Introduction to Indoor GeoNet

Amirreza Farnoosh and Sarah Ostadabbas*
Augmented Cognition Lab (ACLab)
Northeastern University
Boston, MA, USA
ostadabbas@ece.neu.edu *

Abstract

This paper introduces "Indoor GeoNet", a weakly supervised depth and camera pose estimation model targeted for indoor scenes. Our paper is built upon previous works in the field of unsupervised depth and relative camera pose estimation from temporal consecutive video frames through novel view synthesis using deep learning models. However, such works are exclusively trained/tested on a few available outdoor scene datasets and we have shown they are hardly transferable to new scenes, especially to indoor environments, in which estimation requires higher precision and dealing with probable occlusions. In Indoor GeoNet, we used a hybrid learning framework introduced in a recent work called GeoNet, and took advantage of the availability of indoor RGBD datasets collected by human or robot navigators, and added partial (i.e. weak) supervision in depth training into the model. Experimental results showed that our model effectively generalizes to new scenes from different buildings. It demonstrated significant depth and pose estimation error reduction when compared to the original GeoNet, while showing 3 times more reconstruction accuracy in synthesizing novel views in indoor environments.

1. Introduction

Information such as depth and relative camera pose are important in the sense that they can be used together to give a very accurate and detailed representation of an indoor scene [12]. These information could facilitate both navigation and dynamic interaction and also help to reconstruct a unified 3D model of the scene for the purpose of map generation of unknown places [3, 6, 17] or even adding augmented reality features to the scene for a better interaction experience [12]. It can also be used for virtual tours of an indoor scene while the observer looks into the rendered scenes in different views [5].

In computer vision field, there has been extensive research for indoor odometry, scene understanding, and specifically camera pose, depth, and flow extraction from a moving camera (e.g. robot or head mounted), most of which are recently powered with the advances in deep learning (DL) used in visual simultaneous localization and mapping (vSLAM) works [2]. The use of DL in vSLAM can be separated into two categories of supervised [4, 5, 10, 16] and unsupervised learning [7,8,14,20,21,24]. The common methodology behind all of the recent unsupervised methods for vSLAM is warping one image in pairs of related images (either stereo pairs or consecutive frames in a video) to the other view by leveraging the geometry constraints of the problem, in an approach very similar to the idea of autoencoders. Although such unsupervised configurations can be trained on any amount of data without labeling cost, they still fall behind supervised methods in terms of the estimation accuracy, and are hardly generalizable to new scenes which are not seen *apriori* by the network. Besides that, almost all of these works are focused on outdoor scenes. We argue that this approach is not well transferable to indoor scenes, in contrast to outdoor scenes, for several reasons: high relative displacement with respect to the depth range, sharper changes in relative camera view, and need for higher depth precision in indoor scenes. Inspired by the GeoNet framework presented in [23], in this paper we propose our "Indoor GeoNet" model, which is a *weakly* supervised hybrid learning approach for camera pose, depth and flow estimation targeted to indoor scenes. Capitalized on the availability of the inexpensive depth sensing (e.g. Microsoft Kinect and Intel RealSense), we introduce the weak supervision by providing a set of groundtruth depth data into the model during the training. This type of supervision is weak due to the fact that the model is only partially supervised on depth data and the camera pose needs to be learned in an unsupervised fashion. We also believe that this kind of weak supervision for indoor scene understanding has recently become viable since the release of several RGBD open-source datasets collected by human or robot naviga-

[*] Source code available at: https://web.northeastern.edu/ostadabbas/software/

tors such as RSM Hallway, and MobileRGBD datasets [1].

2. Building the Indoor GeoNet

Indoor GeoNet shares the same network structure as original GeoNet [23], in which two sub-networks called DepthNet and PoseNet predict rigid layout of the observed scene including the depth and relative camera pose. The training samples to the network are temporal consecutive frames $I_i(i = 1 \sim N)$ with known camera intrinsics. Typically in a sequence of frames, a reference frame I_r is specified as the reference view, and the other frames are target frames I_t. During training, the DepthNet takes the entire sequence concatenated along batch dimension as input. This allows for single view depth prediction at the test time. In contrast, the PoseNet is naturally fed with the entire sequence concatenated along channel dimension, and outputs all of the relative camera poses. This allows the network to learn the connections between different views in a sequence. Fused with the deep structures of DepthNet and PoseNet, rigid scene geometry equations then will be used to warp a target view to the reference view. Unlike the fully unsupervised approach of original GeoNet, Indoor GeoNet takes advantage of the depth supervision to enhance the transferability of the pose and depth learning to indoor scenes.

2.1. Geometric-Based Hybrid Learning

The rigid optical flow in a sequence of frames is governed by the relative camera motion between the observer and the scene, and can be completely modeled by a collection of depth maps D_i for frame I_i, and the relative camera motion $\mathbf{T}_{r\to t} = [R|T]$ from reference frame I_r to target frame I_t, where $R_{3\times 3}$ and $T_{3\times 1}$ represent the relative rotation and displacement matrices, respectively. Let $p_r = [X, Y, 1]^T$ denote the homogeneous coordinates of a pixel in the reference view, D_{p_r} be its depth value, $\mathcal{P} = [x, y, D_{p_r}, 1]^T$ be its corresponding homogeneous 3D coordinates (referenced on camera pinhole), and $K_{3\times 3}$ be the camera intrinsic matrix. Then, p_r in the image plane is:

$$p_r = D_{p_r}^{-1}\mathbf{K}\big[\mathbf{I}_{3\times 3}|\mathbf{0}\big]\mathcal{P} \tag{1}$$

Moreover, we can obtain the projected coordinates of p_r onto the target view p_t, as:

$$p_t \sim \mathbf{K}\mathbf{T}_{r\to t}\mathcal{P} = \mathbf{K}[R|T]\mathcal{P} \tag{2}$$

Rewriting the Eq. (1) will result in $[x, y, D_{p_r}]^T = D_{p_r}\mathbf{K}^{-1}p_r$ and merging that with the Eq. (2) will give us the corresponding target pixel coordinates p_t in terms of the reference depth map D_{p_r}, reference pixel coordinates p_r, and the relative camera motion $[R|T]$, as:

$$p_t \sim \mathbf{K}\Big[D_{p_r}R\mathbf{K}^{-1}p_r + T\Big] \tag{3}$$

Using Eq. (3), we can synthesize a novel nearby view from a reference frame in non-occluded regions having the depth map of pixels in the reference view as well as the relative camera pose between the views. Therefore, the DepthNet and PoseNet can be trained together through novel view synthesis between any pairs of training samples.

2.2. Weakly Supervised Multi-Objective Training

Let us denote consecutive frames $\{I_1, \ldots, I_r, \ldots, I_N\}$ as a training sequence with the middle frame I_r being the reference view and the rest being the target views, I_t's. Then, $\hat{I}_{t\to r}$ represents the target view I_t warped to the reference coordinate frame by taking the predicted depth $\hat{D}_r$, the predicted camera transformation matrix $\hat{T}_{r\to t}$, and the target view I_t as input. In order to train the Indoor GeoNet in a weakly supervised manner, we define a total loss function $\mathcal{L}_T$ as the weighted summation of multiple losses as:

$$\mathcal{L}_T = \sum_{(r,t)} \lambda_P\mathcal{L}_P + \lambda_D\mathcal{L}_D + \lambda_C\mathcal{L}_C + \lambda_W\mathcal{L}_W \tag{4}$$

where $\mathcal{L}$'s are different loss functions explained in the following sections, λ's are the corresponding loss weights, and (r, t) iterates over all possible pairs of reference I_r and target I_t frames.

2.2.1 Photometric Loss: $\mathcal{L}_P$

The DepthNet and PoseNet networks can be trained by minimizing the photometric loss between the synthesized view (warped target view) $\hat{I}_{t\to r}$ and reference frame I_r:

$$\mathcal{L}_P = \sum_{(r,t)}\sum_{p_r} F_{diss}\big(I_r(p_r), \hat{I}_{t\to r}(p_r)\big) \tag{5}$$

where $\hat{I}_{t\to r}(p_r) = I_t(p_t)$, with warping between p_t and p_r obtained from Eq. (3) using differentiable bilinear sampling [13], and $F_{diss}(.)$ is a dissimilarity measure. We adopted the differentiable photometric dissimilarity measure proposed in [8], which has proven to be successful in measuring perceptual image similarity, and handling occlusions:

$$F_{diss}(I_r, \hat{I}_{t\to r}) = \alpha\frac{1 - \text{SSIM}\big(I_r, \hat{I}_{t\to r}\big)}{2} + (1-\alpha)\big\|I_r - \hat{I}_{t\to r}\big\|_1 \tag{6}$$

where SSIM denotes the structural similarity index [22] and α is taken to be 0.85.

2.2.2 Depth Smoothness Loss: $\mathcal{L}_D$

We used a depth map smoothness loss term $\mathcal{L}_D$ weighted per-pixel by image gradients as proposed in [23] in order to obtain coherent depth maps while allowing depth discontinuities on the edges of the image:

[1] www.bicv.org/datasets/rsm-dataset/, www.mobilergbd.inrialpes.fr/

$$\mathcal{L}_D = \sum_{p_r} |\nabla D_r(p_r)|.\big(\exp(-|\nabla I_r(p_r)|\big)^T \qquad (7)$$

where ∇ is the vector differential operator.

2.2.3 Forward-Backward Consistency Loss: $\mathcal{L}_C$

We applied a forward-backward consistency check as in [23] to enhance our predictions. Pixels for which the forward and backward flows disagree significantly are considered as possible occluded regions and are excluded from both the photometric loss $\mathcal{L}_P$, and the forward-backward flow consistency check. Let us denote $f_{r\to t}(p_r) = p_r - p_t^{\{D_{p_r},\mathbf{T}\}}$ as forward flow (p_t is computed using Eq. (3)), and conversely $f_{t\to r}(p_r) = p_r^{\{D_t,\mathbf{T}^{-1}\}} - p_t$ as backward flow, and $\Delta f_{t,r}(p_r) = f_{r\to t}(p_r) - f_{t\to r}(p_r)$. Then, the geometry consistency is imposed by adding the following loss term:

$$\mathcal{L}_C = \sum_{p_r \in \mathbf{p_r}} \left\|\Delta f_{t,r}(p_r)\right\|_1$$

such that

$$\mathbf{p_r} = \big\{p_r : \|\Delta f_{t,r}(p_r)\|_2 < \max(\alpha, \beta\|f_{r\to t}(p_r)\|_2)\big\} \qquad (8)$$

in which (α, β) are set to be $(3.0, 0.05)$.

2.2.4 Weak Supervision Loss: $\mathcal{L}_W$

In order to enhance the overall performance of the network in prediction of depth and camera pose, we enforced the groundtruth depth maps, D^{gt}, by introducing another loss term on pixel locations for which we have the groundtruth depth values:

$$\mathcal{L}_W = \sum_{i\in r,t} \left\|D_i - D_i^{gt}\right\|_2 \qquad (9)$$

where D_i and D_i^{gt} are the predicted and groundtruth depth maps of training samples, respectively.

2.3. Network Architecture and Implementation Details

The Indoor GeoNet contains two sub-networks, the DepthNet, and the PoseNet, similar to the original GeoNet structure [23]. The DepthNet consists of an encoder part and a decoder part. The encoder part has the structure of ResNet50 [9] and the decoder part uses deconvolution layers to enlarge predicted depth maps to their original resolution (as input) in a multi-scale scheme. The PoseNet has the same architecture as in [23], which consists of 8 convolutional layers followed by a global average pooling layer that outputs the 6-DoF camera poses including rotation and translation. We used batch normalization [11] and ReLU activation functions [19] for all of the convolutional layers except the prediction layers. We considered the training sequence length to be $N = 5$, and resized all the RGB frames to 144×256 pixels, and then trained the network with learning rate of 0.0002, and batch size of 4 for 20 epochs in Tensorflow [1]. We set the loss weights to be $\lambda_P = \lambda_W = 1, \lambda_D = 0.5, \lambda_C = 0.2$, and used Adam optimizer [15] with its parameters set as $\beta_1 = 0.9, \beta_2 = 0.999$ for network training.

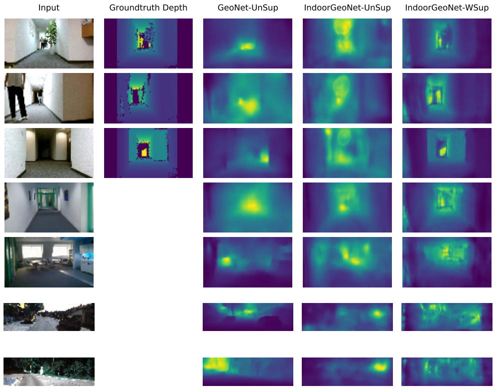

Figure 1. Sample examples of depth image prediction comparing IndoorGeoNet-WSup with other models. The rows 1-3 are samples from MobileRGBD, rows 4-5 are from RSM Hallway, and rows 6-7 are from KITTI datasets.

3. Experimental Results and Evaluation

The weakly supervised Indoor GeoNet (i.e. IndoorGeoNet-WSup) performance is evaluated against two other models, one the original GeoNet trained on KITTI raw and odometry datasets [18] (referred to as GeoNet-UnSup) and the other one the GeoNet trained from scratch in an unsupervised manner on the RSM Hallway dataset (referred to as IndoorGeoNet-UnSup). The performance comparison is done in two aspects: (1) the accuracy of depth and camera pose estimation, (2) the reconstruction accuracy of the novel RGB scene synthesis using different approaches. The quantified results are calculated and reported for the datasets with available groundtruth depth and pose labels.

Table 1. Depth and relative camera pose estimation performance comparison between different methods.

		Depth RMSE		Pose (trajectory) RMSE	
Method	Training Dataset	KITTI Raw	MobileRGBD	KITTI Odometry	MobileRGBD
GeoNet-UnSup	KITTI	4.01	1.24	0.012	0.042
IndoorGeoNet-UnSup	RSM Hallway	13.37	1.14	0.057	0.034
IndoorGeoNet-WSup	MobileRGBD	12.82	**0.72**	0.051	**0.006**

3.1. Depth and Pose Estimation

We reported the depth and relative camera pose root mean squared error (RMSE) on the test set for those datasets for which we have the groundtruth values in Table 1 for the three models. Although GeoNet-UnSup works pretty well on the KITTI datasets, its performance degrades significantly on indoor datasets compared to the IndoorGeoNet-WSup, proving that the model is not generalizable to the indoor scenes. We also depicted some sample figures of depth prediction for the three models side by side in Fig. 1 along with groundtruth depth maps (if available) for comparison. As seen in this figure, although GeoNet-UnSup predicts satisfactory depth maps for the KITTI dataset, its predicted depth maps for MobileRGBD and RSM Hallway samples, hardly give any information about the general geometry of the scene, and the edges are completely lost. On the other hand, IndoorGeoNet-UnSup gives a fair prediction of the global geometry of the scene on sample images of RSM Hallway (on which the model is trained), however, predicted depth maps completely miss the details. The predictions of this model on the MobileRGBD sample images (not seen by the model during training) show that this model also fails to adapt to a new unseen indoor scene.

As seen in Table 1, with IndoorGeoNet-WSup model, depth and pose errors drop significantly for MobileRGBD dataset as compared to other models, since we are adding the depth supervision. Its predicted depth maps on sample images of MobileRGBD dataset clearly demonstrates the effect of supervision (even weak) on the ability of the network to learn detailed maps as shown in Fig. 1. IndoorGeoNet-WSup also shows acceptable depth image estimation on other indoor scene (e.g. RSM Hallway) that were not part of weak supervision. This demonstrates the generalization capability of the proposed IndoorGeoNet-WSup.

3.2. Novel View Reconstruction Estimation

The mean image photometric loss $\mathcal{L}_P$, plus the structural similarity index measure between the reference image and the inverse warped target image are reported in table Table 2 for all of the dataset on our three models. Evident from this table, for GeoNet-UnSup, the reconstruction loss increases significantly on the MobileRGBD and RSM Hallway datasets that are not seen by the network during the training, which proves that the network fails to adapt to the indoor scenes. IndoorGeoNet-UnSup gives the lowest reconstruction $\mathcal{L}_P$ loss on RSM Hallway dataset (on which its network is trained), however, IndoorGeoNet-WSup also gives a comparable SSIM on this dataset, although it has not seen the dataset during the training. As expected, IndoorGeoNet-WSup gives the best reconstruction results on test set of MobileRGBD dataset with which the network is trained in a weak supervision fashion. We also depicted some sample images of novel view reconstruction using the three models in Fig. 2. As seen in this figure, IndoorGeoNet-WSup is able to successfully reconstruct novel nearby view from input images on both sample images of MobileRGBD and RSM Hallway. Although GeoNet-UnSup works well on KITTI sample images, it fails to correctly reconstruct the novel view of indoor scenes. Using the IndoorGeoNet-UnSup model, the reconstructed views of RSM Hallway sample images are acceptable, because as we discussed in the previous subsection, its predicted depth maps give a fair estimation of the global geometry of the scene.

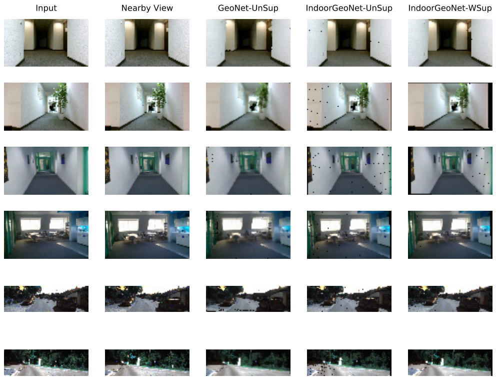

Figure 2. Novel view synthesis samples comparing the reconstruction results of IndoorGeoNet-WSup with other models. The rows 1-2 are samples from MobileRGBD, rows 3-4 are from RSM Hallway, and rows 5-6 are from KITTI datasets.

4. Conclusion

In this work, we proposed "Indoor GeoNet" using a weak supervision in terms of depth to improve both depth and pose predictions for indoor datasets. We believe that such supervision is sensible due to the availability of inexpensive indoor RGB and depth sensors and several open-source indoor datasets. We compared the outcomes of our Indoor GeoNet in terms of depth, camera pose and novel view

Table 2. Novel view synthesis performance comparison between different methods. Please note that while reconstruction error (RMSE) is desired to be low, the reconstruction similarity measure (SSIM) is desired to be close to 1.

Method	Training Dataset	Reconstruction $\mathcal{L}_P$ Loss (RMSE)			Reconstruction Similarity (SSIM)		
		KITTI	RSM Hallway	MobileRGBD	KITTI	RSM hallway	Mobile RGBD
GeoNet-UnSup	KITTI	32.71	28.65	32.92	0.29	0.12	0.20
IndoorGeoNet-UnSup	RSM Hallway	50.20	18.62	26.94	0.21	0.40	0.43
IndoorGeoNet-WSup	MobileRGBD	47.42	23.74	**21.2**	0.24	**0.38**	**0.49**

estimation with the original unsupervied GeoNet models trained on different benchmark datasets. The results revealed that Indoor GeoNet is able to detect more detailed depth maps and also the pose estimation is improved when applied on indoor datasets.

References

[1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. Tensorflow: a system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016. 3

[2] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. J. Leonard. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Transactions on Robotics*, 32(6):1309–1332, 2016. 1

[3] M. Dzitsiuk, J. Sturm, R. Maier, L. Ma, and D. Cremers. De-noising, stabilizing and completing 3d reconstructions on-the-go using plane priors. In *Robotics and Automation (ICRA), 2017 IEEE International Conference on*, pages 3976–3983. IEEE, 2017. 1

[4] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*, pages 2366–2374, 2014. 1

[5] S. A. Eslami, D. J. Rezende, F. Besse, F. Viola, A. S. Morcos, M. Garnelo, A. Ruderman, A. A. Rusu, I. Danihelka, K. Gregor, et al. Neural scene representation and rendering. *Science*, 360(6394):1204–1210, 2018. 1

[6] A. Farnoosh, M. Nabian, P. Closas, and S. Ostadabbas. First-person indoor navigation via vision-inertial data fusion. In *Position, Location and Navigation Symposium (PLANS), 2018 IEEE/ION*, pages 1213–1222. IEEE, 2018. 1

[7] R. Garg, V. K. BG, G. Carneiro, and I. Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *European Conference on Computer Vision*, pages 740–756. Springer, 2016. 1

[8] C. Godard, O. Mac Aodha, and G. J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, volume 2, page 7, 2017. 1, 2

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[10] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *IEEE conference on computer vision and pattern recognition (CVPR)*, volume 2, page 6, 2017. 1

[11] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 3

[12] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison, et al. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 559–568. ACM, 2011. 1

[13] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015. 2

[14] J. Y. Jason, A. W. Harley, and K. G. Derpanis. Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness. In *European Conference on Computer Vision*, pages 3–10. Springer, 2016. 1

[15] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 3

[16] M. Liu, X. He, and M. Salzmann. Geometry-aware deep network for single-image novel view synthesis. *arXiv preprint arXiv:1804.06008*, 2018. 1

[17] R. Maier, R. Schaller, and D. Cremers. Efficient online surface correction for real-time large-scale 3d reconstruction. *In British Machine Vision Conference (BMVC)*, 2017. 1

[18] M. Menze and A. Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3061–3070, 2015. 3

[19] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010. 3

[20] A. Ranjan, V. Jampani, K. Kim, D. Sun, J. Wulff, and M. J. Black. Adversarial collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. *arXiv preprint arXiv:1805.09806*, 2018. 1

[21] S. Vijayanarasimhan, S. Ricco, C. Schmid, R. Sukthankar, and K. Fragkiadaki. Sfm-net: Learning of structure and motion from video. *arXiv preprint arXiv:1704.07804*, 2017. 1

[22] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 2

[23] Z. Yin and J. Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, 2018. 1, 2, 3

[24] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, volume 2, page 7, 2017. 1