IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2023, London, UK. Cite as RA-L paper.

Stereo Plane R-CNN: Accurate scene geometry reconstruction using planar segments and camera-agnostic representation

Jan Wietrzykowski[1] and Dominik Belter[1]

***Abstract*—The article introduces a novel method for planar segments detection and description from a stereo pair of images. The existing systems for planes detection utilize single RGB images and have accuracy- and scale-related problems regarding 3D reconstruction with the obtained planar segments. The proposed approach draws inspiration from deep-learning-based systems for plane detection and depth reconstruction. Firstly, we improve the planes detection in the image. Secondly, we enhance geometry reconstruction accuracy using a stereo setup. To achieve the 3D model of the observed planes, we introduce a novel neural network architecture and training strategy that jointly optimizes the prediction of disparity, normal vectors, and plane parameters. Moreover, the proposed approach utilizes an efficient camera-agnostic representation of the problem. Finally, we show that our system outperforms existing approaches to planar segments detection and parameters estimation and improves the reconstruction accuracy of indoor environments.**

***Index Terms*—Deep Learning for Visual Perception; Mapping; Semantic Scene Understanding**

I. Introduction

A KEY factor in introducing camera-based localization systems to everyday life is their robustness. One way to improve the robustness is to include a relocalization mechanism that uses higher-abstraction-level objects as matched features [1], [2]. Viable alternatives to point features are planar segments because they can be reliably detected and are common in man-made environments [2]. However, precise 3D pose estimation of those segments is crucial, not only for camera localization [3] but also for scene reconstruction [4]. Unfortunately, geometry reconstruction of planar segments using a monocular camera is a difficult task [2], [5] due to problems with metric scale estimation and ambiguity in the orientation of planes. A solution can be to use RGB-D sensors that became widely used in the last years. However, their effective range is limited, which leads to discarding a lot of beneficial information about distant regions of the scene [1]. Promising alternatives are stereo cameras that have a larger effective range than the available RGB-D cameras and enable more accurate geometry reconstruction than a monocular camera [6]. However, most of the stereo-based reconstruction research focuses on dense depth estimation from a pair of images [6], [7] and neglects the role of position and orientation of higher-level planar features, useful in the localization and scene reconstruction.

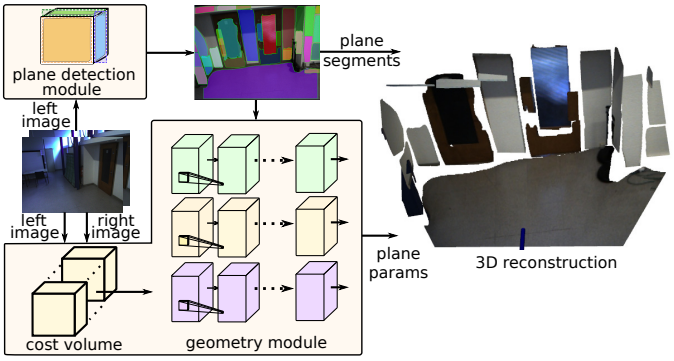

Fig. 1. Architecture of the proposed system. The plane detection module detects planes on a single RGB image, and a novel geometry module utilizes stereo pair of images to estimate the 3D poses of the detected planes.

Almost all state-of-the-art research on planar segments detection is focused on the application of Deep Neural Networks (DNNs) to images from monocular cameras [2], [3], [4]. This group includes the Plane R-CNN [5] that uses an architecture based on Mask R-CNN [8] and achieves state-of-the-art detection performance. However, our tests indicate that the performance of this method on a different dataset drops significantly, especially for distant planes. Also, geometry reconstruction is still unsatisfactory due to problems stemming from using a monocular camera, namely, inaccurate estimates of distances to planes and normal vectors.

Considering the limitations of the existing methods, we propose a new DNN-based system that detects planes and utilizes depth information encoded in a stereo pair of images to estimate 3D plane parameters. We bridge a gap between RGB-based plane detectors and systems that densely estimate depth from stereo pairs of images to obtain an accurate set of planar segments describing the scene (Fig. 1). Our contribution can be summarized as follows[1]:

- An improved plane segmentation method that deals with the problem of suppressed plane segments in the detection

Manuscript received: September 24, 2021; Revised January 12, 2022; Accepted February 3, 2022.

This paper was recommended for publication by Editor Cesar Cadena upon evaluation of the Associate Editor and Reviewers' comments. This work was supported by the National Science Centre (NCN), contract no. UMO-2018/31/N/ST6/00941.

[1]Jan Wietrzykowski and Dominik Belter are with Institute of Robotics and Machine Intelligence, Poznan University of Technology, 60-965 Poznań, Poland name.surname@put.poznan.pl

Digital Object Identifier (DOI): see top of this page.

[1]Implementation and dataset are available at https://github.com/LRMPUT/sprcnn

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2023, London, UK. Cite as RA-L paper.

methods based on Regions Of Interest (ROIs) and Non-Maximum Suppression (NMS).
- A novel neural network architecture that leverages disparity information from a stereo camera to accurately reconstruct scene geometry.
- A camera-agnostic normal vector representation that improves the robustness of the neural network to changes of the camera parameters that naturally arise when deploying a system in a real-life scenario.
- A training procedure that simultaneously utilizes global parameters of planes, pixel-wise normal vectors, and disparity prediction to improve the accuracy of plane parameters estimation.
- A fully automatically generated photorealistic synthetic dataset containing stereo images annotated with planar segments.

II. Related work

A. Pixel-wise depth and normal estimation

Research on scene geometry estimation focuses on recovering pixel-wise depth information from a single image [9], [10]. However, the work by Smolyanskiy *et al.* [6] argues that a 3D precise geometric reconstruction requires a stereo camera. They also propose a semi-supervised method for learning depth prediction. The ground truth data utilizes 3D laser scanner measurements and is augmented by unsupervised photoconsistency evaluation between stereo images. Convolutional neural networks (CNNs) have also been proven to be efficient in estimating normal vectors for each pixel of an RGB image [11]. However, recent work suggests that a coupled estimation of normal vectors and depth values provides more consistent and accurate estimates [12]. Also, Kusupati *et al.* [13] show that joint learning depth and normal vectors and enforcing consistency give significantly better results than separate learning. A generalized approach to consistency learning demonstrated on normal and depth estimation is presented in [14]. Unfortunately, knowledge about depth and surface normals could be only supplementary information for the generation of geometric features. Nonetheless, in this article, we follow the joint learning approach and optimize simultaneously losses related to disparity, pixel-wise normal vectors, and plane parameters estimation to provide more accurate results.

B. Detection of planar segments

Planar segments are a promising alternative to pixel-wise geometry reconstruction. Indoor environments are rich in planar segments and can be described just by a few of these geometric primitives. Another advantage is that the geometric properties of the underlying infinite planes can be easily described by a linear equation with only 4 parameters. Our initial experiments [15] prove that planes are also suitable for the global localization methods. An end-to-end approach to recover 3D planes from a single image is presented in [16], where supervision of learning is done indirectly by using depth ground truth. The parameters of the detected planes are estimated from values extracted from the latent space of the neural network. The limitation of their method is that only five planar segments can be detected in the scene, and learning requires a complete depth map for every training image. A limited number of planar segments can be also processed by the PlaneNet method [17]. The architecture of the neural network proposed in [17] contains separate branches for plane segmentation and for estimation of plane parameters. In our research, we also utilize a separate branch for estimation of plane parameters, but, at the same time, we train dense branches for disparity and normal vector estimation and show that this approach provides more accurate results. The SVPNet method [18] focuses on the binary classification of pixels into planar and non-planar segments and the extraction of embeddings where the same plane instances are close to each other. The planar segments are extracted using the mean-shift algorithm, and the plane parameters are estimated for each pixel in the first processing stage. In [5] the proposed Plane R-CNN method detects planar regions and reconstructs a piecewise planar depth map from a single RGB image. The plane instances are detected using a network based on the Mask R-CNN [8]. Then, a segmentation refinement network improves the consistency of the detected planar segments. The depth image is estimated directly from the RGB image by the decoder connected to the feature pyramid network (FPN) [19]. Estimation of the normal vectors consists of two components. The classifier picks one of the seven anchor normal directions and separately estimates the residual 3D vector. We, however, use a direct normal estimation because it provides better results.

C. 3D reconstruction and applications

Detected planar segments can be used in further scene reconstruction. Park and Yoon [20] show that stereo matching and disparity estimation can be improved by plane hypotheses generation and global optimization with plane hypotheses. In [21], information about planar segments from a single omnidirectional camera image is used to design plane-aware loss that improves normal vectors' predictions accuracy. The Plane R-CNN has been already used to detect and reconstruct planes that are occluded by other objects on a single camera view [22]. Ye *et al.* [2] added a plane description network which is later used to match detected planes between images and estimate the motion of the camera. Also, Xi and Chen [3] show that multi-view regularization of planar segments improves the reconstruction of indoor scenes. Another approach is presented in [4] where the neural networks predict if planes are orthogonal, parallel and if two planes are in contact. In our method, the stereo pair of images is used instead of multiple views and the regularization of two views is embedded in a new DNN architecture designed by us to recover 3D geometry. We focus on the accurate estimation of 3D geometry because the correct poses of planes are crucial for camera localization [2] and scene reconstruction [21].

III. Detection and reconstruction of planar segments

The proposed network consists of two main components: a plane detection module and a geometry module (Fig. 1).

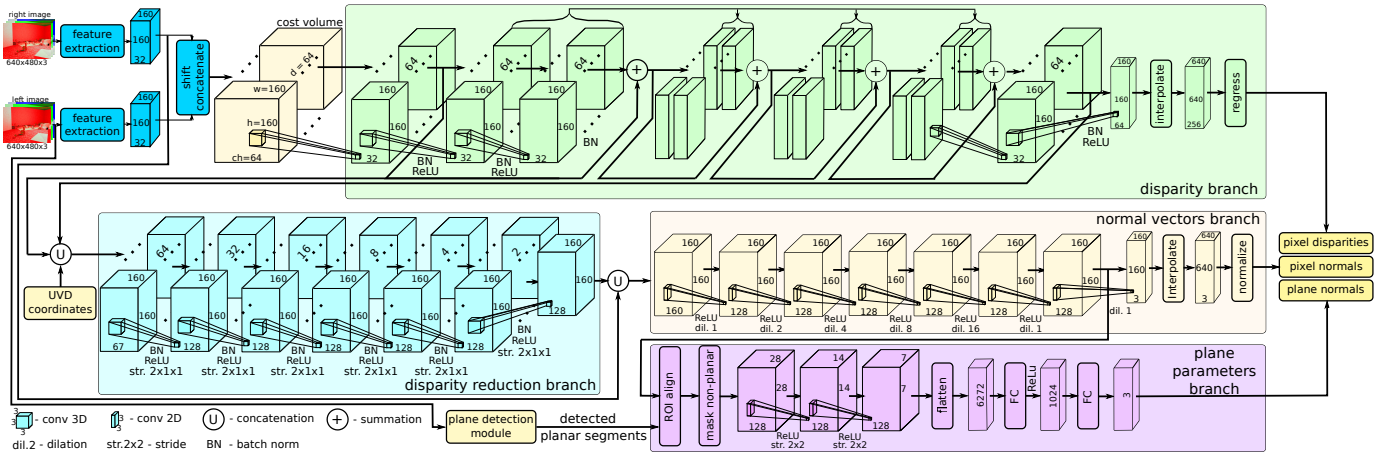

Fig. 2. Architecture of the geometry module in the Stereo Plane R-CNN. Each 3D convolution is 3×3×3 and 2D convolution is 3×3

The plane detection module is inspired by the Plane R-CNN [5] architecture that detects planar segments on a monocular image (the left one in our system). We improved the detection quality of planar segments by applying ROI-aware segmentation during training and by learning on a properly labeled dataset. The geometry module is inspired by the work of Kusupati *et al.* [13] that exploits stereo setup to infer about the geometry of the scene. This module builds a cost volume for a 3D space observed by the sensor and processes neural network embeddings to estimate pixel-wise disparities, normal vectors, and plane parameters for the detected segments. We jointly minimize losses related to all estimated values and use a camera-agnostic normal representation to improve geometry reconstruction performance.

A. Geometry module architecture

Although 3D coordinates of points computed from estimated disparity do not guarantee a precise fitting of a plane model, the geometry module (Fig. 2) utilizes features produced during disparity estimation to estimate normals and plane parameters. The module is based on a cost volume created in a proposed $\widetilde{\text{UVD}}$ space (explained in Sec. III-C), where features from the left and right image are concatenated for every point in that space. A disparity branch (green block in Fig. 2) is based on the Pyramid Stereo Matching Network [7] that uses 3D convolutions to process concatenated features and produce probability distributions of disparities for each pixel. Expected values are computed from those distributions to regress final disparity values. Features from the beginning and the end of the disparity estimation branch are concatenated with $\widetilde{\text{UVD}}$ coordinates and used in a disparity reduction branch (light blue). Using 3D convolutions with stride 2 in the disparity dimension, that halves this dimension's size after each operation, we reduce this dimension of the feature maps to 1. This step effectively removes the disparity dimension, leaving rich 2D normal features. The 2D normal features are concatenated with visual features from the left image and processed using 2D convolutions with various dilations to return three parameters of normals for every pixel. The visual features help to smooth the estimates by providing visual cues about the surface. Moreover, 2D features are also used in a plane parameters subbranch (purple in Fig. 2) that samples them using ROI align according to detected ROIs. Sampled features that do not belong to the segment but are inside the ROI are masked by zeroing their values and such feature map is processed using two convolutional and two fully connected layers to estimate a plane normal.

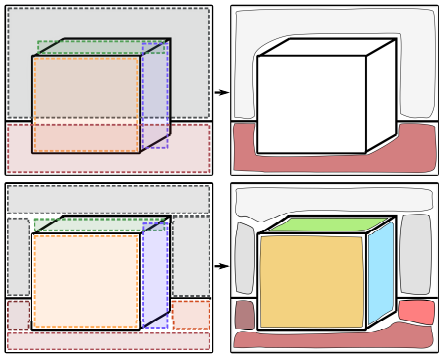

Fig. 3. ROI-aware segmentation: NMS removes some detections for complex scenes where bounding boxes overlap (top). Oversegmentation of the complex shapes (bottom) produces a larger number of smaller detections, but preserves planes that are crucial for scene reconstruction

B. ROI-aware detection and segmentation

We have observed that Plane R-CNN has problems with planar segments occupying a large area of the image, especially the ones also interleaving with other segments (illustrated in Fig. 3). We noticed that it was due to ROI boxes containing multiple segments. Boxes for different segments are overlapping with each other and got suppressed in the Non-Maximum Suppression (NMS) step. The same problem exists for prolonged segments and in general for all segments whose shape is far from square. Therefore, we propose to divide target segments during training into smaller ones with more square-like shapes. It is worth noting that segmentation into planar segments is often arbitrary and can be done in many correct ways. For example, the front of the cupboard can be segmented as one segment or as separate segments for each door. Both segmentations are correct regarding the plane-based localization or navigation [1]. We employed a simple algorithm we call ROI-aware segmentation based on flood fill, that can be executed on the fly when loading training

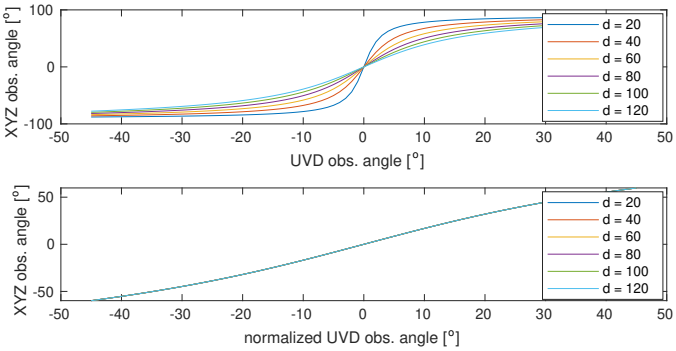

Fig. 4. Angle of observation of a normal in the XYZ space as a function of angle of observation in the UVD space (top) and $\widetilde{\text{UVD}}$ space (bottom) for different disparities, $f_x = f_y = 550$, $c_x = o_x = 320$, $c_y = o_y = 240$, baseline $b = 0.2$, and scale constant $a = 320$. Note values on vertical axes. For $\widetilde{\text{UVD}}$ all lines overlap.

samples. The algorithm starts at a random pixel and floods the segment as long as the ratio of the area of the grown region to the area of the smallest bounding box is above 0.5. If the ratio is below this threshold, the bounding box is much larger than the segment itself and is mostly empty. This implies that the segment's shape is deviating from being square-like and that the ROI of this segment can be overlapping with ROIs of the neighboring segments. Although it is a greedy algorithm that does not guarantee optimal segmentation, we found it works well in practice with an acceptable level of over-segmentation. Therefore, instead of a refining module proposed in [5], we use carefully segmented target masks during learning to obtain good quality detections when testing. Note that it is not necessary to use this mechanism during the inference because a neural network has already learned to produce proposals that are ROI-aware.

C. Scene geometry from stereo camera

A mapping between 3D coordinates of points or normal vectors and pixel coordinates relies heavily on the camera intrinsic parameters. If a black box model (e.g., neural network) is applied to the estimation of 3D coordinates from an image it has to capture this relation. Thus, we make the normal representation camera-agnostic to simplify this problem and to avoid unnecessary transformations that DNN would have to learn. If an input to the DNN is a pair of stereo images, data structures containing those images are organized according to image coordinates (u, v) and disparities (d). Hence, u, v, and d are known to the network for every processed point. What the network does not know, are XYZ coordinates of points because camera parameters are necessary to compute them. Therefore, instead of performing estimation in XYZ space associated with a camera frame and physical dimensions, we exploit disparity-normalized UVD ($\widetilde{\text{UVD}}$) space associated with pixel coordinates and disparity. This mitigates problems with deployment in real-life scenarios that stem from differences between available training datasets and target hardware. The transformation between XYZ space and $\widetilde{\text{UVD}}$ space is linear, so planes remain planes under this transformation. To derive this transformation, let us consider equations of a 3D world point (x, y, z) projection for a calibrated stereo camera with a baseline b:

$$\begin{cases} u = \dfrac{f_x x}{z} + c_x - o_x \\ v = \dfrac{f_y y}{z} + c_y - o_y \\ d = \dfrac{f_x b}{z}, \end{cases} \tag{1}$$

where $(x, y, z)^T$ is a 3D position of a point in a camera frame, f_x, f_y, c_x, c_y are intrinsic camera parameters, and o_x, o_y is an origin of the UVD coordinate frame, which can be chosen arbitrarily to move it from the upper left corner of the image. Transformation of point $\mathbf{p} = \begin{pmatrix} x, & y, & z, & 1 \end{pmatrix}^T$ from XYZ to a point $\mathbf{p}_D$ in UVD using homogeneous coordinates can be written as:

$$\mathbf{p}_D = \mathbf{G}_{D,C}\mathbf{p}, \tag{2}$$

where $\mathbf{G}_{D,C}$ is a matrix following (1). Therefore, plane parameters in UVD $\boldsymbol{\pi}_D = \begin{pmatrix} n_u, & n_v, & n_d, & -r_D \end{pmatrix}^T$ that satisfy $\boldsymbol{\pi}_D \cdot \mathbf{p}_D = 0$ can be transformed to XYZ using:

$$\boldsymbol{\pi} = \mathbf{G}_{D,C}^T \boldsymbol{\pi}_D = \mathbf{G}_{C,D}^{-T} \boldsymbol{\pi}_D, \tag{3}$$

derived using (2), where $\boldsymbol{\pi} = \begin{pmatrix} n_x, & n_y, & n_z, & -r \end{pmatrix}^T$. This transformation is also linear, however has an undesired property that for small disparities (distant objects), a relatively small angular error in normal estimation in UVD propagates as a large error in XYZ. To illustrate this, consider a plane observed at different horizontal angles (rotation around the Y-axis of the camera) in front of the camera. In the top part of Fig. 4 an angle of observation in the XYZ space was plotted as a function of an angle of observation in the UVD space for example camera parameters. It is visible that the smaller the disparity, the sharper the transition and thus the larger the derivative, which is a multiplicative factor in the error propagation. To overcome this problem, we normalize the coordinates in the UVD space with the disparity:

$$\begin{cases} \tilde{u} = \dfrac{u}{d} = \dfrac{f_x}{f_x b} x + \dfrac{c_x - o_x}{f_x b} z \\ \tilde{v} = \dfrac{v}{d} = \dfrac{f_y}{f_x b} y + \dfrac{c_y - o_y}{f_x b} z \\ \tilde{d} = \dfrac{a}{d} = \dfrac{a}{f_x b} z, \end{cases} \tag{4}$$

where a (320 in the experiments) is an arbitrary constant assuring uniform scaling of the space and forcing values in $\widetilde{\text{UVD}}$ to be of the same magnitude. By virtue of this normalization, and using values of o_x, o_y close to c_x, c_y (usually optical centers of cameras do not vary much and are close to image center), relation of observation angles in XYZ and $\widetilde{\text{UVD}}$ is approximately linear and does not depend on d (see bottom part of Fig. 4). Using homogeneous coordinates it can be written as:

$$\mathbf{p}_{\tilde{D}} = \begin{pmatrix} \tilde{u} \\ \tilde{v} \\ \tilde{d} \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{f_x}{f_x b} & 0 & \frac{c_x - o_x}{f_x b} & 0 \\ 0 & \frac{f_y}{f_x b} & \frac{c_y - o_y}{f_x b} & 0 \\ 0 & 0 & \frac{a}{f_x b} & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix} = \mathbf{G}_{\tilde{D},C}\mathbf{p}. \tag{5}$$

This space is camera-agnostic, meaning that to calculate a plane equation in this space it is sufficient to know image coordinates and disparities of points forming this plane, without knowledge about focal lengths f_x, f_y, optical center c_x, c_y, and baseline b. Moreover, the space is scaled similarly to the XYZ space and therefore angular errors are not significantly magnified. To transform plane $\boldsymbol{\pi}_{\tilde{D}} = \begin{pmatrix} n_{\tilde{u}} & n_{\tilde{v}} & n_{\tilde{d}} & -r_{\tilde{D}} \end{pmatrix}^T$ from $\widetilde{\mathrm{UVD}}$ space to XYZ space we use equation analogous to (3):

$$\boldsymbol{\pi} = \mathbf{G}_{C,\tilde{D}}^{-T} \boldsymbol{\pi}_{\tilde{D}}. \tag{6}$$

In the plane parameters branch of the DNN, we estimate only the normal vector of the segment. To estimate the distance to the origin r we use RANSAC and disparity estimates from the disparity branch. In the procedure, we seek the best set of inliers using RANSAC and a threshold on the relative distance $\frac{\mathbf{n}_h \cdot \mathrm{proj}(\mathbf{p})}{r_h} < 0.05$, where $\mathrm{proj}(\mathbf{p})$ is a 3D XYZ point expressed in inhomogeneous coordinates, $\mathbf{n}_h$ is a normal vector of the RANSAC hypothesis, and r_h is a distance to the origin of the RANSAC hypothesis. Finally, r is estimated using all inliers from the best hypothesis, leaving the DNN estimated normal unchanged (RANSAC estimated normal is ignored).

During detection, we use only two classes (planar and non-planar). We found that using anchors for normal directions and dividing planes into classes related to those directions, as in [5], does not improve normal estimation accuracy comparing to direct estimation of 3 normal parameters.

IV. Training

A. Dataset

To the best of our knowledge, there is no large real-world dataset with stereo images and ground truth depth information for the indoor environment. Moreover, the quality of depth information and created mesh models in existing monocular datasets, such as `ScanNet` [23], are insufficient. We examined the labeling of `ScanNet` used by Plane R-CNN [5], which turned out to be of poor quality due to noisy and inaccurate mesh models produced using an RGB-D sensor. Thus, we generated a synthetic dataset called `SceneNet Stereo` to train the neural network. To generate scenes and render photorealistic images, we exploit a method from the `SceneNet RGB-D` dataset [24] by adapting it to produce stereo images. As a result of having perfect information about the geometry of rendered scenes, the training set was very accurate, which is difficult to obtain on real-world images. We generated 200 random scenes with 300 images for each scene. Finally, we selected approximately 35k training images and 2k testing images.

B. Loss and parameters

We use weights pre-trained on the `Coco` dataset for the detection module and weights pre-trained on the `ScanNet` [13] dataset for feature extraction layers and disparity branch of the geometry module. We train the whole model simultaneously, using different loss functions for specific branches. We use a loss from [5], without plane parameters component, to supervise the detection module and denote it as $\mathcal{L}_r$. The disparity estimation is supervised using L_1 smooth loss for all pixels $\mathcal{P}$ that have a valid target depth:

TABLE I
STATISTICS ON DETECTED PLANES FOR TESTING DATASETS

	bin no.	1	2	3	4	5	6
	A [px]	0-50	50-100	100-150	150-200	200-250	250-640
`SceneNet Stereo`	no. of planes	19258	10283	3025	1612	1329	2450
	area [%]	3.9	7.8	7.1	7.6	10.5	39.7
`TERRINet`	no. of planes	138739	70501	16052	4670	3608	4286
	area [%]	6.4	12.2	8.3	5.0	6.6	13.8

$$\mathcal{L}_d = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} f_1(d_p, d_p^*), \tag{7}$$

where f_1 is a smooth L_1 difference function, d_p is a disparity for pixel p, and d_p^* is a target disparity for pixel p. Because we are interested only in pixels belonging to planar segments, we exclude pixels near edges of objects during computation of pixel-wise normal vector loss $\mathcal{L}_n$:

$$\mathcal{L}_n = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} f_1(\mathbf{n}_p, \mathbf{n}_p^*), \tag{8}$$

where $\mathbf{n}_p$ and $\mathbf{n}_p^*$ are an estimated and a reference normal vector, respectively. The plane parameters loss is computed for all detections $\mathcal{D}$ returned by the detection module:

$$\mathcal{L}_p = \frac{1}{|\mathcal{D}|} \sum_{d \in \mathcal{D}} f_1(\mathbf{n}_d, \mathbf{n}_d^*). \tag{9}$$

The final loss is a sum of weighted losses $\mathcal{L}_r$ and (7)–(9):

$$\mathcal{L} = \mathcal{L}_r + w_d \mathcal{L}_d + w_n \mathcal{L}_n + w_p \mathcal{L}_p, \tag{10}$$

where $w_d = 1$, $w_n = 100$, and $w_p = 100$ to accommodate for different scales of values. For a fair comparison, we use the same weights during training of the baseline Plane R-CNN model (note that it is different from the original setup because of the different value of w_p). The training takes 10 epochs using Adam optimizer with a learning rate equal to 10^{-5} and weight decay equal to 10^{-4}. We augment training examples using random color and sharpness manipulation, Gaussian noise, and random cropping. For the baseline solution (Plane R-CNN), we skip augmentation as it worsens results.

V. Experimental verification

We use three metrics that measure geometric aspects of segmentation that are important during localization [1]:

- Detection Error (DE) - measures how planar is the area labeled as one segment. It is computed as RMS of point-to-plane distance in 3D between points belonging to the segment and plane fit into those points using RANSAC. It also measures the quality of segmentation, like metrics evaluating the similarity of pixels clustering, while avoiding problems with the ambiguity of correct segmentation.
- Depth Reconstruction Error (DRE) - measures RMS of depth differences between 3D points belonging to the segment and depthmap induced by a plane estimated by the DNN. We use only points classified as inliers by RANSAC to accommodate for imperfect segmentation

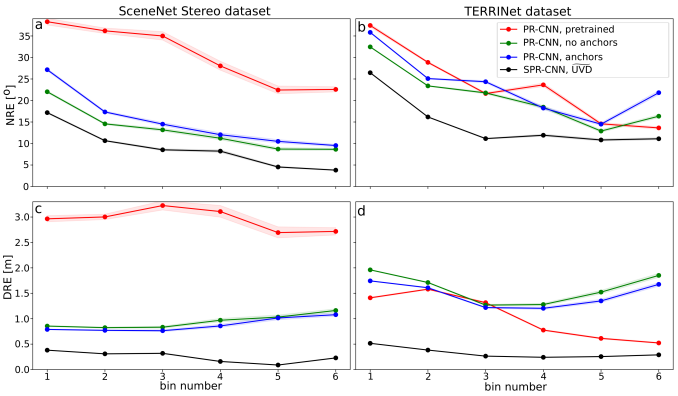

Fig. 5. Dependence between the NRE (a,b) and DRE (c,d) for the `SceneNet Stereo` (a,c) and `TERRINet` (b,d) datasets. Confidence intervals are marked with light-colored regions. Relations between the bin numbers and planar segments sizes are presented in Tab. I

and incorrect ground truth depth at the edges. We also crop plane induced depth to 15 m as we do not consider objects that are further away.

- Normal Reconstruction Error (NRE) - measures RMS of differences between normals found using RANSAC and ones estimated by the DNN.

In all metrics, we use ground truth depth information to precisely estimate plane equations. Robust estimation mitigates situations when segment masks spill over edges of surfaces or depth is incorrect at the edges.

Two different datasets were used during the evaluation to show various aspects of planar segments detection and geometry reconstruction. The first one is a testing part of the synthetic `SceneNet Stereo` dataset with nine different scenes and approximately 2k images. Note that the number of planes used to evaluate the accuracy of the methods is significantly higher. The second one is a real-world `TERRINet` dataset gathered in office and laboratory environments[2]. It contains several indoor sequences of stereo images, Velodyne VLP-16 lidar scans, and ground truth poses from a Qualisys motion capture system. By using ground truth poses and lidar measurements we built a precise point cloud representation of the scenes. Then, the point cloud was used to compute a depth map for every stereo image pair. We used approximately 8.5k images from 3 different environment settings.

To give more insight into the geometry reconstruction of various planar segments, we present results as a function of the segment's area expressed in pixels. We divide segments into six bins, depending on the square root of their area, denoted as A, which can be intuitively compared to the area of a square with a side length equal to A. Statistics regarding bins for used datasets are presented in Tab. I, where $\overline{\text{area}}$ denotes mean area per image.

A. Geometry reconstruction using stereo

The main experiment shows that our system, trained on a photorealistic synthetic dataset, can be used in a real-world scenario and has a superior performance over the baseline Plane R-CNN solution in terms of geometry reconstruction. The Plane R-CNN yields the best results in the literature with other systems only presenting functionalities added on the top of the Plane R-CNN [4], [22] or being closed-source [3]. Those functionalities can be also added on top of Stereo Plane R-CNN if necessary, but would obfuscate the results. The methods used for comparison are shortly characterized below:

- *Plane R-CNN (PR-CNN), pre-trained* - baseline version presented in [5] with anchors for normal estimation and refinement module, trained by the authors on the `ScanNet` dataset, using left image only.
- *PR-CNN, no anchors* - baseline version without anchors and the refinement module, trained on the `SceneNet Stereo` dataset, using left image only.
- *PR-CNN, anchors* - baseline version with anchors but without the refinement module, trained on the `SceneNet Stereo` dataset, using left image only.
- *RANSAC, SGBM depth* - the method that uses non-learned Semi-Global Block Matching stereo depth estimation and performs classic plane fitting using RANSAC.
- *RANSAC, DNN depth* - the method that uses learned stereo depth estimation architecture from [7] trained on our dataset and performs RANSAC plane fitting.
- *Stereo Plane R-CNN (SPR-CNN), $\widetilde{UVD}$* - the proposed method described in Sec. III.

The results of the experiment are presented in Tab. II. To eliminate the influence of segment detection on the results, we used the same detections for all tests. Detections were generated and saved by one version of the system and are loaded in all test cases, except the pre-trained Plane R-CNN due to the presence of the refinement module. Table II presents qualitative results, while detailed results on `TERRINet` dataset are visible in Fig. 5, where performance as a function of the square root of the area A is presented. Both learned stereo versions perform significantly better in terms of depth errors, which suggests that it is crucial to precise geometry reconstruction. However, the classic approach to stereo depth estimation does not provide enough valid points to precisely fit a plane. As for normal errors, Stereo Plane R-CNN outperforms other systems by a large margin. Results also indicate that using anchors does not improve normal estimation accuracy which is why we do not use this technique. Please also note that the pre-trained version of Plane R-CNN detects less distant segments (mean distance to points is 3.39 m, compared to 4.21 m for detections used for other versions), which explains better results of the pre-trained version compared to the version trained by us when testing on the `TERRINet` dataset. Figure 7 shows visualizations of example scenes for the baseline Plane R-CNN without ROI-aware segmentation and Stereo Plane R-CNN.

B. Detection using ROI-aware segmentation

This experiment aims at showing the effects of the proposed adaptation to the ROI-based processing. We use the DE score, which is designed to measure the quality of detection, to show the differences between methods. We employ the synthetic dataset only because it has precise depth information for all

[2]Dataset collection was supported by the TERRINet project funded by EU H2020 under GA No.730994

TABLE II
GEOMETRIC ACCURACY (DRE AND NRE) FOR BOTH DATASETS

	SceneNet Stereo		TERRINet	
	DRE [m]	NRE [°]	DRE [m]	NRE [°]
PR-CNN, pretrained [5]	2.82	25.64	1.00	20.75
PR-CNN, no anchors	1.05	11.13	1.66	21.34
PR-CNN, anchors	0.98	12.81	1.52	24.12
RANSAC, SGBM depth	0.44	11.28	2.17	30.84
RANSAC, DNN depth	**0.22**	10.13	0.38	22.88
SPR-CNN, $\widetilde{\text{UVD}}$ (full arch.)	0.24	**7.09**	**0.34**	**15.07**

TABLE III
DETECTION ERRORS FOR THE SceneNet Stereo DATASET

bin no.		1	2	3	4	5	6	all
SPR-CNN	DE [m]	0.180	0.206	0.185	0.187	0.152	**0.118**	0.147
w/o ROI-aware	$\overline{\text{area}}$ [%]	1.3	5.5	6.3	8.0	11.5	43.8	76.5
SPR-CNN	DE [m]	**0.148**	**0.127**	**0.123**	**0.166**	**0.102**	0.136	**0.134**
w. ROI-aware	$\overline{\text{area}}$ [%]	4.1	9.4	8.5	8.3	10.6	39.0	79.8

pixels. Quantitative results are presented in Tab. III, where despite a larger area of detected segments, DNN trained with ROI-aware segmentation performs significantly better. However, the most notable differences can be seen in Fig. 7, where visual comparison is presented.

C. Ablation study

To justify our design choices we conduct an ablation study comparing different versions of Stereo Plane R-CNN:

- *SPR-CNN normal vec. only* - version without plane parameters branch, plane normals are estimated by averaging pixel-wise values from the normal vector branch, using $\widetilde{\text{UVD}}$.
- *SPR-CNN, plane param. only* - version without the supervision of pixel-wise normals in the normal vector branch, using $\widetilde{\text{UVD}}$.
- *SPR-CNN, XYZ (full arch.)* - estimates normals in the XYZ space, instead of the $\widetilde{\text{UVD}}$ space.
- *SPR-CNN, UVD (full arch.)* - estimates normals in the UVD space, instead of the $\widetilde{\text{UVD}}$ space.
- *SPR-CNN, $\widetilde{\text{UVD}}$ (full arch.)* - proposed method.

Results are presented in Tab. IV and suggest that having a specialized branch for plane parameters estimation boosts performance significantly. However, supervision of normals at the level of pixels and $\widetilde{\text{UVD}}$ representation also contribute to the final result notably. Additionally, it is clearly visible that regular UVD space (without normalization) is not suitable for normal estimation as it yields the worst results as far as NRE is concerned.

D. Robustness to camera parameters change

The goal of the last experiment is to show that the proposed camera-agnostic representation performs well when camera parameters change. Because there is no real-world dataset that contains images from different cameras with different parameters, we use the synthetic SceneNet Stereo dataset in this experiment. Test sequences were rendered once more with fixed lighting and with different camera parameters. We were changing diagonal field of view (FoV, change of, both, f_x and f_y), vertical FoV (change of f_y), horizontal FoV (change of f_x), and baseline (b). We do not consider different c_x and c_y values, without a loss of generality, because their change only shifts image left/right and up/down. In this experiment, we compare monocular Plane R-CNN that estimates normals in XYZ, Stereo Plane R-CNN that estimates normals in XYZ, and the proposed method that estimates normals in $\widetilde{\text{UVD}}$. The results are summarized in Fig. 6. The increase of NRE for the version exploiting camera-agnostic representation is lower than for the version using XYZ representation, which supports the thesis that such a representation is beneficial to assure robustness to camera parameters change. However, despite using camera-agnostic representation, the whole model is not completely camera-agnostic because of changing incidence relations when f_x, f_y, or b change. The model seems to be more sensitive to f_x change (change of diagonal and horizontal FoV) than to f_y and b change (change of vertical FoV and baseline). It is worth noting that changing the diagonal FoV in the monocular system slightly lowers the normal estimation error. This phenomenon comes from the fact that changing, both, f_x and f_y only scales the image. Moreover, with the wider camera field of view, a broader context is captured and the normal estimation error decreases. Nonetheless, results for the monocular system are still worse than for the stereo ones. In terms of DRE, the performance of both stereo versions slightly deteriorates, which can be again attributed to changing incidence relations. However, changing camera parameters introduces the scale change and increases significantly the DRE value for the monocular version when the diagonal FoV changes.

TABLE IV
ABLATION STUDY OF DIFFERENT VERSIONS OF STEREO PLANE R-CNN

	SceneNet Stereo		TERRINet	
	DRE [m]	NRE [°]	DRE [m]	NRE [°]
SPR-CNN, normal vec. only	0.29	8.75	0.38	18.77
SPR-CNN, plane param. only	0.28	8.28	0.37	16.09
SPR-CNN, XYZ (full arch.)	0.36	7.21	0.71	16.07
SPR-CNN, UVD (full arch.)	0.25	10.41	0.48	21.30
SPR-CNN, $\widetilde{\text{UVD}}$ (full arch.)	**0.24**	**7.09**	**0.34**	**15.07**

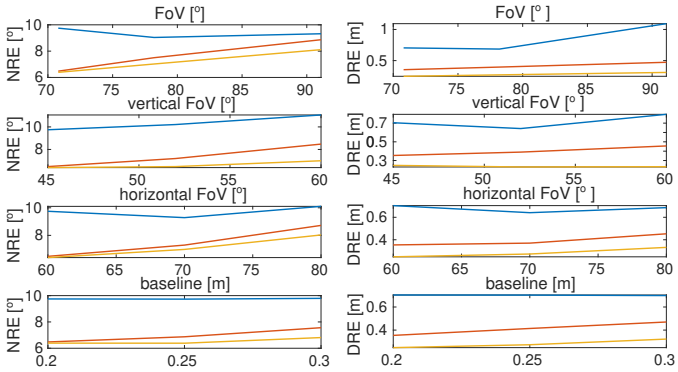

Fig. 6. Dependence between NRE (left) DRE (right) and camera parameters change for the PR-CNN (blue), SPR-CNN XYZ (red), and SPR-CNN, $\widetilde{\text{UVD}}$ (orange). The most left values on horizontal axes are used during learning.

VI. CONCLUSIONS

In this article, we propose the Stereo Plane R-CNN method that detects and computes the parameters of planar segments from stereo pairs of images. The system is trained on the

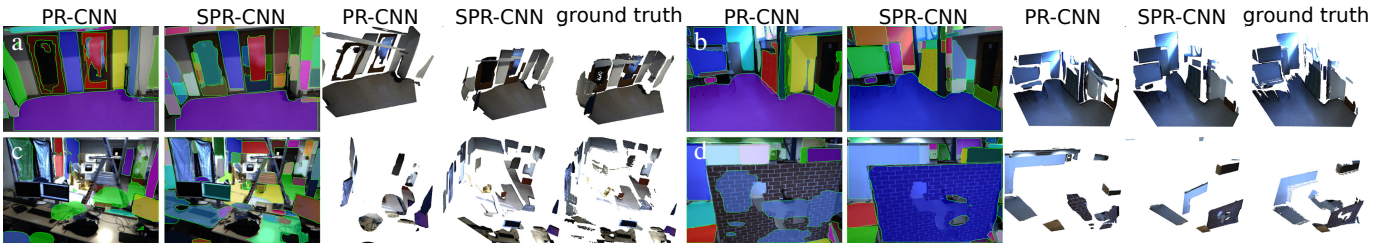

Fig. 7. Comparison of segment detection and scene geometry reconstruction performance on the four scenes (a,b,c,d) from the TERRINet dataset. Note scale problems of the monocular version for the first example (a).

synthetic dataset that provides accurate information about the depth of the scene, segmentation, and plane parameters. The system is verified on the dataset obtained in the indoor unstructured and challenging environment. The proposed method is compared to other the state-of-the-art methods. Moreover, we provide ablation studies on the contributions of main components in our system to justify our design choices and to show the performance of the proposed method. The results presented in Tab. II show that the proposed problem representation and neural network architecture outperform other approaches. The mean inference time for our solution is 0.419 s on RTX 3090 with batch size 1, which is approximately twice as much as Plane R-CNN and is caused by a larger amount of computations needed to process the cost volume. However, the obtained computation time is sufficient for the global localization task [1].

In particular, we show that improved image segmentation deals with the suppression problems of methods based on ROIs and NMS (results in Tab. III). We also propose a novel neural network architecture that leverages disparity information from a stereo camera to precisely reconstruct scene geometry (Fig. 7). The neural network uses the proposed camera-agnostic normal representation $\widetilde{\text{UVD}}$ that improves robustness to camera parameters change (Fig. 6). Finally, we propose a training procedure that simultaneously utilizes parameters of planes, pixel-wise normal vectors, and disparity prediction to improve the accuracy of reconstruction (results in Tab. II and Fig. 5). In the future, we are going to integrate the Stereo Plane R-CNN with our global relocalization system [1] to improve localization in the indoor environment.

REFERENCES

[1] J. Wietrzykowski and P. Skrzypczyński, "PlaneLoc: Probabilistic global localization in 3-D using local planar features," *Robotics and Autonomous Systems*, vol. 113, pp. 160–173, 2019.

[2] W. Ye, H. Li, T. Zhang, X. Zhou, H. Bao, and G. Zhang, "SuperPlane: 3D plane detection and description from a single image," in *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, 2021, pp. 207–215.

[3] W. Xi and X. Chen, "Reconstructing piecewise planar scenes with multi-view regularization," *Computational Visual Media*, vol. 5, p. 337–345, 2019.

[4] Y. Qian and Y. Furukawa, "Learning pairwise inter-plane relations for piecewise planar reconstruction," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 330–345.

[6] N. Smolyanskiy, A. Kamenev, and S. Birchfield, "On the importance of stereo for accurate depth estimation: An efficient semi-supervised deep neural network approach," in *CVPR 2018 Workshop on Autonomous Driving*, Salt Lake City, 2018.

[5] C. Liu, K. Kim, J. Gu, Y. Furukawa, and J. Kautz, "PlaneRCNN: 3D plane detection and reconstruction from a single image," in *IEEE Conf. on Comp. Vis. and Pattern Recog. (CVPR)*, 2019, pp. 4445–4454.

[7] J.-R. Chang and Y.-S. Chen, "Pyramid Stereo Matching Network," in *IEEE Conf. on Comp. Vis. and Pattern Recog. (CVPR)*, 2018, pp. 5410–5418.

[8] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *IEEE Int. Conf. on Computer Vision (ICCV)*, 2017, pp. 2980–2988.

[9] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep Ordinal Regression Network for Monocular Depth Estimation," in *IEEE Conf. on Comp. Vis. and Pattern Recog. (CVPR)*, 2018.

[10] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," *ArXiv preprint*, 2021.

[11] X. Wang, D. F. Fouhey, and A. Gupta, "Designing deep networks for surface normal estimation," in *IEEE Conf. on Comp. Vis. and Pattern Recog. (CVPR)*, 2015, pp. 539–547.

[12] X. Qi, R. Liao, Z. Liu, R. Urtasun, and J. Jia, "Geonet: Geometric neural network for joint depth and surface normal estimation," in *IEEE Conf. on Comp. Vis. and Pattern Recog. (CVPR)*, 2018, pp. 283–291.

[13] U. Kusupati, S. Cheng, R. Chen, and H. Su, "Normal assisted stereo depth estimation," in *IEEE Conf. on Comp. Vis. and Pattern Recog. (CVPR)*, 2020, pp. 2186–2196.

[14] A. R. Zamir, A. Sax, N. Cheerla, R. Suri, Z. Cao, J. Malik, and L. J. Guibas, "Robust learning through cross-task consistency," in *IEEE Conf. on Comp. Vis. and Pattern Recog. (CVPR)*, 2020, pp. 11 194–11 203.

[15] J. Wietrzykowski and P. Skrzypczyński, "A probabilistic framework for global localization with segmented planes," in *2017 European Conference on Mobile Robots (ECMR)*, 2017, pp. 1–6.

[16] F. Yang and Z. Zhou, "Recovering 3D Planes from a Single Image via Convolutional Neural Networks," in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018, pp. 87–103.

[17] C. Liu, J. Yang, D. Ceylan, E. Yumer, and Y. Furukawa, "PlaneNet: Piece-wise planar reconstruction from a single RGB image," in *IEEE Conf. on Comp. Vis. and Pattern Recog. (CVPR)*, 2018, pp. 2579–2588.

[18] Z. Yu, J. Zheng, D. Lian, Z. Zhou, and S. Gao, "Single-image piece-wise planar 3D reconstruction via associative embedding," in *IEEE Conf. on Comp. Vis. and Pattern Recog. (CVPR)*, 2019, pp. 1029–1037.

[19] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *IEEE Conf. on Comp. Vis. and Pattern Recog. (CVPR)*, 2017, pp. 936–944.

[20] M.-G. Park and K.-J. Yoon, "As-planar-as-possible depth map estimation," *Comp. Vision and Image Underst.*, vol. 181, pp. 50–59, 2019.

[21] M. Eder, P. Moulon, and L. Guan, "Pano popups: Indoor 3D reconstruction with a plane-aware network," in *2019 International Conference on 3D Vision (3DV)*, 2019, pp. 76–84.

[22] Z. Jiang, B. Liu, S. Schulter, Z. Wang, and M. Chandraker, "Peek-a-boo: Occlusion reasoning in indoor scenes with plane representations," in *IEEE Conf. on Comp. Vis. and Pattern Recog. (CVPR)*, 2020, pp. 110–118.

[23] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3D reconstructions of indoor scenes," in *IEEE Conf. on Comp. Vis. and Pattern Recog. (CVPR)*, 2017.

[24] J. McCormac, A. Handa, S. Leutenegger, and A. J. Davison, "SceneNet RGB-D: Can 5M synthetic images beat generic ImageNet pre-training on indoor segmentation?" in *IEEE Int. Conf. on Computer Vision (ICCV)*, 2017, pp. 2697–2706.