MV-Map: Offboard HD Map Generation with Multi-view Consistency

Ziyang Xie*[1,2] Ziqi Pang*[1] Yu-Xiong Wang[1]
[1]University of Illinois Urbana-Champaign [2]Fudan University
{ziyang8, ziqip2, yxw}@illinois.edu

Abstract

While bird's-eye-view (BEV) perception models can be helpful in building high-definition maps (HD maps) with less human labor, their results are often unreliable and demonstrate noticeable inconsistencies in the predicted HD maps from different viewpoints. This is because BEV perception is typically set up in an "onboard" manner, which restricts the computation and prevents algorithms from simultaneously reasoning multiple views. This paper overcomes these limitations and advocates a more practical "offboard" HD map generation setup that removes the computation constraints, based on the fact that HD maps are commonly reusable infrastructures built offline in data centers. To this end, we propose a novel offboard pipeline called MV-Map that capitalizes multi-view consistency and can handle an arbitrary number of frames with the key design of a "region-centric" framework. In MV-Map, the target HD maps are created by aggregating all the frames of onboard predictions, weighted by the confidence scores assigned by an "uncertainty network." To further enhance multi-view consistency, we augment the uncertainty network with the global 3D structure optimized by a voxelized neural radiance field (Voxel-NeRF). Extensive experiments on nuScenes show that our MV-Map significantly improves the quality of HD maps, further highlighting the importance of offboard methods for HD map generation. Our code and model are available at https://github.com/ZiYang-xie/MV-Map.

1. Introduction

High-definition maps (HD maps) play a crucial role in ensuring the safe navigation of autonomous vehicles, by providing essential positional and semantic information about road elements. Ideally, one would expect the process of constructing HD maps to be as simple as collecting numerous sensory data while driving and then utilizing an *automatic* algorithm to extract the road elements, as illustrated in Fig. 1. However, the mainstream solutions generally involve human annotators, as seen in widely-used datasets [3, 4, 8, 39]. This design is based on the consideration of the infrastructure role and high re-usability of HD maps, which can serve autonomous vehicles for virtually *infinite* times after a *single* construction process before any environmental changes. Even so, the expense of manual annotation obstructs the expansion of autonomous driving to new locations, and we aim to develop reliable algorithms that can decrease or replace the need for human labor in HD map construction.

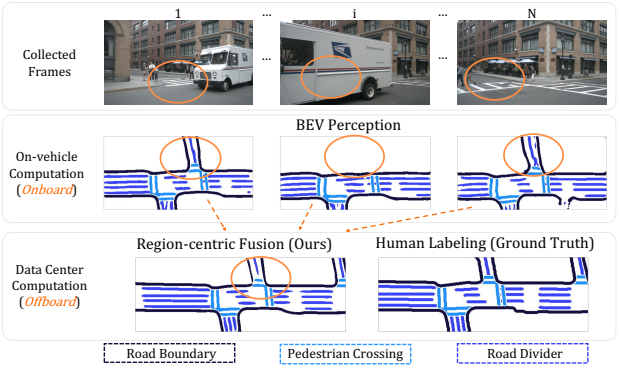

Figure 1: Current *onboard* methods generate unreliable HD map predictions that are inconsistent across multiple views, due to occlusions or viewpoint changes. By contrast, our *offboard* pipeline constructs a unified and multi-view consistent HD map with clearer lanes. Our key design is a *region-centric* framework that aggregates single-frame information for each target HD map region.

Towards this goal, there have been recent attempts that automatically generate HD maps using bird's-eye-view (BEV) perception [11, 13, 15]. However, their results are often unreliable, as illustrated by noticeable inconsistencies in the predicted HD maps from different viewpoints (a representative example is in Fig. 1). We argue that *multi-view consistency is an intrinsic property of HD maps*, because the rigid and static HD maps shouldn't change significantly after simply switching viewpoints. The violations of consistency arise from the fact that existing BEV perception

*Equal contribution.

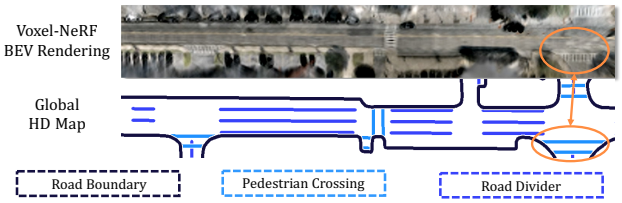

Figure 2: Our Voxel-NeRF reconstructs high-quality 3D structure of the scene. As in the rendering result, the lanes and pedestrian crossings (highlighted) are clear.

algorithms do not account for all the views explicitly and thus do not align their predictions. This issue further boils down to their *onboard* setting, where the models are only allowed to access computing devices *onboard* in autonomous vehicles and can only handle a single frame or a few neighboring frames.

Given such limitations of the *onboard* setup, we underline a critical yet under-explored *offboard* setup that removes the computation constraints. Our offboard setting aligns well with the *infrastructure* role of HD maps: constructing HD maps can and should utilize powerful data centers to maximize the fidelity of predictions, thus ensuring the safety and reliability of the virtually infinite usages of HD maps. By aggregating information from diverse viewpoints and enhancing consistency, our offboard generation provides a natural improvement. As shown in Fig. 1, having multiple views of a shared region offers richer geometric and semantic cues, as well as improves the completeness of scene understanding, particularly regarding frequent occlusions in urban traffic.

With HD map construction primarily relying on vision information to infer the semantics, which is different from the previous offboard studies depending on point clouds, we are the first to explore a vision-oriented offboard perception framework to our best knowledge. To this end, we propose *Multi-view Map (MV-Map)* that leverages information from every frame's viewpoint and generates a unified HD map consistent with all of them. In contrast to the *frame-centric* design in current onboard methods that merges a fixed number of frames on the input level, we propose a *region-centric* design inspired by "offboard 3D detection" [30] to fully utilize the data from diverse views. Notably, our design can connect every HD map region with an arbitrary number of input frames covering its area. The pipeline of our framework involves extracting all the HD map patches predicted by an off-the-shelf onboard model related to that HD map region, and then fusing the patches into a final result that agrees with all the views, as illustrated in the arrows in Fig. 1. To give more weight to reliable frames, such as those with a clear view of the target region, we introduce an "*uncertainty network*" as a key component, which assigns confidence scores to onboard results and performs a weighted average of HD map patches guided by the confidence.

We further enhance the consensus among all the frames by augmenting the uncertainty network with cross-view consistency information. Our key insight is to learn a coherent 3D structure from diverse views and provide it as an auxiliary input to the uncertainty network. For this purpose, we exploit neural radiance fields (NeRFs) [22], a state-of-the-art approach that represents 3D structures of scenes. As shown in Fig. 2, our NeRF model synthesizes a high-quality scene structure. Compared with alternative 3D reconstruction strategies like structure from motion (*e.g.*, COLMAP [34]), NeRF is more preferred from a practical perspective, because its runtime grows linearly with the frame number, whereas COLMAP increases quadratically. Moreover, NeRF is *fully self-supervised* and does not require additional annotations, unlike multi-view stereo methods such as MVSNet [44]. To improve the scalability of NeRF, we leverage a voxelized variant of NeRF (Voxel-NeRF) to promote efficiency and propose loss functions that implicitly guide the concentration of NeRF on the near-ground geometry related to HD map generation. Additionally, we highlight NeRF's flexibility and scalability to an arbitrary number of views, making it critical in offboard HD map generation.

To summarize, we make the following contributions:

1. We are the *first* to study the problem of learning to generate HD maps *offboard*, and we are also the first *vision-oriented* offboard study to our best knowledge.
2. We propose an effective *region-centric* framework MV-Map that can generate a multi-view consistent HD map from an arbitrarily large number of frames.
3. We introduce and extend Voxel-NeRF to encode the 3D structure from all frames for HD map generation tasks, further guiding the fusion for multi-view consistency.

Large-scale experiments on nuScenes [3] show that MV-Map significantly improves HD map quality. Notably, MV-Map can effectively utilize an increasing number of input frames, making it attractive for real-world applications.

2. Related Work

Offboard 3D perception. The need for large-volume training data encourages developing offboard algorithms. Existing studies mainly focus on predicting 3D bounding boxes [24, 28, 30, 43]. The most representative work on "offboard 3D detection" [30] extracts multi-frame point clouds in object tracks and refines the 3D bounding boxes with the "4D" data. Its success heavily relies on the absolute 3D positions of point clouds, where simply overlaying LiDAR points can construct denser surfaces of objects. However, in HD map generation that relies on images, it is not straightforward to accumulate imagery data in the 3D space. To overcome this limitation, we propose region-centric fu-

sion to aggregate multi-frame information and utilize multi-view reconstruction, *e.g.* NeRF, to encode global geometry. Our study is also the *first vision-oriented offboard* pipeline.

BEV segmentation and HD map construction. Onboard HD map construction is closely related to BEV segmentation, as in HDMapNet [13]. The major challenge in BEV segmentation is to map the image features to the 3D world. The conventional approach leverages inverse perspective warping [1, 2, 6, 27, 31]. BEV perception methods either apply attention to capture the transformation [15, 19], incorporate depth information [12, 14, 26, 29], or directly query the features from voxels [11]. To better support downstream applications, some recent methods [16, 18] have developed special decoders to generate vectorized HD maps. Unlike these *onboard* methods, our proposition is a general *offboard* pipeline that utilizes any off-the-shelf segmentation models as an internal component and refines its results with multi-view consistent fusion. In this sense, neural map prior [41] and NeMO [47] also propose to perform long-term temporal fusion, but they primarily focus on an onboard setting and cannot fully leverage the multi-view consistency from long video sequences. Compared with them, our MV-Map is an offboard framework and explicitly accounts for the geometry of diverse views by reconstructing the 3D structure of the scene.

Neural radiance fields. NeRF [22] has shown outstanding capability in 3D reconstruction. Recent work [20, 32, 37, 40] has extended NeRF into large unbounded scenes, such as city-scale NeRF with ego-centric camera settings [32, 37, 40] and improvement from depth-supervised methods [5, 25, 38, 42]. With NeRF's ability to optimize 3D structures from numerous views, it becomes an ideal method to enforce multi-view consistency for offboard perception. However, as we are the first to adapt NeRF for HD map generation, some important modifications are made. First, we adopt voxel-based NeRF [10, 17, 23, 35, 36] to accelerate the NeRF training by voxelizing the space and encoding the parameters for each position in the voxels. This allows us to reconstruct a huge scene from nuScenes within minutes. In addition, we propose a "total-variance loss" to enhance NeRF's concentration on the near-ground geometry, which also reflects the shift of concentration from pixel quality to downstream HD map generation.

3. Offboard HD Map Generation

Given a sequence of sensory data, the goal of HD map generation is to predict the positions and semantics of road elements in the BEV space, including road dividers, road boundaries, and pedestrian crossings.

Problem statement. We consider the input of HD map generation as $\mathcal{D} = \{(I_i, P_i)\}_{i=1}^{N}$, where I_i denotes the i-th sensor frame, P_i is the set of associated sensor poses, and N is the total number of frames in the database representing diverse views of a scene captured by a moving ego vehicle. The output is denoted as $\mathcal{M} = \{M_i\}_{i=1}^{N}$, where M_i is the HD map for the region nearby the ego vehicle on frame i. Following HDMapNet [13], we define M_i as a local semantic map on BEV. Note that the aforementioned formulation is agnostic to sensor types. In the main paper, we mainly focus on *vision-oriented* HD map generation, and we extend it to leveraging additional LiDAR data in Sec. 5.6. Specifically, every frame I_i contains $K = 6$ RGB images $\{I_{i,j}\}_{j=1}^{K}$ on nuScenes [3], and $P_i = \{P_{i,j}\}_{j=1}^{K}$ comprises of the intrinsic and extrinsic matrices of corresponding cameras.

Offboard vs. onboard settings. Compared with the conventional onboard setup, our offboard setup offers greater flexibility in terms of speed and computation resources. Onboard HD map generation algorithms are often constrained by efficiency requirements and cannot use all the N frames *in a single run*. By contrast, offboard algorithms are allowed to have access to all the N frames, and can then leverage the offline setting and abundant computation resources to generate HD maps of higher quality.

From *frame-centric* to *region-centric* designs. There are different strategies to utilize the temporal data from N frames, similar to offboard 3D detection [30]. A direct solution is *frame-centric* [30], in which we naïvely increase the number of frames for existing *onboard* HD map construction methods, typically BEV segmentation models, and extend them to long sequences. While previous work [13, 15] has illustrated the benefit of longer temporal horizons, a multi-frame BEV segmentation model can only handle a fixed number of input frames, and increasing the frame number requires a linear growth in GPU capacity. Therefore, simply scaling up the input frames of existing onboard models is not an effective way of exploiting the offboard data, which often have varying and large frame numbers.

To overcome the limitations of the frame-centric design, we propose a novel *region-centric* design that adaptively aggregates information from an arbitrary number of available frames for each HD map region. Our design is inspired by the *object-centric* notion in 3D detection [30], but extends to the task of HD map construction. Doing so enables the consensus across frames captured from different viewpoints.

4. Method: Multi-view Map

Overview. Fig. 3 illustrates the overall framework of our Multi-view Map (MV-Map). An onboard HD map model processes every frame (I_i, P_i) and generates its corresponding BEV feature map F_i and HD map semantics S_i (Sec. 4.1). Then, an uncertainty network assesses the reliability of the single-frame information F_i for every region on the HD map (Sec. 4.2). Meanwhile, a voxelized NeRF

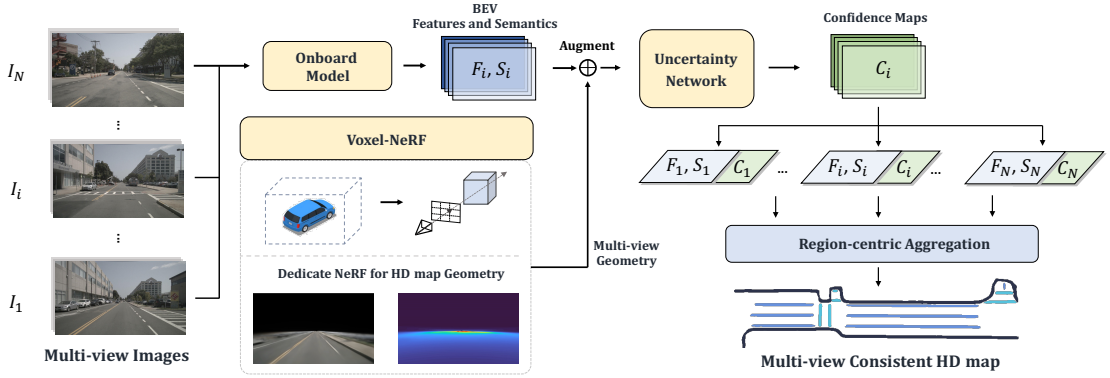

Figure 3: Offboard pipeline of MV-Map. Given an arbitrary number of input frames, MV-Map first leverages an off-the-shelf *onboard model* to generate BEV features and semantic maps for each frame. Then an *uncertainty network* predicts their corresponding confidence maps and guides the region-centric aggregation of a unified HD map. Our pipeline further develops a Voxel-NeRF tailored to 3D structures related to HD maps to augment MV-Map with multi-view geometry.

f_{NeRF} optimizes a global 3D structure from all N frames and provides multi-view consistency information to the uncertainty network (Sec. 4.3). The final prediction for every region on the HD map is produced by a weighted average of the single-frame semantics S_i, which enables handling an arbitrary number of frames.

4.1. Onboard Model

The onboard model is the entry point of our pipeline. Most existing HD map generation methods follow an encoder-decoder design. The encoder generates a BEV feature map F_i from the input (I_i, P_i) as $\texttt{Encoder}(I_i, P_i) \rightarrow F_i$, and the decoder converts the feature map F_i into a semantic map S_i as $\texttt{Decoder}(F_i) \rightarrow S_i$.

Since our pipeline only requires the BEV feature map F_i to activate the subsequent modules, MV-Map is agnostic to specific encoder-decoder designs. Without the loss of generality, we mainly adopt the encoder in SimpleBEV [11] and use a lightweight convolutional decoder. Results based on additional models are in Table D (Supplementary).

Encoder. For each frame, a convolutional backbone first converts K images $\{I_{i,j}\}_{j=1}^{K}$ into 2D image feature maps $\{F_{i,j}^{\text{2D}}\}_{j=1}^{K}$. The features are then *lifted* into the 3D world, through a set of voxels that are pre-defined by the encoder with shape $X \times Y \times Z$ centered around the ego vehicle: the 2D features are bi-linearly sampled for every voxel based on their projected locations on the images, leading to a voxelized 3D feature map F_i^{3D}. Finally, reducing the Z-axis of F_i^{3D} produces a BEV feature map F_i with shape $X \times Y \times C$, where C is the feature dimension.

Decoder. Our decoder is a fully-convolutional segmentation head that predicts the logits of semantics from every BEV grid in F_i. It generates the surrounding HD map as the semantic segmentation result S_i with shape $X \times Y$.

Region-centric extension. Our *region-centric* design considers each BEV grid as an HD map region. If a grid is covered in N' frames, it receives N' features and predictions from different viewpoints. MV-Map then fuses information from N' view-specific frames to create a multi-view consistent feature for this region, detailed as below.

4.2. Global Aggregation via Uncertainty Network

Region-centric uncertainty-aware fusion. Our region-centric offboard pipeline learns to aggregate the N frames of independent HD map predictions $\{S_i\}_{i=1}^{N}$ into a multi-view consistent prediction for each region. Our key design is to introduce an *uncertainty network*. For the HD map predictions from all viewpoints, the uncertainty network assigns a confidence score to each BEV grid, resulting in $N \times X \times Y$ scores that reflect the pairwise reliability of a viewpoint contributing to an HD map region. Specifically, the uncertainty network takes the BEV features $\{F_i\}_{i=1}^{N}$ as input and generates the confidence maps $\{C_i\}_{i=1}^{N}$, with C_i of shape $X \times Y$. In Sec. 4.3, we will describe how we further incorporate global geometry encoded by Voxel-NeRF into the uncertainty network. The architecture of our uncertainty network is illustrated in Fig. 4, adopting a UNet-like [33] structure for predicting densely on every voxel.

We then aggregate the per-frame semantics and confidences into a final HD map. Suppose an arbitrary target position (x^w, y^w) is specified in the world coordinate system, we transform it to the local coordinate system of every frame with poses $\{P_i\}_{i=1}^{N}$ and sample the semantic maps and confidence scores at corresponding locations. Finally, the prediction for (x^w, y^w) is obtained by a weighted average of per-frame semantics according to their confidences.

KL-divergence loss for enhanced uncertainty learning. In addition to generating confidence scores, we add a multi-layer perceptron (MLP) head to the uncertainty network to infer the KL-divergence between the predicted and ground truth semantics, as shown in Fig. 4. Intuitively, learning to

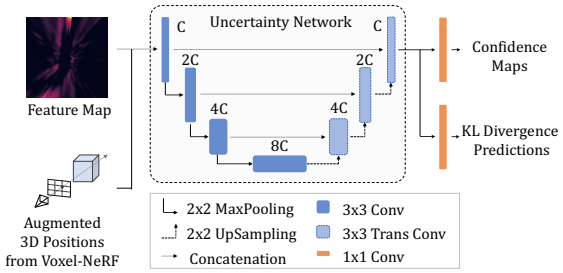

Figure 4: Architecture of **uncertainty network**, which takes as input the feature map and the augmented 3D positions from NeRF for each voxel (Sec. 4.3). It outputs the confidence maps for region-centric fusion and, optionally, the predicted KL-divergence for the KL-divergence loss (detailed in Sec. 4.2).

regress the KL-divergence value between the predicted and ground truth HD maps augments the confidence score of the uncertainty network, because a smaller KL-divergence value indicates better quality of map construction and the uncertainty network should assign higher confidence accordingly. During training, we encourage the inferred divergence KL^U to be close to the true divergence KL^G between semantics S_i and S_i's ground truth, formally,

$$\mathcal{L}_{\mathrm{KL}} = \frac{1}{XY}\sum_{x=1}^{X}\sum_{y=1}^{Y} \|\mathrm{KL}^G[x,y] - \mathrm{KL}^U[x,y]\|_2^2. \quad (1)$$

We train the uncertainty network with both a cross-entropy loss between the fusion result and the ground truth semantics at each location (x^w, y^w) [13] and our auxiliary KL-divergence loss in Eqn. 1. Given that the weighted average operation is differentiable, the gradients from both of the loss terms can be back-propagated to the confidence scores for updating the uncertainty network.

4.3. Voxel-NeRF for Multi-view Consistency

MV-Map further leverages a voxelized NeRF to effectively construct a *unified* 3D structure of the scene from the N frames, which is incorporated with the uncertainty network to improve the multi-view consistency of HD maps.

Voxel-NeRF for traffic scenes. NeRF [22] represents a 3D scene as a continuous function $f_{\mathrm{NeRF}} : (\mathbf{x}, \theta) \rightarrow (\mathbf{c}, \sigma)$, which maps every point x in the 3D space to its color c and density σ, relative to the viewing direction θ. By explicitly encoding camera projection in the neural rendering process, the learned NeRF model f_{NeRF} encodes the 3D geometry of the corresponding scene from input images. Despite the success of the vanilla NeRF, applying it to autonomous driving datasets poses significant challenges, because of the unbounded nature of the scenes and the huge quantities of data involved (*e.g.*, 850 scenes on nuScenes [3]). Therefore, we introduce a voxelized NeRF based on DVGO [36] for better training speed and scalability. Our Voxel-NeRF captures multi-view consistent geometry for outdoor scenes, instead of small objects as in conventional NeRFs. To achieve this, we initialize voxel grids with shape $X_s \times Y_s \times Z_s$ to cover the entire scene, which is larger than $X \times Y \times Z$ used in onboard models for single-frame areas. For each camera ray, the neural rendering operation in Voxel-NeRF *concurrently* queries every voxel intersected by the ray. This concurrent querying of voxels significantly accelerates the training of our Voxel-NeRF, *from hours to minutes* for any scene in nuScenes, enabling a reasonable computation budget for MV-Map. More details can be found in Sec. 5.

Augmenting uncertainty network. Conceptually, the predicted semantics at a position is more reliable when it resides on the object surfaces. Once NeRF produces a multi-view consistent structure, we can compute the distance between each voxel center and its closest surface. Such a clue can be exploited to evaluate the reliability of semantic maps. To this end, for an arbitrary (x, y) on BEV, we first recover all the voxel center locations at the BEV coordinate (x, y) as $L_{\mathrm{Voxel}} = \{(x, y, z_i)\}_{i=1}^{Z}$ and then compute their corresponding pixel locations on the images $L_{\mathrm{Image}} = \{(x_i^p, y_i^p\}_{i=1}^{Z}$. By volume rendering along the camera rays crossing these pixels, f_{NeRF} reconstructs the 3D positions of these pixels, denoted as $L_{\mathrm{NeRF}} = \{(x_i^R, y_i^R, z_i^R)\}_{i=1}^{Z}$, which are generally the intersections between their camera rays and surfaces. (Ray casting details are in Sec. B.2 (Supplementary).) By calculating $\Delta L_{\mathrm{N/V}} = \{(x_i^R - x, y_i^R - y, z_i^R - z_i)\}_{i=1}^{Z}$, we assess the consistency between voxel centers and the global 3D structure. Finally, we employ an MLP upon $\Delta L_{\mathrm{N/V}}$ and concatenate its output with the BEV feature, which is then used as the augmented input to the uncertainty network, which is the "augmented 3D positions" in Fig. 4.

Dedicating NeRF for HD maps with total-variance loss. Note that *our objective is to facilitate HD map generation, rather than optimizing rendering quality*. With the majority of HD map elements situated on the ground, we modify the NeRF to focus less on the quality of pixels in the air. To this end, we introduce a simple yet effective "*total-variance loss*" that guides the optimization of near-ground geometry *implicitly*. This total-variance loss $\mathcal{L}_{\mathrm{TV}}$ is obtained by accumulating the total-variance $\mathrm{TV}(\cdot)$ at each BEV position:

$$\mathcal{L}_{\mathrm{TV}} = -\frac{1}{X_s Y_s}\sum_{x=1}^{X_s}\sum_{y=1}^{Y_s} \mathrm{TV}(x,y). \quad (2)$$

Here the total-variance $\mathrm{TV}(\cdot)$ is defined as the L2-norm of the differences of occupancies along the Z-axis, given by

$$\mathrm{TV}(x,y) = \|O[x,y,2\!:\!Z_s] - O[x,y,1\!:\!Z_s-1]\|_2, \quad (3)$$

where $O[x, y, z]$ represents the density of voxel (x, y, z) predicted by NeRF and $\|\cdot\|_2$ denotes the L2-norm.

Table 1: Comparison with state-of-the-art vision-based HD map generation methods on nuScenes [3]. "*" means the results reported in HDMapNet [13]. "Average Fusion" is an offboard baseline explained in Sec. 5.2. The quantitative results indicate that our MV-Map has significant benefits to HD map generation and outperforms offboard baseline approaches.

Setup	Method	mIoU (Short-range / Long-range)			
		Divider	Ped Crossing	Boundary	All
Onboard	IPM(B)*	25.5 / -	12.1 / -	27.1 / -	21.6 / -
	IPM(B+C)*	38.6 / -	19.3 / -	39.3 / -	32.4 / -
	VPN*	36.5 / -	15.8 / -	35.6 / -	29.3 / -
	Lift-Splat-Shoot [29]*	38.3 / -	14.9 / -	39.3 / -	30.8 / -
	HDMapNet [13]	40.6 / 33.9	18.7 / 19.4	39.5 / 34.9	32.9 / 29.4
	Onboard Model (Ours)	46.4 / 39.3	29.7 / 26.4	48.1 / 39.1	41.4 / 35.0
Offboard	Average Fusion	48.86 / 42.83	31.55 / 24.75	51.98 / 43.91	44.13 / 37.16
	MV-Map (Ours)	**50.87 / 48.15**	**34.52 / 33.34**	**55.64 / 50.28**	**47.01 / 43.92**

We emphasize the "negative" sign in Eqn. 2. It indicates "maximizing" the variance, because an accurate ground plane has a *peak* distribution of voxel occupancy on the Z-axis instead of a *uniform* one. TV-loss enables Voxel-NeRF to assign larger densities to the ground plane than transient objects, leading to high-quality 3D structures as in Fig. 2.

4.4. Training and Inference

The procedure of our offboard pipeline follows three steps: (1) we adopt an existing onboard model, (2) train Voxel-NeRF on sequences, and (3) train and infer the uncertainty network. We describe these steps in order and leave detailed configurations in Sec. G (Supplementary).

Onboard model. As MV-Map is agnostic to the choice of onboard models (Sec. 4.1), here we adopt an *off-the-shelf* BEV segmentation model and freeze its parameters during both training and inference stages of the offboard pipeline.

Voxel-NeRF. We train the Voxel-NeRF for all sequences in our training and validation datasets, using both conventional photometric loss and our total-variance loss (Sec. 4.3). Note that our NeRF training is entirely *self-supervised and does not require any annotations*.

Uncertainty network. The *region-centric* design enables the uncertainty network to handle varying frame numbers of offboard data. In practice, however, the GPU capacities and batching during training limit the network to a fixed and restricted frame number. To overcome this issue, we adopt the solution from video-based tasks (*e.g.*, 2D multi-object tracking [21, 46]), where models are trained on *short video clips* but are inferred iteratively on *unbounded sequences*.

Similarly, given the input N frames of a scene, the uncertainty network is trained with samples containing $M (M < N)$ adjacent frames to fit into limited GPU memory. The loss is a weighted sum of our KL-divergence loss and a BEV segmentation loss (Sec. 4.2). During inference, we apply the uncertainty network to all the N frames independently and use region-centric aggregation to fuse single-frame semantics into a unified HD map.

5. Experiments

5.1. Dataset and Implementation Details

Dataset. We conduct experiments on a large-scale autonomous driving dataset: nuScenes [3]. It contains 850 videos with 28,130 and 6,019 frames for training and validation, respectively. On each timestamp, six surrounding cameras collect high-resolution images as input.

Evaluation metrics. Following prior work [13], we compute the intersection-over-union (IoU) for HD map categories: divider, pedestrian crossing, and road boundaries. To highlight the challenge of predicting a scene-scale HD map, our evaluation adopts both a *short-range* setting [13] covering 60m $\times$ 30m and a new *long-range* setting covering 100m $\times$ 100m, which aligns with the common perception range in self-driving [7, 9]. Without further mentioning, we conduct our ablation studies under the more challenging long-range setting.

Implementation details. We follow the training and inference settings in Sec. 4.4 and discuss the details in Sec. G (Supplementary). We emphasize that MV-Map is scalable to large volumes of offboard data. Within 15 minutes on a single A40 GPU, our Voxel-NeRF can optimize the 3D structure from each nuScenes sequence, which typically has over 1k images covering regions with an average length of $\sim$300m, less than 1 second per frame. In comparison, the common multi-view stereo baseline of COLMAP [34] may take several hours or even days. Notably, this is because COLMAP spent most of the time in feature matching, which is pairwise across frames and $\mathcal{O}(N^2)$ regarding the frame number N. In comparison, the time complexity of NeRF is $\mathcal{O}(N)$. Moreover, our uncertainty network is trained on the samples with $M = 5$ adjacent frames to fit into our GPU memory, but it can jointly handle all the frames ($\sim$40) in a nuScenes sequence during the inference stage, as explained in Sec. 4.4.

5.2. Comparison with State-of-the-Art Methods

As our work represents the *first* study on offboard HD map generation, there are no readily available competing

Table 2: MV-Map significantly improves vectorized HD map quality over both the onboard VectorMapNet [18] model and the recent temporal fusion method Neural Map Prior [41] (NMP). († means the performance reported in NMP; ‡ means the performance from VectorMapNet's officially released checkpoint.)

Methods	mAP			
	Divider	Ped Crossing	Boundary	All
VectorMapNet†	47.3	36.1	39.3	40.9
+ NMP	49.6	42.9	41.9	44.8
ΔmAP	+2.3	+6.8	+2.6	+3.9
VectorMapNet‡	47.7	39.8	38.9	42.1
+MV-Map	**55.0**	**46.2**	**45.5**	**48.9**
Δ mAP	**+7.3**	**+6.4**	**+6.6**	**+6.7**

methods. Additionally, our MV-Map can utilize any off-the-shelf onboard model as its internal component. To ensure a meaningful and fair comparison, we organize the experimental results and analysis in Table 1 as follows.

First, our onboard model adopts the simple-yet-effective design from SimpleBEV [11]. As shown in the "onboard" rows, our onboard model already *consistently* outperforms previous baselines in both short-range and long-range settings. Second, our MV-Map brings a significant improvement of ∼7% mIoU compared with our already effective onboard model. Notably, our offboard method is better than HDMapNet [13] by around 50% with over 15% IoU increase on all the categories. Finally, we develop an offboard baseline algorithm called "Average Fusion." It does not consider the quality of different viewpoints and performs region-centric aggregation by equally averaging the single-frame semantic maps. Compared with "Average fusion," our MV-Map still improves the HD map quality by a large margin of over ∼7% mIoU under the long-range setting.

5.3. Comparison on Vectorized HD Maps

We further extend our MV-Map to vectorized HD map generation to demonstrate its wide compatibility. In Table 2, we apply MV-Map to VectorMapNet [18], one of the state-of-the-art onboard models generating vectorized HD maps. Specifically, our MV-Map performs uncertainty-aware fusion on the frozen BEV features of an open-sourced VectorMapNet model. The results show that MV-Map improves the HD Map generation quality by a large margin of over ∼7 mAP, proving its effectiveness in constructing high-quality HD maps. We further compare MV-Map with the recent neural map prior (NMP) [41], which also capitalizes long-term temporal fusion for global HD map generation in Table 2.[1] In the experiments, We adopt the same metric (mAP) as VectorMapNet [18]. As clearly illustrated, our offboard MV-Map pipeline is better in both absolute (48.9 mAP vs. 44.8 mAP) and relative improvements (6.8 mAP vs. 3.9 mAP).

[1]NMP is not included for semantic HD map comparison in Sec. 5.2 because it adopts a different resolution for rasterized HD maps.

Table 3: The components in MV-Map effectively improve HD map generation step by step. We analyze the uncertainty network (UN) and KL-divergence loss ($\mathcal{L}_{\text{KL}}$) discussed in Sec. 4.2, and Voxel-NeRF (NeRF) and total-variance loss ($\mathcal{L}_{\text{TV}}$) discussed in Sec. 4.3.

ID	Offboard Component				mIoU			
	UN	$\mathcal{L}_{\text{KL}}$	NeRF	$\mathcal{L}_{\text{TV}}$	Divider	Crossing	Boundary	All
1	Onboard Model				39.30	26.44	39.10	34.95
2	Average fusion				42.83	24.75	43.91	37.16
3	✓				46.90	30.30	49.07	42.09
4	✓	✓			47.38	31.11	49.53	42.67
5	✓	✓	✓		47.64	32.36	49.67	43.22
6	✓		✓	✓	48.01	32.65	50.12	43.59
7	✓	✓	✓	✓	**48.15**	**33.34**	**50.28**	**43.92**

5.4. Ablation Studies

MV-Map Components. We quantify the improvement from each offboard module in Table 3. (1) **Region-centric fusion baseline.** Beginning from the onboard model (row 1), we first apply average fusion (row 2) to it (discussed in Sec. 5.2) as a baseline. The improvement indicates that our region-centric design indeed helps by fusing numerous frames into a unified HD map. (2) **Uncertainty network.** Replacing the average fusion (row 2) with the uncertainty network (row 3) enables larger contributions from more reliable frames and the ∼5% increase in mIoU validates that assessing quality is critical for better HD map quality. (3) **KL-divergence loss.** The ∼0.5% mIoU on using KL-divergence loss or not (row 3 and row 4) supports that explicitly supervising the uncertainty network with KL-divergence effectively improves the region-centric fusion to rely on frames with higher quality. (4) **Voxel-NeRF.** Adding NeRF to a full-fledged uncertainty network further improves the mIoU (row 4 and row 5). In the category-level analysis, we highlight that NeRF is critical for the fusion *especially on the challenging structures with smaller regions*, *e.g.*, pedestrian crossings. This evidence validates the importance of global geometry in multi-view consistency. (5) **Total-variance loss.** Utilizing it further boosts the performance in all the scenarios, validating our effort to dedicate NeRFs for the downstream HD map generation.

Scaling to more frames. We demonstrate that our fusion strategy can handle and significantly benefit from a larger number of frames, which is critical for offboard HD map generation. We evaluate our offboard framework under varied input frames in Fig. 5. MV-Map can utilize all the keyframes (40 frames) in nuScenes and this number is only bounded by the sequence length. As clearly shown in the blue curve of Fig. 5, MV-Map benefits from more frames, indicating its *scalability* for offboard scenarios, especially compared with the average fusion baseline, whose performance drops after using more than 15 frames. This indicates that our region-centric fusion strategy is able to

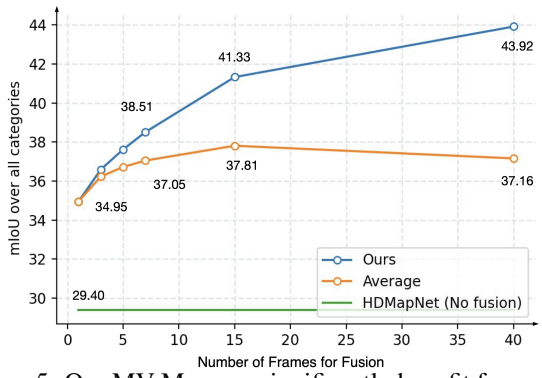

Figure 5: Our MV-Map can significantly benefit from more input frames, which is attractive for offboard applications. Notably, the performance of the "average fusion" baseline saturates and even decreases with more input frames.

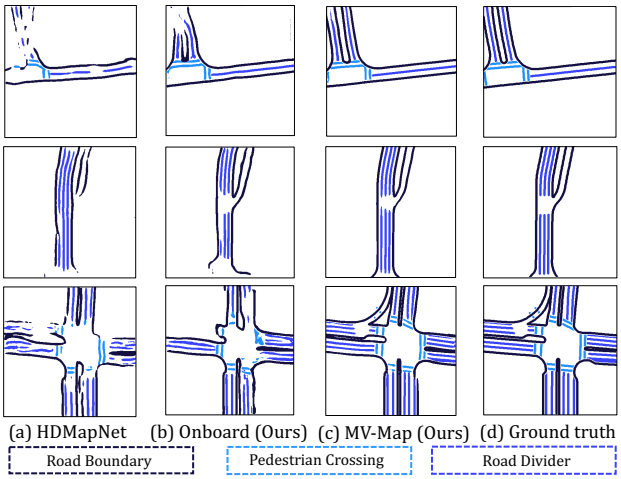

Figure 6: Qualitative comparison in the long-range settings. HD map generated offboard has significantly better quality by fixing the artifacts of the onboard model.

reason the complementary regions among the frames, instead of blindly averaging them all.

Qualitative comparison. We visualize the generated HD maps in Fig. 6. As clearly shown, MV-Map corrects the artifacts from onboard models and achieves better completeness and details. In addition, the HD maps generated offboard have high fidelity compared with the ground truth, especially in the center regions covered by more frames.

Analyzing KL-divergence and confidence scores. We empirically analyze the output of the uncertainty network in Fig. 7. As for the confidence scores, we observe that they indeed decrease the contributions of unreliable regions, such as the part with occlusion in Fig. 7a, highlighted with solid circles. The invisible area has much smaller confidence than its nearby regions. Additionally, we transfer the KL-divergence prediction head to the validation set and find the predictions reasonably correlate with the confidence scores,

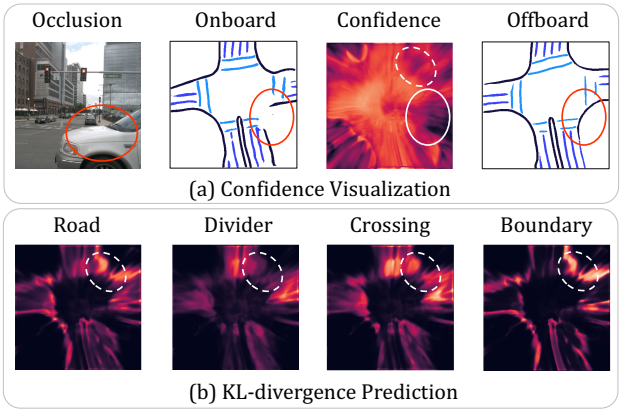

Figure 7: **(a)** The confidence scores predicted by the uncertainty network *capture the challenging regions* (solid circles) as the occluded region has significantly smaller confidence values than its nearby regions. Darker colors indicate smaller confidence values. **(b)** Predicted KL divergence between the prediction and ground truth label captures the connection with the predicted confidence scores (dashed circles). The region with a dashed circle has a much larger predicted KL-divergence value than its nearby regions. Accordingly, this area exhibits smaller confidence scores. Darker colors mean smaller KL-divergence values.

as in Fig. 7b. We notice that the regions with higher KL-divergence values (Fig. 7b) also have lower confidences (Fig. 7a), highlighted with the dashed circles.

Using geometric information from data-driven priors. Our Voxel-NeRF offers geometric information in a *fully self-supervised* manner. Meanwhile, our MV-Map framework is general and can leverage alternative approaches for providing geometric information, such as *learning data-driven priors* from large-scale datasets. We investigate this type of approach here and consider representative monocular depth estimators that are learned off-the-shelf in a *supervised* manner. Specifically, we replace the rendering process of Voxel-NeRF with the results from *NeWCRFs* [45] (details in Sec. G, Supplementary). As in Table 4, monocular depth can improve the uncertainty fusion as well (row 2 and row 3). We further notice that NeRF performs slightly better because it *encodes multiple views consistently in a shared 3D structure*, while monocular depth considers each view independently and suffers from scale variation across frames. Encouraged by the benefits of these two distinct types of geometric information, future work is to combine NeRF with learnable priors into our framework.

5.5. Globally Consistent HD Map Generation

Our offboard MV-Map can handle numerous frames. Its application is to expand the range of HD map generation from a local region around the ego-vehicle to a global re-

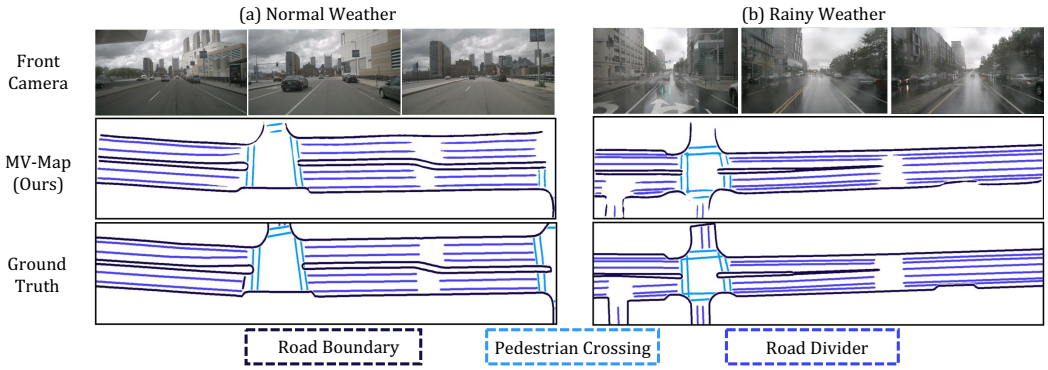

Figure 8: Visualization of a **unified, scene-scale HD map** through our MV-Map. It can fuse numerous frames and generate global-scale HD maps with high quality. It is also *robust* under different weather conditions and complex road topology.

Table 4: MV-Map is a general framework: In addition to NeRF, MV-Map can benefit from data-driven priors of geometric information; here we use monocular depth estimation [45] as an example. "UN-Only:" using the uncertainty network without augmentation of 3D structural information. Then we separately incorporate mono-depth or NeRF to it.

Methods	mIoU			
	Divider	Ped Crossing	Boundary	All
Average Fusion	42.83	24.75	43.91	37.16
MV-Map (UN-Only)	47.38	31.11	49.53	42.67
MV-Map (Mono-Depth)	48.04	32.96	50.08	43.69
MV-Map (NeRF)	**48.15**	**33.34**	**50.28**	**43.92**

Table 5: Performance of incorporating the *LiDAR modality* in MV-Map, evaluated under the long-range setting on the validation set. As a *general* framework, MV-Map can exploit multi-modality as input and improve the performance consistently and significantly.

Methods	mIoU (Long-range)			
	Divider	Ped Crossing	Boundary	All
Onboard	41.63	27.13	41.65	36.80
MV-Map	**50.72**	**32.99**	**54.63**	**46.11**

gion covering all the input frames, which saves the labor in stitching multiple local predictions in the real world. Our global maps in Fig. 8 demonstrate high fidelity for complex topology in two challenging scenes. While some regions do not match the ground truth, we argue that these regions fall outside the collected frames and perception ranges, which are beyond the scope of offboard algorithms.

5.6. Generalizability of MV-Map with LiDAR

To demonstrate the generalizability of our framework, we analyze incorporating the LiDAR sensor into MV-Map. In the main paper, we focused on cameras, because they contribute primarily to HD map generation as shown in HDMapNet [13]. On the other hand, LiDARs are also widely used for their accurate distance sensing and their ability to enhance localization, which motivates our investigation here. We first describe the design of utilizing the extra LiDAR modality and then analyze the results. Extra details are explained in Sec. C (Supplementary).

6. Conclusion

Regarding the infrastructure role of HD maps, we propose a novel *offboard* HD map generation setup to address the unreliability of *onboard* BEV perception. By removing the computation constraints, the models are allowed to reason all the frames altogether and construct multi-view consistent HD maps. Concretely, we propose an offboard HD map generation framework called MV-Map. To address numerous frames, MV-Map designs region-centric aggregation to unify the HD maps from all the frames. The key design is an uncertainty network that weighs the contribution of different frames and utilizes a Voxel-NeRF to provide multi-view consistent 3D structural information. Experiments validate that MV-Map is scalable to large volumes of offboard data and significantly improves the HD map quality. We hope that our framework can become an effective augmentor for onboard algorithms and also inspire future research on offboard problems.

Limitations and future work. Although our Voxel-NeRF improves the offboard pipeline in a scalable way, several challenges still present, including moving objects in traffic scenes and exploiting data-driven priors for better geometric information. In addition, we seek to connect our work with auto-labeling and compare it with human annotation quality, so as to explore more potential applications such as autonomous vehicle navigation and urban planning.

Acknowledgement. This work was supported in part by NSF Grant 2106825, NIFA Award 2020-67021-32799, the Jump ARCHES endowment, the NCSA Fellows program, the Illinois-Insper Partnership, and the Amazon Research Award. This work used NVIDIA GPUs at NCSA Delta through allocations CIS220014 and CIS230012 from the ACCESS program.

References

[1] Mohamed Aly. Real time detection of lane markers in urban streets. In *IEEE Intelligent Vehicles Symposium*, 2008. 3

[2] Massimo Bertozzi and Alberto Broggi. Real-time lane and obstacle detection on the GOLD system. In *IEEE Intelligent Vehicles Symposium*, 1996. 3

[3] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020. 1, 2, 3, 5, 6

[4] Ming-Fang Chang, John W Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, and James Hays. Argoverse: 3D tracking and forecasting with rich maps. In *CVPR*, 2019. 1

[5] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised NeRF: Fewer views and faster training for free. In *CVPR*, 2022. 3

[6] Liuyuan Deng, Ming Yang, Hao Li, Tianyi Li, Bing Hu, and Chunxiang Wang. Restricted deformable convolution-based road scene semantic segmentation using surround view cameras. In *ITSC*, 2020. 3

[7] Hao Dong, Xianjing Zhang, Xuan Jiang, Jun Zhang, Jintao Xu, Rui Ai, Weihao Gu, Huimin Lu, Juho Kannala, and Xieyuanli Chen. SuperFusion: Multilevel LiDAR-camera fusion for long-range HD map generation and prediction. *arXiv preprint arXiv:2211.15656*, 2022. 6

[8] Scott Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhao, Sabeek Pradhan, Yuning Chai, Ben Sapp, Charles R. Qi, Yin Zhou, Zoey Yang, Aur'elien Chouard, Pei Sun, Jiquan Ngiam, Vijay Vasudevan, Alexander McCauley, Jonathon Shlens, and Dragomir Anguelov. Large scale interactive motion forecasting for autonomous driving: The Waymo open motion dataset. In *ICCV*, 2021. 1

[9] Lue Fan, Yuxue Yang, Feng Wang, Naiyan Wang, and Zhaoxiang Zhang. Super sparse 3D object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 6

[10] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *CVPR*, 2022. 3

[11] Adam W Harley, Zhaoyuan Fang, Jie Li, Rares Ambrus, and Katerina Fragkiadaki. Simple-BEV: What really matters for multi-sensor BEV perception? In *ICRA*, 2023. 1, 3, 4, 7

[12] Anthony Hu, Zak Murez, Nikhil Mohan, Sofía Dudas, Jeffrey Hawke, Vijay Badrinarayanan, Roberto Cipolla, and Alex Kendall. FIERY: Future instance prediction in Bird's-Eye View from surround monocular cameras. In *ICCV*, 2021. 3

[13] Qi Li, Yue Wang, Yilun Wang, and Hang Zhao. HDMapNet: An online hd map construction and evaluation framework. In *ICRA*, 2022. 1, 3, 5, 6, 7, 9

[14] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. BEVDepth: Acquisition of reliable depth for multi-view 3D object detection. In *AAAI*, 2023. 3

[15] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. BEVFormer: Learning Bird's-Eye-View representation from multi-camera images via spatiotemporal transformers. In *ECCV*, 2022. 1, 3

[16] Bencheng Liao, Shaoyu Chen, Xinggang Wang, Tianheng Cheng, Qian Zhang, Wenyu Liu, and Chang Huang. MapTR: Structured modeling and learning for online vectorized HD map construction. In *ICLR*, 2023. 3

[17] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. In *NeurIPS*, 2020. 3

[18] Yicheng Liu, Yuan Yuantian, Yue Wang, Yilun Wang, and Hang Zhao. VectorMapNet: End-to-end vectorized HD map learning. In *ICML*, 2023. 3, 7

[19] Zhijian Liu, Haotian Tang, Alexander Amini, Xingyu Yang, Huizi Mao, Daniela Rus, and Song Han. BEVFusion: Multi-task multi-sensor fusion with unified Bird's-Eye View representation. In *ICRA*, 2023. 3

[20] Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. NeRF in the wild: Neural radiance fields for unconstrained photo collections. In *CVPR*, 2021. 3

[21] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. TrackFormer: Multi-object tracking with transformers. In *CVPR*, 2022. 6

[22] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 2, 3, 5

[23] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 2022. 3

[24] Mahyar Najibi, Jingwei Ji, Yin Zhou, Charles R Qi, Xinchen Yan, Scott Ettinger, and Dragomir Anguelov. Motion inspired unsupervised perception and prediction in autonomous driving. In *ECCV*, 2022. 2

[25] Thomas Neff, Pascal Stadlbauer, Mathias Parger, Andreas Kurz, Joerg H. Mueller, Chakravarty R. Alla Chaitanya, Anton S. Kaplanyan, and Markus Steinberger. DONeRF: Towards real-time rendering of compact neural radiance fields using depth oracle networks. In *Computer Graphics Forum*, 2021. 3

[26] Mong H. Ng, Kaahan Radia, Jianfei Chen, Dequan Wang, Ionel Gog, and Joseph E Gonzalez. BEV-Seg: Bird's Eye View semantic segmentation using geometry and semantic point cloud. *arXiv preprint arXiv:2006.11436*, 2020. 3

[27] Bowen Pan, Jiankai Sun, Ho Yin Tiga Leung, Alex Andonian, and Bolei Zhou. Cross-view semantic segmentation for sensing surroundings. *IEEE Robotics and Automation Letters*, 2020. 3

[28] Ziqi Pang, Zhichao Li, and Naiyan Wang. Model-free vehicle tracking and state estimation in point cloud sequences. In *IROS*, 2021. 2

[29] Jonah Philion and Sanja Fidler. Lift, Splat, Shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3D. In *ECCV*, 2020. 3, 6

[30] Charles R Qi, Yin Zhou, Mahyar Najibi, Pei Sun, Khoa Vo, Boyang Deng, and Dragomir Anguelov. Offboard 3D object detection from point cloud sequences. In *CVPR*, 2021. 2, 3

[31] Lennart Reiher, Bastian Lampe, and Lutz Eckstein. A Sim2Real deep learning approach for the transformation of images from multiple vehicle-mounted cameras to a semantically segmented image in Bird's Eye View. In *ITSC*, 2020. 3

[32] Konstantinos Rematas, Andrew Liu, Pratul P. Srinivasan, Jonathan T. Barron, Andrea Tagliasacchi, Tom Funkhouser, and Vittorio Ferrari. Urban radiance fields. In *CVPR*, 2022. 3

[33] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015. 4

[34] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 2, 6

[35] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *CVPR*, 2022. 3

[36] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Improved direct voxel grid optimization for radiance fields reconstruction. *arXiv preprint arXiv:2206.05085*, 2022. 3, 5

[37] Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul Srinivasan, Jonathan T. Barron, and Henrik Kretzschmar. Block-NeRF: Scalable large scene neural view synthesis. In *CVPR*, 2022. 3

[38] Yi Wei, Shaohui Liu, Yongming Rao, Wang Zhao, Jiwen Lu, and Jie Zhou. NerfingMVS: Guided optimization of neural radiance fields for indoor multi-view stereo. In *ICCV*, 2021. 3

[39] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, Deva Ramanan, Peter Carr, and James Hays. Argoverse 2: Next generation datasets for self-driving perception and forecasting. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS Datasets and Benchmarks 2021*, 2023. 1

[40] Ziyang Xie, Junge Zhang, Wenye Li, Feihu Zhang, and Li Zhang. S-NeRF: Neural radiance fields for street views. In *ICLR*, 2023. 3

[41] Xuan Xiong, Yicheng Liu, Tianyuan Yuan, Yue Wang, Yilun Wang, and Hang Zhao. Neural map prior for autonomous driving. In *CVPR*, 2023. 3, 7

[42] Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, and Ulrich Neumann. Point-NeRF: Point-based neural radiance fields. In *CVPR*, 2022. 3

[43] Bin Yang, Min Bai, Ming Liang, Wenyuan Zeng, and Raquel Urtasun. Auto4D: Learning to label 4D objects from sequential point clouds. *arXiv preprint arXiv:2101.06586*, 2021. 2

[44] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. MVSNet: Depth inference for unstructured multi-view stereo. In *ECCV*, 2018. 2

[45] Weihao Yuan, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, and Ping Tan. NeWCRFs: Neural window fully-connected CRFs for monocular depth estimation. In *CVPR*, 2022. 8, 9

[46] Fangao Zeng, Bin Dong, Yuang Zhang, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. MOTR: End-to-end multiple-object tracking with transformer. In *ECCV*, 2022. 6

[47] Xi Zhu, Xiya Cao, Zhiwei Dong, Caifa Zhou, Qiangbo Liu, Wei Li, and Yongliang Wang. Nemo: Neural map growing system for spatiotemporal fusion in bird's-eye-view and bdd-map benchmark. *arXiv preprint arXiv:2306.04540*, 2023. 3