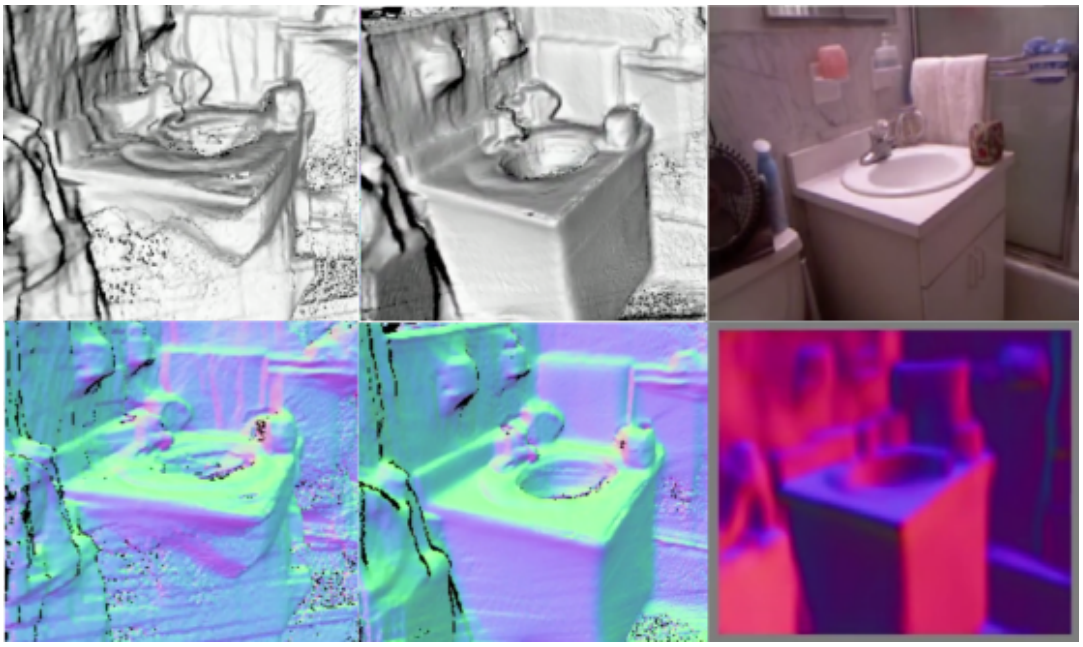

Fig. 5. Qualitative results on NYU raw dataset 'bathroom_0003' test sequence. Phong shaded fused reconstruction using smoothness prior (top left), phong shaded fused reconstruction using normals prior (top middle), a rgb keyframe image in the sequence (top right), surface normal rendering of fused smoothness-prior reconstruction (bottom left), surface normal rendering of fused normal-prior reconstruction (bottom middle), corresponding normal predictions for rgb keyframe image (bottom right). Note the more accurate reconstruction of textureless regions like the inside of the round sink using the normal-prior. A live comparison video is available at https://youtu.be/BRLN-1MTZtw.

Fig. 6. Qualitative results on NYU raw dataset 'bedroom_0048' test sequence. Input rgb image (top left), reconstructed keyframe depth map with smoothness regularizer (top middle), fused reconstruction using smoothness regularizer (top right), keyframe surface normal prediction (bottom left), reconstructed keyframe depth map with normal-based regularizer (bottom middle), fused reconstruction using normal-based regularizer (bottom right). Note the more accurate reconstruction of the wall and floor overall when using the normal prior.

Fig. 7. Qualitative results on TUM dataset 'fr2_desk' sequence. From left-to-right are fused reconstruction using smoothness prior, fused reconstruction using normals prior, a rgb keyframe image in the sequence, and corresponding normal predictions for rgb keyframe image.

		Error (lower is better)				Accuracy (higher is better)		
		rms (m)	log	abs.rel	sq.rel	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
NYU-D V2 Raw 25 Test Scenes	CNN Depth [12]	0.637	0.226	0.163	0.135	0.738	0.937	0.982
	P.E. + Smoothness	0.522	0.206	0.123	0.111	0.834	0.949	0.979
	P.E. + Normals	**0.449**	**0.174**	**0.086**	**0.076**	**0.893**	**0.964**	**0.985**
TUM dataset 'fr2_desk'	CNN Depth [12]	1.141	0.368	0.227	0.261	0.543	0.820	0.923
	P.E. + Smoothness	0.678	0.254	0.132	0.127	0.788	0.889	**0.963**
	P.E. + Normals	**0.654**	**0.242**	**0.119**	**0.115**	**0.829**	**0.898**	**0.963**
ICL-NUIM dataset 'lr kt0'	CNN Depth [12]	0.829	0.426	0.295	0.261	0.472	0.781	0.905
	P.E. + Smoothness	0.322	0.175	0.123	0.058	0.828	0.966	**0.998**
	P.E. + Normals	**0.221**	**0.118**	**0.073**	**0.024**	**0.936**	**0.991**	**0.998**

TABLE I

QUANTITATIVE RESULTS ON 25 RAW NYU-D V2 DATASET TEST SEQUENCES, TUM DATASET 'FR2_DESK' SEQUENCE, AND ICL-NUIM DATASET 'LR KT0' SEQUENCE. P.E. = PHOTOMETRIC ERROR. THE AVERAGE ERRORS AND ACCURACY ARE FOR KEYFRAME RECONSTRUCTIONS AGAINST KINECT DEPTH MAPS (WHERE VALID DEPTHS ARE AVAILABLE). THE RESULTS HERE ARE SHOWN FOR THE OPTIMAL LAMBDA VALUES FOR NORMALS AND SMOOTHNESS REGULARIZER BASED ON FIG. 4, BUT WITH HIGHER NUMBER OF ITERATIONS WHICH ALLOWED FOR HIGHER ACCURACY IN THE RECONSTRUCTIONS.

depth respectively of a pixel location corresponding to p.

rms: $\sqrt{\frac{1}{|P|}\sum_{p\in P}\|d_p - d_p^{gt}\|^2}$

log rms : $\sqrt{\frac{1}{|P|}\sum_{p\in P}\|log(d_p) - log(d_p^{gt})\|^2}$

abs. rel: $\frac{1}{|P|}\sum_{p\in P}\frac{|d_p - d_p^{gt}|}{d_p^{gt}}$

sq. rel: $\frac{1}{|P|}\sum_{p\in P}\frac{\|d_p - d_p^{gt}\|^2}{d_p^{gt}}$

Accuracies: $\%$ of d_p s.t. $max(\frac{d_p}{d_p^{gt}}, \frac{d_p^{gt}}{d_p}) = \delta < thr$

The errors are computed at locations where both Kinect raw depth data is available and where depth regularization is performed (regions excluding the small border where predictions are not made). The regularized depth maps and CNN depth predictions are bilinearly upsampled to 640x480 resolution prior to evaluating against the raw Kinect depth maps. Note that the same optimisation and cost-volume-related parameters were used for comparing the two regularizer types. We follow the same θ scheduling policy as [4] with similar choice of parameters. The table, in particular the low threshold accuracy column, help validate that the normal-prior helps in recovering the fine details in the scene. Qualitative comparisons are shown for two NYU raw test sequences in Figures 5 and 6. The improvements in reconstruction in terms of both fine detail and global scene structure are apparent, especially in textureless regions.

The same experiments were carried out on the TUM dataset [38] and the living room sequence 'lr kt0' in the ICL-NUIM dataset [39]. Quantitative results for these sequences are also shown in Table I, and qualitative results for the TUM sequence is shown in Fig 7. Again a similar trend to that observed before can be seen. It can also be seen that CNN depth predictions do not generalize to new scene types as well as the other two methods.

While our experiments were limited to reconstructing indoor environments, the same framework in theory can be used for building dense maps of outdoor scenes, given the large depth range covered by the cost volume and large-scale volumetric fusion capabilities of [37]. However, the neural network (which is trained on indoor scenes) will likely require finetuning to adapt – this is yet to be validated. The main difficulty here is in acquiring densely labelled outdoor depth maps (required for generating ground truth normals) for training, although an unsupervised learning scheme similar to [40] should help in this regard.

V. CONCLUSION

In this work we presented a simple yet efficient solution that jointly exploits low-level geometry-based photometric evidence and high-level scene information captured from a multi-scale CNN architecture in the form of surface normals, for improving the accuracy of dense reconstructions in cases where otherwise there is very little photometric evidence. It was seen that incorporating learnt surface orientations enabled smooth and accurate reconstructions especially in areas with little photometric evidence to guide the solution. Deep learning has enabled prediction of geometry of objects and scenes directly from a single image and this alleviates the need for prior assumptions about scene structure, and handcrafted scene priors that are otherwise required for dense reconstruction. It was also seen that these networks are capable of generalizing to new types of environments well enough for practical use. We believe this work is a step forward in unifying the two complementary tasks of 3D reconstruction and scene understanding, aiding purely vision-based autonomous robots.

REFERENCES

[1] G. Klein and D. Murray, "Parallel tracking and mapping for small AR workspaces," in *Proc. Sixth IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR'07)*, Nara, Japan, November 2007.

[2] J. Engel, T. Schöps, and D. Cremers, "LSD-SLAM: Large-scale direct monocular SLAM," in *European Conference on Computer Vision (ECCV)*, September 2014.

[3] R. Mur-Artal, J. M. M. Montiel, and J. D. Tards, "Orb-slam: A versatile and accurate monocular slam system," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, Oct 2015.

[4] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison, "Dtam: Dense tracking and mapping in real-time," in *Proceedings of the 2011 International Conference on Computer Vision*, ser. ICCV '11. Washington, DC, USA: IEEE Computer Society, 2011, pp. 2320–2327.

[5] J. Stühmer, S. Gumhold, and D. Cremers, "Real-time dense geometry from a handheld camera," in *Proceedings of the 32Nd DAGM Conference on Pattern Recognition*. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 11–20.

[6] A. Flint, D. Murray, and I. Reid, "Manhattan scene understanding using monocular, stereo, and 3d features," in *Computer Vision (ICCV), 2011 IEEE International Conference on*, Nov 2011, pp. 2228–2235.

[7] A. Concha, M. W. Hussain, L. Montano, and J. Civera, "Manhattan and piecewise-planar constraints for dense monocular mapping," in *Robotics: Science and Systems X, University of California, Berkeley, USA, July 12-16, 2014*, 2014.

[8] A. Concha and J. Civera, "Dpptam: Dense piecewise planar tracking and mapping from a monocular sequence," in *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*, Sept 2015, pp. 5686–5693.

[9] A. Dame, V. Prisacariu, C. Ren, and I. Reid, "Dense reconstruction using 3d object shape priors," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, June 2013, pp. 1288–1295.

[10] S. Bao, M. Chandraker, Y. Lin, and S. Savarese, "Dense object reconstruction with semantic priors," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, June 2013, pp. 1264–1271.

[11] D. Herrera C., J. Kannala, L. Ladický, and J. Heikkila, "Depth map inpainting under a second-order smoothness prior," in *Image Analysis*. Springer Berlin Heidelberg, 2013, vol. 7944, pp. 555–566.

[12] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," *CoRR*, vol. abs/1411.4734, 2014.

[13] F. Liu, C. Shen, and G. Lin, "Deep convolutional neural fields for depth estimation from a single image," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2015. [Online]. Available: http://arxiv.org/abs/1411.6387

[14] R. Zhang, P.-S. Tsai, J. E. Cryer, and M. Shah, "Shape from shading: A survey," *IEEE Transactions on Pattern Analaysis and Machine Intelligence*, vol. 21, no. 8, pp. 690–706, 1999.

[15] C. Hane, L. Ladicky, and M. Pollefeys, "Direction matters: Depth estimation with a surface normal classifier," in *CVPR*, 2015, pp. 381–389.

[16] P. K. Nathan Silberman, Derek Hoiem and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *ECCV*, 2012.

[17] M. Pizzoli, C. Forster, and D. Scaramuzza, "Remode: Probabilistic, monocular dense reconstruction in real time," in *Robotics and Automation (ICRA), 2014 IEEE International Conference on*. IEEE, 2014, pp. 2609–2616.

[18] V. Pradeep, C. Rhemann, S. Izadi, C. Zach, M. Bleyer, and S. Bathiche, "Monofusion: Real-time 3d reconstruction of small scenes with a single web camera," in *ISMAR*, 2013.

[19] A. Concha and J. Civera, "Using superpixels in monocular slam," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*, May 2014, pp. 365–372.

[20] S. Y. Bao and S. Savarese, "Semantic structure from motion: A novel framework for joint object recognition and 3d reconstruction," in *Proceedings of the 15th International Conference on Theoretical Foundations of Computer Vision: Outdoor and Large-scale Real-world Scene Analysis*, Berlin, Heidelberg, 2012, pp. 376–397.

[21] L. Ladický, P. Sturgess, C. Russell, S. Sengupta, Y. Bastanlar, W. Clocksin, and P. Torr, "Joint optimization for object class segmentation and dense stereo reconstruction," *International Journal of Computer Vision*, vol. 100, no. 2, pp. 122–133, 2012.

[22] C. Hane, C. Zach, A. Cohen, R. Angst, and M. Pollefeys, "Joint 3d scene reconstruction and class segmentation," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, June 2013, pp. 97–104.

[23] A. Kundu, Y. Li, F. Dellaert, F. Li, and J. Rehg, "Joint semantic segmentation and 3d reconstruction from monocular video," in *Computer Vision ECCV 2014*, ser. Lecture Notes in Computer Science, 2014, vol. 8694, pp. 703–718.

[24] D. Hoiem, A. A. Efros, and M. Hebert, "Recovering surface layout from an image," *Int. J. Comput. Vision*, vol. 75, no. 1, pp. 151–172, Oct. 2007.

[25] L. Ladický, B. Zeisl, and M. Pollefeys, "Discriminatively trained dense surface normal estimation," in *ECCV*, 2014, vol. 8693, pp. 468–484.

[26] D. F. Fouhey, A. Gupta, and M. Hebert, "Data-driven 3d primitives for single image understanding," in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 3392–3399.

[27] X. Wang, D. F. Fouhey, and A. Gupta, "Designing deep networks for surface normal estimation," in *CVPR*, 2015.

[28] J.-D. Durou, Y. Quéau, and J.-F. Aujol, "Normal Integration – Part I: A Survey," June 2016. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01334349

[29] K. Kolev, T. Pock, and D. Cremers, "Anisotropic minimal surfaces integrating photoconsistency and normal information for multiview stereo," in *Proceedings of the 11th European Conference on Computer Vision Conference on Computer Vision: Part III*, ser. ECCV'10, 2010, pp. 538–551.

[30] C. Hane, N. Savinov, and M. Pollefeys, "Class specific 3d object shape priors using surface normals," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, 2014.

[31] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning," *ArXiv e-prints*, June 2015.

[32] R. Rockafellar, *Convex Analysis*, ser. Princeton landmarks in mathematics and physics. Princeton University Press, 1997.

[33] F. Steinbrücker, T. Pock, and D. Cremers, "Large displacement optical flow computation without warping," in *2009 IEEE 12th International Conference on Computer Vision*. IEEE, 2009, pp. 1609–1614.

[34] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *Journal of Mathematical Imaging and Vision*, vol. 40, no. 1, pp. 120–145, 2010.

[35] http://www.nvidia.com.

[36] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.

[37] V. A. Prisacariu, O. Kahler, M. M. Cheng, C. Y. Ren, J. Valentin, P. H. S. Torr, I. D. Reid, and D. W. Murray, "A Framework for the Volumetric Integration of Depth Images," *ArXiv e-prints*, 2014.

[38] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of rgb-d slam systems," in *Proc. of the International Conference on Intelligent Robot Systems (IROS)*, Oct. 2012.

[39] A. Handa, T. Whelan, J. McDonald, and A. Davison, "A benchmark for RGB-D visual odometry, 3D reconstruction and SLAM," in *IEEE Intl. Conf. on Robotics and Automation, ICRA*, Hong Kong, China, May 2014.

[40] R. Garg, V. Kumar, G. Carneiro, and I. Reid, "Unsupervised cnn for single view depth estimation: Geometry to the rescue," in *European Conference on Computer Vision, (ECCV)*, 2016.