PreSight: Enhancing Autonomous Vehicle Perception with City-Scale NeRF Priors

Tianyuan Yuan[1], Yucheng Mao[1,2], Jiawei Yang[3], Yicheng Liu[1], Yue Wang[3], and Hang Zhao[1,4✉]

[1]IIIS, Tsinghua University [2]University of Science and Technology Beijing
[3]University of Southern California [4]Shanghai Qi Zhi Institute
{yuanty22@mails, hangzhao@mail}.tsinghua.edu.cn

Abstract. Autonomous vehicles rely extensively on perception systems to navigate and interpret their surroundings. Despite significant advancements in these systems recently, challenges persist under conditions like occlusion, extreme lighting, or in unfamiliar urban areas. Unlike these systems, humans do not solely depend on immediate observations to perceive the environment. In navigating new cities, humans gradually develop a preliminary mental map to supplement real-time perception during subsequent visits. Inspired by this human approach, we introduce a novel framework, **PreSight**, that leverages past traversals to construct static prior memories, enhancing online perception in later navigations. Our method involves optimizing a city-scale neural radiance field with data from previous journeys to generate neural priors. These priors, rich in semantic and geometric details, are derived without manual annotations and can seamlessly augment various state-of-the-art perception models, improving their efficacy with minimal additional computational cost. Experimental results on the nuScenes dataset demonstrate the framework's high compatibility with diverse online perception models. Specifically, it shows remarkable improvements in HD-map construction and occupancy prediction tasks, highlighting its potential as a new perception framework for autonomous driving systems. Our code will be released at https://github.com/yuantianyuan01/PreSight.

Keywords: Autonomous Driving · Vision-Based Perception · Neural Implicit Field

1 Introduction

Vision-based perception modules serve as the cornerstone of autonomous vehicle navigation, ensuring the safety and reliability of these systems. These modules support a variety of perception tasks, such as 3D object detection [14, 15, 21, 27, 41, 42], online map construction [19, 25, 28, 45, 50], and occupancy prediction [17, 20, 24, 38], each focusing on different aspects of the vehicle's surroundings. Perception tasks can be broadly classified into dynamic object perception,

✉: Corresponding Author.

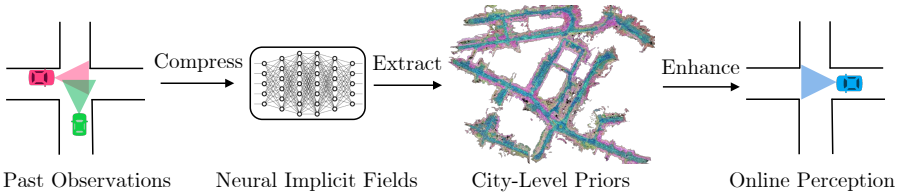


Fig. 1: Pipeline of PreSight. Leveraging historical traversal data, PreSight enhances online perception through NeRF, compressing extensive observations into implicit fields. A ray-marching algorithm extracts structured city-level priors, facilitating improved perception in subsequent visits.

which deals with moving entities like vehicles and pedestrians, and static environment perception, focusing on invariant features such as lane markers, vegetation, and obstacles. While dynamic objects require real-time observation, static environments can be pre-mapped and annotated, enhancing navigational safety. Conventionally, static environments are captured in high-definition (HD) maps using Simultaneous Localization and Mapping (SLAM) techniques [7, 34, 35, 52], supplemented with manual annotations. Despite their precision, expanding these methods to cover new geographic regions or incorporate new attributes presents considerable challenges, primarily due to the extensive reliance on manual labor for annotations and updates. In response, online perception models powered by deep learning have been developed [19, 22, 25, 28], utilizing only onboard sensor data to infer the surrounding environment. However, these models often struggle with occlusions, adverse weather, or extreme lighting conditions, and when deployed in unfamiliar urban settings not included in their training datasets, their reliability is compromised [26, 50].

Consider the innate human capability to navigate through cities. Individuals inherently develop preliminary mental maps of urban landscapes, with varying degrees of accuracy. These priors serve as valuable supplements to real-time sensory perception during subsequent explorations. Inspired by this human capability, we propose a novel framework, **PreSight**, illustrated in Figure 1, that utilizes historical traversal data—gathered through crowdsourcing or personal journeys—to automatically generate static priors without manual annotation efforts. These unsupervised, semantically and geometrically rich priors significantly enhance the efficacy of online perception models in subsequent navigations. Our framework includes city-scale neural radiance fields (NeRF) [31] priors and a straightforward integration module for various online perception models. Initially, we employ city-scale NeRFs to reconstruct urban environments using camera images and poses captured from past traversals, without requiring LiDAR input. Given the extensive spatial expanse of these cities, we divide them into tiles, each encompassing approximately 1 km^2. For each tile, several Instant-NGP [32] sub-fields are deployed and optimized in parallel for detailed coverage. Semantic information is distilled into the NeRFs by leveraging features from pre-trained 2D visual foundation models with semantic supervision. Structured

priors are then extracted using a ray marching algorithm inspired by EmerNeRF [47], selecting occupied voxels and gathering their semantic features from the constructed NeRFs. These featurized prior voxels serve as robust supplements to online perception. To integrate with prevalent online models [14, 22, 25, 50], we encode these priors into BEV or 3D features, using straightforward CNNs to merge features from both priors and real-time observations. The details of each part of our framework are provided in Section 3.

We validate our framework on the widely used nuScenes [4] dataset, focusing on two important online perception tasks: local HD-map construction and occupancy prediction. To illustrate our method's efficacy in enhancing the adaptability of online perception models to novel environments—and to avoid data leakage—we re-split the training-testing set to eliminate any geometric overlaps. The experimental results demonstrate that our framework significantly boosts the performance of several perception models across both tasks. For an in-depth discussion on the dataset and the experimental setup, please see Section 4.

To summarize, our contributions are as follows:

- We introduce a novel framework, **PreSight**, that leverages historical data to automatically construct powerful priors using NeRF.
- We demonstrate that the generated priors can be seamlessly integrated into various perception models and tasks, significantly enhancing their performance with minimal modification.

2 Related Work

2.1 Conventional 3D Reconstruction Techniques

Various 3D reconstruction techniques have been developed to create 3D models from 2D images. Structure from Motion (SfM) uses the relative motion between different views in a set of 2D images to reconstruct 3D structures [9,30]. It is commonly applied for estimating camera positions and generating 3D point clouds of scenes. However, SfM typically produces sparse and coarse geometric reconstructions. It also faces challenges in complex scenarios with occlusions [1,6,36]. Visual SLAM is regarded as a specialized variant of the SfM problem, tailored for robotics and autonomous driving applications. By incorporating pose information from GPS or IMU, vSLAM achieves more precise camera pose estimation and better reconstruction quality. Nonetheless, it encounters similar limitations in dynamic, complex, and large-scale environments [2, 10, 11, 53]. Additionally, infusing semantic information into the reconstructed models is not inherently straightforward with conventional 3D reconstruction techniques.

2.2 Large-Scale Neural Radiance Fields

The application of Neural Radiance Field (NeRF) in urban scenarios has attracted significant interest. Block NeRF [37] attempts to reconstruct city-scale scenes using traditional MLP-based NeRF from densely captured camera images,

but suffers from prolonged training durations. SUDS [40] employs hash grids [32] for faster training and divides large areas into multiple sub-fields for detailed coverage. StreetSurf [12], focusing on static scenarios in autonomous driving, delivers high-quality novel view synthesis and geometry but overlooks dynamic objects. UniSim [48] and NeuRAD [39] build neural simulators using NeRFs, and use SDF to model the surfaces. EmerNeRF [47] distinguishes between dynamic and static elements in a self-supervised manner and leverages vision foundation models for knowledge distillation. However, these approaches often depend on LiDAR depth data or dense observational views, limiting scalability in new regions via crowdsourcing. OccNeRF [51] and SelfOcc [16] explore using NeRF for autonomous driving scene occupancy estimation but neglect dynamic entities and underperform compared to online perception models, questioning their suitability as ground truth. MV-Map [44] leverages NeRF for offboard HD-map generation. Our methodology, in contrast, focuses on generating priors to enhance various online perception models.

2.3 Priors by Multi-Traversals

Recent work has explored on the nuScenes dataset [4] to enhance online perception in autonomous driving using priors from previous traversals. HindSight [49] utilizes LiDAR point clouds from past journeys to aid online detection tasks and proposes to re-split the nuScenes dataset to provide multi-traversal samples without geometrical overlaps, mitigating data leakage risks. Yet, these historical point clouds lack semantic depth, narrowing their utility. Neural Map Prior [45] employs neural representations to build and update a global map prior, boosting online map perception. Nonetheless, the substantial memory requirement for its neural representation limits scalability. Both approaches, while innovative, are confined to particular perception tasks and underscore the need for versatile priors that can support a broader range of tasks.

2.4 Visual Online Perception Model

Visual online perception models leverage camera image input to predict the surrounding environment in real-time, including a range of tasks from 3D object detection [14, 15, 21, 23, 29, 46] to online HD map construction [8, 19, 25, 28, 50] and occupancy prediction [13, 14, 22]. Particularly, we focus on online HD map construction and occupancy prediction. Online HD map construction is crucial for autonomous navigation, utilizing onboard sensors to create local HD maps. These maps can be vectorized [25, 28, 50] or rasterized [33, 45], providing detailed semantic guidance to ensure vehicular safety. This approach offers a cost-effective, scalable alternative to traditional, manually annotated HD maps by minimizing the need for human labor. The occupancy prediction task, pioneered by Occ3D [38], involves predicting the occupancy and semantic label of each voxel in a scene from camera images. This task delivers comprehensive information about the vehicle's surrounding environment but presents significant challenges in achieving accurate geometric and semantic predictions.

3 Methodologies

Our framework leverages self-supervised Neural Radiance Fields (NeRFs) [31] to generate static city priors from historical data and integrates these priors into online perception models, as demonstrated in Figure 1. This section is structured as follows: We begin by defining the problem in Section 3.1, proceed to describe the construction of city-scale NeRFs fused with semantic features in Section 3.2, introduce the extraction of structured priors in Section 3.3, and conclude with our approach to integrating these priors with state-of-the-art online perception models in Section 3.4.

3.1 Problem Formulation

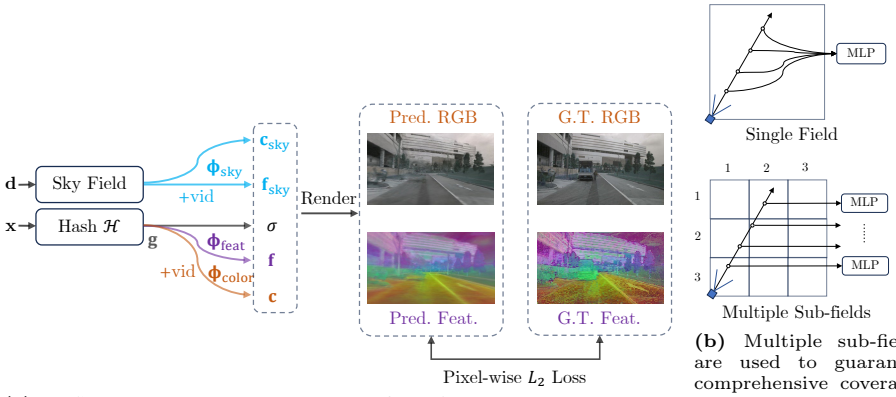
Consider a set of training samples X_{train} and their corresponding annotations Y_{train}. Typically, an online perception model M is trained on the dataset (X_{train}, Y_{train}). At deployment, this model is expected to effectively generalize to unseen samples X_{test}, including those from entirely new locations—a scenario where many advanced perception models encounter difficulties. Suppose we also have access to a set of unlabeled samples X_{prior}, which geographically overlap with X_{test} and can be readily acquired through crowdsourcing. Our objective is to effectively utilize the unlabeled X_{prior} to autonomously generate priors P_{prior}, thereby enhancing the performance of model M on X_{test}.

More specifically, during the training phase, we generate priors P_{train} based on X_{train}. An enhanced perception model M', equipped with an integration module, is then trained on the combined dataset $(X_{train} + P_{train}, Y_{train})$. Before evaluation, we generate priors P_{prior} using the unlabeled X_{prior}. The model M' is subsequently evaluated on the dataset $(X_{test} + P_{prior}, Y_{test})$. It is crucial to note that X_{train} is geographically distinct from $X_{test} \cup X_{prior}$ to prevent data leakage. For more details about the experimental setup, please refer to Section 4.1.

3.2 Building City-Scale NeRF

Inputs. Our method begins by taking sequences of RGB images from videos and their associated camera poses. To enhance these inputs, we utilize pre-trained vision foundation models for generating additional descriptors on pixels. Specifically, we employ DINO [5] for camera-view semantic feature generation and SegFormer [43] for producing camera-view semantic segmentation masks.

Benefits of Using NeRF in Our Framework. Neural Radiance Fields (NeRF) [32] excel at generating continuous, consistent, and high-fidelity 3D representations from sparse and irregularly spaced images, setting them apart from traditional 3D reconstruction techniques. Unlike methods that depend on specialized inputs such as LiDAR data or stereo camera imagery, NeRF operates effectively with standard 2D images. Moreover, NeRF's volumetric rendering approach allows for the straightforward exclusion of dynamic objects from scenes and the integration of pixel-wise semantic features [47]. These features can be extracted from powerful vision foundation models and seamlessly incorporated into the 3D



(a) PreSight's NeRF constructs distinct fields for nearby scenes and the sky, utilizing video embeddings to accommodate variations in lighting conditions across scenes. Semantic features are merged similar RGB color blending, enabling the distillation of semantic knowledge from DINO [5].

(b) Multiple sub-fields are used to guarantee comprehensive coverage. Volumetric rendering directly integrates sample points across these sub-fields.

Fig. 2: Overview of PreSight's scene representation.

field. In our framework, we use NeRF not merely as a tool for synthesizing novel views but as an efficient, self-supervised learner of environmental priors. Our focus is on the quality of geometric and semantic reconstructions rather than on the fidelity of novel view generation, leveraging NeRF's strengths to enhance the robustness and accuracy of online perception models in static environment tasks.

Scene Representation. Given the vast spatial extent of input videos throughout the city, we partition the city into tiles, each approximately 1 km^2. Each tile is independently optimized using video data overlapping with its designated area. We apply a straightforward application of K-Means clustering on all camera poses to determine tile boundaries.

For each tile, we construct a large static scene representation consisting of multiple sub-fields $\mathbf{F}^i$ to ensure comprehensive coverage. The query outputs are aggregated from these sub-fields along the ray during the rendering process. Similar to the initial clustering, we employ K-Means clustering on camera poses to establish centroids for these sub-fields. Each sub-field mimics the structure found in Instant-NGP [32], utilizing multi-resolution hash grids $\mathcal{H}$ coupled with a small MLP g to encode the implicit spatial features. Separate heads $\phi_{\text{color}}, \phi_{\text{feat}}, \phi_{\text{sky}}$ are utilized for predictions of color, semantic information, and sky appearance, respectively. The queried inputs include $(\mathbf{x}, \mathbf{d}, vid)$, where $\mathbf{x}$ represents the 3D location, $\mathbf{d}$ the view direction, and vid the video identifier:

$$\mathbf{g}, \sigma = g\left(\mathcal{H}(\mathbf{x})\right) \tag{1}$$

$\mathbf{g}, \sigma$ represent the per-point feature and density, respectively. For color predictions:

$$\mathbf{c} = \phi_{\text{color}}\left(\mathbf{g}, \gamma(\mathbf{d}), \mathbf{V}(vid)\right) \tag{2}$$

$\gamma(\cdot)$ denotes spherical harmonic encoding for direction, and $\mathbf{V}(\cdot)$ represents per-video embeddings to account for video-specific lighting conditions. For semantic feature predictions:

$$\mathbf{f} = \phi_{\text{feat}}\left(\mathbf{g}\right) \tag{3}$$

This process models direction-invariant semantic features. Overall, the output for each query is given by:

$$\sigma, \mathbf{c}, \mathbf{f} = \mathbf{F}(\mathbf{x}, \mathbf{d}, vid) \tag{4}$$

For sky features, a position-independent MLP ϕ_{sky} predicts both color and semantic features:

$$\mathbf{c}_{\text{sky}}(\mathbf{d}), \mathbf{f}_{\text{sky}}(\mathbf{d}) = \phi_{\text{sky}}\left(\gamma(\mathbf{d}), \mathbf{V}(vid)\right) \tag{5}$$

Rendering. To render a camera ray $\mathbf{r} = (\mathbf{o}, \mathbf{d})$ within a video vid (where o denotes the origin), we sample N points $\left\{\mathbf{x}^i\right\}$ along the direction $\mathbf{d}$. Each sample point is assigned to the nearest sub-field. The assignment process is defined as:

$$\delta(\mathbf{x}^i) = \arg\min_j \left\|\mathbf{x}^i - \mathbf{centroid}^j\right\|_2 \tag{6}$$

where $\mathbf{centroid}^j$ denotes the centroid of j-th sub-field. After assignment, outputs for each sample point are queried exclusively from the assigned sub-field:

$$\sigma^i, \mathbf{c}^i, \mathbf{f}^i = \mathbf{F}^{\delta(\mathbf{x}^i)}(\mathbf{x}^i, \mathbf{d}, vid) \tag{7}$$

The final rendering is an integration along the ray, blending color and semantic features:

$$\hat{C}(\mathbf{r}) = \sum_{i=1}^{N} T^i \alpha^i \mathbf{c}^i + \left(1 - \sum_{i=1}^{N} T^i \alpha^i\right) \mathbf{c}_{\text{sky}}(\mathbf{d}) \tag{8}$$

$$\hat{F}(\mathbf{r}) = \sum_{i=1}^{N} T^i \alpha^i \mathbf{f}^i + \left(1 - \sum_{i=1}^{N} T^i \alpha^i\right) \mathbf{f}_{\text{sky}}(\mathbf{d}) \tag{9}$$

α^i and T^i denote the accumulated transmittance and piece-wise opacity, respectively, calculated as follows:

$$\alpha^i = 1 - \exp\left(-\sigma^i(\mathbf{x}^{i+1} - \mathbf{x}^i)\right) \tag{10}$$

$$T^i = \prod_{j=1}^{i-1}(1 - \alpha^j) \tag{11}$$

Optimization. We employ an L2 loss for color $\mathcal{L}_{\text{rgb}}$ and semantic feature reconstruction $\mathcal{L}_{\text{feat}}$, along with a binary cross-entropy loss for sky features $\mathcal{L}_{\text{sky}}$ following approach introduced in [12, 47]. Inspired by Mip-NeRF 360 [3], we use two proposal networks, integrating interlevel and distortion losses. The comprehensive loss function is represented as:

$$\mathcal{L} = \mathcal{L}_{\text{rgb}} + \lambda_{\text{feat}}\mathcal{L}_{\text{feat}} + \lambda_{\text{sky}}\mathcal{L}_{\text{sky}} + \lambda_{\text{inter}}\mathcal{L}_{\text{inter}} + \lambda_{\text{dist}}\mathcal{L}_{\text{dist}} \tag{12}$$

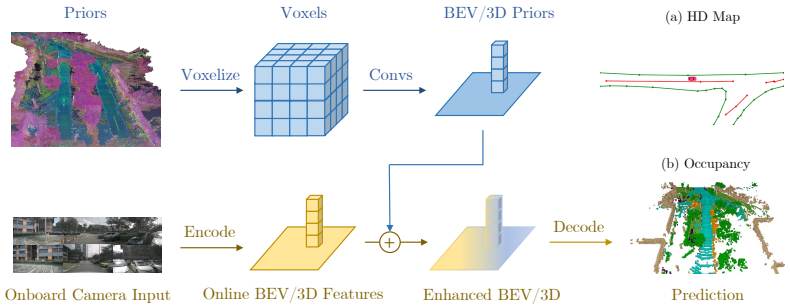


Fig. 3: The integration module. Point-based priors are voxelized and then encoded into BEV or 3D features using convolutional layers. These features are fused with online features from perception models, enhancing the overall features for decoding.

To focus on static environment features, we mask out pixels associated with potentially moving objects, such as vehicles and pedestrians, using segmentation masks derived from SegFormer [43].

3.3 Prior Extraction

After the optimization phase, the NeRFs encapsulate rich information within their hash grids and MLPs. To make this dense, unstructured information inherent in NeRFs usable for online perception models, it requires conversion into a structured form. We employ a ray marching algorithm to identify the voxels that are occupied and to capture their associated features. The process begins by casting rays originating from the training cameras. Similar to the training procedure, we sample N sample points $\{\mathbf{x}^i\}$ along each ray. The surface point is determined by locating the first point where the cumulative transmittance and opacity exceed a threshold:

$$j = \min_i \left(\sum_{i=1}^{N} T^i \alpha^i > 0.5 \right) \tag{13}$$

Here, T^i and α^i are defined in Equation (10) and Equation (11). We collect the identified surface point $\mathbf{x}^j$ and its corresponding semantic feature $\phi_{\text{feat}}(\mathbf{g})$, as previously defined in Equation (3). Following the aggregation of surface points from all training views, we apply voxel-based downsampling to minimize the points, averaging the features within each voxel. These feature-enriched voxels then act as robust priors for enhancing online perception models. It is worth noting that the extraction is only performed once after building the city-scale NeRF, and the extracted priors are stored. Neither the slow rendering process or the huge size of NeRFs impact the online perception models. During deployment, only the extracted priors are stored and used by the online perception models.

3.4 Integrating Priors to Online Models

Online perception models commonly transform image inputs into a canonical representation, such as BEV or 3D space features. To ensure our framework's broad compatibility, we adapt our pre-built priors to BEV or 3D space, simplifying their integration with various state-of-the-art online perception models [14, 22, 25, 50]. As illustrated in Figure 3, our integration strategy employs several simple CNN layers to transform the voxel features into BEV or 3D space, which are then concatenated with the existing online features from the perception models. This module introduces negligible computational overhead and requires no alterations to the original model architectures, efficiently enhancing the models with rich environmental priors for improved real-time perception capabilities.

4 Experiments

In this section, we evaluate our **PreSight** framework on the nuScenes [4] dataset, specifically targeting perception tasks involving static targets: online HD map construction and occupancy prediction. These tasks are chosen because our framework is designed to leverage priors to enhance static environment perception, a crucial part for these applications. The dataset are introduced in Section 4.1, followed by NeRF implementation details in Section 4.2. Section 4.3 showcases our framework's enhancements to baseline models. Our priors are compared to alternatives in Section 4.4, and Section 4.5 presents some ablation studies of our methods.

4.1 Setup

We evaluate our framework using the nuScenes [4] dataset, a leading benchmark in autonomous driving, collected across four regions: *Boston Seaport*, *Singapore Onenorth*, *Singapore Queenstown*, and *Singapore Holland Village*, featuring 1000 scenes with six synchronized cameras and vehicle poses. The official dataset split allocates 700 scenes for training and 150 for validation. However, recent studies have identified data leakage issues within this official split, highlighting the need for a revised partitioning to accurately assess static perception models' generalization capabilities. Following our problem setup in Section 3.1, we assign scenes from *Boston Seaport* and *Singapore Queenstown* to our training set (X_{train}, Y_{train}). For testing, one-third of the scenes from *Singapore Onenorth* and *Singapore Holland Village* are used as X_{prior} for generating priors, with the remaining two-thirds forming the testing set (X_{test}, Y_{test}). Therefore, following the problem setup detailed in Section 3.1, we set all samples in *Boston Seaport* and *Singapore Queenstown* as training set (X_{train}, Y_{train}). This distribution strategy mirrors the practical scenario of scaling and generalizing an autonomous driving system: training is conducted using data from specific cities, followed by deployment in previously unencountered cities, supported by a limited set of observations X_{prior} sourced through crowdsourcing. Overall, we established

Table 1: PreSight enhancements across different online HD map construction models: StreamMapNet and MapTR use vectorized predictions, evaluated by Average Precision (AP), while BEVFormer's rasterized predictions are assessed using Intersection over Union (IoU).

Model	Metric	w. Prior	Ped Crossing	Divider	Boundary	All	Runtime (FPS)
StreamMapNet	AP	×	10.19	11.26	11.87	11.10	**22.4**
		✓	**21.11**	**23.73**	**32.31**	**25.72 (+14.62)**	21.9
MapTR	AP	×	4.97	8.20	9.83	7.67	**25.2**
		✓	**16.18**	**19.04**	**34.14**	**23.12 (+15.45)**	23.2
BEVFormer	IoU	×	14.90	29.88	32.74	25.84	**15.5**
		✓	**16.37**	**34.82**	**51.66**	**34.28 (+8.44)**	14.3

a split of 614 training scenes, 82 prior scenes, and 154 testing scenes, with an 80% geographical overlap between the prior and testing sets. Further details on scene allocation can be found in the supplemental material.

4.2 NeRF Implementation

This section introduces the implementation details for constructing city-scale Neural Radiance Fields (NeRFs). We utilize 12Hz image sequences from six cameras provided by the nuScenes dataset, yielding approximately 1,400 images for each 20-second scene. following the approach of EmerNeRF [47], we compress the dimension of DINO features from 764 to 64. Dynamic objects including *person, rider, car, truck, bus, train, motorcycle,* and *bicycle* are masked using segmentation results from SegFormer [43]. As detailed in Section 3.2, urban areas are divided into tiles of roughly 1 km^2. Specifically, we assign 8, 4, 4, and 2 tiles to *Boston Seaport*, *Onenorth*, *Queenstown*, and *Holland Village*, respectively, based on their sizes. Each tile contains 8 or 16 sub-fields, depending on the tile's area. NeRF training for each tile involves 100,000 iterations with a batch size of 2^{16}, requiring about 20 hours on a single A100 GPU.

4.3 Experimental Results on Online models

Online HD Map construction. Online HD maps construction takes as input camera images to generate local HD maps, which can be either vectorized [25, 50] or rasterized [45]. Our research demonstrates that the integration of proposed priors enhances the performance of state-of-the-art models. Specifically, for vectorized mapping, we employed two models with distinct decoders, namely MapTR [25] and StreamMapNet [50]. For rasterized mapping, we adopted a BEVFormer-based [21] model, following Neural Map Prior [45]. These models utilize BEV features as an intermediate representation, which are fused with our pre-built priors using our integration module. We trained and evaluated each model based on their original codebase, with our proposed training-testing split. The perception range was set to 100×50 meters, aligning with StreamMapNet's specifications. Given that map construction is inherently a static perception task,

Table 2: PreSight boosts state-of-the-art occupancy prediction models, distinguishing between dynamic and static targets for clarity. The *trailer* class is excluded due to its absence in our proposed testing set.

					Dynamic										**Static**						
Method	w. Priors	mIoU	Dynamic	**Static**	others	barrier	bicycle	bus	car	constr. vehicle	motorcycle	pedestrian	traffic cone	truck	drive surface	other flat	sidewalk	terrain	manmade	vegetation	Runtime (FPS)
BEVDet	×	29.3	24.4	38.2	**1.5**	**42.4**	11.0	**43.0**	47.1	**19.1**	23.3	23.4	**19.5**	**37.8**	72.9	11.6	30.9	48.6	32.7	32.5	**5.1**
	✓	**33.7**	**24.4**	**50.5**	1.2	40.1	**14.8**	42.1	**48.3**	15.7	**26.4**	**24.4**	18.7	37.2	**81.8**	**15.2**	**40.3**	**60.5**	**50.4**	**54.9**	4.9
FB-Occ	×	30.0	25.1	39.2	9.2	37.2	**21.8**	**41.6**	43.4	15.8	27.3	25.4	**23.8**	**30.3**	74.7	17.3	33.0	50.6	28.2	31.1	**9.1**
	✓	**34.3**	**25.4**	**50.7**	**9.3**	**38.3**	21.0	40.3	**45.0**	**15.9**	**29.9**	**26.0**	23.8	30.2	**82.3**	**18.5**	**39.1**	**61.2**	**48.0**	**54.7**	8.6

our framework significantly elevates the quality of the resulting HD maps across all models Table 1. The performance metrics for the baseline models are lower than those originally reported in their respective papers due to our use of a more challenging dataset split without geographical overlaps, a phenomenon similarly observed by Lilja *et al.* [26] and Yuan *et al.* [50] More implementation details are provided in the supplemental materials.

3D Occupancy Prediction. 3D occupancy prediction task involves estimating both the occupancy state and semantic label of each voxel in a scene, using camera images as input. We benchmark our approach using the widely recognized Occ3D dataset [38] on nuScenes. For baselines, we selected FB-OCC [22] and BEVDet [14], two leading state-of-the-art models with distinct architectures, integrating our pre-built priors into their BEV features. Each model was trained and evaluated on their original codebase, with our proposed training-testing split. The results, presented in Table 2, demonstrate our framework's ability to enhance the baseline models' performance. Notably, we specifically highlight static categories such as *drive surface*, *other flat*, *sidewalk*, *terrain*, *manmade*, and *vegetation*. Our framework significantly improves outputs for these targets, without impacting dynamic categories, aligning with our goal of enhancing static perception accuracy.

4.4 Comparison with Other Priors

This section compares our generated priors with alternative approaches to evaluate their effectiveness.

Historical Point Clouds. A straightforward method for utilizing historical data involves collecting LiDAR-generated point clouds from previous traversals [49]. Despite our framework's independence from LiDAR, we assessed whether our priors could outperform these point clouds, known for their geometric precision. We replaced our priors with historical LiDAR point clouds to evaluate performance in HD map construction and occupancy prediction tasks.

Neural Map Prior. Neural Map Prior [45] enhances online map perception models with automatically generated priors, aligning with our objectives. We

Table 3: PreSight outperforms other priors in HD map construction and achieves similar results to LiDAR priors in occupancy prediction, using only camera inputs.

(a) HD Map Construction

Model	Metric	Prior Type	mAP/mIoU
StreamMapNet	AP	×	11.10
		LiDAR	16.87
		ours	**25.72**
MapTR	AP	×	7.67
		LiDAR	15.93
		ours	**23.12**
BEVFormer	IoU	×	25.84
		LiDAR	33.01
		NMP	26.50
		ours	**34.42**

(b) Occupancy Prediction

Model	Prior Type	mIoU	Dynamic	Static
BEVDet	×	29.3	24.4	38.2
	LiDAR	33.1	21.7	**53.9**
	ours	**33.7**	**24.4**	50.5
FB-OCC	×	30.0	25.1	39.2
	LiDAR	34.3	23.4	**54.4**
	ours	**34.3**	**25.4**	50.7

Table 4: Ablation study reveals the impact of distilled semantic features: Priors enhance online perception on their own, but performance significantly improves with the addition of distilled features.

(a) HD Map Construction

Model	Prior Type	mAP
StreamMapNet	×	11.10
	no feats.	11.45
	full	**25.72**

(b) Occupancy Prediction

Model	Prior Type	mIoU	Dynamic	Static
BEVDet	×	29.3	24.4	38.2
	no feats.	32.0	24.1	46.4
	full	**33.7**	**24.4**	**50.5**

adapted this method to our experimental setup by training it with our dataset (X_{train}, Y_{train}) and generating neural priors from X_{prior} for evaluation on (X_{test}, Y_{test}), making best use of its framework's potential. This comparison conducted on rasterized map construction task, following the official implementation. Notably, our priors are significantly more compact, requiring only 3.5 GB for *Boston Seaport*, compared to the 38 GB needed for the rasterized BEV features in Neural Map Prior.

The results in Table 3 highlight our framework's superiority in both perception tasks. For HD map construction, our priors offer richer semantic details. In occupancy prediction, our approach matches historical LiDAR point clouds. While LiDAR slightly enhances static target perception, it fails to differentiate dynamic from static objects, causing potential inaccuracies. Our method maintains performance without compromising dynamic object perception.

4.5 Ablation Studies

This section explores the impact of various components of our framework on the final outcomes. Specifically, we investigate the role of distilled DINO [5] features and the effect of varying the proportion of X_{prior} utilized for constructing priors.

Table 5: Performance improves as as we priors scenes scale up. × represents baseline models. 0% denotes that no priors are used during testing, whereas 50% indicates that only half of the scenes in X_{prior} are used for testing. A 100% ratio means all scenes in X_{prior} are used.

(a) HD Map Construction

Model	Prior Type	mAP
StreamMapNet	×	11.10
	0%	12.03
	50%	21.62
	100%	**25.72**

(b) Occupancy Prediction

Model	Prior Type	mIoU	Dynamic	Static
BEVDet	×	29.3	24.4	38.2
	0%	28.7	24.6	36.2
	50%	32.0	24.1	46.4
	100%	**33.7**	**24.4**	**50.5**

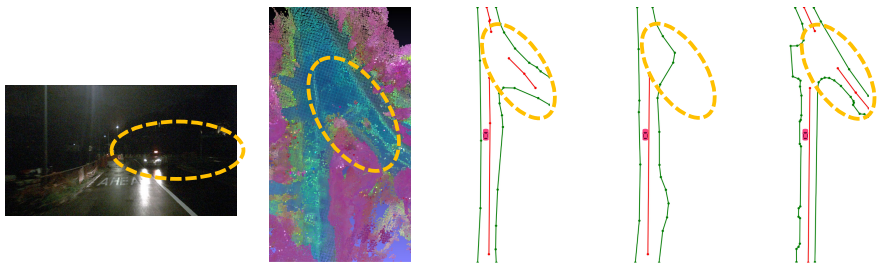


Fig. 4: Visualization of StreamMapNet prediction with and without priors. The model with priors accurately predicts road structure, while baseline model without priors fails.

Distilled DINO Features. A crucial element of our proposed priors is the distilled semantic knowledge from the robust vision foundation model, DINO. As demonstrated in Table 4, removing these distilled semantic features leads to a notable decline in performance for both tasks, highlighting the importance of integrating distilled semantic knowledge. Particularly, HD map construction benefits greatly from semantic features for identifying map elements, whereas occupancy prediction depends more on geometrical information.

Proportions of Used Prior Scenes. The effectiveness of our framework increases with the amount of historical traversal data utilized. To illustrate this, we evaluated the trained model using varying proportions of prior scenes. The outcomes are summarized in Table 5. Performance improves as more scenes are utilized to create priors, showcasing our framework's scalability. Importantly, models trained with priors do not depend exclusively on them and still achieve performance comparable to baseline models with no priors at test time.

4.6 Qualitative Analysis

In this section, we present qualitative results from our framework, underscoring the value of incorporating priors. Figure 4 depicts a low-light scenario at an intersection that is hardly visible in the front-view image. Here, the baseline

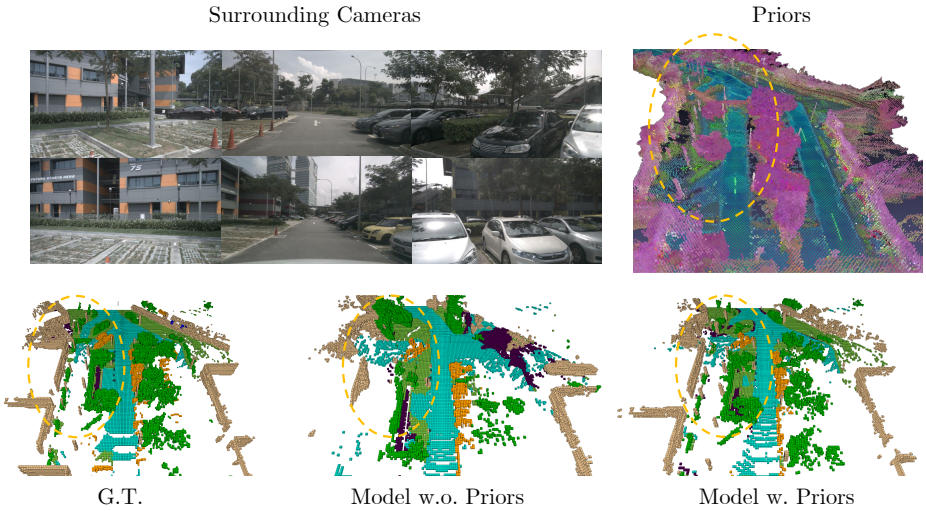


Fig. 5: Visualization of BEVDet predictions and the pre-built priors. Self-supervised priors reconstruct fine-grained geometries even more denser than groundtruth, enabling the enhanced BEVDet model to deliver more precise occupancy predictions than baseline models lacking priors.

model without priors misinterprets the road layout, whereas the model equipped with priors accurately reconstructs it. The visualization of prior features reveals the detailed road structure, underscoring how priors provide crucial information for generating consistent and reliable HD maps critical for autonomous vehicle safety. Figure 5 highlights the enhancement in occupancy prediction provided by our priors. Notably, within the highlighted region, our priors accurately capture the detailed structure of a light pole and the ground surface—details even absent in the ground truth. Consequently, the model with priors precisely predicts occupancy with fine details, in contrast to the baseline model's failure.

5 Dicussion, Conclusion & Acknowledgement

Limitations. Our approach relies on accurate ego-vehicle poses and camera sensors, usually provided by GPS and IMU systems, which might not be readily available in crowdsourced data. Future efforts could aim to reduce the prior memory size or exploring alternative self-supervised reconstruction methods like 3D Gaussian Splatting [18], offering promising directions for further research.
Conclusion. In this paper, we introduced PreSight, a novel framework that autonomously constructs static, city-scale priors for enhancing online perception in subsequent traversals. Our experiments on the nuScenes dataset demonstrate the framework's effectiveness and its wide applicability across various perception models.
Acknowledgement This work is supported by Tsinghua University Dushi Program.

References

1. Arie-Nachimson, M., Kovalsky, S.Z., Kemelmacher-Shlizerman, I., Singer, A., Basri, R.: Global motion estimation from point matches. In: 2012 Second International Conference on 3D Imaging, Modeling, Processing, Visualization & Transmission. pp. 81–88 (2012). https://doi.org/10.1109/3DIMPVT.2012.46
2. Aulinas, J., Petillot, Y., Salvi, J., Llado, X.: The slam problem: a survey. vol. 184, pp. 363–371 (01 2008). https://doi.org/10.3233/978-1-58603-925-7-363
3. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Mip-nerf 360: Unbounded anti-aliased neural radiance fields. CVPR (2022)
4. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11621–11631 (2020)
5. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the International Conference on Computer Vision (ICCV) (2021)
6. Crandall, D., Owens, A., Snavely, N., Huttenlocher, D.: Discrete-continuous optimization for large-scale structure from motion. In: CVPR 2011. pp. 3001–3008 (2011). https://doi.org/10.1109/CVPR.2011.5995626
7. Dellaert, F.: Factor graphs and gtsam: A hands-on introduction. Tech. rep., Georgia Institute of Technology (2012)
8. Ding, W., Qiao, L., Qiu, X., Zhang, C.: Pivotnet: Vectorized pivot learning for end-to-end hd map construction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3672–3682 (2023)
9. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM **24**(6), 381–395 (jun 1981). https://doi.org/10.1145/358669.358692, https://doi.org/10.1145/358669.358692
10. Fuentes-Pacheco, J., Ascencio, J., Rendon-Mancha, J.: Visual simultaneous localization and mapping: A survey. Artificial Intelligence Review **43** (11 2015). https://doi.org/10.1007/s10462-012-9365-8
11. Goldstein, T., Hand, P., Lee, C., Voroninski, V., Soatto, S.: Shapefit and shapekick for robust, scalable structure from motion (08 2016)
12. Guo, J., Deng, N., Li, X., Bai, Y., Shi, B., Wang, C., Ding, C., Wang, D., Li, Y.: Streetsurf: Extending multi-view implicit surface reconstruction to street views. arXiv preprint arXiv:2306.04988 (2023)
13. Hong, Y., Liu, Q., Cheng, H., Ma, D., Dai, H., Wang, Y., Cao, G., Ding, Y.: Univision: A unified framework for vision-centric 3d perception (2024)
14. Huang, J., Huang, G.: Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. arXiv preprint arXiv:2203.17054 (2022)
15. Huang, J., Huang, G., Zhu, Z., Yun, Y., Du, D.: Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. arXiv preprint arXiv:2112.11790 (2021)
16. Huang, Y., Zheng, W., Zhang, B., Zhou, J., Lu, J.: Selfocc: Self-supervised vision-based 3d occupancy prediction. arXiv preprint arXiv:2311.12754 (2023)
17. Huang, Y., Zheng, W., Zhang, Y., Zhou, J., Lu, J.: Tri-perspective view for vision-based 3d semantic occupancy prediction. arXiv preprint arXiv:2302.07817 (2023)
18. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4) (July 2023), https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/

19. Li, Q., Wang, Y., Wang, Y., Zhao, H.: Hdmapnet: A local semantic map learning and evaluation framework. arXiv preprint arXiv:2107.06307 (2021)
20. Li, Y., Yu, Z., Choy, C., Xiao, C., Alvarez, J.M., Fidler, S., Feng, C., Anandkumar, A.: Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
21. Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Qiao, Y., Dai, J.: Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. arXiv preprint arXiv:2203.17270 (2022)
22. Li, Z., Yu, Z., Austin, D., Fang, M., Lan, S., Kautz, J., Alvarez, J.M.: FB-OCC: 3D occupancy prediction based on forward-backward view transformation. arXiv:2307.01492 (2023)
23. Li, Z., Yu, Z., Wang, W., Anandkumar, A., Lu, T., Alvarez, J.M.: FB-BEV: BEV representation from forward-backward view transformations. In: IEEE/CVF International Conference on Computer Vision (ICCV) (2023)
24. Li, Z., Zhang, C., Ma, W.C., Zhou, Y., Huang, L., Wang, H., Lim, S., Zhao, H.: Voxelformer: Bird's-eye-view feature generation based on dual-view attention for multi-view 3d object detection (2023)
25. Liao, B., Chen, S., Wang, X., Cheng, T., Zhang, Q., Liu, W., Huang, C.: Maptr: Structured modeling and learning for online vectorized hd map construction. In: International Conference on Learning Representations (2023)
26. Lilja, A., Fu, J., Stenborg, E., Hammarstrand, L.: Localization is all you evaluate: Data leakage in online mapping datasets and how to fix it (2023)
27. Lin, X., Lin, T., Pei, Z., Huang, L., Su, Z.: Sparse4d v2: Recurrent temporal fusion with sparse model (2023)
28. Liu, Y., Yuan, T., Wang, Y., Wang, Y., Zhao, H.: Vectormapnet: End-to-end vectorized hd map learning (2023)
29. Liu, Z., Tang, H., Amini, A., Yang, X., Mao, H., Rus, D., Han, S.: Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In: IEEE International Conference on Robotics and Automation (ICRA) (2023)
30. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International journal of computer vision **60**, 91–110 (2004)
31. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: ECCV (2020)
32. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM Trans. Graph. **41**(4), 102:1–102:15 (Jul 2022). https://doi.org/10.1145/3528223.3530127, https://doi.org/10.1145/3528223.3530127
33. Roddick, T., Cipolla, R.: Predicting semantic map representations from images using pyramid occupancy networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11138–11147 (2020)
34. Segal, A., Hähnel, D., Thrun, S.: Generalized-icp. In: Robotics: science and systems (06 2009). https://doi.org/10.15607/RSS.2009.V.021
35. Shan, T., Englot, B.: Lego-loam: Lightweight and ground-optimized lidar odometry and mapping on variable terrain. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 4758–4765. IEEE (2018)
36. Snavely, N., Seitz, S.M., Szeliski, R.: Photo tourism: exploring photo collections in 3D. Association for Computing Machinery, New York, NY, USA, 1 edn. (2023), https://doi.org/10.1145/3596711.3596766

37. Tancik, M., Casser, V., Yan, X., Pradhan, S., Mildenhall, B., Srinivasan, P.P., Barron, J.T., Kretzschmar, H.: Block-nerf: Scalable large scene neural view synthesis (Feb 2022), http://arxiv.org/abs/2202.05263v1
38. Tian, X., Jiang, T., Yun, L., Wang, Y., Wang, Y., Zhao, H.: Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. arXiv preprint arXiv:2304.14365 (2023)
39. Tonderski, A., Lindström, C., Hess, G., Ljungbergh, W., Svensson, L., Petersson, C.: Neurad: Neural rendering for autonomous driving. arXiv preprint arXiv:2311.15260 (2023)
40. Turki, H., Zhang, J.Y., Ferroni, F., Ramanan, D.: Suds: Scalable urban dynamic scenes. In: Computer Vision and Pattern Recognition (CVPR) (2023)
41. Wang, T., Zhu, X., Pang, J., Lin, D.: FCOS3D: Fully convolutional one-stage monocular 3d object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops (2021)
42. Wang, Y., Guizilini, V.C., Zhang, T., Wang, Y., Zhao, H., Solomon, J.: Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In: Conference on Robot Learning. pp. 180–191. PMLR (2022)
43. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. In: Neural Information Processing Systems (NeurIPS) (2021)
44. Xie, Z., Pang, Z., Wang, Y.X.: Mv-map: Offboard hd-map generation with multi-view consistency. arXiv (2023)
45. Xiong, X., Liu, Y., Yuan, T., Wang, Y., Wang, Y., Hang, Z.: Neural map prior for autonomous driving. In: CVPR (2023)
46. Yang, C., Chen, Y., Tian, H., Tao, C., Zhu, X., Zhang, Z., Huang, G., Li, H., Qiao, Y., Lu, L., Zhou, J., Dai, J.: Bevformer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision. ArXiv (2022)
47. Yang, J., Ivanovic, B., Litany, O., Weng, X., Kim, S.W., Li, B., Che, T., Xu, D., Fidler, S., Pavone, M., Wang, Y.: Emernerf: Emergent spatial-temporal scene decomposition via self-supervision. arXiv preprint arXiv:2311.02077 (2023)
48. Yang, Z., Chen, Y., Wang, J., Manivasagam, S., Ma, W.C., Yang, A.J., Urtasun, R.: Unisim: A neural closed-loop sensor simulator. In: CVPR (2023)
49. You, Y., Luo, K.Z., Chen, X., Chen, J., Chao, W.L., Sun, W., Hariharan, B., Campbell, M., Weinberger, K.Q.: Hindsight is 20/20: Leveraging past traversals to aid 3d perception. arXiv preprint arXiv:2203.11405 (2022)
50. Yuan, T., Liu, Y., Wang, Y., Wang, Y., Zhao, H.: Streammapnet: Streaming mapping network for vectorized online hd map construction. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 7356–7365 (January 2024)
51. Zhang, C., Yan, J., Wei, Y., Li, J., Liu, L., Tang, Y., Duan, Y., Lu, J.: Occnerf: Self-supervised multi-camera occupancy prediction with neural radiance fields. arXiv preprint arXiv:2312.09243 (2023)
52. Zhang, J., Singh, S.: Loam : Lidar odometry and mapping in real-time. Robotics: Science and Systems Conference (RSS) pp. 109–111 (01 2014)
53. Özyeşil, O., Voroninski, V., Basri, R., Singer, A.: A survey of structure from motion. Acta Numerica **26**, 305–364 (2017). https://doi.org/10.1017/S096249291700006X