Multi-Task Planar Reconstruction with Feature Warping Guidance

Luan Wei[1], Anna Hilsmann[1] and Peter Eisert[1,2]
[1]*Fraunhofer Heinrich Hertz Institute, Berlin, Germany*
[2]*Humboldt University, Berlin, Germany*

Keywords: Planar Reconstruction, Real-Time, Neural Network, Segmentation, Deep Learning, Scene Understanding.

Abstract: Piece-wise planar 3D reconstruction simultaneously segments plane instances and recovers their 3D plane parameters from an image, which is particularly useful for indoor or man-made environments. Efficient reconstruction of 3D planes coupled with semantic predictions offers advantages for a wide range of applications requiring scene understanding and concurrent spatial mapping. However, most existing planar reconstruction models either neglect semantic predictions or do not run efficiently enough for real-time applications. We introduce SOLOPlanes, a real-time planar reconstruction model based on a modified instance segmentation architecture which simultaneously predicts semantics for each plane instance, along with plane parameters and piece-wise plane instance masks. We achieve an improvement in instance mask segmentation by including multi-view guidance for plane predictions in the training process. This cross-task improvement, training for plane prediction but improving the mask segmentation, is due to the nature of feature sharing in multi-task learning. Our model simultaneously predicts semantics using single images at inference time, while achieving real-time predictions at 43 FPS. Code is available at: https://github.com/fraunhoferhhi/SOLOPlanes.

1 INTRODUCTION

Estimating the 3D structure of a scene holds importance across a variety of domains, including robotics, virtual reality (VR), and augmented reality (AR). The demand for real-time applications in these areas increases over time as such technologies proliferate. Man-made architectures and indoor environments, where the application end-users spend a significant amount of time, often consist of regular structures like planar surfaces, aligning well with the Manhattan world assumption that such surfaces typically exist on a regular 3D grid (Coughlan and Yuille, 2003). Estimating plane parameters directly can reduce noise for areas lying on a planar surface, which can be particularly useful for indoor scenes dominated by planar surfaces. It also holds relevance for outdoor scenarios, such as self-driving cars and outdoor AR applications, where streets and buildings often adhere to similar geometric principles.

Several methods have been proposed to use deep learning to recover planes of indoor scenes from a single image (Liu et al., 2018a; Yu et al., 2019; Liu et al., 2019; Xie et al., 2021b,a). While existing works have made strides in predicting piece-wise instance masks and plane parameters, they often ignore the added information from scene semantics. Incorporating semantics provides an added layer of scene understanding, which can be useful in many applications. For instance, the semantic label for a planar surface can help a service robot in determining the correct

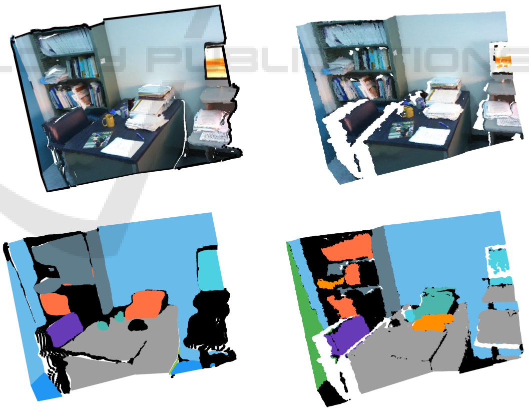
(a) Ours. (b) GT.

Figure 1: Comparison of SOLOPlanes output with ground truth (GT). 3D projections using predicted plane parameters (left) and GT depth (right). Textures use RGB (top), predicted semantics (bottom left), and GT semantics (bottom right).

Wei, L., Hilsmann, A. and Eisert, P.
Multi-Task Planar Reconstruction with Feature Warping Guidance.
DOI: 10.5220/0012396200003660
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2024) - Volume 2: VISAPP*, pages 652-661
ISBN: 978-989-758-679-8; ISSN: 2184-4321

behaviour (*e.g.* mopping floor vs. wiping table), or AR/VR experiences could offer semantics-dependent retexturization. Some models predict semantics along with plane parameters but are often too computationally intensive to meet the real-time requirements of practical applications (Liu et al., 2022).

Multi-task learning, the technique of using a single model to learn multiple tasks concurrently, has shown promise in terms of data efficiency and improved generalization (Caruana, 1997). However, recent studies indicate that there is also added difficulty in jointly learning multiple tasks. While some tasks may benefit from being learned together, thereby boosting accuracy, others may interfere with each other, leading to worse performance (Standley et al., 2020).

Our aim was to create a data-efficient model with improved run-time efficiency compared to existing models for planar reconstruction with semantics. We achieve the desired outcome via our model, SOLO-Planes (SOLOP), where we make use of multi-view guidance for improved data usage when acceptable ground truth plane segments differ across views, and made adjustments to the base architecture for improved efficiency. Multi-view warping is done in feature space, by warping plane features from neighbour to source view, decoding, then transforming the decoded plane parameters to the source view camera view for comparison with ground truth data during training. This additional warping guidance for plane features positively impacts the learning of segmentation masks, particularly when using a more limited dataset, while only requiring a single view at inference time.

In the context of our work, we found that multi-view guidance using plane features leads to a notable improvement in segmentation results. We attribute this enhancement to our multi-task architecture and the use of a shared trunk, meaning a global feature extractor that is common to all tasks (Crawshaw, 2020). This architecture allows for loss propagation through shared features and common base networks, and may be particularly relevant in the case of incomplete or varying data across overlapping views.

Our contributions include the following:

1. An empirical demonstration of cross-task improvement using multi-view guidance by feature warping, with particular relevance in cases where ground truth data may be incomplete across neighboring views.
2. A single-image planar reconstruction model, that can concurrently predict semantics for planar segments while achieving the best efficiency compared to other known planar reconstruction methods at a processing speed of 43 FPS.

Our approach may be a helpful method for other multi-task models limited in some forms of ground truth training data. The efficiency of the model makes it suitable for a range of real-world applications.

2 RELATED WORK

Planar Reconstruction. Early works in planar reconstruction using a single image predicted a set number of planes per scene without using ground truth plane annotations by employing a plane structure-induced loss (Yang and Zhou, 2018). Another early end-to-end planar reconstruction network from a single image is PlaneNet, which uses separate branches for plane parameter, mask, and non-planar area depth estimation (Liu et al., 2018a). Two major subsequent models serve as the foundation for several later works. PlaneRCNN is an extension of the two-stage instance segmentation model, Mask-RCNN (He et al., 2017), and predicts the plane instance normal and depth map, then jointly process plane parameters along with segmentation masks through a refinement module (Liu et al., 2019). PlaneAE predicts per-pixel parameters and associative embeddings and employs efficient mean clustering to group the pixel embeddings to plane instances (Yu et al., 2019). PlaneTR uses geometric guidance by generating and tokenizing line segments, giving the input additional structural information (Tan et al., 2021). Additional contributions in this area include post-processing refinement networks that enforce interplane relationships, via predicting the contact line or geometric relations between adjacent planes (Qian and Furukawa, 2020). More recent works follow the approach of using an instance segmentation model base. PlaneSegNet is based on a real-time instance segmentation architecture and introduces an efficient Non-Maximum Suppression (NMS) technique to reduce redundant proposals (Xie et al., 2021a). PlaneRecNet predicts per-pixel depth and plane segmentation masks, then use classical methods like PCA or RANSAC to recover plane parameters (Xie et al., 2021b). A number of single-image plane reconstruction models use some form of instance segmentation baseline. However, most single-image models focus solely on spatial parameters and largely ignore the task of recovering semantics.

Multi-View Approaches. The task of predicting 3D plane parameters from a single image is inherently ambiguous and challenging. Thus, several works have incorporated multi-view information, either as a loss guidance or by using multiple image inputs at inference time. PlanarRecon (Xie et al., 2022) is a real-time model using multiple image frames which makes predictions directly in 3D by using a coarse-to-fine approach for building a sparse feature volume, then clustering occupied voxels for instance planes, and uses a tracking and fusion module to get a global plane representation. PlaneMVS (Liu et al., 2022) is the first to apply a deep multi-view stereo (MVS) approach to plane parameters. Although it achieves state-of-the-art results and also predicts class semantics, it is less computationally efficient due to the use of 3D convolutions and requires generation of plane hypotheses. PlaneRCNN incorporates a multi-view warping loss module that enforces consistency with nearby views by projecting the predictions to 3D and calculating the distance after transforming to the same camera coordinates (Liu et al., 2019). Unlike our approach, their warping module is applied directly on the predictions rather than in feature space. Another work enhances the PlaneAE model with multi-view regularization by warping the plane embedding feature maps and using associative embeddings from multiple views to improve the final instance segmentation (Xi and Chen, 2019).

Feature Warping. Feature warping is commonly done in deep Multi-View Stereo (MVS) approaches, as it was found that creating the cost volume using features is as effective for artificial neural networks and more computationally efficient due to reduced size (Im et al., 2019; Yao et al., 2018). While some approaches use a similarity function on the features, others simply concatenate the warped feature with the original and let the model learn the relation rather than calculate an explicit cost volume (Chen et al., 2020; Yao et al., 2018). The latter approach is used by PlaneMVS to construct a Feature/Cost volume, which is then processed by a 3D CNN to get the plane parameters. Deep MVS methods are more commonly used for depth estimation, and their application to plane parameter estimation is relatively novel. Other research suggests that calculating a feature error between frames is more robust than a photometric error (Guo et al., 2021). However, this cannot directly be applied to plane reconstruction, as the plane features contain information in different camera views when considering a video dataset. Takanori et al. use multi-frame attention via feature warping for the task of drone crowd tracking (Asanomi et al., 2023). Ding et al. take MVS as a feature matching task and use a Transformer model to aggregate long-range global context using warped feature maps (Ding et al., 2022).

In order to ensure differentiability, the warp to another view using depth values and camera parameters must be backprojected using bilinear interpolation. Most existing works involving feature warping do not specifically deal with plane features, which require transformation to the correct view when decoded. Additionally, the majority of planar reconstruction models do not offer semantic predictions for the scene.

The majority of existing works primarily focus on the geometric accuracy of planes without holistically addressing the more practical requirements of speed and semantic understanding of planar scenes. Our work aims to fill this gap by offering a unified framework for semantic planar reconstruction. We improve data efficiency during training and achieve cross-task improvement using multi-view guidance for plane features, while maintaining an inference speed that is suitable for real-time applications.

3 METHOD

Our objective is to develop a real-time framework for the task of 3D semantic planar reconstruction. This section is organized as follows: Section 3.1 provides details of the framework, Section 3.2 elaborates on the loss terms, and Section 3.3 introduces our multi-view guidance.

3.1 Framework

Our framework is built on a light version of the SOLOv2 instance segmentation model (Wang et al., 2020b) using a ResNet-50 backbone (He et al., 2016). SOLOv2 is a single-stage instance segmentation model that predicts instance masks and labels based on their spatial location. It achieves an execution speed of around 31 FPS using a single V100 GPU card (Wang et al., 2020b). The model employs dynamic convolution to generate the final segmentation mask, leveraging multi-scale features from the Feature Pyramid Network (FPN) (Lin et al., 2017a). Each level of the FPN output features are used to predict mask kernels and class semantics, with the features reshaped to square grids of varying sizes, with each responsible for predictions at a different scale. Each grid location predicts a kernel and semantic category scores. The mask feature is obtained through feature fusion of the first four levels of the FPN outputs via bilinear upsampling and convolution layers, and the final segmentation masks are obtained via convolution using the predicted kernels, with redundant masks suppressed using matrix Non-Maximum Supression (NMS) (Wang et al., 2020b). The mask and kernel features receive spatial awareness information from concatenated normalized coordinate, a method from Liu et al. (2018b).

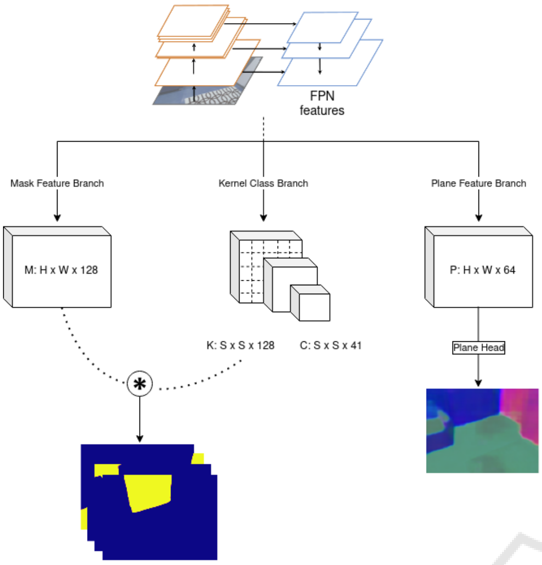

Figure 2: Simplified overview of SOLOPlanes architecture.

We extend the base architecture by introducing a plane feature branch that fuses the first two levels of the feature map, along with a plane prediction head that outputs per-pixel plane parameters via a convolution layer (see Fig. 2). This prediction is supervised by a set of loss functions that leverage geometrical constraints and ground truth depth information (detailed in Section 3.2). The original architecture predicts the kernels and semantic categories using all five feature map levels of the FPN. Based on the findings of Chen et al. (2021), a divide-and-conquer approach is more crucial than leveraging multi-scale features for task-specific predictions, we experimented with using different feature levels and found that using fewer feature levels not only maintained comparable performance in multi-task planar segmentation but also improved the overall efficiency of the model.

Our final architecture takes a single RGB image, $I \in \mathbb{R}^{H\times W\times C}$, as input during inference, and outputs an arbitrary number of plane instance masks along with instance level semantics and per-pixel plane parameters. We obtain the final result by pooling per-pixel parameter prediction using the predicted masks, and retaining per-pixel predictions in areas without a plane instance. The model is trained on the large-scale public ScanNet dataset containing indoor scenes from Dai et al. (2017), supplemented with ground truth plane annotations from Liu et al. (2019).

3.2 Losses

Mask & Category. We retain the original loss functions from Wang et al. (2020b) for mask and category predictions. The Dice Loss, L_M, guides mask prediction with the original loss weight $w_M = 3$, and focal loss, L_C, for semantic category prediction (Lin et al., 2017b). For full details, we refer readers to (Wang et al., 2020a). In order to address class imbalances due to dominating negative samples, we modified L_C to only consider grid locations containing an instance.

Plane Parameters. Plane parameters are represented by the normal and offset of the plane, denoted as $\mathbf{p} = (\mathbf{n}, d)$, which we combine into a single parameter $\mathbf{p} = \mathbf{n} * d \in \mathbb{R}^3$, with $\mathbf{n}$ normalized to unit length. Due to the complexity of predicting plane parameters containing both normal and depth information, we employ multiple loss functions for supervising per-pixel plane predictions. We use L_1 loss for direct comparison with ground truth plane parameters:

$$L_{plane} = \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{p}_i - \mathbf{p}_i^*\|. \qquad (1)$$

An asterisk is used to denote predicted values, and N represents the total number of pixels. The cosine distance, denoted $L_{surface}$, is used to guide the learning of surface normals. Due to the way we represent plane parameter $\mathbf{p}$, we get the equivalent result calculating cosine similarity on plane parameters directly.

$$sim_i = \frac{\mathbf{p}_i \cdot \mathbf{p}_i^*}{\|\mathbf{p}_i\|\|\mathbf{p}_i^*\|}, \qquad L_{surface} = \frac{1}{N}\sum_{i=1}^{N} 1 - sim_i \quad (2)$$

Due to noisy and incomplete ground truth plane annotations, we also make use of ground truth depth data, $D \in \mathbb{R}^{H\times W}$, for additional supervision. We calculate the plane induced depth at pixel location i by

$$D_i^* = \frac{d_i^*}{\mathbf{n}_i^{*T} \cdot K^{-1}\mathbf{q}_i}, \qquad (3)$$

where K represents the ground truth camera intrinsics of the scene and q_i is the x and y index for pixel location i. The plane induced depth loss, L_{depth}, is formulated as:

$$L_{depth} = \frac{1}{N}\sum_{i=1}^{N} |D_i - D_i^*|. \qquad (4)$$

We use the plane structure induced loss, first introduced by (Yang and Zhou, 2018) and which we denote by L_{geom}, based on the principle that the dot product of a 3D point on a plane with the normal equals the offset, $n^T Q = d$. We use ground truth depth and camera intrinsics to retrieve the 3D point at each pixel location. $Q_i = D_i K^{-1}\mathbf{q}_i$ obtains the 3D point projected at one location.

$$L_{geom} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{n}_i^{*T} \cdot Q_i - d_i^* \tag{5}$$

Gradient Weighting. We add gradient edge weighting as a model variation, weighting L_{depth} and L_{geom} to emphasize learning at edges, areas which are typically more difficult to learn. We choose to use the gradient of the image, $G \in \mathbb{R}^{H\times W}$ rather than depth, in order to better capture edges. Despite more noise at non-edge areas, it can capture more plane edges as some plane instances can have the same depth but still represent different surfaces (*e.g.* picture frame on a wall). This addition results in cross-task improvements for segmentation mask prediction in the case of the multi-view model (see Section 3.3).

$$L_{depth,geom} = \frac{1}{N}\sum_{i=1}^{N} G_i * L_i \tag{6}$$

The total loss for plane guidance is

$$L_P = L_{plane} + L_{surface} + L_{geom} + L_{depth}, \tag{7}$$

and the final combined losses:

$$L_{total} = L_M * w_M + L_C + L_P. \tag{8}$$

3.3 Multiview Plane Feature Guidance

In this section, we introduce our multi-view guidance approach, depicted in Fig 3. We take neighbouring image pairs, which we denote by source and neighbouring view (I_s, I_n), and extract the corresponding 2D features. The two finest pyramid feature maps are fused to generate plane features $f \in \mathbb{R}^{\frac{1}{4}H\times\frac{1}{4}W\times C}$. We backproject the neighbouring feature f_N to the corresponding location of the source view using bilinear interpolation. This process uses the ground truth depth, intrinsic parameters, and the relative transform between the views to obtain the warped 2D coordinates, from which we obtain the out-projection mask. We then decode the warped neighbouring feature $\hat{f}_N$ with the plane prediction head to get the corresponding plane parameters. It is important to note that $\hat{f}_N$ contains plane information of the neighbouring view, under the camera coordinates of I_n. Therefore, we transform the decoded plane parameters to the source view's camera coordinates before comparing to ground truth. This transformation is given by:

$$\hat{\mathbf{n}}_\mathbf{s} = \mathrm{R}\,\mathbf{n}_\mathbf{n}, \qquad \hat{d}_s = d_n + \mathbf{n}_\mathbf{n}^T \cdot \mathbf{t}, \tag{9}$$

where $(R, \mathbf{t})$ represents the rotation matrix and translation vector from neighbour to source view, and $(\mathbf{n}_\mathbf{n}, d_n)$ are the normal and offset in the neighbouring view. We then calculate an additional plane loss L_P using the transformed plane parameters decoded from the warped feature, excluding from the loss areas that are occluded or fall outside of the 2D image coordinates using the out-projection mask.

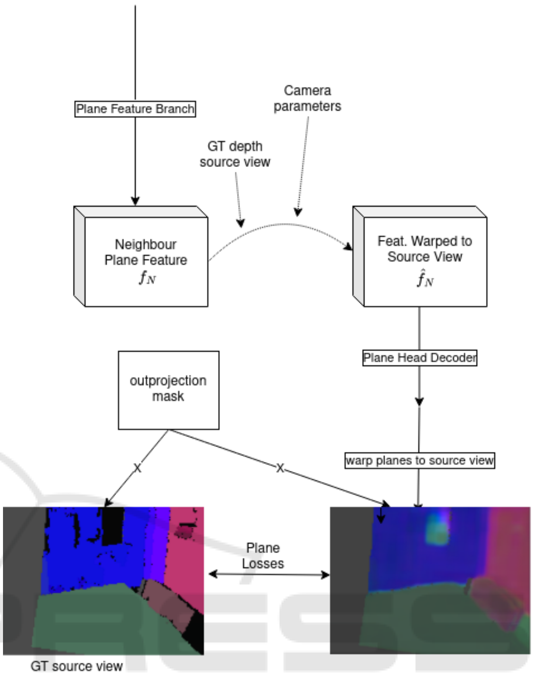

Figure 3: Overview of the feature warping guidance.

3.4 Instance Plane Soft-Pooling

To obtain the final instance level plane parameters, we use a soft-pooling technique which only considers per-pixel parameters within the area of the predicted instance. We found that restricting the pooling to this binary area yields better results compared to using soft-pooling across all pixel locations. We opted to not use an instance level plane loss as it negatively impacts the learning of mask segmentation. We generate a binary segmentation mask by applying a threshold to the predicted soft mask, denoted as $m^* \in [0,1]$. The instance level parameter can be retrieved by

$$\mathbf{p}_{ins} = \frac{\sum_{i=1}^{M} m_i^* \, \mathbf{p}_i^*}{\sum_{i=1}^{M} m_i^*}, \tag{10}$$

where M represents all the pixels falling within the region indicated by the binary segmentation mask, and $\mathbf{p}_i^*$ the predicted plane parameter at the corresponding location.

4 EXPERIMENTS

We evaluate various configurations of our model as well as comparison models. The nomenclature for our model versions is as follows: SOLOP-5lvls is a single view version using the original 5 feature levels for prediction, SOLOP-SV refers to the single-view model trained on 60,000 samples, SOLOP-MV is the multi-view model trained on 30,000 pairs, and SOLOP-MV-gw incorporates gradient edge loss weighting into the multi-view model. Qualitative results are obtained using the last configuration, as it achieved the best performance.

4.1 Setup & Training Details

For comparison between different model versions, we train a base model initialized with a pretrained ResNet-50 backbone and employ a data augmentation scheme where each sample has a 15% chance of undergoing one of several augmentations, such as a) jitter of brightness, contrast, hue, saturation, b) Planckian jitter (Zini et al., 2023), c) Gaussian noise, or d) motion blur. We use learning rate warm-up for the first 2000 steps starting from a learning rate of 1e−6 and increases until 2e−4. After the initial warm-up period, the learning rate is reduced by a factor of 0.1 given no improvement to the validation loss. For quicker and more fair comparison of model variations, a base model with the best validation loss was saved at epoch 9 and used as initialization to our main models, which were trained for 11 additional epochs. We employ early stopping if validation loss fails to improve for 5 consecutive epochs and save the model with best validation performance as well as the last checkpoint. For evaluation, we take the best of either saved model. The additional models trained using the base model initialization do not use data augmentation, and have 500 steps of learning rate warm-up starting from 1e−6 to 1e−5. We use a batch size of 32 for the single view model with gradient accumulation to mitigate the higher instability associated with multi-task learning. We train the models on a single NVIDIA Ampere A100 GPU. For evaluation and FPS calculation, we use a single NVIDIA GeForce RTX 3090 GPU for all models.

4.2 Dataset

For training and evaluation, we use the ScanNet dataset which contains RGB-D images from video sequences, totalling 2.5 million views complete with camera parameters and instance level semantics (Dai et al., 2017). The ground truth plane instance annotations for instance masks and plane parameters are generated by the authors of PlaneRCNN, and we follow the same process for filtering and preprocessing the planes (Liu et al., 2019). We also obtain the corresponding plane instance semantics from the metadata of the plane annotations. The ground truth plane data often exhibited issues such as over-segmented, rough edges, or missing plane instances, as planes with a depth error above a 0.1 meter threshold were omitted. For multi-view guidance training, we take sample pairs which are 10 time-steps away. In some cases, a neighbouring ground truth plane image might contain a segment which is missing in the source view, and vice versa. For the single-view model, we use 60,000 random samples from the training set and 10,000 from the validation set. For the multi-view model, we use 30,000 neighboring pairs for training and 5,000 pairs for validation.

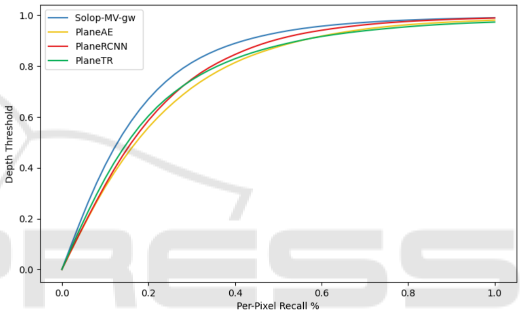

Figure 4: Per-pixel recall at varying depth thresholds in meters.

4.3 Comparison

Our model is most comparable to the PlaneAE model from Yu et al. (2019) and PlaneTR model from Tan et al. (2021), primarily due to the speed of prediction and the fact that they predict plane parameters directly using a single image as input. Given the inconsistent quality of ground truth plane data, the authors of PlaneMVS manually selected stereo pairs for the test set, which contained samples with more complete plane annotations (Liu et al., 2022). We run our evaluations using the same test set. For a fair comparison, we train the PlaneAE model for a total of 20 epochs using a ResNet-50 backbone and the same data with an input size of 480 x 640. The original model was trained using an input size of 192 x 256, resulting in a higher FPS. To align with our training regimen, we train PlaneAE for 11 epochs using 60,000 samples and an additional 9 epochs with 100,000 random samples. We retain the original training configuration of the authors (Yu et al., 2019). We use the same approach for retraining the PlaneTR model, and gen-

Table 1: Model comparison results on ScanNet dataset for variations of SOLOP model and other single-image planar reconstruction methods.

Method	**Depth Metrics**							**Detection Metrics**		**FPS**
	AbsRel↓	SqRel↓	RMSE↓	log_RMSE↓	$\delta < 1.25 \uparrow$	$\delta^2 < 1.25 \uparrow$	$\delta^3 < 1.25 \uparrow$	AP	mAP	
PlaneAE	0.181	0.092	0.325	0.208	0.746	0.931	0.983	-	-	17
PlaneTR	0.178	0.133	0.365	0.215	0.768	0.930	0.977	-	-	15
PlaneRCNN	0.165	0.070	0.278	0.187	0.780	0.954	0.991	0.193	-	7
SOLOP-5lvls*	0.143	0.059	0.276	0.185	0.813	0.960	0.990	0.416	0.314	38
SOLOP-SV*	0.134	**0.052**	**0.259**	0.178	0.832	**0.964**	0.991	0.389	0.267	**43**
SOLOP-MV*	0.136	0.054	0.261	**0.177**	0.832	0.962	0.991	0.427	0.344	**43**
SOLOP-MV-gw*	**0.133**	**0.052**	**0.259**	**0.177**	**0.833**	**0.964**	**0.992**	**0.434**	**0.347**	**43**

* = Ours

erate the required line segments using HAWPv3 (Xue et al., 2023). While PlaneRCNN also takes a single image at inference time, its slower inference speed makes it a less direct comparison. We run evaluations on the provided model from authors Liu et al. (2019).

4.4 Evaluation Metrics

We follow previous methods (Yu et al., 2019; Liu et al., 2019) and calculate the per-pixel depth recall at varying thresholds in meters, shown in Fig. 4. We also calculate standard depth and detection metrics for a comprehensive evaluation of model performance. Average Precision (AP) is used to assess the quality of the predicted masks, and Mean Average Precision (mAP) takes into account the semantic labels by averaging AP across class categories. For depth metrics, we use Absolute Relative Difference (AbsRel), Squared Relative Difference (SqRel), Root Mean Squared Error (RMSE), log RMSE, and delta accuracy (Eigen et al., 2014). We also calculate model efficiency using Frames Per Second (FPS). The results of these evaluations are summarized in Table 1, which shows a marked improvement using our architecture.

4.5 Results

The task of segmentation becomes more challenging when predicting multiple classes, as overlapping masks from different classes are less likely to be suppressed. The oversegmentation issue appears to be more pronounced in the single view model, whereas multi-view guidance using plane features helped to produce more complete and less oversegmented masks. This improvement is likely attributable to feature sharing and the correlation between ground truth plane instance masks and plane parameters.

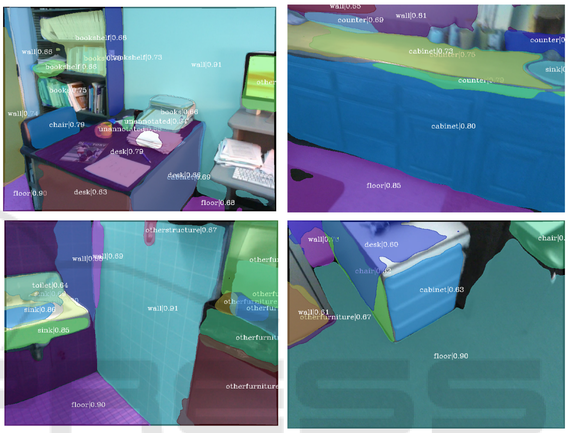

Figure 5: Visualization of semantic predictions using the SOLOP-MV-gw model.

Despite using multi-view guidance on plane predictions, we observe an objective improvement in prediction of segmentation masks. We hypothesize that this is especially effective when adjacent views have disparate ground truth data, such as in the case of missing annotations. This would explain the similar performance with regards to depth metrics between SOLOP-SV and multiview variants, as the ground truth depth is fairly stable across views. Ground truth mask completeness can differ across neighbouring views due to lower quality segments being filtered out. Even though the variants using multi-view guidance saw a lower diversity of scenes compared to the single view version, it nevertheless outperforms the single view variant on the task of mask segmentation.

Quantitative. All results are obtained using the selected test set chosen by the authors of PlaneMVS. The authors Liu et al. (2022) manually selected a higher quality set to evaluate on due to the incomplete and imprecise nature of the ground truth plane annotations. The resulting test set contains 949 image pairs. Our quantitative findings from model comparisons, summarized in Table 1, indicate that our multi-

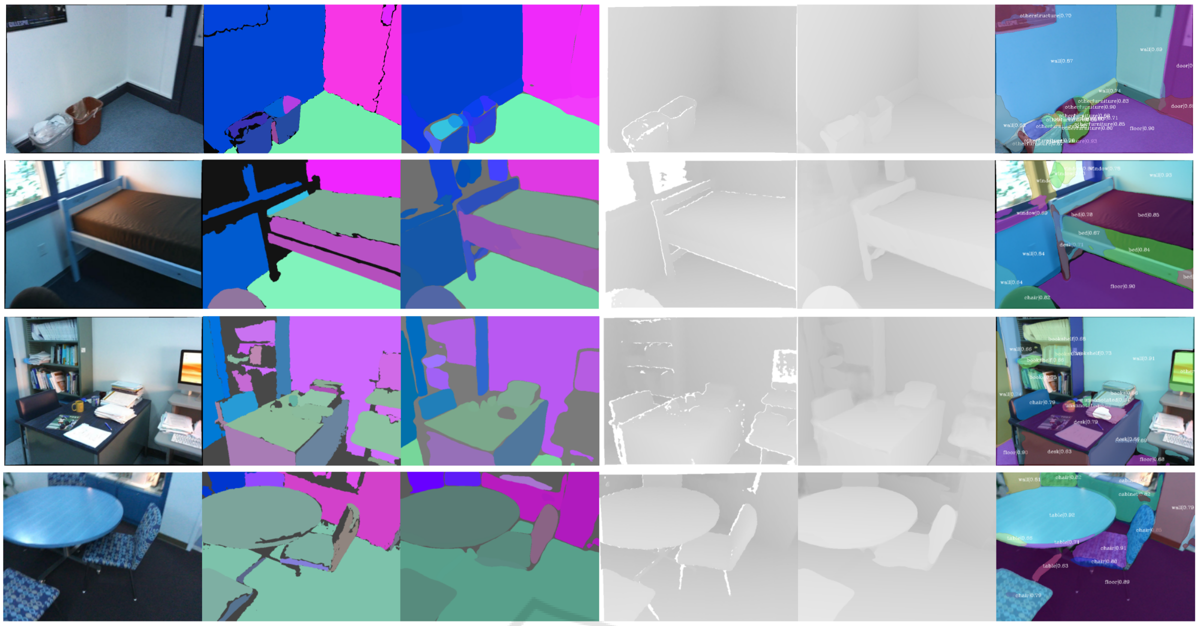

Figure 6: Qualitative results of instance plane and semantic prediction using model with best performance, SOLOP-MV-gw. From left to right: Input image, GT planes, predicted planes, GT depth, predicted depth, predicted semantics.

view model variant not only matches the performance of the single-view model in depth metrics, but also shows a significant improvement in detection metrics. This demonstrates the efficacy and improved data efficiency in using multi-view guidance via warping in feature space, at least in the case of using shared features for multitask learning. Since all SOLOP variants use a single image at inference time, the FPS result is the same for the versions of the model using 3 feature levels (SOLOP-SV, SOLOP-MV, SOLOP-MV-gw), but significantly reduced for the version with the original 5 level architecture (SOLOP-5lvls). SOLOP-MV-gw achieves better depth recall comparatively (see Fig. 4), while all SOLOP variants outperform the comparison models on standard metrics.

Qualitative. We display different types of visual results from our best model in Figures 1, 5, and 6. In contrast to previous works that predicted a binary plane indication, the incorporation of multi-class semantics introduces an added complexity. The change made to the focal loss for category predictions (see Section 3.2) leads to more confident scoring as well as a potential increase in false positives, which is already exacerbated in the case of multi-class predictions. However, we found that raising the score threshold for the final masks partially mitigated this issue. See Fig. 6 for visual results. The structure of the scene is easier to predict than the exact depth, a challenge presented when using a single image for inference. Sample visualizations of the semantic predictions can be found in Fig. 5. Cases of oversegmentation can occur due to prediction of different classes, or different plane orientation, as each mask represents a planar segment associated with a class label. Overall, our model demonstrates robust performance both visually and quantitatively for the task of planar reconstruction with semantic labels.

5 DISCUSSION AND FUTURE WORK

In this work, we introduce SOLOPlanes, a real-time semantic planar reconstruction network which shows cross-task improvement when using multi-view guidance in feature space. The task of predicting plane parameters from a single image is non-trivial, and the complexity is further compounded by multi-task learning. Despite these challenges, our model competes favorably with other, less efficient methods in planar reconstruction that do not offer semantic predictions. To the best of our knowledge, our model also outperforms all other planar reconstruction models in computational efficiency, measured using FPS. Our work advances semantic plane instance segmentation without sacrificing computational efficiency, striking a balance between efficiency and performance. We hope it will serve as an inspiration or stepping stone for further research geared towards applications with real-world impact.

REFERENCES

Asanomi, T., Nishimura, K., and Bise, R. (2023). Multi-frame attention with feature-level warping for drone crowd tracking. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1664–1673.

Caruana, R. (1997). Multitask learning. *Machine learning*, 28:41–75.

Chen, Q., Wang, Y., Yang, T., Zhang, X., Cheng, J., and Sun, J. (2021). You only look one-level feature. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13039–13048.

Chen, R., Han, S., Xu, J., and Su, H. (2020). Visibility-aware point-based multi-view stereo network. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3695–3708.

Coughlan, J. M. and Yuille, A. L. (2003). Manhattan World: Orientation and Outlier Detection by Bayesian Inference. *Neural Computation*, 15(5):1063–1088.

Crawshaw, M. (2020). Multi-task learning with deep neural networks: A survey. *ArXiv*, abs/2009.09796.

Dai, A., Chang, A. X., Savva, M., Halber, M., Funkhouser, T., and Nießner, M. (2017). Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*.

Ding, Y., Yuan, W., Zhu, Q., Zhang, H., Liu, X., Wang, Y., and Liu, X. (2022). Transmvsnet: Global context-aware multi-view stereo network with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8585–8594.

Eigen, D., Puhrsch, C., and Fergus, R. (2014). Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27.

Guo, E., Chen, Z., Zhou, Y., and Wu, D. O. (2021). Unsupervised learning of depth and camera pose with feature map warping. *Sensors*, 21(3).

He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). Mask r-cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Im, S., Jeon, H.-G., Lin, S., and Kweon, I.-S. (2019). Dpsnet: End-to-end deep plane sweep stereo. In *7th International Conference on Learning Representations, ICLR*.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. (2017a). Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 936–944.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017b). Focal loss for dense object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007.

Liu, C., Kim, K., Gu, J., Furukawa, Y., and Kautz, J. (2019). Planercnn: 3d plane detection and reconstruction from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Liu, C., Yang, J., Ceylan, D., Yumer, E., and Furukawa, Y. (2018a). Planenet: Piece-wise planar reconstruction from a single rgb image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Liu, J., Ji, P., Bansal, N., Cai, C., Yan, Q., Huang, X., and Xu, Y. (2022). Planemvs: 3d plane reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8665–8675.

Liu, R., Lehman, J., Molino, P., Petroski Such, F., Frank, E., Sergeev, A., and Yosinski, J. (2018b). An intriguing failing of convolutional neural networks and the coordconv solution. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Qian, Y. and Furukawa, Y. (2020). Learning pairwise inter-plane relations for piecewise planar reconstruction. In *European Conference on Computer Vision*.

Standley, T., Zamir, A., Chen, D., Guibas, L., Malik, J., and Savarese, S. (2020). Which tasks should be learned together in multi-task learning? In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9120–9132. PMLR.

Tan, B., Xue, N., Bai, S., Wu, T., and Xia, G.-S. (2021). Planetr: Structure-guided transformers for 3d plane recovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4186–4195.

Wang, X., Kong, T., Shen, C., Jiang, Y., and Li, L. (2020a). Solo: Segmenting objects by locations. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 649–665. Springer.

Wang, X., Zhang, R., Kong, T., Li, L., and Shen, C. (2020b). Solov2: Dynamic and fast instance segmentation. *Advances in Neural information processing systems*, 33:17721–17732.

Xi, W. and Chen, X. (2019). Reconstructing piecewise planar scenes with multi-view regularization. *Computational Visual Media*, 5(4):337–345.

Xie, Y., Gadelha, M., Yang, F., Zhou, X., and Jiang, H. (2022). Planarrecon: Real-time 3d plane detection and reconstruction from posed monocular videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6219–6228.

Xie, Y., Rambach, J., Shu, F., and Stricker, D. (2021a). Planesegnet: Fast and robust plane estimation using a single-stage instance segmentation cnn. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13574–13580. IEEE.

Xie, Y., Shu, F., Rambach, J. R., Pagani, A., and Stricker, D. (2021b). Planerecnet: Multi-task learning with cross-task consistency for piece-wise plane detection and reconstruction from a single rgb image. In *British Machine Vision Conference*.

Xue, N., Wu, T., Bai, S., Wang, F.-D., Xia, G.-S., Zhang, L., and Torr, P. H. S. (2023). Holistically-attracted wireframe parsing: From supervised to self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12):14727–14744.

Yang, F. and Zhou, Z. (2018). Recovering 3d planes from a single image via convolutional neural networks. In Ferrari, V., Hebert, M., Sminchisescu, C., and Weiss, Y., editors, *Computer Vision – ECCV 2018*, pages 87–103, Cham. Springer International Publishing.

Yao, Y., Luo, Z., Li, S., Fang, T., and Quan, L. (2018). Mvsnet: Depth inference for unstructured multi-view stereo. *European Conference on Computer Vision (ECCV)*.

Yu, Z., Zheng, J., Lian, D., Zhou, Z., and Gao, S. (2019). Single-image piece-wise planar 3d reconstruction via associative embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1029–1037.

Zini, S., Gomez-Villa, A., Buzzelli, M., Twardowski, B., Bagdanov, A. D., and van de Weijer, J. (2023). Planckian jitter: countering the color-crippling effects of color jitter on self-supervised training. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.