PlaneDepth: Self-supervised Depth Estimation via Orthogonal Planes

Ruoyu Wang1, Zehao Yu2, Shenghua Gao1,3,4*
1ShanghaiTech University 2University of Tübingen
3Shanghai Engineering Research Center of Intelligent Vision and Imaging
4Shanghai Engineering Research Center of Energy Efficient and Custom AI IC
{wangry3, gaoshh}@shanghaitech.edu.cn, zehao.yu@uni-tuebingen.de

Abstract

Multiple near frontal-parallel planes based depth representation demonstrated impressive results in self-supervised monocular depth estimation (MDE). Whereas, such a representation would cause the discontinuity of the ground as it is perpendicular to the frontal-parallel planes, which is detrimental to the identification of drivable space in autonomous driving. In this paper, we propose the PlaneDepth, a novel orthogonal planes based presentation, including vertical planes and ground planes. PlaneDepth estimates the depth distribution using a Laplacian Mixture Model based on orthogonal planes for an input image. These planes are used to synthesize a reference view to provide the self-supervision signal. Further, we find that the widely used resizing and cropping data augmentation breaks the orthogonality assumptions, leading to inferior plane predictions. We address this problem by explicitly constructing the resizing cropping transformation to rectify the predefined planes and predicted camera pose. Moreover, we propose an augmented self-distillation loss supervised with a bilateral occlusion mask to boost the robustness of orthogonal planes representation for occlusions. Thanks to our orthogonal planes representation, we can extract the ground plane in an unsupervised manner, which is important for autonomous driving. Extensive experiments on the KITTI dataset demonstrate the effectiveness and efficiency of our method. The code is available at https://github.com/svip-lab/PlaneDepth.

1. Introduction

Monocular depth estimation (MDE) is an important task in computer vision and it has tremendous potential applications, such as autonomous driving. However, the expensive data and labels acquisition process restricts the data scale in supervised MDE [3,4,9,16,26,27]. Thus, researchers turn to solve the data constraints in supervised MDE with the self-supervised MDE framework by leveraging videos or stereo image pairs.

Most of the early works in self-supervised MDE leverage a regression module to estimate pixel-wise depth map [11, 12, 15, 29, 36, 40] and warp the reference view image to the target view based on the estimated depth. Then, a photometric consistency loss is used to guide the learning of the depth regression module. However, these methods usually encounter the local minimum issue because of the locality of bilinear interpolation on the reference view. To avoid this issue, rather than using simple depth regression, multiple frontal-parallel planes based depth representation is introduced where depth space is divided into a fixed number of frontal-parallel planes, and the depth network learns to classify which predefined plane each pixel belongs to [13, 14]. It has been shown that such representation could produce much sharper depth on the edges of the object. However, they are insufficient to represent the ground because the ground plane is perpendicular to these predefined frontal-parallel planes. As shown in Fig. 1, such vertical depth planes only solution would lead to discontinuity on the ground, which is obviously detrimental to the identification of drivable space in autonomous driving. Further, photometric consistency loss is applied to the weighted composition of each plane-warped image, which is sub-optimal as the combination of different weights may lead to exactly the same color image [30], resulting in ambiguous solutions for depth plane classification.

Considering the ground is perpendicular to the frontal-parallel planes, in this paper, we propose to leverage orthogonal planes to represent the scene where the ground planes favor the depth estimation in the ground region. Further, we propose to model the depth as a mixture of Laplace distributions of orthogonal planes [38], where each Laplacian is centered at one plane. We compute the photometric consistency loss independently on the color image warped by each plane, resulting in a more deterministic and less ambiguous optimization objective compared with the weighted com-
Figure 1. Monocular Depth Estimation Results. The zoomed-in visualizations and bird-eye-view colored point cloud show that our method can predict continuous depth in the ground region while preserving the sharp edges of objects. Compared with PladeNet [13], our prediction has smoother depth in the ground region. Compared with DepthHint [40], our method addresses the occlusion problem, which can be seen in figure that our prediction have eliminated depth artifacts in the edges of street lights and cars.

2. Related Works

2.1. Self-supervised Depth Estimation

Self-supervised learning of depth estimation from video sequences or stereo image pairs has been proposed to ease the demand for large-scale labeled data. Since the seminal work by Zhou et al. [46] that demonstrated inspiring results by jointly optimizing depth and pose networks with an image reconstruction loss. Many following work further improve the performance through occlusion modeling [12, 33, 34, 41], addressing scale ambiguity [45], modeling moving objects [2, 15, 24], enforcing depth consistency [11, 19], designing better network architectures [17, 31, 49], designing better network architectures [17, 31, 49], incorporating additional semantic information [5, 6, 23, 28, 50], exploiting relative object size cues [14, 29], or extending to indoor scenes [21, 25, 44]. In this paper, we propose to utilize orthogonal planes for better depth representation in driving scenarios and demonstrate improved performance with orthogonal planes.

2.2. Plane-based Scene Representation

Benefiting from the simplicity of planes and the efficiency of differentiable homography warping, multi-plane image (MPI) is widely used in novel view synthesis [16, 25, 47]. The idea is also introduced in the depth estimation problem is cast as the optimizing the mixture of Laplace distribution, which avoids the ambiguity in color expectation based depth estimation and leads to more stable depth estimation.

3. An orthogonality-preserved data augmentation strategy is proposed, which improves the robustness of network training.

4. We combine post-processing with self-distillation by our augmented self-distillation, which improves both efficiency and accuracy.
3. Methods

Previous works [13, 14] use planes parallel to the camera to represent the scene which would lead to a discontinuity of the ground. In this paper, we propose to model the depth distribution with a mixture of Laplacian based on a set of orthogonal planes consisting of the vertical and ground planes. Then orthogonality-preserved data augmentation is introduced by taking the resized cropping transformation and Neural Positional Encoding. We also propose an augmented self-distillation with the supervision of the bilateral occlusion mask to improve the robustness of our PlaneDepth to occlusions.

3.1. Orthogonal Planes based Representation

To improve the capability of the network in predicting the depth for both ground and objects at a distance, we propose to represent the scene with a set of vertical planes as well as ground planes, which are orthogonal.

**Plane Definition.** Take the target view as the origin, the planes are defined as:

\[
\mathbf{n}_i^T \mathbf{w} - \delta_i = 0, \quad i = 0, 1, \ldots, N - 1
\]

where \(\mathbf{n}\) is the normal of the plane, \(\mathbf{w}\) is a point in the world coordinate system, \(\delta_i\) is the distance from the \(i\)-th plane to the origin, and \(N\) is the number of planes.

**Vertical Planes.** Following [14], we initialize \(N^v\) vertical planes with \(\mathbf{n}^v = [0 \ 0 \ 1]^T\), and sample \(\delta_i^v\) in exponential
disparity space:
\[
\delta \hat{i} = \frac{B f_x}{d_i} - d_i = d_{\text{max}} \left( \frac{d_{\text{min}}}{d_{\text{max}}} \right)^{i+1} 
\]
where \( B \) is the baseline of the stereo pair and \( f_x \) is the horizontal focal length, \( d_{\text{min}} \) and \( d_{\text{max}} \) are the minimum and maximum disparity, respectively. \( r_i^g \) are predicted by the network to account for discretization errors.

**Ground Planes.** Similarly, we initialize \( N^g \) ground planes with \( \mathbf{n}^g = [0 \ 1 \ 0]^T \) and different \( \delta_i^g \) linearly in camera height space:
\[
\delta_i^g = h_{\text{min}} + \frac{j + r_i^g}{N_g - 1} (h_{\text{max}} - h_{\text{min}}),
\]
where \( h_{\text{min}} \) and \( h_{\text{max}} \) are the minimum and maximum height of the ground planes, respectively. Again, \( r_i^g \) are predicted by the network to account for discretization errors.

**Depth from Orthogonal Planes.** We model the depth as a mixture of Laplace distributions of orthogonal planes, with each Laplacian centered at one plane. Therefore, PlaneDepth also predicts weights \( \pi_i \) and scales \( \sigma_i \) for each Laplace distribution at each pixel, where \( \pi_i \) is normalized by a softmax function \( S \) of logit \( l_i \). Plane residuals \( r_i^p \) and \( r_i^g \) are shared by all pixels, and each pixel has its own parameters for mixture Laplace distributions. We omit the subscript of pixel for simplicity. Now we can compute the depth \( \hat{D} \) for each pixel via composition:
\[
\hat{D} = \frac{1}{Z} \sum_{i=0}^{N-1} p_i D_i, \quad p_i = \sum_{j=0}^{N-1} \frac{\pi_j e^{-|D_i - D_j|/\sigma_j}}{2\sigma_j}
\]
where \( p_i \) is the probability that the pixel belongs to \( i \)-th plane, \( Z = \sum_{i=0}^{N-1} p_i \) is a normalization constant and \( D_i \) is rendered depth of \( i \)-th plane.

**Homography Warping.** Given relative camera pose \( R \) and \( t \), either from calibrated stereo pairs or predicted by a pose network, between the target image and the reference image, we construct homography mappings \( H_i \) for the \( i \)-th plane:
\[
H_i = K (R + \frac{1}{\delta_i} (\mathbf{t} \mathbf{n}_i^T)) K^{-1},
\]
where \( K \) is the camera intrinsic matrix. Following [13, 14, 49], we warp the target image \( \mathbf{I}_t \) together with pixel-wise plane distributions \( \pi_i, \sigma_i \) to the reference view via bilinear sampling \( G \):
\[
\hat{\mathbf{I}}_i, \hat{\mathbf{l}}_i, \hat{\sigma}_i = G(H_i, \mathbf{I}_t, \mathbf{t}_i, \sigma_i) \quad \hat{\sigma} = S([\hat{\mathbf{l}}_i]_{i=0}^{N-1}).
\]
Note that our warped weights \( \hat{\sigma} \) is obtained by warping logit \( \mathbf{l} \) followed by softmax \( S \) [14]. Then we can synthesis a reference image \( \mathbf{I}_r \) via composition:
\[
\hat{\mathbf{I}}_r = \frac{1}{Z} \sum_{i=0}^{N-1} \hat{p}_i \hat{\mathbf{I}}_i, \quad \hat{p}_i = \sum_{j=0}^{N-1} \frac{\pi_j e^{-|D_j - D_i|/\sigma_j}}{2\sigma_j},
\]
where \( \hat{Z} = \sum_{i=0}^{N-1} \hat{p}_i \) and \( D_j^r \) is rendered depth of the \( i \)-th plane in the reference view.

**Optimization.** We use perceptual loss [22] to constrain the feature-level similarity of the synthesis reference view:
\[
L_{\text{pc}} = ||\phi_l(\mathbf{I}_r) - \phi_l(\mathbf{I}_t)||_2^2;
\]
where \( \phi_l \) is the first \( l \) maxpool layers of a VGG19 [37] pre-trained on the ImageNet [7]. Further, we propose a novel photometric consistency based on Laplace distributions:
\[
L_{\text{Mll}} = - \log N \sum_{i=0}^{N-1} \frac{\frac{1}{\pi_i} e^{-\frac{||\hat{\mathbf{l}}_i - \mathbf{l}_i||_1}{\sigma_i}}}{2\sigma_i}.
\]
Compared to \( L_1 \) loss which supervise the final composite image \( \mathbf{I}_r \), our \( L_{\text{Mll}} \) supervise each plane warped image \( \mathbf{I}_i \), resulting in a more deterministic optimization objective.

We also use smooth loss \( L_{\text{ds}} \) for depth map \( \hat{D} \) as it has been shown effective in many recent works [12-14, 25, 36, 40]. Therefore, our final loss is:
\[
L = L_{\text{Mll}} + \lambda_1 L_{\text{pc}} + \lambda_2 L_{\text{ds}}
\]
which is averaged over pixels, views and batches. \( \lambda_1 \) and \( \lambda_2 \) are hyper-parameters to balance different loss terms.

### 3.2. Orthogonality-preserved Data Augmentation

Resizing and cropping is a widely used data augmentation strategy, which encourages the MDE to exploit relative size cue [2]. However, when the crop does not occur at the center height of the image, the ground will tilt, as shown in Fig. 3, and contradict our orthogonal planes assumptions. Consequently, our PlaneDepth network struggles to predict accurate parameters for such augmented data. We address this issue by calculating the transformation matrix of the augmented data explicitly.

We assume that when the size of the object becomes \( f_s \) times larger, its depth is reduced by a factor \( f_s \). Given an image corresponding to a camera intrinsic with focal length \( (f_x, f_y) \) and principal point \( (c_x, c_y) \), when it is scaled with a factor \( f_s \) and cropped with a grid centered at \( (p_x, p_y) \), the resizing and cropping lead to a transformation \( R_C \) which transforms the original world coordinates \( \mathbf{w} \) to the new coordinates \( \mathbf{w} \) originated at the augmented image:
\[
\mathbf{w} = R_C \mathbf{w} \quad R_C = \begin{bmatrix} 1 & \frac{c_y - p_y}{f_y} & \frac{f_x - p_x}{f_s} \\ \frac{c_y - p_y}{f_y} & 1 & \frac{c_x - p_x}{f_s} \end{bmatrix}.
\]

Similarly, the normal \( \mathbf{n} \) and distance \( \delta \) of planes in the augmented image are also transformed as:
\[
\hat{\mathbf{n}} = R_C^{-T} \mathbf{n} \quad \hat{\mathbf{d}} = \frac{\delta}{||R_C^{-T} \mathbf{n}||}.
\]
As a result, the vertical planes $\vec{n}^u$ and $\vec{n}^q$ after transformation are not orthogonal\(^1\). To maintain the orthogonality of our predefined planes, we use Eq. (12) to rectify ground planes during training such that they are always perpendicular to the ground after resizing and cropping.

Further, after resizing and cropping augmentation, it is desirable that the networks would predict the depth, disparity and pose as\(^2\):

$$\tilde{d} = \frac{d}{f_s}, \quad \tilde{D} = f_s D, \quad \tilde{R} = R_C R_R C^{-1}, \quad \tilde{t} = R_C t \quad (13)$$

However, $\tilde{R}$ is not orthogonal and hence not a canonical rotation matrix. Therefore, $\tilde{R}$ cannot be predicted with 3 degrees of freedom as existing pose estimation methods [42]. In order to be compatible with these methods, we explicitly rectify the rotation matrix $R$ predicted by the pose network to $\tilde{R}$ according to Eq. (13), which ease the learning of the pose network.

**Neural Positional Encoding.** After adjusting the planes and camera rotation, the network still has difficulty in segmenting the ground because it needs to predict the degree of inclination caused by resize cropping before predicting the ground height $\theta^g$. In order to help the network to better segment the ground, we encode the $\mathbf{g}$ with the neural positional encoding (NPE) [13], where the grid $\mathbf{g} \in [-1, 1]^2$ is represented as the relative position of each pixel in the original image, after the resizing and cropping transformation. Since our encoder is designed to extract image features based on the relative sizes of objects, we only feed the NPE of the grid into each layer of the decoder, as shown in Fig. 2. Similarly for the pose network, in order to avoid the network predicting changes from $\tilde{R}$ to $R$, we also feed NPE of $\mathbf{g}$ into the pose network.

### 3.3. Augmented Self-distillation

Self-distillation is proven to be effective for getting rid of occlusion effects [13, 14, 49]. Since pixels that are not visible in the right view exist only on the left side of objects in the left view, which means that the photometric matching failure caused by occlusion only occurs on the left side of the object in the left view. Hence, a network that takes

---

\(^1\)More details are provided in the supplementary material.

\(^2\)The derivation of Eq. (11), Eq. (12) and Eq. (13) are provided in the supplementary material.
In this way, we can obtain occlusion-aware self-distillation supervision combined with post-processing as:

\[ d_{sd} = M^L_{RL}(M^L_{LR}d_{pp} + (1 - M^L_{LR})d) + (1 - M^L_{RL})d_{fl} \]  

(16)

where \(d_{fl}\) is the flipped disparity map of the flipped input, as shown in Fig. 4. We use \(L_1\) loss on the predicted disparity in the self-distillation stage as \(L_{sd} = ||d - d_{sd}||_1\).

We also modify our mixture Laplace and perceptual loss in the self-distillation stage using the occlusion mask in the right view as in [14]. We get the right view occlusion mask \(M^R\) by using Eq. (14) to warp the weight \(\pi\) from the left view to the right. Thus we use \(M^R\) to mask out the loss of the occluded areas in the right view, getting \(L_{M^R_{MLL}}\) and \(L_{M^R_{pc}}\).

Therefore, our final loss in the self-distillation stage is:

\[ L_{distill} = L_{M^R_{MLL}} + \lambda_1 L_{M^R_{pc}} + \lambda_2 L_{ds} + \lambda_3 L_{sd} \]  

(17)

which is averaged over pixel, view and batch. \(\lambda_1\), \(\lambda_2\) and \(\lambda_3\) are the weight hyper-parameters.

4. Experiments

4.1. Training Strategy

We first train our model with \(L\) for 50 epochs with a batch size of 8, where half of input images are obtained by flipping the other half. Then we fine-tune the network with another epoch at high resolution without resizing and cropping data augmentation, which allows the model to utilize two monocular depth cues simultaneously. We then train another 10 epochs with the self-distillation \(L_{distill}\) with a batch size of 4 still without resizing and cropping at high resolution. More implementation details are provided in the supplementary material.

4.2. Dataset

Following previous work [29, 36, 40, 42, 48], we mainly conducted our experiments on the widely used KITTI [10] dataset, which provides sequential stereo images and sparse point clouds using stereo cameras and LiDAR mounted on a moving car. We followed the train test splits used in [12]. During stereo training, we adopt the Eigen training split [8], which contains 22,600 training and 888 validation stereo pairs. As for monocular training, we use the Zhou’s split [46], which removes static frames and contains 19,905 training and 2,212 validation stereo pairs. We evaluate our models on the 697 Eigen raw test images [8] and 652 Eigen improved test images [39] using the metrics proposed in [10], where Eigen improved test set [39] is obtained by warping the ground truth depth of 11 adjacent frames into one frame and filtering by SGM [20]. During both Eigen evaluations, all ground truth and predictions are clipped to within 80m and cropped by as in Eigen crop, and evaluated using the metrics proposed in [10].

3More details are provided in the supplementary material.

4.3. Performance Comparisons

We compare our PlaneDepth against existing state-of-the-art MDE methods on the raw and improved KITTI Eigen test set [8], and the quantitative results are shown in Tab. 2. In stereo setting, our method outperforms all competing methods on the Improved Eigen test set [39] even without using post-processing [11], which is two times slower because the network needs to make an additional forward for the flipped images. Further, our method performs similarly with or without post-processing thanks to our augmented self-distillation loss. It is worth noting that our proposed augmented self-distillation label further improves performance, but it is slightly less efficient compared to post-processing. Our method also demonstrated consistent improvements when using monocular videos and stereo pairs for training (MS in the table). We also show qualitative comparison against previous state-of-the-art methods in Fig. 5 and Fig. 1. Our method predicts continuous depth in the ground region while preserving sharp object edges, recovering more fine details and having fewer depth artifacts around the object boundary compared to other methods. It is also interesting that all methods, including ours, perform slightly worse in the MS setting than in the stereo setting. We attribute it to dynamic objects and inaccurate pose predictions in monocular videos.

4.4. Ablation Studies

We have shown improved performance compared to existing methods, now we conduct experiments to validate the effectiveness of each component of our method.

Orthogonal Planes Prediction. We have tried various combinations of plane predictions and show the results in Table 1. While the quantitative results changed slightly when adding more vertical planes or ground planes, using ground planes brings additional benefits that we can extract the ground plane without supervision and predict continuous depth as shown in Fig. 6 and Fig. 1, respectively, which is important for autonomous driving.

<table>
<thead>
<tr>
<th>#VP</th>
<th>#GP</th>
<th>Abs Rel ↓</th>
<th>Sq Rel ↓</th>
<th>Rmse ↓</th>
<th>Al ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>49</td>
<td>0</td>
<td>0.090</td>
<td>0.590</td>
<td>4.147</td>
<td>0.899</td>
</tr>
<tr>
<td>63</td>
<td>0</td>
<td>0.090</td>
<td>0.615</td>
<td>4.166</td>
<td>0.900</td>
</tr>
<tr>
<td>49</td>
<td>14</td>
<td>0.089</td>
<td>0.598</td>
<td>4.175</td>
<td>0.900</td>
</tr>
</tbody>
</table>

Table 1. Comparison of different plane combinations, where #VP and #GP are the number of vertical and ground planes, respectively. We also evaluate a model using planes with 3-axis normals.

Mixture Laplace Loss. Quantitative results in Table 3 show that our mixture Laplace loss outperforms simple L1 loss used in baseline methods [13, 14]. To further quantify the effects of different losses, we propose a new metric: Mean Maximum Probability (MMP)\(^4\) which reflects the

\(^4\)Definition is provided in the supplementary material.
Table 2. Comparison of performance on KITTI Eigen test set [8]. The best is in bold and the second best is underlined in each metric. S stands for stereo training using left and right views and a fixed baseline and MS stands for both stereo and monocular training adding front and rear frames of the left view without camera poses. ResNet-50-A represents U-Net [35] with DenseASPP module [43] using ResNet-50 [18] as the backbone. PP with checkmark refers to post-processing [46]. *: Due to the limitation of memory usage for high-resolution training, our MS training is only performed at low-resolution in stage one without self-distillation. †: Post-processing with our method for generating self-distillation labels.

Table 3. Comparison of different photometric loss, where MLL denotes our mixture Laplace loss. MMP stands for Stand Maximum Probability. VP means only using the vertical planes, OP means using our orthogonal planes. The result shows that MLL is effective for plane-based methods and makes predictions more focused on a specific plane.

Table 4. Comparison of ground prediction. Smoother ground depth with the help of \( R_C \) and NPE brings a performance boost.

Without NPE, the network has difficulty in predicting the tilt of the ground plane, resulting in discontinuous depth patches. In contrast, with the help of \( R_C \) and NPE, our method can utilize ground planes easily, resulting in more accurate ground segmentation and smoother ground depth.

**Augmented Self-distillation.** We first compare the accuracy of pseudo labels generated by different methods. As shown in Tab. 6, our proposed bilateral mask combined with post-processing generates more accurate labels than other methods, which is beneficial for self-distillation. Further, we compare quantitative results with or without self-distillation loss in Tab. 7, which verifies that the performance improvement is due to our self-distillation loss rather than the longer training time.
Figure 5. Qualitative results on the KITTI dataset. Our network predicts smooth depth for the ground while preserving thin structures and sharp object edges with fewer depth artifacts.

Table 5. Comparison of different training strategies. LR and HR denote low resolution (640×192) and high resolution (1280×384) training, respectively. The results show that our approach of combining LR training with one additional HR epoch, utilizing both monocular depth cues, significantly improves performance.

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Abs Rel</th>
<th>Sq Rel</th>
<th>Rmse</th>
<th>A1↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR</td>
<td>0.090</td>
<td>0.616</td>
<td>4.188</td>
<td>0.898</td>
</tr>
<tr>
<td>HR</td>
<td>0.090</td>
<td>0.630</td>
<td>4.270</td>
<td>0.898</td>
</tr>
<tr>
<td>Ours</td>
<td>0.086</td>
<td>0.581</td>
<td>4.094</td>
<td>0.906</td>
</tr>
</tbody>
</table>

Table 6. Comparison of different distillation label, where PP denotes post-processing [11], UM denotes unilateral occlusion mask [14], and PP+BM denotes our bilateral mask combined with post-processing. Our proposed method generate more accurate self-distillation labels.

<table>
<thead>
<tr>
<th>Source</th>
<th>Abs Rel</th>
<th>Sq Rel</th>
<th>Rmse</th>
<th>A1↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>PP</td>
<td>0.085</td>
<td>0.557</td>
<td>4.020</td>
<td>0.909</td>
</tr>
<tr>
<td>UM</td>
<td>0.084</td>
<td>0.552</td>
<td>3.982</td>
<td>0.912</td>
</tr>
<tr>
<td>PP+BM</td>
<td>0.083</td>
<td>0.540</td>
<td>3.940</td>
<td>0.913</td>
</tr>
</tbody>
</table>

Table 7. Ablation study on self-distillation. Result shows that using self-distillation loss $\mathcal{L}_{distill}$ significantly improves performance compared to naively increasing training time.

<table>
<thead>
<tr>
<th>Loss</th>
<th>Abs Rel</th>
<th>Sq Rel</th>
<th>Rmse</th>
<th>A1↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathcal{L}$</td>
<td>0.086</td>
<td>0.584</td>
<td>4.132</td>
<td>0.905</td>
</tr>
<tr>
<td>$\mathcal{L}_{distill}$</td>
<td>0.085</td>
<td>0.563</td>
<td>4.023</td>
<td>0.910</td>
</tr>
</tbody>
</table>

5. Conclusion

We propose PlaneDepth, a novel orthogonal planes based depth representation for self-supervised depth estimation, which enables segmenting and predicting a continuous ground plane compared to existing frontal-parallel planes based representations. The depth of the scene is represented as a mixture of Laplacian of the orthogonal planes. Considering that traditional data augmentation may break the orthogonality constraint of the planes, we solve this explicitly by calculating the resizing cropping transformation as well as the neural position encoding. We further propose to improve the final prediction by incorporating post-processing into the self-distillation objective via our bilateral occlusion masks. Extensive experiments on KITTI validate the effectiveness and efficiency of our proposed approach.

Limitations. We do not constrain the choice of different types of planes and the inclination caused by car bumping, so our unsupervised ground segmentation may not be robust enough to be directly used in safety-critical applications. Based on this, combining with other unsupervised ground segmentation methods, or explicitly modeling car tilt may be interesting future directions.

Acknowledgements. This work was supported by National Key R&D Program of China (2018AAA0100704), NSFC #61932020, #62172279, Science and Technology Commission of Shanghai Municipality (Grant No.20ZR1436000), Program of Shanghai Academic Research Leader, and "Shuguang Program" supported by Shanghai Education Development Foundation and Shanghai Municipal Education Commission.
References


