Signal, Image and Video Processing
https://doi.org/10.1007/s11760-021-01996-1

ORIGINAL PAPER


# Improving outdoor plane estimation without manual supervision

Furkan Eren Uzyıldırım[1] · Mustafa Özuysal[1]

Received: 3 May 2021 / Revised: 9 July 2021 / Accepted: 25 July 2021


**Abstract**
Recently, great progress has been made in the automatic detection and segmentation of planar regions from monocular images of indoor scenes. This has been achieved thanks to the development of convolutional neural network architectures for the task and the availability of large amounts of training data usually obtained with the help of active depth sensors. Unfortunately, it is much harder to obtain large image sets outdoors partly due to limited range of active sensors. Therefore, there is a need to develop techniques that transfer features learned from the indoor dataset to segmentation of outdoor images. We propose such an approach that does not require manual annotations on the outdoor datasets. Instead, we exploit a network trained on indoor images and an automatically reconstructed point cloud to estimate the training ground truth on the outdoor images in an energy minimization framework. We show that the resulting ground truth estimate is good enough to improve the network weights. Moreover, the process can be repeated multiple times to further improve plane detection and segmentation accuracy on monocular images of outdoor scenes.

**Keywords** Deep learning · Outdoor plane estimation · Transfer learning · Weakly supervised learning

## 1 Introduction

Most scenes include planar regions that provide information about the geometric structure of the scene and their automatic segmentation is a long standing goal of Computer Vision. Early methods for this task required images of multiple views ( [7,8,16]). Generally, they first reconstruct 3D point clouds from the images and generate plane candidates with 3D segmentation methods such as robust plane fitting via RANSAC [5]. They do not always give an accurate segmentation since the plane boundary in the 3D point cloud carries high uncertainty. Moreover, they require textured surfaces so that matching can successfully perform the reconstruction.

In recent years, Convolutional Neural Network (CNN)-based approaches ( [11,12,17,18]) allowed piece-wise planar segmentation without explicit reconstruction of a 3D point cloud. These approaches outperform traditional methods in terms of indoor reconstruction accuracy. However, most of them do not perform well for outdoor scenes.

✉ Mustafa Özuysal
mustafaozuysal@iyte.edu.tr

Furkan Eren Uzyıldırım
furkanuzyildirim@iyte.edu.tr

[1] Department of Computer Engineering, İzmir Institute of Technology, İzmir, Turkey

Indoor scenes have the advantage of easy depth sensing with the aid of active sensors. Since such sensors have limited operating range, manual annotation of images of outdoor scenes is a time-consuming task. Lack of large training sets can be dealt with transfer learning in which features learned from the indoor datasets are transferred to the outdoor domain. However, this too requires manually annotating outdoor images one by one to provide the best possible transfer performance. Therefore, it is desirable to transfer features from the existing networks to a suitably collected set of outdoor images without requiring manual annotation.

In this paper, we propose such an approach that requires a training set of outdoor images that can be processed by a structure from motion system to produce a dense 3D point cloud. Instead of directly extracting plane segmentation maps, we exploit the point cloud as a weak supervision signal much like approaches that exploit a small amount of user input [6] to improve segmentation quality. Unlike these, the point cloud reconstruction is fully automated. Our approach is based on the premise that combining such a point cloud and the output from a state-of-the-art plane segmentation network trained on indoor images such as PlaneRCNN [11], we can obtain an approximate but high-quality estimate of the ground truth annotations on the outdoor images. This estimate can then be used as training data to improve the network

Published online: 07 August 2021

weights, achieving transfer without manual labor. Moreover, we show that this process can be repeated multiple times to further improve plane segmentation accuracy. Once the training and the feature transfer is completed, during test time, our approach can detect and segment planar regions on a given monocular outdoor image with a much greater accuracy than a network trained on indoor images.

Our main contributions can be summarized as follows:

- We combine the traditional and recent approaches for the task of piece-wise plane reconstruction by achieving feature transfer under the guidance of 3D dense point cloud at training time with the initialization provided by a network trained on indoor images.
- We formulate an approximate and iterative transfer scheme that alternates between estimating ground truth labels and improving network weights in the target domain.
- We demonstrate that PlaneRCNN can be adapted to outdoor scenes for the task of piece-wise planar reconstruction without requiring manual annotations.

In the following, we briefly overview state-of-the-art in plane estimation and segmentation. We formalize our approach in Sect. 3 and provide experimental results on both structure-from-motion and simultaneous localization and mapping data sets. We provide an ablation study to highlight the contribution of different energy terms to the improved segmentation quality and briefly summarize our results and current limitations of the proposed approach.

## 2 Related work

Traditional piece-wise planar reconstruction methods ( [7,8,16]) for outdoor scenes require images of multiple views. [7] reconstructs 3D-oriented points with the aid of a multiview stereo approach and then generates plane candidates with heuristics and with Markov Random Fields (MRF) optimization. [8] generates plane hypotheses with RANSAC from a set of depth maps. These are refined with the MRF framework to obtain the final result. [16] generates a 3D sparse point cloud with a Structure-from-Motion (SfM) approach and extracts 3D line segments which are used in a graph cut formulation. [3] proposes a piece-wise planar model and obtains a 3D sparse point cloud from a set of images and solves the reconstruction problem with graph cuts by assigning plane labels to superpixels under the guidance of the point cloud. However, they do not use machine learning to regularize the estimation in textureless regions and initialization with planes fitted to superpixels is not robust.

Recently, deep neural architectures are trained for piece-wise planar reconstruction from a single image. PlaneNet ([12]) is designed for and trained with the images of indoor scenes. PlaneRecover( [17]) presents an unsupervised learning approach instead of having manual annotations for 3D plane parameters. It is trained with a synthetic outdoor dataset. Both PlaneNet and PlaneRecover limit the maximum possible number of planar regions which degrades applicability in general scenarios. To overcome this, [18] proposes an approach based on associative embedding. Recently, PlaneRCNN ( [11]) improves upon PlaneNet. Although, PlaneRCNN is designed for and trained with the images of indoor scenes like PlaneNet, piece-wise planar segmentation accuracy for outdoor images is slightly better. However, as we will demonstrate its performance is still constrained by the features learned on indoor images.

We combine ideas from both traditional approaches and neural network architectures. An automatically reconstructured point cloud is exploited only during training to estimate ground truth segmentations on the outdoor datasets. Once training is completed, we do not require multiple images and our approach reconstructs planes even in less textured areas. Moreover, since it adapts to the image features on outdoor datasets, its plane segmentation performance surpasses the existing networks that were trained on indoor data.

Finally, [19] showed that integrating geometric cues such as vanishing points and lines to constrain the plane segmentation results improves both segmentation quality and estimated plane parameters. Our approach relies on similar reasoning to exploit geometry of the scene to ease domain transfer from indoor to outdoor imagery. It might be possible to fuse our approach with that of [19], either to improve the output of our method using perspective cues or to provide stronger training for the approach of [19] by integrating larger amounts of unlabelled outdoor training data.

## 3 Transfer learning without manual supervision

The conventional approach to train a neural network for plane estimation on outdoor images requires collecting a set of training outdoor images $\mathtt{D}^{\mathtt{out}}$ with annotated ground truth training targets $\mathtt{T}^{\mathtt{out}}$. For plane estimation, the training targets are given by a tuple of segmentation masks $\mathtt{S}^{\mathtt{out}}$ and plane equations $\mathtt{5}^{\mathtt{out}}$, $\mathtt{T}^{\mathtt{out}} = \left(\mathtt{S}^{\mathtt{out}}, \mathtt{5}^{\mathtt{out}}\right)$. Then the neural network can be trained by searching for a set of weights $\mathbf{w}$ that minimizes a loss function $\mathscr{L}(\mathbf{w})$ tying together the network output to the ground truth values as

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \mathscr{L}\left(f(\mathtt{D}^{\mathtt{out}} \mid \mathbf{w}),\ \mathtt{T}^{\mathtt{out}}\right),$$

where $\mathscr{L}(\mathbf{w})$ measures the discrepancy between $\mathtt{T}^{\mathtt{out}}$ and the network output $f(\mathtt{D}^{\mathtt{out}} \mid \mathbf{w})$ on the training images.

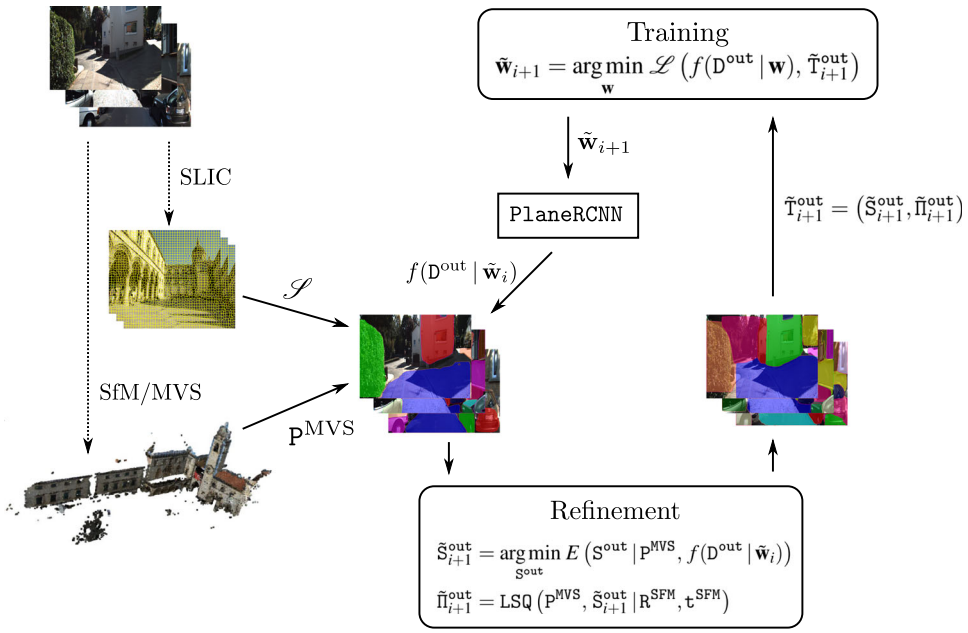



**Fig. 1** Proposed iterative transfer learning approach. We preprocess the outdoor images to extract a point cloud $\mathtt{P}^{\mathtt{MVS}}$ and a set of SLIC superpixels $\mathscr{S}$. Using these data, we initialize segmentation maps based on the current network output with weights $\tilde{\mathbf{w}}_i$. An energy-based minimization problem is solved to refine this crude initialization into an estimated set of training targets $\tilde{\mathtt{T}}^{\mathtt{out}}_{i+1}$. The network is then trained to minimize a loss function on these training targets, yielding improved network weights $\tilde{\mathbf{w}}_{i+1}$. These new weights provide a better initialization, so we can repeat the process multiple times. Note that the whole process is automated and the point cloud is only used in the training phase

We assume that the ground truth training targets $\mathtt{T}^{\mathtt{out}}$ are not available, but we have access to network weights $\mathbf{w}_{\mathtt{PlaneRCNN}}$ trained on indoor images. Moreover, we assume that it is possible to obtain a point cloud $\mathtt{P}^{\mathtt{MVS}}$ by using a state-of-the-art Structure from Motion (SfM) and Multiview Stereo (MVS) system such as COLMAP ( [14,15]) on the collected outdoor image set.

Our approach is based on the idea that the above minimization problem can be approximately solved by first estimating the training targets based on $\mathbf{w}_{\mathtt{PlaneRCNN}}$ and $\mathtt{P}^{\mathtt{MVS}}$, and then improving the weights $\mathbf{w}$ by using this estimated $\tilde{\mathtt{T}}^{\mathtt{out}}$ which provides weak-supervision:

$$\begin{aligned}\tilde{\mathtt{T}}^{\mathtt{out}} &= \arg\min_{\mathtt{T}^{\mathtt{out}}} E\left(\mathtt{T}^{\mathtt{out}} \mid \mathtt{P}^{\mathtt{MVS}}, f(\mathtt{D}^{\mathtt{out}} \mid \mathbf{w}_{\mathtt{PlaneRCNN}})\right)\\ \tilde{\mathbf{w}}^{*} &= \arg\min_{\mathbf{w}} \mathscr{L}\left(f(\mathtt{D}^{\mathtt{out}} \mid \mathbf{w}), \tilde{\mathtt{T}}^{\mathtt{out}}\right),\end{aligned}$$

where $E\left(\mathtt{T}^{\mathtt{out}}\right)$ measures the mismatch between the estimated training targets and the information provided by the 3D point cloud $\mathtt{P}^{\mathtt{MVS}}$ and the network output using pretrained weights $f(\mathtt{D}^{\mathtt{out}} \mid \mathbf{w}_{\mathtt{PlaneRCNN}})$. It also ensures that the segmentation masks are smooth. The exact formulation of the energy function is given in more detail in Sect. 3.2.

By solving the two minimization problems above, we expect to get better weights and consequently a better set of segmentation masks and plane parameters on outdoor images than what is initially possible with $\mathbf{w}_{\mathtt{PlaneRCNN}}$. As a result, we may attempt to re-estimate the training targets, taking advantage of the newly estimated weights. This leads to the following iterative approach that alternates between the estimation of training targets and the optimization of the network weights based on the last estimate:

$$\begin{aligned}\tilde{\mathbf{w}}_0 &= \mathbf{w}_{\mathtt{PlaneRCNN}}\\ \tilde{\mathtt{T}}^{\mathtt{out}}_{i+1} &= \arg\min_{\mathtt{T}^{\mathtt{out}}} E\left(\mathtt{T}^{\mathtt{out}} \mid \mathtt{P}^{\mathtt{MVS}}, f(\mathtt{D}^{\mathtt{out}} \mid \tilde{\mathbf{w}}_i)\right)\\ \tilde{\mathbf{w}}_{i+1} &= \arg\min_{\mathbf{w}} \mathscr{L}\left(f(\mathtt{D}^{\mathtt{out}} \mid \mathbf{w}), \tilde{\mathtt{T}}^{\mathtt{out}}_{i+1}\right),\end{aligned}$$

Finally, it is easy to estimate the plane parameters based on the information provided by the 3D point cloud and the segmentation masks. We just assign each 3D point to one of the planes based on its projection into the segmentation masks with the camera rotation matrices and translations estimated during the SfM stage. Once a set of 3D points are determined for each plane, we employ robust least square fitting to estimate a better set of plane equations. Therefore, the energy minimization stage in the iterative scheme can be replaced by a minimization involving only the segmentation masks and subsequent geometric fitting for the plane equations as follows:

$$\begin{aligned}\tilde{\mathbf{w}}_0 &= \mathbf{w}_{\mathtt{PlaneRCNN}}\\ \tilde{\mathtt{S}}^{\mathtt{out}}_{i+1} &= \arg\min_{\mathtt{S}^{\mathtt{out}}} E\left(\mathtt{S}^{\mathtt{out}} \mid \mathtt{P}^{\mathtt{MVS}}, f(\mathtt{D}^{\mathtt{out}} \mid \tilde{\mathbf{w}}_i)\right)\\ \tilde{\mathtt{S}}^{\mathtt{out}}_{i+1} &= \mathtt{LSQ}\left(\mathtt{P}^{\mathtt{MVS}}, \tilde{\mathtt{S}}^{\mathtt{out}}_{i+1} \mid \mathtt{R}^{\mathtt{SFM}}, \mathtt{t}^{\mathtt{SFM}}\right)\\ \tilde{\mathtt{T}}^{\mathtt{out}}_{i+1} &= \left(\tilde{\mathtt{S}}^{\mathtt{out}}_{i+1}, \tilde{\mathtt{S}}^{\mathtt{out}}_{i+1}\right)\\ \tilde{\mathbf{w}}_{i+1} &= \arg\min_{\mathbf{w}} \mathscr{L}\left(f(\mathtt{D}^{\mathtt{out}} \mid \mathbf{w}), \tilde{\mathtt{T}}^{\mathtt{out}}_{i+1}\right),\end{aligned}$$

As illustrated by Fig. 1, we propose to iteratively train a network for outdoor plane estimation and segmentation without requiring annotated ground truth. We assume the availability of trained network weights for the same task but obtained from a training set of indoor images. The iterative scheme that we propose is a form of transfer learning

that relies on auxiliary weak supervision instead of manually annotated ground truth to refine the network weights for outdoor images. The advantage of this approach is that the auxiliary supervision signal (the point cloud $\mathtt{P}^{\mathrm{MVS}}$) can be computed automatically under some mild assumptions about the training set, such as overlap of viewpoints and presence of textured regions. In the experiments section, we show that this approximate scheme successfully improves the quality of estimated plane segmentation masks. In the following, we present the details of each stage in the proposed approach.

### 3.1 Estimation of the initial segmentation masks

Instead of requiring ground truth annotations, we process the training images by an MVS pipeline to obtain a point cloud $\mathtt{P}^{\mathrm{MVS}}$ that will provide the necessary information to refine the network weights. Given $\mathtt{P}^{\mathrm{MVS}}$, it is possible to estimate the planar surfaces in the scene by robust geometric fitting using standard least squares estimation and RANSAC.

Although the reconstructed point cloud $\mathtt{P}^{\mathrm{MVS}}$ is informative, it has two problems that prevents computation of the segmentation masks and plane equations directly by fitting planes to the 3D points. Firstly, the estimated planes have infinite extent and direct computation of the segmentation boundaries solely based on 3D points is a challenging problem. Secondly, this point cloud is only partial since MVS pipelines produce data only for textured image regions. Regions of uniform intensity in the image do not contain enough texture. Consequently, we avoid directly computing the segmentation masks from $\mathtt{P}^{\mathrm{MVS}}$ by casting the estimation of segmentation masks as an energy minimization problem. Once a good segmentation mask is estimated, it is easy to refine the plane equations using the 3D points. Hence, the information provided by $\mathtt{P}^{\mathrm{MVS}}$ is exploited in the formulation of the energy terms for segmentation mask estimation and the subsequent plane parameter estimation.

The estimation of segmentation masks is most commonly formulated as a min-cut problem to be solved by graph cuts and we also make use of this formulation ( [4]). Instead of estimating a segmentation label for each pixel, we extract superpixels ( [1,2,13]) and solve for the superpixel labels. As a result, the segmentation mask $\mathtt{S}^{\mathtt{out}}$ is given as a set of label assignments $\{l_s : \forall s \in \mathscr{S}\}$, where $\mathscr{S}$ is the set of extracted superpixels and the labels $l_s \in \{-1, 0, 1, \ldots, K-1\}$ are selected from a set of $K+1$ possibilities. The labels with non-negative indices $j$, $j = 0, \ldots, K-1$ represent assignment to one of the possible planes $\pi^j$ in the network output and the label $-1$ represents non-planar regions that we denote as $\pi^{-1}$ for notational convenience. The formulation using superpixels allows us to adjust the granularity of the estimated ground truth to the density of the reconstructed point cloud and the resolution of the training images.

Before we estimate the segmentation labels by energy minimization, we calculate an initial set of labels as follows. We first project each 3D point $p \in \mathtt{P}^{\mathrm{MVS}}$ into the image and assign it to the superpixel it falls into based on superpixel boundaries. Similarly, we assign each projected point to a plane using the per-pixel segmentation masks estimated by the current network weights $f(\mathtt{D}^{\mathtt{out}} \mid \tilde{\mathbf{w}}_i)$. Within each superpixel $s$, each projected point votes for its assigned plane. Each superpixel is assigned an initial label $\hat{l}_s$ corresponding to the plane $\pi^{\hat{l}_s}$ that received the majority of the votes. The energy formulation described below ensures that the initial assignments $\{\hat{l}_s\}$ are also taken into account.

### 3.2 Updating the segmentation masks by energy minimization

Given the point cloud $\mathtt{P}^{\mathrm{MVS}}$ and the network weights $\tilde{\mathbf{w}}_i$ obtained in the last iteration, the energy of a superpixel segmentation mask $\mathtt{S}^{\mathtt{out}} = \{l_s\}$ is defined as follows

$$E\left(\mathtt{S}^{\mathtt{out}}\right) = \sum_{s \in \mathscr{S}} E_d\left(l_s \mid \mathtt{P}^{\mathrm{MVS}},\ f(\mathtt{D}^{\mathtt{out}} \mid \tilde{\mathbf{w}}_i)\right) + \lambda_s \sum_{(s,t) \in \mathscr{N}_{\mathscr{S}}} E_s(l_s, l_t \mid \mathtt{P}^{\mathrm{MVS}},\ f(\mathtt{D}^{\mathtt{out}} \mid \tilde{\mathbf{w}}_i)),$$

where $\mathscr{N}_{\mathscr{S}}$ is the set of neighboring superpixels.

The energy data term $E_d\left(l_s \mid \mathtt{P}^{\mathrm{MVS}},\ f(\mathtt{D}^{\mathtt{out}} \mid \tilde{\mathbf{w}}_i)\right)$ measures the discrepancy between a given superpixel label $l_s$ and the point cloud $\mathtt{P}^{\mathrm{MVS}}$. It depends on two components $E_{\mathrm{support}}(l_s \mid \mathtt{P}^{\mathrm{MVS}})$ and $E_{\mathrm{distance}}(l_s \mid \mathtt{P}^{\mathrm{MVS}})$. These components are combined in a weighted fashion as follows:

$$E_d(l_s) = \left(\alpha_1 + \delta(l_s - \hat{l}_s)\right) E_{\mathrm{support}}(l_s \mid \mathtt{P}^{\mathrm{MVS}}) + \left(\alpha_2 + \delta(l_s - \hat{l}_s)\right) E_{\mathrm{distance}}(l_s \mid \mathtt{P}^{\mathrm{MVS}}),$$

where $\alpha_1$ and $\alpha_2$ are scalar constants. $\delta(l_s - \hat{l}_s)$ is equal to zero whenever the new label is the same as the initial assignment, it is equal to one otherwise. The delta terms increase the cost of assignments that change the initial labels. This ensures that after minimization, the labels will change only when the 3D points of $\mathtt{P}^{\mathrm{MVS}}$ consistently get assigned to a plane other than the one indicated by the initial label.

$E_{\mathrm{support}}(l_s \mid \mathtt{P}^{\mathrm{MVS}})$ measures the ratio of projected 3D points assigned to the same plane as the label $l_s$. It is computed as

$$E_{\mathrm{support}}(l_s \mid \mathtt{P}^{\mathrm{MVS}}) = \frac{n_t - n_s}{n_t},$$

where $n_s$ is the number of projected points in the superpixel assigned to plane $\pi^{l_s}$ and $n_t$ is the total number of projected points in the superpixel.

$E_{\text{distance}}(l_s \mid \mathtt{P}^{\mathtt{MVS}})$ measures the average distance of projected 3D points to the plane corresponding to the label $l_s$. It is computed as

$$E_{\text{distance}}(l_s \mid \mathtt{P}^{\mathtt{MVS}}) = \frac{1}{n_t} \sum_{p \in s} d(\pi^{l_s}, p),$$

where $d(\pi^{l_s}, p)$ is the 3D Euclidean distance between the 3D point $p$ and plane $\pi^{l_s}$.

The smoothness data term $E_s(l_s, l_t \mid \mathtt{P}^{\mathtt{MVS}},\ f(\mathtt{D}^{\mathtt{out}} \mid \tilde{\mathbf{w}}_i))$ ensures that the estimated labels respect image color and depth information which regularizes the problem by constraining the labels of superpixels with a small number of projected 3D points. It is calculated as

$$\begin{aligned} E_s(l_s, l_t) = {} & E_{\text{color}}\left(l_s, l_t \mid \mathtt{P}^{\mathtt{MVS}},\ f(\mathtt{D}^{\mathtt{out}} \mid \tilde{\mathbf{w}}_i)\right) \\ & + \alpha_3 E_{\text{depth}}\left(l_s, l_t \mid \mathtt{P}^{\mathtt{MVS}},\ f(\mathtt{D}^{\mathtt{out}} \mid \tilde{\mathbf{w}}_i)\right), \end{aligned}$$

where $E_{\text{color}}(l_s, l_t)$ penalizes label changes over smooth intensity regions and $E_{\text{depth}}(l_s, l_t)$ penalizes label changes over regions of similar depth. They are calculated as

$$E_{\text{color}}(l_s, l_t) = \exp(-\Delta_c), \text{ and } E_{\text{depth}}(l_s, l_t) = \exp(-\Delta_d),$$

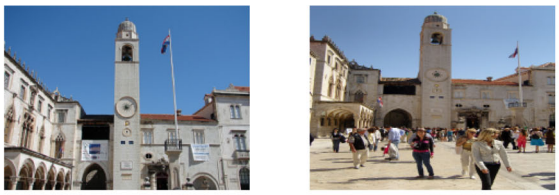
where $\Delta_c$ is the difference between mean intensity values (average of color channels) over superpixels $s$ and $t$, and $\Delta_d$ is the difference between mean depth values of 3D points projected into superpixels $s$ and $t$.

By finding a set of superpixel labels $\mathtt{S}^{\mathtt{out}} = \{l_s : \forall s \in \mathscr{S}\}$ that minimize the energy, we recover a new segmentation mask $\mathtt{S}^{\mathtt{out}}_{i+1}$ for planar regions and the non-planar areas. For each planar region, a set of updated parameters $\tilde{5}^{\mathtt{out}}_{i+1}$ are calculated by robust plane fitting to 3D points projecting onto the corresponding segment. The combined set of estimated segmentation mask and plane parameters $\tilde{\mathtt{T}}^{\mathtt{out}}_{i+1}$ can now be used in training. In the next section, we show that by repeating this process, we can improve the outdoor plane estimation performance of the network trained on indoor images to a large extent. We also provide a detailed analysis of the contribution from each energy term described in this section to the aforementioned performance increase.

## 4 Experiments

We have performed a set of experiments to verify that our approximate training approach is effective in improving the outdoor plane estimation performance of a state-of-the-art network trained on indoor data. Since our approach requires an outdoor dataset that is suitable for the geometric estimation, we have used popular structure from motion (SfM) and simultaneous localization and mapping (SLAM) benchmark datasets in our experiments. Both SfM and SLAM benchmarks provide suitable imagery of structured urban scenes that contain the necessary textured surfaces and viewpoint overlap that our approach relies upon. They also contain many planar surfaces and typical scenes for which outdoor plane estimation applications are likely to operate on.

We perform quantitative experiments that demonstrate a single iteration of our approach improves overall plane estimation quality over the baseline. Furthermore, we show that as proposed in Sect. 3, repeated iterations further improve performance. We also present qualitative results that demonstrate the improved segmentation performance as the transfer learning iterations progress. Finally, we present results of an



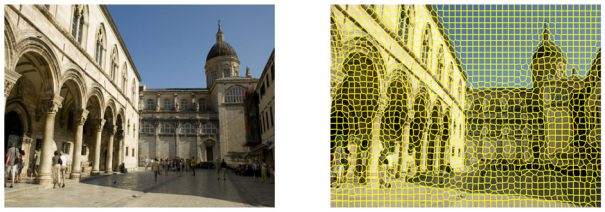
**(a)** Example images from Part I

**(b)** Example images from Part II

**(c)** Example images from Part III

**(d)** Point cloud computed for Part I

**Fig. 2** Dubrovnik dataset splits. **a–c** Three parts of the dataset where each consists of images that belong to a different part of the city. We form the training, validation, and the test sets with different parts to ensure spatial separation of the images in the splits. **d** The COLMAP output point cloud extracted from images in Part I



**(a)** Input Image **(b)** SLIC superpixels from (a)

**Fig. 3** Our training approach estimates a plane segmentation map for each image based on the reconstructed point cloud and the current network output. Since the point cloud is not dense enough to cover each pixel, we compute SLIC superpixels and estimate segmentation labels per superpixel. This ensures that the energy data terms for most of the superpixels depend on several projected 3D points that fall into the corresponding superpixel boundary

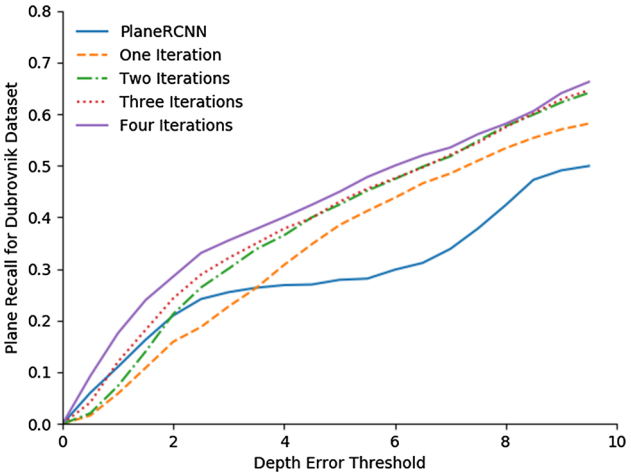



**Fig. 4** Average plane recall for the `Dubrovnik` dataset as the number of training iterations is increased

ablation study that measures the contribution from each of the energy terms detailed in Sect. 3.2.

### 4.1 Experiments on a structure-from-motion dataset

We have used the `Dubrovnik` dataset( [10]) as a primary set of images to test our approach. It includes 6844 city images taken by different cameras and from varying viewpoints. In order to have different training, validation and test splits, the Dubrovnik dataset is grouped into three parts each consisting of images that depict different city regions. This split ensures that our approach does not overfit to the textures of particular buildings in the same region. Example images corresponding to each part are shown in Fig. 2.

Measuring the plane estimation quality requires ground-truth data. Since the `Dubrovnik` dataset does not contain plane segmentation ground-truth, we manually annotated the plane boundaries in 50 images in each part for validation and test purposes. We emphasize that the manual annotations are exclusively used in evaluating test results and they are not used in the proposed training algorithm. The training set size is 150 for each part and no ground truth data is required for this set.

Each part is separately processed by COLMAP to compute the 3D point clouds that act as the weak-supervision training signal. We also compute SLIC( [1,2]) superpixels as shown in Fig. 3. We extract 1500 superpixels from each training image.

Since we have three splits, we perform six different experiments, where each part forms the training set twice with the other two alternating between validation and test splits. For each experimental run, we initialize the PlaneRCNN network with pretrained weights obtained by using the indoor dataset and evaluate its initial plane segmentation quality on the test set. We estimate a refined set of segmentation masks and plane parameters as described in Section 3 to act as the training targets on the outdoor dataset. We then retrain the PlaneRCNN layers that belong to the mask head, box head, classifier head and the depthmap decoder of the plane detection network for 60 epochs using the estimated training targets. The set of weights that gives the best piece-wise segmentation performance for the validation set of images is kept as the training result. Similarly, the scalar weights in the energy function, $\lambda_s$, $\alpha_1 - \alpha_3$ are set by a grid search that maximize the validation set performance.

In order to evaluate piece-wise plane segmentation accuracy, we measure the *Plane Recall* which is defined as the ratio of the number of estimated planes that have at least 0.5 Intersection over Union (IOU) score with one of the ground-truth planes to the number of ground-truth planes. The IOU score is measured with a varying depth error threshold from 0 to 1 meters with an increment of 0.05m for indoor images of PlaneRCNN. We set the depth error threshold to vary from 0 to 10 meters with an increment of 0.5m for outdoor images.

Figure 4 shows the performance of our approach for the `Dubrovnik` dataset experiments averaged over the six experimental runs described above. As the figure shows, our approach performs better than PlaneRCNN even with a single training iteration especially at larger depth error thresholds. As number of iterations increases, its performance becomes significantly better than the PlaneRCNN trained on indoor images for all depth thresholds. This shows that the automatically estimated training targets, plane segmentation boundaries and plane parameters are accurate enough to improve the network weights. Moreover, multiple iterations of our approach is able to improve the network performance on outdoor images showing its ability to transfer to this new domain without a strong supervision signal.

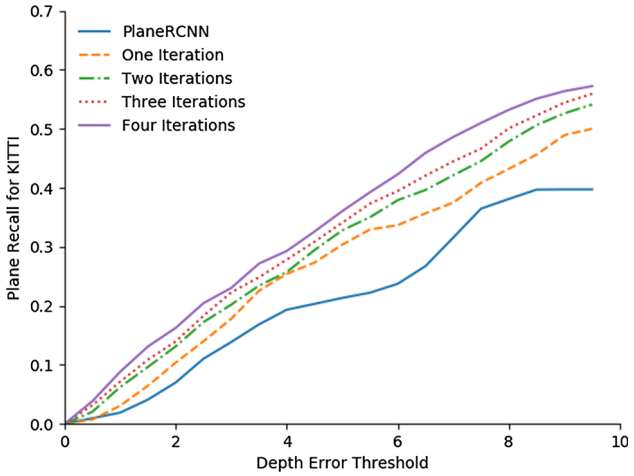



**Fig. 5** Average plane recall for the KITTI dataset test sequences. Even with a single iteration, our approach performs better than the PlaneRCNN-trained indoors despite being trained on images from the Dubrovnik dataset. Plane detection and segmentation performance improved for each of the depth error intervals as the number of iterations increase

## 4.2 Experiments on a SLAM dataset

To further test our approach and show its ability to improve outdoor plane segmentation accuracy, we perform another set of experiments on a SLAM dataset. For this purpose, we use the images from the KITTI dataset ( [9]) that depict urban scenes captured from a car travelling around city blocks for test purposes. We use the images in the Dubrovnik dataset to form the training and validation sets. Since the style of buildings and the distribution of viewpoints is very different between the datasets, this is a more stringent test setup than the previous set of experiments on Dubrovnik dataset.

KITTI dataset contains eleven test sequences numbered from 11 to 21. We randomly select 50 test images from all test sequences except those numbered 12, 14, 17, and 21 since these are taken outside the urban areas and do not contain planar regions except the ground plane. We manually annotate these test images, keep the experimental methodology the same as before and measure the plane recall for the initial network and after each iteration. Figure 5 shows the overall results averaged over all test sequences of the KITTI dataset. Despite the large visual differences between the images from two datasets, our approach trained on the Dubrovnik dataset significantly improves the outdoor plane estimation performance on the images of the KITTI dataset. As iterations progress this improvement slows down, so we stop the iterations at iteration four.

Figure 6 shows plane segmentation results for different test images from both the Dubrovnik and KITTI datasets for a qualitative comparison. The results from our approach correspond to network output after four iterations. PlaneRCNN trained for indoor images detects a few planar objects in the outdoor scenes with inaccurate segmentation boundaries and fails to segment most of the planar regions. The same architecture retrained on outdoor images by the proposed approach is able to detect most of the planes in the scene with more accurate boundaries. This improvement is achieved without providing detailed segmentation maps on the outdoor images of the training set.

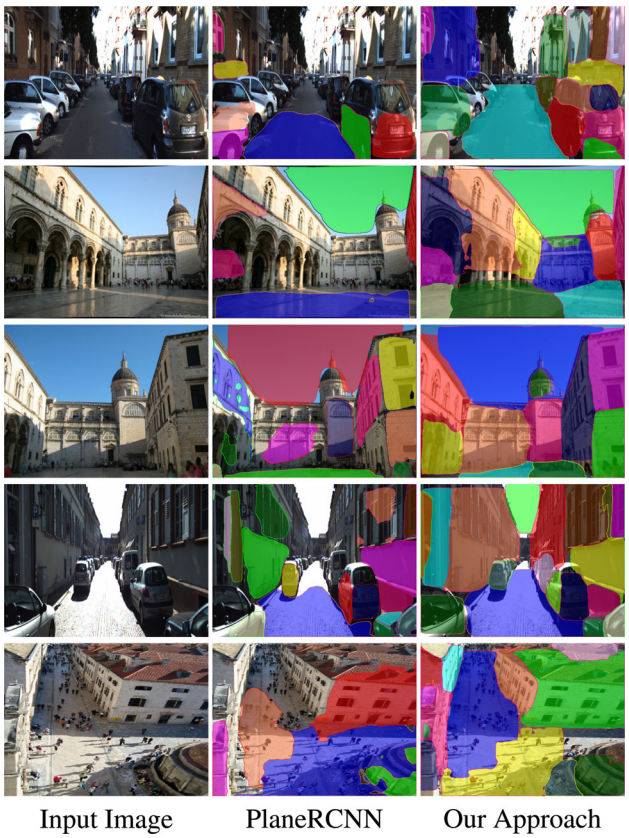



**Fig. 6** Comparison of piece-wise planar segmentation maps for different test images from both Dubrovnik and KITTI datasets. Results of our approach belong to estimations obtained after fourth training iteration. PlaneRCNN-trained indoors misses most of the planar regions and undersegments the detected ones. Despite this, the ground truth estimates automatically obtained with our energy minimization formulation are accurate enough to improve both plane detection rates and segmentation accuracy

## 4.3 Ablation study

In order to show the contribution of each individual data term of the energy function described in Sect. 3.2, we perform an additional experiment on the Dubrovnik dataset. We use Part I as training and test the plane recall obtained on Part II. As before, PlaneRCNN is retrained in four iterations for varying terms in the segmentation data cost function $E_d(l_s)$. The results of Table 1 show that the term measuring support from projected 3D points and the one measuring 3D

**Table 1** Plane recall values as the data term $E_d(l_s)$ varies. To better understand the effect of different parts of the segmentation energy data cost, we gradually add more complex terms and measure the plane recall for each variation. Adding terms for both $E_{\text{support}}$ and $E_{\text{distance}}$ improve results over using either term. Using the additive $\delta(l_s - \hat{l}_s)$ factor also boost results by increasing the label cost changes when there is less evidence from the point cloud

| $E_d(l_s)$ | Depth Error Threshold 0.0–2.5 | 2.5–5.0 | 5.0–7.5 | 7.5–10 |
|---|---|---|---|---|
| $\alpha_1 E_{\text{support}}$ | 0.101 | 0.226 | 0.261 | 0.282 |
| $\alpha_2 E_{\text{distance}}$ | 0.063 | 0.141 | 0.172 | 0.207 |
| $\alpha_1 E_{\text{support}} + \alpha_2 E_{\text{distance}}$ | 0.124 | 0.276 | 0.313 | 0.357 |
| $\left(\alpha_1 + \delta(l_s - \hat{l}_s)\right) E_{\text{support}}$ | 0.171 | 0.495 | 0.578 | 0.621 |
| $\left(\alpha_2 + \delta(l_s - \hat{l}_s)\right) E_{\text{distance}}$ | 0.227 | 0.499 | 0.551 | 0.593 |
| $\left(\alpha_1 + \delta(l_s - \hat{l}_s)\right) E_{\text{support}}$ + $\left(\alpha_2 + \delta(l_s - \hat{l}_s)\right) E_{\text{distance}}$ | **0.316** | **0.614** | **0.667** | **0.701** |

Results in bold indicate the best plane recall values obtained in each column

plane distance both contribute to plane segmentation performance. Moreover, the $\delta(l_s - \hat{l}_s)$ terms that increase the cost of label changes positively affect the network training. Best results are obtained when the data term matches the final form described in Sect. 3.2.

## 5 Conclusion

We have proposed an approach for increasing outdoor plane detection and segmentation accuracy without requiring manual annotations. Our approach is able to improve the weights of a network trained on indoors images with the sole requirement that the training dataset can be processed by an SfM/MVS pipeline. We have shown that the weak supervision supplied by the initial network and the point cloud is enough to effectively estimate the ground truth labels on the outdoor images. This enables improving both the weights and consequently the estimate of the outdoor ground truth. Repeated applications of this idea further improve the detection and segmentation accuracy on several images of outdoor scenes. As a result, we have shown that a network trained on indoor images can be tuned to operate on outdoor imagery without requiring manual training annotations.

Currently, our approach is limited to training data that can be processed by SfM pipelines and the point cloud data are extracted only for textured surfaces. Some of these limitations might be overcome by integrating lidar point cloud data. Since the proposed iterative framework is quite generic, it might be possible to integrate domain cues such as known vanishing lines into the energy terms such as in [19].

## References


1. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua P., Süsstrunk, S.: SLIC superpixels compared to state-of-the-art superpixel methods. IEEE Trans. Pattern Anal. Mach. Intell. **34**(11), 2274–2282 (2012). https://doi.org/10.1109/TPAMI.2012.120
2. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S.: Slic superpixels compared to state-of-the-art superpixel methods. IEEE Trans. Pattern Anal. Mach. Intell. **34**(11), 2274–2282 (2012)
3. Bódis-Szomorú, A., Riemenschneider, H., Van Gool, L.: Fast, approximate piecewise-planar modeling based on sparse structure-from-motion and superpixels. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 469–476 (2014)
4. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. IEEE Trans. Pattern Anal. Mach. Intell. **23**(11), 1222–1239 (2001)
5. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM **24**(6), 381–395 (1981)
6. Freedman, D.: An improved image graph for semi-automatic segmentation. SIVP **6**, 533–545 (2012)
7. Furukawa, Y., Curless, B., Seitz, S.M., Szeliski, R.: Manhattan-world stereo. In: 2009 IEEE conference on computer vision and pattern recognition, pp. 1422–1429. IEEE (2009)
8. Gallup, D., Frahm, J.M., Pollefeys, M.: Piecewise planar and non-planar stereo for urban scene reconstruction. In: 2010 IEEE computer society conference on computer vision and pattern recognition, pp. 1418–1425. IEEE (2010)
9. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? The kitti vision benchmark suite. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2012)
10. Li, Y., Snavely, N., Huttenlocher, D.P.: Location recognition using prioritized feature matching. In: European conference on computer vision, pp. 791–804. Springer (2010)
11. Liu, C., Kim, K., Gu, J., Furukawa, Y., Kautz, J.: Planercnn: 3d plane detection and reconstruction from a single image. In: The IEEE conference on computer vision and pattern recognition (CVPR) (2019)
12. Liu, C., Yang, J., Ceylan, D., Yumer, E., Furukawa, Y.: Planenet: Piece-wise planar reconstruction from a single rgb image. In: The IEEE conference on computer vision and pattern recognition (CVPR) (2018)
13. Liu, G., Duan, J.: RGB-D image segmentation using superpixel and multi-feature fusion graph theory. SIVP **14**, 1171–1179 (2020)

14. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Conference on computer vision and pattern recognition (CVPR) (2016)
15. Schönberger, J.L., Zheng, E., Pollefeys, M., Frahm, J.M.: Pixelwise view selection for unstructured multi-view stereo. In: European Conference on Computer Vision (ECCV) (2016)
16. Sinha, S., Steedly, D., Szeliski, R.: Piecewise planar stereo for image-based rendering (2009)
17. Yang, F., Zhou, Z.: Recovering 3d planes from a single image via convolutional neural networks. In: The European Conference on Computer Vision (ECCV) (2018)
18. Yu, Z., Zheng, J., Lian, D., Zhou, Z., Gao, S.: Single-image piecewise planar 3d reconstruction via associative embedding. CoRR **abs/1902.09777** (2019)
19. Zeng, Z., Wu, M., Zeng, W., Fu, C.-W.: Deep recognition of vanishing-point-constrained building planes in urban street views. IEEE Trans. Image Process. **29**, 5912–5923 (2020)