PlaneDepth: Self-supervised Depth Estimation via Orthogonal Planes

Ruoyu Wang[1] Zehao Yu[2] Shenghua Gao[1,3,4*]
[1]ShanghaiTech University [2]University of Tübingen
[3]Shanghai Engineering Research Center of Intelligent Vision and Imaging
[4]Shanghai 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 compo-

*Corresponding Author.

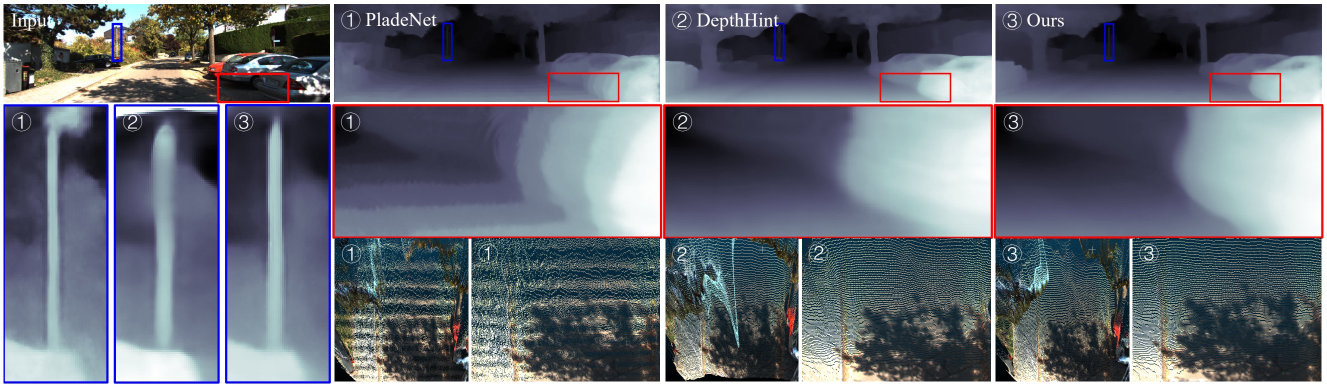

Figure 1. **Monocular Depth Estimation Results**. The zoomed-in visualizations and bird-eye-view colored point cloud show that our method can predict continuous depth in the ground region while preserving the sharp edges of objects. Compared with PladeNet [13], our prediction has smoother depth in the 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.

sition strategy mentioned above [13, 14], consequently leading to better depth estimation results, as shown in Fig. 1. Moreover, we can extract the ground plane in an unsupervised manner thanks to our orthogonal planes representation.

Resizing and cropping are widely used as a data augmentation strategy in the stereo training setting of MDE [13, 14]. However, it would destroy the orthogonality of our predefined planes, leading to inferior plane distributions estimation. To remedy this issue, in this paper, we deeply analyze the effects of resizing and cropping on the world coordinates system. We explicitly compute the resizing and cropping transformation and use it to rectify the predefined planes and the predicted camera rotation, which eases the learning of plane distributions. We further propose to use neural positional encoding (NPE) for the resizing and cropping parameter and incorporate it into our PlaneDepth network, which improves the robustness of network training.

The self-distillation strategy is commonly used to solve occlusion problems [13, 14] and improve the depth prediction results [29]. Post-processing is widely used to improve the final prediction [11], which can be used naturally to generate more accurate self-distillation labels. In this paper, we propose to combine post-processing with self-distillation by using a bilateral occlusion mask to generate more accurate supervision of network training, which improves both accuracy and efficiency of our method.

We summarize our contributions as follows:

1. We propose the PlaneDepth, a novel orthogonal plane-based monocular depth estimation network, which favours the representation of both vertical objects and ground. Such representation leads to a much smoother depth for ground regions and would facilitate the identification of drivable regions.
2. The depth within the scene is modeled by a mixture of Laplacian distributions, and the depth classification 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.

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], 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 estima-

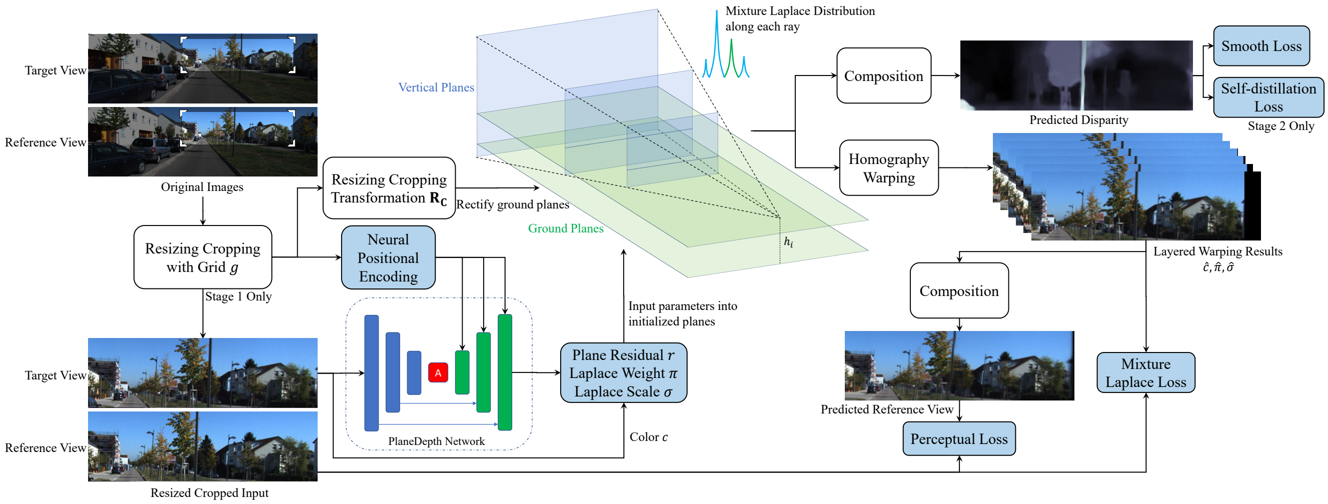

Figure 2. **PlaneDepth Pipeline**. We model the depth as a mixture of Laplace distribution of orthogonal planes for self-supervised depth estimation. Our PlaneDepth network predicts distribution parameters for these planes, which together with the target view image are warped to the reference view. The plane-warped images are composed to synthesize a predicted reference view which is supervised with perceptual loss. We further supervise each individual plane-warped image with the photometric consistency loss between the reference view weighted by the Laplace distributions. We address the non-orthogonality issue due to resizing and cropping by explicitly rectifying predefined planes and neural positional encoding from the resizing cropping grid. A self-distillation loss is used to further improve the performance. We use double-channel ResNet18 [18] following Godard *et al*. [12] and add NPE to predict the relative pose of two views in our monocular training experiments. The blue and white boxes indicate the differentiable variables and the hard operations, respectively.

tion [13, 14] recently and has been shown effective in self-supervised depth estimation since it addresses the local minima problem caused by local searching in the reference view. Zhou and Dong [48] further address the depth discontinuity problem by regressing a depth residual for each plane. However, these MPI-based networks will predict vertical-prone depth in the ground region which is the dominant region in the outdoor driving scenario. In contrast, we propose an orthogonal planes based representation that can model continuous depth and free space in the ground region while preserving the sharp edges of objects.

2.3. Self-distillation for Depth Estimation

Self-distillation is a technique commonly used in self-supervised depth estimation that involves using a pretrained teacher network to generate labels for training a better student network. For instance, Peng *et al*. [29, 32] synthesized multiple predictions from teacher network to generate pseudo labels, which are then used to train the student network. Self-distillation can also be used to reduce the uncertainty of the prediction [1, 42], or solve the artifact problem caused by occlusion [14, 48]. Inspired by these works and the widely-used post-processing (PP) technique, which inputs both an original image and its flipped version into a depth network and averages the predictions for more accurate depth estimation, we propose to utilize a bilateral occlusion mask into PP technique to synthesis pseudo labels which is used as an additional distillation loss to further improve our depth network.

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:

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

where $\mathbf{n}$ is the normal of the plane, $\mathbf{w}$ is a point in the world coordinate system, δ_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 δ_i^v in exponential disparity space:

$$\delta_i^v = \frac{Bf_x}{d_i^v} \qquad d_i^v = d_{\text{max}}(\frac{d_{\text{min}}}{d_{\text{max}}})^{\frac{i+r_i^v}{N_v-1}} \tag{2}$$

where B is the baseline of the stereo pair and f_x is the horizontal focal length, d_{min} and d_{max} are the minimum and maximum disparity, respectively. r_i^v 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 δ_i^g linearly in camera height space:

$$\delta_i^g = h_{\text{min}} + \frac{i + r_i^g}{N_g - 1}(h_{\text{max}} - h_{\text{min}}), \tag{3}$$

where h_{min} and h_{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 π_i and scales σ_i for each Laplace distribution at each pixel, where π_i is normalized by a softmax function $\mathcal{S}$ of logit l_i. Plane residuals r_i^v 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:

$$\hat{D} = \frac{1}{Z}\sum_{i=0}^{N-1} p_i D_i \qquad p_i = \sum_{j=0}^{N-1} \frac{\pi_j e^{\frac{-|D_i - D_j|}{\sigma_j}}}{2\sigma_j} \tag{4}$$

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 $\mathbf{R}$ and $\mathbf{t}$, either from calibrated stereo pairs or predicted by a pose network, between the target image and the reference image, we construct homography mappings $\mathbf{H}_i$ for the i-th plane:

$$\mathbf{H}_i = \mathbf{K}(\mathbf{R} + \frac{1}{\delta_i}(\mathbf{t}\mathbf{n}_i^T))\mathbf{K}^{-1}, \tag{5}$$

where $\mathbf{K}$ is the camera intrinsic matrix. Following [13, 14, 49], we warp the target image $\mathbf{I}_t$ together with pixel-wise plane distributions π_i, σ_i to the reference view via bilinear sampling $\mathcal{G}$:

$$\hat{\mathbf{I}}_i, \hat{l}_i, \hat{\sigma}_i = \mathcal{G}(\mathbf{H}_i, \mathbf{I}_t, l_i, \sigma_i) \qquad \hat{\pi} = \mathcal{S}([\hat{l}_i]_{i=0}^{N-1}). \tag{6}$$

Note that our warped weights $\hat{\pi}$ is obtained by warping logit l followed by softmax $\mathcal{S}$ [14]. Then we can synthesis a reference image $\hat{\mathbf{I}}_r$ via composition:

$$\hat{\mathbf{I}}_r = \frac{1}{\hat{Z}}\sum_{i=0}^{N-1} \hat{p}_i \hat{\mathbf{I}}_i \qquad \hat{p}_i = \sum_{j=0}^{N-1} \frac{\hat{\pi}_j e^{\frac{-|D_i^r - D_j^r|}{\hat{\sigma}_j}}}{2\hat{\sigma}_j}, \tag{7}$$

where $\hat{Z} = \sum_{i=0}^{N-1} \hat{p}_i$ and D_i^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:

$$\mathcal{L}_{\text{pc}} = ||\phi_l(\mathbf{I}_r) - \phi_l(\hat{\mathbf{I}}_r)||_2^2, \tag{8}$$

where ϕ_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:

$$\mathcal{L}_{\text{MLL}} = -\log \sum_{i=0}^{N-1} \frac{\hat{\pi}_i e^{\frac{-\frac{1}{3}||\mathbf{I}_r - \hat{\mathbf{I}}_i||_1}{\hat{\sigma}_i}}}{2\hat{\sigma}_i}. \tag{9}$$

Compared to L_1 loss which supervise the final composite image $\hat{\mathbf{I}}_t$, our $\mathcal{L}_{\text{MLL}}$ supervise each plane warped image $\hat{\mathbf{I}}_i$, resulting in a more deterministic optimization objective.

We also use smooth loss $\mathcal{L}_{\text{ds}}$ 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:

$$\mathcal{L} = \mathcal{L}_{\text{MLL}} + \lambda_1\mathcal{L}_{\text{pc}} + \lambda_2\mathcal{L}_{\text{ds}} \tag{10}$$

which is averaged over pixels, views and batches. λ_1 and λ_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 $\mathbf{R_C}$ which transforms the original world coordinates w to the new coordinates $\tilde{w}$ originated at the augmented image:

$$\tilde{\mathbf{w}} = \mathbf{R_C}\mathbf{w} \qquad \mathbf{R_C} = \begin{bmatrix} 1 & & \frac{c_x - p_x}{f_x} \\ & 1 & \frac{c_y - p_y}{f_y} \\ & & f_s \end{bmatrix}. \tag{11}$$

Similarly, the normal $\mathbf{n}$ and distance δ of planes in the augmented image are also transformed as:

$$\tilde{\mathbf{n}} = \frac{\mathbf{R_C}^{-T}}{||\mathbf{R_C}^{-T}\mathbf{n}||}\mathbf{n} \qquad \tilde{\delta} = \frac{\delta}{||\mathbf{R_C}^{-T}\mathbf{n}||}. \tag{12}$$

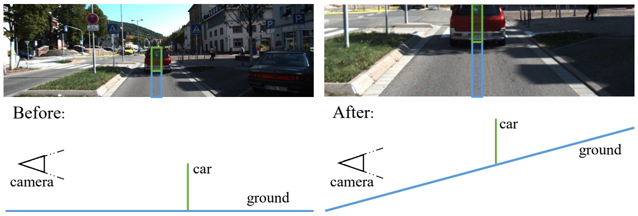

Figure 3. Resizing cropping causes planes to be non-orthogonal.

As a result, the vertical planes $\tilde{\mathbf{n}}^v$ and $\tilde{\mathbf{n}}^g$ 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]:

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

However, $\widetilde{\mathbf{R}}$ is not orthogonal and hence not a canonical rotation matrix. Therefore, $\widetilde{\mathbf{R}}$ 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 $\mathbf{R}$ predicted by the pose network to $\widetilde{\mathbf{R}}$ 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 δ^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 $\widetilde{\mathbf{R}}$ to $\mathbf{R}$, 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 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 only the left view as input and finds matchings in the right view will produce artifacts to the left of objects. Gonzalez *et al.* [14] exploits this property to generate pseudo labels for self-distillation with a **unilateral** occlusion mask, which corresponds to the areas on left view that are not visible in the right view due to occlusion. It can be computed as:

$$\mathbf{M}^{\mathrm{L}}_{\mathrm{RL}} = \min(\sum_{i=0}^{N-1} \mathcal{W}_{\mathrm{R}\to\mathrm{L}}(\hat{\pi}^{\mathrm{R}}_i, d_i), 1) \quad (14)$$

$$\hat{\pi}^{\mathrm{R}} = \mathcal{S}([\mathcal{W}_{\mathrm{L}\to\mathrm{R}}(l^L_i, d_i)]^{N-1}_{i=0}) \quad (15)$$

where $\mathcal{W}(\cdot, d)$ is a warping using disparity d and l_i is the logit of the i-th plane. L and R represent left and right, respectively. Taking advantage of the property that the network only predicts artifacts on the left side of objects, they flip the left image and flip the predicted disparity map back, getting a disparity map with only artifacts on the right side of the objects d_{ff}, where the disparity to the left of the objects is accurate. Thus they can supervise artifacts areas on the left side of the object through the unilateral mask $\mathbf{M}^L_{RL}$ and the accurate disparity on the left side of the object d_{ff}.

Post-processing is a method commonly used in depth estimation to improve the final prediction [11], which inputs both the original image and its flipped image to a depth network and then averages the prediction as $d_{\mathrm{pp}} = \frac{1}{2}(d + d_{\mathrm{ff}})$, where d and d_{ff} are the predicted disparity map of the original image and its flipped image, respectively.

A natural idea is to use d_{pp} to further improve the accuracy of the self-distillation label. However, as aforementioned, the left-side artifact of d and the right-side artifact of d_{ff} cause artifacts to exist on both sides of the objects in d_{pp}. In this paper, rather than using unilateral occlusion masks, we leverage the **bilateral** occlusion masks to filter both sides artifacts in d_{pp}. By assuming the left view to be the right view, the occlusion mask on the right side of objects $\mathbf{M}^{\mathrm{L}}_{\mathrm{LR}}$ can be obtained simply by swapping $\mathcal{W}_{\mathrm{L}\to\mathrm{R}}$ and $\mathcal{W}_{\mathrm{R}\to\mathrm{L}}$ in Eq. (14) and Eq. (15).

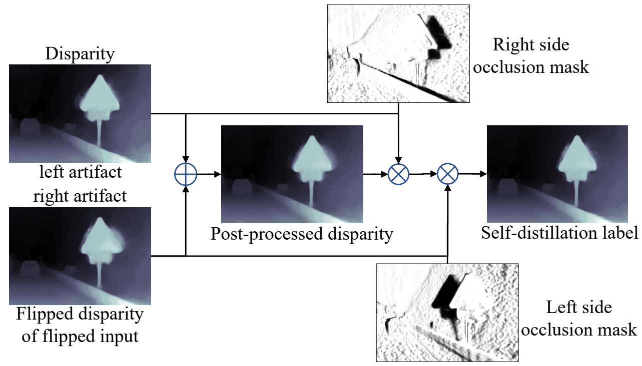

Figure 4. **Obtain the self-distillation label**. Our final self-distillation labels are obtained by filtering the artifacts on both sides of the post-processed disparity map using the corresponding side mask and accurate disparity respectively.

[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:

$$d_{\text{sd}} = \mathbf{M}^{\text{L}}_{\text{RL}}(\mathbf{M}^{\text{L}}_{\text{LR}} d_{\text{pp}} + (1 - \mathbf{M}^{\text{L}}_{\text{LR}})d) + (1 - \mathbf{M}^{\text{L}}_{\text{RL}})d_{\text{ff}} \quad (16)$$

where d_{ff} 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 $\mathcal{L}_{\text{sd}} = ||\hat{d} - d_{\text{sd}}||_1$.

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 $\mathbf{M}^{\text{R}}$ by using Eq. (14) to warp the weight π from the left view to the right. Thus we use $\mathbf{M}^{\text{R}}$ to mask out the loss of the occluded areas in the right view, getting $\mathcal{L}^{M}_{\text{MLL}}$ and $\mathcal{L}^{M}_{\text{pc}}$[3].

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

$$\mathcal{L}_{\text{distill}} = \mathcal{L}^{M}_{\text{MLL}} + \lambda_1 \mathcal{L}^{M}_{\text{pc}} + \lambda_2 \mathcal{L}_{\text{ds}} + \lambda_3 \mathcal{L}_{\text{sd}} \quad (17)$$

which is averaged over pixel, view and batch. λ_1, λ_2 and λ_3 are the weight hyper-parameters.

4. Expariments

4.1. Training Strategy

We first train our model with $\mathcal{L}$ 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 $\mathcal{L}_{\text{distill}}$ 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].

[3]More 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.

#VP	#GP	Abs Rel↓	Sq Rel↓	Rmse↓	A1↑
49	0	0.090	**0.590**	4.147	0.899
63	0	0.090	0.615	4.166	**0.900**
49	14	**0.089**	0.598	4.175	**0.900**
3-axis planes		0.090	0.598	**4.109**	**0.900**

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.

Methods	PP	Network	Resolution	Train	Abs Rel↓	Sq Rel↓	RMSE↓	RMSE log↓	A1↑	A2↑	A3↑
Raw Eigen Test Set [8]											
Monodepth2 [12]	✓	ResNet-50	1024×320	S	0.097	0.793	4.533	0.181	0.896	0.965	0.983
DepthHint [40]	✓	ResNet-50	1024×320	S	0.096	0.710	4.393	0.185	0.890	0.962	0.981
EPCDepth [29]	✓	ResNet-50	1024×320	S	0.091	0.646	4.207	0.176	0.901	0.966	0.983
OCFDNet [48]	✓	ResNet-50-A	1280×384	S	0.090	0.564	4.007	0.172	0.903	0.967	0.984
FALNet [14]		FALNet	1280×384	S	0.097	0.590	3.991	0.177	0.893	0.966	0.984
FALNet [14]	✓	FALNet	1280×384	S	0.093	0.564	**3.973**	0.174	0.898	0.967	**0.985**
PLADENet [13]		PLADENet	1280×384	S	0.092	0.626	4.046	0.175	0.896	0.965	0.984
PLADENet [13]	✓	PLADENet	1280×384	S	0.089	0.590	4.008	0.172	0.900	0.967	**0.985**
Ours		ResNet-50-A	1280×384	S	0.085	0.563	4.023	0.171	0.910	**0.968**	0.984
Ours	✓	ResNet-50-A	1280×384	S	**0.084**	**0.549**	3.981	**0.169**	**0.911**	**0.968**	0.984
Ours†	✓	ResNet-50-A	1280×384	S	0.083	0.533	3.919	0.167	0.913	0.969	0.985
Monodepth2 [12]	✓	ResNet-18	1024×320	MS	0.104	0.775	4.562	0.191	0.878	0.959	0.981
DepthHint [40]	✓	ResNet-18	1024×320	MS	0.098	0.702	4.398	0.183	0.887	**0.963**	**0.983**
FeatureNet [36]		ResNet-50	1024×320	MS	0.099	0.697	4.427	0.184	0.889	**0.963**	0.982
Ours*		ResNet-50-A	1280×384	MS	0.092	0.601	4.188	0.184	0.893	0.961	0.981
Ours*	✓	ResNet-50-A	1280×384	MS	**0.090**	**0.584**	**4.130**	**0.182**	**0.896**	0.962	0.981
Improved Eigen Test Set [39]											
Monodepth2 [12]	✓	ResNet-50	1024×320	S	0.077	0.455	3.489	0.119	0.933	0.988	0.996
DepthHint [40]	✓	ResNet-50	1024×320	S	0.074	0.364	3.202	0.114	0.936	0.989	0.997
OCFDNet [48]	✓	ResNet-50-A	1280×384	S	0.070	0.262	2.786	0.103	0.951	0.993	0.998
FALNet [14]	✓	FALNet	1280×384	S	0.071	0.281	2.912	0.108	0.943	0.991	0.998
PLADENet [13]	✓	PLADENet	1280×384	S	0.066	0.272	2.918	0.104	0.945	0.992	0.998
Ours		ResNet-50-A	1280×384	S	**0.063**	0.245	2.718	0.096	0.959	0.994	0.998
Ours	✓	ResNet-50-A	1280×384	S	**0.063**	**0.236**	**2.674**	**0.095**	**0.960**	**0.994**	**0.999**
Ours†	✓	ResNet-50-A	1280×384	S	0.062	0.227	2.609	0.093	0.961	0.995	0.999

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.

Loss	Plane	Abs Rel↓	Sq Rel↓	Rmse↓	A1↑	MMP
L1	VP	0.093	0.624	4.402	0.891	0.574
L1	OP	0.095	0.606	4.293	0.893	0.543
MLL	VP	0.090	**0.590**	**4.147**	0.899	0.610
MLL	OP	**0.089**	0.598	4.175	**0.900**	0.607

Table 3. Comparison of different photometric loss, where MLL denotes our mixture Laplace loss. MMP stands for Mean 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.

$\mathbf{R_C}$	NPE	Abs Rel↓	Sq Rel↓	Rmse↓	A1↑
		0.092	0.596	4.143	0.898
	✓	0.090	**0.593**	4.186	0.899
✓		0.094	**0.593**	**4.141**	0.898
✓	✓	**0.089**	0.598	4.175	**0.900**

Table 4. Comparison of ground prediction. Smoother ground depth with the help of $\mathbf{R_C}$ and NPE brings a performance boost.

confidence of the network's prediction on certain planes. As shown in Table 3, using MML results in higher accuracy and MMP, indicating that MML is more deterministic and less ambiguous compared to using L1 loss on the composite image.

Orthogonality-preserved Data Augmentation. We show the quantitative ablation of transformation matrix $\mathbf{R_C}$ and neural positional encoding (NPE) in Tab. 4 and qualitative comparison on ground prediction in Fig. 6. Without using $\mathbf{R_C}$ to rectify predefined ground planes, the tilt of the ground makes the network relies on vertical planes completely, resulting in discontinuity in the ground region. 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 $\mathbf{R_C}$ 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.

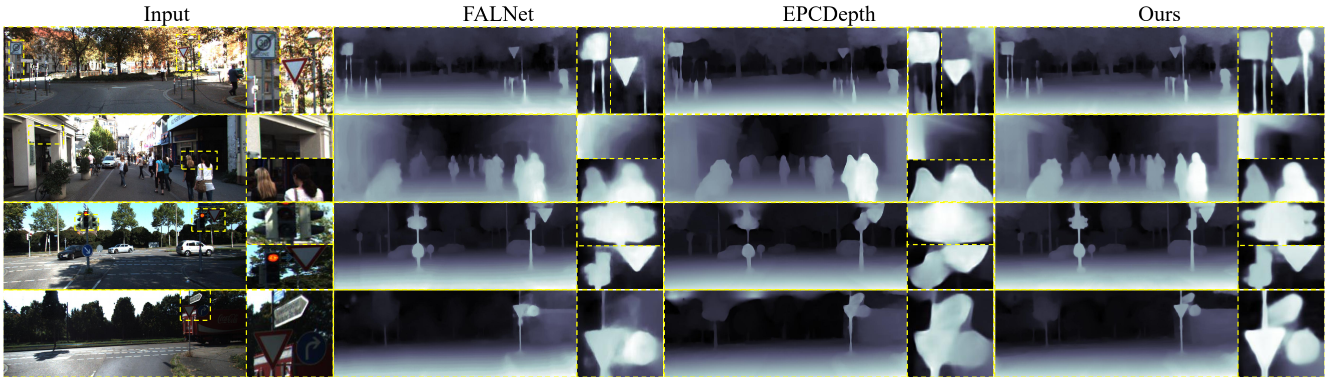

Figure 5. **Qualitative results on the KITTI dataset.** Our network predicts smooth depth for the ground while preserving thin structures and sharp object edges with fewer depth artifacts.

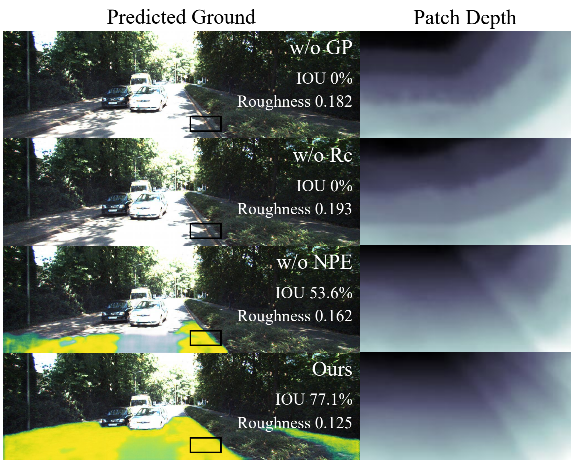

Figure 6. **Predicted Ground**. We use IOU and L2 gradient of disparity map to evaluate the segmentation and roughness of ground on the Eigen test set, respectively. With the help of $\mathbf{R_C}$ and NPE, our method can predict more accurate segmentation and smoother depth for the ground.

Strategy	Abs Rel↓	Sq Rel↓	Rmse↓	A1↑
LR	0.090	0.616	4.188	0.898
HR	0.090	0.630	4.270	0.898
Ours	**0.086**	**0.581**	**4.094**	**0.906**

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 improvements performance.

Source	Abs Rel↓	Sq Rel↓	Rmse↓	A1↑
PP	0.085	0.557	4.020	0.909
UM	0.084	0.552	3.982	0.912
PP+BM	**0.083**	**0.540**	**3.940**	**0.913**

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.

Loss	Abs Rel↓	Sq Rel↓	Rmse↓	A1↑
$\mathcal{L}$	0.086	0.584	4.132	0.905
$\mathcal{L}_{\text{distill}}$	**0.085**	**0.563**	**4.023**	**0.910**

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

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

[1] Juan Luis Gonzalez Bello and Munchurl Kim. Self-supervised deep monocular depth estimation with ambiguity boosting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 3

[2] Juan Luis Gonzalez Bello, Jaeho Moon, and Munchurl Kim. Positional information is all you need: A novel pipeline for self-supervised svde from videos. *arXiv preprint arXiv:2205.08851*, 2022. 2, 4

[3] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4009–4018, 2021. 1

[4] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Localbins: Improving depth estimation by learning local distributions. In *European Conference on Computer Vision*, pages 480–496. Springer, 2022. 1

[5] Po-Yi Chen, Alexander H Liu, Yen-Cheng Liu, and Yu-Chiang Frank Wang. Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2624–2632, 2019. 2

[6] Xingyu Chen, Ruonan Zhang, Ji Jiang, Yan Wang, Ge Li, and Thomas H Li. Self-supervised monocular depth estimation: Solving the edge-fattening problem. *arXiv preprint arXiv:2210.00411*, 2022. 2

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 4

[8] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014. 6, 7

[9] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2002–2011, 2018. 1

[10] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 6

[11] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 270–279, 2017. 1, 2, 5, 6, 8

[12] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3828–3838, 2019. 1, 2, 3, 4, 6, 7

[13] Juan Luis Gonzalez and Munchurl Kim. Plade-net: Towards pixel-level accuracy for self-supervised single-view depth estimation with neural positional encoding and distilled matting loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6851–6860, 2021. 1, 2, 3, 4, 5, 6, 7

[14] Juan Luis GonzalezBello and Munchurl Kim. Forget about the lidar: Self-supervised depth estimators with med probability volumes. *Advances in Neural Information Processing Systems*, 33:12626–12637, 2020. 1, 2, 3, 4, 5, 6, 7, 8

[15] Ariel Gordon, Hanhan Li, Rico Jonschkowski, and Anelia Angelova. Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8977–8986, 2019. 1, 2

[16] Shir Gur and Lior Wolf. Single image depth estimation trained via depth from defocus cues. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7683–7692, 2019. 1, 2

[17] Wencheng Han, Junbo Yin, Xiaogang Jin, Xiangdong Dai, and Jianbing Shen. Brnet: Exploring comprehensive features for monocular depth estimation. In *European Conference on Computer Vision*, pages 586–602. Springer, 2022. 2

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3, 7

[19] Mu He, Le Hui, Yikai Bian, Jian Ren, Jin Xie, and Jian Yang. Ra-depth: Resolution adaptive self-supervised monocular depth estimation. In *European Conference on Computer Vision*, pages 565–581. Springer, 2022. 2

[20] Heiko Hirschmuller. Accurate and efficient stereo processing by semi-global matching and mutual information. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 807–814. IEEE, 2005. 6

[21] Pan Ji, Runze Li, Bir Bhanu, and Yi Xu. Monoindoor: Towards good practice of self-supervised monocular depth estimation for indoor environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12787–12796, 2021. 2

[22] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 4

[23] Hyunyoung Jung, Eunhyeok Park, and Sungjoo Yoo. Fine-grained semantics-aware representation enhancement for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12642–12652, 2021. 2

[24] Marvin Klingner, Jan-Aike Termöhlen, Jonas Mikolajczyk, and Tim Fingscheidt. Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance. In *European Conference on Computer Vision*, pages 582–600. Springer, 2020. 2

[25] Jiaxin Li, Zijian Feng, Qi She, Henghui Ding, Changhu Wang, and Gim Hee Lee. Mine: Towards continuous depth mpi with nerf for novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12578–12588, 2021. 2, 4

[26] Zhenyu Li, Xuyang Wang, Xianming Liu, and Junjun Jiang. Binsformer: Revisiting adaptive bins for monocular depth estimation. *arXiv preprint arXiv:2204.00987*, 2022. 1

[27] Yue Luo, Jimmy Ren, Mude Lin, Jiahao Pang, Wenxiu Sun, Hongsheng Li, and Liang Lin. Single view stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 155–163, 2018. 1

[28] Jingyuan Ma, Xiangyu Lei, Nan Liu, Xian Zhao, and Shiliang Pu. Towards comprehensive representation enhancement in semantics-guided self-supervised monocular depth estimation. In *European Conference on Computer Vision*, pages 304–321. Springer, 2022. 2

[29] Rui Peng, Ronggang Wang, Yawen Lai, Luyang Tang, and Yangang Cai. Excavating the potential capacity of self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15560–15569, 2021. 1, 2, 3, 6, 7

[30] Rui Peng, Rongjie Wang, Zhenyu Wang, Yawen Lai, and Ronggang Wang. Rethinking depth estimation for multi-view stereo: A unified representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8645–8654, 2022. 1

[31] Sudeep Pillai, Rareş Ambruş, and Adrien Gaidon. Superdepth: Self-supervised, super-resolved monocular depth estimation. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 9250–9256. IEEE, 2019. 2

[32] Andrea Pilzer, Stephane Lathuiliere, Nicu Sebe, and Elisa Ricci. Refine and distill: Exploiting cycle-inconsistency and knowledge distillation for unsupervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9768–9777, 2019. 3

[33] Matteo Poggi, Fabio Tosi, and Stefano Mattoccia. Learning monocular depth estimation with unsupervised trinocular assumptions. In *2018 International conference on 3d vision (3DV)*, pages 324–333. IEEE, 2018. 2

[34] Michael Ramamonjisoa, Yuming Du, and Vincent Lepetit. Predicting sharp and accurate occlusion boundaries in monocular depth estimation using displacement fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14648–14657, 2020. 2

[35] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 7

[36] Chang Shu, Kun Yu, Zhixiang Duan, and Kuiyuan Yang. Feature-metric loss for self-supervised learning of depth and egomotion. In *European Conference on Computer Vision*, pages 572–588. Springer, 2020. 1, 4, 6, 7

[37] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4

[38] Fabio Tosi, Yiyi Liao, Carolin Schmitt, and Andreas Geiger. Smd-nets: Stereo mixture density networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8942–8952, 2021. 1

[39] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. In *2017 international conference on 3D Vision (3DV)*, pages 11–20. IEEE, 2017. 6, 7

[40] Jamie Watson, Michael Firman, Gabriel J Brostow, and Daniyar Turmukhambetov. Self-supervised monocular depth hints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2162–2171, 2019. 1, 2, 4, 6, 7

[41] Alex Wong and Stefano Soatto. Bilateral cyclic constraint and adaptive regularization for unsupervised monocular depth prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5644–5653, 2019. 2

[42] Hongbin Xu, Zhipeng Zhou, Yali Wang, Wenxiong Kang, Baigui Sun, Hao Li, and Yu Qiao. Digging into uncertainty in self-supervised multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6078–6087, 2021. 3, 5, 6

[43] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3684–3692, 2018. 7

[44] Zehao Yu, Lei Jin, and Shenghua Gao. Pnet: Patch-match and plane-regularization for unsupervised indoor depth estimation. In *European Conference on Computer Vision*, pages 206–222. Springer, 2020. 2

[45] Sen Zhang, Jing Zhang, and Dacheng Tao. Towards scale-aware, robust, and generalizable unsupervised monocular depth estimation by integrating imu motion dynamics. In *European Conference on Computer Vision*, pages 143–160. Springer, 2022. 2

[46] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1851–1858, 2017. 2, 6, 7

[47] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018. 2

[48] Zhengming Zhou and Qiulei Dong. Learning occlusion-aware coarse-to-fine depth map for self-supervised monocular depth estimation. *arXiv preprint arXiv:2203.10925*, 2022. 3, 6, 7

[49] Zhengming Zhou and Qiulei Dong. Self-distilled feature aggregation for self-supervised monocular depth estimation. In *European Conference on Computer Vision*, pages 709–726. Springer, 2022. 2, 4, 5

[50] Shengjie Zhu, Garrick Brazil, and Xiaoming Liu. The edge of depth: Explicit constraints between segmentation and depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13116–13125, 2020. 2