# Semi-Supervised 3D Structural In-variance for World Coordinates Prediction for realistic AR object placement

Akankshya Kar Anand Bhoraskar
{akankshk,abhorask}@andrew.cmu.edu

## PROBLEM STATEMENT

Recent works focusing on Indoor scenes try to predict 3D world coordinates and place objects in real world for Augmented reality applications, specially for Monocular hand held phones and devices. This is challenging because indoor scenes have walls, corridors and featureless planar objects. Height and Uprightness Invariance for 3D Prediction from a Single View [1] in CVPR 2020 tries to solve this by predicting the camera intrinsics and the camera position with respect to the ground. It then predicts the 3D position for each pixel. However the 3D structural accuracy of this method is worse compared to DORN [4] which was the earlier state of the art in monocular depth estimation. Geometry based monocular SLAM methods which work well in outdoor environments for self-driving cars, fail indoors due to large texture-less regions such as walls. In order to solve this, we borrow ideas from [5] [10] which use global and local planar constraints by modifying loss functions and as this paper performs poorly than DORN, we will modify the network to enable Local Planar guidance from BTS [6]. Finally, using this we will find the world coordinates for the system with accurate structure and project realistic placement of AR objects in indoor scenes

## PROPOSED METHOD

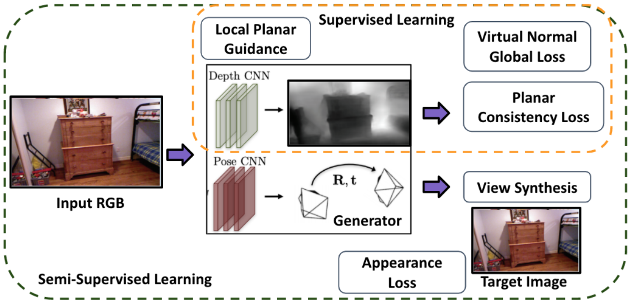

**Figure 1: Pipeline**

The pipeline we proposed is described in Figure 1. We will follow a semi-supervised approach with separate loss functions based on ground truth depth and photo-metric consistency.

For supervised learning, we will predict the depth map, intrinsics and extrinsics with respect to ground. This will enable us to use the regression loss on world coordinates similar to [1]. Additionally, we will change the depth architecture from Megadepth [7] to follow BTS architecture which encapsulates local planar Guidance. In terms of loss functions for supervised setting from either [10] or from [9] for accurate depth prediction. See figure 2 for more information on BTS constraints and Re-porojection error loss.

For the self-supervised learning, we will use the neighbouring frames in the video for an appearance based loss. For this we will predict the pose between the frame and each neighbouring frame and reproject pixels from the source frame using the depth map. We will include a photometric error term similar to [5].

The deliverable for the method is the DepthNet which will be used to predict camera parameters and the depth map which will enable AR object placement.

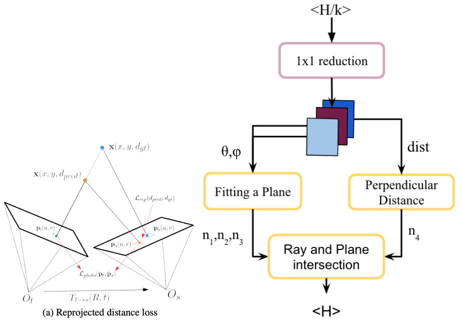
$\mathbf{x}(x, y, d_{gt})$
$\hat{\mathbf{x}}(x, y, d_{pred})$
$\mathcal{L}_{rep}(d_{pred}, d_{gt})$
$\mathbf{p}_t(u, v)$
$\mathbf{p}_s(u, v)$
$\mathbf{p}_s(\hat{u}, \hat{v})$
$\mathcal{L}_{photo}(\mathbf{p}_t, \hat{\mathbf{p}}_s)$
$O_t$
$T_{t \to s}(R, t)$
$O_s$
(a) Reprojected distance loss
<H/k>
1x1 reduction
θ,φ
dist
Fitting a Plane
Perpendicular Distance
$n_1, n_2, n_3$
$n_4$
Ray and Plane intersection
<H>

**Figure 2: Local planar constraint in Depth Module**

We desire a dataset with annotated ground truth for supervised learning, video sequences for self-supervised learning and comparison with state-of-the-art depth prediction networks. Scannet[3], Matterport3D [2] NYUv2[8] are datasets which satisfy our conditions. We will decide on the basis of ease of access.

## REFERENCES

[1] Manel Baradad and Antonio Torralba. 2020. Height and Uprightness Invariance for 3D Prediction From a Single View. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 491–500.

[2] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. 2017. Matterport3D: Learning from RGB-D Data in Indoor Environments. *International Conference on 3D Vision (3DV)* (2017).

[3] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. 2017. ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*.

[4] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. 2018. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 2002–2011.

[5] Vitor Guizilini, Jie Li, Rares Ambrus, Sudeep Pillai, and Adrien Gaidon. 2020. Robust Semi-Supervised Monocular Depth Estimation With Reprojected Distances. In *Conference on Robot Learning*. PMLR, 503–512.

[6] Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh. 2019. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326* (2019).

[7] Zhengqi Li and Noah Snavely. 2018. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 2041–2050.

[8] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. 2012. Indoor Segmentation and Support Inference from RGBD Images. In *ECCV*.

[9] Wei Yin, Yifan Liu, Chunhua Shen, and Youliang Yan. 2019. Enforcing geometric constraints of virtual normal for depth prediction. In *Proceedings of the IEEE International Conference on Computer Vision.* 5684–5693.

[10] Zehao Yu, Lei Jin, and Shenghua Gao. 2020. P{2} Net: Patch-match and Plane-regularization for Unsupervised Indoor Depth Estimation. *arXiv preprint arXiv:2007.07696* (2020).