BEVSeg2TP: Surround View Camera Bird's-Eye-View Based Joint Vehicle Segmentation and Ego Vehicle Trajectory Prediction

Sushil Sharma[1,2], Arindam Das[1], Ganesh Sistu[1], Mark Halton[1] and Ciarán Eising[1,2]

[1]*Department of Electronic & Computer Engineering, University of Limerick, Ireland*
[2]*SFI CRT Foundations in Data Science, University of Limerick, Ireland*
{firstname.lastname}@ul.ie

Keywords: Surrounded-View Camera, Encoder-Decoder Transformer, Segmentation, Trajectory Prediction.

Abstract: Trajectory prediction is, naturally, a key task for vehicle autonomy. While the number of traffic rules is limited, the combinations and uncertainties associated with each agent's behaviour in real-world scenarios are nearly impossible to encode. Consequently, there is a growing interest in learning-based trajectory prediction. The proposed method in this paper predicts trajectories by considering perception and trajectory prediction as a unified system. In considering them as unified tasks, we show that there is the potential to improve the performance of perception. To achieve these goals, we present BEVSeg2TP - a surround-view camera bird's-eye-view-based joint vehicle segmentation and ego vehicle trajectory prediction system for autonomous vehicles. The proposed system uses a network trained on multiple camera views. The images are transformed using several deep learning techniques to perform semantic segmentation of objects, including other vehicles, in the scene. The segmentation outputs are fused across the camera views to obtain a comprehensive representation of the surrounding vehicles from the bird's-eye-view perspective. The system further predicts the future trajectory of the ego vehicle using a spatiotemporal probabilistic network (STPN) to optimize trajectory prediction. This network leverages information from encoder-decoder transformers and joint vehicle segmentation. The predicted trajectories are projected back to the ego vehicle's bird's-eye-view perspective to provide a holistic understanding of the surrounding traffic dynamics, thus achieving safe and effective driving for vehicle autonomy. The present study suggests that transformer-based models that use cross-attention information can improve the accuracy of trajectory prediction for autonomous driving perception systems. Our proposed method outperforms existing state-of-the-art approaches on the publicly available nuScenes dataset. This link is to be followed for the source code: https://github.com/sharmasushil/BEVSeg2TP/.

1 INTRODUCTION

Accurate trajectory prediction is a critical capability for autonomous driving systems, playing a pivotal role in enhancing safety, efficiency, and driving policies. This technology is increasingly vital as autonomous vehicles become more prevalent on public roads, as it enables these vehicles to anticipate the movements of various road users, including pedestrians, cyclists, and other vehicles. By doing so, autonomous vehicles can proactively plan and execute safe manoeuvres, reducing the risk of potential collisions (Li and Guo, 2021; Cheng et al., 2019) and effectively navigating through complex traffic scenarios. Moreover, trajectory prediction empowers autonomous vehicles to optimise their driving behaviour, enabling smoother lane changes (Chen et al., 2020) and seamless merging to improve overall traffic flow and reduce congestion (Wei et al., 2021). Furthermore, trajectory prediction also plays a crucial role in facilitating effective communication and interaction between autonomous vehicles, human drivers, and pedestrians. By behaving predictably, autonomous vehicles can earn the trust of other road users (Liu et al., 2021; Yang et al., 2021) and support other extended applications in the ADAS perception stack, such as pedestrian detection (Das et al., 2023; Dasgupta et al., 2022), and pose estimation (Das et al., 2022).

In this paper, we introduce an approach called BEVSeg2TP for joint vehicle segmentation and ego vehicle trajectory prediction, leveraging a bird's-eye-view perspective from surround-view cameras. Our proposed system employs a network trained on

Sharma, S., Das, A., Sistu, G., Halton, M. and Eising, C.
BEVSeg2TP: Surround View Camera Bird's-Eye-View Based Joint Vehicle Segmentation and Ego Vehicle Trajectory Prediction.
DOI: 10.5220/0012321700003660

In *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2024) - Volume 4: VISAPP*, pages 25-34
ISBN: 978-989-758-679-8; ISSN: 2184-4321

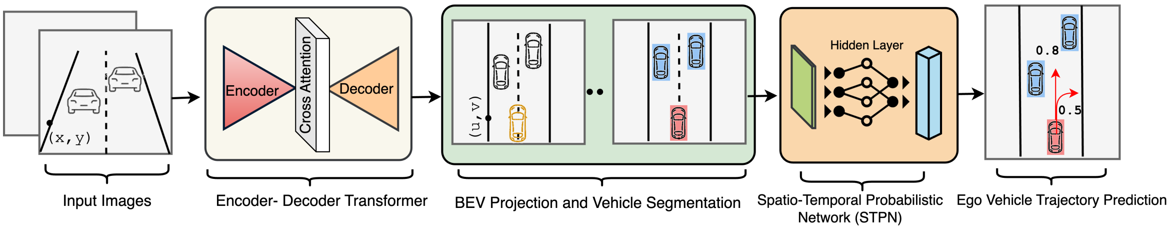
(x,y)
Encoder
Cross Attention
Decoder
(u,v)
Hidden Layer
0.8
0.5
Input Images
Encoder- Decoder Transformer
BEV Projection and Vehicle Segmentation
Spatio-Temporal Probabilistic Network (STPN)
Ego Vehicle Trajectory Prediction

Figure 1: Our proposed **BEVSeg2TP framework - surround-view camera joint vehicle segmentation and ego vehicle trajectory prediction in bird's-eye-view** approach consists of an encoder-decoder transformer, BEV projection module followed by segmentation outputs fed to the spatio-temporal probabilistic network to produce ego vehicle trajectory prediction.

surround-view or multi-camera view from the host vehicle, which it transforms into bird's-eye-view imagery of the surrounding context. These images undergo deep learning-driven processes to perform semantic segmentation on objects, including neighboring vehicles within the scene. The segmentation outcomes are then amalgamated across camera perspectives to generate a comprehensive representation of the surrounding vehicles from a bird's-eye-view perspective (Zhou and Krähenbühl, 2022). Building upon this segmented data, the proposed system also anticipates the future trajectories of the host vehicle using a spatio-temporal probabilistic network (STPN) (Cui et al., 2019). The STPN learns the spatiotemporal patterns of vehicle motion from historical trajectory data. The predicted trajectories are then projected back to the ego vehicle's bird's-eye-view perspective to provide a holistic understanding of the surrounding traffic dynamics. Figure 1 represents the overarching depiction of our approach. Our principal contributions to the BEVSeg2TP proposal are:

- Our proposed deep architecture offers an approach to jointly accomplish vehicle segmentation and ego vehicle trajectory prediction tasks by combining and adapting the works of (Zhou and Krähenbühl, 2022; Phan-Minh et al., 2020; Cui et al., 2019).
- We propose enhancements to the capabilities of the current encoder-decoder transformer used in the spatio-temporal probabilistic network (STPN) for optimizing trajectory prediction.
- We implemented an end-to-end trainable surround-view camera bird's-eye-view-based network that achieves state-of-the-art results on the nuScenes dataset (Caesar et al., 2020) when jointly trained with segmentation.

2 PRIOR ART

Joint vehicle segmentation and ego vehicle trajectory prediction using a surround or multi-camera bird's-eye view is currently an emerging area of research with several motivating factors. Firstly, working on this problem could help advance the field and contribute to the development of more effective and accurate autonomous driving systems. The potential uses of precise vehicle segmentation and predictions for ego vehicle trajectories are vast, encompassing domains such as self-driving vehicles, intelligent transportation systems, and automated driving systems, among others.

Moreover, this problem is complex and challenging, requiring the integration of information from multiple sensors and camera views. Addressing the technical challenges of this problem, such as designing effective deep learning models or developing efficient algorithms, could be a motivating factor for researchers interested in solving complex and challenging problems. Our primary focus is on enhancing map-view segmentation. It is undeniable that extensive research has been conducted in this field, which lies at the convergence of 3D recognition (Ma et al., 2019; Lai et al., 2023; Manhardt et al., 2019), depth estimation (Eigen et al., 2014; Godard et al., 2019; Ranftl et al., 2020; Zhou et al., 2017), and mapping (Garnett et al., 2019; Sengupta et al., 2012; Zhu et al., 2021).

These are the key areas that can facilitate segmentation construction and improvement. While trajectory prediction or motion planning for autonomous systems is crucial, we acknowledge the need to consider various aspects of the vehicle state, such as current position and velocity, road geometry (Lee and Kim, 2016; Wiest et al., 2020; Wu et al., 2017), other vehicles, environmental factors, and driver behaviour (Zhang et al., 2020; Abbink et al., 2017; McDonald and Mazumdar, 2020). The architecture previously described by the authors (Sharma et al., 2023) explores the utilization of the CNN-LSTM model for predicting trajectories, covering unique scenarios like pedestrians crossing roads. While the model adeptly comprehends these scenarios, it adheres to a model-driven methodology, thereby carrying inherent limitations. In our pursuit to address these limitations and devise an alternative approach, we propose the integration of a transformer-based model into our trajectory prediction methodology. Our strategy entails a partial adoption of the principles from CoverNet (Phan-Minh et al., 2020), albeit with notable distinctions. CoverNet's trajectory prediction relies on raster maps, whereas our model pivots towards real-time map view representations.

3 PROPOSED METHODOLOGY

In this section, we present BEVSeg2TP - our proposed deep architecture designed to efficiently achieve both vehicle segmentation and ego vehicle trajectory prediction tasks simultaneously. The proposed method, as depicted in Figure 2, utilizes multiple cameras to create a comprehensive view of the environment around the ego vehicle, improving ego vehicle and object segmentation, based on the work presented by (Zhou and Krähenbühl, 2022). We extend this transformer technique to incorporate trajectory prediction using a spatio-temporal probabilistic network to calculate path likelihoods, as presented in (Phan-Minh et al., 2020; Cui et al., 2019). This approach combines multiple sources of information for more accurate future trajectory predictions, enhancing self-driving car safety and performance by jointly learning the segmentation and the trajectory prediction.

3.1 Surround-View Camera Inputs

The dataset used in this paper is nuScenes (Caesar et al., 2020). It consists of six cameras located on the vehicle, providing a 360° field of view. All cameras in each scene have extrinsic (R,t) and intrinsic K calibration parameters provided at every timestamp; the intrinsic parameters remain unchanged with time. Other perception sensors in the nuScenes dataset (radar and lidar) are not used in this work.

3.2 Image Encoder

We use the simple and effective encoder-decoder architecture for map-view semantic segmentation from (Zhou and Krähenbühl, 2022). In summary, the authors proposed an image encoder that generates a multi-scale feature representation $\{\phi\}$ for each input image, which is then combined into a shared map-view representation using a cross-view cross-attention mechanism. This attention mechanism utilizes a positional embedding $\{\delta\}$ to capture both the geometric structure of the scene, allowing for accurate spatial alignment, and the sequential information between different camera views, facilitating temporal understanding and context integration. All camera-aware positional embeddings are presented as a single key vector $\delta = [\delta_1, \delta_2 ...... \delta_6]$. Image features are combined into a value vector $\phi = [\phi_1, \phi_2 .....]$. Both are merged to create a comparison of attention keys and subsequently, a softmax-cross attention is used (Vaswani et al., 2017).

3.3 Cross Attention

As illustrated in Figure 2, the cross-view transformation component aims to establish a connection between a map view and image features, as presented by (Zhou and Krähenbühl, 2022). To summarise, precise depth estimation is not learned; rather, the transformer learns a depth proxy through positional embedding $\{\delta\}$ (x^{world} remains ambiguous). The cosine similarity is used to express the geometric relationship between the world and unprojected image coordinates:

$$\cos(\theta) = \frac{\left(R_k^{-1} K_k^{-1} x^{image}\right) \cdot \left(x^{world} - t_k\right)}{\|R_k^{-1} K_k^{-1} x^{image}\| \|x^{world} - t_k\|} \quad (1)$$

where denoted as $x^{image} \in \mathbb{P}^3$ is a homogeneous image point for a given world coordinate $x^{world} \in \mathbb{R}^3$. The cosine similarity traditionally relies on precise world coordinates.

However, in this approach, the cosine similarity is augmented with positional embeddings, thus having the capability to learn both geometric and appearance features (Zhou and Krähenbühl, 2022). Direction vectors $d_{k,i} = R_k^{-1} K_k^{-1} x_i^{image}$ are created for each image coordinate x_i^{image}, serving as a reference point in world coordinates. An MLP is used to convert the direction vector $d_{k,i}$ into a D-dimensional positional embedding denoted as $\delta_{k,i} \in \mathbb{R}^D$ (Per (Zhou and Krähenbühl, 2022), we have set the value of D to 128).

3.4 Joint Vehicle Segmentation

To enhance the vehicle segmentation, we have designed our segmentation head to be simple, utilizing a series of convolutions on the bird's-eye view (BEV) feature. Specifically, it consists of four 3×3 convolutions followed by a 1×1 convolution, resulting in

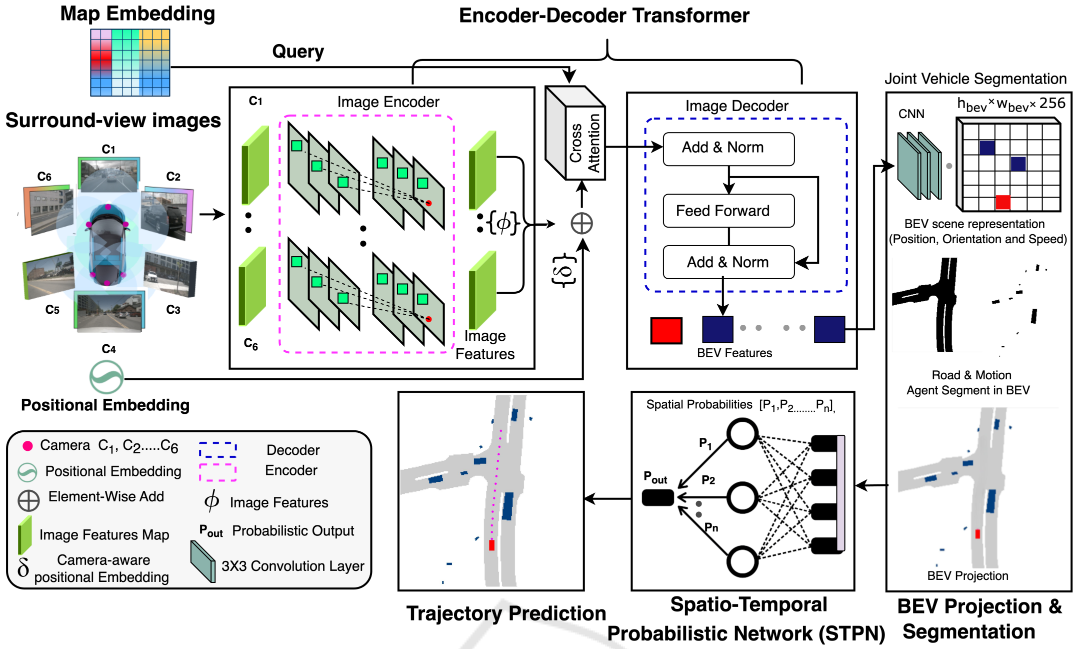
Map Embedding
Encoder-Decoder Transformer
Query
Surround-view images
C1
C2
C3
C4
C5
C6
Image Encoder
C1
C6
Image Features
:{ϕ}
{δ}
Cross Attention
Image Decoder
Add & Norm
Feed Forward
Add & Norm
BEV Features
Joint Vehicle Segmentation
CNN
$h_{bev} \times w_{bev} \times 256$
BEV scene representation (Position, Orientation and Speed)
Road & Motion Agent Segment in BEV
Positional Embedding
Camera C1, C2.....C6
Positional Embedding
Element-Wise Add
Image Features Map
δ Camera-aware positional Embedding
Decoder
Encoder
ϕ Image Features
Pout Probabilistic Output
3X3 Convolution Layer
Spatial Probabilities [P1,P2........Pn],
P1
P2
Pn
Pout
BEV Projection
Trajectory Prediction
Spatio-Temporal Probabilistic Network (STPN)
BEV Projection & Segmentation

Figure 2: **Our proposed BEVSeg2TP architecture:** Joint vehicle segmentation and ego vehicles trajectory prediction involves extracting image features $\{\phi\}$ at multiple scales and using a camera-aware positional embedding $\{\delta\}$ to account for perspective distortion. We then use map-view positional embedding and cross-attention layers to capture contextual information from multiple views and refine the vehicle segmentation. This segmentation information is then used as input to a spatio-temporal probabilistic network (STPN) for trajectory prediction based on the surrounding environment.

a BEV tensor of size $h \times w \times n$, where n represents the number of categories. In our case, we set n to 1, as we focus solely on the vehicles and other agents related to it following the approach used in the cross-view transformer (Zhou and Krähenbühl, 2022). To enhance road and vehicle segmentation in the dataset using an encoder-decoder transformer, we employ the following equation:

$$y = f(X1, X2)$$

where y is the output segmentation map, $X1$ is the input image from one sensor modality (e.g., camera), and $X2$ is the input image from another sensor modality (e.g., map information). f is the cross-view transformer, which learns to combine the information from the two modalities to produce a more accurate segmentation map. The cross-attention mechanism can be implemented using the following equation:

$$M = \text{softmax}\left(\frac{Q.(K^T)}{\sqrt{d_k}}\right)V \tag{2}$$

where Q, K, and V are the queries, keys, and values, respectively, for each modality. The dot product between the queries and keys is present in the form of $Q.(K^T)$ is divided by the square root of the dimensionality of the key vectors (d_k) to prevent the dot product from becoming too large. Subsequently, the obtained attention weights are employed to weigh the values associated with each modality. These weighted values are then combined to generate the output feature map M.

3.5 Spatio-Temporal Probabilistic Network (STPN)

This section describes the Spatio-temporal probabilistic network for trajectory prediction of the future states of an ego vehicle and a high-definition map, assuming that we have access to the state outputs of an object detection and tracking system of sufficient quality for autonomous vehicles, based on (Phan-Minh, 2021). The agents that an ego vehicle interacts with at time t are denoted by the set I_t, and s_t^i represents the state of agent $i \in I_t$ at time t. The discrete-time trajectory of agent i for times $t = (m, \ldots\ldots, n)$ is denoted by $s_{m:n}^i = [s_m^i, \ldots\ldots, s_n^i]$, where $m < n$ and $i \in I_t$.

Additionally, we presume that the high-definition map, as depicted in our proposed method, will be accessible. This includes lane geometry, crosswalks, drivable areas, and other pertinent information. The scene context over the past m steps, which includes the map and partial history of ego vehicles, is denoted by $C = \left\{\bigcup_i s_{t-m:t}^i ; \texttt{Map Information}\right\}$.

Our architecture follows the trajectory prediction layer with the approach presented in (Cui et al., 2019). To achieve effectiveness in this domain, we employ ResNet-50 (Table:1) (He et al., 2016), as recommended by previous research (Cui et al., 2019; Chai et al., 2019). Although our network currently generates predictions for one agent at a time, our approach has the potential to predict for multiple agents simultaneously in a manner similar to (Chai et al., 2019). However, we limit our focus to single-agent predictions (as in (Cui et al., 2019)) to streamline the paper and emphasize our primary contributions. To represent probabilistic trajectory predictions in multiple modes, we utilize a classification technique that selects the relevant trajectory set based on the agent of interest and scene context C. The softmax distribution is employed, as is typical in classification literature. Specifically, the probability of the k-th trajectory is expressed as follows:

$$p(s^k_{t:t+N}|x) = \frac{\exp f_k(x)}{\sum_i \exp f_i(x)} \tag{3}$$

where $f_i(x) \in \mathbb{R}$ is the output of the network of probabilistic layer. We have implemented Multi-Trajectory Prediction (MTP) (Cui et al., 2019) with adjustments made for our datasets. This model forecasts a set number of trajectories (modes) and determines their respective probabilities. Note that we are now focusing on single trajectory prediction (STP)(Djuric et al., 2020).

3.6 Loss Function

The loss function employed for vehicle segmentation in our transformer-based model is defined as follows:

$$\mathcal{L}_{\text{seg}}(\mathbf{m}, \hat{\mathbf{m}}) = -\frac{1}{N}\sum_{i=1}^{N}\big[m_i \cdot \log(p(\hat{m}_i)) + (1-m_i)\cdot \log(1-p(\hat{m}_i))\big] \tag{4}$$

where, $\mathcal{L}_{\text{seg}}(\mathbf{m}, \hat{\mathbf{m}})$ is the binary cross-entropy loss (Jadon, 2020) for vehicle segmentation, $\mathbf{m}$ is the input tensor, and $\hat{\mathbf{m}}$ is the target tensor for all N points. This loss function is particularly valuable for binary classification challenges where our model generates logits (unbounded real numbers) as output. It facilitates the computation of the binary cross-entropy loss concerning binary target labels $\hat{\mathbf{m}}$, ensuring effective training and performance evaluation for vehicle segmentation in our transformer-based approach.

In terms of trajectory prediction, the loss function we are considering is one of the most commonly used: the mean squared error (MSE). This loss function typically involves measuring the dissimilarity between the predicted and the ground-truth trajectories.

$$\mathcal{L}_{traj} = \frac{1}{N}\sum_{i=1}^{N} ||\hat{y}_i - y_i||_2^2 \tag{5}$$

Here, N is the number of training examples, $\hat{y}_i$ is the predicted trajectory for ego vehicle i, and y_i is the corresponding ground truth trajectory. The squared difference between the two trajectories is calculated element-wise and then averaged across all elements in the trajectory. The resulting value is the mean squared error loss, which measures the overall performance of the model in predicting the trajectories for the ego vehicle.

Our final loss function $\mathcal{L}_{\text{total}}$ constitutes two components, as shown in the equation below.

$$\mathcal{L}_{\text{total}} = \alpha\mathcal{L}_{\text{seg}} + \beta\mathcal{L}_{\text{traj}} \tag{6}$$

Gradients are mutually shared by both tasks till the initial layers of the network. In the above equation, α and β are the hyperparameters to balance between segmentation and trajectory prediction losses.

4 EXPERIMENTAL SETUP

4.1 Dataset

Experiments are carried out on the nuScenes dataset (Caesar et al., 2020), which comprises 1000 video sequences gathered in Boston and Singapore. The dataset is composed of scenes that have a duration of 20 seconds and consist of 40 frames each, resulting in a total of 40k samples. The dataset is divided into training, validation, and testing sets, with 700, 150, and 150 scenes respectively. The recorded data provides a comprehensive 360° view of the surrounding area around the ego-vehicles and comprises six camera perspectives. Note that we are employing identical train-test-validation splits as those used in the previous works (Zhou and Krähenbühl, 2022; Philion and Fidler, 2020) for comparison.

4.2 Transformer Architecture and Implementation Details

The initial step of the network involves creating a camera-view representation for each input image. To achieve this, we utilize EfficientNet-B4 (Tan and Le, 2019) as the feature extractor and input each image I_i to obtain a multi-resolution patch embedding $\{\delta^1_1, \delta^2_1, \delta^3_1, ...\delta^R_n\}$, where R denotes the number of resolutions that are taken into account.

According to our experimental findings, accurate results can be achieved when using $R = 1$ resolution. However, if we were to increase the value of R to 2, as suggested by CVT in (Zhou and Krähenbühl, 2022), the camera-view representation for each input image in the network would incorporate additional information, such as BEV features. While this has the potential to result in a more detailed representation of the input images, it also comes with drawbacks, including increased computational requirements and a higher risk of overfitting.

The processing for each resolution is carried out individually, beginning with the lowest resolution. We employ cross-view attention to map all image features to a map-view and refine the map-view embedding, repeating this procedure for higher resolutions. In the end, we employ three up-convolutional layers to produce the output at full resolution. Once we obtain the full-resolution output, we input the ego vehicle features, which have a resolution of $h_{bev} \times w_{bev} \times 256$, into the probabilistic function for trajectory forecasting, resulting in the set of trajectories $[p_1, p_2, p_3, ..., p_n]$. Subsequently, we refine and obtain the probabilistic value, which represents our final trajectory.

To implement the architecture, we employ a pretrained EfficientNet-B4 (Tan and Le, 2019) that we fine-tune. The two scales, $(28, 60)$ and $(14, 30)$, correspond to an 8× and 16× downscaling, respectively. For the initial map view positional embedding, we use a tensor of learned parameters with dimensions $w \times h \times D$, where D is set to 128. To ensure computational efficiency, we limit the grid size to $w = h = 25$, as the cross-attention function becomes quadratic in growth with increasing grid size. The encoder comprises two cross-attention blocks, one for each scale of patch features, which utilize multi-head attention with 4 heads and an embedding size of $d_{head} = 64$.

The decoder includes three layers of bilinear upsampling and convolution, each of which increases the resolution by a factor of 2 up to the final output resolution of 200×200, corresponding to a 100×100 meter area around the ego-vehicle. The map-view representation obtained through the cross-attention transformer is passed through the joint vehicle segmentation module to accurately identify the vehicle's segmentation. This segmentation is then utilized as input to the Spatial-Temporal Probabilistic Network (STPN), which offers probabilistic predictions. Instead of providing a single deterministic trajectory, the network offers a probability distribution over possible future trajectories. This information aids in identifying the motion planning of the ego vehicle. Precisely segmenting the pixels corresponding to the ego vehicle enables the system to more accurately estimate its position, speed, and orientation in relation to other objects in the environment. This, in turn, facilitates improved decision-making during navigation. Figure 2 offers a comprehensive overview of this architecture.

5 ABLATION STUDY

We perform a detailed ablation experiment to assess the influence of several factors on the functionality of our segmentation model. We specifically examined the impacts of various backbone models and loss functions.

Table 1: Comparison study of **different standard backbone models employed for trajectory prediction** on nuScenes dataset (Caesar et al., 2020).

Backbone	# Params. (M)	Features	MSE ↓
EfficientNet-80	1.9	1280	0.3385
DenseNet-121	1.7	1024	0.2079
ResNet-50	**1.4**	**512**	**0.1062**

We performed an ablation on different backbone models to investigate their impact on the performance of our target task on the nuScenes dataset, as presented in Table 1. Notably, the ResNet-50 backbone, with 1.4 million trainable parameters and a feature size of 512, demonstrated promising results, achieving the lowest MSE of 0.1062. It is likely that ResNet-50 works well for trajectory prediction on the nuScenes dataset, as its model parameters align well with the characteristics of that dataset.

Table 2: Ablation on **different loss functions for segmentation task** on the nuScenes dataset (Caesar et al., 2020).

Loss Function	No. of Class	Loss ↓
Binary Cross Entropy	2	0.1848
Binary Focal Loss	2	0.2758

In our task, we utilize the binary cross-entropy loss function, which aligns well with the inherent characteristics of our standard binary classification problem. Additionally, we explore and compare alternative loss functions, including binary focal loss. However, our findings indicate that the binary cross-entropy loss function yields superior results, as presented in Table 2. This is primarily attributed to the balanced distribution of classes within our dataset, which favors the effectiveness of binary cross-entropy in accurately modeling the classification problem.

Table 3: Comparison of **visibility-based methods for Setting 1 and Setting 2**, where our method achieves the highest visibility rate among those with visibility greater than 40%.

	Visibility $> 40\%$	
Method	Setting 1	Setting 2
LSS (Philion and Fidler, 2020)	-	32.1
CVT (Zhou and Krähenbühl, 2022)	37.5	36.0
BEVSeg2TP (Ours)	**37.8**	**37.9**

Table 4: Comparison of **vehicle segmentation performance on the nuScene dataset** using different methods, including LSS, CVT, and our proposed method. Results are presented in terms of Intersection over Union (IoU) scores.

Method	Resolution **R**	Vehicle ↑
LSS (Philion and Fidler, 2020)	-	32.1
CVT (Zhou and Krähenbühl, 2022)	2	36.0
BEVSeg2TP (Ours)	1	**37.9**

Table 5: Comparison of the **Minimum Average Prediction Error (MinADE) and Final Displacement Error (MinFDE)** for Competing Methods on the nuScenes Dataset, over a Prediction Horizon of 6 Seconds.

Method	$\text{MinADE}_5 \downarrow$	$\text{MinADE}_{10} \downarrow$	$\text{MinADE}_{15} \downarrow$	$\text{MinFDE}_5 \downarrow$	$\text{MinFDE}_{10} \downarrow$	$\text{MinFDE}_{15} \downarrow$
Const Vel and Yaw	4.61	4.61	4.61	11.21	11.21	11.21
Physics oracle	3.69	3.69	3.69	9.06	9.06	9.06
CoverNet (Phan-Minh et al., 2020)	2.62	1.92	1.63	11.36	-	-
Trajectron++ (Salzmann et al., 2020)	1.88	1.51	-	-	-	-
MTP (Cui et al., 2019)	2.22	1.74	1.55	4.83	3.54	3.05
MultiPath (Chai et al., 2019)	1.78	1.55	1.52	**3.62**	2.93	2.89
BEVSeg2TP (Ours)	**1.63**	**1.29**	**1.15**	3.85	**2.13**	**1.65**

6 RESULTS

We evaluate the BEV map representation and trajectory planning of the BEVSeg2TP model on the publicly available nuScenes dataset. The evaluation is conducted in two different settings - 'Setting 1' refers to a $100m \times 50m$ grid with a $25cm$ resolution, while 'Setting 2' refers to a $100m \times 100m$ grid with a $50cm$ resolution. During training and validation, vehicles with a visibility level above the predefined threshold of 40% are considered. Table 3 demonstrates the comparison of our proposed approach with other existing works such as LSS (Philion and Fidler, 2020) and CVT (Zhou and Krähenbühl, 2022).

First, we compare the BEV segmentation obtained from various methods, including LSS and CVT with the results from our proposed BEVSeg2TP. Accurately predicting the future motion of vehicles is critical, as it helps the model gain a comprehensive understanding of the environment by capturing the spatial relationships among pedestrians, vehicles, and obstacles. However, our second contribution focuses on improving map-view segmentation of vehicles. Our experimental findings show that employing a resolution of $R = 1$ yields promising results. However, increasing the value of R to 2, as recommended by CVT, would lead to the camera-view representation for each input image in the network losing information, such as BEV features. We conducted further evaluations using various methods, as illustrated in Table 4.

As shown in Table 5, the ablation study has been evaluated by comparing it with four baselines: (Cui et al., 2019) (Chai et al., 2019) (Phan-Minh et al., 2020) and (Salzmann et al., 2020) and two physics-based approaches. These four baselines are a recently proposed model which is considered to be the current state-of-the-art for multimodel trajectory prediction. This comparison aims to assess the effectiveness and accuracy of our model in predicting trajectories in comparison to existing models. The goal is to determine if our model performs better than or at least as well as the state-of-the-art baseline model. By doing so, we can gain insight into the strengths and weaknesses of our model and identify areas for further improvement. To evaluate the performance of our model on the nuscenes dataset, we first obtained the output trajectories $\left[y_1, y_2, y_3, \ldots. y_n\right]$. We evaluated the performance of the model on this specific dataset for different values of K, where K was set to 5, 10, and 15 respectively.

$$\mathbf{MinADE_k} = \min_{i \in \{1 \ldots K\}} \frac{1}{T_f} \sum_{t=1}^{T_f} \left\| y_t^{\text{gt}} - y_t^{(i)} \right\|_2 \tag{7}$$

To train the model, we minimized the minimum

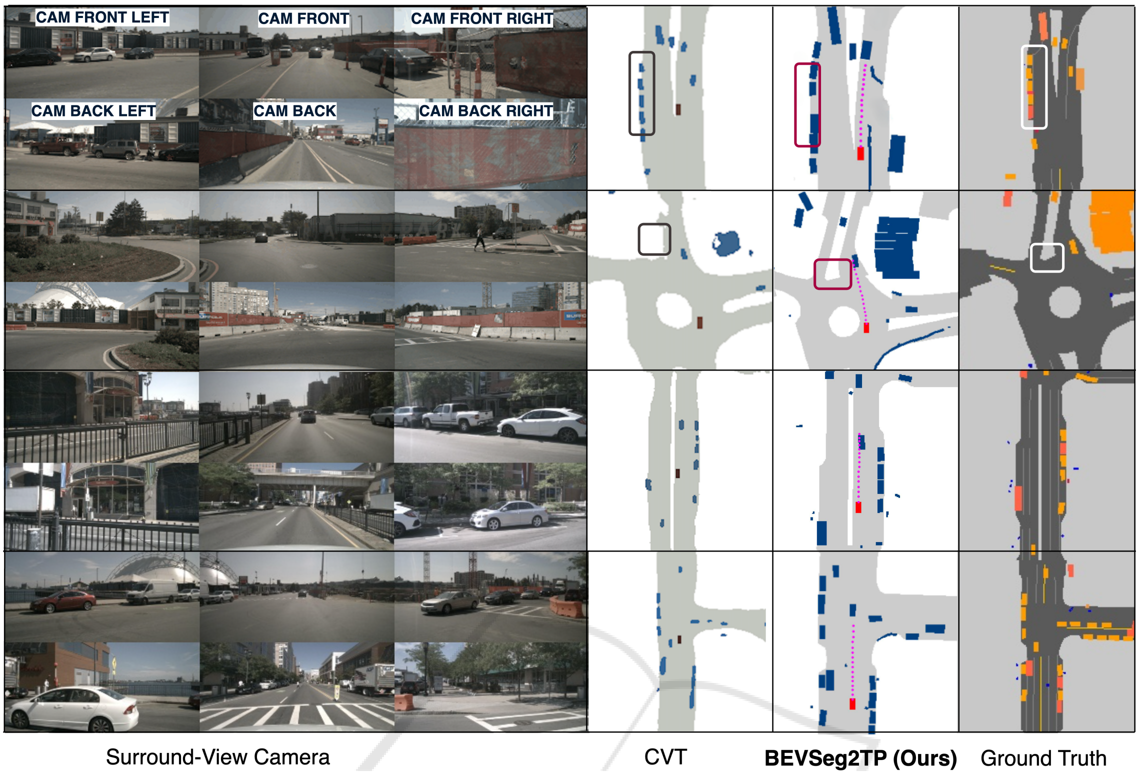

Figure 3: **Qualitative results of BEVSeg2TP model for joint vehicle segmentation and ego vehicle trajectory prediction:** Six camera views around the vehicle (top three facing forward, bottom three facing backwards) with ground truth segmentation on the right. Our trajectory prediction with improved map-view segmentation (second from right) compared to the CVT method (third from right).

average displacement error over K (MinADE$_k$) on the training set. In other words, we aimed to reduce the error between the predicted trajectories and the actual trajectories by minimizing the minimum distance between them for each of the K time steps. This method allowed us to improve the accuracy of our model's predictions and ensure that it performs well on the nuScenes dataset. Here, y_t^{gt} represents the ground truth position of the object at the final time step T, and $y_t^{(i)}$ represents the predicted position of the object at the final time step T for the i_{th} trajectory in the set of K trajectories.

We took the output trajectories $\left[y_1, y_2, y_3, \ldots . y_n\right]$ and we used $K = 15$ for nuscenes datasets. we minimize the minimum over K average displacement error (MinADE$_k$) over the training set. As depicted in Figure 3, on the left-hand side of the image, there are six camera views surrounding the vehicle. The top three views are oriented forward, while the bottom three views face backwards. On the right side of the image, there is ground truth segmentation for reference. Moving from right to left, the second image from the right displays our trajectory prediction, along with improved map-view segmentation for vehicles. Lastly, the third image from the right illustrates the CVT (Zhou and Krähenbühl, 2022) method, which we use to conduct a comparison and present the results.

The **black** color corresponds to the results obtained using a model called CVT, the red color corresponds to the results obtained using our model, and the white color corresponds to the nuScenes ground truth, which is the true segmentation of the images. The purpose of the comparison was to evaluate the performance of the other model and compare it with our model. Figure 3 reveals that our model performs well compared to the other model in both vehicle and road segmentation tasks. When it comes to vehicle segmentation, our model demonstrates a high level of accuracy in identifying the precise positions of vehicles within the image. In contrast, the other model exhibits a slightly lower level of accuracy in this regard. This distinction is clearly visible in the accompanying figure, where the red markings, representing the outcomes produced by our model, closely align with the green markings, representing the ground truth, in comparison to the black markings, which correspond to the results generated by the other model. Similarly, with regard to road segmentation, our model also exhibits decent performance. To gain further insights, additional results can be explored via the following link: https://youtu.be/FNBMEUbM3r8.

7 CONCLUSION

In this paper, we propose BEVSeg2TP - a surround-view camera bird's-eye-view-based joint vehicle segmentation and ego vehicle trajectory prediction using encoder-decoder transformer-based techniques that have shown promising results in achieving safe and effective driving for autonomous vehicles. The system processes images from multiple cameras mounted on the vehicle, performs semantic segmentation of objects in the scene, and predicts the future ego vehicle trajectory of surrounding vehicles using a combination of transformer and spatio-temporal probabilistic network (STPN) to calculate the trajectory. The predicted trajectories are projected back to the ego vehicle's bird's-eye-view perspective, providing a comprehensive understanding of the surrounding traffic dynamics. Our findings underscore the potential benefits of employing transformer-based models in conjunction with spatio-temporal networks, highlighting their capacity to significantly enhance trajectory prediction accuracy. Ultimately, these advancements contribute to the overarching goal of achieving a safer and more efficient autonomous driving experience.

While the camera configuration of nuScenes is important, it is not a typical commercially deployed surround-view system. Commercial surround view systems, used for both viewing and vehicle automation and perception tasks (Kumar et al., 2023; Eising et al., 2022), typically employ a set of four fisheye cameras around the vehicle. In the future, we intend to apply the methods discussed here to fisheye surround-view camera systems.

ACKNOWLEDGEMENTS

This publication has emanated from research supported in part by a grant from Science Foundation Ireland under Grant number 18/CRT/6049. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

REFERENCES

Abbink, D. A., Mulder, M., and de Winter, J. C. (2017). Driver behavior in automated driving: Results from a field operational test. *Transportation Research Part F: Traffic Psychology and Behaviour*, 45:93–106.

Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., and Beijbom, O. (2020). nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631.

Chai, Y., Sapp, B., Bansal, M., and Anguelov, D. (2019). Multipath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction. *arXiv preprint arXiv:1910.05449*.

Chen, D., Jiang, L., Wang, Y., and Li, Z. (2020). Autonomous driving using safe reinforcement learning by incorporating a regret-based human lane-changing decision model. In *2020 American Control Conference (ACC)*, pages 4355–4361. IEEE.

Cheng, H., Wang, Y., and Wu, J. (2019). Research on the design of an intelligent vehicle collision avoidance system. In *Journal of Physics: Conference Series*, volume 1239, page 012096.

Cui, H., Radosavljevic, V., Chou, F.-C., Lin, T.-H., Nguyen, T., Huang, T.-K., Schneider, J., and Djuric, N. (2019). Multimodal trajectory predictions for autonomous driving using deep convolutional networks. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 2090–2096. IEEE.

Das, A., Das, S., Sistu, G., Horgan, J., Bhattacharya, U., Jones, E., Glavin, M., and Eising, C. (2022). Deep multi-task networks for occluded pedestrian pose estimation. *Irish Machine Vision and Image Processing Conference*.

Das, A., Das, S., Sistu, G., Horgan, J., Bhattacharya, U., Jones, E., Glavin, M., and Eising, C. (2023). Revisiting modality imbalance in multimodal pedestrian detection. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 1755–1759.

Dasgupta, K., Das, A., Das, S., Bhattacharya, U., and Yogamani, S. (2022). Spatio-contextual deep network-based multimodal pedestrian detection for autonomous driving. *IEEE transactions on intelligent transportation systems*, 23(9):15940–15950.

Djuric, N., Radosavljevic, V., Cui, H., Nguyen, T., Chou, F.-C., Lin, T.-H., Singh, N., and Schneider, J. (2020). Uncertainty-aware short-term motion prediction of traffic actors for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2095–2104.

Eigen, D., Puhrsch, C., and Fergus, R. (2014). Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27.

Eising, C., Horgan, J., and Yogamani, S. (2022). Near-field perception for low-speed vehicle automation using surround-view fisheye cameras. *IEEE Transactions on Intelligent Transportation Systems*, 23(9):13976–13993.

Garnett, N., Cohen, R., Pe'er, T., Lahav, R., and Levi, D. (2019). 3d-lanenet: end-to-end 3d multiple lane detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2921–2930.

Godard, C., Mac Aodha, O., Firman, M., and Brostow, G. J. (2019). Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Jadon, S. (2020). A survey of loss functions for semantic segmentation. In *2020 IEEE conference on computational intelligence in bioinformatics and computational biology (CIBCB)*, pages 1–7. IEEE.

Kumar, V. R., Eising, C., Witt, C., and Yogamani, S. K. (2023). Surround-view fisheye camera perception for automated driving: Overview, survey & challenges. *IEEE Transactions on Intelligent Transportation Systems*, 24(4):3638–3659.

Lai, X., Chen, Y., Lu, F., Liu, J., and Jia, J. (2023). Spherical transformer for lidar-based 3d recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17545–17555.

Lee, J. and Kim, J. (2016). Road geometry recognition for intelligent vehicles: a survey. *International Journal of Automotive Technology*, 17(1):1–10.

Li, Y. and Guo, Q. (2021). Intelligent vehicle collision avoidance technology and its applications. *Journal of Advanced Transportation*, 2021:6623769.

Liu, Y., Li, X., Li, X., Li, Z., Wu, C., and Li, J. (2021). Autonomous vehicles and human factors: A review of the literature. *IEEE Access*, 9:38416–38434.

Ma, X., Wang, Z., Li, H., Zhang, P., Ouyang, W., and Fan, X. (2019). Accurate monocular 3d object detection via color-embedded 3d reconstruction for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Manhardt, F., Kehl, W., and Gaidon, A. (2019). Roi-10d: Monocular lifting of 2d detection to 6d pose and metric shape. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

McDonald, M. and Mazumdar, S. (2020). Drivers' perceived benefits and barriers of advanced driver assistance systems (adas) in the uk. *Transportation Research Part F: Traffic Psychology and Behaviour*, 73:1–16.

Phan-Minh, T. (2021). *Contract-based design: Theories and applications*. PhD thesis, California Institute of Technology.

Phan-Minh, T., Grigore, E. C., Boulton, F. A., Beijbom, O., and Wolff, E. M. (2020). Covernet: Multimodal behavior prediction using trajectory sets. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14074–14083.

Philion, J. and Fidler, S. (2020). Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*. Springer.

Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., and Koltun, V. (2020). Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637.

Salzmann, T., Ivanovic, B., Chakravarty, P., and Pavone, M. (2020). Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 683–700. Springer.

Sengupta, S., Sturgess, P., Torr, P., et al. (2012). Automatic dense visual semantic mapping from street-level imagery. in 2012 ieee. In *RSJ International Conference on Intelligent Robots and Systems*, pages 857–862.

Sharma, S., Sistu, G., Yahiaoui, L., Das, A., Halton, M., and Eising, C. (2023). Navigating uncertainty: The role of short-term trajectory prediction in autonomous vehicle safety. In *Proceedings of the Irish Machine Vision and Image Processing Conference*.

Tan, M. and Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*. PMLR.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Wei, Y., Cheng, S., Wu, Y., and Liu, Y. (2021). Traffic congestion prediction and control using machine learning: A review. *IEEE Transactions on Intelligent Transportation Systems*, 22(7):4176–4195.

Wiest, J., Omari, S., Köhler, J., Lützenberger, M., and Ziegler, J. (2020). Learning to predict the effect of road geometry on vehicle trajectories for autonomous driving. *IEEE Robotics and Automation Letters*, 5(2):2426–2433.

Wu, C., Li, X., Li, X., and Guo, K. (2017). Road geometry modeling and analysis for vehicle dynamics control. *Mechanical Systems and Signal Processing*.

Yang, Y., Chen, Y., and Zhang, J. (2021). A survey on human-autonomous vehicle interaction: Past, present and future. *IEEE Transactions on Intelligent Vehicles*, 6(2):141–154.

Zhang, Y., Liu, H., Shen, S., and Wang, D. (2020). Multimodal trajectory prediction with maneuver-based motion prediction and driver behavior modeling. *IEEE Robotics and Automation Letters*, 5(4):5461–5468.

Zhou, B. and Krähenbühl, P. (2022). Cross-view transformers for real-time map-view semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13760–13769.

Zhou, T., Brown, M., Snavely, N., and Lowe, D. G. (2017). Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Zhu, M., Zhang, S., Zhong, Y., Lu, P., Peng, H., and Lenneman, J. (2021). Monocular 3d vehicle detection using uncalibrated traffic cameras through homography. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE.