TBP-Former: Learning Temporal Bird's-Eye-View Pyramid for Joint Perception and Prediction in Vision-Centric Autonomous Driving

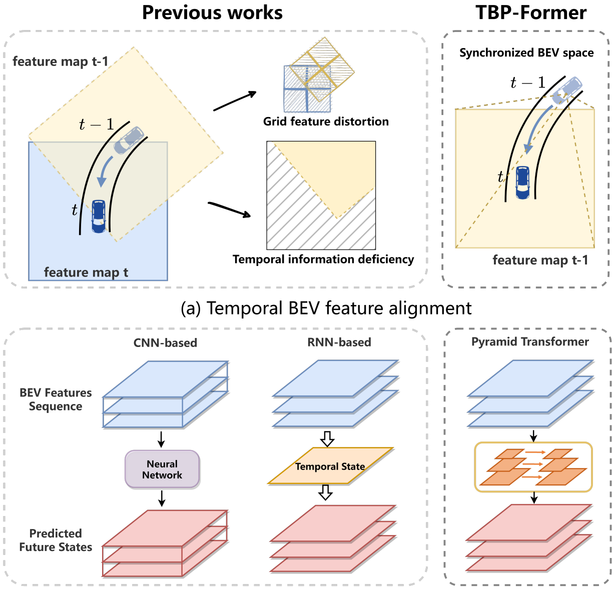
Shaoheng Fang[1]* Zi Wang[1]* Yiqi Zhong[2] Junhao Ge[1] Siheng Chen[1,3]†
[1]Cooperative Medianet Innovation Center, Shanghai Jiao Tong University
[2]Department of Computer Science, University of Southern California [3]Shanghai AI Laboratory
[1]{shfang, w4ngz1, cancaries, sihengc}@sjtu.edu.cn [2]{yiqizhon}@usc.edu

Abstract

Vision-centric joint perception and prediction (PnP) has become an emerging trend in autonomous driving research. It predicts the future states of the traffic participants in the surrounding environment from raw RGB images. However, it is still a critical challenge to synchronize features obtained at multiple camera views and timestamps due to inevitable geometric distortions and further exploit those spatial-temporal features. To address this issue, we propose a temporal bird's-eye-view pyramid transformer (TBP-Former) for vision-centric PnP, which includes two novel designs. First, a pose-synchronized BEV encoder is proposed to map raw image inputs with any camera pose at any time to a shared and synchronized BEV space for better spatial-temporal synchronization. Second, a spatial-temporal pyramid transformer is introduced to comprehensively extract multi-scale BEV features and predict future BEV states with the support of spatial priors. Extensive experiments on nuScenes dataset show that our proposed framework overall outperforms all state-of-the-art vision-based prediction methods. Code is available at: https://github.com/MediaBrain-SJTU/TBP-Former

1. Introduction

As one of the most fascinating engineering projects, autonomous driving has been an aspiration for many researchers and engineers for decades. Although significant progress has been made, it is still an open question in designing a practical solution to achieve the goal of full self-driving. A traditional and common solution consists of a sequential stack of perception, prediction, planning, and control. Despite the idea of divide-and-conquer having achieved tremendous success in developing software systems, a long stack could cause cascading failures in an autonomous system. Recently, there is a trend to combine multiple parts in an autonomous system to be a joint module, cutting down the stack. For example, [25, 46] consider joint perception and prediction and [5, 43] explore joint prediction and planning. This work focuses on joint perception and prediction.



(a) Temporal BEV feature alignment

(b) Spatial-temporal feature modeling for prediction

Figure 1. Two major challenges in vision-based perception and prediction are (a) how to avoid distortion and deficiency when aggregating features across time and camera views; and (b) how to achieve spatial-temporal feature learning for prediction. Our Pose-Synchronized BEV Encoder can precisely map the visual features into synchronized BEV space, and Spatial-Temporal Pyramid Transformer extracts feature at multiple scales.

The task of joint perception and prediction (PnP) aims to predict the current and future states of the surrounding environment with the input of multi-frame raw sensor

*These authors contributed equally to this work.
†Corresponding author.

data. The output current and future states would directly serve as the input for motion planning. Recently, many PnP methods are proposed based on diverse sensor input choices. For example, [4, 25, 34] take multi-frame LiDAR point clouds as input and achieve encouraging 3D detection and trajectory prediction performances simultaneously. Recently, the rapid development of vision-centric methods offers a new possibility to provide a cheaper and easy-to-deploy solution for PnP. For instance, [1, 16, 17] only uses RGB images collected by multiple cameras to build PnP systems. Meanwhile, without precise 3D measurements, vision-centric PnP is more technically challenging. Therefore, this work aims to advance this direction.

The core of vision-centric PnP is to learn appropriate spatial-temporal feature representations from temporal image sequences. It is a crux and difficult from three aspects. First, since the input and the output of vision-centric PnP are supported in camera front-view (FV) and bird's-eye-view (BEV) respectively, one has to deal with distortion issues during geometric transformation between two views. Second, when the vehicle is moving, the view of the image input is time-varying and it is thus nontrivial to precisely map visual features across time into a shared and synchronized space. Third, since information in temporal image sequences is sufficiently rich for humans to accurately perceive the environment, we need a powerful learning model to comprehensively exploit spatial-temporal features.

To tackle these issues, previous works on vision-centric PnP consider diverse strategies. For example, [16, 56] follows the method in [38] to map FV features to BEV features, then synchronizes BEV features across time via rigid transformation, and finally uses a recurrent network to exploit spatial-temporal features. However, due to the image discretization nature and depth estimation uncertainty, simply relying on rigid geometric transformations would cause inevitable distortion; see Fig. 1. Some other work [49] transforms the pseudo feature point cloud to current ego coordinates and then pools the pseudo-lidar to BEV features; however, this approach encounters deficiency due to the limited sensing range in perception. Meanwhile, many works [16, 17, 56] simply employ recurrent neural networks to learn the temporal features from multiple BEV representations, which is hard to comprehensively extract spatial-temporal features.

To promote more reliable and comprehensive feature learning across views and time, we propose the temporal bird's-eye-view pyramid transformer (TBP-Former) for vision-centric PnP. The proposed TBP-Former includes two key innovations: i) pose-synchronized BEV encoder, which leverages a pose-aware cross-attention mechanism to directly map a raw image input with any camera pose at any time to the corresponding feature map in a shared and synchronized BEV space; and ii) spatial-temporal pyramid transformer, which leverages a pyramid architecture with Swin-transformer [28] blocks to learn comprehensive spatial-temporal features from sequential BEV maps at multiple scales and predict future BEV states with a set of future queries equipped with spatial priors.

Compared to previous works, the proposed TBP-Former brings benefits from two aspects. First, previous works [16, 17, 24, 56] consider FV-to-BEV transformation and temporal synchronization as two separate steps, each of which could bring distortion due to discrete depth estimation and rigid transformation; while we merge them into one step and leverage both geometric transformation and attention-based learning ability to achieve spatial-temporal synchronization. Second, previous works [16, 53] use RNNs or 3D convolutions to learn spatial-temporal features; while we leverage a powerful pyramid transformer architecture to comprehensively capture spatial-temporal features, which makes prediction more effective.

To summarize, the main contributions of our work are:

- To tackle the distortion issues in mapping temporal image sequences to a synchronized BEV space, we propose a pose-synchronized BEV encoder (PoseSync BEV Encoder) based on cross-view attention mechanism to extract quality temporal BEV features.

- We propose a novel Spatial-Temporal Pyramid Transformer (STPT) to extract multi-scale spatial-temporal features from sequential BEV maps and predict future BEV states according to well-elaborated future queries integrated with spatial priors.

- Overall, we propose TBP-Former, a vision-based joint perception and prediction framework for autonomous driving. TBP-Former achieves state-of-the-art performance on nuScenes [2] dataset for the vision-based prediction task. Extensive experiments show that both PoseSync BEV Encoder and STPT contribute greatly to the performance. Due to the decoupling property of the framework, both proposed modules can be easily utilized as alternative modules in any vision-based BEV prediction framework.

2. Related Work

2.1. Joint Perception and Prediction

As the two core system modules of autonomous driving, how to conduct perception and prediction tasks jointly has received a lot of attention. Traditional approaches [3, 4, 25, 34, 39] formulate this joint task as a trajectory prediction problem that relies on the perception outputs of 3D object detection and tracking. The dependency on intermediate results tends to accumulate errors and lacks the capacity to perceive unknown objects [52, 53]. Subsequently,

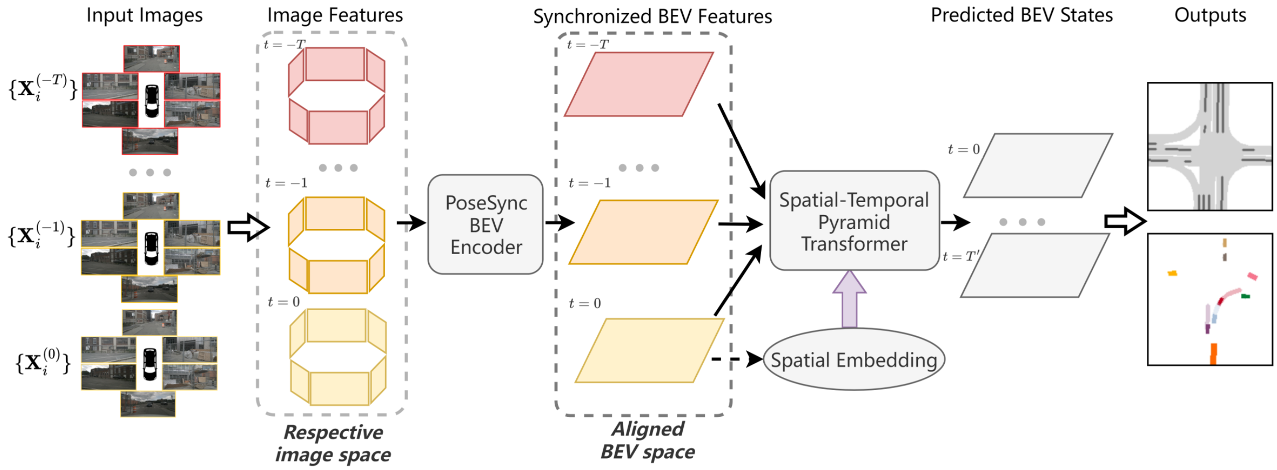


Figure 2. An overview of TBP-Former architecture. Taking consecutive surrounding camera images as inputs, TBP-Former first generates image-space features and uses the PoseSync BEV Encoder to map front-view features to BEV features in a shared and synchronized BEV space. Then the BEV features from multiple frames are processed by the Spatial-Temporal Pyramid Transformer to extract BEV spatial-temporal features and predict future BEV states in order. In this process, high-level scene representations are generated from the last frame BEV feature as spatial priors to guide the prediction. Finally, the well-predicted future states are sent to decoder heads for joint perception and prediction tasks.

instance-free methods [9,21,33,43,45,53] that predict dense future semantic occupancy and flow have become a growing trend to simplify the understanding of dynamic scenes. Also, several recent works [16, 17, 56] explore joint perception and prediction in the form of dense occupancy and flow using only surrounding camera input.

In many previous works [4, 25, 39, 43], raster HD (high-definition) maps play an important role as input of the frameworks. HD maps can provide strong priors to guide the predicted results to follow the traffic lanes. However, in practice, HD maps are laborious and costly to produce and require frequent maintenance. Instead of using off-the-peg HD maps, we follow the philosophy of [5, 7, 23] in predicting online HD maps but propose to learn high-level scene geometry representations from real-time sensor inputs and take these representations as priors for the prediction task.

2.2. BEV Representations

BEV representations provide a unified and physical-interpretable way to represent the rich information of road, moving objects and occlusion in a traffic scene, which can be easily utilized for downstream tasks such as motion prediction, planning and control, etc. For camera-based methods, how to solve the problem of projecting features from perspective view to BEV is a major challenge. Some learnable methods use MLP [23, 36, 42] or transformer network [37, 57] to implicitly reason the relationship between two different views. LSS [38] proposes the approach of predicting depth distribution per pixel on 2D features, then 'lifting' the 2D features according to the corresponding depth distribution to BEV space. Numerous works, aiming at tasks of BEV perception [19, 20, 54], motion prediction [1, 16, 17, 56], lidar-camera fusion [26], etc., follow this form to generate BEV representations. Also, some methods [6, 13, 24] explicitly establish the correspondence from BEV location to image-view pixel using homography between image and BEV plane and achieve attractive performance in diverse tasks.

However, when dealing with temporal information, most methods [1, 16, 17, 24, 56] warp history BEV representations according to the variation of ego poses. Due to the pre-defined fixed range and size of the BEV grid, rotation and translation operations may cause distortion and out-of-range problems when aligning history BEV maps to current ego coordinates. Though [40] introduces a similar operation to us to integrate historical information into the current frame, the design of their model is unable to predict future states. To alleviate these issues, we propose a PoseSync BEV Encoder module based on deformable attention to generate pose-synchronized BEV representations from temporally consecutive image-view input.

2.3. Spatial-Temporal Modeling

In the BEV prediction field, how to design a temporal model to aggregate spatial-temporal information is a critical problem. Existing modeling methods can be classified into three categories: RNN-based, CNN-based and transformer-based. RNN-based methods [1, 16–18, 43, 56] utilize recurrent models such as LSTM [15], GRU [8] to predict the future latent states. Though the recurrent model is powerful to model temporal relationships, it is time-consuming for constraints in the parallelization of computation. Besides,

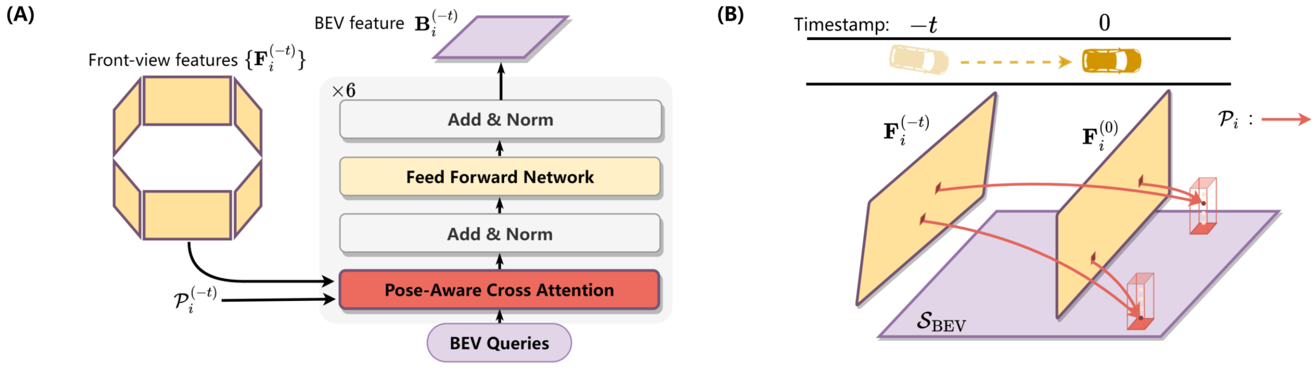


Figure 3. The PoseSync BEV Encoder (A) takes front-view features and camera poses as input and then maps to BEV space. The core to generate BEV features in a synchronized way is the Pose-Aware Cross Attention. Its cross-view attention mechanism is depicted in (B), where front-view features from different frames of a dynamic vehicle are projected into a uniform BEV space.

some CNN-based methods [5, 34, 50, 53] concatenate BEV features in the time dimension and take advantage of 3D convolution to extract spatial-temporal features.

Due to the great power of transformer [48] in sequence modeling, it has shown promise in many temporal modeling tasks such as trajectory prediction [11, 35, 55], object tracking [22], video prediction [12, 41, 51], video interpolation [10, 32, 47], etc. For BEV perception, [24, 27] utilize self-attention to model temporal information from multiple frames to boost perception task. [22] leverage self-attention to aggregate spatial information and cross-attention to exploit affinities among sequence frames. To explore the capacity of transformer model in BEV spatial-temporal modeling, we propose a novel Spatial-Temporal Pyramid Transformer (STPT) architecture with future queries for BEV spatial-temporal features extraction and BEV future states prediction.

3. Methodology

3.1. Overview Architecture

The overall architecture of the proposed TBP-Former is illustrated in Fig. 2. It takes the input of multi-view images with the corresponding camera poses at consecutive T timestamps. The final output includes BEV map segmentation for current scene understanding and occupancy flow for motion prediction. The whole TBP-Former can be decoupled into three parts: (i) pose-synchronized BEV encoder, which maps raw image sequences into feature maps in a spatial-temporal-synchronized BEV space; (ii) spatial-temporal pyramid transformer, which achieves comprehensive feature learning at multiple spatial and temporal scales; and (iii) a multi-head decoder, which takes the spatial and temporal features to achieve scene understanding and motion prediction. We will elaborate on each part in the following subsections.

3.2. Pose-Synchronized BEV Encoder

Given images collected at multiple time stamps and from various camera poses, we aim to generate the corresponding feature maps in a shared and synchronized BEV space. Different from many previous works that synchronize spatial and temporal information in two separate steps, the proposed pose-synchronized BEV encoder leverage both geometric prior and learning ability to achieve one-step synchronization, alleviating distortion effects. Following the previous transformer-based method [24], this encoder adopts a transformer architecture whose core is a novel cross-view attention operation.

Front-view feature map. Let $\mathcal{X} = \{\mathbf{X}_i^{(-t)}\}_{i=1,t=0}^{N,T}$ be the input multi-frame multi-view images, where N is the number of cameras, T is the number of historical timestamps and $\mathbf{X}_i^{(-t)} \in \mathbb{R}^{H\times W\times 3}$ is the RBG image captured by the ith camera at historical time stamp t. Note that each front-view image is associated with a different camera pose. Let $\mathcal{S}_i^{(-t)} = \{(u_i^{(-t)}, v_i^{(-t)})\}_{1,1}^{H,W}$ be the pixel indices of the ith camera's front-view space, whose image size is $H \times W$. We feed each RBG image $\mathbf{X}_i^{(-t)}$ into a shared backbone network (our implementation uses ResNet-101 [14]) and obtain the corresponding front-view feature map $\mathbf{F}_i^{(-t)} \in \mathbb{R}^{H'\times W'\times C}$ with C the channel number, which is also supported on the front-view space $\mathcal{S}_i^{(-t)}$.

BEV queries. Let $\mathcal{S}_{\mathrm{BEV}} = \{(x,y)\}_{x=1,y=1}^{X,Y}$ be the BEV grid indices, reflecting the $X \times Y$ BEV grid space based on the vehicle-ego pose at the current timestamp. Note that $\mathcal{S}_{\mathrm{BEV}}$ is the only BEV space we work with in this paper. Let $\mathbf{Q} \in \mathbb{R}^{X\times Y\times C}$ be the trainable BEV queries whose element $\mathbf{Q}_{x,y} \in \mathbb{R}^C$ is a C-dimensional query feature at the (x,y)th geo-location in the BEV space $\mathcal{S}_{\mathrm{BEV}}$. We use $\mathbf{Q}$ as the input to query from the front-view feature map $\mathbf{F}_i^{(-t)}$ to produce the corresponding BEV feature map.

Cross-view attention. As the key operation in the pose-synchronized BEV encoder, the proposed cross-view attention constructs a feature map in the BEV space $\mathcal{S}_{\text{BEV}}$ by absorbing information from the corresponding pixels in the front-view feature map.

Let $\mathcal{P}_i^{(-t)} : \mathcal{S}_{\text{BEV}} \times \mathcal{Z} \to \mathcal{S}_i^{(-t)}$ be a project operation that maps a BEV index with a specific height index to a pixel index in the ith camera's front view at historical timestamp t; that is,

$$(u_i^{(-t)}, v_i^{(-t)}) = \mathcal{P}_i^{(-t)}\Big((x, y, z)\Big),$$

where $z \in \mathcal{Z} = \{1, \cdots, Z\}$. The project operation $\mathcal{P}_i^{(-t)}$ builds the geometric relationship between the BEV and a front view. The implementation of $\mathcal{P}_i^{(-t)}$ works as

$$z_i^{(-t)} \cdot \begin{bmatrix} u_i^{(-t)} \\ v_i^{(-t)} \\ 1 \end{bmatrix} = \mathcal{T}_{\mathcal{S}_{\text{BEV}} \to \mathcal{S}_i^{(-t)}} \cdot \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix},$$

where $\mathcal{T}_{\mathcal{S}_{\text{BEV}} \to \mathcal{S}_i^{(-t)}} \in \mathbb{R}^{3\times 4}$ is a transformation matrix that can be calculated by camera's intrinsic/ extrinsic parameters and vehicle-ego pose.

Based on the project operation, the cross-view attention can trace visual features in the front view through a BEV index. Let $\mathbf{B}_i^{(-t)} \in \mathbb{R}^{X\times Y\times C}$ be the BEV feature map associated with the RBG image $\mathbf{X}_i^{(-t)}$. The (x, y)th element of the BEV feature map is obtained as

$$(\mathbf{B}_i^{(-t)})_{x,y} = \sum_z f_{\text{DA}}\left(\mathbf{Q}_{x,y}, \mathcal{P}_i^{(-t)}(x,y,z), \mathbf{F}_i^{(-t)}\right), \tag{1}$$

where, $f_{\text{DA}}(\cdot)$ represents the deformable attention operation [58]. It allows BEV query $Q_{x,y}$ only to interact with the front-view feature $\mathbf{F}_i^{(-t)}$ within its regions of interest, which is sampled around the reference point calculated by $\mathcal{P}_i^{(-t)}$. Since one BEV index might lead to multiple pixel indices in the front-view image because of various height possibilities in the 3D space. We thus sum over all possible heights in (1). To further aggregate BEV feature maps across all the N camera views, we simply take the average; that is, the BEV feature map at historical timestamp t is $\mathbf{B}^{(-t)} = \frac{1}{N}\sum_i \mathbf{B}_i^{(-t)}$. Note that all the front-view features across time and from multiple cameras are synchronized into the same BEV space in one step (1), leading to less information distortion or deficiency issues.

We can successively apply the cross-view attention followed by feed forward networks and normalization layers for multiple times. Finally, we order BEV feature maps at multiple timestamps and obtain a temporal BEV feature map $\mathcal{B} = [\mathbf{B}^{(0)}, \mathbf{B}^{(-1)}, ..., \mathbf{B}^{(-T)}] \in \mathbb{R}^{(T+1)\times X\times Y\times C}$.

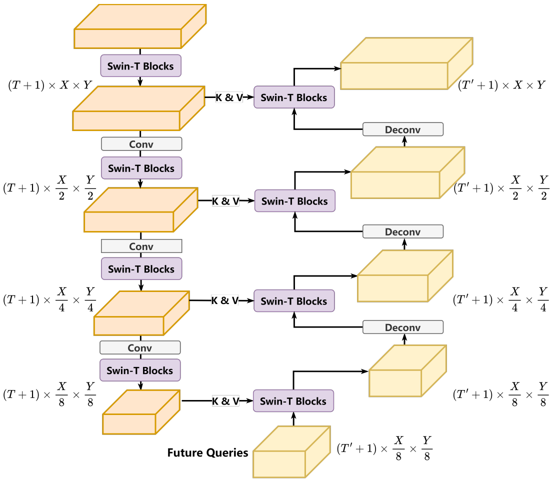

Swin-T Blocks
$(T+1)\times X\times Y$
K & V
Swin-T Blocks
$(T'+1)\times X\times Y$
Deconv
Conv
Swin-T Blocks
$(T+1)\times \frac{X}{2}\times \frac{Y}{2}$
K & V
Swin-T Blocks
$(T'+1)\times \frac{X}{2}\times \frac{Y}{2}$
Deconv
Conv
Swin-T Blocks
$(T+1)\times \frac{X}{4}\times \frac{Y}{4}$
K & V
Swin-T Blocks
$(T'+1)\times \frac{X}{4}\times \frac{Y}{4}$
Deconv
Conv
Swin-T Blocks
$(T+1)\times \frac{X}{8}\times \frac{Y}{8}$
K & V
Swin-T Blocks
$(T'+1)\times \frac{X}{8}\times \frac{Y}{8}$
Future Queries
$(T'+1)\times \frac{X}{8}\times \frac{Y}{8}$

Figure 4. The network architecture of Spatial-Temporal Pyramid Transformer (STPT). Each encoder layer consists of an optional convolutional block for downsampling and Swin Transformer Blocks, while each decoder layer contains Swin Transformer Blocks and a deconvolutional block for upsampling. In the decoding process, we pre-define a set of future queries to represent future BEV states and query the features from encoders.

3.3. Spatial-Temporal Pyramid Transformer

We further propose a novel spatial-temporal pyramid transformer (STPT) to learn spatial-temporal features more comprehensively and produce the future BEV states. The detailed structure of STPT is depicted in Fig. 4.

Temporal BEV pyramid feature learning. We encode the input temporal BEV feature map $\mathcal{B}$ with four hierarchical layers. Each encoder layer is composed of an optional convolution layer with stride 2 to downsample the features and a swin transformer encoder, which is a stack of Swin Transformer blocks [28]. In our implementation, the window size of Swin-T blocks is set as $(4, 4)$. We can then obtain multi-scale spatial-temporal features $\mathcal{B}_s \in \mathbb{R}^{(T+1)\times\frac{X}{2^s}\times\frac{Y}{2^s}\times C}$, $s = 0, 1, 2, 3$.

Future BEV queries. Future BEV queries are defined to represent future BEV states and query the generated multi-scale spatial-temporal features. There is a set of learnable future queries $\{\mathbf{Q}^{(t)}\}$, $t = 0, \ldots, T'$, where $\mathbf{Q}^{(t)}$ has the same spatial dimension $\frac{X}{8} \times \frac{Y}{8}$ as $\mathcal{B}_3$. Separate learning embeddings are employed for future queries to differentiate the predicted BEV states over time. Additionally, a map feature generator is applied to generate high-dimensional features from $\mathbf{B}^{(T)}$ in order to extract information about the scene's geometry. To be specific, the same structure and parameters of the hdmap decoder head (excluding the last linear layer) are reused. The resulting map feature is added to all future queries to provide spatial information priors.

Future BEV state prediction. The decoding process contains corresponding four hierarchical layers as the en-

Method	RGB	Future semantic seg.		Future instance seg.		FPS
	Resolution	IoU (Short)	IoU (Long)	VPQ (Short)	VPQ (Long)	
FIERY [16]	224×480	59.4	36.7	50.2	29.9	1.56
StretchBEV [1]	224×480	55.5	37.1	46.0	29.0	1.56
ST-P3 [17]	224×480	-	38.9	-	32.1	1.43
BEVerse [56]	256×704	60.3	38.7	52.2	33.3	1.96
TBP-Former	224×480	**64.7**	**41.9**	**56.7**	**36.9**	**2.44**

Table 1. **Prediction results on nuScenes [2] validation set.** Intersection-over-Union (IoU) is used for future semantic segmentation and Video Panoptic Quality (VPQ) for future instance segmentation. Results are reported under two settings: short ($30m \times 30m$) range and long ($100m \times 100m$) range. Frame Per Second (FPS) means the inverse of inference time. All methods are tested under the same settings on a single NVIDIA A100. Our TBP-Former achieves SOTA performance and is still more computationally efficient than other methods.

Method	Temp.	Veh. IoU	Ped. IoU
VED [31]		23.3	11.9
VPN [36]		28.2	10.3
PON [42]		27.9	13.9
LSS [38]		34.6	15.0
CVT [57]		36.0	-
Image2Map [44]		40.2	-
BEVFormer [24]		44.4	-
IVMP [49]	✓	36.8	17.4
FIERY [16]	✓	38.2	17.2
ST-P3 [17]	✓	40.1	14.5
TBP-Former *static*		44.8	17.2
TBP-Former	✓	**46.2**	**18.6**

Table 2. **Perception results on nuScenes [2] validation set.** Results of vehicles and pedestrians are compared by segmentation IoU. Temp. indicates whether temporal information is involved.

coding process. Unlike the encoding process, $\{\mathbf{Q}^{(t)}\}$ is used as the query input of Swin-T block and performs cross attention with the encoded features $\mathbf{E}_3$. After the first decoding layer, the output of each layer is used as the query input of the next layer. Similar to encoding layers, the deconvolution layer is optionally applied to upsample the decoded features. The simplified process can be written as

$$\mathcal{D}_s = \begin{cases} \mathrm{SwinT}(\mathcal{B}_3, \{\mathbf{Q}^{(t)}\}), & s = 3 \\ \mathrm{SwinT}(\mathcal{B}_s, \mathrm{DeConv}(\mathbf{D}_{s+1})), & s = 0, 1, 2 \end{cases}$$

where $\mathcal{D}_s \in \mathbb{R}^{(T'+1) \times \frac{X}{2^s} \times \frac{Y}{2^s} \times C}$, $s = 0, 1, 2, 3$ are decoded features. The future temporal BEV feature map at the 0th scale is a temporal sequence of final predicted BEV states; that is, $\mathcal{D}_0 = [\mathbf{B}_*^{(0)}, \mathbf{B}_*^{(1)}, ..., \mathbf{B}_*^{(T')}]$, where $\mathbf{B}_*^{(t)} \in \mathbb{R}^{X \times Y \times C}$ is the BEV state at future time stamp t.

3.4. Multi-head decoder

The future temporal BEV feature map is fed into the multi-task decoder heads to generate various outputs for dynamic scene understanding, see Fig. 5. We follow the output setting in [16] that predicts BEV semantic segmentation, instance center, instance offset, and future flow for joint perception and prediction. Meanwhile, we set up an additional HD map decoder head to predict basic traffic scene elements including drivable areas and lanes. The map decoder head can not only provide scene information for subsequent planning and control modules but also give guidance to the prediction process, see sec. 3.3.

4. Experiments

4.1. Dataset and settings

We use nuScenes [2] datasets to evaluate our approach. NuScenes contains 1000 scenes, each of which has 20 seconds annotated at 2Hz. In nuScenes, the images are captured by 6 cameras with a small overlap in the field of view, which guarantees the cameras cover the full 360° field of view. For model input, raw camera images with the size of 900 × 1600 are resized and cropped to a resolution of 224 × 480. We follow the training and evaluating settings used in previous methods [1,16,17,56] for fair comparisons, which use 1.0 second past states and current state to predict 2.0 seconds of the future states. It corresponds to predicting 4 future frames based on 3 observed frames. The size of the generated BEV grid map is 200 × 200. Each grid has a range of $0.5m \times 0.5m$, which means the perception and prediction range is $100m \times 100m$.

For training, we use AdamW [30] with a weight decay 0.01 to optimize the models. The learning rate is initialized as 10^{-4} and decays with a cosine annealing scheduler [29]. All models are trained on 4 NVIDIA A100 GPUs for 10 epochs.

4.2. Metrics

Following previous works [1, 16, 17, 56], we mainly use two metrics for evaluation. The first is Intersection over Union (IoU), which measures the quality of segmentation at each frame. The second is Video Panoptic Quality (VPQ), which is used to measure the consistency of the detected instances over time and the accuracy of the segmentation.

Exp.	Warp.	Sync.	SLQ	SPE	Future semantic seg. Short (IoU)	 Long (IoU)	Future instance seg. Short (VPQ)	 Long (VPQ)
1	✓				58.7	38.4	50.6	31.8
2	✓		✓		60.8	38.6	52.4	33.4
3	✓		✓	✓	62.0	40.7	53.2	34.3
4		✓			63.0	40.8	54.1	34.3
5		✓	✓		63.8	41.1	55.8	35.8
6		✓	✓	✓	**64.7**	**41.9**	**56.7**	**36.9**

Table 3. **Ablation of our proposed architecture.** Ablation results for our PoseSync BEV Encoder (Sync.), the learnable future queries, and the spatial embedding are presented. Exp. 1-3 use the traditional warping methods to align temporal BEV features. Separate learnable queries (SLQ) represent using separate learnable future queries instead of utilizing the same query with temporal positional encoding. Spatial positional embedding (SPE) represents using spatial scene representations in future prediction queries.

Temporal model	IoU	VPQ	VRQ	VSQ
MotionNet† [53]	35.4	30.6	43.1	71.1
FIERY† [16]	38.3	32.1	45.4	70.7
BEVerse† [56]	40.2	34.0	48.0	70.9
TBP-Former	**41.9**	**36.9**	**51.5**	**72.6**

Table 4. **Ablation for the prediction model.** †: We use MotionNet, FIERY and BEVerse to replace our prediction model for comparison, and the BEV encoder and task heads are the same. Besides IoU and VPQ, we also use Video Recognition Quality (VRQ) and Video Segmentation quality (VSQ) for evaluation.

Augmentation Cam	 BEV	Perception Veh.	 Ped.	Prediction IoU	 VPQ
		45.0	17.7	40.5	34.4
✓		44.8	18.5	40.9	35.3
	✓	45.3	18.6	41.4	35.6
✓	✓	**46.2**	**18.6**	**41.9**	**36.9**

Table 5. **Ablation for data augmentation strategies.** Perception of Vehicles and Pedestrians with different data augmentation strategies are evaluated on segmentation IoU. Prediction results are evaluated on segmentation IoU and Video Panoptic Quality.

The formula is shown below:

$$\mathrm{VPQ} = \sum_{t=0}^{H} \frac{\sum_{(p_t,q_t)\in TP_t} \mathrm{IoU}(p_t, q_t)}{|TP_t| + \frac{1}{2}|FP_t| + \frac{1}{2}|FN_t|}$$

where H is the sequence length, TP_t represents the set of true positives, FP_t represents the set of false positives and FN_t represents the set of false negtives at timestamp t.

4.3. PnP results

Perception and Prediction. Table 1 compares TBP-Former with other methods of perception and prediction task based on multi-view cameras. We see that i) we achieve state-of-the-art performance and exceed previous methods by a large margin. ii) Even though BEVerse has larger RGB resolutions, TBP-Former still surpasses their performance on IoU by **7.3%/8.3%** for short/long settings, respectively. TBP-Former also improves the VPQ by **12.1%/10.8%**. iii) Apart from the performance improvement, TBF-Former also has a larger FPS compared to other methods. Its inference speed is 25% faster than BEVerse's.

Fig. 5 shows the visualization results of our proposed method. We see that i) almost all the objects are detected correctly except for those occluded ones. ii) TBP-Former is capable of capturing the motion information in past frames and precisely predicting the vehicles' trajectories by occupancy and flow. Compared with FIERY [16], TBP-Former is closer to the ground truth. iii) TBP-Former does a better job than FIERY when predicting vehicles' turning.

Perception Only. Table 2 compares the results of plenty of state-of-the-art methods on perception (segmentation) task. We see that our static model, which does not contain temporal information, can achieve 44.8 and 17.2 IoU of vehicles and pedestrians. With the input of temporal sequences, the performance improves further since auxiliary information is provided for better perception. The state-of-the-art results prove the effectiveness of the novel design of our Pose-synchronized BEV encoder.

4.4. Ablation

Effectiveness of PoseSync View Projection. The Exp. 1&4, 2&5, 3&6 in Table 3 compare the proposed PoseSync View Projection and the existing feature warping methods. We see that the proposed method always achieves better performance when other settings remain the same. The reasons are that: i) PoseSync View Projection based on Deformable Attention can guarantee the precise correspondence between BEV grids and image features. ii) Our projection method can alleviate distortion and our-of-range issues when synchronizing sequential BEV features.

Effectiveness of the designed future queries. In Exp. 1&3 in Table 3, we utilize the identical query with temporal positional encoding for future queries. Exp. 2&4 in Table 3 demonstrate that using separate learnable embedding for future queries can achieve better performance. Exp. 3&6 in Table 3 validate the efficacy of the proposed spatial priors for future queries. The generated high-

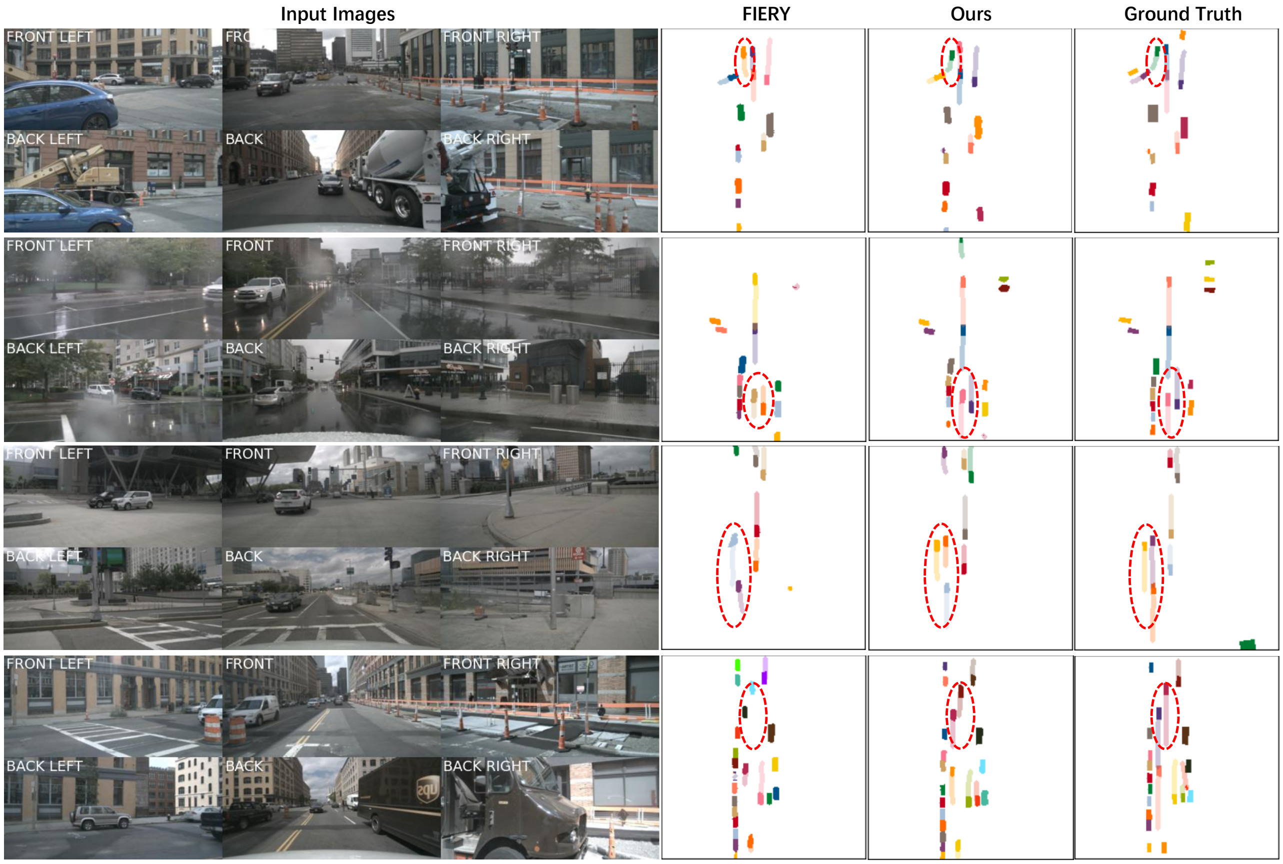


Figure 5. Demonstration of our results compared with FIERY and Ground Truth. Different vehicles are assigned with different colors in order to make a distinction. The darker parts represent the perception of the current frame, and the lighter parts represent the prediction of the vehicles in future frames. The visualization is based on the predicted occupancy and flow.

dimensional map features provide the prediction model with useful geographic information. The additional spatial information can aid the prediction and lead to better scene forecasting.

Effectiveness of STPT. Table 4 compares STPT with popular CNN-based [53] and RNN-based [16, 56] methods. We keep all the settings the same except for temporal modeling. To be specific, the size of input images (224 × 480), image backbones and BEV feature extractor are the same. And then we plug their temporal models into our architecture. We see that i) STPT model performs better in all four metrics, including semantic segmentation IoU and three instance segmentation metrics from the video prediction area. ii) Our reproduced temporal models achieve higher performance than the original implements. This further validates the effectiveness and power of our BEV feature extractor.

Data Augmentation. We perform both image-view and BEV augmentations. The image-view augmentations include random scaling, rotation and flip of the input images. The BEV augmentations include similar operations on both BEV representations and corresponding ground truth labels. Table 5 compares the results of different data augmentation strategies. We see that i) both augmentation methods improve the performance when used separately. ii) The combination of two methods works better than any single approach. Introducing data augmentation strategies is beneficial to the model's robustness and generalization ability.

5. Conclusion

This paper proposes a novel TBP-Former for vision-centric joint perception and prediction. We design a pose-synchronized BEV encoder module using a cross-view attention mechanism to solve the distortion issues in previous works. Furthermore, we propose a powerful spatial-temporal pyramid transformer for BEV feature extraction and BEV state prediction. Experiments show that i) TBP-Former improves the prediction performance over state-of-the-art methods significantly; and ii) both PoseSync BEV Encoder and STPT contribute to better performances.

Acknowledgement This research is partially supported by National Natural Science Foundation of China under Grant 62171276 and the Science and Technology Commission of Shanghai Municipal under Grant 21511100900 and 22DZ2229005.

References

[1] Adil Kaan Akan and Fatma Güney. Stretchbev: Stretching future instance prediction spatially and temporally. *arXiv preprint arXiv:2203.13641*, 2022. 2, 3, 6

[2] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 2, 6

[3] Sergio Casas, Cole Gulino, Renjie Liao, and Raquel Urtasun. Spagnn: Spatially-aware graph neural networks for relational behavior forecasting from sensor data. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9491–9497. IEEE, 2020. 2

[4] Sergio Casas, Wenjie Luo, and Raquel Urtasun. Intentnet: Learning to predict intention from raw sensor data. In *Conference on Robot Learning*, pages 947–956. PMLR, 2018. 2, 3

[5] Sergio Casas, Abbas Sadat, and Raquel Urtasun. Mp3: A unified model to map, perceive, predict and plan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14403–14412, 2021. 1, 3, 4

[6] Li Chen, Chonghao Sima, Yang Li, Zehan Zheng, Jiajie Xu, Xiangwei Geng, Hongyang Li, Conghui He, Jianping Shi, Yu Qiao, et al. Persformer: 3d lane detection via perspective transformer and the openlane benchmark. *arXiv preprint arXiv:2203.11089*, 2022. 3

[7] Siheng Chen, Baoan Liu, Chen Feng, Carlos Vallespi-Gonzalez, and Carl Wellington. 3d point cloud processing and learning for autonomous driving: Impacting map creation, localization, and perception. *IEEE Signal Processing Magazine*, 38(1):68–86, 2020. 3

[8] Rahul Dey and Fathi M Salem. Gate-variants of gated recurrent unit (gru) neural networks. In *2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS)*, pages 1597–1600. IEEE, 2017. 3

[9] Artem Filatov, Andrey Rykov, and Viacheslav Murashkin. Any motion detector: Learning class-agnostic scene dynamics from a sequence of lidar point clouds. In *2020 IEEE international conference on robotics and automation (ICRA)*, pages 9498–9504. IEEE, 2020. 3

[10] Zhicheng Geng, Luming Liang, Tianyu Ding, and Ilya Zharkov. Rstt: Real-time spatial temporal transformer for space-time video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17441–17451, 2022. 4

[11] Roger Girgis, Florian Golemo, Felipe Codevilla, Martin Weiss, Jim Aldon D'Souza, Samira Ebrahimi Kahou, Felix Heide, and Christopher Pal. Latent variable sequential set transformers for joint multi-agent motion prediction. In *International Conference on Learning Representations*, 2021. 4

[12] Agrim Gupta, Stephen Tian, Yunzhi Zhang, Jiajun Wu, Roberto Martín-Martín, and Li Fei-Fei. Maskvit: Masked visual pre-training for video prediction. *arXiv preprint arXiv:2206.11894*, 2022. 4

[13] Adam W Harley, Zhaoyuan Fang, Jie Li, Rares Ambrus, and Katerina Fragkiadaki. A simple baseline for bev perception without lidar. *arXiv e-prints*, pages arXiv–2206, 2022. 3

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[15] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 3

[16] Anthony Hu, Zak Murez, Nikhil Mohan, Sofía Dudas, Jeffrey Hawke, Vijay Badrinarayanan, Roberto Cipolla, and Alex 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*, pages 15273–15282, 2021. 2, 3, 6, 7, 8

[17] Shengchao Hu, Li Chen, Penghao Wu, Hongyang Li, Junchi Yan, and Dacheng Tao. St-p3: End-to-end vision-based autonomous driving via spatial-temporal feature learning. In *European Conference on Computer Vision*, pages 533–549. Springer, 2022. 2, 3, 6

[18] Yue Hu, Siheng Chen, Ya Zhang, and Xiao Gu. Collaborative motion predication via neural motion message passing. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[19] Yue Hu, Shaoheng Fang, Zixing Lei, Yiqi Zhong, and Siheng Chen. Where2comm: Communication-efficient collaborative perception via spatial confidence maps. In *Thirty-sixth Conference on Neural Information Processing Systems (Neurips)*, November 2022. 3

[20] Junjie Huang, Guan Huang, Zheng Zhu, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021. 3

[21] Kuan-Hui Lee, Matthew Kliemann, Adrien Gaidon, Jie Li, Chao Fang, Sudeep Pillai, and Wolfram Burgard. Pillarflow: End-to-end birds-eye-view flow estimation for autonomous driving. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2007–2013. IEEE, 2020. 3

[22] Peixuan Li and Jieyu Jin. Time3d: End-to-end joint monocular 3d object detection and tracking for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3885–3894, 2022. 4

[23] Qi Li, Yue Wang, Yilun Wang, and Hang Zhao. Hdmapnet: An online hd map construction and evaluation framework, 2021. 3

[24] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. *arXiv preprint arXiv:2203.17270*, 2022. 2, 3, 4, 6

[25] Ming Liang, Bin Yang, Wenyuan Zeng, Yun Chen, Rui Hu, Sergio Casas, and Raquel Urtasun. Pnpnet: End-to-end perception and prediction with tracking in the loop. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11553–11562, 2020. 1, 2, 3

[26] Tingting Liang, Hongwei Xie, Kaicheng Yu, Zhongyu Xia, Zhiwei Lin, Yongtao Wang, Tao Tang, Bing Wang, and Zhi

Tang. Bevfusion: A simple and robust lidar-camera fusion framework. *arXiv preprint arXiv:2205.13790*, 2022. 3

[27] Yingfei Liu, Junjie Yan, Fan Jia, Shuailin Li, Qi Gao, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petrv2: A unified framework for 3d perception from multi-camera images. *arXiv preprint arXiv:2206.01256*, 2022. 4

[28] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 2, 5

[29] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 6

[30] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[31] Chenyang Lu, Marinus Jacobus Gerardus van de Molengraft, and Gijs Dubbelman. Monocular semantic occupancy grid mapping with convolutional variational encoder–decoder networks. *IEEE Robotics and Automation Letters*, 4(2):445–452, 2019. 6

[32] Liying Lu, Ruizheng Wu, Huaijia Lin, Jiangbo Lu, and Jiaya Jia. Video frame interpolation with transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3532–3542, 2022. 4

[33] Chenxu Luo, Xiaodong Yang, and Alan Yuille. Self-supervised pillar motion learning for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3183–3192, 2021. 3

[34] Wenjie Luo, Bin Yang, and Raquel Urtasun. Fast and furious: Real time end-to-end 3d detection, tracking and motion forecasting with a single convolutional net. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 3569–3577, 2018. 2, 4

[35] Jiquan Ngiam, Benjamin Caine, Vijay Vasudevan, Zhengdong Zhang, Hao-Tien Lewis Chiang, Jeffrey Ling, Rebecca Roelofs, Alex Bewley, Chenxi Liu, Ashish Venugopal, et al. Scene transformer: A unified multi-task model for behavior prediction and planning. *arXiv e-prints*, pages arXiv–2106, 2021. 4

[36] Bowen Pan, Jiankai Sun, Ho Yin Tiga Leung, Alex Andonian, and Bolei Zhou. Cross-view semantic segmentation for sensing surroundings. *IEEE Robotics and Automation Letters*, 5(3):4867–4873, 2020. 3, 6

[37] Lang Peng, Zhirong Chen, Zhangjie Fu, Pengpeng Liang, and Erkang Cheng. Bevsegformer: Bird's eye view semantic segmentation from arbitrary camera rigs. *arXiv preprint arXiv:2203.04050*, 2022. 3

[38] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *European Conference on Computer Vision*, pages 194–210. Springer, 2020. 2, 3, 6

[39] John Phillips, Julieta Martinez, Ioan Andrei Bârsan, Sergio Casas, Abbas Sadat, and Raquel Urtasun. Deep multi-task learning for joint localization, perception, and prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4679–4689, 2021. 2, 3

[40] Zequn Qin, Jingyu Chen, Chao Chen, Xiaozhi Chen, and Xi Li. Uniformer: Unified multi-view fusion transformer for spatial-temporal representation in bird's-eye-view. *arXiv preprint arXiv:2207.08536*, 2022. 3

[41] Ruslan Rakhimov, Denis Volkhonskiy, Alexey Artemov, Denis Zorin, and Evgeny Burnaev. Latent video transformer. *arXiv preprint arXiv:2006.10704*, 2020. 4

[42] Thomas Roddick and Roberto Cipolla. Predicting semantic map representations from images using pyramid occupancy networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11138–11147, 2020. 3, 6

[43] Abbas Sadat, Sergio Casas, Mengye Ren, Xinyu Wu, Pranaab Dhawan, and Raquel Urtasun. Perceive, predict, and plan: Safe motion planning through interpretable semantic representations. In *European Conference on Computer Vision*, pages 414–430. Springer, 2020. 1, 3

[44] Avishkar Saha, Oscar Mendez, Chris Russell, and Richard Bowden. Translating images into maps. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 9200–9206. IEEE, 2022. 6

[45] Marcel Schreiber, Vasileios Belagiannis, Claudius Gläser, and Klaus Dietmayer. Dynamic occupancy grid mapping with recurrent neural networks. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6717–6724. IEEE, 2021. 3

[46] Meet Shah, Zhiling Huang, Ankit Laddha, Matthew Langford, Blake Barber, Sidney Zhang, Carlos Vallespi-Gonzalez, and Raquel Urtasun. Liranet: End-to-end trajectory prediction using spatio-temporal radar fusion. *arXiv preprint arXiv:2010.00731*, 2020. 1

[47] Zhihao Shi, Xiangyu Xu, Xiaohong Liu, Jun Chen, and Ming-Hsuan Yang. Video frame interpolation transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17482–17491, 2022. 4

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[49] Hengli Wang, Peide Cai, Yuxiang Sun, Lujia Wang, and Ming Liu. Learning interpretable end-to-end vision-based motion planning for autonomous driving with optical flow distillation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13731–13737. IEEE, 2021. 2, 6

[50] Yunlong Wang, Hongyu Pan, Jun Zhu, Yu-Huan Wu, Xin Zhan, Kun Jiang, and Diange Yang. Be-sti: Spatial-temporal integrated network for class-agnostic motion prediction with bidirectional enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17093–17102, 2022. 4

[51] Dirk Weissenborn, Oscar Täckström, and Jakob Uszkoreit. Scaling autoregressive video models. *arXiv preprint arXiv:1906.02634*, 2019. 4

[52] Kelvin Wong, Shenlong Wang, Mengye Ren, Ming Liang, and Raquel Urtasun. Identifying unknown instances for au-

tonomous driving. In *Conference on Robot Learning*, pages 384–393. PMLR, 2020. 2

[53] Pengxiang Wu, Siheng Chen, and Dimitris N Metaxas. Motionnet: Joint perception and motion prediction for autonomous driving based on bird's eye view maps. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11385–11395, 2020. 2, 3, 4, 7, 8

[54] Enze Xie, Zhiding Yu, Daquan Zhou, Jonah Philion, Anima Anandkumar, Sanja Fidler, Ping Luo, and Jose M Alvarez. Mˆ 2bev: Multi-camera joint 3d detection and segmentation with unified birds-eye view representation. *arXiv preprint arXiv:2204.05088*, 2022. 3

[55] Chenxin Xu, Maosen Li, Zhenyang Ni, Ya Zhang, and Siheng Chen. Groupnet: Multiscale hypergraph neural networks for trajectory prediction with relational reasoning. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 4

[56] Yunpeng Zhang, Zheng Zhu, Wenzhao Zheng, Junjie Huang, Guan Huang, Jie Zhou, and Jiwen Lu. Beverse: Unified perception and prediction in birds-eye-view for vision-centric autonomous driving. *arXiv preprint arXiv:2205.09743*, 2022. 2, 3, 6, 7, 8

[57] Brady Zhou and Philipp Krähenbühl. 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, 2022. 3, 6

[58] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 5