Monitored Distillation for Positive Congruent Depth Completion SUPPLEMENTARY MATERIALS

Tian Yu Liu[1⋆], Parth Agrawal[1⋆], Allison Chen[1⋆], Byung-Woo Hong[2], and Alex Wong[3]

[1] UCLA Vision Lab, Los Angeles, CA 90095, USA
{tianyu139,parthagrawal24,allisonchen2}@ucla.edu
[2] Chung-Ang University, Heukseok-Dong, Dongjak-Gu, Seoul, 06973, Korea
hong@cau.ac.kr
[3] Yale University, New Haven, CT 06511, USA
alex.wong@yale.edu

1 Summary of Contents

In Sec. 2 we provide our implementation details, hyperparameters, learning rate schedule, and augmentations used during training. In Sec. 3, we provide a sensitivity study on the effect of various density levels in the sparse depth input. In Sec. 4, we compare Monitored Distillation to methods operating under the non-blind ensemble setting and show that distilling from a blind ensemble using our method improves over directly distilling from any single teacher – even the best teacher. In Sec. 5 we make qualitative comparisons against the state-of-the-art unsupervised (Fig. 2, 3) and supervised methods (Fig. 4, 5) and show that our method achieves comparable performance to the top supervised methods while using significantly fewer parameters. We further include a discussion regarding the error modes of teacher models and show that Monitored Distillation is able to avoid distilling the error modes of individual teachers. Lastly, we conclude with a discussion on the limitations of our method in Sec. 6. Code available at: https://github.com/alexklwong/mondi-python.

2 Implementation Details

We implement our approach in PyTorch and optimized our networks using Adam [13] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We trained for a total of 200 epochs on KITTI [27], and 75 epochs on VOID [33]. We use a batch size of 8 and choose $w_{md} = 1.0$, $w_{ph} = 0.15$, $w_{st} = 0.85$, $w_{sm} = 0.1$ and temperature parameters $\lambda = 0.10$ for both KITTI and VOID, $\alpha = 0.001$ for KITTI and $\alpha = 0.10$ for VOID. We detail our learning rate schedule for each dataset in Table 1. We employ a sparse-to-dense module from [36], and the min and max pool kernel sizes are detailed in Table 2.

⋆ denotes equal contribution.

Table 1: **Learning Rate Schedule**. Presented for KITTI (outdoors) and VOID (indoors) depth completion benchmark datasets.

Epochs	Learning Rate
KITTI [27]	
0 to 30	5×10^{-4}
30 to 50	2×10^{-4}
50 to 90	5×10^{-5}
90 to 100	2×10^{-5}
100 to 120	5×10^{-5}
120 to 200	2×10^{-5}
VOID [33]	
0 to 20	2×10^{-4}
20 to 75	5×10^{-5}

Table 2: **Min Pool and Max Pool Kernel Sizes.** Used in our sparse-to-dense module. Kernel sizes for VOID [36] are larger because the point cloud generated from VIO [6] is much sparser than that of the LIDAR used in KITTI [27].

Dataset	Min Pool	Max Pool
KITTI [27]	5, 7, 9, 11, 13	15, 17
VOID [33]	15, 17, 19, 21, 23	27, 29

For data augmentations, we performed random horizontal and vertical crops to the image and depth maps of size 768×320 for KITTI and 576×448 for VOID. We randomly removed between 60% to 70% of the sparse points for KITTI and 60% to 95% of the sparse points for VOID. For both KITTI and VOID, we performed random color shifts, saturation and contrast adjustments between 0.80 and 1.20 in the input. Each augmentation has a 50% chance of being applied. Augmentations are enabled 100% of the time for VOID; for KITTI, augmentations are enabled 100% of the time until the 100th epoch, after which it reduces to 50% for the remaining 50 epochs.

For the ease of training, we preprocess both datasets by running inference on the training sets using each teacher model (except for sparse depth maps, which are given) and load them during training. We note that teachers can also be used for training online as we require $<$ 8GB of GPU memory for training. It will take longer as training time scales with the number of teachers, but even if teacher inference is done sequentially, we will only use an extra 3265MiB of memory for largest teacher (NLSPN) and at most 0.20s per image (total inference time for all

Table 3: **Inference Time and GPU Memory.** Measured for a single image (480×640) taken from VOID dataset. Training online requires an extra 3265MiB of memory for largest teacher (NLSPN) and at most 0.20s per image (total inference time for all teachers used) on standard 11GB GPU.

	KBNet	ENet	PENet	NLSPN	MSG-CHN	FusionNet	ScaffNet
Time (ms)	15	15	24	112	6	24	6
GPU (MiB)	1043	3263	3265	2471	1095	1067	1047

Table 4: **Error metrics.** d_{gt} denotes ground truth depth.

Metric	Definition
MAE	$\frac{1}{\lvert\Omega\rvert} \sum_{x \in \Omega} \lvert\hat{d}(x) - d_{gt}(x)\rvert$
RMSE	$\left(\frac{1}{\lvert\Omega\rvert} \sum_{x \in \Omega} \lvert\hat{d}(x) - d_{gt}(x)\rvert^2\right)^{1/2}$
iMAE	$\frac{1}{\lvert\Omega\rvert} \sum_{x \in \Omega} \lvert 1/\hat{d}(x) - 1/d_{gt}(x)\rvert$
iRMSE	$\left(\frac{1}{\lvert\Omega\rvert} \sum_{x \in \Omega} \lvert 1/\hat{d}(x) - 1/d_{gt}(x)\rvert^2\right)^{1/2}$

teachers used) on a standard 11GB GPU. In Table 3, we present inference times and memory usage for a single image from VOID using each of our teachers. For the student baseline (without teachers), we emulate the training procedure of [36]. To obtain pose, we trained a pose network jointly with our depth model by minimizing Eqn. 10 from the main text.

In Table 4 we present four metrics that we use to evaluate our models. These are the metrics reported on the KITTI and VOID benchmark datasets.

3 Sensitivity to Various Input Densities

To demonstrate robustness against varying levels of sparsity in the input, we evaluate our method on VOID using input sparse point clouds of varying densities: 150, 500, and 1500 points which correspond to densities of approximately 0.05%, 0.15%, and 0.5% respectively over the image space. Compared to datasets such as KITTI, sparse point clouds on VOID can be $100\times$ more sparse, making sparse to dense depth completion even more challenging.

As expected, qualitative results shown in Fig. 1 demonstrate that our performance improves as density increases. We note that as the density of the point cloud decreases, more errors occur in far, homogeneous regions that tend to lack sparse points. In which case, we observe that our model is biased towards outputting farther depths. Quantitative results against naive blind ensembling baselines are provided in Table 5, where we restored the best checkpoint trained on 0.5% density for each model and evaluated on the VOID test set of 0.05%, 0.15%, and 0.5%. Our method consistently outperforms ensembling baselines on an average of 18.08% across all metrics, demonstrating greater improvement for lower densities. This suggests our method is robust to sparseness and less reliant

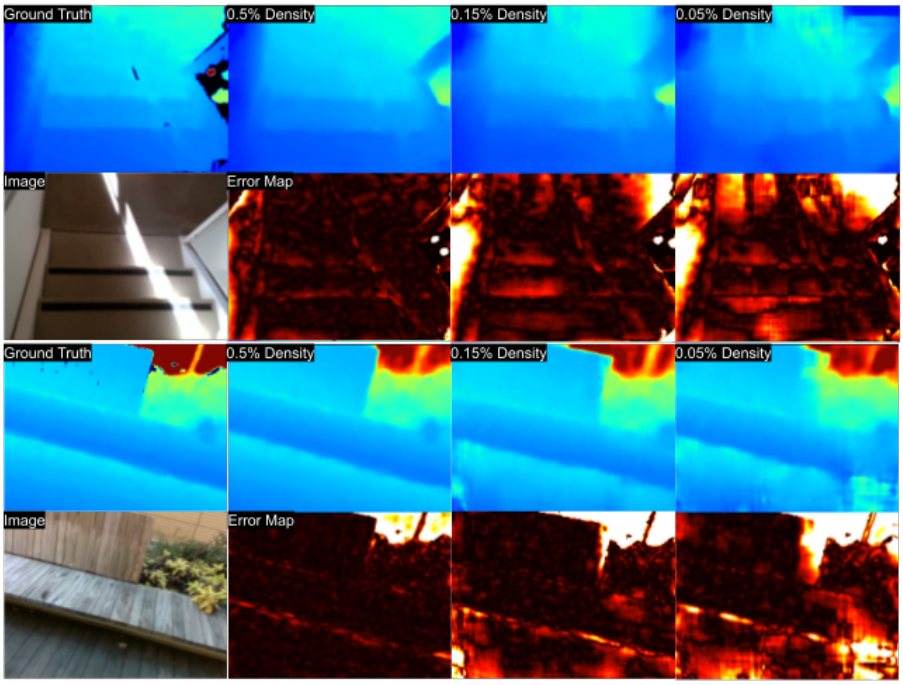

Fig. 1: **Qualitative Results for Density Sensitivity Study.** For two different samples (top half and bottom half), the left column shows the ground truth depth prediction and original image. The right three columns represent our method's output predictions given sparse depth maps at 0.5%, 0.15% and 0.05% density. As observed, error increases with decreasing densities.

on sparse depth assistance. The use case is for densifying point clouds produced by VIO systems, where locally there are few or no points. Thus, learning a prior on the shapes of objects populating the scene becomes critical as the model must depend more on the information from the image.

However, we must note that, while our method beats others for each tested density levels, our model is also sensitive to the input density. Specifically, our mean error doubles ($\times 1.87$) when density decreases to 0.15% and more than triples ($\times 3.04$) when density decreases to 0.05% (i.e. decreases by 10x). This is shown quantitatively in Table 5 and qualitatively in Fig. 1 where far regions that are largely homogeneous become increasingly corrupted. This is because there are usually fewer or no points tracked by the VIO system in those areas.

4 Comparison to Non-blind Ensemble Baselines

We compare our results to naive ensembling baselines without the blind-ensemble assumption (i.e. the best performing model is known). We compare against the

Table 5: **Sensitivity Study for Sparse Depth Density on VOID.** We train a single model on VOID using monocular video and corresponding sparse depth maps of 0.50% density and evaluate it on 0.50%, 0.15%, 0.05% density test sets. On average, we outperform naive ensembling by 17.22% at 0.50% density, 17.44% at 0.15% density, and 29.58% at 0.05% density. Across all densities, we outperform naive ensembling on average by 18.08%.

Distillation Method	MAE	RMSE	iMAE	iRMSE
0.50% Density				
Mean	35.791	84.780	18.651	42.899
Median	33.889	85.245	17.296	40.401
Random	43.638	94.384	24.741	50.265
Ours	**29.666**	**79.775**	**14.838**	**37.875**
0.15% Density				
Mean	75.969	169.259	35.502	76.108
Median	74.192	177.365	33.046	71.460
Random	78.819	166.750	39.183	77.690
Ours	**61.370**	**146.569**	**27.963**	**64.356**
0.05% Density				
Mean	139.676	281.677	64.233	119.177
Median	139.276	306.621	57.785	109.511
Random	129.900	259.239	62.354	111.820
Ours	**104.966**	**225.604**	**48.440**	**96.786**

baseline approach of simply training using the best performing teacher, and show that we perform better despite operating under the blind ensemble setting. We note that while many methods have proposed distilling from a single teacher, in many cases this is not practical. To determine which teacher to distill from, one must have a measure error; existing methods relied on ground truth to select the teacher model. Yet, in reality, ground truth is often not available and when available it is expensive to obtain. So without ground truth i.e. the blind ensemble setting, it becomes non-trivial to "find" the best teacher. For this particular scenario in Table 6, we assume competing methods are able to choose the best teacher and distill from them. This also serves as baseline for how well a student model distilling from any particular top method will perform.

In Table 6, we compare against using only unsupervised losses (baseline, row 1), the naive mean blind ensembling method with unsupervised loss (row 2), distilling from a single teacher (rows 3-5), and Monitored Distillation (last

Table 6: **Comparisons to Non-Blind Ensembles on KITTI Validation Set.** Row 1 is trained on standard photometric reprojection loss. Distilling from the mean of the ensemble (row 2) yields union of the error modes. Single teacher distillation baselines (rows 3-5) improves upon the mean ensemble. The best performing method is our full model (row 6), where our monitored distillation boosts performance of the best model (NLSPN) even though we operate in the blind ensemble setting.

Distillation Method	MAE	RMSE	iMAE	iRMSE
Unsupervised Loss Only	333.865	1374.013	1.315	4.260
Mean	232.481	851.285	0.963	2.405
Distill E-Net [11]	228.356	831.737	0.952	2.278
Distill PE-Net [11]	226.058	819.46	0.964	2.316
Distill NLSPN [20]	221.077	841.952	0.921	2.234
Ours	**218.222**	**815.157**	**0.910**	**2.184**

row). We observe that learning from an ensemble of teachers using the mean prediction (row 2) with unsupervised losses yield the union of error modes. In fact, distilling from the mean of the ensemble performs *worse* than distilling from any single teachers across all metrics. Furthermore, while knowledge distillation with a single teacher (rows 3-5) improves the baseline and also distilling from the mean of the ensemble, none of them produce the results that outperforms our method that distills from a blind ensemble since the student model may still propagate the teacher's error modes.

We demonstrate the effectiveness of Monitored Distillation in row 6 of Table 6, where our model performs significantly better than the distilling from any of the individual teachers – even the best one, NLSPN [20]. Specifically, our monitor allows the model to adaptively choose the teachers that best minimize reconstruction residual (for calibrated images, the photometric reprojection error is a well-supported measure of reconstruction quality) and to fall back on unsupervised losses to learn the correct correspondences when the teachers fail to yield low reconstruction residuals.

5 Qualitative Comparisons on KITTI

Here, we provide qualitative comparisons across the spectrum of supervision. First, we compare against top unsupervised methods on the KITTI benchmark, where we show that our method is able to better recover the complex and homogeneous structures. After that, we show head-to-head comparisons against the top supervised and unsupervised methods that we distill from and demonstrate that we yield positive congruent training as we avoid distilling from the error modes of individual teachers.

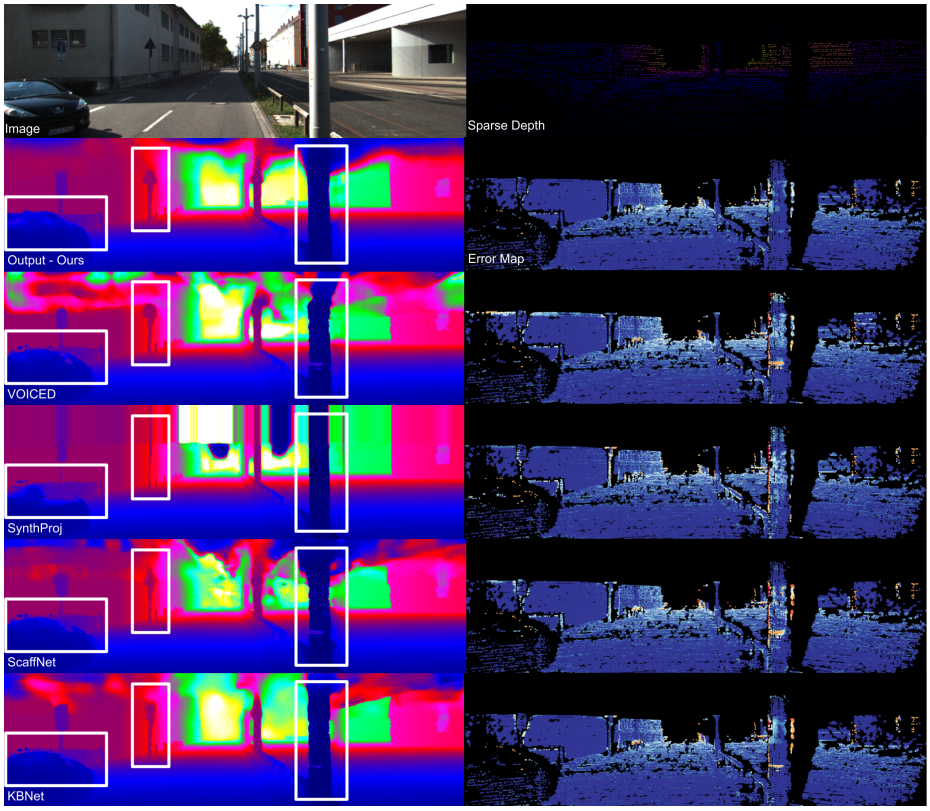

Fig. 2: **KITTI - Comparison to Unsupervised Methods #1.** Monitored Distillation, VOICED [33], ScaffNet [31], SynthProj [18], KBNet [36]. The first row shows the input image I_t (left) and sparse depth z (right). Rows 2-5 are the respective models' dense depth maps (left) and error maps w.r.t ground truth (right). The distilled regularization learnt by our approach improves accuracy for transparent/translucent regions like car windows, and structures such as poles. This is a known error mode of unsupervised methods due to the ambiguity of homogeneous surfaces.

5.1 Against Unsupervised Methods

We qualitatively compare our results against top unsupervised methods in Fig. 2 and Fig. 3. We provide head-to-head comparisons against VOICED [33], ScaffNet [31], SynthProj [18], KBNet [36]. As demonstrated in our figures, the distilled regularization learned by our approach yields higher model accuracy overall, especially in transparent or translucent regions such as car windows, and largely homogeneous and thin structures like poles and trees. This shows that our Monitored Distillation approach can effectively distill priors learnt by the complex teacher networks to our lightweight student model. Compared against the state

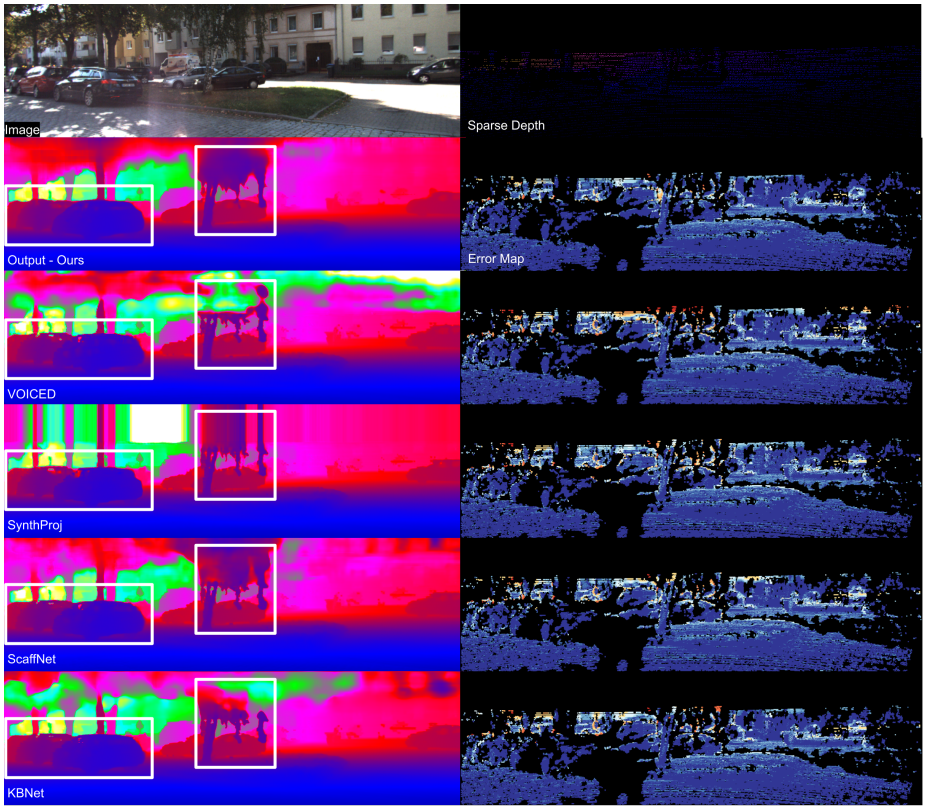

Fig. 3: **KITTI - Comparison to Unsupervised Methods #2.** Monitored Distillation, VOICED [33], ScaffNet [31], SynthProj [18], KBNet [36]. The first row shows the input image I_t (left) and sparse depth z (right). Rows 2-5 are the respective models' dense depth maps (left) and error maps w.r.t ground truth (right). The distilled regularization learnt by our approach improves accuracy for transparent/translucent regions like car windows, and complex structures like trees. This is a known error mode of unsupervised methods due to the ambiguity of homogeneous surfaces.

of the art [36], our method consistently yields lower error in vehicles, highlighted in white, where we do not suffer from the lidar artifacts leaving "holes" in the cars. In general, our method learns to produce consistent depths within an object for instance the pole and wall in the left image of Fig. 2 – [36] predicted a break in the pole whereas our method produces a continuous surface.

5.2 Positive Congruent Training

In Fig. 4 and Fig. 5, we further compare the output of our method against the top supervised and unsupervised methods from which we distill our regularities:

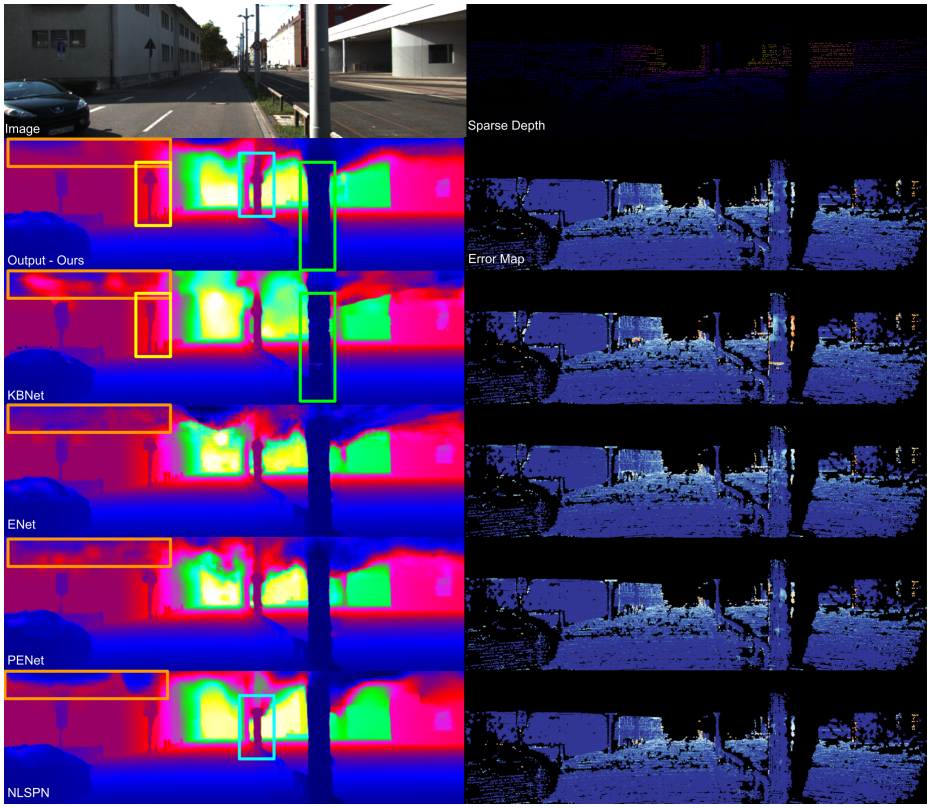

Fig. 4: **KITTI - Comparison to Teacher Methods #1.** Monitored Distillation, KBNet [36], PENet [11], ENet [11], and NLSPN [20]. The first row shows the input image I_t (left) and sparse depth z (right). Rows 2-5 are the respective models' dense depth maps (left) and error maps w.r.t ground truth (right). We show that our method fixes (highlighted) error modes present in the teachers.

KBNet [36], PENet [11], ENet [11], and NLSPN [20]. Each teacher has some error modes. For instance, as highlighted in Fig. 4, KBNet fails to reconstruct the pole and leftmost street sign, NLSPN predicts the wrong shape for the middle street sign, and ENet, PENet, and NLSPN fails to reconstruct the top left building region. In Fig. 5, KBNet and ENet fail to reconstruct the tree, and NLSPN and PENet fail to predict a smooth surface for the bottom right building. In our predictions, we show that our method is able to address the (highlighted) error modes present in the various teachers and avoid distilling them (see Sec. 3 on main paper for details). This results in positive congruent training, where we distill from a teacher only when it yields low reconstruction errors and avoid the error modes of individual teachers.

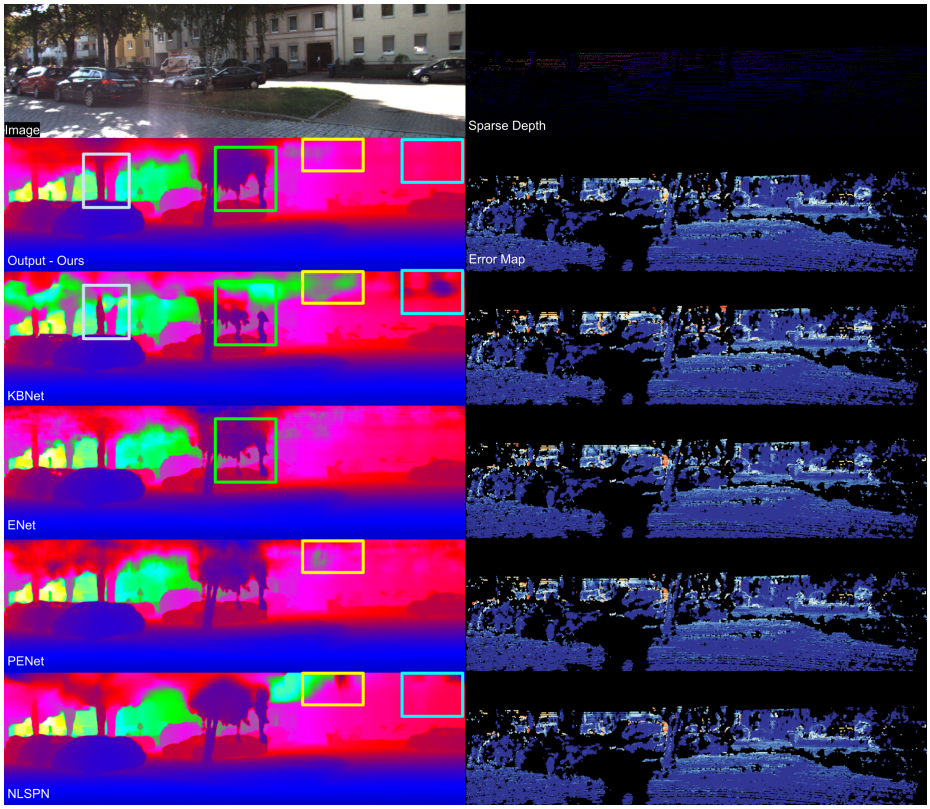

Fig. 5: **KITTI - Comparison to Teacher Methods #2.** Monitored Distillation, KBNet [36], PENet [11], ENet [11], and NLSPN [20]. The first row shows the input image I_t (left) and sparse depth z (right). Rows 2-5 are the respective models' dense depth maps (left) and error maps w.r.t ground truth (right). We show that our method fixes (highlighted) error modes present in the teachers.

6 Limitations

As noted in our discussion (Sec. 5 from main paper), learning distilled regularities from teachers imposes several risks and limitations. The effectiveness of Monitored Distillation is lower bounded by training using the unsupervised loss, and upper bounded by the performance of teachers and their error modes. In particular, if all teachers yield high photometric reprojection errors on certain regions due to inaccurate depth values, the student model will have to rely on unsupervised losses rather than the distilled depth, which lower bounds our performance.

Our approach also depends on unsupervised photometric and structural losses that are limited by parallax. In stereo settings with insufficient baseline, or in monocular settings where there is insufficient movement between image frames,

the photometric reprojection error would be limited in conveying information about the 3D scene layout for distant regions. Our approach is further limited by the identifiability of shape from the reprojection error, and relies on generic priors to resolve the aperture problem and blank-wall effects.

Lastly, our method struggles to explicitly handle non-Lambertian surfaces as we rely on photometric reprojection error for our ensembling method. However, we know from [12] that the domain coverage of specularities and translucency is sparse due to the sparsity of primary illuminants (rank of the reflectance tensor is deficient and typically small). So, explicitly modeling deviations from diffuse Lambertian reflection is likely to yield modest returns in accuracy of the reconstruction. Nevertheless, we account for such surfaces to a certain extent by additionally incorporating sparse depth constraints.

Nonetheless, this is the first work to introduce Monitored Distillation for depth completion in the blind ensemble setting. Not only that, by leveraging Monitored Distillation, we are able to compress the student such that it can run in real-time, unlike the teachers. Our framework is general and we believe it can be formulated to be applied to a number of tasks outside of depth completion [11,17,19,20,31,32,33,36,38,41], including but not limited to unsupervised learning of geometry, i.e. stereo [2,3,5,22,37,34], optical flow [1,14,15,16,25,26], multi-view stereo [4,8,28,39,40], monocular depth prediction [6,7,21,23,24,29,30,35], and adaptive regularization [9,10,35].

References

1. Aleotti, F., Poggi, M., Tosi, F., Mattoccia, S.: Learning end-to-end scene flow by distilling single tasks knowledge. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 10435–10442 (2020)
2. Berger, Z., Agrawal, P., Liu, T.Y., Soatto, S., Wong, A.: Stereoscopic universal perturbations across different architectures and datasets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15180–15190 (2022)
3. Chang, J.R., Chen, Y.S.: Pyramid stereo matching network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5410–5418 (2018)
4. Chen, R., Han, S., Xu, J., Su, H.: Point-based multi-view stereo network. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1538–1547 (2019)
5. Duggal, S., Wang, S., Ma, W.C., Hu, R., Urtasun, R.: Deeppruner: Learning efficient stereo matching via differentiable patchmatch. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4384–4393 (2019)
6. Fei, X., Wong, A., Soatto, S.: Geo-supervised visual depth prediction. IEEE Robotics and Automation Letters **4**(2), 1661–1668 (2019)
7. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2019)
8. Gu, X., Fan, Z., Zhu, S., Dai, Z., Tan, F., Tan, P.: Cascade cost volume for high-resolution multi-view stereo and stereo matching. In: Proceedings of the

IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2495–2504 (2020)
9. Hong, B.W., Koo, J.K., Dirks, H., Burger, M.: Adaptive regularization in convex composite optimization for variational imaging problems. In: German Conference on Pattern Recognition. pp. 268–280. Springer (2017)
10. Hong, B.W., Koo, J., Burger, M., Soatto, S.: Adaptive regularization of some inverse problems in image analysis. IEEE Transactions on Image Processing (2019)
11. Hu, M., Wang, S., Li, B., Ning, S., Fan, L., Gong, X.: Penet: Towards precise and efficient image guided depth completion. arXiv preprint arXiv:2103.00783 (2021)
12. Jin, H., Soatto, S., Yezzi, A.J.: Multi-view stereo beyond lambert. In: 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings. vol. 1, pp. I–I. IEEE (2003)
13. Kingma, D.P., Ba, J.L.: Adam: A method for stochastic gradient descent. In: ICLR: International Conference on Learning Representations (2015)
14. Lao, D., Sundaramoorthi, G.: Minimum delay moving object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4250–4259 (2017)
15. Lao, D., Sundaramoorthi, G.: Extending layered models to 3d motion. In: Proceedings of the European conference on computer vision (ECCV). pp. 435–451 (2018)
16. Lao, D., Sundaramoorthi, G.: Minimum delay object detection from video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5097–5106 (2019)
17. Liu, T.Y., Agrawal, P., Chen, A., Hong, B.W., Wong, A.: Monitored distillation for positive congruent depth completion. arXiv preprint arXiv:2203.16034 (2022)
18. Lopez-Rodriguez, A., Busam, B., Mikolajczyk, K.: Project to adapt: Domain adaptation for depth completion from noisy and sparse sensor data. In: Proceedings of the Asian Conference on Computer Vision (2020)
19. Merrill, N., Geneva, P., Huang, G.: Robust monocular visual-inertial depth completion for embedded systems. In: International Conference on Robotics and Automation (ICRA). IEEE (2021)
20. Park, J., Joo, K., Hu, Z., Liu, C.K., Kweon, I.S.: Non-local spatial propagation network for depth completion. In: European Conference on Computer Vision, ECCV 2020. European Conference on Computer Vision (2020)
21. Poggi, M., Aleotti, F., Tosi, F., Mattoccia, S.: On the uncertainty of self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3227–3237 (2020)
22. Poggi, M., Aleotti, F., Tosi, F., Zaccaroni, G., Mattoccia, S.: Self-adapting confidence estimation for stereo. In: European Conference on Computer Vision. pp. 715–733. Springer (2020)
23. Poggi, M., Tosi, F., Aleotti, F., Mattoccia, S.: Real-time self-supervised monocular depth estimation without gpu. IEEE Transactions on Intelligent Transportation Systems (2022)
24. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12179–12188 (2021)
25. Sun, D., Yang, X., Liu, M.Y., Kautz, J.: Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8934–8943 (2018)
26. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: European conference on computer vision. pp. 402–419. Springer (2020)

27. Uhrig, J., Schneider, N., Schneider, L., Franke, U., Brox, T., Geiger, A.: Sparsity invariant cnns. In: 2017 International Conference on 3D Vision (3DV). pp. 11–20. IEEE (2017)
28. Wang, F., Galliani, S., Vogel, C., Speciale, P., Pollefeys, M.: Patchmatchnet: Learned multi-view patchmatch stereo. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14194–14203 (2021)
29. Watson, J., Firman, M., Brostow, G.J., Turmukhambetov, D.: Self-supervised monocular depth hints. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2162–2171 (2019)
30. Wong, A., Cicek, S., Soatto, S.: Targeted adversarial perturbations for monocular depth prediction. Advances in Neural Information Processing Systems **33** (2020)
31. Wong, A., Cicek, S., Soatto, S.: Learning topology from synthetic data for unsupervised depth completion. IEEE Robotics and Automation Letters **6**(2), 1495–1502 (2021)
32. Wong, A., Fei, X., Hong, B.W., Soatto, S.: An adaptive framework for learning unsupervised depth completion. IEEE Robotics and Automation Letters **6**(2), 3120–3127 (2021)
33. Wong, A., Fei, X., Tsuei, S., Soatto, S.: Unsupervised depth completion from visual inertial odometry. IEEE Robotics and Automation Letters (2020)
34. Wong, A., Mundhra, M., Soatto, S.: Stereopagnosia: Fooling stereo networks with adversarial perturbations. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 2879–2888 (2021)
35. Wong, A., Soatto, S.: Bilateral cyclic constraint and adaptive regularization for unsupervised monocular depth prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5644–5653 (2019)
36. Wong, A., Soatto, S.: Unsupervised depth completion with calibrated backprojection layers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12747–12756 (2021)
37. Xu, H., Zhang, J.: Aanet: Adaptive aggregation network for efficient stereo matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1959–1968 (2020)
38. Yang, Y., Wong, A., Soatto, S.: Dense depth posterior (ddp) from single image and sparse range. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3353–3362 (2019)
39. Yao, Y., Luo, Z., Li, S., Fang, T., Quan, L.: Mvsnet: Depth inference for unstructured multi-view stereo. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 767–783 (2018)
40. Yao, Y., Luo, Z., Li, S., Shen, T., Fang, T., Quan, L.: Recurrent mvsnet for high-resolution multi-view stereo depth inference. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5525–5534 (2019)
41. Zhu, Y., Dong, W., Li, L., Wu, J., Li, X., Shi, G.: Robust depth completion with uncertainty-driven loss functions. arXiv preprint arXiv:2112.07895 (2021)