

# Supplementary Materials: Domain Shared and Specific Prompt Learning for Incremental Monocular Depth Estimation

Anonymous Authors

## 1 OVERVIEW

In the supplementary materials, we provide extra information for the submitted manuscript.

- We introduce datasets involved in our experiments. (Section 2.1)
- We provide more results of different incremental domain orders on three incremental domain sequences. (Section 2.2)
- We further investigate the impact of the weights of inter-domain alignment loss function and intra-domain orthogonality loss function on the model performance. (Section 2.3)
- We exhibit visualization comparisons of our DSSP against other methods in more incremental domain sequences. (Section 2.4)

## 2 EXPERIMENTS

### 2.1 Experimental Setup

**Datasets.** We employ unified data argumentation for indoor and outdoor datasets respectively. Specifically, we first resize the images to  $320 \times 240$  pixels and subsequently center crop them into  $304 \times 228$  for indoor datasets. And for outdoor datasets, we random crop the images to  $480 \times 320$  during training phase and  $640 \times 480$  during test phase.

**NYU\_v2** [8] has 464 indoor scenes with a resolution of  $640 \times 480$ . Among them, 249 scenes are used for training, and the rest 215 scenes are used for testing. We use the pre-processed data by [5] with about 50k samples.

**ScanNet** [4] is a large-scale indoor RGB-D dataset that contains 2.5 million RGB-D images. According to [5], we use a subset of 50k samples from the training splits of 1513 scenes for training and evaluate the models on the test set of another 100 scenes with 17k samples. The resolution of RGB images is  $1296 \times 968$ .

**KITTI** [9] is collected by car-mounted cameras and a LIDAR sensor which is an outdoor dataset. We use the official KITTI depth prediction dataset with the official split of scenes for training and validation. The training and validation set has 138 and 18 driving sequences, respectively. The resolution is about  $1216 \times 352$  for most images.

**Virtual Kitti\_v2 (vKITTI\_v2)** [1] is a synthetic dataset for urban driving outdoor environments. It contains 6 different scenes in monocular videos with ground truth depth and different weather conditions. The total training set contains 85k images. Following [2], we use a subset of 12k images for training and 600 for testing. The original resolution is  $1242 \times 375$ .

**CityScapes** [3] is an outdoor autonomous driving dataset that is recorded in 50 cities in Germany and bordering regions with 5000 clear images. The original resolution is  $2048 \times 1024$ .

**CityScapes\_Foggy** [7] has 8975 images for training and 500 for testing. The original resolution of all images is  $2048 \times 1024$ . We compute the depth error metrics using the provided disparity maps.

**CityScapes\_Rainy** [6] has 9432 images for training and 1188 for testing. Same as CityScapes\_Foggy, the original resolution of all images is  $2048 \times 1024$ .

**DAOD** [10] is a large-scaled outdoor dataset with different weather conditions, including regular, sunny, cloudy, foggy, and rainy. For the regular subset, there are around 174k images for training and 7.7K for testing while the remaining subsets only have 500 RGB-D pairs. Therefore, we randomly select 400 pairs for training and 100 pairs for testing.

### 2.2 Ablation of the Incremental Domain Orders

To further investigate the impact of incremental domain sequencing on model adaptability and resistance to forgetting, we train the model by various incremental domain orders, and the results are presented in Table 1, Table 2 and Table 3. We report the average values of corresponding metrics across three datasets along with the forgetting metrics.

**Table 1: Results of different incremental domain orders between NYU\_v2, ScanNet, and KITTI datasets.**

Domain Orders	Abs_Rel	RMSE	$\delta_{1.25}$	Forgetting ↓
NYU_v2 → ScanNet → KITTI	0.204	1.919	69.693	0.255
NYU_v2 → KITTI → ScanNet	0.180	1.817	75.052	0.240
ScanNet → NYU_v2 → KITTI	0.222	1.846	69.021	0.279
ScanNet → KITTI → NYU_v2	0.201	1.805	73.588	0.211
KITTI → ScanNet → NYU_v2	0.172	2.259	74.778	0.027
KITTI → NYU_v2 → ScanNet	0.198	2.436	70.964	0.804

**Table 2: Results of different incremental domain orders between vKITTI\_v2, KITTI, and CityScapes datasets.**

Domain Orders	Abs_Rel	RMSE	$\delta_{1.25}$	Forgetting ↓
vKITTI_v2 → KITTI → CityScapes	7.918	9.963	50.865	4.279
vKITTI_v2 → CityScapes → KITTI	6.634	18.0903	49.810	29.39842
CityScapes → vKITTI_v2 → KITTI	13.464	8.965	75.110	2.798
CityScapes → KITTI → CityScapes	8.684	7.318	79.996	-2.140
KITTI → CityScapes → vKITTI_v2	3.590	7.686	75.909	-0.096
KITTI → vKITTI_v2 → CityScapes	5.149	8.145	69.240	0.519

**Table 3: Results of different incremental domain orders between CityScapes, CityScapes\_Foggy, and CityScapes\_Rainy datasets.**

Domain Orders	Abs_Rel	RMSE	$\delta_{1.25}$	Forgetting ↓
CityScapes → CS_Foggy → CS_Rainy	14.805	16.283	56.551	-0.295
CityScapes → CS_Rainy → CS_Foggy	20.584	21.412	43.912	15.799
CS_Foggy → CityScapes → CS_Rainy	17.383	22.531	41.199	19.204
CS_Foggy → CS_Rainy → CityScapes	13.0459	19.120	44.306	8.121
CS_Rainy → CityScapes → CS_Foggy	11.142	13.274	66.440	-5.548
CS_Rainy → CS_Foggy → CityScapes	11.616	13.954	63.510	-4.530

### 2.3 Searching for Loss Weights

We perform a grid search of the weights of inter-domain alignment and intra-domain orthogonal loss functions. As illustrated in Figure 1, we verify  $\beta = 0.1$  and  $\gamma = 10$  is the optimal choice.

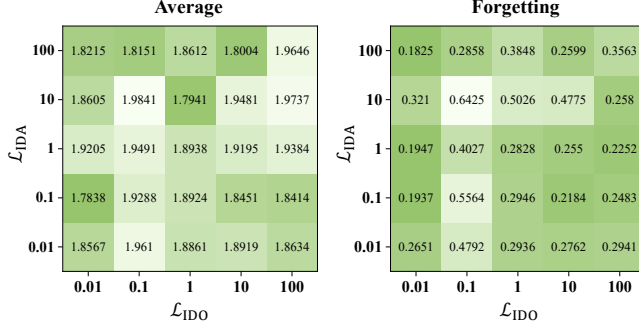


Figure 1: Grid search results of  $\beta$  for  $L_{IDO}$  and  $\gamma$  for  $L_{IDA}$ . We report the average metric and the forgetting metric of RMSE on the "NYU\_v2→ScanNet→KITTI" incremental domain sequence.

### 2.4 Qualitative Results

We show more qualitative results under the other three additional incremental settings in Figure 2, Figure 3, and Figure 4. It can be observed that DSSP exhibits excellent performance across multiple incremental domain sequences. For the KITTI and vKITTI\_v2 datasets, in order to achieve BETTER visualization results, we present the predictions of all methods of the original resolution.

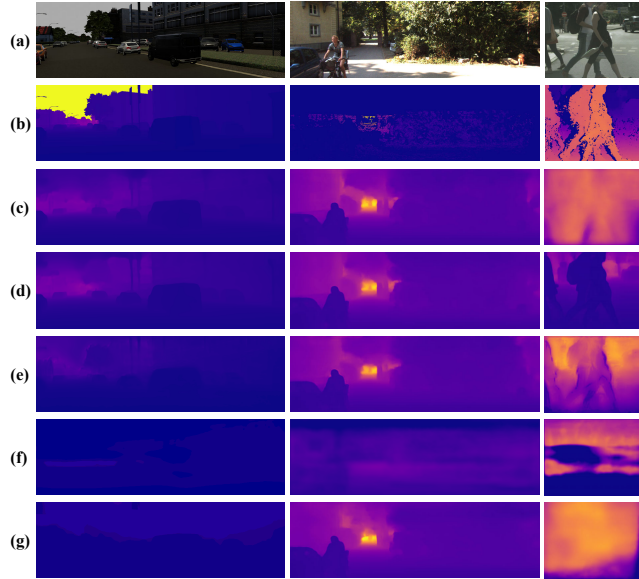


Figure 2: Qualitative comparison with SOTA methods in the learning order of vKITTI\_v2 → KITTI → CityScapes. (a) RGB. (b) Ground truths. (c) FT. (d) JDT. (e) EWC. (f) LL-MonoDepth. (g) Ours.

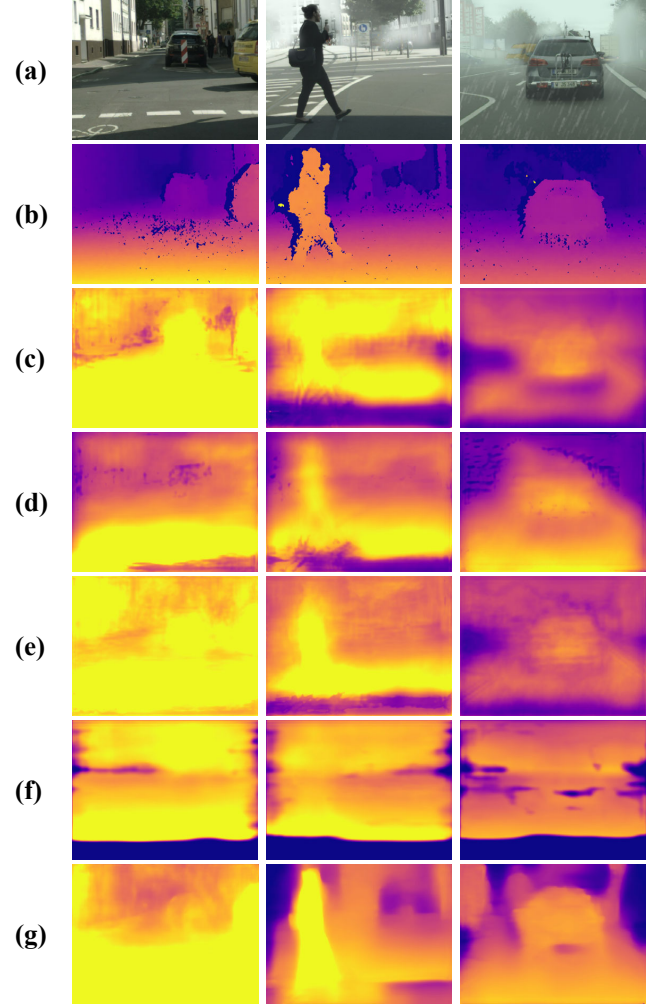
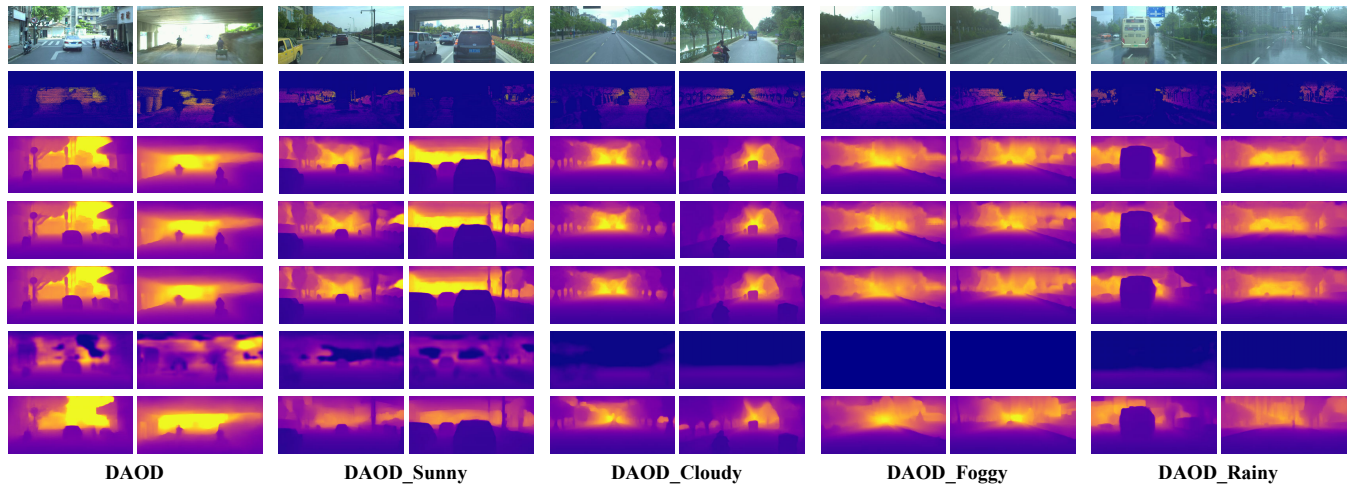


Figure 3: Qualitative comparison with SOTA methods in the learning order of CityScapes → CS\_Foggy → CS\_Rainy. (a) RGB. (b) Ground truths. (c) FT. (d) JDT. (e) EWC. (f) LL-MonoDepth. (g) Ours.



**Figure 4: Qualitative comparison with SOTA methods. From top to bottom are RGB, Ground truths, results of FT, JDT, EWC, LL-MonoDepth, and Ours.**

## REFERENCES

- [1] Yohann Cabon, Naila Murray, and Martin Humenberger. 2020. Virtual kitti 2. *arXiv preprint arXiv:2001.10773* (2020).
- [2] Hemang Chawla, Arnav Varma, Elahe Arani, and Bahram Zonooz. 2024. Continual Learning of Unsupervised Monocular Depth from Videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 8419–8429.
- [3] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 2016. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3213–3223.
- [4] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5828–5839.
- [5] Junjie Hu, Chenyou Fan, Liguang Zhou, Qing Gao, Honghai Liu, and Tin Lun Lam. 2023. Lifelong-MonoDepth: Lifelong Learning for Multi-Domain Monocular Metric Depth Estimation. *arXiv preprint arXiv:2303.05050* (2023).
- [6] Xiaowei Hu, Chi-Wing Fu, Lei Zhu, and Pheng-Ann Heng. 2019. Depth-attentional features for single-image rain removal. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*. 8022–8031.
- [7] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. 2018. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision* 126 (2018), 973–992.
- [8] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. 2012. Indoor segmentation and support inference from rgbd images. *ECCV (5)* 7576 (2012), 746–760.
- [9] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. 2017. Sparsity invariant cnns. In *2017 international conference on 3D Vision (3DV)*. IEEE, 11–20.
- [10] Guorun Yang, Xiao Song, Chaoqin Huang, Zhidong Deng, Jianping Shi, and Bolei Zhou. 2019. Drivingstereo: A large-scale dataset for stereo matching in autonomous driving scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 899–908.