TLCFuse: Temporal Multi-Modality Fusion Towards Occlusion-Aware Semantic Segmentation

Gustavo Salazar-Gomez, Wenqian Liu, Manuel Alejandro Diaz-Zapata, David Sierra González, Christian Laugier

TLCFuse: Temporal Multi-Modality Fusion Towards Occlusion-Aware Semantic Segmentation

Gustavo Salazar-Gomez[†1], Wenqian Liu[†1], Manuel Diaz-Zapata[1,2], David Sierra-Gonzalez[3], Christian Laugier[1]

***Abstract*— In autonomous driving, addressing occlusion scenarios is crucial yet challenging. Robust surrounding perception is essential for handling occlusions and aiding navigation. State-of-the-art models fuse LiDAR and Camera data to produce impressive perception results, but detecting occluded objects remains challenging. In this paper, we emphasize the crucial role of temporal cues in reinforcing resilience against occlusions in the bird's eye view (BEV) semantic grid segmentation task. We proposed a novel architecture that enables the processing of temporal multi-step inputs, where the input at each time step comprises the spatial information encoded from fusing LiDAR and camera sensor readings. We experimented on the real-world nuScenes dataset and our results outperformed other baselines, with particularly large differences when evaluating on occluded and partially-occluded vehicles. Additionally, we applied the proposed model to downstream tasks, such as multi-step BEV prediction and trajectory forecasting of the ego-vehicle. The qualitative results obtained from these tasks underscore the adaptability and effectiveness of our proposed approach.**

***Index Terms*— Semantic Segmentation, Spatio-Temporal, Multi-Sensor Fusion, Deep Learning, Autonomous Vehicles**

I. Introduction

In the realm of autonomous driving, the ability to navigate complex and dynamic environments is contingent on a vehicle's capacity to perceive its surroundings accurately. However, there are instances when objects become partially or entirely obstructed by other elements, causing them to vanish from the ego-vehicle's perspective. Developing robust perception methods to deal with these occluded objects is crucial yet challenging, as it directly impacts the safety, efficiency, and reliability of navigation for autonomous driving.

One notable advancement in perception is the fusion of multiple sensor modalities, which leads to impressive results in object detection [1]–[5], bird's eye view (BEV) semantic grid segmentation[6]–[9] and the other related tasks. Nevertheless, these state-of-the-art approaches still often struggle to detect occluded objects effectively and the research towards occlusion-aware models remains limited[8], [10]–[13].

Unlike previous studies that utilize temporal data solely for predicting future time steps, our work demonstrates that incorporating temporal cues can enhance how autonomous vehicles perceive their surroundings, particularly in complex scenarios involving object occlusion. We propose our innovative TLCFuse designed to process sequences of multi-modal sensor data inputs, capturing the temporal memory of the scene and objects to enhance occlusion-aware semantic BEV grid segmentation. By employing a fixed-size low-dimensional latent representation, our approach extracts spatial features from fused LiDAR and camera sensor readings at each time step throughout the temporal horizon. This representation serves as a stable spatio-temporal memory bank for feature extraction across time, ensuring coherence and efficiency in processing. Moreover, TLCFuse is flexible to be applied to downstream tasks such as one-shot multi-step future BEV grids forecasting and ego-vehicle trajectory prediction. In addition, TLCFuse is differentiable, allowing it to be integrated into an end-to-end training framework. One example for illustration can be seen in Fig. 1.

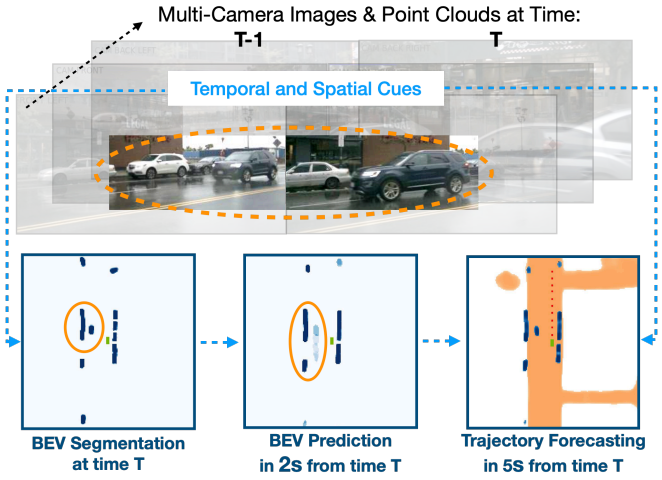


Fig. 1: TLCFuse takes temporal input LiDAR and Camera data, accurately predicts vehicles' locations in the BEV at the reference time, even in highly occluded scenarios. It can also forecast the surrounding traffic scenes up to 2 seconds. As illustrated in the example, the white car visible at $T-1$ becomes occluded by the black SUV at T. TLCFuse captures the presence of the white car in its BEV segmentation at time T (circled), and predicts the future trajectory of the moving black SUV (light blue trace within the circle). Leveraging these output BEV maps, TLCFuse forecasts the ego-vehicle's trajectory for the next 5 seconds.

In summary, the contributions of our paper are:

1) **Novel Architecture:** TLCFuse integrates temporal multi-step spatial information derived from the fusion

[†] G. Salazar-Gomez and W. Liu contributed equally to this work.

This work was supported by Toyota Motor Europe.

[1]The authors are with Inria, Univ. Grenoble-Alpes, Grenoble, France {firstname.lastname}@inria.fr

[2] Manuel Diaz-Zapata is with CITILab, INSA Lyon.

[3] This work was conducted when David Sierra-Gonzalez was affiliated with Inria.

of LiDAR and camera sensor readings, to enhance occlusion-aware semantic BEV grid segmentation.
2) **Accuracy:** Our experimental results show that TLCFuse outperforms existing state-of-the-art methods, especially when evaluating occluded and partially-occluded vehicles on the nuScenes dataset[14].
3) **Flexibility:** Experiments on downstream tasks are conducted by applying TLCFuse to predict multi-step future BEV grids in a one-shot manner and to forecast the ego-vehicle's trajectory.
4) **Differentiability:** We offer insights on integrating TLCFuse within an end-to-end (E2E) trainable framework for full-stack driving tasks.

II. Related Work

BEV perception methods have become popular in recent years in autonomous driving. Camera-based approaches try to find the correspondence between image pixels and cell locations in the BEV grids. LSS [15] projects 2D features to 3D by inferring an implicit depth distribution over pixels. Saha et al. [16] use attention operations to associate image columns with their respective frustum projections in the BEV. PON [17] and VPN [18] use MLPs to learn correspondences across multiple scales from the image space to the BEV. Learning these associations can be computationally costly and prone to overfitting. Methods like BEVFormer [19] and CVT [20] adopt cross-attention to find correspondences between images and BEV grids. LaRa [11] uses cross-attention to efficiently aggregate camera images and camera geometries in a encoder-decoder manner. BAEFormer[13] further improves the performance by cross-integrating multi-scale input image information. Nevertheless, camera-only methods lack real depth information for accurate results, while Lidar data can compensate for this. Lidar-based methods such as PillarSegNet [21] associate the input data to each BEV cell using an orthographic projection, only taking into account the geometric structure of the scene. Yet, the sparse LiDAR points can't provide high-resolution information that is available in images. Consequently, the study of fusing multiple modalities becomes popular and promising. BEVFusion [6] processes images and point clouds separately and uses the BEV as the shared space to fuse them together. Fishing Net [7] incorporates camera, radar and LiDAR in the input to predict future BEV semantic grids. TransFuseGrid [8] also encodes images and point clouds separately into a BEV by fusing them with multi-scale self-attention transformers. LAPTNet [9] projects the point cloud to the image plane to get depth information for the projection of image features to the BEV. Building upon these foundations, this paper introduces a novel approach that seamlessly integrates temporal sequences of multi-modal data processing. This integration not only facilitates the fusion of spatial information from LiDAR and camera sources but also enhances the incorporation of temporal cues, contributing to robust perception of the surrounding scene.

Following BEVdet4D[22] and SoloFusion[23]'s success in object detection by fusing temporal image features, BEVFormer [12] performs temporal self-attention with aligned BEV features to enhance perception results. FIERY [24] performs temporal fusion of BEV features to predict vehicles' future states, became the first to combine perception and prediction in one network. Followed by BEVerse[25] that generates iterative flows for future state prediction and jointly reason object detection. Asghar et al.[26] explored the integration of prior knowledge derived from DOGMs to predict agents' motion within BEV grids. Unlike these works, we leverage temporal information to tackle challenges in occlusion scenarios for semantic segmentation. Moreover, our approach can handle future scene forecasting and ego-car trajectory prediction.

III. Approach

In this section, we introduce TLCFuse, a novel architecture for occlusion-aware semantic BEV segmentation. TLCFuse is a transformer-based encoder-decoder network that leverages spatio-temporal information acquired through sequential multi-modal sensor fusion. Unlike existing methods that rely on either recurrent networks or external memory structures for capturing temporal dependencies, TLCFuse employs a temporal-augmented attention-based encoder to create a compact representation of input data over time. To the best of our knowledge, this paper is the first to propose such an architecture. An overview of the pipeline is illustrated in Fig. 2.

A. Temporal Augmented Attention-based Encoder

Attention-based transformer models have demonstrated their efficacy in capturing spatial relationships among multiple modalities[6], [8], [9], [27], providing accurate spatial cues for scene understanding. Nonetheless, there are known limitations hindering our integration of sequential sensor readings. Transformer models struggle with large inputs like point clouds due to their quadratic scaling with input size. Additionally, their fixed-length input sequences restrict them by a fixed-length memory to capture long-term dependencies for future predictions. Some existing models try to use recurrence or convolutional layers to explicitly model temporal dependencies from the input, or use external memory structures to store and retrieve long-term information[12], [24], [28]. However, these techniques increase even more the computational cost.

In response to the aforementioned limitations, we designed a novel attention-based encoder for TLCFuse, augmenting the comprehensive spatial feature extraction capabilities of transformers by integrating temporal information. We first initialize a fixed-sized low-dimensional latent-array $L \in \mathbb{R}^{N \times M}$ with N latent vectors of dimension M, and where N is much smaller than the input feature size. This latent representation will function as a spatio-temporal memory bank, facilitating feature extraction across time through cross-attention operations. The encoder receives a temporal sequence of consecutive feature tensors as input, where each feature tensor comprises concatenated features from LiDAR, Camera and egomotion. At each time step, L is utilized to

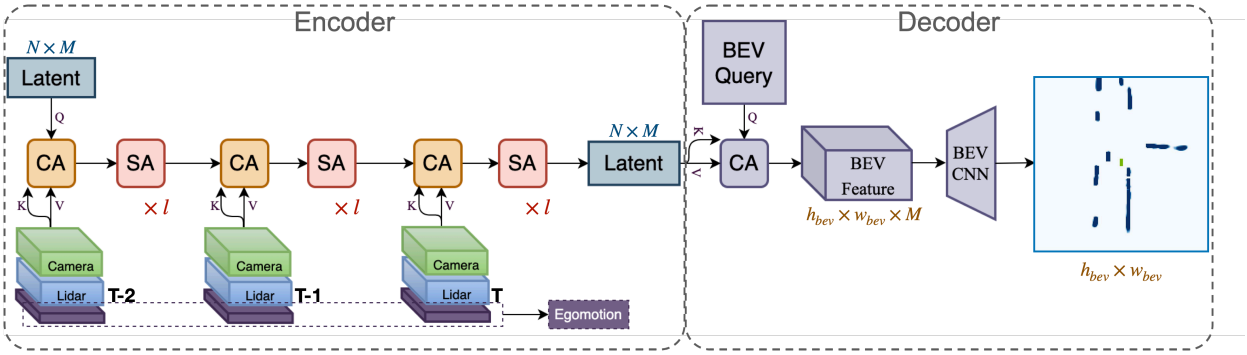


Fig. 2: Overview of our proposed approach. At the encoding stage, a sequence of consecutive feature tensors at times $T-2$, $T-1$ and T is input, where each feature tensor comprises concatenated domain-specific features from LiDAR, Camera and egomotion. A low-dimensional latent representation L is utilized to extract spatio-temporal information from the input through a cross-attention (CA) layer and a few self-attention (SA) layers sequentially. At the decoding stage, a BEV query is employed to extract information from L, which is then stored in the BEV feature vector through a cross-attention operation. Subsequently, a CNN module is applied to refine the BEV feature into the target semantic BEV grid.

capture intricate interactions across modalities and to extract spatial relationships among objects in the scene through a cross-attention layer, followed by a few self-attention layers. Subsequently, we proceed to the features of the next time step. The temporal information flow in the encoder mimics the progression of time in the real world, and such temporal awareness is crucial in maintaining a coherent understanding of object movements and relationships, addressing challenges associated with occlusions.

B. Multi-Modal Fusion

TLCFuse's encoder takes as its input sequential multi-modal fused feature tensors. At each time step, an input feature tensor comprises three geometrically informed modalities: multiple RGB camera frames, LiDAR point cloud, and the egomotion corresponding to the given reference time. We exploit the strong contextual ability of attention-based TLCFuse model to fuse these modalities in a trainable manner.

Given a set of K camera frames $F_T^k \in \mathbb{R}^{H\times W\times 3}{}_{k=1}^{K}$ at each time T, we adopt a pretrained EfficientNet [29] backbone to extract camera features $e_T^{cam} \in \mathbb{R}^{K\times H/d_f\times W/d_f\times c}$ where d_f is the down-sized factor and c is the number of feature channels. Drawing inspiration from [11], we initialize a set of positional embeddings which contain the intrinsic and the extrinsic information of the k^{th} camera, and to transform the homogeneous coordinates of each feature pixel in the corresponding k^{th} batch of e_T^{cam}, to project e_T^{cam} from the image 2D coordinate system to the 3D ego-frame coordinate system. Then, we concatenate the K transformed embeddings and pass it through a 2-layer MLP module [30] to yield the final positional embedding $\emptyset_T^{cam} \in \mathbb{R}^{K\times H/d_f\times W/d_f\times c}$, and to project e_T^{cam} to the BEV coordinates.

Given the point cloud $P_t \in \mathbb{R}^{4\times D}$ with D points taken from a LiDAR sensor on the ego-vehicle, we adopt a PointPillar [31] backbone to group the points in the 3D space into a grid of 'pillars' in the BEV plane. Within each pillar, the information of all the points that fall inside are combined to yield the LiDAR feature map $e_T^{pc} \in \mathbb{R}^{h\times w\times c}$ where (h, w) is the spatial shape of the BEV plane and c is the same number of feature channels as used in the camera features e_T^{cam}. Additionally, we concatenate a Fourier positional embedding $\emptyset_T^{pc} \in \mathbb{R}^{h\times w\times c}$ to the feature tensor e_T^{pc}, to strengthen the geometric alignment between the 3D point cloud coordinates and the BEV coordinates.

On top of these modalities, we incorporate the egomotion embedding $\emptyset_{T-h\to T}^{ego} \in \mathbb{R}^{6\times c}, h = 1, ..., H$ into our input feature tensor at each time step, where H is the temporal horizon. The egomotion represents the rotation and translation of the ego-vehicle from a past time step $T-h$ to the reference time T. For time T, the egomotion embedding $\emptyset_T^{ego}$ is a zero-filled tensor. As the ego position of the autonomous vehicle evolves over time, egomotion embedding plays a crucial role in aiding TLCFuse's temporal mechanism to dynamically adapt to changes in perspective.

C. TLCFuse's Decoder

TLCFuse's decoder consists of a BEV query and a BEV refinement network based on ResNet18 [32]. The BEV query $Q \in \mathbb{R}^{h_{bev}\times w_{bev}\times c_{bev}}$ is initialized to the spacial size of the target output BEV grid. At the decoding stage, Q attends to the latent representation L through a cross-attention, and passes the obtained BEV feature map $\mathbb{R}^{h_{bev}\times w_{bev}\times M}$ to the CNN module to be decoded into the output grid. Simply by extending the output BEV channel to $\mathbb{R}^{h_{bev}\times w_{bev}\times f}$, we can obtain a sequence of BEV predictions from time T to $T+f$ in a one-shot manner.

D. Trajectory Forecasting with TLCFuse

TLCFuse can be applied to downstream task to anticipate the future trajectory of the ego-vehicle within the scene. As shown in Fig. 3, we utilize TLCFuse to generate 5 BEV predictions for the vehicles and 1 BEV grid for the Drivable Area. We concatenate these grids as the input to a Feature Pyramid Network (FPN)[33] followed by a 1×1 convolution layer. The extracted feature receives a skip connection from the input Drivable Area map to boost the memory of the driving lane. Finally, a ResNet [32] is adopted to predict

the final trajectory and heading $\mathbb{R}^{p\times 3}$ of the ego-vehicle, where p is the prediction time horizon. In this paper, we adopt pretrained TLCFuse model in generating inputs for the Predictor. However, ongoing work is underway towards end-to-end trainable TLCfuse for motion planning to be discussed in future works.

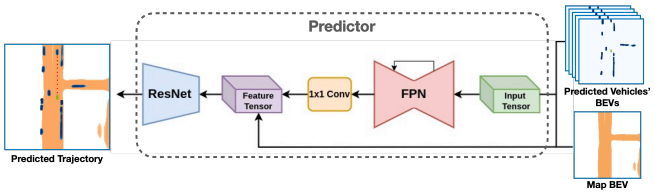


Fig. 3: We introduce a trajectory forecasting network designed to anticipate the future trajectory of the ego-vehicle. 5 BEV grid predictions of the Vehicle and 1 BEV grid of the Drivable Area serve as inputs to the predictor. The network analyzes these inputs to discern the most probable route for the ego-vehicle over the next 5 seconds.

IV. Experiments

A. Dataset

We use nuScenes [14] and nuPlan [34] datasets to evaluate our proposed framework. NuScenes consists of 1000 scenes captured from different locations and times of the day; each scene's duration is approximately 20 seconds providing data from an entire sensor suite, including LiDAR and cameras. The 6 cameras provide a 360° view field around the ego-vehicle with overlaps. The synchronized keyframes(images, LiDAR) are annotated at 2Hz and all objects come with attributes such as visibility level in the camera images. NuPlan provides more than 1200 hours of realistic human driving data with distinct and diverse traffic behaviors. It is the world's first large-scale standardized dataset for autonomous vehicle motion planning.

We use the nuScenes dataset to train and test TLCFuse for semantic grid prediction tasks. For trajectory forecasting, we first train and test our motion predictor on the nuPlan dataset, then evaluate the full pipeline end-to-end on the nuScenes validation set.

B. Metrics

For semantic grid predcition tasks, we use Intersection Over Union (IoU) to evaluate our model's performance. IoU is used to quantify the amount of overlapping between the generated semantic grids and the corresponding ground truth. To better show TLCFuse's performance under different occluded scenarios, we incorporate a set of visibility levels of the objects provided by the nuScenes dataset while computing IoUs for comparison. The visibility percentage specifies the fraction of visible pixels for an object over the whole camera rig. For nuScenes, the visibility levels are organized in 4 bins: 0-40%, 40-60%, 60-80%, 80-100%.

C. Implementation Details

We analyze sequences of input data with the temporal horizon extending over 1 second, covering three consecutive time steps from the past, aligning with the 2Hz frequency of the nuScenes dataset. At each time step, the input modalities include 6 RGB camera images with sizes scaled to 224×480 and 10 sweeps of LiDAR point cloud data.

In the encoder, the latent-array L is initialized to the size of 256×256 with zero mean, standard deviation 0.02 and value limits of (-2, 2). In the decoder, we initialize the BEV Query to the size of $200\times 200\times 256$. The target output BEV grid is 200×200 in cells, with a resolution of $50cm$ per cell, corresponding to a 100×100 meters area centered around the ego-vehicle. When predicting the future BEV grids, we set prediction horizon $f = 5$, corresponding to a 2 seconds into the future. We train TLCFuse on nuScenes training set for 25 epochs with a batch of 2, and test on the nuScenes validation set. In motion forecasting experiment, we train the proposed motion forecasting network independently on the nuPlan dataset for 76 epochs. Subsequently, we conduct an evaluation on the nuPlan validation set using TLCFuse without fine-tuning. This involves generating 5 BEV grid predictions for the vehicles and 1 BEV grid for the drivable area. These grids are then input into the motion planner network to forecast the trajectory of the ego-vehicle for 5 seconds.

All our experiments are conducted in a computer with 8 GPUs Nvidia Tesla V100.

V. Results

A. Qualitative Results

Qualitative results in Fig. 4 illustrate TLCFuse's performance in occluded scenarios. In this instance, the scene captured by the back-left camera of the ego-vehicle reveals that a black car (highlighted by the yellow circle) parked alongside the road becomes concealed behind a moving silver SUV at time T Comparing the BEV segmentation results generated by LSS [15], FIERY [24], LaRa [11], and TLCFuse, our method stands out as the only one capable of accurately and clearly recovering the occluded black car in its generated BEV map (indicated by the green box). In addition, TLCFuse also excels at perceiving objects located far from the ego-vehicle and accurately mapping them onto the Bird's-Eye View (BEV) grids (as evident within the green circles in TLCFuse's result). From this figure, it is apparent that TLCFuse's BEV semantic segmentation maps align more closely with the actual ground truth than those of other state-of-the-art methods.

B. Quantitative Results

In Tables I and II, we present quantitative results of TLCFuse in the BEV segmentation task, comparing it to previous works using the nuScenes dataset. For BEV grid segmentation, nuScenes provides multiple semantic categories, including Vehicle, Drivable Area, Human, and Walkway. Not all existing BEV segmentation approaches operate on the same categories. In Table I, we present the IoU scores specifically computed for the Vehicle category. Our scores are reported across three different visibility levels provided by nuScenes. Visibility greater than 0% (second column

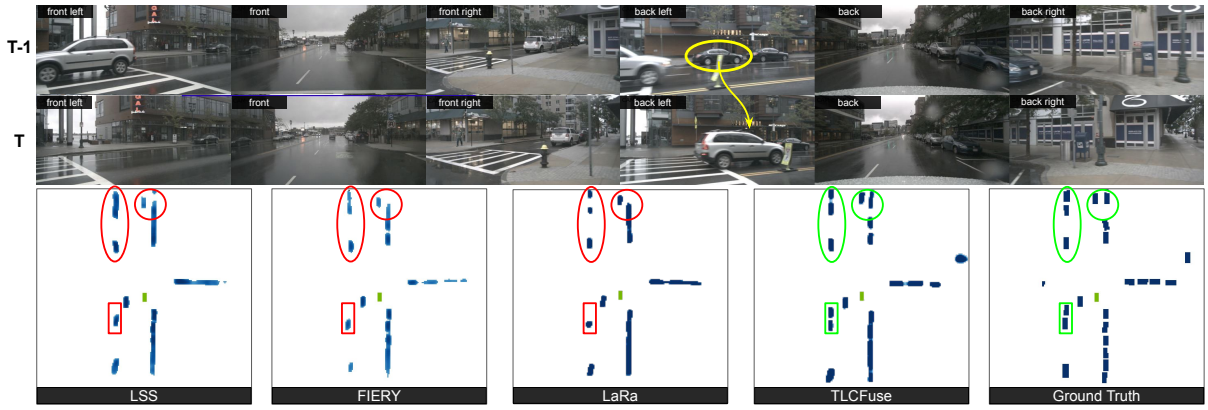


Fig. 4: Qualitative example of occlusion-aware BEV segmentation. The back-left camera of the ego-vehicle records two black cars being occluded by a silver car. Prior works LSS [15], FIERY [24] and LaRa [11] fail to generate the locations of the occluded cars in their semantic maps (indicated by a red box). While TLCFuse correctly locates the occluded vehicles (indicated by a green box). Additionally, TLCFuse generate the clearest and sharpest semantic map than the others (see the areas indicated by green circles).

Method	Modalities	vis≥0%	vis>40%	0%<vis<40%	FPS
LSS [15]	C	32.1	34.81	8.65	25
BEVFormer[19]	C	43.2	-	-	2
CVT [10]	C	-	36.0	-	48
FIERY Static[24]	C	35.8	39.26	8.54	8
LaRa[11]	C	35.81	37.92	9.41	-
BAEFormer[13]	C	36.0	38.9	-	45
TFGrid [8]	C+L	35.88	-	-	18.3
LAPTNet [9]	C+L	40.13	-	-	43.8
TLCFuse	C+L+T	**43.65**	**46.40**	**13.34**	9

TABLE I: **Vehicle BEV segmentation with different visibility levels.** TLCFuse was trained on nuScenes with all vehicles within the scenes, and different visibility levels are only considered for validation.

Method	Modalities	Human	Drivable Area	Walkway	FPS
Pyramid Occupancy Network [17]	C	8.2	60.4	31.0	22.3
LSS [15]	C	9.99	72.9	51.03	25
M2BEV [35]	C	-	75.9	-	4.3
Translating Images into Maps [16]	C	8.7	72.6	32.4	-
BEVFormer [12]	C	-	77.5	-	2
UniAD [28]	C+T	-	69.1	-	2
PointPillars baseline [31]	L	0.0	58.44	33.66	-
TFGrid [8]	C+L	-	78.87	50.98	18.3
BEVFusion [6]	C+L	-	**85.5**	**67.6**	-
LAPTNet [9]	C+L	13.8	79.43	57.25	**43.8**
TLCFuse	C+L+T	**15.24**	76.79	46.37	9

TABLE II: **BEV segmentation qualitative results on nuScenes dataset for other semantic classes**: Human, Drivable Area and Walkway. TLCFuse outperforms other models in the Human category and performs comparably with state-of-the-art models in the Drivable Area and Walkway categories.

in the table) represents all the vehicles within the scenes. Visibility greater than 40% (third column in the table) means that only the visible vehicles (not significantly occluded) in the scenes are considered. And visibility between 0% and 40% (fourth column in the table) only consists of partially or completely occluded vehicles in the scenes. We compare to the state-of-the-art methods that produce BEV segmentation results for the Vehicle category. We observe that TLCFuse consistently achieves the best scores compared to others across all different visibility levels. Our outstanding performance under visibility levels between 0% and 40% further proves that TLCFuse can provide robust perception under occluded situations.

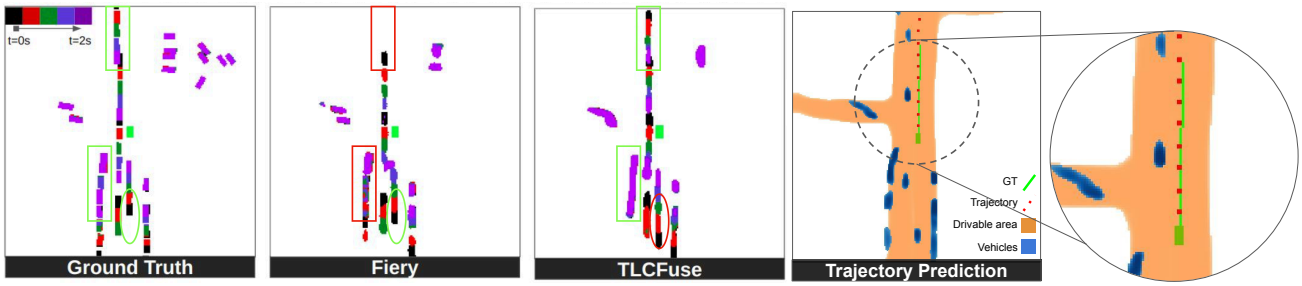
We transition from evaluating the vehicle category to assessing other categories. Specifically, we focus on three of the most commonly adopted categories: Human, Drivable Area, and Walkway. It is important to note the absence of a visibility mask for the Drivable Area and Walkway categories in nuScenes. Unlike vehicle category, drivable areas, and walkways are typically considered accessible and unobstructed for vehicle navigation. Therefore, we generate Table II without any visibility mask and compare it to state-of-the-art methods that include these three categories. From the table, we observe that TLCFuse performs admirably across the Drivable Area and Walkway categories and achieves the best results for the Human category. Additionally, we would like to note that the state-of-art model BEVFusion[6] achieves the best scores in Drivable Area and Walkway categories by processing higher-resolution camera images and focusing specific on road map generation. We would like to adopt the same strategy in TLCFuse in our future work.

C. Ablation Study

We conducted an ablation study to elucidate the individual contributions of different input modalities, with the results presented in Table III. This table provides insights into the impact of varying input modalities on the performance of our proposed network.

ID	Modality			IoU
	Camera	LiDAR	Temporal	
1	✓			35.61
2		✓		31.09
3	✓	✓		42.56
4	✓		✓	35.67
5		✓	✓	32.01
6	✓	✓	✓	**43.65**

TABLE III: **Ablation on different modalities.** We show here the BEV segmentation IoU scores using different combinations of input modalities.



(a) (b)

Fig. 5: (a) Qualitative results of one-shot multi-step future BEV grid prediction on the nuScenes dataset are presented. We compare TLCFuse's results against the state-of-the-art model Fiery[24] and the ground truth. The predictions are color-coded for 2 seconds into the future, with green boxes and circles indicating accurate predictions compared to the ground truth, and red ones representing mistakes. TLCFuse produces qualitatively satisfying predictions, showcasing its effective performance despite its naive prediction mechanism. (b) Qualitative results of trajectory forecasting on nuScenes dataset. The predicted 5seconds trajectory of the ego-vehicle is shown in red dots overlayed onto the semantic maps of the drivable area and surrounding vehicles at time 0s. The groundtruth future trajectory is drew in a green line. We see that our model predicts reasonable and accurate trajectory for the ego-vehicle.

In this table, we provide IoU scores for BEV grid segmentation on the nuScenes dataset using all the vehicles in the scene. We compare the performances given by different modality combinations: 1) only using multiple camera images; 2) only using LiDAR point cloud; 3) using camera and LiDAR fusion; 4) using a sequence of camera images; 5) using a sequence of LiDAR data; and 6) using temporally fused camera and LiDAR data. From the table, the best score was achieved by fusing temporal LiDAR and camera data, namely TLCFuse. While the fusion of LiDAR and camera already improves performance compared to using a single modality, adding temporal information further boosts the performance.

D. Motion Prediction

We applied TLCFuse to two downstream motion prediction tasks: one-shot multi-step BEV grid prediction and trajectory forecasting for the ego-vehicle. These tasks not only validate the effectiveness of our model but also highlight its flexible applicability across a wide range of autonomous driving scenarios.

We first employed TLCFuse in the BEV grid prediction task, as explained in Section III-C. Existing motion prediction approaches, such as FIERY[24] and BEVerse[25], often rely on recurrent networks (RNNs, LSTMs) to predict accurate future instances **in an iterative** fashion. However, these recurrent networks may increase the network's complexity and inference latency. In contrast, TLCFuse can be used to generate multiple predictions **in one shot** by simply adjusting its decoder stage's output channel. This allows TLCFuse to achieve a 9 FPS inference latency, which is much faster than FIERY[24] and BEVerse[25], with an average speed of 2 FPS.

We present a qualitative example in Fig. 5a. In this illustration, multi-step BEV predictions are generated in one shot. The predictions are color-coded to visualize the time range from 0s to 2s, progressing from black to purple, and are overlaid on the map of the Drivable Area generated by TLCFuse. When compared to the ground truth and the state-of-the-art model Fiery[24] on its left, TLCFuse yields qualitatively satisfying predictions. In particular, TLCFuse provides more accurate predictions for both moving and static objects in the scene than Fiery, as indicated in the green/red boxes in the figure. However, as circled in the figure, TLCFuse also makes mistakes in predicting the accurate locations of this moving vehicle.

The second downstream task we undertake involves predicting the future trajectory of the ego-vehicle. We have detailed the network and experiment designs in Sections III-D and IV-C. We pursued this task due to its significant relevance in showcasing our model's capacity to comprehend surrounding information and guide the ego-vehicle effectively. Furthermore, this task aligns with our overarching research goal of transitioning our approach towards an end-to-end framework.

One qualitative example is shown in Fig. 5b. Our model predicted the trajectory for the ego-vehicle 5seconds into the future as indicated by the red dots. To enhance result visualization, we drew the groundtruth future trajectory in a green line, and overlaid the predicted trajectory on the drivable area map and surrounding vehicles at time 0s. The model reasonably predicts the ego-vehicle following the current lane on the road with a constant speed.

For more experimental results, please refer to `https://github.com/gsg213/TLCFuse`.

VI. Conclusions and Future Work

In this paper, we present TLCFuse. A novel, flexible and differentiable architecture that extract spatio-temporal cues for occlusion-aware BEV segmentation. We conducted extensive experiments on the nuScenes datase, and our experimental results show that TLCFuse outperforms existing methods, especially under occluded scenarios. In addition, we also demonstrate the flexibility of TLCFuse by applying it to two downstream tasks: multi-step BEV prediction in

a one-shot manner and trajectory forecasting of the ego-vehicle. In future work, we will continue the ongoing work of a new end-to-end trainable pipeline integrating TLCFuse and the proposed motion planner network.

VII. Acknowledgment

Experiments presented in this paper were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see https://www.grid5000.fr).

References

[1] Y. Chen, S. Liu, X. Shen, and J. Jia, "Dsgn: Deep stereo geometry network for 3d object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 12 536–12 545.

[2] M. Liang, B. Yang, S. Wang, and R. Urtasun, "Deep continuous fusion for multi-sensor 3d object detection," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 641–656.

[3] Y. Li, A. W. Yu, T. Meng, B. Caine, J. Ngiam, D. Peng, J. Shen, Y. Lu, D. Zhou, Q. V. Le *et al.*, "Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 182–17 191.

[4] C. Wang, C. Ma, M. Zhu, and X. Yang, "Pointaugmenting: Cross-modal augmentation for 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 794–11 803.

[5] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3d object detection network for autonomous driving," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 1907–1915.

[6] Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. Rus, and S. Han, "Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," *arXiv preprint arXiv:2205.13542*, 2022.

[7] N. Hendy, C. Sloan, F. Tian, P. Duan, N. Charchut, Y. Xie, C. Wang, and J. Philbin, "Fishing net: Future inference of semantic heatmaps in grids," *arXiv preprint arXiv:2006.09917*, 2020.

[8] G. Salazar-Gomez, D. Sierra-Gonzalez, M. Diaz-Zapata, A. Paigwar, W. Liu, O. Erkent, and C. Laugier, "Transfusegrid: Transformer-based lidar-rgb fusion for semantic grid prediction," in *2022 17th International Conference on Control, Automation, Robotics and Vision (ICARCV)*. IEEE, 2022, pp. 268–273.

[9] M. A. Diaz-Zapata, D. S. González, Ö. Erkent, J. Dibangoye, and C. Laugier, "Laptnet-fpn: Multi-scale lidar-aided projective transform network for real time semantic grid prediction," *arXiv preprint arXiv:2302.06414*, 2023.

[10] B. Zhou and P. Krähenbühl, "Cross-view transformers for real-time map-view semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 760–13 769.

[11] F. Bartoccioni, E. Zablocki, A. Bursuc, P. Perez, M. Cord, and K. Alahari, "Lara: Latents and rays for multi-camera bird's-eye-view semantic segmentation," in *6th Annual Conference on Robot Learning*, 2022. [Online]. Available: https://openreview.net/forum?id=abd_D-iVjk0

[12] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Q. Yu, and J. Dai, "Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," *arXiv preprint arXiv:2203.17270*, 2022.

[13] C. Pan, Y. He, J. Peng, Q. Zhang, W. Sui, and Z. Zhang, "Baeformer: Bi-directional and early interaction transformers for bird's eye view semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 9590–9599.

[14] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *CVPR*, 2020.

[15] J. Philion and S. Fidler, "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d," in *European Conference on Computer Vision*. Springer, 2020, pp. 194–210.

[16] A. Saha, O. Mendez Maldonado, C. Russell, and R. Bowden, "Translating images into maps," *2022 IEEE International Conference on Robotics and Automation (ICRA)*, 2022.

[17] T. Roddick and R. Cipolla, "Predicting semantic map representations from images using pyramid occupancy networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 138–11 147.

[18] B. Pan, J. Sun, H. Y. T. Leung, A. Andonian, and B. Zhou, "Cross-view semantic segmentation for sensing surroundings," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 4867–4873, 2020.

[19] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Y. Qiao, and J. Dai, "Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*. Springer, 2022, pp. 1–18.

[20] B. Zhou and P. Krähenbühl, "Cross-view transformers for real-time map-view semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 760–13 769.

[21] J. Fei, K. Peng, P. Heidenreich, F. Bieder, and C. Stiller, "Pillarsegnet: Pillar-based semantic grid map estimation using sparse lidar data," in *2021 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2021, pp. 838–844.

[22] J. Huang and G. Huang, "Bevdet4d: Exploit temporal cues in multi-camera 3d object detection," *arXiv preprint arXiv:2203.17054*, 2022.

[23] J. Park, C. Xu, S. Yang, K. Keutzer, K. Kitani, M. Tomizuka, and W. Zhan, "Time will tell: New outlooks and a baseline for temporal multi-view 3d object detection," *arXiv preprint arXiv:2210.02443*, 2022.

[24] A. Hu, Z. Murez, N. Mohan, S. Dudas, J. Hawke, V. Badrinarayanan, R. Cipolla, and A. Kendall, "Fiery: future instance prediction in bird's-eye view from surround monocular cameras," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 273–15 282.

[25] Y. Zhang, Z. Zhu, W. Zheng, J. Huang, G. Huang, J. Zhou, and J. Lu, "Beverse: Unified perception and prediction in birds-eye-view for vision-centric autonomous driving," *arXiv preprint arXiv:2205.09743*, 2022.

[26] R. Asghar, M. Diaz-Zapata, L. Rummelhard, A. Spalanzani, and C. Laugier, "Vehicle motion forecasting using prior information and semantic-assisted occupancy grid maps," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 49–54.

[27] A. Jaegle, S. Borgeaud, J.-B. Alayrac, C. Doersch, C. Ionescu, D. Ding, S. Koppula, D. Zoran, A. Brock, E. Shelhamer *et al.*, "Perceiver io: A general architecture for structured inputs & outputs," *arXiv preprint arXiv:2107.14795*, 2021.

[28] Y. Hu, J. Yang, L. Chen, K. Li, C. Sima, X. Zhu, S. Chai, S. Du, T. Lin, W. Wang *et al.*, "Planning-oriented autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17 853–17 862.

[29] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[30] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.

[31] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 697–12 705.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[33] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[34] K. T. e. a. H. Caesar, J. Kabzan, "Nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles," in *CVPR ADP3 workshop*, 2021.

[35] E. Xie, Z. Yu, D. Zhou, J. Philion, A. Anandkumar, S. Fidler, P. Luo, and J. M. Alvarez, "Mˆ 2bev: Multi-camera joint 3d detection and segmentation with unified birds-eye view representation," *arXiv preprint arXiv:2204.05088*, 2022.