Mask2Map:
Vectorized HD Map Construction Using Bird's Eye View Segmentation Masks

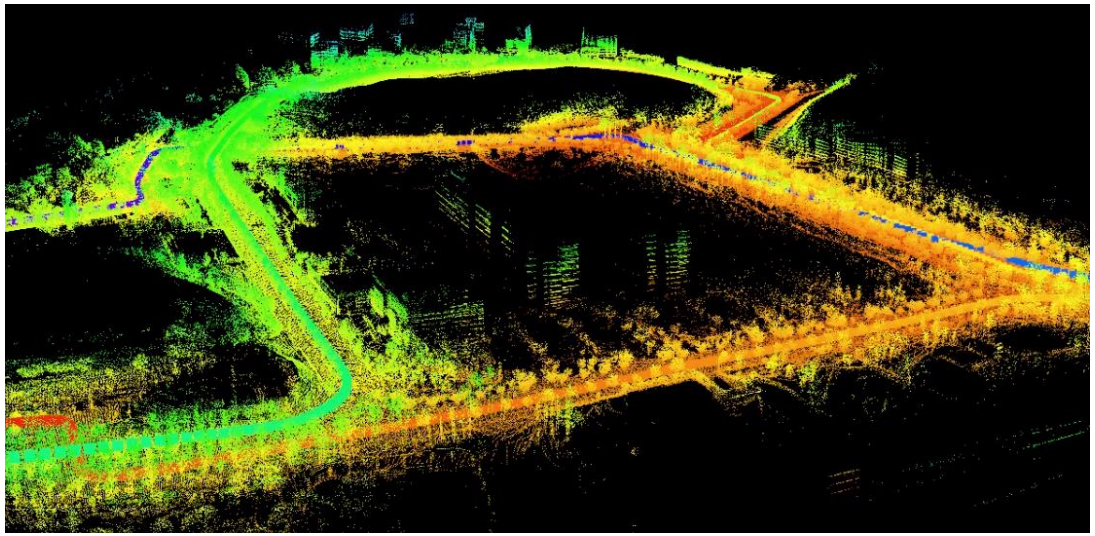
Sehwan Choi*, Jungho Kim*, Hongjae Shin, Jun Won Choi†
Hanyang University, Seoul National University
* Equal contribution, †Corresponding Author

SPALAB
SIGNAL PROCESSING & ARTIFICIAL-INTELLIGENCE LABORATORY

ECCV
EUROPEAN CONFERENCE ON COMPUTER VISION
MILANO
2024

Online HD Map Construction

- HD maps are crucial for autonomous driving
- Offline SLAM-based methods are **costly and limited in timely updates**
- Early approaches used semantic segmentation, which is not suitable for downstream tasks
- Recent works focus on **online vectorized HD map construction** using vehicle sensor data

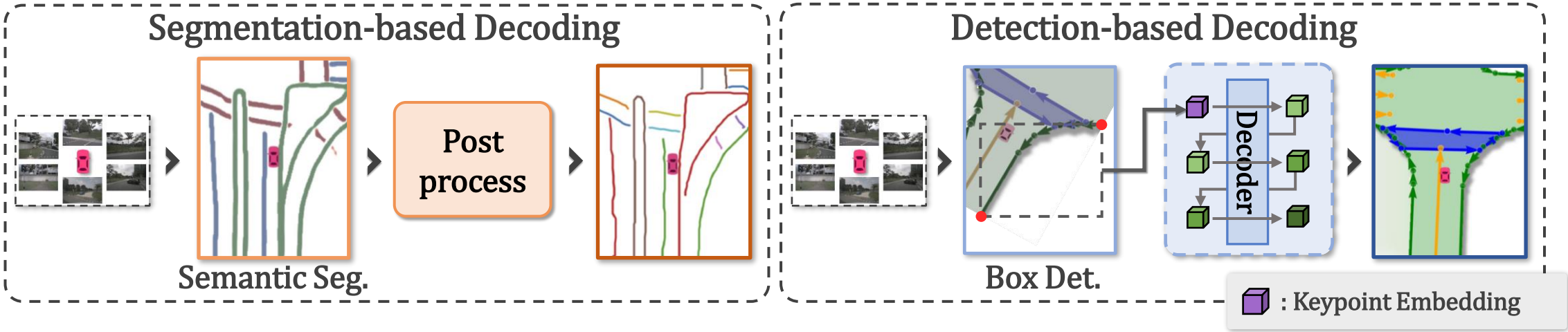


[SLAM-based HD map]

[Online HD map construction]

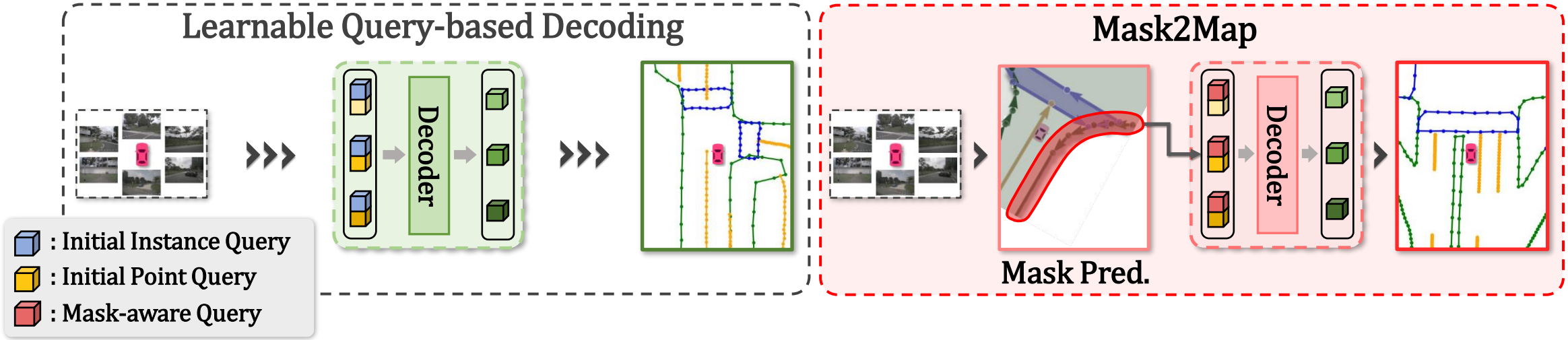
Online HD Map Construction Approaches

- **Segmentation-based:** Require heuristic post-processing, making it time-consuming
- **Detection-based:** Rely only on key points in 2D boxes, limiting the capture of diverse shapes
- **Learnable query-based:** Fail to capture the semantic and geometric information of map instances in complex scenes
- **Mask2Map:** Construct the fine-grained map components using **semantic features of instances** derived from a global perspective



Online HD Map Construction Approaches

- **Segmentation-based**: Require heuristic post-processing, making it time-consuming
- **Detection-based:** Rely only on key points in 2D boxes, limiting the capture of diverse shapes
- **Learnable query-based:** Fail to capture the semantic and geometric information of map instances in complex scenes
- **Mask2Map:** Construct the fine-grained map components using **semantic features of instances** derived from a global perspective



Contribution

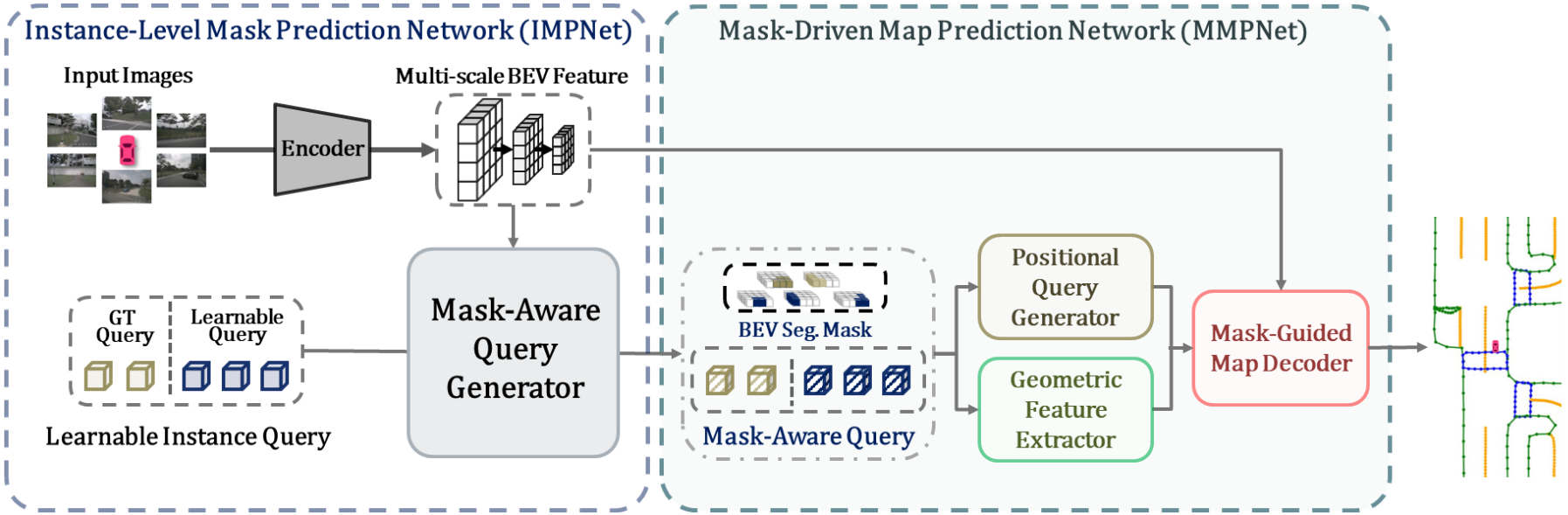
- Present **Mask2Map**, a new framework for online HD map construction
  - Capture semantic information at the instance-level and use it to generate fine-grained map components

- Design a **mask-guided hierarchical feature extraction** architecture
  - Encode instance-level and point-level features for spatial context and geometric information

- Present an **Inter-network Denoising Training strategy** that uses noisy GT queries and perturbed GT Segmentation Masks
  - Ensure inter-network consistency between IMPNet and MMPNet

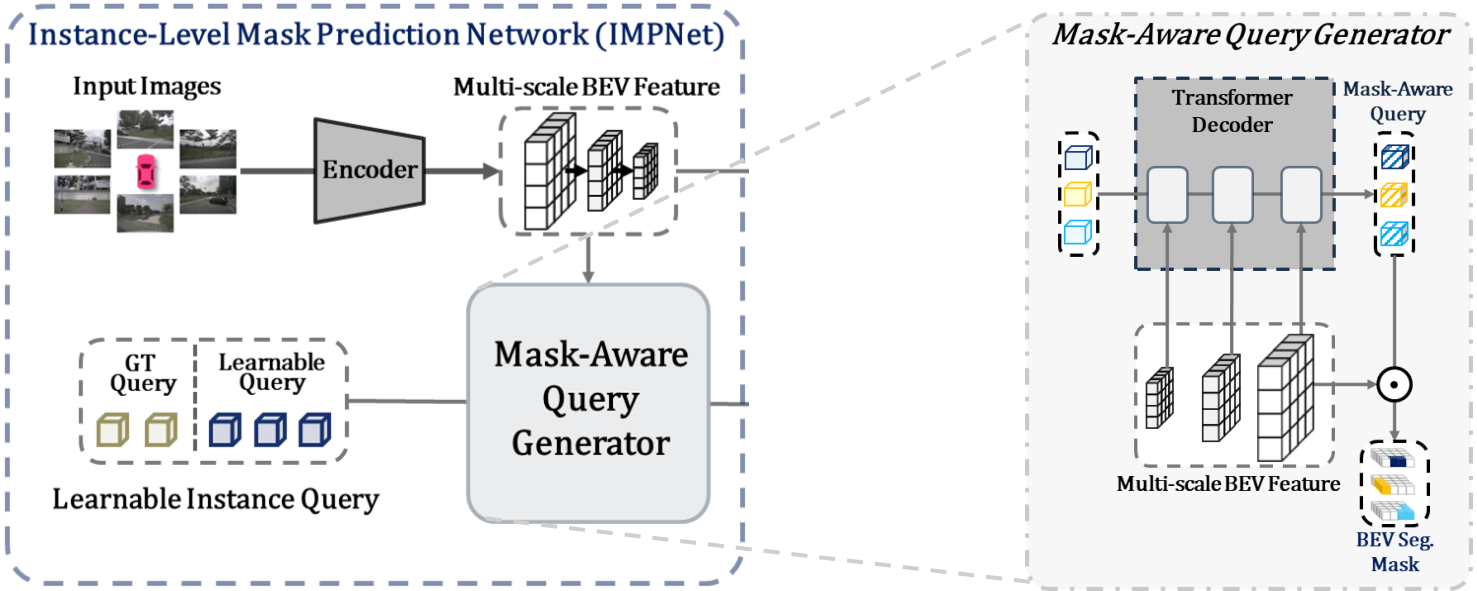
Overall Architecture

- Mask2Map architecture comprises two networks: *IMPNet* and *MMPNet*
- **Instance-Level Mask Prediction Network (IMPNet)**
  - Generate a Mask-Aware Query to capture the semantic features from a global perspective
- **Mask-Driven Map Prediction Network (MMPNet)**
  - Construct the vectorized HD map components from a local perspective using Mask-Aware Query



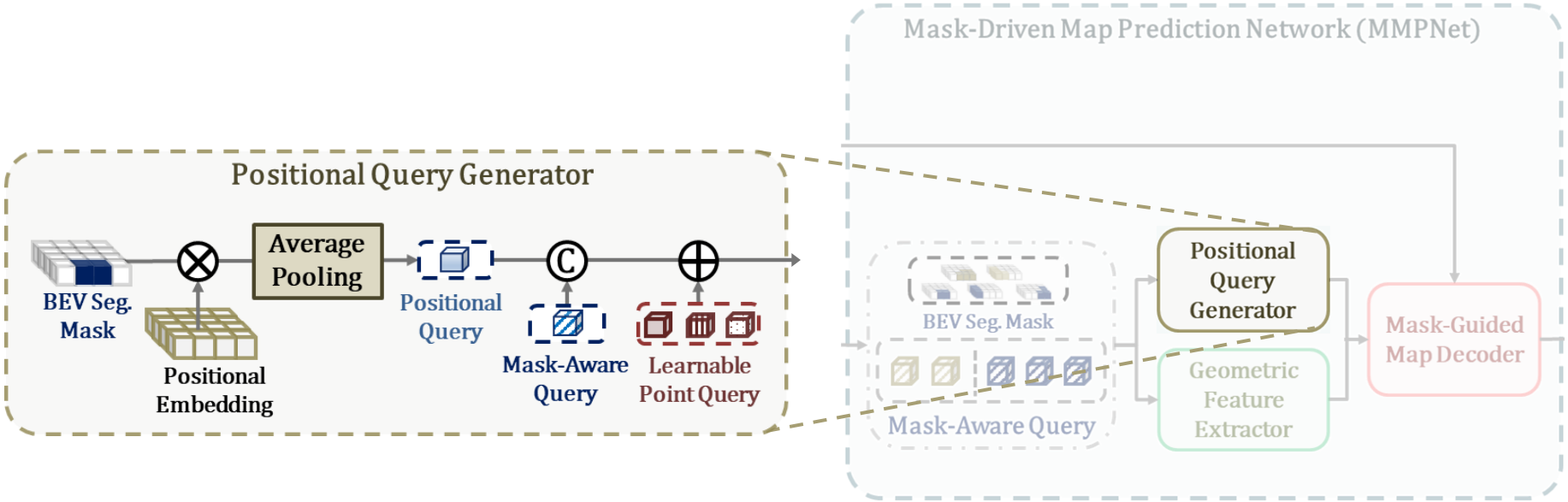
Instance-Level Mask Prediction Network (IMPNet)

- IMPNet extracts **multi-scale BEV features** from input sensor data
- **Mask-Aware Query Generator**
  - Employ a Mask Transformer to generate **Mask-Aware Queries** and predict **BEV Seg. Masks**



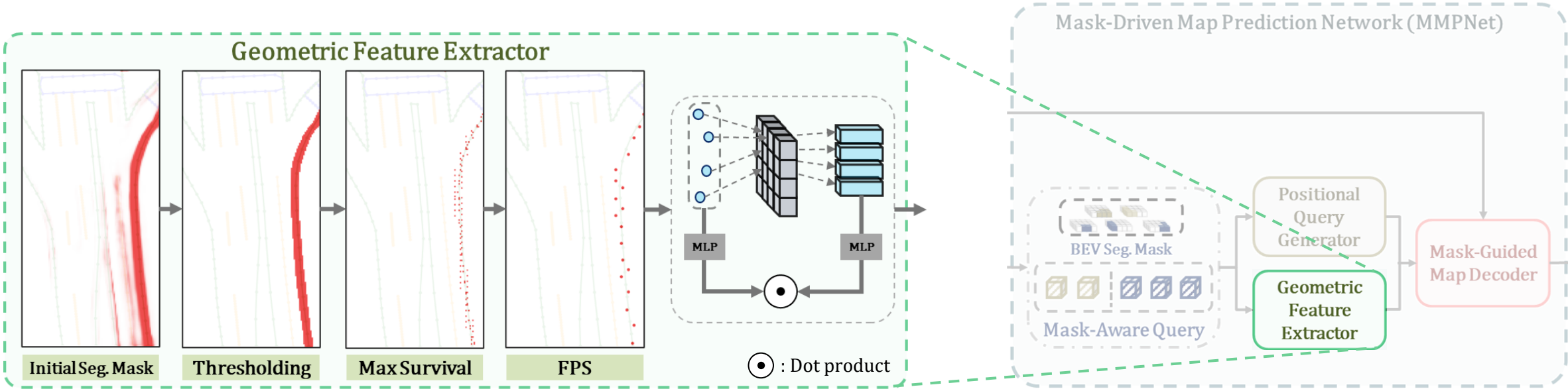
Mask-Driven Map Prediction Network (MMPNet)

- **Positional Query Generator (PQG)**
  - Generate **Positional Query** by average pooling positional embedding within the BEV Seg. Mask
  - Concatenate the **Positional Query** and **Mask-Aware Query**
  - Add the **Learnable Point Query** to convert the query at point level for generating PQG Queries



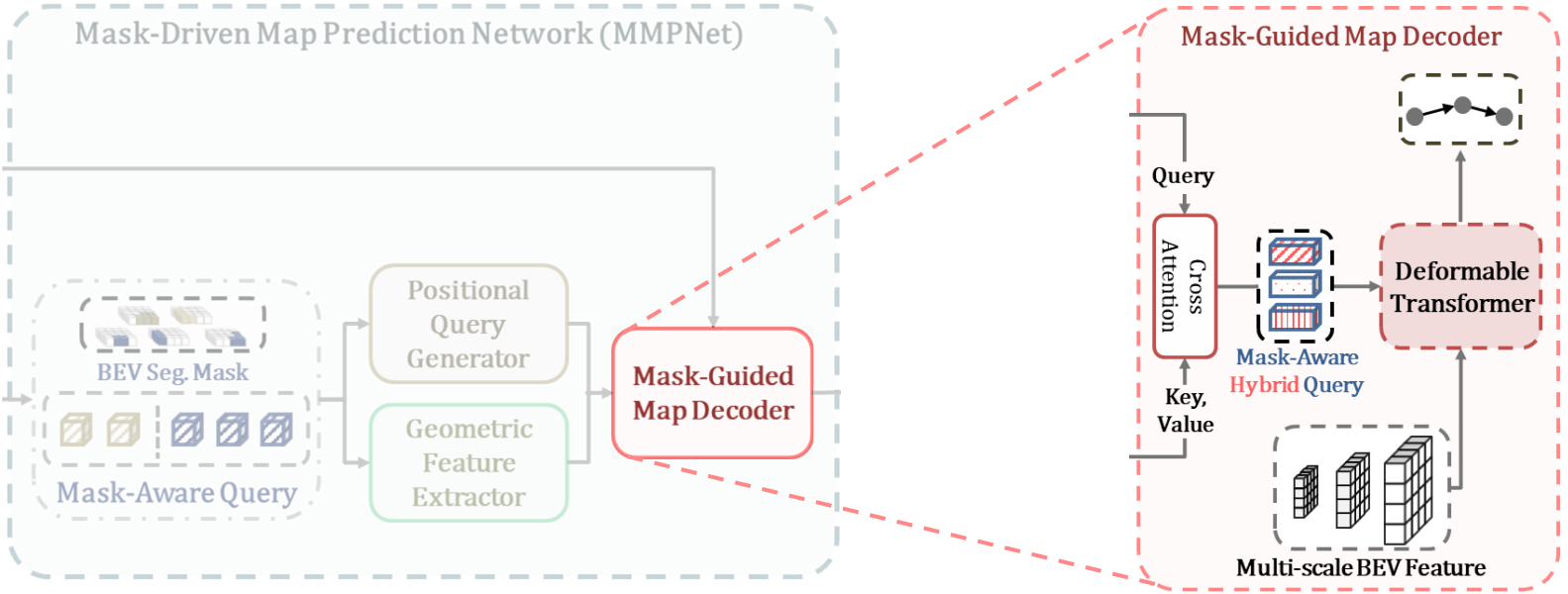
Mask-Driven Map Prediction Network (MMPNet)

- **Geometric Feature Extractor (GFE)**
  - Generate a sparsified BEV mask from the BEV Segmentation Mask by using a **threshold**
  - Use the **Max Survival method** to select the strongest pixel within a sliding window
  - Sample key points using **Farthest Point Sampling**
  - Produce GFE features using coordinates and BEV features from sampled key points



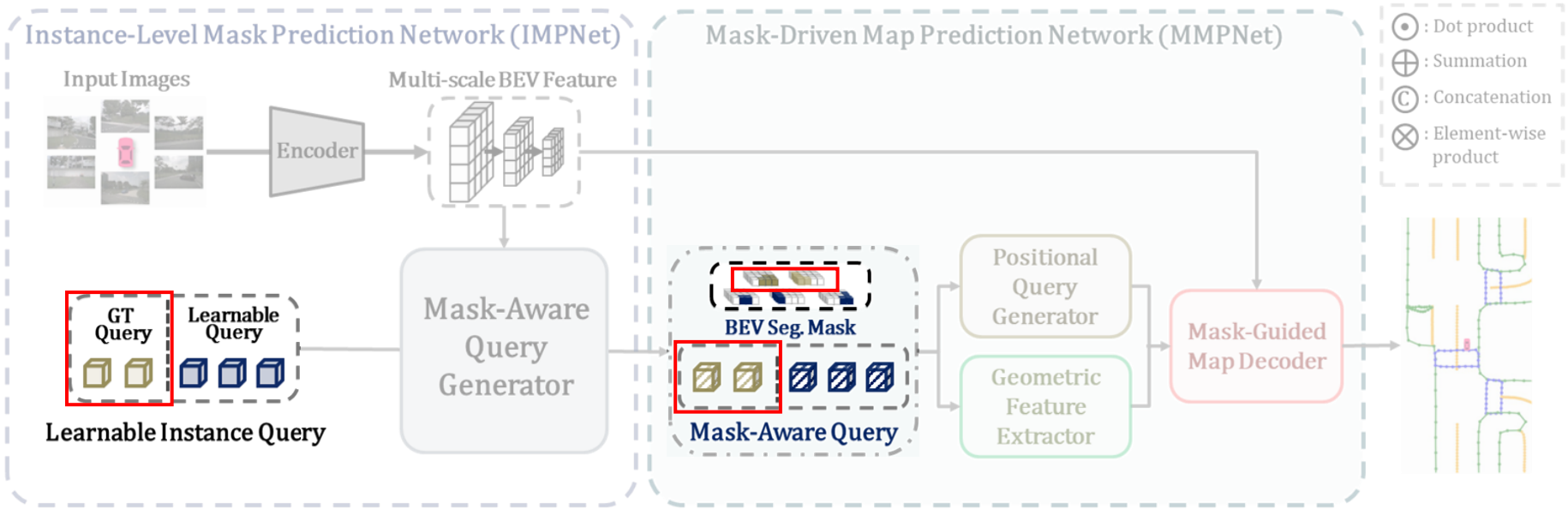
Mask-Driven Map Prediction Network (MMPNet)

- **Mask-Guided Map Decoder (MMD)**
  - Encode PQG Queries through cross-attention using GFE features as keys and values
  - Predict **class scores** and **normalized BEV coordinates** by Deformable Transformer



Inter-network Denoising Training

- Mask2Map passes Mask-Aware Queries from IMPNet to MMPNet
- Inter-network inconsistency occurs when IMPNet and MMPNet queries match different GT instances
- To solve this, we merge noisy GT Queries into learnable queries
  - Our model is trained to denoise GT queries by directly matching them with their corresponding GTs
  - Generate perturbed GT Segmentation Masks alongside GT Queries, replacing BEV Masks for IMPNet



Comparison with State-of-the-Art Methods

- mAP : Chamfer distance-based mAP
- Mask2Map achieves remarkable performance improvements over previous state-of-the-art methods, with gains of 10.1% mAP and 4.1% mAP on nuScenes and Argoverse2, respectively

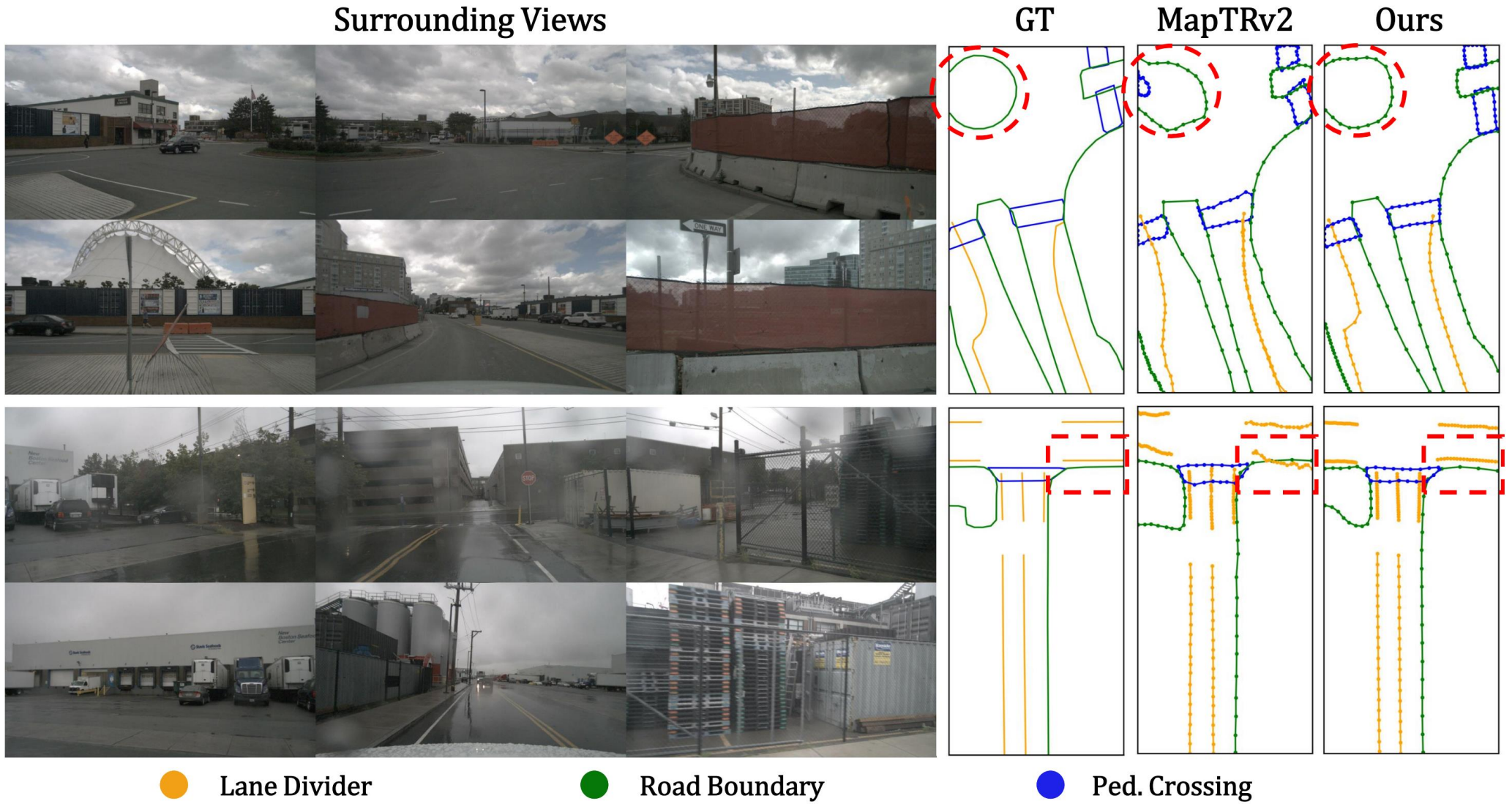
Method	AP_{ped}	$AP_{divider}$	$AP_{boundary}$	mAP
MapVR	47.7	54.4	51.4	51.2
PivotNet	56.2	56.5	60.1	57.6
BeMapNet	57.7	62.3	59.4	59.8
MapTRv2	59.8	62.4	62.4	61.5
Ours	**70.6**	**71.3**	**72.9**	**71.6**

[Comparison with SOTA on *nuScenes* validation set]

Method	AP_{ped}	$AP_{divider}$	$AP_{boundary}$	mAP
VectorMapNet	38.3	36.1	39.2	37.9
MapTR	54.7	58.1	56.7	56.5
MapVR	54.6	60.0	58.0	57.5
MapTRv2	62.9	72.1	67.1	67.4
Ours	**68.1**	**72.7**	**73.7**	**71.5**

[Comparison with SOTA on *Argoverse2* validation set]

Qualitative Results



Thank you

Paper

Website

E-mail

SPALAB

SIGNAL PROCESSING & ARTIFICIAL-INTELLIGENCE LABORATORY

ECCV

EUROPEAN CONFERENCE ON COMPUTER VISION

MILANO 2024