Supplementary Document of CenterLineDet

Zhenhua Xu, *Student Member, IEEE*, Yuxuan Liu, *Student Member, IEEE*,
Yuxiang Sun, *Member, IEEE*, Ming Liu, *Senior Member, IEEE*, and Lujia Wang, *Member, IEEE*

I. Introduction

In this supplementary document, we provide more detailed explanations of concepts and visualizations of CenterLineDet due to the manuscript page limit.

II. Road Lane Centerline

A. What is Lane CenterLine?

In this work, we aim to detect the graph of road lane centerlines for HD map creation. Road lane centerlines are virtual lines that vehicles drive on. They are critical for the prediction and planning algorithms of autonomous vehicles but do not physically exist. Besides, lane centerlines can have very complicated topology near lane intersections, lane splits and lane merges, which makes our task even more challenging. The diagram showing the definition of lane centerline is shown in Fig. 1.

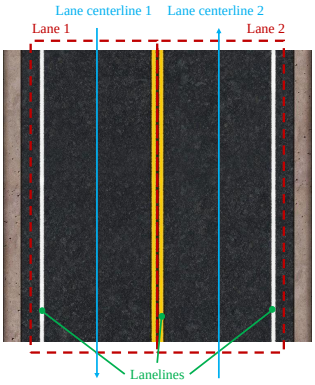


Fig. 1: Road elements. Each red rectangle crops a lane area. Green lines mark lanelines that draw the boundary of lane areas. Both lane and lanelines are physical existing elements. Blue lines are lane centerlines, which are virtual lines defined by humans.

B. Why CenterLine Matters?

Please see Fig. 2 for information.

III. How baselines work?

Previous works mainly focus on the detection task of relatively simpler road elements (e.g., laneline and road boundary), which usually have no overlapping or complicated topological structure. In HDMapNet [1], the authors can simple project detection results of each frame into the world coordinate system and merge them into a global map, which is visualized in Fig. 3.

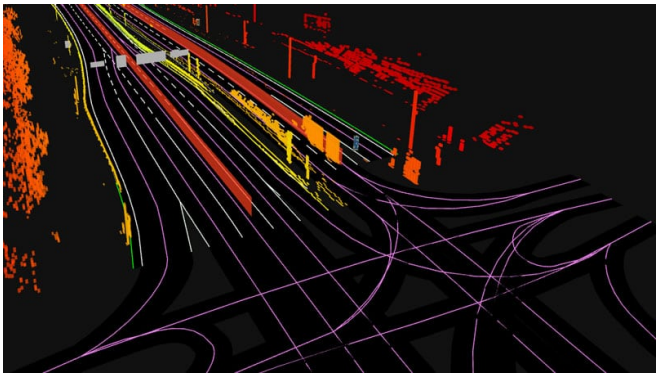


Fig. 2: Visualization of commonly used HD map. Different from road elements such as lanelines and boudanries, lane centerlines (pink lines in this figure) are high-level layer of the HD map. They define the path that vehicles can drive on, thus they can be directly used to assist the planning and prediction task of vehicles. Therefore, the detection of lane centerlines is important for the automatic create of HD maps. (Image url: https://www.automotiveworld.com/articles/hd-maps-the-hidden-sensors-that-help-autonomous-vehicles-see-round-corners/)

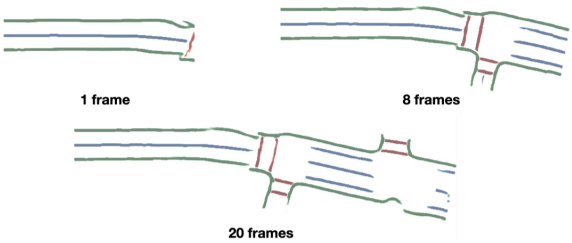
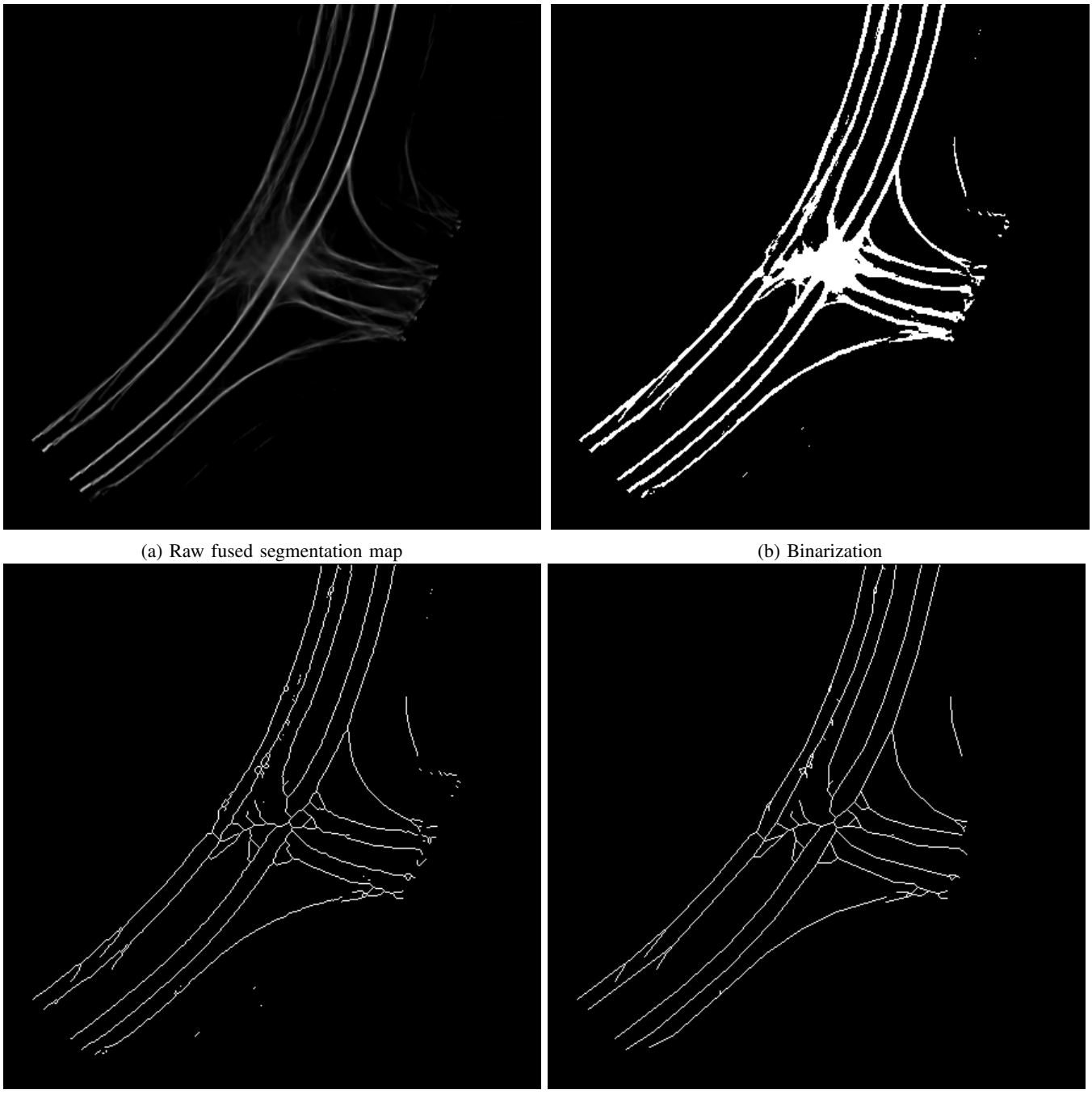


Fig. 3: HDMapNet long-term temporal accumulation (i.e., multi-frame mapping). This naive multi-frame fusion method works fine for simple road elements without overlapping or topology changes.

However, for lane centerlines that have much more complicated topology and severe overlapping, this method tends to fail. The pipeline of how baselines detect lane centerlines is visualized in Fig. 4.

The problems of previous for HD map mapping can be briefly summarized as: (1) they output rasterized results while HD map requires vectorized data; (2) they cannot handle overlapping centerline instances; (3) they are not designed for long-term multi-frame mapping purpose, thus their systems do not have corresponding optimizations for HD map mapping. Most previous works only focus on the detection of objects in a single frame.



(a) Raw fused segmentation map

(b) Binarization

(c) Skeleonization

(d) Skeleton refinement

Fig. 4: Pipeline of baselines for lane centerline detection. (a) Raw segmentation map in the world coordinate system fused by multiple frames. (b) Binarization result. (c) Extract the skeleton of the binary map. (d) Refine the skeleton by connecting closed endpoints and filter short line segments. To the best of our knowledge, all previous works at most do frame fusion only (i.e., they only do step (a)), since steps (b)-(d) are steps for vectorized results. From the results, we can see that since the complicated topology and overlapping of lane centerlines, baselines methods cannot handle intersection areas or topology changes very well. Baselines methods output noisy results, which have unsatisfactory topology correctness. Please zoom in for details.

IV. CenterLineDet

A. How does CenterLineDet solve problems?

To handle the aforementioned problems of previous works, we propose CenterLineDet, which utilizes imitation learning

and a DETR-like decision-making network for HD map creation. The working pipeline of CenterLineDet in a single frame is visualized in Fig. 5, and the final obtained graph is visualized in Fig. 6.

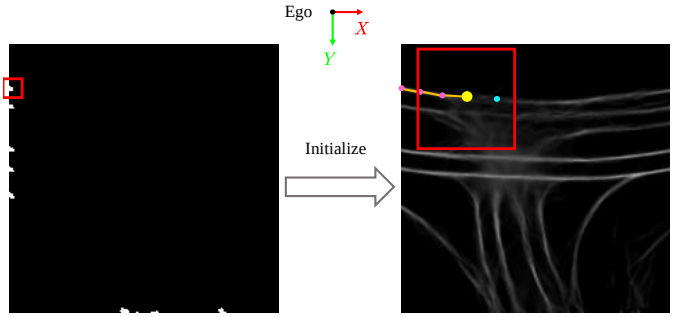


Fig. 5: Pipeline of CenterLineDet for lane centerline detection. In each ego frame, we first predict candidate initial vertices (left figure). Then, CenterLineDet is initialized with one initial vertex (in the red rectangle in left figure). Next, starting from the initial vertex, CenterLineDet iteratively generates the graph by predicting vertices in the next step. In the right figure, pink points are generated vertices, orange lines are generated edges, yellow point is where the agent is currently locating. The agent will predict vertices in the next step based on the visual information within the red rectangle. The cyan point represents the predicted vertex in the next step. The CenterLineDet agent keeps repeating the above processes to detect the centerline graph. It should be noted that, CenterLineDet agent makes decisions based on ego frame information while it is recorded in the world coordinate system. Thus it can directly generate the final required global centerline graph. Please zoom in for details.

The advantages of CenterLineDet can be briefly summaried here: (1) it directly outputs vectorized results; (2) it can better handle overlapping instance and complicate topology than baselines; (3) it records the graph in the world coordinate system, thus it can better merge the detection results of multiple frames. In conclusion, we claim that CenterLineDet presents superiority upon baseline models for the centerline HD map mapping task.

B. Inverse Perspective Mapping (IPM)

First, We obtain the coordinates of the the center of the BEV maps with respect to the base link of the vehicle as $X_{grid} = (x_{grid}, y_{grid}, z_{grid})$. For each image from a camera m, we project these points on to each camera with the extrinsic matrix and intrinsic matrics to obtain $X_{cam} = K(RX_{grid} + T)$. We select points inside of FOV of the camera, and we sample feature vectors from the front view to the BEV features $F^m_{bev}(X_{grid}) = I^m(X_{cam})$ using bilinear interpolation. Finally, we aggregate all the features from six cameras to get $F_{BEV} = \sum_m (F^m_{BEV})$. We will modify the manuscript to provide more information of IPM.

C. Fuse of neighboring frames

In our experiments, there are intersection regions between neighboring frames. For an intersection region, the segmentation heatmap of different frames could be different. Suppose we are at frame T now. To handle inconsistent perspective transformation results of neighboring frames, we fuse neighboring frames (from $T - \tau$ to $T + \tau$) by warping and averaging. The visualization of the fusing process is visualized in Fig. 7.

D. Candidate initial vertices

At each frame T, there is a set of candidate initial vertices as a set $\mathbb{S} = \{s_k\}_{k=1}^K$ to initialize the agent. There are two kinds of candidate initial vertices: (1) points extracted by finding local peaks in the segmentation heatmap $\mathcal{H}_I$, and (2) endpoints of the previous frame. Candidate initial vertices are visualized in Fig. 8.

V. Experimental Settings

In our experiments, images from six cameras and point cloud captured by the top LiDAR are utilized. NuScenes also releases the HD map of road lane centerlines, consisting of vertices and edges. After warping, modifying, and cropping the source HD map in NuScenes, we obtain the ground truth lane centerline graphs G^* that can be used to generate expert demonstrations in our task.

The size of the predicted BEV is 200p×200p, corresponding to a 50m×50m-sized square region centering on the vehicle in the real world.

We train FusionNet with a learning rate as 10^{-3}, and train the transformer network with a learning rate as 10^{-4}. The decay rate of both networks is 10^{-4}. For better performance and faster convergence, the FusionNet and the transformer are trained separately. The FusionNet and other baseline perspective transformation networks are trained for 50 epochs. After obtaining the training data by the proposed behavior-cloning sampling algorithm, we train the transformer network for 75 epochs. All experiments are conducted on a PC with 4 RTX-3090 GPUs.

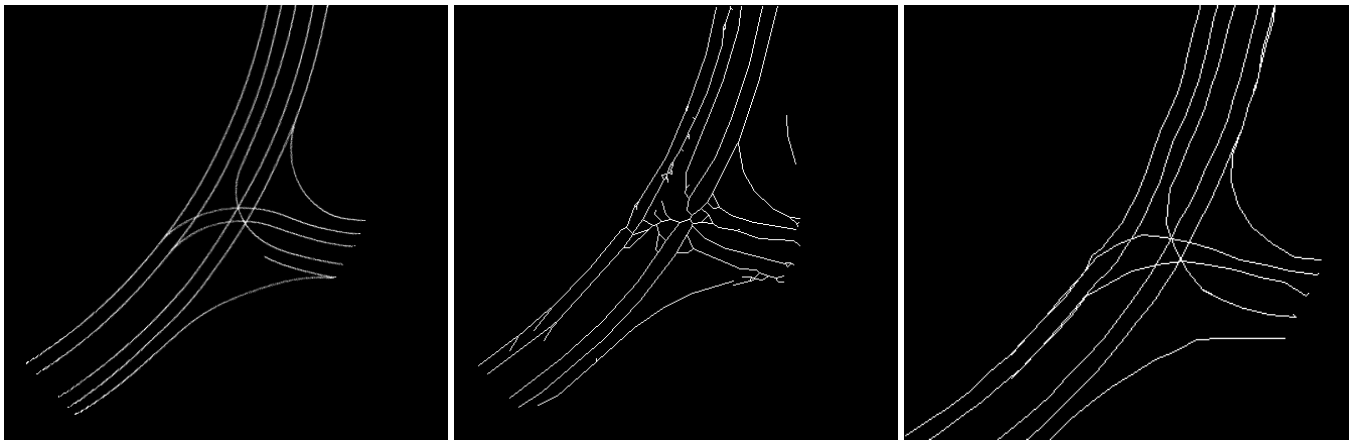
VI. Additional visualizations

More qualitative visualizations of our approach results are shown in Fig. 9 and Fig. 10. CenterLineDet presents more accurate results, **especially in the road intersection areas where centerlines severely overlap with each other. Please zoom in to check the intersection areas.**

References

[1] Q. Li, Y. Wang, Y. Wang, and H. Zhao, "Hdmapnet: A local semantic map learning and evaluation framework," 2021.

[2] Y. B. Can, A. Liniger, D. P. Paudel, and L. Van Gool, "Topology preserving local road network estimation from single onboard camera image," *arXiv preprint arXiv:2112.10155*, 2021.



(a) Ground truth (b) Centerline graph obtained by baselines (c) Centerline graph obtained by CenterLineDet

Fig. 6: Final results comparison. Compared with baselines, CenterLineDet can handle more complicate cases based on the DETR-like decision-making network. Thus, it has much better ability to handle complicate centerline topology and overlapping issues. Please zoom in for details.

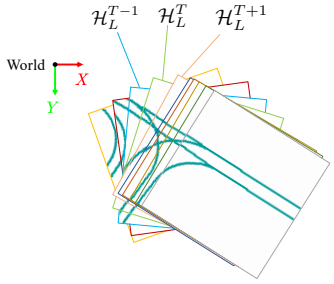


Fig. 7: Visualization of fusing neighboring frames. For better visualization, we use the ground truth lane centerline (cyan lines) to represent the predicted BEV heatmap $\mathcal{H}_L^T$. Suppose $\tau = 1$. (1) We project neighboring frame BEV maps to $\mathcal{H}_L^T$ (i.e., green frame in this figure) based on vehicle poses. (2) Within $\mathcal{H}_L^T$, we sum $\mathcal{H}_L^T$ with projected neighboring BEV maps and average. (3) The sum-averaged BEV map $\tilde{\mathcal{H}}_L^T$ is used as the new BEV segmentation heatmap that fuses neighboring frames at the current step T. The afterward decision-making transformer controls the agent based on the visual information provided by $\tilde{\mathcal{H}}_L^T$.

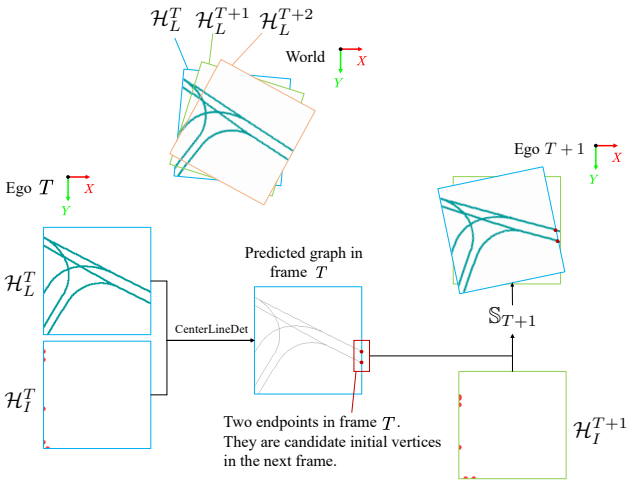
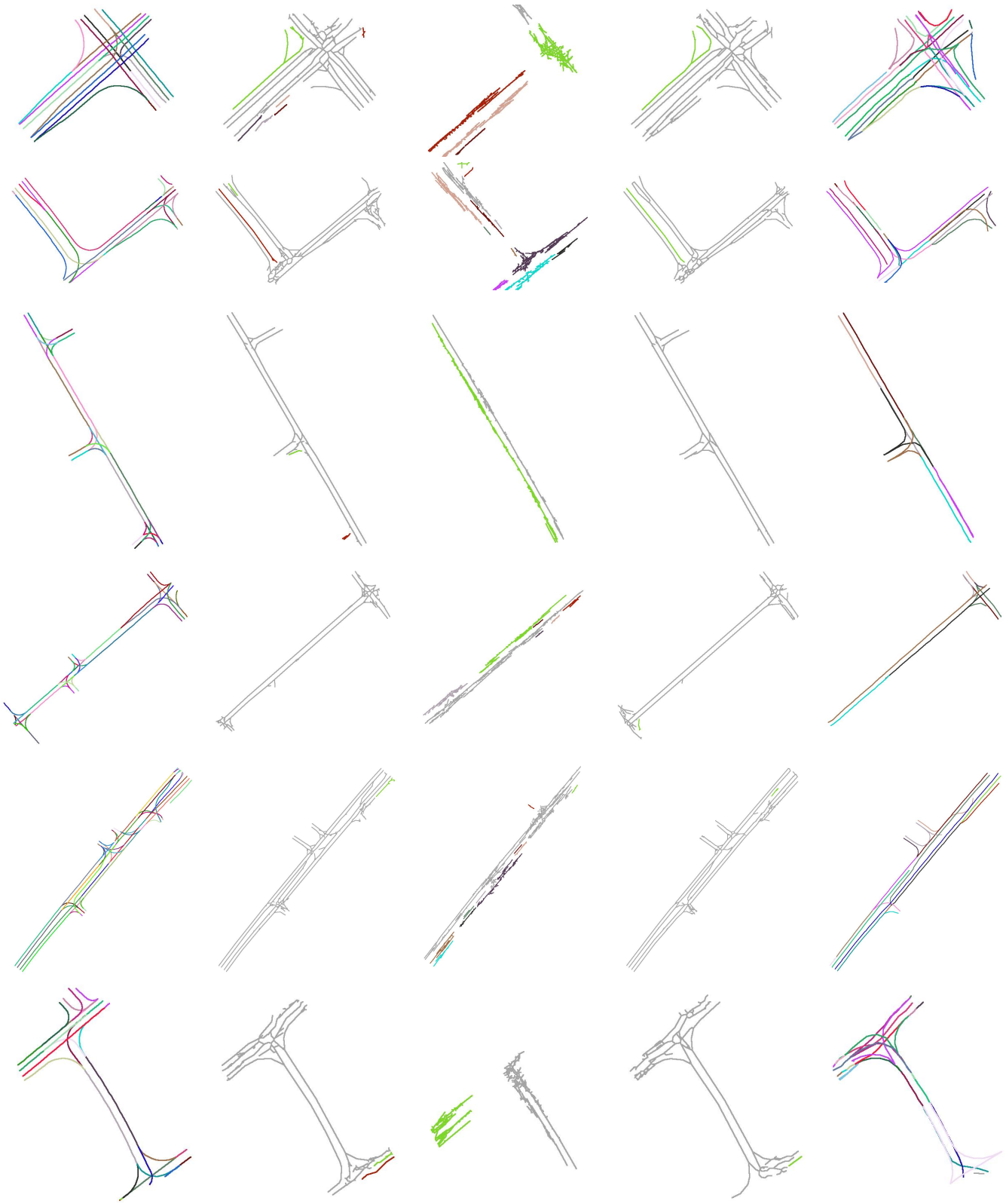


Fig. 8: Candidate initial vertices of frame $T+1$. For better visualization, we use the ground truth lane centerline (cyan lines) to represent the fused predicted BEV heatmap $\tilde{\mathcal{H}}_L^T$. The candidate initial vertices of frame $T+1$ come from (1) local peaks of the heatmap $\mathcal{H}_I^{T+1}$ and (2) endpoints of frame T. Endpoints are vertices at which the agent decides to stop in the previous frame. This figure is best viewed in color. Please zoom in for details.



(a) GT (b) HDMapNet [1] (c) TopoRoad [2] (d) FusionNet (e) CenterLineDet

Fig. 9: Qualitative visualizations. Different colors represent different road centerline instances. CenterLineDet is the only approach that can detect and distinguish multiple instances. All baselines mess up different instances, which leads to incorrect topology, especially in the intersection area. For better visualization, graphs are widened but they are actually of one-pixel width. **Zoom in for details, especially intersection areas.**

(a) GT (b) HDMapNet [1] (c) TopoRoad [2] (d) FusionNet (e) CenterLineDet

Fig. 10: Qualitative visualizations. Different colors represent different road centerline instances. CenterLineDet is the only approach that can detect and distinguish multiple instances. All baselines mess up different instances, which leads to incorrect topology, especially in the intersection area. For better visualization, graphs are widened but they are actually of one-pixel width. **Zoom in for details, especially intersection areas.**