Supplementary Materials for Surface Normals in the Wild

Weifeng Chen[2] Donglai Xiang[1*] Jia Deng[2]
[1]Tsinghua University, Beijing, China [2]University of Michigan, Ann Arbor, USA
xdl13thu@gmail.com, {wfchen,jiadeng}@umich.edu

Training	Method	λ
NYU Subset	d	1
	d_n_al	1
	d_n_al_M	1
	d_n_dl	100
	d_n_dl_M	100
NYU Full	**d_n_al_F**	1
	d_n_al_F_M	1
	d_n_dl_F	100
	d_n_dl_F_M	100

Table 1. λ used in the NYU experiments.

Experiment Pair	RMSE	RMSE (log)	log RMSE (s.inv)	absrel	sqrrel	LS RMSE
d *vs.* d_n_dl	0.000000	0.000006	0.000026	0.005282	0.000063	0.000000
d_F *vs.* d_n_dl_F	0.059211	0.866385	0.820841	0.749617	0.167718	0.000017

Table 2. P-values for each pair of NYU experiments obtained from the paired t-test.

1. More Details on Implementation

Here we provide additional details on the choice of parameters used in our experiments.

For all NYU images, we use the one ground truth focal length shared by all images. For all SNOW images, we assume the same focal length as NYU. Focal length is not needed to predict depth at test time. In training, for SNOW images focal length is often available in EXIF, which could be used to further improve our results.

The parameter λ in Eq.(1) of the paper is selected by performing hyper-parameter optimization on a held-out validation set. They are listed in Tab. 1

The parameter τ in Eq.(3) of the paper equals $ln(1.02)$. It is chosen to be consistent with the threshold used in [4] to generate the ground truth relative depth. We use the same τ for all experiments including those on SNOW.

2. Paired t-test on NYU Results

We quantify uncertainty in our NYU result table using the paired t-test, comparing the performances of two methods on a per-image basis. The results are in Tab. 2. For d *vs.* d_n_dl, all the p-value are less than 0.01, showing that the improvement from normals is significant. For d_F *vs.* d_n_dl_F, the p-value for the LS-RMSE metric is also less than 0.01, again showing that the improvement in LS-RMSE is significant.

*Work done while a visiting student at the University of Michigan.

Crop	Method	RMSE	RMSE (log)	log RMSE (s.inv)	absrel	sqrrel	LS RMSE
Eigen	d	8.09	2.54	2.34	0.43	3.77	6.53
	d_n_al	7.63	1.91	1.75	0.37	2.97	6.04
	d_n_dl	7.39	1.72	1.60	0.36	2.79	5.70
	Godard [3]	**5.74**	**0.24**	**0.22**	**0.13**	**1.14**	**5.17**
Garg	d	7.45	2.58	2.41	0.43	3.51	6.30
	d_n_al	6.86	1.75	1.63	0.36	2.58	5.69
	d_n_dl	6.64	1.82	1.71	0.35	2.51	5.40
	Godard [3]	**5.21**	**0.22**	**0.20**	**0.11**	**0.89**	**4.73**

Table 3. Metric depth error evaluated on the KITTI dataset.

Method	WKDR	$WKDR^{=}$	$WKDR^{\neq}$
d	25.96%	21.57%	26.93%
d_n_al	**22.42%**	**19.80%**	**23.18%**
d_n_dl	25.22%	22.02%	26.07%
Godard [3]	25.84%	26.17%	26.21%

Table 4. Ordinal error evaluated on the KITTI dataset.

3. Experiment on KITTI

For completeness, we provide experimental results on the KITTI dataset. Following [3], we evaluate our methods on two sub-regions of the KITTI test images (i.e., the *Garg_Crop* and *Eigen_Crop* as described in [3]), and use the test/train split of [2].

The relative depth annotations for both training and testing are generated in the same way as described in [4]. As the ground truth surface normals are not provided in the official KITTI dataset, we train on the surface normals generated by Eq.(6) of the paper, and only provide qualitative results of surface normal prediction on the test set. During training, we provide 5,000 surface normal annotations per image.

We test and compare these 3 models: **(1)** a model trained with relative depth only (d); **(2)** a model trained with relative depth and surface normals using the angle-based loss

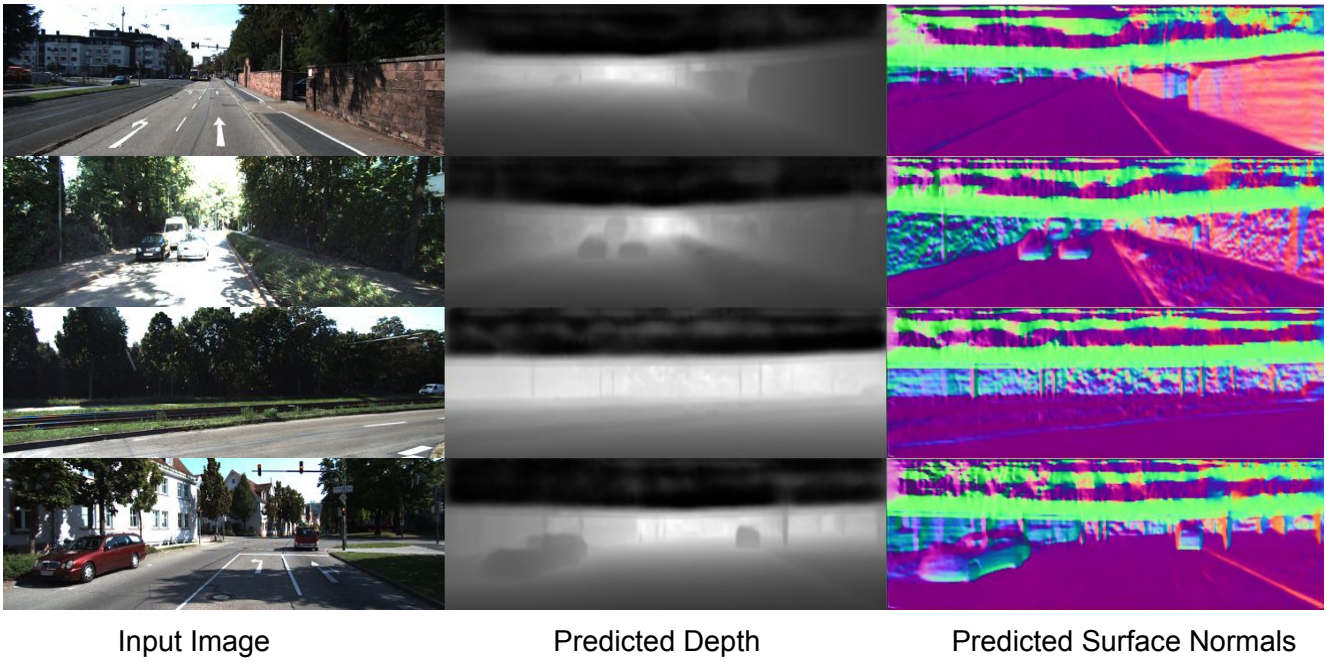

Figure 1. Qualitative results of the KITTI test set.

(*d_n_al*); **(3)** a model trained with relative depth and surface normals using depth-based loss (*d_n_dl*). We use the same network as used in the NYU experiment, with $\tau = ln(1.02)$ and $\lambda = 1$. The input to our network is a 128×416 image and the output is a depth map of the same size. Although ground truth depth values are only available on the lower part of the image, we feed the entire image into the network as is done in [2]. All the metric errors except the LS_RMSE are calculated by first normalizing the depth map to have the same mean and standard variation as the training set. However, some depth maps may contain negative depth value after normalization, and we replace those negative values with the minimum of the non-negative depth values of that depth map when calculating the RMSE (log) and log RMSE (s.inv) metric. For comparison, we also show the state-of-the-art depth-prediction results of Godard et al. [3], which exploits epipolar geometry constraints to train monocular depth-prediction networks (we show the results from their ***Ours resnet pp*** model, which is their best performing model).

We show the results in Tab. 3, 4. Some qualitative results are shown in Fig. 1. Models trained with surface normals (*d_n_al*, *d_n_dl*) consistently outperform the depth-only model (*d*) in both metric error and ordinal error. Training with depth-based loss yields the most significant improvement in metric error while the improvement in ordinal error is the most significant for the angle-based loss model. These results once again show that surface normals can help improve depth predictions in the absence of ground truth depth in training.

References

[1] A. Bansal, B. Russell, and A. Gupta. Marr revisited: 2d-3d alignment via surface normal prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5965–5974, 2016. 4

[2] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NIPS*, 2014. 1, 2

[3] C. Godard, O. Mac Aodha, and G. J. Brostow. Unsupervised monocular depth estimation with left-right consistency. *arXiv preprint arXiv:1609.03677*, 2016. 1, 2

[4] D. Zoran, P. Isola, D. Krishnan, and W. T. Freeman. Learning ordinal relationships for mid-level vision. In *ICCV*, 2015. 1

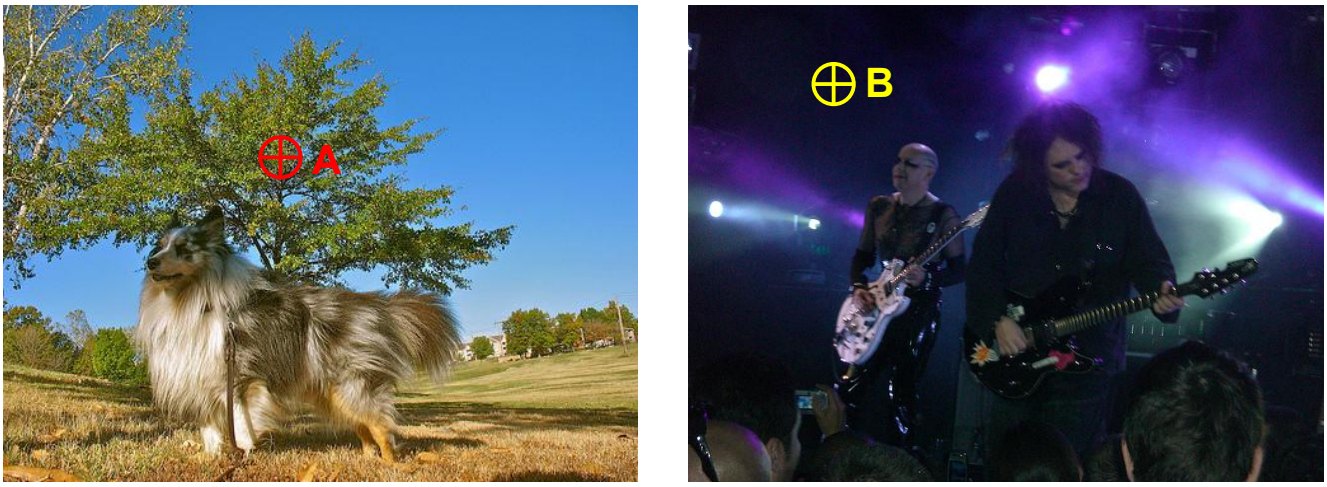

Figure 2. Some examples of the very difficult cases where the surface normal is hard to infer from the image. Point A is on tree leaves, which are small and cluttered. Point B is on a dark background where nothing can be seen clearly. In these case, the worker can indicate that the surface normal is hard to tell. Please view in color.

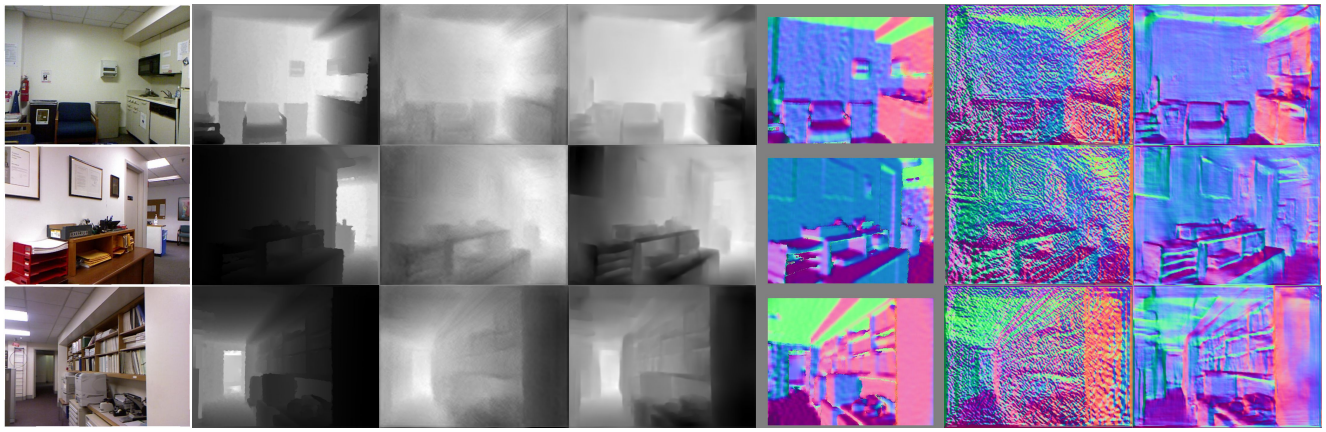

Figure 3. Qualitative results of the NYU test set. Here we show example outputs of the networks trained with or without surface normals on the NYU Subset.

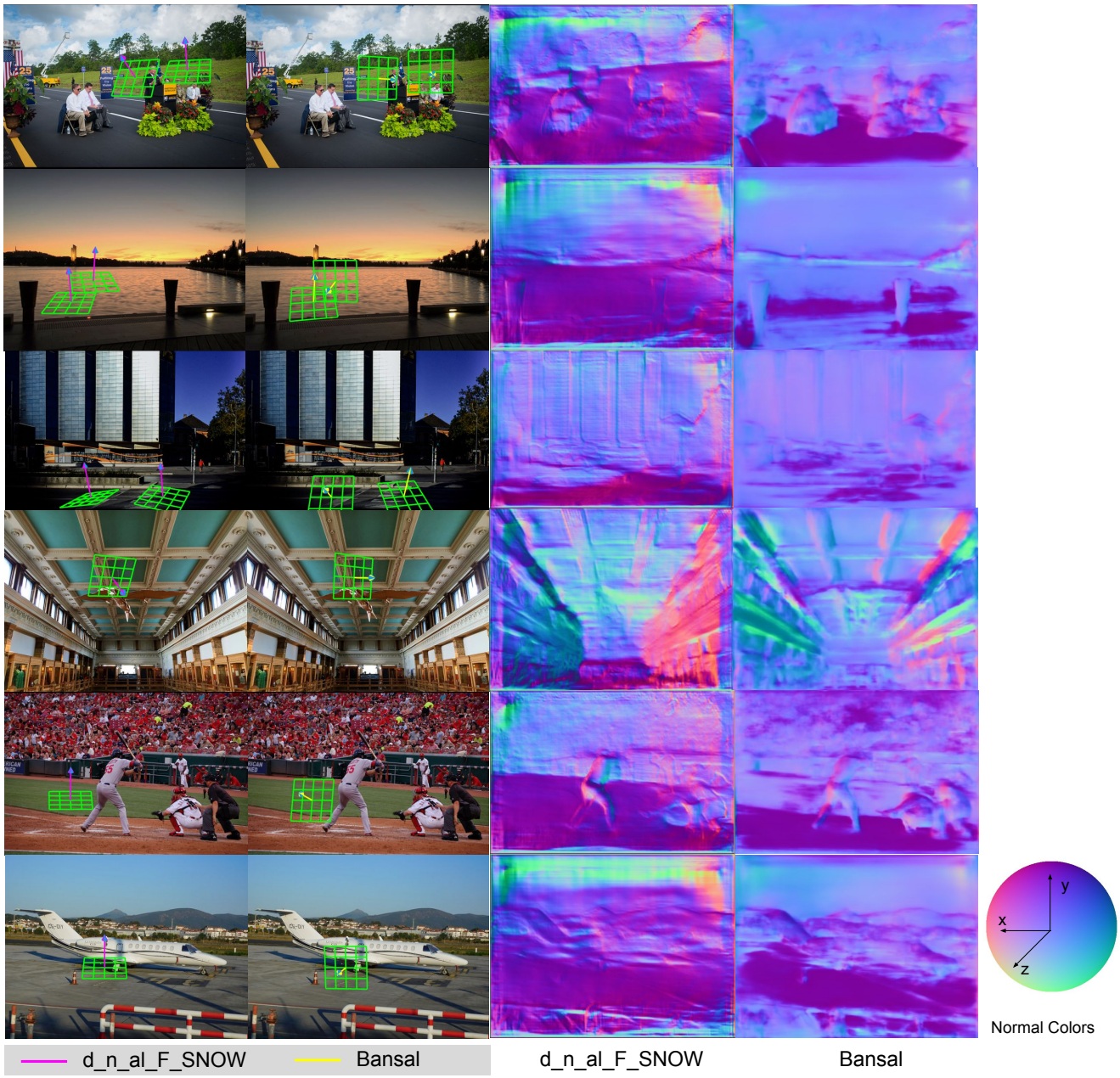

Figure 4. Additional qualitative results on SNOW produced by our model and Bansal [1]. The left two columns visualize some predicted normal vectors from the two methods. The other two columns are the full normal maps.