A benchmark with decomposed distribution shifts for 360 monocular depth estimation

Anonymous Author(s)
Affiliation
Address
email

Abstract

In this work we contribute a distribution shift benchmark for a computer vision task; monocular depth estimation. Our differentiation is the decomposition of the wider distribution shift of uncontrolled testing on in-the-wild data to three distinct distribution shifts. Specifically, we generate data via synthesis and analyze them to produce covariate (color input), prior (depth output) and concept (their relationship) distribution shifts. We also synthesize combinations and show how each one is indeed a different challenge to address, as stacking them produces increased performance drops and cannot be addressed horizontally using standard approaches.

1 Introduction

Data-driven methods are conditioned on the data which are available during the model development time but are to be applied to in-the-wild real world data. Considering that the former data distribution is $P_s = \sigma(P_r)$, which is the result of applying a sampling function $\sigma(\cdot)$ to the real world distribution P_r. Typically, P_s is separated into different splits P_{trn} and P_{val}, used to train the model and validate its behaviour respectively, with the latter process driving model selection. A data distribution shift can be described as the condition where the joint distribution P of inputs and outputs differs between the training and test stages [1].

This is an actual problem that many practical applications face, affecting their overall performance, robustness and reliability. This phenomenon is more prominent in tasks where annotated data collection is difficult. This problem has been generally addressed in the literature as the domain shift [2] or the generalization [3] of data-driven models, or otherwise as out-of-distribution robustness [4]. More information about out-of-distribution (OOD) learning and generalization can be found in recent surveys [5, 6].

Up to now most works approach this problem in its general setting via zero-shot cross dataset transfer experiments that aim at assessing model performance under a general distribution shift considering two different samplings $P_{s^1} = \sigma_1(P_r)$ and $P_{s^2} = \sigma_2(P_r)$. A recent benchmark [7] provided simultaneously data for subpopulation shift, a special case of distribution shifts, as well as a generic domain generalization shift across a number of datasets and tasks.

In this work, we contribute a novel benchmark for distribution shift performance assessment for a computer vision task notorious for its complex data collection processes, monocular depth estimation. The novelty of our benchmark is the decomposition of distribution shifts into different testing splits that offer distribution shifts of various types either in isolated settings, or combined.

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

2 The Shifted Panoramas Dataset

Our benchmark relies on two recent 3D building scan datasets, Matterport3D (M3D) [8] and GibsonV2 (GV2) [9], as well as modern synthesis approaches, to produce high quality spherical panoramas coupled with spherical depth maps. Specifically, we use M3D as a traditional in-distribution model development dataset, and GV2 as a zero-shot cross dataset transfer and out-of-distribution benchmark dataset.

For M3D we consider its standard partitioning into train P_{trn}^{M3D}, validation P_{val}^{M3D}, and test P_{tst}^{M3D} splits. The GV2 splits represent another sampling of the real world domain, or otherwise a zero-shot cross dataset transfer experiment. Nonetheless, GV2 itself is partitioned into different splits, the *tiny* P_{tiny}^{GV2}, *medium* P_{med}^{GV2}, *full* P_{full}^{GV2} and *fullplus* P_{full+}^{GV2} splits[1]. After synthesizing coupled color and depth panoramas for all splits of both datasets, and analyzing the different splits we observe that it is possible to decompose them into three core distribution shifts.

A **covariate distribution shift** represents a shift of the input domain, which in our case is the color image's domain. As we rely on a synthesis (*i.e.* raytracing) approach to generate our data, we are also in control of the camera color transfer function. Consequently, we generate a shifted input distribution P_{cov} using the M3D test split P_{tst}^{M3D}, where only the color domain has been shifted, illustrated in the first row of Figure 1.

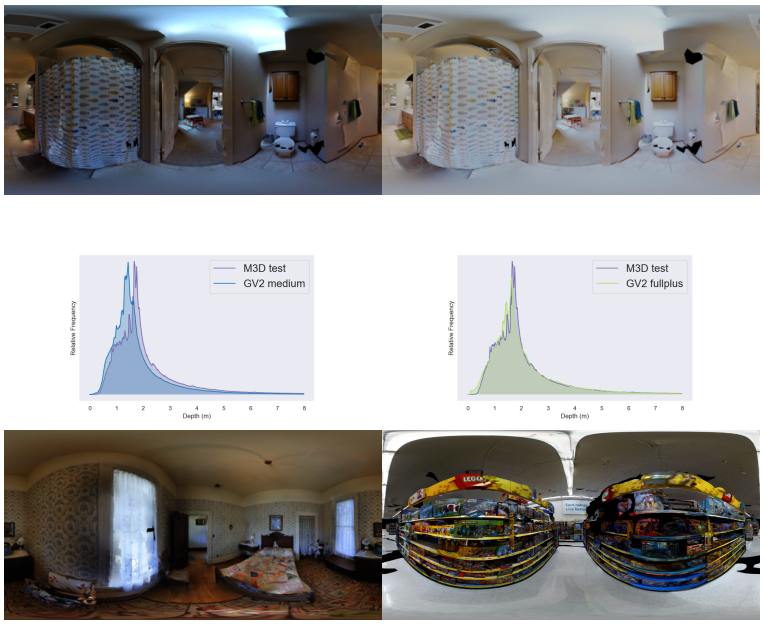

Figure 1: The first row depicts the covariate distribution shift,while the second one the differences related to the depth distribution values.The third row depicts the difference regarding the context of the scenes(i.e. residential vs industrial).

After examining the different splits' statistics we also observe a **prior probability distribution shift** manifesting at the tiny, P_{tiny}^{GV2}, and medium, P_{med}^{GV2}, splits which corresponds to P_{prior}, meaning that the output depth distribution has shifted from the training one P_{trn}^{M3D}. The different depth distributions between these splits are presented in Figure 1 in the second row. Yet, the input (color) distribution is similar as the color camera transfer function is the same, and the context is also preserved to residential scenes.

Finally, analysing *fullplus* split, we observe a **concept distribution shift**, which is the shifted context of the depicted scenes as evident in the third row of Figure 1. While Matterport3D (*i.e.* P_{trn}^{M3D}) only contains indoor residential scenes, the *fullplus* split P_{full+}^{GV2} presents varying scenes like supermarkets, garages, under construction buildings, etc., corresponding to P_{conc}. At the same time though, the

[1]For the remainder of the document we ignore the *full* split, which is kept aside for future training purposes.

input (color) and output (depth) distributions are preserved between P^{M3D}_{trn} and P_{conc}. Notably, our benchmark decomposes the wider domain shift into three distinct distribution shifts. But since we rely on synthesis processes, it is straightforward to combine distribution shifts, producing P^{cov}_{prior} and P^{cov}_{conc} by re-rendering the corresponding splits with a shifted color transfer function, essentially adding a covariant shift to the prior and concept ones. This provides two extra combined distribution shift splits, with only the simultaneous prior and concept shifts missing. All our data are publicly available at http://anonymized.url.

3 Analysis

We support our benchmark with a set of zero-cross dataset transfer experiments across the different distribution shifts. We use a standard UNet [10] architecture training a supervised model with a complex objective:

$$\mathcal{L} = \lambda_{L1}\mathcal{L}_{L1} + \lambda_{cos}\mathcal{L}_{cos} + \lambda_{grad}\mathcal{L}_{grad} + \lambda_{vnl}\mathcal{L}_{vnl}, \tag{1}$$

where $\mathcal{L}_{L1}$ is an L1 loss, $\mathcal{L}_{cos}$ is an angular loss defined on the surface normals, $\mathcal{L}_{grad}$ is the multi-scale gradient matching loss from [11], and $\mathcal{L}_{vnl}$ is the virtual normal loss [12]. All the independent term weights are equally weighted, *i.e.* $\lambda_i = 1.0 \, \forall \, i$ We initialize our model using [13] and optimize it using a batch size of 4 and the Adam optimizer [14], using a learning rate of 0.0002 and its default momentum parameter values.

When training we only use P^{M3D}_{trn} and for all experiments we calculate standard metrics for depth estimation [15], as well as boundary [16, 17] and normals RMSE and accuracies [18]. We aggregate performance across the different traits (direct depth, boundary and smoothness) them using a set of indicators:

$$i_d = ((1 - \delta_{1.25})RMSE)^{-1}, \tag{2}$$

$$i_b = ((1 - \frac{(F_{0.25} + F_{0.5} + F_{1.0})}{3})dbe^{acc})^{-1}, \tag{3}$$

$$i_s = ((1 - \frac{(\alpha_{11.25^o} + \alpha_{22.5^o} + \alpha_{30^o})}{3})RMSE^o)^{-1}, \tag{4}$$

where $RMSE$ and $RMSE^o$ are the depth and normal angular errors respectively, dbe^{acc} is the accuracy boundary error from [16], F_t is the F1-score for different edge thresholds from [17], and δ_d and α_a are the accuracy under thresholds for the depth and surface normals from [15] and [18] respectively.

Through these indicators we present an holistic view of how task performance is affected from the different distribution shifts. In the following subsections, we examine isolated distribution shifts as well as some of their combinations.

Decomposed Shifts: Varying the input, output and combined domains. After training a supervised model on M3D's train split, we examine its performance on the different distribution shifts we have generated compared to that of the in-distribution test set. Figure 2 illustrates the results using the indicators from Eq. (2). We observe a performance drop for all distribution shifts, with the covariate (magenta box) and prior (orange box) being at about the same level, while the concept shift (cyan box) presents the largest performance loss. At the same time, combining two distribution shifts hurts performance even more, as shown by the combined distribution shifts (violet box).

Photometric augmentations for covariate shift. Next, we examine the effect of photometric augmentations (*i.e.* brightness, contrast, hue shifts, and gamma corrections) on the different distribution shifts, in isolation and in combination. Figure 3 presents the results of a model trained with such augmentations when tested on the our benchmarks different splits. Even though the fact that photometric augmentations address camera domain shifts, or color transfer function shifts is already acknowledged in the literature, through these experiments where the performance gains are only observed in the splits where covariate shifts manifests, we verify the decomposition of our benchmark's shifts.

Pretraining for generalization boost. It is also usually acknowledged that pretraining on large image datasets like ImageNet [19] usually helps with the domain shift. We perform another experiment, using the PNAS model [20], with all hyperparameters preserved, and train a model initialized using

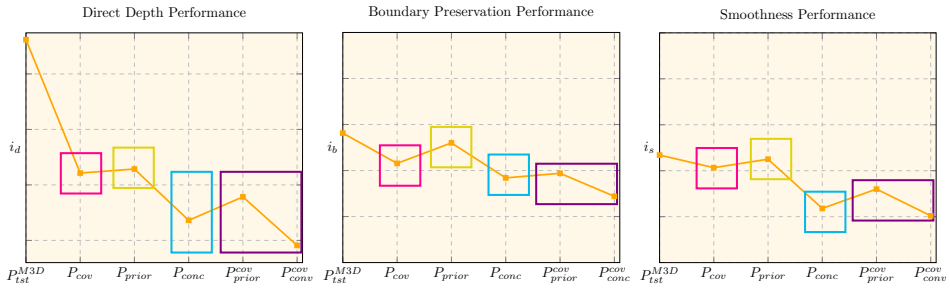

Figure 2: The effect of each distribution shift and their combinations.

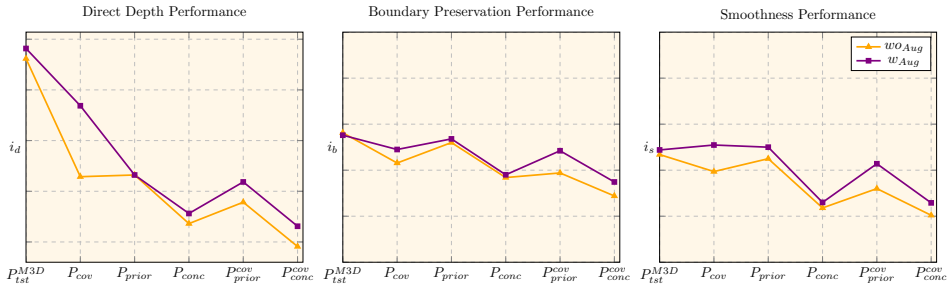

Figure 3: Photometric augmentations effect on the different distribution shifts.

pretrained weights, and one initialized using [13]. Figure 4 presents the results of these experiments when tested on our benchmark's different splits and distribution shifts. Interestingly, we observe a performance boost in the splits where only a single distribution shift is present, where in contrast, the ones with two simultaneous distribution shifts show minimal gains. This indicates that pretraining does not necessarily offer generalization – in the form of more transferable initial features – but instead only helps as a good parameter initializer, compared to random initializations.

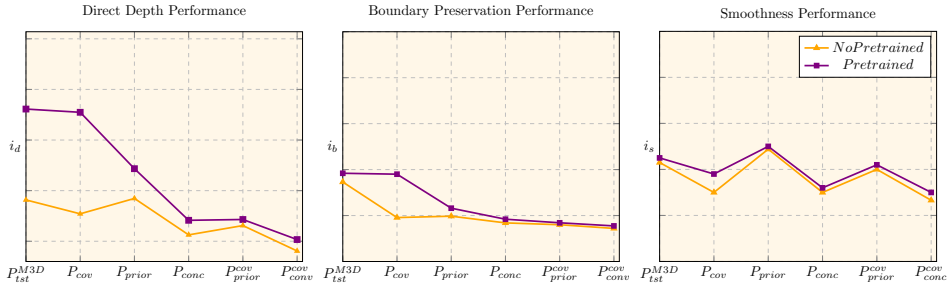

Figure 4: The effect of pretraining on our benchmarks different distribution shifts.

4 Conclusion

Distribution shifts are very important to the application of data-driven methods. In this work, we contribute a distribution shift benchmark for an ill-posed dense computer vision task with difficult data collection processes. Seeking to facilitate future research towards addressing this challenging problem, we decompose distribution shifts to input (covariate), output (prior) and their relationship (concept), and provide a baseline to build upon. The baselines are accompanied by an analysis for the simpler covariate shift case, using photometric augmentations, as well as the general case using pretrained models.

References

[1] Joaquin Quiñonero-Candela, Masashi Sugiyama, Neil D Lawrence, and Anton Schwaighofer. *Dataset shift in machine learning*. Mit Press, 2009.

[2] Karin Stacke, Gabriel Eilertsen, Jonas Unger, and Claes Lundström. A closer look at domain shift for deep learning in histopathology. In *COMPAY19: 2nd MICCAI workshop on Computational Pathology, Shenzhen, China, October 13 2019*, 2019.

[3] Abhimanyu Dubey, Vignesh Ramanathan, Alex Pentland, and Dhruv Mahajan. Adaptive methods for real-world domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14340–14349, 2021.

[4] Anders Andreassen, Yasaman Bahri, Behnam Neyshabur, and Rebecca Roelofs. The evolution of out-of-distribution robustness throughout fine-tuning. *arXiv preprint arXiv:2106.15831*, 2021.

[5] Ali Geisa, Ronak Mehta, Hayden S Helm, Jayanta Dey, Eric Eaton, Carey E Priebe, and Joshua T Vogelstein. Towards a theory of out-of-distribution learning. *arXiv preprint arXiv:2109.14501*, 2021.

[6] Zheyan Shen, Jiashuo Liu, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*, 2021.

[7] Pang Wei Koh, Shiori Sagawa, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021.

[8] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niebner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. In *2017 International Conference on 3D Vision (3DV)*, pages 667–676. IEEE Computer Society, 2017.

[9] Fei Xia, William B Shen, Chengshu Li, Priya Kasimbeg, Micael Edmond Tchapmi, Alexander Toshev, Roberto Martín-Martín, and Silvio Savarese. Interactive gibson benchmark: A benchmark for interactive navigation in cluttered environments. *IEEE Robotics and Automation Letters*, 5(2):713–720, 2020.

[10] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[11] Rene 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.

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

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[15] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in Neural Information Processing Systems*, 27:2366–2374, 2014.

[16] Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco Korner. Evaluation of cnn-based single-image depth estimation methods. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[17] Junjie Hu, Mete Ozay, Yan Zhang, and Takayuki Okatani. Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1043–1051. IEEE, 2019.

[18] Rui Wang, David Geraghty, Kevin Matzen, Richard Szeliski, and Jan-Michael Frahm. Vplnet: Deep single view normal estimation with vanishing points and lines. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 689–698, 2020.

[19] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[20] Chenxi Liu, Barret Zoph, Maxim Neumann, Jonathon Shlens, Wei Hua, Li-Jia Li, Li Fei-Fei, Alan Yuille, Jonathan Huang, and Kevin Murphy. Progressive neural architecture search. In *Proceedings of the European conference on computer vision (ECCV)*, pages 19–34, 2018.