

PlaneRecTR: Unified Query Learning for 3D Plane Recovery from a Single View

Jingjia Shi Shuaifeng Zhi* Kai Xu*
 National University of Defense Technology, China

Abstract

3D plane recovery from a single image can usually be divided into several subtasks of plane detection, segmentation, parameter estimation and possibly depth estimation. Previous works tend to solve it by either extending the RCNN-based segmentation network or the dense pixel embedding-based clustering framework. However, none of them tried to integrate above related subtasks into a unified framework but treated them separately and sequentially, which we suspect is potentially a main source of performance limitation for existing approaches. Motivated by this finding and the success of query-based learning in enriching reasoning among semantic entities, in this paper, we propose PlaneRecTR, a Transformer-based architecture, which for the first time unifies all subtasks related to single-view plane recovery with a single compact model. Extensive quantitative and qualitative experiments demonstrate that our proposed unified learning achieves mutual benefits across subtasks, obtaining a new state-of-the-art performance on public ScanNet and NYUv2-Plane datasets. Codes are available at <https://github.com/SJingjia/PlaneRecTR>.

1. Introduction

Immersing in a virtual 3D world involves efficient reasoning about surrounding scenes, whose properties need to be frequently updated. Though tremendous efforts have been paid to create authentic 3D geometry from multi-view image observations or even a single image, a trade-off exists between the reconstruction quality and efficiency, depending on the underlying scene representations [7, 21, 37, 38, 39]. 3D maps composed of sparse primitives such as point clouds are light-weight to maintain but lack topological structures, while dense geometry like volumetric grids and meshes are computational intensive to acquire and maintain. In this spectrum, planar representation has been proven to be a reliable alternative, which is compact, efficient, expressive and generalizable enough to be deployed ubiquitously as well. Therefore, in practice it would always

¹Shuaifeng Zhi and Kai Xu are corresponding authors.

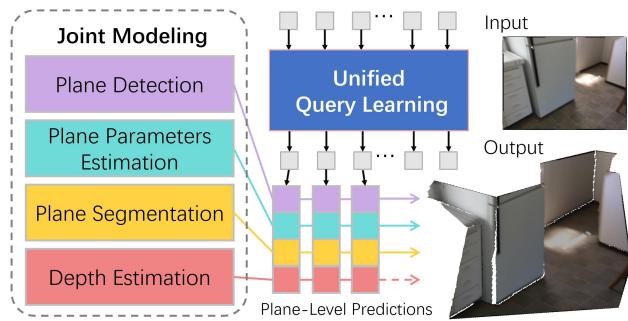


Figure 1. Illustration of PlaneRecTR. Our core idea is to use unified query learning to jointly model all subtasks of single-view plane recovery and output plane-level predictions, which achieves mutual benefits among tasks and enables higher performance.

be ideal to infer planar information purely from a video sequence, or even a single image. Being a challenging and yet fundamentally ill-posed computer vision problem, single view plane recovery/reconstruction has been extensively researched and focused so far. There have been early attempts using image processing to extract low-level primitives, such as line segments and vanishing points, to extract planar structures from the input image [8, 15].

As Convolutional Neural Networks (CNNs) have become the major approach to tackle vision problems in the past few years, their excelling in performance gradually spreads to the plane estimation task. Pioneering works such as PlaneNet [17] and PlaneRCNN [16] propose CNN-based efficient solutions to piece-wise planar structure recovery from a single image in a top-down manner. There have also been bottom-up solutions such as PlaneAE [35] and PlaneTR [23] which obtain plane-level masks by a post-clustering procedure on top of learned deep pixel-wise embedding space. Recently, Transformers [25], as another major type of fundamental vision models, have made great progress on various vision tasks [26, 9, 33, 24, 28]. The success of vision Transformers does not merely comes from its realization of global/long-range interaction across images via attention mechanisms, another important aspect is the design of query-based set predictions, initially proposed in Detection Transformer (DETR) [3] to enable reasoning between detected instances and global context. The introduction of query learning technique to vision Transformers

has been further proven to be effective in several high-level vision tasks including instance and semantic segmentation [5, 4, 32], video panoptic segmentation [29, 36], etc. PlaneTR [23] is an early attempt of using such ideas in single-view plane reconstruction. Motivated by the structured-guided learning, PlaneTR integrate additional geometric cues like line segments into its training process, leading to state-of-the-art performance on the ScanNet dataset and unseen NYUv2-Plane dataset.

However, all of above mentioned single-view plane recovery methods somewhat dis-entangle the prediction of principle components required for plane reconstruction. Specifically, PlaneRCNN [16] learns to predict plane offsets from a monocular depth prediction branch while other attributes such as plane masks and normals are estimated separately from colour images. PlaneTR [23] also learns to predict a monocular depth map apart from the Transformer module, which is later used for acquiring plane segmentation masks via associate embedding clustering [35]. Under such cases, these closely related prediction subtasks are usually interleaved and we conjecture this could be one performance bottleneck for existing data-driven approaches.

Motivated by this finding, we seek to borrow recent advance in query-based learning and aim to design a single, compact and unified model to jointly learn all plane-related tasks. We expect such design would achieve a mutual benefits among tasks and improve the existing performance of single view plane reconstruction. Extensive experimental results on two public benchmark datasets show that *without using neighbouring views* [16] nor extra structural cues [23] during training, our unified querying learning model, named PlaneRecTR, achieves new state-of-the-art performance with a concise structure. Additionally, we have also found such framework can well take the advancements in fundamental vision models and is able to further increase its performance given stronger backbones.

To summarize, our contribution are as follows:

- We proposed a first single unified framework to address the single view plane reconstruction task, where all related subtasks including plane detection, plane parameter estimation, plane segmentation and depth prediction are jointly inferred in a multi-task manner motivated by query-based learning.
- Extensive numerical evaluation and visual comparisons demonstrate the benefits of our proposed unified query learning, which has been proven to take full advantage of plane semantic and geometric cues to achieve mutual benefits.
- We have achieved state-of-the-art single image plane recovery performance on the public ScanNet and NYUv2-Plane datasets, surpassing existing CNN-based and Transformer-based baselines. In addition, we have shown that our model could keep benefiting

from stronger backbone vision models.

2. Related Work

2.1. 3D Plane Recovery from a Single Image

Traditional methods for 3D plane recovery from a single image often rely on strong assumptions of scenes [2, 8, 10](e.g., the Manhattan world assumption), or require manual extraction of primitives [8, 10, 15], such as superpixels, vanishing points and line segments, which may not be applicable to actual scenes. In contrast, recent learning-based approaches have alleviated such limitations leveraging large-scale training corpus, which have witnessed exciting progress in single view plane recovery.

PlaneNet [17] is the first to propose an end-to-end learning framework for this task, it also generates a large dataset of planar depth maps by utilizing the ScanNet dataset [6]. PlaneRecover [34] presents an unsupervised learning approach that specifically targets outdoor scenes. Both PlaneNet and PlaneRecover can only predict a fixed number of planes. PlaneRCNN [16] tackles the limitation and extracts an arbitrary number of planes with their parameters and segmentation masks by using a proposal-based instance segmentation framework, i.e., Mask R-CNN [11]. It also proposes a segmentation refinement network and a warping loss between frames to improve performance. The proposal-based method requires multiple steps to successively tackle subtasks of 3D plane recovery including plane detection, plane segmentation, plane normal estimation, global depth estimation, plane offset calculation and plane refinement, therefore the overall pipeline are decoupled where planar attributes are usually inferred given predicted plane segmentation.

On the other hand, PlaneAE [35] leverages a proposal-free instance segmentation approach, which uses mean shift clustering to group embedding vectors within planar regions. PlaneTR [23] inherits the design of DETR [3] to detect possible plane instances and estimate plane parameters in parallel, followed by plane segmentation generated by pixel clustering like PlaneAE [35]. Specifically, its Transformer branch only predicts instance-level plane information, thus post-processing like clustering is still required to carry out pixel-wise segmentation. The global depth is inferred by another convolution branch.

Therefore, existing advanced methods, whether based on direct CNN prediction or embedding clustering, still divide the whole 3D plane recovery task into several stages.

2.2. Vision Transformers

Transformer [25] has recently become a popular fundamental vision model to solve computer vision problems [26, 9, 33, 24, 28]. Researchers have employed vision Transformers to develop new methods, e.g., DETR

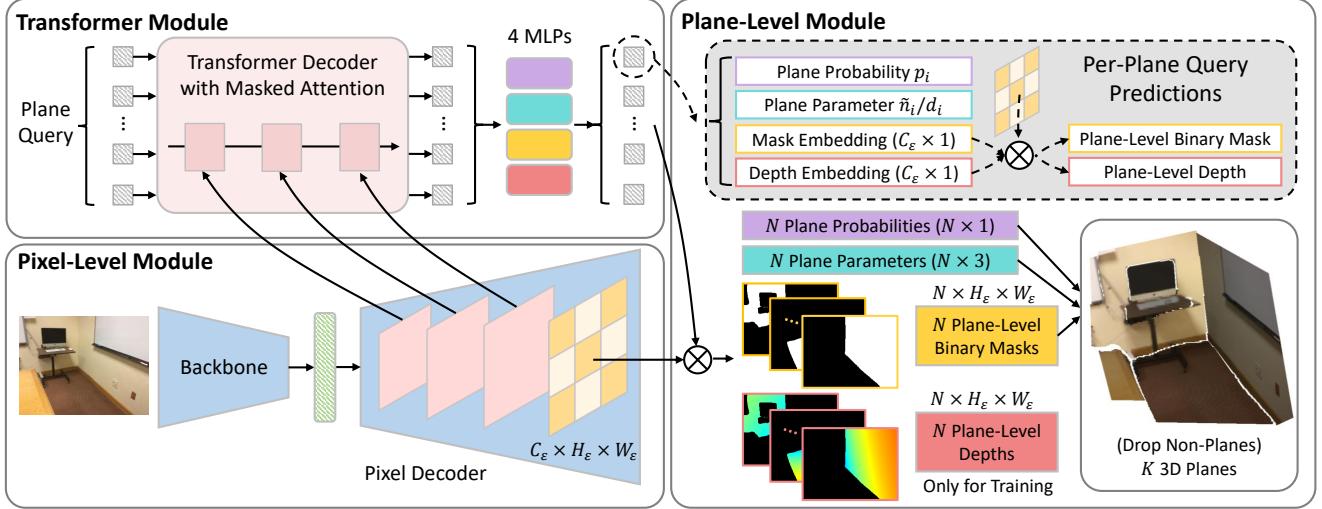


Figure 2. Overview of our proposed PlaneRecTR. Our model consists of three main modules: (1) A pixel-level module to extract dense pixel-wise image features; (2) A Transformer module to jointly predict 4 plane-related properties from each plane query, including plane classification probability, plane parameter, mask and depth embedding; (3) A plane-level module to calculate dense plane-level binary masks/depths, then filter non-plane predictions and produce the final 3D plane recovery.

[3], which treats object detection as a direct set prediction problem using a Transformer encoder-decoder architecture. DETR uses a fixed number of learned object queries to reason about object relations and global image context, enabling to output the final set of predictions in parallel. Building on the success of DETR, Maskformer [5] and Mask2former [4] have introduced a pixel-level decoder to solve both semantic-level and instance-level segmentation tasks in a unified manner. Besides, there have been other inspiring works [36, 14, 13, 31] to tackle different 3D vision tasks through vision Transformers. For example, Polyphonicformer [36] brings the merits of query-based reasoning into joint depth and panoptic prediction to solve Depth-aware Video Panoptic Segmentation (DVPS) task. MonoDTR [14] proposes a depth-aware Transformer network to integrates context and depth features for monocular 3D object detection. Our method is also a Transformer-based architecture and leverages query learning to unify all subtasks of plane recovery.

3. Method

In this section, we present architecture details and training configurations of PlaneRecTR. We start by first introducing the overall architecture of PlaneRecTR in Section 3.1, and then describe the inference process of recovering 3D planes in Section 3.2. Finally, we discuss the training process and its objective functions in Section 3.3.

3.1. Unified Query Learning for Plane Recovery

Inspired by the recent successes of DETR [3] and Mask2Former [4] in object detection and instance segmentation, we find that, it is not impossible to tackle the chal-

lenging monocular planar reconstruction task using a *single, compact* and *unified* framework, thanks to the merits of query-based reasoning to enable joint modeling of multiple tasks.

As shown in Figure 2, our PlaneRecTR consists of three main modules: (1) A pixel-level module to learn dense pixel-wise deep embedding of the input colour image. (2) A Transformer-based unified query learning module to jointly predict four target properties for each of N learnable plane queries, including plane classification probability p_i , plane parameter n_i , mask embedding, and depth embedding ($i \in [1, 2, \dots, N]$). Specifically, p_i is the probability to judge whether the i^{th} query corresponds to a plane or not; $n_i = \tilde{n}_i/d_i \in \mathbb{R}^3$, where $\tilde{n}_i \in \mathbb{R}^3$ is its plane normal and d_i is the distance from the i^{th} plane to camera center, i.e., offset. (3) A plane-level module to generate plane-level mask m_i and plane-level depth d_i through mask and depth embedding ($i \in [1, 2, \dots, N]$). And then we remove non-plane query hypothesis while combining the remaining ones for final image-wise plane recovery. These three modules will be described in details below.

Pixel-Level Module. Given an image of size $H \times W$ as input, we use the pre-trained ResNet-50 [12] as the backbone model to extract image feature maps. Similar to Mask2Former, a multi-scale convolutional pixel decoder is used to produce a set of dense feature maps with four scales, denoted as follows:

$$\mathbb{F} = \{F_1 \in \mathbb{R}^{C_1 \times H/32 \times W/32}, F_2 \in \mathbb{R}^{C_2 \times H/16 \times W/16}, F_3 \in \mathbb{R}^{C_3 \times H/8 \times W/8}, \mathcal{E}_{\text{pixel}} \in \mathbb{R}^{C_{\text{pixel}} \times H_{\text{pixel}} \times W_{\text{pixel}}}\}. \quad (1)$$

Among these, the first three feature maps $\{F_1, F_2, F_3\}$

are fed to the Transformer module, while the last one $\mathcal{E}_{\text{pixel}}$, a dense per-pixel embedding of resolution $H_{\mathcal{E}} = H/4$ and $W_{\mathcal{E}} = W/4$, is used for computing plane-level binary masks and plane-level depths.

Transformer Module. We use the Transformer decoder with masked attention proposed in [4], which computes N plane-level embeddings from above mentioned multi-scale feature maps $\{F_1, F_2, F_3\}$ and N learnable plane queries. The predicted embeddings are then independently projected to four target properties by four different multi-layer perceptrons (MLPs). Overall, the Transformer module predicts required planar attributes through N plane queries (upper left part in Figure 2).

Plane-Level Module. As shown in upper right part of Figure 2, we obtain a dense plane-level binary mask $m_i \in [0, 1]^{H_{\mathcal{E}} \times W_{\mathcal{E}}}$ /depth prediction $d_i \in \mathbb{R}^{H_{\mathcal{E}} \times W_{\mathcal{E}}}$ by a dot product between the i^{th} mask/depth embedding and the dense per-pixel embedding $\mathcal{E}_{\text{pixel}}$ from previous two modules, respectively.

Please note that we have also investigated the idea of learning two individual pixel decoders for semantics and depth specific dense embeddings to predict planar masks and depths separately. However, we did not observe a clear performance improvement, and therefore stick to current set-up with a shared $\mathcal{E}_{\text{pixel}}$ for sake of efficiency.

We finally obtain N plane-level predictions $\{y_i = (p_i, n_i, m_i, d_i)\}_{i=1}^N$ by the plane-level module, and each plane-level prediction contains all the information needed to recover a possible 3D plane.

3.2. 3D Plane Recovery during Inference

In this section, we briefly discuss how to recover 3D planes from our plane-level predictions. For the N plane-level predictions $\{y_i\}_{i=1}^N$ predicted by the network, we first drop non-plane components according to the plane classification probability p_i , leading to a subset of K planes (*i.e.*, $K \leq N$). For each pixel in the planar regions, we calculate $\arg \max_i \{m_i\}_{i=1}^K$ to obtain the most likely plane index as the final global segmentation mask.

In this simple and efficient manner, we can use the network predictions to create 3D plane reconstruction of the input frame. Note that plane-level depths are not involved during inference and we use plane parameters and segmentation to infer planar depths. We experimentally found that this design also leads to more structural and smooth geometric predictions than that relying on direct depth predictions.

3.3. Network Training

Plane-level Depth Training. Previous methods tend to predict a global depth by separate network branches to calculate plane offset [16] or use depth as additional cues to formulate segmentation masks [35, 23]. In contrast, our

method tries to achieve mutual benefits between planar semantic and geometric reasoning, we leverage joint query learning to unify all components of plane recovery in a concise multi-task manner. As a result, we explicitly predict dense plane-level depths, binary-masks, plane probabilities and parameters from a shared feature space, which is produced and refined via attention mechanism of the Transformer.

Bipartite Matching. During training, one important step is to build optimal correspondences between N predicted planes and M ground truth planes. Following bipartite matching of [4, 23], we search for a permutation $\hat{\sigma}$ by minimizing a matching defined cost function D :

$$\hat{\sigma} = \arg \min_{\sigma} \sum_{i=1}^M D(\hat{y}_i, y_{\sigma(i)}) , \quad (2)$$

$$\begin{aligned} D = & -\omega_1 p_{\sigma(i)}(\hat{p}_i) + \mathbb{1}_{\{\hat{p}_i=1\}} \omega_2 L_1(\hat{n}_i, n_{\sigma(i)}) \\ & + \mathbb{1}_{\{\hat{p}_i=1\}} \omega_3 L_1(\hat{d}_i, d_{\sigma(i)} \hat{m}_i) \\ & + \mathbb{1}_{\{\hat{p}_i=1\}} \omega_4 L_{ce} \\ & + \mathbb{1}_{\{\hat{p}_i=1\}} \omega_5 L_{dice}, \end{aligned} \quad (3)$$

where $\hat{y}_i = (\hat{p}_i, \hat{n}_i, \hat{m}_i, \hat{d}_i)$ are the i^{th} ground-truth plane attributes; $\sigma(i)$ indicates the matched index of the predict planes to the ground truth \hat{y}_i ; $\omega_1, \omega_2, \omega_3, \omega_4$ and ω_5 are weighting terms and set to 2, 1, 2, 5, 5, respectively. Here we additional consider the influence of mask and depth quality using a mask binary cross-entropy loss L_{ce} [4], a mask dice loss L_{dice} [20] and an L_1 depth loss, respectively.

Loss Functions. After bipartite matching, the final training objectives, *i.e.*, the total loss L is formulated by four parts as follows:

$$\mathcal{L} = \sum_{i=1}^M \left(\lambda \mathcal{L}_{\text{cls}}^{(i)} + \mathcal{L}_{\text{param}}^{(i)} + \mathcal{L}_{\text{mask}}^{(i)} + \lambda \mathcal{L}_{\text{depth}}^{(i)} \right), \quad (4)$$

where λ is a weight to balance the magnitude for loss terms and is set to 2 in this paper. \mathcal{L}_{cls} and $\mathcal{L}_{\text{param}}$ are a plane classification loss and a plane parameter loss, in a similar form to previous work [23].

Different from [23], the left two loss terms in our paper $\mathcal{L}_{\text{mask}}$ and $\mathcal{L}_{\text{depth}}$ are designed to explicitly learn dense planar masks and depths. Specifically, we have plane segmentation mask prediction loss as a combination of cross-entropy loss and dice loss:

$$\begin{aligned} \mathcal{L}_{\text{mask}}^{(i)} = & \mathbb{1}_{\{\hat{p}_i=1\}} \beta_1 L_{ce} + \\ & \mathbb{1}_{\{\hat{p}_i=1\}} \beta_2 L_{dice}, \end{aligned} \quad (5)$$

where $\beta_1 = 5$ and $\beta_2 = 5$. The depth prediction loss is in a typical L_1 form:

$$\mathcal{L}_{\text{depth}}^{(i)} = \mathbb{1}_{\{\hat{p}_i=1\}} L_1(\hat{d}_i, d_{\sigma(i)} \hat{m}_i). \quad (6)$$

4. Experiments

In this section, we conduct experiments to evaluate the performance of the proposed PlaneRecTR on two public datasets. Then, we ablate structural designs of PlaneRecTR to demonstrate that multiple subtasks of plane recovery can benefit each other through unified query learning.

Datasets. We train and evaluate PlaneRecTR on two popular benchmarks: ScanNet [6] and NYUv2-Plane [22].

ScanNet dataset is a large-scale RGB-D video collection of 1,513 indoor scenes. We use piece-wise planar ground-truth generated by PlaneNet [17], which contains 50,000 training and 760 testing images with resolution 256×192 .

NYUv2-Plane dataset is a planar variant of the original NYUv2 dataset [22] provided by [17]. It has 654 testing images with resolution 256×192 . Please note that NYUv2-Plane is adopted to highlight generalization capability of plane recovery methods without any further fine-tuning.

Evaluation Metrics. For the entire task of 3D plane recovery, we adopt per-plane and per-pixel recalls to evaluate our PlaneRecTR and baselines, following [23, 35]. The per-plane/pixel recall metric is defined as the percentage of the correctly predicted ground truth planes/pixels. A plane is considered being correctly predicted if its segmentation Intersection over Union (IoU), depth normal and plane offset errors satisfy pre-defined thresholds. Specifically, the segmentation threshold of IoU is set to 0.5, while the error thresholds of depth/surface normal vary from $0.05m/2.5^\circ$ to $0.6m/30^\circ$, with an increment of $0.05m/2.5^\circ$.

Additionally, for evaluating plane segmentation, we apply three popular metrics[1, 34]: variation of information (VI), rand index (RI) and segmentation covering (SC). For evaluating plane parameter, we obtained the best match by minimizing the L_1 cost between K predicted plane parameters and M ground truth plane parameters, and then separately compute the average errors of plane normal and offset. As to NYUv2-Plane dataset, depth accuracy are evaluated on structured planar regions to highlight plane recovery performance [17, 23].

Network Variants. We note a few important architectural differences w.r.t. model designs and learning tasks. To reflect the contributions of our key design choices, we consider (1) explicit plane-level binary mask prediction (M), (2) plane-level depth prediction (D), (3) plane parameter estimation (P), and thus denote network varieties in Table 1.

4.1. Implementation Detail

Our network is implemented with Detectron2 [30]. We train our network on the ScanNet training set with total of 34 epochs on a single NVIDIA TITAN V GPU. We use AdamW optimizer [19] with an initial learning rate of 1×10^{-4} and a weight decay of 0.05. The batch size is set to 16.

Name	Backbone	Task 'M'	Task 'P'	Task 'D'
PlaneRecTR	ResNet-50	✓	✓	✓
PlaneRecTR (-D)	ResNet-50	✓	✓	-
PlaneRecTR (-M-D)	ResNet-50	-	✓	-
PlaneRecTR (-P-D)	ResNet-50	✓	-	-
PlaneRecTR (HRNet-32)	HRNet-32 [27]	✓	✓	✓
PlaneRecTR (Swin-B)	Swin-B [18]	✓	✓	✓

Table 1. Our PlaneRecTR variants during experiments. ('-X' indicates the network is trained without task 'X')

4.2. Results on the ScanNet Dataset

Qualitative Results. In this section, we present qualitative results of our single-view plane reconstruction on a variety of unseen ScanNet scenes in both 2D and 3D domains. As can be seen in Figure 3, our PlaneRecTR is able to predict accurate plane segmentation masks and plane parameters, as well as reasonable 3D plane reconstructions.

We further show detailed visual comparisons to PlaneTR [23], up-to-date the strongest single-view plane recovery approach, in Figure 4. Although PlaneTR has integrated extra structural cues like line segments of scenes, it is exciting to see that our approach, mainly benefiting from the explicit joint modelling of plane geometry and segmentation, is able to predict crisper plane segmentation mask and discriminate planes with similar normals (columns (b) and (c) of Figure 4), with more complete and holistic structures.

Quantitative Results. In this section, we conduct extensive quantitative evaluations towards previous state-of-the-art learning-based plane recovery methods: PlaneNet [17], PlaneRCNN [16], PlaneAE [35] and PlaneTR [23].

Like [23], PlaneRCNN is shown here mainly as a reference because of its learning with a different ScanNet training set-up. We thus focus on numerical comparison towards other recent baselines. We use public implementation of PlaneTR [23] and its provided pre-trained weights to report corresponding performance.

In terms of segmentation accuracy, Table 2 shows that on challenging ScanNet we achieve a new state-of-the-art plane segmentation performance, outperforming leading PlaneTR with a relative large margin especially in the VI metric. To further demonstrate the flexibility of our method, we have shown improved versions of PlaneRecTR by replacing ResNet-50 backbone [12] with the same HRNet-32 backbone [27] as PlaneTR or a SwinTransformer-B model [18]. The performance gap widens and shows that our framework could benefit from ongoing research in developing more powerful fundamental vision models.

As to performance of entire plane recovery task, we display per-pixel and per-plane recalls of depth and plane normal on the ScanNet dataset, respectively (see Figure 5). With varying thresholds from 0.05 to 0.6 meters and from 2.5 to 30 degrees for depth and normal evaluations, our method consistently outperforms all baselines, indicating more precise predictions of plane parameters, segmentation

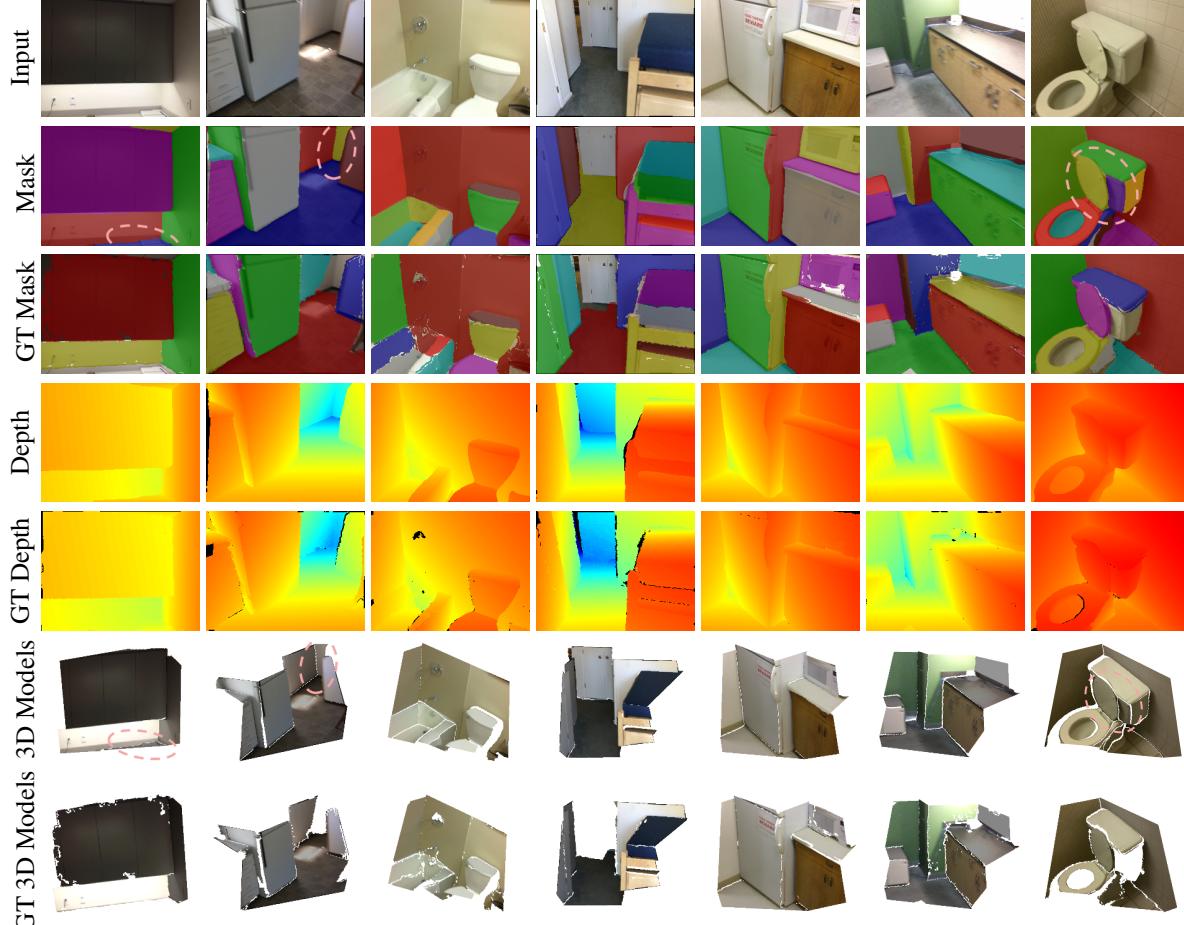


Figure 3. 3D plane recovery results of PlaneRecTR on the ScanNet dataset. See Appendix for 3D segmentaion visualization.

mask and planar depths. We want to further highlight the significant improvement w.r.t. per-plane recall, our methods could efficiently discover structural planes of various scales, in contrast to [23] which tends to miss planes with small areas or sharing similar geometry (see Figure 4).

We own the improvements to the explicit joint modelling of plane detection, plane parameter estimation, plane segmentation and depth prediction as a multi-task querying learning problem, and we will give detailed ablation studies in Section 4.4.

4.3. Results on the NYUv2-Plane Dataset

The NYUv2-Plane dataset is chosen here mainly to verify the generalization capability of our method on unseen novel indoor scenes. As shown in Table 2, our method still achieves leading plane segmentation accuracy in all metrics without any fine-tuning. Please check the bottom part of Figure 4 for more detailed visual comparisons. In Table 3, we focus on pixel-wise depth accuracy in planar regions, thus we treat PlaneTR [23] as the targeting baseline for a fair comparison and others as reference. Our base pipeline

Method	ScanNet			NYUv2-Plane		
	VI↓	RI↑	SC↑	VI↓	RI↑	SC↑
PlaneNet [17]	1.259	0.858	0.716	1.813	0.753	0.558
PlaneRCNN [16]	1.337	0.845	0.690	1.596	0.839	0.612
PlaneAE [35]	1.025	0.907	0.791	1.393	0.887	0.681
PlaneTR [23]	0.767	0.925	0.838	1.163	0.891	0.712
PlaneRecTR	0.698	0.936	0.854	1.130	0.905	0.722
PlaneRecTR (HRNet-32)	0.679	0.937	0.857	1.049	0.912	0.738
PlaneRecTR (Swin-B)	0.651	0.943	0.866	1.045	0.915	0.745

Table 2. Comparison of plane segmentation results on the ScanNet dataset and NYUv2-Plane dataset.

Method	Rel↓	\log_{10} ↓	RMSE↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
PlaneNet [17]	0.236	0.124	0.913	53.0	78.3	90.4
PlaneAE [35]	0.205	0.097	0.820	61.3	87.2	95.8
PlaneRCNN [16]	0.183	0.076	0.619	71.8	93.1	98.3
PlaneTR [23]	0.199	0.100	0.700	59.6	86.6	96.3
PlaneRecTR	0.202	0.100	0.729	59.5	87.4	96.0
PlaneRecTR (HRNet-32)	0.178	0.087	0.635	66.6	91.1	97.7
PlaneRecTR (Swin-B)	0.157	0.073	0.547	74.2	94.2	99.0

Table 3. Depth accuracy comparison on the NYUv2 dataset.

performs on-par with PlaneTR and outperforms PlaneNet [17] and PlanAE [35]. When our backbone is replaced from ResNet-50 [12] to the same HRNet-32 [27] as PlaneTR, our performance is significantly better than PlaneTR. PlaneR-

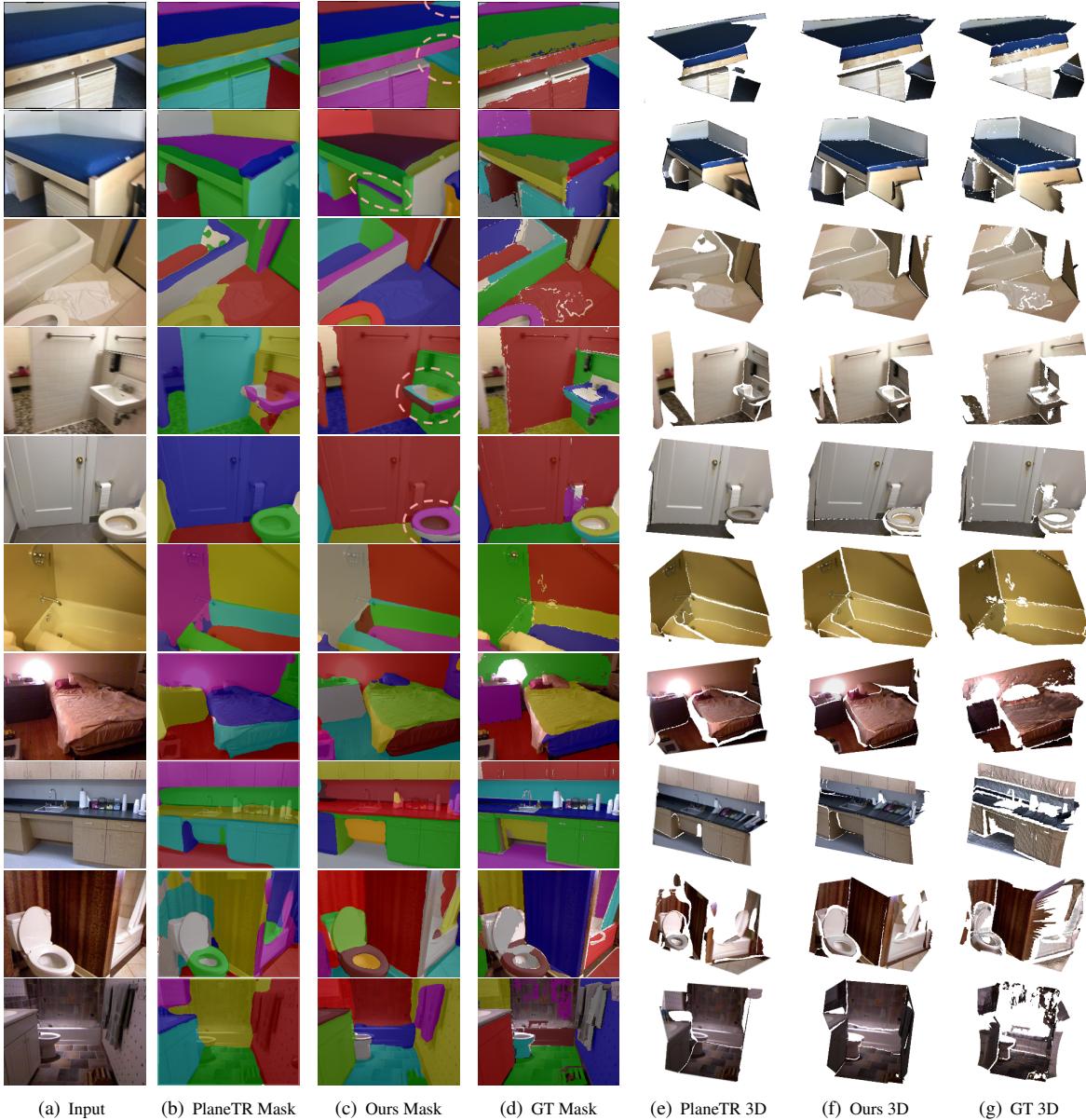


Figure 4. Comparison of plane reconstruction results on the ScanNet (top 6 rows) and NYUv2-Plane (bottom 4 rows) datasets. See Appendix for comparison of depth and 3D segmentation.

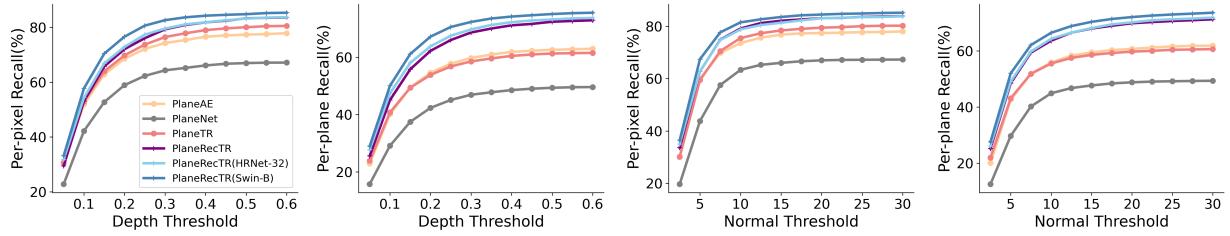


Figure 5. Per-pixel and per-plane recalls on the ScanNet dataset. See Appendix for more details.

CNN [16], with a ResNet-101 backbone, achieves better performance, partly caused by its utilization of neighbouring multi-view information and higher resolution images

during training, while ours is trained with a single image without using extra cues [16, 23]. Nevertheless, we have found that the benefit using more discriminative backbones

Method	Per-Pixel/Per-Plane Recalls \uparrow				Plane Parameters	
	Depth		Normal		Estimation Errors \downarrow	
	@0.10 m	@0.60 m	@5°	@30°	Normal (°)	Offset (mm)
PlaneRecTR (-M-D)	-	-	-	-	10.97	177.69
PlaneRecTR(-D)	50.28/43.36	82.87/72.29	61.42/47.80	83.44/71.11	10.27 (-0.70)	166.85 (-10.84)
PlaneRecTR	53.07/45.07	83.60/72.84	62.75/48.48	83.85/71.33	10.23 (-0.74)	165.13(-12.56)

Table 4. Ablation studies of PlaneRecTR’s unified settings on the ScanNet dataset.

(e.g., HRNet-32, Swin-B) also transfers to our generalization ability in novel scenes, resulting in a huge performance boost to reach a leading planar depth accuracy (see bottom two rows of Table 3). We present more evaluation results of 3D plane recovery on the NYUv2-Plane dataset in Appendix.

4.4. Ablation Studies

In this section, we conduct detailed ablation studies on the ScanNet dataset and show the effectiveness of each key design choice in our method.

Effects of Jointly Predicting Plane-Level Depths. One key difference to previous Transformer-based [23] and CNN-based baselines [17, 35, 16] is that, instead of learning monocular (planar) depth prediction in an individual branch/network, we predict plane-level depths based on shared representation for plane detection, segmentation and parameter prediction. It turns out that this simple design choice leads to promising performance increment, achieve mutual benefits among these tasks. Bottom two rows of Table 4 show that augmenting depth prediction task to query learning affects per-pixel and per-plane recalls of depth and normal, plane parameter errors (normal and offset). It is worth to note the significant recall rates gain especially under the strict (lower) thresholds.

Effects of Predicting Dense Plane-Level Masks. Compared to our closest Transformer-based baseline PlaneTR, we also differs in getting plane-level masks via a dense prediction instead of post-clustering from dense embeddings and depth information. Rightmost two columns of Table 4 have shown that explicitly driving to learn per-plane masks significantly boosts the accuracy of plane parameter estimation of our method, even when direct depth supervision is not available.

Plane Segmentation and Geometry Prediction. We also investigate whether the powerful segmentation framework could further benefit from learning plane geometry (plane parameters and depths). The experiment results shows that there are marginal numerical gains in terms of plane segmentation task comparing our complete model to the variant PlaneRecTR (-P-D). Fortunately, we do observe clear qualitative improvements. As shown in Figure 6, our complete model manages to distinguish detailed planar structures and

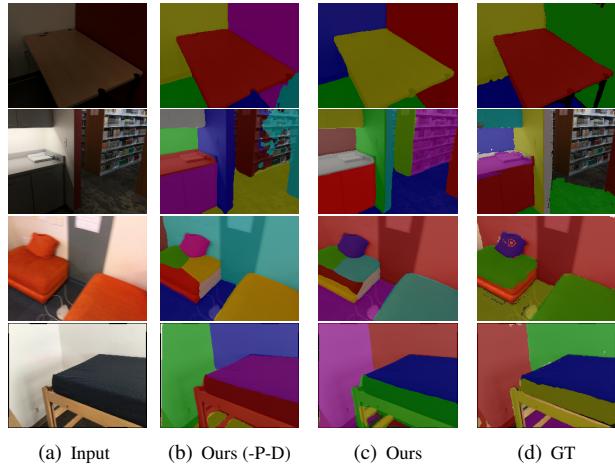


Figure 6. Plane segmentation comparison of PlaneRecTR(-P-D) and PlaneRecTR.

even predict more fine-grained and sound plane segmentation than ground truth annotation, which, however, could possibly be penalized during quantitative evaluation (bottom row of Figure 6 where bedstead labelling is missing).

5. Conclusion

Our proposed PlaneRecTR unifies plane related tasks including plane detection, plane parameter estimation, plane segmentation and depth estimation into a single Transformer-based module, its learnable plane queries implement explicit joint modelling and mutual benefits of plane-level semantic and geometric information in challenging single-view plane recovery task. We have shown that the simple yet effective design leads to obtain SOTA performance on the public ScanNet and NYUv2-Plane datasets. Ablation studies and visual examples validate our motivation and key design choices. More importantly, we have observed that the continually developing backbone vision models could further push the boundary of our performance, allowing our framework to be updated with new findings in such area with a plug-and-play manner. However, our method could only work on planar regions of a single image while it is more practical to build accurate planar reconstructions from a sparse set of views, possibly holistic geometric reasoning as well. We see this as an exciting venue to explore in the recent future.

6. Acknowledgements

We thank the anonymous reviewers and AC for their valuable comments, and authors of PlaneTR [23] for the helpful comments on evaluations. This work was supported in part by the National Key Research and Development Program of China (2018AAA0102200), NSFC (62325211, 62132021, 62201603) and Research Program of National University of Defense Technology (Grant No. ZK22-04).

References

- [1] Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 33(5):898–916, 2010. 5
- [2] Olga Barinova, Vadim Konushin, Anton Yakubenko, KeeChang Lee, Hwasup Lim, and Anton Konushin. Fast automatic single-view 3-d reconstruction of urban scenes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2008. 2
- [3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2020. 1, 2, 3
- [4] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3, 4
- [5] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021. 2, 3
- [6] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3d reconstructions of indoor scene. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 5
- [7] A. J. Davison. FutureMapping: The computational structure of Spatial AI systems. *arXiv preprint arXiv:1803.11288*, 2018. 1
- [8] Erick Delage, Honglak Lee, and Andrew Y Ng. Automatic single-image 3d reconstructions of indoor manhattan world scenes. In *Proceedings of the International Symposium on Robotics Research (ISRR)*, 2007. 1, 2
- [9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 2
- [10] David Ford Fouhey, Abhinav Gupta, and Martial Hebert. Unfolding an indoor origami world. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014. 2
- [11] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2017. 2
- [12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3, 5, 6
- [13] Xuan He, Fan Yang, Jiacheng Lin, Haolong Fu, Jin Yuan, Kailun Yang, and Zhiyong Li. Ssd-monodtr: Supervised scale-constrained deformable transformer for monocular 3d object detection. *arXiv preprint arXiv:2305.07270*, 2023. 3
- [14] Kuan-Chih Huang, Tsung-Han Wu, Hung-Ting Su, and Winston H. Hsu. Monodtr: Monocular 3d object detection with depth-aware transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3
- [15] David C Lee, Martial Hebert, and Takeo Kanade. Geometric reasoning for single image structure recovery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 1, 2
- [16] Chen Liu, Kihwan Kim, Jinwei Gu, Yasutaka Furukawa, and Jan Kautz. Planercnn: 3d plane detection and reconstruction from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2, 4, 5, 6, 7, 8
- [17] Chen Liu, Jimei Yang, Duygu Ceylan, Ersin Yumer, and Yasutaka Furukawa. Planenet: Piece-wise planar reconstruction from a single rgb image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2, 5, 6, 8
- [18] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 5
- [19] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5
- [20] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *Proceedings of the International Conference on 3D Vision (3DV)*, 2016. 4
- [21] David M Rosen, Kevin J Doherty, Antonio Terán Espinoza, and John J Leonard. Advances in inference and representation for simultaneous localization and mapping. *Annual Review of Control, Robotics, and Autonomous Systems*, 4:215–242, 2021. 1
- [22] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from RGBD images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2012. 5
- [23] Bin Tan, Nan Xue, Song Bai, Tianfu Wu, and Gui-Song Xia. Planetr: Structure-guided transformers for 3d plane recovery. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 1, 2, 4, 5, 6, 7, 8
- [24] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021. 1, 2
- [25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Neural Information Processing Systems (NeurIPS)*, 2017. 1, 2
- [26] Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Axial-deeplab: Standalone axial-attention for panoptic segmentation. In *Pro-*

- ceedings of the European Conference on Computer Vision (ECCV)*, 2020. 1, 2
- [27] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *CoRR*, abs/1908.07919, 2019. 5, 6
- [28] Wen Wang, Jing Zhang, Yang Cao, Yongliang Shen, and Dacheng Tao. Towards data-efficient detection transformers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 1, 2
- [29] Mark Weber, Jun Xie, Maxwell Collins, Yukun Zhu, Paul Voigtlaender, Hartwig Adam, Bradley Green, Andreas Geiger, Bastian Leibe, Daniel Cremers, et al. Step: Segmenting and tracking every pixel. *arXiv preprint arXiv:2102.11859*, 2021. 2
- [30] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. <https://github.com/facebookresearch/detectron2>, 2019. 5
- [31] Zizhang Wu, Yuanzhu Gan, Lei Wang, Guilian Chen, and Jian Pu. Monopgc: Monocular 3d object detection with pixel geometry contexts. *arXiv preprint arXiv:2302.10549*, 2023. 3
- [32] Shilin Xu, Xiangtai Li, Jingbo Wang, Guangliang Cheng, Yunhai Tong, and Dacheng Tao. Fashionformer: A simple, effective and unified baseline for human fashion segmentation and recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 2
- [33] Yufei Xu, Qiming Zhang, Jing Zhang, and Dacheng Tao. Vitae: Vision transformer advanced by exploring intrinsic inductive bias. *Advances in Neural Information Processing Systems*, 34:28522–28535, 2021. 1, 2
- [34] Fengting Yang and Zihan Zhou. Recovering 3d planes from a single image via convolutional neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2, 5
- [35] Zehao Yu, Jia Zheng, Dongze Lian, Zihan Zhou, and Shenghua Gao. Single-image piece-wise planar 3d reconstruction via associative embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2, 4, 5, 6, 8
- [36] Haobo Yuan, Xiangtai Li, Yibo Yang, Guangliang Cheng, Jing Zhang, Yunhai Tong, Lefei Zhang, and Dacheng Tao. Polyphonicformer: unified query learning for depth-aware video panoptic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 582–599, 2022. 2, 3
- [37] Shuaifeng Zhi, Michael Bloesch, Stefan Leutenegger, and Andrew J Davison. SceneCode: Monocular dense semantic reconstruction using learned encoded scene representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1
- [38] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J Davison. In-Place Scene Labelling and Understanding with Implicit Scene Representation. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 1
- [39] Shuaifeng Zhi, Edgar Sucar, Andre Mouton, Iain Haughton, Tristan Laidlow, and Andrew J Davison. iLabel: Revealing objects in neural fields. *IEEE Robotics and Automation Letters (RA-L)*, 8(2):832–839, 2022. 1