

# Generation of Vectorized Maps Beyond Vehicle View

Clara Gomez<sup>1</sup>, Alberto Jaenal<sup>2</sup>, Antonio Artuñedo<sup>1</sup>, Jorge Godoy<sup>1</sup>, Jorge Villagra<sup>1</sup>

**Abstract**—Autonomous driving relies on High Definition (HD) maps for safe navigation. Traditional HD maps construction is costly in hardware, data and human resources, which together with its update limitations hinders scalability. Recent works have proposed online alternatives for HD vectorized mapping from onboard sensors. However, sensor field of view is limited, and the range of the reconstructed maps ahead of the vehicle is insufficient for safe planning. This paper aims to address this limitation by proposing the novel *beyond-view* vectorized map generation problem: given vectorized maps of the area sensed by the vehicle (*in-view*), to generate plausible map continuations. To experimentally assess its feasibility, we propose BeyondFormer, which, to the best of our knowledge, is the first work designed towards beyond-view map generation. Given the novelty of the problem, we generate the first dataset specifically designed for it and evaluate the proposed approach. The results demonstrate consistent performance across diverse scenarios, establishing learning-based methods as a promising direction for map forecasting in autonomous driving. Beyond demonstrating the feasibility of the task, we provide an extensive discussion of the method's limitations and identify key future research directions for scaling it to more complex driving conditions. Code is available at <https://git-autopia.car.upm-csic.es/beyondformer>.

## I. INTRODUCTION

Recent developments in autonomous driving have been boosted by High Definition (HD) maps [1]. HD maps are highly accurate maps that provide detailed and vectorized representations of road elements such as lanes and pedestrian crossings. They have become a widely adopted mapping solution enabling safe and reliable operation for commercial autonomous vehicles. However, traditional HD map construction is prohibitively expensive, requiring highly sensorized vehicles, time-consuming human annotations and costly data storage. They also require frequent updates as they quickly become outdated. These requirements make HD maps difficult to scale to large environments and long time horizons [2].

Learning-based methods have recently emerged to reduce dependence on offline map creation by constructing online vectorized HD maps around the vehicle using onboard sensors [3], [4]. These approaches are inherently limited by the sensor range and do not provide information beyond the observable region. In contrast, classical HD maps do not have this restriction enabling long-range planning, which motivates studying whether road structure beyond the sensor field of view can be inferred from partial observations or *in-view* maps. This is an unexplored gap for vectorized road representations.

<sup>1</sup> Centre for Automation and Robotics, CSIC-UPM, Spain. {clara.gomez, antonio.artunedo, jorge.godoy, jorge.villagra}@csic.es

<sup>2</sup> Aragon Institute for Engineering Research (I3A), University of Zaragoza, Spain. ajaenal@unizar.es

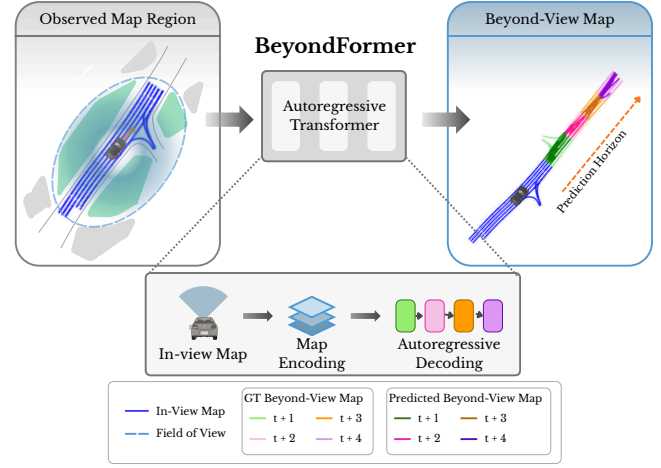


Fig. 1. Given vectorized in-view maps of the observed region, BeyondFormer generates plausible continuations autoregressively.

To address this gap, we formulate the *beyond-view* vectorized map generation problem: consuming in-view maps to predict vectorized map continuations or beyond-view maps, as depicted in Fig. 1. To experimentally assess this problem, we propose a learning based model, BeyondFormer, which generates plausible vectorized maps beyond the sensed area by the vehicle. This model adopts a multi-head encoder-decoder architecture: a Transformer and a simple MLP are used to encode the in-view map and learn feature representations that are decoded with a Transformer decoder using a geometry head and an auxiliary topology head.

BeyondFormer is evaluated on map samples extracted from 3DHD CityScenes Dataset [5]. Given the exploratory nature of this work, our goal is to introduce a novel problem formulation and establish a clear and reproducible first step for beyond-view map generation, rather than a highly optimized model. Accordingly, the experimental evaluation focuses on controlled settings and ablation studies, aiming to analyze the model behavior and identify key challenges and limitations.

To summarize, our contributions are:

- The introduction and study of the novel problem of beyond-view vectorized map generation in autonomous driving.
- To the best of our knowledge, this is the first work to generate vectorized representations of road environments beyond the sensed area by the vehicle, establishing a first step towards beyond-view map generation.
- Given the novelty of this work, no dataset matches the requirements for this problem. We release a dataset with in-view and beyond-view map samples from 3DHD CityScenes Dataset, and evaluate BeyondFormer to analyze its capability for beyond-view map continuation.

## II. RELATED WORK

Current learning-based approaches for autonomous driving aim to reconstruct and forecast different aspects of the driving environment directly from sensor observations and structural priors reducing the dependence on offline maps. These works can be grouped into in-view vectorized map reconstruction and beyond-view trajectory and world model generation.

**In-view HD map reconstruction.** Most of the work on HD map generation has been developed for in-view map prediction in which learning-based solutions transform on-board sensor readings (mainly images and point clouds) into vectorized Bird’s-Eye View (BEV) representations [3], [6], [7], [8]. By encoding the scene as vectorized BEVs, the representation remains compact, memory-efficient, sensor-agnostic and invariant to viewpoint, supporting consistent environment modeling. HDMaNet [3] pioneers in solving this problem, although they first obtain a rasterized map (which is highly resource demanding) to then extract the vectorized representation. VectorMapNet [7] is the first work to directly produce vectorized in-view maps. Of most interest for this work are previous works that, in addition to on board sensor readings, integrate SD maps as input for the reconstruction [4], [9], [10], [11]. SMERF [4] encodes the SD map as a structure prior with a Transformer encoder that gets cross-attended with the observed evidence (BEV constructor) to reconstruct the in-view map. P-MapNet [9] and StreamMapNet [12] already identify that previous work prediction ahead of the vehicle is insufficient in distance, and point to more intelligent decisions when far-seeing solutions are available. However, they remain constrained to the observable region, extending the effective range of in-view reconstruction rather than predicting road structure beyond the sensor field of view. P-MapNet improves the long-range prediction by specific SD map and HD map encoders to aid the decoding process, and StreamMapNet by integrating a memory buffer that better captures the temporal dimension of the sequences to map. An interesting approach is also proposed in [13] where in-view map prediction is produced with a teacher-student model in which the teacher model has privilege access to future information. Although the student model does not predict into the future, it benefits from ahead-looking capabilities.

**Beyond-view generation.** Generation of future views of the vehicle has been addressed to predict other users’ movement through trajectory prediction [14] and video or world model generation [15]. Works on the latter aim to forecast pixel-level scene evolution [16], [17], [18], [19]. However, these works do not explicitly model vectorized road geometry and topology beyond the observed region and, as discussed by VectorMapNet, predicting pixels adds an extra computationally intensive process to the already demanding learning process of vectorization unsuitable for vectorized map generation.

Recently, similar ideas have been explored in indoor environments for structural scene completion and layout extrapolation [20], [21]. However, transferring this idea to outdoor driving scenes introduces additional challenges due to signifi-

cantly higher structural variability and topological complexity (multiple lanes and directions, roundabouts, lane splits, etc). Given the wide variability in driving situations, predicting beyond-view maps presents a significant challenge.

Although inspiration and technical solutions can be borrowed from the related work, to the best of the author’s knowledge, this work is the first to study beyond-view vectorized map generation for autonomous driving.

## III. METHOD

### A. Problem formulation

The problem of beyond-view vectorized map generation aims to infer the continuation of the road structure outside of the sensor field of view from a partial map observation. At each time step  $t$ , we assume that a perfect in-view map of the area perceived by the vehicle is built from sensor readings. These maps are represented as a set of vectorized instances, each instance containing a segment of a lane centerline. The complete set of lanes for the in-view map covers approximately 70m around the vehicle (which aligns to the area that most in-view map generation methods provide [10], [11]). The objective is to increase the vehicle’s horizon by predicting plausible continuations of the centerlines in the driving direction, extending the map for another 100 – 150m.

### B. In-view input map and ground truth map

Every data sample in the in-view input map consists of the lane centerline segments in a  $R$  radius around a global

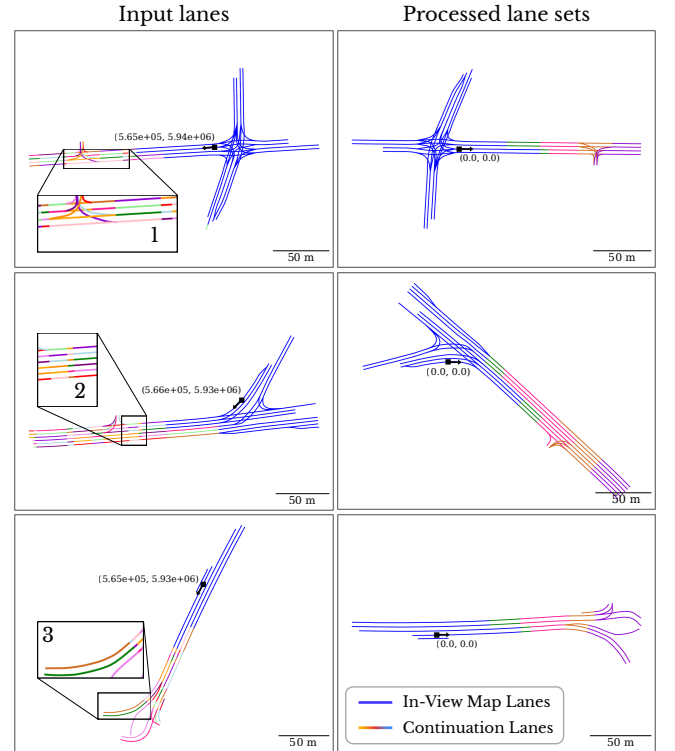


Fig. 2. Examples of input data samples (left) and processed for autoregressive map prediction (right). In addition to the effect of normalization to vehicle pose, (1) shows unrealistically long lane segments, (2) shows multiple adjacent short lane segments and (3) shows imbalanced adjacent lane lengths. Processed lanes offer a more balanced and realistic representation.

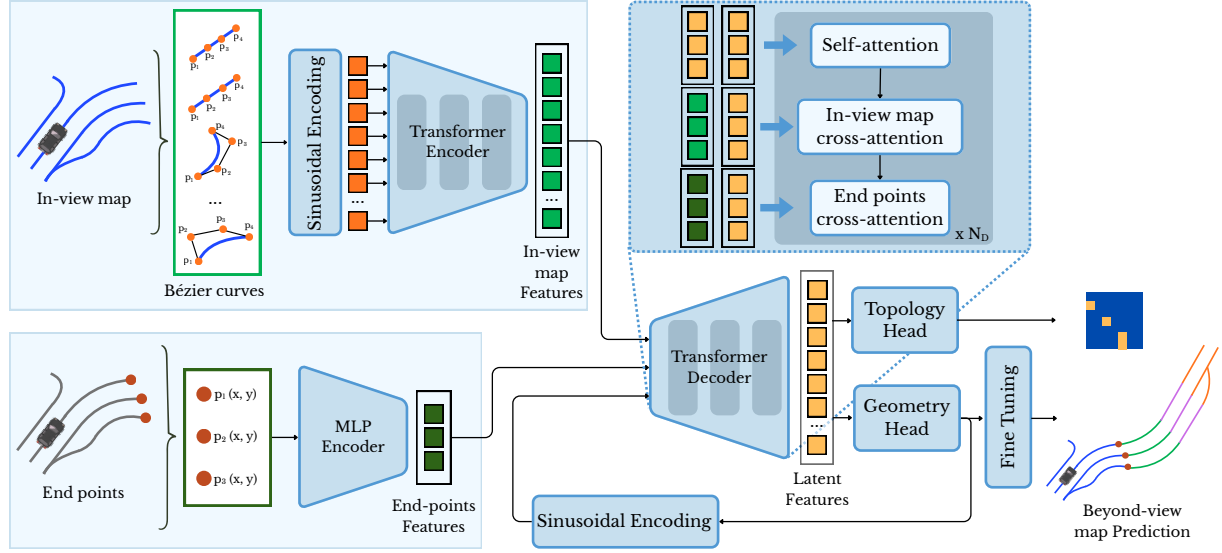


Fig. 3. Architecture of BeyondFormer comprising three main components: the in-view map transformer encoder that processes maps as Bézier curves, the end points encoder that uses a simple MLP encoder to extract end point features, and the decoding block formed by a transformer decoder with self-attention and cross-attention to the in-view map and end point features, a topology head and a geometry head that autoregressively generates the map continuation.

UTM coordinate (shown in blue in Fig. 2). The ground-truth map consists of the lane centerline segments in a  $R$  radius for the subsequent global coordinates of the trajectory up to  $D$  distance. Extracted data samples are shown in the left column of Fig. 2. We unify lane centerline segments lengths and normalize centerline points using the vehicle's position and orientation, resulting in the processed lanes shown in the right column of Fig. 2. Processed lanes solve several data problems as shown in the zoomed areas.

Typically, lane centerline segments consist of a variable number of points depending on the complexity of the curve. As this is not suitable for learning-based algorithms, we turn each lane centerline segment into a cubic Bézier curve [6]. Cubic Bézier curves are convenient as they can represent curves up to two inflection points using only 4 control points.

The resulting in-view input map encompasses  $M$  connected Bézier curves, where each curve corresponds to a lane centerline segment. The resulting ground truth map encompasses  $N$  sets of Bézier curves, where each set corresponds to the new centerline segments that the vehicle will observe in the next generation step, until the prediction horizon,  $t + N$ . The map is autoregressively generated using these sets.

### C. System Overview

The pipeline of BeyondFormer is shown in Fig. 3 and it consists of four building blocks: the in-view map encoder, the end points encoder, the multi-head map decoder and fine tuning. The input of our model are the in-view input map and the end points in the direction of travel from the in-view map. First we encode them into a feature representation, using a Transformer for the former and a MLP for the latter. Then, we apply cross-attention between both feature representations to decode a latent space through a transformer decoder. The latent space is processed with a geometry head to decode the

beyond-view map and an auxiliary topology head to learn lane connectivity. Up to here, the pipeline is trained end-to-end, afterwards an additional fine tuning step performs smoothing and connectivity checks to the predictions.

### D. In-view map Encoder

A Transformer encoder processes the in-view map following the standard pipeline consisting of an input embedding, a positional encoding and  $N_e$  encoder blocks with multi-head self-attention. The input embedding maps each of the four control points  $\mathbf{P}_i^{\text{in}} = \{\mathbf{p}_{i,j}^{\text{in}}\}_{j=1}^4$  with  $\mathbf{p}_{i,j}^{\text{in}} = (x_{i,j}, y_{i,j})$  of the  $M$  Bézier curves into an embedding of size  $d_{\text{model}}$ . Then, we employ a sinusoidal positional encoding to map the sequentiality of the data. This enhances sensitivity to positional variations, enabling the model to effectively reason about the structure of the curves. The sinusoidal encoding is formulated based on the definition in [22], as:

$$\begin{aligned} \text{PE}(p, 2k) &= \sin\left(p \cdot 10000^{-\frac{2k}{d_{\text{model}}}}\right) \\ \text{PE}(p, 2k+1) &= \cos\left(p \cdot 10000^{-\frac{2k}{d_{\text{model}}}}\right), \end{aligned} \quad (1)$$

where  $k$  is the point index in the embedding dimension and  $p$  is the position in the input sequence.

The embedded data is then processed by the encoder blocks using  $H$  self-attention heads to extract and encode the global geometry of the lanes in the in-view map.

### E. End points Encoder

End points of the last lane segments in the direction of travel are used to guide the decoder to the coordinates to start the prediction. As for the in-view map, points  $\mathbf{p}_i^{\text{end}} = (x_i, y_i)$  are transformed into embeddings of size  $d_{\text{model}}$ . Given the simplicity and low dimensionality of the end points, a 3-layer MLP encoder is suitable to build their feature representation.

### F. Multi-head Map Decoder

The Transformer decoder produces a latent space of the lane map around the vehicle from the encoded feature representations of the in-view map and the end points. The geometry head predicts beyond-view map sections and the auxiliary topology head predicts the connectivity between lanes.

The decoder receives the encoder outputs (feature representations for in-view map and end points) and the input data for each generation step in  $N$ . The input data is processed by an input embedding and a sinusoidal positional encoding.  $N_d$  decoder blocks with multi-head attention are used to produce the latent space. First, self-attention is used to iteratively refine the prediction. Then, BeyondFormer fuses the in-view map features and the end point features with its own prediction by leveraging cross-attention. A 3-layer MLP is used as geometry head to produce the decoder output. The auxiliary topology head consists of a 3-layer MLP lane embedding and a 2-layer MLP pairwise edge logits. Edges are computed for every pair of lane features. The output of the auxiliary topology head is the learned adjacency matrix. Finally, a projection layer transforms the decoder output from  $d_{model}$  to Bézier curve points,  $\mathbf{P}_i^{\text{out}} = \{\mathbf{p}_{i,j}^{\text{out}}\}_{j=1}^4$  with  $\mathbf{p}_{i,j}^{\text{out}} = (x_{i,j}, y_{i,j})$ .

The prediction is generated autoregressively, i.e., the model sequentially predicts the next lanes set given its own past predictions. We perform  $N$  autoregressive generation steps according to the sets of data that we prepared from the ground truth. For example, in Fig. 3 there are 3 of such steps. A scheduled training [23] strategy is used which gently changes the training process from a fully guided scheme using the true previous set, towards an unguided scheme which uses the generated set instead. This strategy has shown to be more robust to correct its own mistakes at inference. Practically, we start decoding using as input data a learned beginning-of-sequence element and then append the next ground-truth set or predicted set according to a scheduled-training ratio,  $r_{st}$ .

### G. Fine tuning

Fine tuning is performed on the generated samples to ensure lane connectivity. Nearest neighbour matching is applied between generated lane starts and closest generated lane ends.

### H. Loss Function

The loss function combines the geometry loss function, the topology loss function and a coupling loss:

$$\mathcal{L} = \lambda_{geom}\mathcal{L}_{geom} + \lambda_{topo}\mathcal{L}_{topo} + \lambda_{coupling}\mathcal{L}_{coupling} . \quad (2)$$

The geometry loss function combines four losses:

$$\mathcal{L}_{geom} = \lambda_{vec}\mathcal{L}_{vec} + \lambda_{dist}\mathcal{L}_{dist} + \lambda_{curv}\mathcal{L}_{curv} + \lambda_{cont}\mathcal{L}_{cont} , \quad (3)$$

for which we discretize the ground truth and predicted Bézier curves in  $K$  points. The vectorized Huber loss  $\mathcal{L}_{vec}$  minimizes the error between matched centerline points. The distance loss  $\mathcal{L}_{dist}$  supervises the accumulated length of each lane and its corresponding ground-truth lane. The curvature loss  $\mathcal{L}_{curv}$

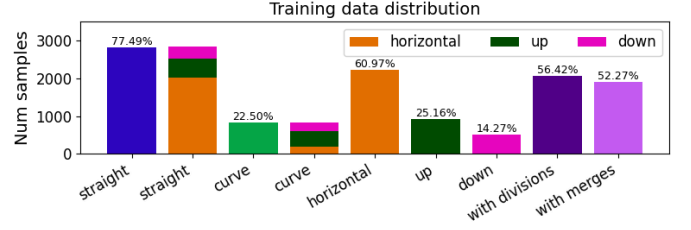


Fig. 4. Distribution of road samples according to their direction (after normalization), shape and complexity.

supervises the difference in logarithmic curvature between the predicted lanes and ground truth. Curvature is calculated for the sampled points and accumulated over the lane. Logarithmic curvature is used to avoid loss explosion for straight lanes. The continuity loss  $\mathcal{L}_{cont}$  supervises the connectivity between lane sections by minimizing the distance between each lane start point and the end points of any previous lane section.

For the topology loss function  $\mathcal{L}_{topo}$ , we extract the ground truth adjacency matrix from the connected lanes and compute the binary cross-entropy with logits. The coupling loss  $\mathcal{L}_{coupling}$  enforces consistency between the auxiliary head and the main head by minimizing the geometric end-to-start point distance between topological adjacent lane pairs.

## IV. EXPERIMENTAL EVALUATION

This section validates BeyondFormer, a proof of concept for beyond-view map generation, demonstrating its feasibility through quantitative and qualitative results.

### A. Experimental setup

**Dataset.** We evaluate BeyondFormer in 3DHD CityScenes Dataset [5], a real-world LiDAR-based 3D dataset of driving sequences in Hamburg. In addition to point clouds, it provides geo-localized vectorized representations of road elements. 3DHD CityScenes Dataset contains 57510 geo-localized samples for training, 8087 for validation and 13061 for test. The samples in each partition belong to disjoint road sections. We selected 3608 training, 171 validation and 104 test samples in simplified conditions (they do not contain roundabouts or intersections) for our experiments. We use  $R = 70\text{m}$  as radius for the in-view and ground-truth map of each sample and  $D = 150\text{m}$  as ground-truth distance. The distribution of road samples is presented in Fig. 4. Our processed dataset is available at <https://zenodo.org/records/22256829>.

**Hardware and training details.** Our model trains on 1 NVIDIA H200 GPU with a batch size of 64 using AdamW optimizer and an adaptative learning rate consisting in 2.5% warm-up epochs and a cosine scheduler. Initial learning rate is set to  $1 \times 10^{-4}$  and dropout to 0.1. As samples consist of different lane numbers, we use zero padding and masking, and filtering of uninitialized lane predictions before metric computation. We train the model for 10k epochs with a generation horizon of  $N = 4$  steps, and  $H = 8$ ,  $N_e = N_d = 6$ ,  $d_{model} = 128$  and  $r_{st} = 2 \times 10^{-4}$ . The dimensions of MLPs are [512, 256, 128] for the end points encoder, [128, 128, 128] for the geometry head and for the topology head lane embedding and [256, 128] for the edges. The hyperparameters

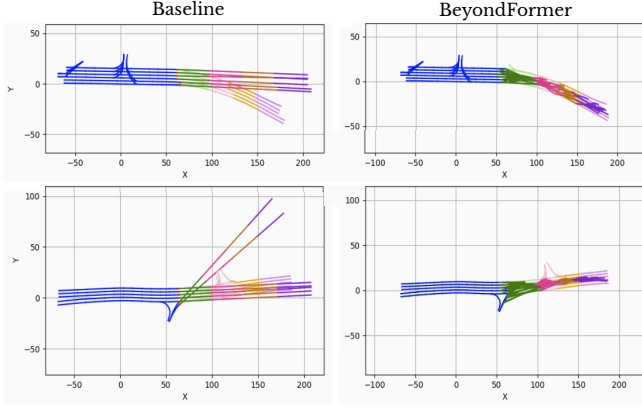


Fig. 5. Examples of beyond-view map generation with baseline and BeyondFormer. The maps are color-coded as follows: blue for in-view maps, light colors (green, pink, orange, and purple) for ground-truth beyond-view maps, and dark colors for predicted outputs.

for loss calculation are:  $K = 100$ ,  $\lambda_{geom} = 1$ ,  $\lambda_{topo} = 6$ ,  $\lambda_{coupling} = 5$ ,  $\lambda_{vec} = 1$ ,  $\lambda_{dist} = 5$ ,  $\lambda_{curv} = 350$  and  $\lambda_{cont} = 50$ . BeyondFormer comprises  $\sim 8M$  learnable parameters.

We provide a classical baseline by matching the endpoint tangent and second derivative (local curvature) of the previous segment while extending the Bézier curve by a fixed length.

**Metrics.** We adopt common metrics from in-view map generation works [4], [7], [9]. We compute Root Mean Square Error (RMSE), Chamfer Distance (CD) and Fréchet Distance (FD) for discretized curves from the predicted and ground-truth Béziers. RMSE measures average pointwise deviation, CD measures average nearest-neighbor distance between two point sets, quantifying bidirectional geometric similarity, and FD measures the minimum coupling cost between two curves respecting the ordering of their points.

### B. Beyond-view map generation performance

Qualitative results of beyond-view map generation using the baseline and BeyondFormer are shown in Fig. 5. Cleaner lines are obtained with the baseline, although it is unable to predict curvature and direction changes. BeyondFormer achieves correct overall lane prediction in terms of curvature, length and distance, although failing to generate smooth parallel lanes. Both methods miss predictions of road splits.

Tab. I and Tab. II show the quantitative performance of the method and the classical baseline in the test data split. Tab. I provides mean  $\pm$  standard deviation values for RMSE, CD and FD in meters. BeyondFormer outperforms the baseline in terms of geometric precision with an overall average decrease in error of 3.14m. As predictions are generated for variable lane length and number, the percentage of error over the total prediction length (approx., 120m and 4 lanes) is also provided. Below 2% error demonstrates the feasibility of the method, although absolute errors indicate room for improvement before operational performance. Additionally, Tab. II shows RMSE values for the 4 autoregressive steps. In both methods, errors increase with prediction horizon as each step relies on previous predictions, with a larger increase for the baseline.

TABLE I  
BEYONDFORMER QUANTITATIVE EVALUATION

|                  | RMSE (m) ↓                        | CD (m) ↓                          | FD (m) ↓                          |
|------------------|-----------------------------------|-----------------------------------|-----------------------------------|
| Baseline         | $8.78 \pm 2.16$                   | $12.23 \pm 4.97$                  | $10.80 \pm 2.79$                  |
| BeyondFormer     | <b><math>7.83 \pm 3.66</math></b> | <b><math>7.08 \pm 4.94</math></b> | <b><math>7.48 \pm 2.95</math></b> |
| BeyondFormer (%) | 1.4277                            | 1.2910                            | 1.3639                            |

TABLE II  
RMSE OVER AUTOREGRESSIVE STEPS (M) ↓

|              | $t + 1$                           | $t + 2$                           | $t + 3$                           | $t + 4$                           |
|--------------|-----------------------------------|-----------------------------------|-----------------------------------|-----------------------------------|
| Baseline     | <b><math>3.22 \pm 0.64</math></b> | $8.12 \pm 1.52$                   | $10.17 \pm 7.53$                  | $12.72 \pm 3.75$                  |
| BeyondFormer | $5.90 \pm 1.98$                   | <b><math>6.34 \pm 2.34</math></b> | <b><math>7.34 \pm 4.28</math></b> | <b><math>9.06 \pm 6.89</math></b> |

### C. Ablation study

We provide ablation studies for two key aspects in learning-based approaches: the loss function and model components. For the loss function ablation study, we consider only the geometry loss. The results in Tab. III show that RMSE, CD and FD can be reduced 23.60%, 30.66% and 24.29% respectively, by adding  $\mathcal{L}_{dist}$ ,  $\mathcal{L}_{curv}$  and  $\mathcal{L}_{cont}$  to  $\mathcal{L}_{vec}$ .

Regarding the components ablation study, in Tab. IV, we show BeyondFormer only with the in-view map encoder, adding the end points encoder, including the auxiliary topology head, and fine tuning the output. We observe how the additions help in better guiding the prediction, reducing RMSE errors in 16.97%, 38.55% and 12.90%, respectively.

### D. Real-time Evaluation

Inference results are measured in a NVIDIA GeForce RTX 4060 GPU with batch size 1. Peak GPU memory consumption during inference is 221MB and average inference time is  $15.3 \pm 1.2$ ms, showing the real-time potential of the method.

## V. CONCLUSIONS

We introduced the problem of beyond-view vectorized map generation for autonomous driving, being the first study to describe and explore the generation of vectorized map continuations beyond the observed field of view. Unlike existing

TABLE III  
ABLATION ON GEOMETRY LOSS FUNCTION

| Method                 | RMSE (m) ↓                         | CD (m) ↓                          | FD (m) ↓                          |
|------------------------|------------------------------------|-----------------------------------|-----------------------------------|
| $\mathcal{L}_{vec}$    | $13.64 \pm 7.67$                   | $10.47 \pm 7.12$                  | $8.11 \pm 4.36$                   |
| + $\mathcal{L}_{dist}$ | $11.03 \pm 6.69$                   | $8.18 \pm 6.74$                   | $6.69 \pm 4.29$                   |
| + $\mathcal{L}_{curv}$ | $12.18 \pm 6.06$                   | $9.98 \pm 6.22$                   | $7.95 \pm 3.92$                   |
| + $\mathcal{L}_{cont}$ | <b><math>10.42 \pm 6.17</math></b> | <b><math>7.26 \pm 5.07</math></b> | <b><math>6.14 \pm 3.31</math></b> |

TABLE IV  
ABLATION ON ENCODERS & FINE TUNING

| Method          | RMSE (m) ↓                        | CD (m) ↓                          | FD (m) ↓                          |
|-----------------|-----------------------------------|-----------------------------------|-----------------------------------|
| In-view map     | $17.62 \pm 10.59$                 | $16.27 \pm 12.88$                 | $13.63 \pm 8.42$                  |
| + End points    | $14.63 \pm 9.77$                  | $12.59 \pm 10.95$                 | $10.65 \pm 6.87$                  |
| + Topology head | $8.99 \pm 4.64$                   | $11.65 \pm 7.21$                  | <b><math>7.25 \pm 3.89</math></b> |
| + Fine tuning   | <b><math>7.83 \pm 3.66</math></b> | <b><math>7.08 \pm 4.94</math></b> | $7.48 \pm 2.95$                   |

works, which focus on reconstructing maps within the visible region, our formulation addresses the prediction of the map in unobserved areas. This problem aims to bridge the gap between short-sighted online methods and long-range offline ones by extending the vehicle horizon with plausible predictions. The experimental results show that this task is feasible under simplified road configurations, while also highlighting the challenges that remain for reliable map forecasting.

As an initial approach to this problem, we presented BeyondFormer, a Transformer-based implementation that provides a reproducible starting point for studying this task. Rather than targeting deployment-ready performance, BeyondFormer demonstrates the feasibility of the presented problem, identifies the main limitations, and provides a foundation for future research toward more complex driving scenarios.

#### A. Limitations and Future Work

As a first study of beyond-view vectorized map generation, this work is intentionally limited to controlled scenarios in order to isolate the core challenges of the proposed task. The experimental evaluation focuses on simple road layouts and assumes accurate in-view maps, leaving the study of complex intersections, roundabouts and robustness for future work.

Although BeyondFormer is able to predict coherent road continuations in simple scenarios, the generated maps remain inaccurate at finer geometric and topological levels. In particular, the model struggles to correctly infer lane splits and merges, and often produces cluttered lane structures. These observations suggest that current architectures primarily capture local continuation patterns and that richer structural reasoning is required to model complex road topology.

Future work will investigate extensions that incorporate additional structural priors, such as SD map encoders similar to recent in-view map generation methods [4], [9], together with improved loss formulations and lane-aware representations. Extending the representation to richer HD map elements, including lane widths, road boundaries, and lane semantics, constitutes another promising research direction.

#### ACKNOWLEDGMENTS

This work was primarily funded by the European Union's Horizon Europe MSCA-2022-COFUND-01 under the Marie Skłodowska-Curie Grant 101126626. Additional support was provided by the MCIN/AEI/10.13039/501100011033 through the TRUEMAP project under Grant PID2024-162510B-I00, also by MICIU/AEI/10.13039/501100011033 and the FSE+ under Grant JDC2024-055088-I.

#### REFERENCES

- [1] H. G. Seif and X. Hu, "Autonomous driving in the iCity—HD maps as a key challenge of the automotive industry," *Engineering*, vol. 2, no. 2, pp. 159–162, 2016.
- [2] G. Elghazaly, R. Frank, S. Harvey, and S. Safko, "High-definition maps: Comprehensive survey, challenges, and future perspectives," *IEEE Open Journal of Intelligent Transportation Systems*, vol. 4, pp. 527–550, 2023.
- [3] Q. Li, Y. Wang, Y. Wang, and H. Zhao, "HDMapNet: An online HD map construction and evaluation framework," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 4628–4634.
- [4] K. Z. Luo, X. Weng, Y. Wang, S. Wu, J. Li, K. Q. Weinberger, Y. Wang, and M. Pavone, "Augmenting lane perception and topology understanding with standard definition navigation maps," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 4029–4035.
- [5] C. Plachetka, B. Sertolli, J. Fricke, M. Klingner, and T. Fingscheidt, "3DHD CityScenes: High-definition maps in high-density point clouds," in *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*, 2022, pp. 627–634.
- [6] L. Qiao, W. Ding, X. Qiu, and C. Zhang, "End-to-end vectorized HD-map construction with piecewise Bezier curve," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13 218–13 228.
- [7] Y. Liu, T. Yuan, Y. Wang, Y. Wang, and H. Zhao, "VectorMapNet: End-to-end vectorized HD map learning," in *International Conference on Machine Learning*. PMLR, 2023, pp. 22 352–22 369.
- [8] B. Liao, S. Chen, Y. Zhang, B. Jiang, Q. Zhang, W. Liu, C. Huang, and X. Wang, "MapTRv2: An end-to-end framework for online vectorized HD map construction," *International Journal of Computer Vision*, vol. 133, no. 3, pp. 1352–1374, 2025.
- [9] Z. Jiang, Z. Zhu, P. Li, H.-a. Gao, T. Yuan, Y. Shi, H. Zhao, and H. Zhao, "P-MapNet: Far-seeing map generator enhanced by both SDMap and HDMap priors," *IEEE Robotics and Automation Letters*, 2024.
- [10] T. Monninger, Z. Zhang, S. Staab, and S. Ding, "NavMapFusion: Diffusion-based fusion of navigation maps for online vectorized HD map construction," *arXiv preprint arXiv:2512.03317*, 2025.
- [11] R. Sun, L. Yang, D. Lingrand, and F. Precioso, "Mind the map! accounting for existing maps when estimating online HDmaps from sensors," in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2025, pp. 1671–1681.
- [12] T. Yuan, Y. Liu, Y. Wang, Y. Wang, and H. Zhao, "StreamMapNet: Streaming mapping network for vectorized online HD map construction," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 7356–7365.
- [13] R. Li, X. Li, M. Xie, H. Shan, S. Qiu, X. Chang, Y. Fan, F. Xiong, H. Jiang, Y. Ren *et al.*, "AMap: Distilling future priors for ahead-aware online HD map construction," *arXiv preprint arXiv:2512.19150*, 2025.
- [14] T. Zhang, M. Fu, Y. Yang, W. Song, and T. Liu, "Edge-enriched graph transformer for multiagent trajectory prediction with relative positional semantics," *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1–12, 2024.
- [15] L. Rowe, R. Girgis, A. Gosselin, L. Paull, C. Pal, and F. Heide, "Scenario dreamer: Vectorized latent diffusion for generating driving simulation environments," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 17 207–17 218.
- [16] W. Zheng, W. Chen, Y. Huang, B. Zhang, Y. Duan, and J. Lu, "Occworld: Learning a 3D occupancy world model for autonomous driving," in *European conference on computer vision*. Springer, 2024, pp. 55–72.
- [17] Z. Yang, Z. Liu, Y. Lu, L. Hou, C. Miao, S. Peng, B. Feng, X. Bai, and H. Zhao, "GenieDrive: Towards physics-aware driving world model with 4D occupancy guided video generation," *arXiv preprint arXiv:2512.12751*, 2025.
- [18] Y. Yang, J. Mei, Y. Ma, S. Du, W. Chen, Y. Qian, Y. Feng, and Y. Liu, "Driving in the occupancy world: Vision-centric 4D occupancy forecasting and planning via world models for autonomous driving," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 9, 2025, pp. 9327–9335.
- [19] K. Zhang, Z. Tang, X. Hu, X. Pan, X. Guo, Y. Liu, J. Huang, L. Yuan, Q. Zhang, X.-X. Long *et al.*, "Epona: Autoregressive diffusion world model for autonomous driving," *arXiv preprint arXiv:2506.24113*, 2025.
- [20] L. Ericson and P. Jensfelt, "Beyond the frontier: Predicting unseen walls from occupancy grids by learning from floor plans," *IEEE Robotics and Automation Letters*, vol. 9, no. 8, pp. 6832–6839, 2024.
- [21] H. Shah, J. Xing, N. Messikommer, B. Sun, M. Pollefeys, and D. Scaramuzza, "ForesightNav: Learning scene imagination for efficient exploration," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 5236–5245.
- [22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
- [23] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer, "Scheduled sampling for sequence prediction with recurrent neural networks," *Advances in neural information processing systems*, vol. 28, 2015.