545 | 0.0650 | 0.405 | 0.129  |
|                | MegaDepth [17]  | 0.220 | 0.632 | 0.515 | 0.417 | 0.0874 | 0.0285 | 0.200 | 0.107  |
|                | VNL [16]        | 0.072 | 0.938 | 0.306 | 0.527 | 0.126 | 0.166 | 0.400 | 0.290  |
| Self-supervised| Monodepth [20]  | 0.148 | 0.803 | 0.436 | 0.455 | 0.0475 | 0.0213 | 0.198 | 0.104  |
| Stereo Training| adareg [21]     | 0.126 | 0.840 | 0.507 | 0.405 | 0.0630 | 0.0474 | 0.178 | 0.0137 |
|                | Monodepth [22]  | 0.096 | 0.890 | 0.487 | 0.389 | 0.286 | 0.0871 | 0.414 | 0.160  |
| Self-supervised| Monocular       |          |        |        |        |        |        |        |        |
| Monocular      | SfMLearner [23] | 0.181 | 0.733 | 0.360 | 0.495 | 0.0801 | 0.0628 | 0.269 | 0.182  |
| Video Training | PackNet [24]    | 0.116 | 0.865 | 0.722 | 0.421 | 0.187 | 0.0705 | 0.186 | 0.155  |
|                | Monodepth2 [25] | 0.106 | 0.874 | 0.256 | 0.624 | 0.0311 | 0.0532 | 0.235 | 0.229  |
|                | CC [26]         | 0.140 | 0.826 | 0.648 | 0.479 | 0.223 | 0.0881 | 0.280 | 0.241  |
|                | SfMLearner [27] | 0.113 | 0.879 | 0.648 | 0.480 | 0.0987 | 0.0498 | 0.197 | 0.169  |
|                | FSRE-Depth [31] | 0.105 | 0.886 | 0.256 | 0.624 | 0.0288 | 0.0283 | 0.227 | 0.158  |
|                | CADDepth-Net [32] | 0.105 | 0.892 | 0.257 | 0.625 | 0.0447 | 0.0725 | 0.265 | 0.278  |
|                | VADepth [30]    | 0.104 | 0.892 | 0.230 | 0.667 | 0.0158 | 0.0215 | 0.205 | 0.179  |
| Syn-to-real    | Atapour et al. [33] | 0.110 | 0.923 | 0.687 | 0.300 | 0.224 | 0.0220 | 0.231 | 0.0622 |
| Domain         | T2Net [34]      | 0.169 | 0.769 | 0.827 | 0.391 | 0.399 | 0.0799 | 0.286 | 0.146  |
| Adaptation     | GASDA [35]      | 0.143 | 0.836 | 0.438 | 0.411 | 0.121 | 0.0665 | 0.271 | 0.145  |

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
Fig. 2. Results on SeasonDepth dataset under 12 different environments with dates. The shadows indicate error bars around mean values with $0.2 \times \text{Standard Deviation}$ for more clarity.

![RGB, Ground Truth, VNL [16], adareg [21], Monodepth2 [25]](image)

Fig. 3. Qualitative comparison results with illumination or vegetation changes. The conditions from top to down are C+MF, Nov. 22, LS+MF, Nov. 3, C+MF, Nov. 22, and C+F, Oct. 1. Green blocks indicate good performance while red blocks are for bad results.

C. Influence of Challenging Environments

In this section, we further investigate which environment is more difficult to the current depth prediction models. The abbreviations of environments in Fig. 2 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. From Fig. 2, we can see that dusk scenes in LS+MF, Nov. 3 and snowy scenes in LS+NF+Sn, Dec. 21 pose great challenge for most algorithms, which points out directions for future research and safe applications. Besides, the consistent error bar in Fig. 2 shows such adverse environments always result in large deviations for all algorithms.

Under these adverse environmental conditions, promising algorithms can also be found. For the dusk or snowy scenes, some domain adaptation methods [33], [34] present impressive robustness under adverse scenes due to the various appearances of synthetic images. For the snowy scenes, self-supervised models are less influenced compared to supervised methods. Qualitative experimental results in Fig. 3 show how extreme illumination or vegetation changes affect depth prediction. From the top two rows, it can be seen that the illumination change of low sun makes the depth prediction of tree trunks less clear under the same vegetation condition as green and red blocks show. Also, no foliage tends to make telephone poles and tree trunks less distinguishable by comparing red and green blocks from the last two rows, while the depth prediction of heavy vegetation is difficult as red blocks show on the fourth row given the same illumination and weather condition.

V. Conclusion

In this paper, a new dataset SeasonDepth is built for monocular depth prediction under different environments, and supervised and self-supervised state-of-the-art open-source algorithms are evaluated. From the experimental results, we find that there is still a long way to go to achieve robustness for long-term depth prediction and several promising avenues are given, pointing out self-supervised methods are more robust to changing environments. Through studying how adverse environments influence, our findings via this dataset and benchmark will impact the research on long-term robust perception and related applications.

REFERENCES

[1] D. Eigen, C. Puhrsch, and R. Fergus, “Depth map prediction from a single image using a multi-scale deep network,” in Advances in neural information processing systems, pp. 2366–2374, 2014.
[2] T. Sattler, W. Maddern, C. Toft, A. Torii, L. Hammarstrand, E. Stenborg, D. Safari, M. Okutomi, M. Pollefeys, J. Sivic, et al., “Benchmarking 6dof outdoor visual localization in changing conditions,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8601–8610, 2018. https://www.visuallocalization.net/.
[3] H. Hu, H. Wang, Z. Liu, C. Yang, W. Chen, and L. Xie, “Retrieval-based localization based on domain-invariant feature learning under changing environments,” in 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2019.
[4] H. Hu, Z. Qiao, M. Cheng, Z. Liu, and H. Wang, “Dasgill: Domain adaptation for semantic and geometric-aware image-based localization,” IEEE Transactions on Image Processing, vol. 30, pp. 1342–1353, 2020.
[5] H. Hu, H. Wang, Z. Liu, and W. Chen, “Domain-invariant similarity activation map contrastive learning for retrieval-based long-term visual localization,” IEEE/CAA Journal of Automatica Sinica, vol. 9, no. 2, pp. 313–328, 2021.
