2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Distortion-Aware PETR for BEV Object Detection with Mixed Pinhole–Fisheye Cameras

Xiangzhong Liu

***Abstract*— Fisheye cameras are widely deployed in autonomous driving perception suites for their low cost and full-coverage field of view (FOV), yet their potential remains under-leveraged in 3D object detection. Severe radial distortion challenges most BEV detectors by violating the fundamental assumption of uniform sampling. To bridge this gap, we propose Distortion-Aware PETR (DAPETR), a projection-free detector tailored for mixed pinhole–fisheye camera setups. DAPETR incorporates two key learned-adaptive modules: a unified distortion-aware positional embedding that harmonizes positional encodings for image representations with fisheye geometry, and a bidirectional feature-geometry co-modulation module that mutually adapts image features and 3D positional embeddings. In our experiments on a converted KITTI-360 benchmark, we systematically compare our learned-adaptive approach against PETR in polar coordinates (PolarPETR). We find that while both methods improve over the baseline, our learned modules achieve superior performance. Crucially, we uncover a negative interaction when combining both strategies, revealing that learned adaptation and explicit geometric re-parameterization can conflict. Our final DAPETR model significantly advances the research and benchmark for fisheye BEV detection, providing critical insights into effective distortion-aware 3D perception design other than image rectification.**

I. INTRODUCTION

Modern autonomous driving systems increasingly rely on mixed camera configurations with pinhole and fisheye cameras with ultra-wide FOV coverage to achieve full 360° perception at manageable cost [1]. Despite the wide deployment of fisheye cameras, their potential for 3D object detection remains largely under-exploited. The fundamental challenge lies in the severe radial distortion that violates the uniform sampling assumption underlying most BEV-based detection frameworks [2]. Benchmarks like nuScenes [3] have substantially advanced BEV object detection with standard pinhole setups. However, there is still no widely adopted real-world benchmark for BEV detection that incorporates fisheye cameras.

Projection-free methods like PETR [4] bypass explicit geometric transformations and instead employ learned spatial priors encoded via 3D position embeddings. Implicit BEV approaches become difficult to implement facing complex sensor calibration for fisheye cameras. When camera parameters describe severe distortion, the fundamental correspondence between 2D image features and 3D spatial locations deteriorates. This calibration sensitivity and geometric breakdown lead to substantial performance degradation, preventing projection-free architectures from capitalizing on their computational efficiency advantages in mixed-camera systems.

Xiangzhong Liu is with Machine Learning Group, fortiss GmbH, Guerickestraße 25, 80805 Munich, Germany xiangzhong.liu@tum.de

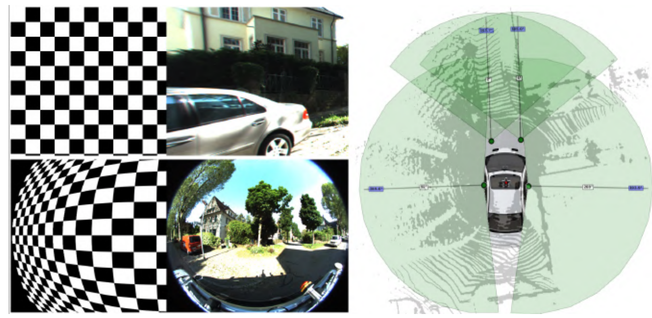

Fig. 1. Sensor configuration and full FOV coverage. Left: KITTI-360 provides both pinhole and fisheye cameras. The distortion characteristics are visualized by their respective checkerboard patterns(uniform grid spacing v.s. severe radial warping). Right: Top-down view of the KITTI-360 sensor setup with 2 front-seeing pinhole cameras and 2 side fisheye cameras overlapping with roof LiDAR, providing complete 360° surround coverage.

Polar BEV representations have shown promise for better alignment with camera frustum geometry [5], [6], [7], and naturally match the non-uniform distribution of scene content. However, their application has been restricted to improve BEV performance for unified pinhole cameras. All three methods employ explicit projection-based view transformation modules (VTMs), yet no work has explored adapting projection-free detectors to polar space. Specifically, polar representation aligns with fisheye geometry by preserving angular consistency across camera frustums and enforcing a more uniform sampling density in radial and angular dimensions. Combining projection-free efficiency with this polar geometric alignment could simultaneously achieve distortion robustness and computational efficiency, yet remains a critical unexplored research direction for fisheye 3D perception.

In this work, we introduce Distortion-Aware PETR (DAPETR), a projection-free BEV detector specifically designed for mixed pinhole–fisheye camera setups. We employ unified distortion-aware positional embeddings for both 2D pixel tokens and 3D position-aware features with a unified camera model. In addition, a bidirectional feature–geometry co-modulation module allows distortion-aware image features and 3D positional embeddings to refine one another for cross-attention stages. As an insightful comparison, we reformulate PETR to operate in polar coordinates as PolarPETR. We evaluate our method on the KITTI-360 dataset [8], which provides an ideal testbed with a combination of 2 front pinhole and 2 side fisheye cameras.

Our main contributions are:

- We develop unified distortion-aware positional embed-

dings for both 2D pixel tokens and 3D position-aware features through the MEI [9] camera model, harmonizing image representations with fisheye geometry for robust cross-attention under severe distortion.
- We introduce bidirectional feature-geometry co-modulation that mutually adapts distortion-aware image representations and 3D positional embeddings, enabling enhanced appearance-geometry alignment through joint spatial reasoning.
- We conduct the first systematic comparison between explicit geometric alignment (PolarPETR) and learned feature adaptation, revealing a negative interaction that provides critical insights for future distortion-aware model design.

II. RELATED WORK

A. BEV Multi-View 3D Detection

The BEV representation has become the de facto standard for multi-view 3D object detection, providing a unified space for sensor fusion and temporal modeling [2]. Current methods primarily differ in their view transformation module (VTM), which maps features from image space to the BEV representations.

Forward projection methods, pioneered by LSS [10], explicitly predict a depth distribution for each image pixel and lift 2D features into 3D space. BEVDet [11] and its successors [12] optimized this approach, but its accuracy relies heavily on the quality of the intermediate depth prediction, which is intractable for distorted fisheye images.

Backward projection approaches, such as BEVFormer [13] and DETR3D [14], pre-define a set of BEV reference/queries and project them back to 2D image planes to sample features with deformable attention. It avoids explicit depth prediction but restricts the receptive field to local area with significant computational cost.

Projection-free models like PETR [4] offer an alternative bypassing explicit feature projection. PETR enriches 2D image features with 3D positional embeddings derived from camera intrinsics and extrinsics. A decoder-only transformer then attends to spatial-aware features to directly predict 3D bounding boxes. While efficient, PETR's reliance on learned geometric priors makes it sensitive to camera configurations and calibrations. Our work is the first to address this limitation by adapting PETR to handle severe lens distortion.

B. Polar Coordinate Representation

The standard Cartesian BEV grid is ill-suited to the radial nature of the camera projections, leading to sparse and inefficient representations. Recent works have demonstrated that polar/cylindrical coordinates better match the non-uniform distribution of scene content and exploit view symmetry in surround-view systems.

PolarDETR [7] explicitly parameterizes 3D objects in polar coordinates (radial distance and angle) and decomposes velocities into radial/tangential components, achieving faster convergence and better accuracy than DETR3D. PolarBEVDet [6] adapts the BEVDet4D framework by replacing the standard Cartesian BEV with polar grids using angular-radial bins. It reformulates all core components, including view transformation, temporal fusion, and detection head, demonstrating that polar BEV representation better aligns with image information density (dense near vs. sparse far). PolarFormer [5] similarly adapted forward projection methods with transformers to polar BEV space for improved efficiency and uniformity.

These successes motivate our investigation into adapting projection-free detectors to a polar space handling fisheye distortion, combining geometric alignment advantages with computational efficiency of projection-free designs for fisheye 3D perception.

C. Distortion-Aware Modeling

Handling the severe distortion of fisheye lenses is a long-standing problem in computer vision. Traditional methods rely on pinhole or approximate rectification, which projects the fisheye image onto a virtual pinhole plane. However, this process introduces computational overhead to get projection mappings and information loss due to the limited FOV.

Distortion-Aware Architectures: More recent work has focused on building distortion-aware models. F2BEV [15] adapted BEVFormer with a fisheye 3D-to-2D projection with MEI, but was limited to a backward-projection architecture and synthetic fisheye images. Yogamani et al. [16] studied BEV semantic segmentation from surround-view fisheye cameras, showing that simple cylindrical undistortion [17] is suboptimal compared to distortion-aware learnable BEV pooling with occlusion reasoning. FisheyeDetNet [18] addresses 2D fisheye object detection through a data-centric approach for performance improvements, replacing standard rectangular boxes with rotated polygons for annotations in polar coordinates, but does not fundamentally adapt the model architecture for fisheye geometry. Beyond that, spherical transformer architectures [19] extend Vision Transformers to spherical projections, while DarSwin [20] presents a distortion-aware encoder-only architecture, offering generalized distortion modeling.

Camera-Aware Feature Learning: Another approach incorporates camera models directly into networks. CamConvs [21] appends per-pixel FOV and normalized pixel coordinates to image feature tensors, enabling depth networks to generalize across different intrinsics. Reichert et al. [22] extend this to diverse lenses using the unified camera model by distorting FOV maps according to lens parameters to yield distortion-robust features without explicit distortion inversion. RectConv [23] leverages kernel adjustments such that the convolutional filters see rectified patches while retaining the full FOV of fisheye images. Calibrated Convolutions [24] proposes deforming convolution kernels using the calibration parameters to adapt standard CNNs to fisheye distortion. While effective for single-image tasks, per-view adjustments cannot guarantee the consistent cross-camera spatial alignment required for multi-view BEV detection with heterogeneous sensor configurations.

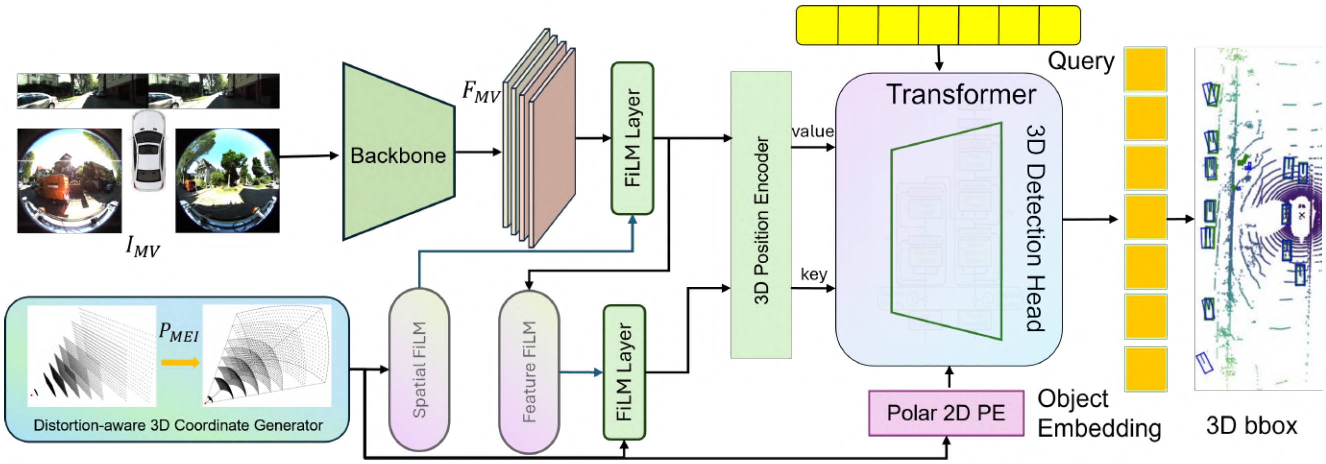

Fig. 2. An overview of the Distortion-Aware PETR pipeline. Multi-view images from mixed pinhole and fisheye cameras are fed into an image backbone. MEI camera model is applied for generating 3D coordinates for 3D positional encoding. We introduce a polar positional embedding (PE) for the 2D image features. A co-modulation module (e.g., FiLM) refines the image features based on the 3D query positions before the attention mechanism. The decoder then outputs 3D bounding box predictions.

D. Feature Modulation for Camera Adaptation

Feature-wise modulation techniques such as FiLM [25] and Squeeze-and-Excitation blocks [26] have proven effective for adapting visual features to contextual cues in language-guided vision and domain adaptation. These modulation networks generate scale and shift parameters that rescale intermediate activations based on the injected auxiliary information. However, their application to camera geometry and distortion handling remains unexplored. Different from the distortion-aware architectures, feature modulation offers a lightweight and unified approach to adapt features by dynamically conditioning intermediate representations on fisheye parameters and distortion maps. This enables view-specific adaptation while preserving consistency for feature extraction and cross-view fusion in BEV.

Our work pioneers the integration of FiLM-style conditioning with projection-free BEV detection, extending modulation beyond decoder features to include 3D positional embeddings of object queries. This creates the first end-to-end, geometry-aware solution for multi-camera 3D detection that explicitly addresses fisheye distortion through feature modulation.

III. METHODOLOGY

Our core contribution is a projection-free BEV detection framework, Distortion-Aware PETR, designed to handle the geometric challenges of mixed pinhole and fisheye camera systems. An overview of our method is shown in Fig. 2. The framework enhances the PETR architecture with learned distortion-adaptive modules, including unified distortion-aware positional embeddings for both 2D image features and 3D coordinates, and a bidirectional co-modulation mechanism that mutually refines image features and positional embeddings, creating a unified distortion-aware system while preserving PETR's projection-free efficiency.

A. Distortion Modeling in 3D Position Encoding

We adopt the unified MEI camera model [9] to handle mixed pinhole and fisheye lens distortion in KITTI-360. Unlike traditional pinhole cameras that assume rectilinear projection, fisheye cameras exhibit significant radial distortion requiring specialized modeling. The MEI model provides a mathematically elegant framework unifying pinhole and fisheye cameras through a single parameter set, enabling seamless integration across different camera types within the same multi-view system [15].

For a 3D point (X,Y,Z) in camera coordinates, the MEI model first projects the point to a unit sphere, then applies perspective projection with mirror parameter ξ, followed by radial distortion correction and image plane projection with $\mathbf{K}_f$:

$$\begin{aligned}
\mathbf{P}_s &= \mathbf{P}/\|\mathbf{P}\| \\
\mathbf{P}_c &= \left(\frac{X_s}{Z_s+\xi}, \frac{Y_s}{Z_s+\xi}\right) \\
r^2 &= X_c^2 + Y_c^2 \\
\mathbf{P}_d &= (1+k_1 r^2 + k_2 r^4) \times \mathbf{P}_c \\
\mathbf{P}_I &= \mathbf{K}_f \mathbf{P}_d
\end{aligned} \tag{1}$$

This unified formulation reduces to pinhole projection when $\xi = 0$ and $k_i = 0$, enabling consistent processing across mixed camera configurations. PETR employs 3D position encoding to establish spatial correspondences between image features and 3D queries. We replace 3D coordinate generation with a distortion-aware ray generation in the position-encoding module to account for non-linear distortion characteristics:

$$\mathbf{R}_{\mathrm{f}}(u,v,d) = \mathrm{UnprojectMEI}([u,v], \mathbf{K}_{\mathrm{f}}, \xi, k_1, k_2) \times \mathbf{D}_{\mathrm{u}} \tag{2}$$

where $\mathrm{UnprojectMEI}(\cdot)$ implements the inverse unified camera model transformation along uniformly distributed depths $\mathbf{D}_{\mathrm{u}}$. Following PETR's 3D Position Encoder [4], the unprojected 3D coordinates are transformed to 3D position embedding by a multi-layer perceptron (MLP) network $\psi(.)$.

Then the 2D features are flattened and added with position embedding to formulate 3D position-aware features for the Transformer decoder.

$$\mathbf{PE}_i^{3d}(t) = \psi\big(\mathbf{R}_{\mathrm{f}}(u,v,d)\big), \tag{3}$$

B. Feature-Positional Embedding Co-Modulation

In the original PETR, the 3D positional embedding is simply added to the image features. To create a more powerful fusion of geometric and appearance information, we introduce a bidirectional co-modulation (FPECoM) module that enables mutual refinement between image features and position embeddings. Our co-modulation approach operates in two stages using Feature-wise Linear Modulation (FiLM) [25]. First, we modulate the image features based on the distortion map derived from unprojecting image pixels to 3D coordinates (x_s, y_s, z_s) on a unit sphere. The distortion-derived scaling factor γ_d and bias factor β_d are computed from the converted spherical coordinate in (ρ, ϕ, z_s) format:

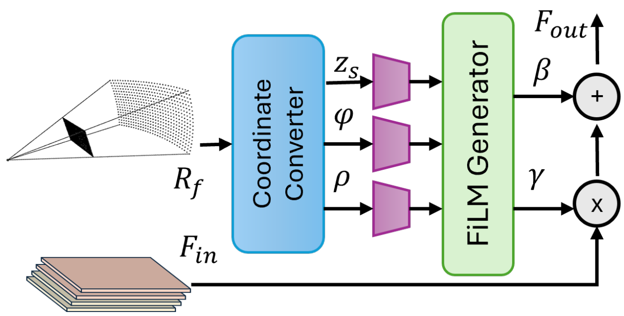

Fig. 3. The architecture of our spatial FiLM module. Unprojected rays (x_s, y_s, z_s) are converted to spherical coordinates (e.g., radial and elevation angles) as inputs. These are processed by separate encoders and a FiLM generator network to produce the spatial and channel-wise modulation parameters β and γ as outputs, which are then applied to the image features.

$$\mathbf{F}_{mod} = \gamma_d \odot F_{img} + \beta_d, \tag{4}$$

where F_{img} represents the original image features. Second, the 3D positional embedding is further guided by these distortion-modulated image features, where feature-aware factors γ_a and β_a are derived from F_{mod}:

$$PE_{out} = \gamma_a \odot PE_{in} + \beta_a, \tag{5}$$

where PE_{in} represents the distortion-aware 3D positional embedding. This bidirectional co-modulation allows the attention mechanism to learn how geometric distortion and visual appearance jointly influence spatial reasoning, creating a more robust feature-geometry correspondence under varying distortion conditions.

Similarly, PETRv2 [27] employs feature-guided positional encoding (FPE) with SE layers to enhance the positional embeddings to capture input dependency. Unlike SE layers, which merely reweight channels through a global gating vector, our FiLM-based co-modulation enables spatially adaptive scaling and shifting of features and positional embeddings. This richer affine transformation enables the network not only to emphasize embedding dimensions, as in FPE, but also to reposition them in the latent space based on local distortion and appearance cues. By jointly adapting appearance features and geometric priors, the decoder achieves better query–feature correspondence. In our experiments, we demonstrate that this mutual modulation approach significantly outperforms unidirectional alternatives like standard FiLM on features only and SE layers.

C. Polar 2D Positional Encoding

To make the 2D image features aware of the underlying fisheye geometry, we introduce a hybrid positional encoding scheme. This encoding augments the standard 2D positional information with distortion-aware polar coordinates derived from the camera's MEI model.

For each pixel (u, v) in an image from camera n, we generate a multi-component positional embedding. First, following standard practice in detection transformers, we compute sinusoidal embeddings for the pixel's grid position (x, y) and the camera index n. This provides the network with basic spatial and view-identity information.

Second, to encode the geometric distortion, we unproject each pixel to its corresponding 3D ray on the unit sphere, yielding a vector $\mathbf{P}_s = (x_s, y_s, z_s)$. From this, we derive polar coordinates on the sensor plane projection:

$$r = \sqrt{x_s^2 + y_s^2}, \quad \theta = \mathrm{atan2}(y_s, x_s) \tag{6}$$

where r is the radial distance from the principal point on the unit sphere projection and θ is the azimuth angle. These two components explicitly describe each pixel's position relative to the lens's optical center, directly encoding the non-linear nature of the fisheye projection.

Following [28], all components are converted into high-frequency embeddings using a sinusoidal function

$$\begin{aligned} \mathrm{PE}(p, 2i) &= \sin(p/T^{2i/d}), \\ \mathrm{PE}(p, 2i+1) &= \cos(p/T^{2i/d}) \end{aligned} \tag{7}$$

where p is the input position (x, y, n, r, θ), T is the temperature, d is the feature dimension, and i is the dimension index. By combining grid-based coordinates with geometry-derived polar coordinates, we provide the cross-attention mechanism with a rich, hybrid representation that preserves both the image's raster structure and the camera's intrinsic geometric properties, improving feature-query correspondence under severe distortion.

IV. EXPERIMENTS

Our work leverages the KITTI-360 dataset [8], a comprehensive dataset featuring both forward-facing pinhole cameras and two 190° fisheye cameras, making it ideal for mixed-camera 3D object detection. To enable standardized evaluation with the existing BEV detection frameworks, we utilize the KITTI-360 to nuScenes conversion pipeline introduced in [29]. We refer readers to [29] for comprehensive details regarding annotation alignment, calibration adaptation, and evaluation protocol adjustments necessary for this conversion.

TABLE I

COMPREHENSIVE EVALUATION ON FULL KITTI-360 DATASET.

Model	Backbone	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓	AP_{car} ↑	AP_{ped} ↑	AP_{bus} ↑
BEVDet	ResNet-50	0.121	0.159	0.736	0.481	1.031	1.017	0.462	0.096	0.000
Polar-BEVDet	ResNet-50	0.153	0.219	0.655	0.384	0.979	**0.846**	0.517	0.153	0.002
BEVFormer	ResNet-101	0.184	0.216	0.716	0.378	0.847	1.216	0.468	0.131	0.103
BEVFormer(F2BEV)	ResNet-101	0.167	0.236	0.720	0.370	**0.780**	0.982	0.508	0.136	0.071
PETR	VovNet-99	0.272	0.290	0.629	0.340	0.834	0.961	0.578	0.227	0.140
PETR(CAMConv)	VovNet-99	0.279	0.299	0.608	**0.339**	0.905	0.874	0.602	0.277	**0.146**
PolarPETR	VovNet-99	0.280	0.288	0.598	0.347	0.875	0.993	0.602	**0.296**	0.104
DAPETR	VovNet-99	**0.286**	**0.306**	**0.585**	0.342	0.816	0.951	**0.603**	0.270	0.139

We conduct a comprehensive set of experiments on the converted KITTI-360 benchmark to validate the effectiveness of DAPETR. Evaluation is performed on 10 classes: [car, truck, trailer, bus, bicycle, motorcycle, pedestrian, pole, object, traffic sign] using nuScenes benchmark metrics without average attribute error (AAE) and rebalanced weights for NDS calculation. For the ablation studies, we sampled approximately 11k out of 56k frames of 258 scenes for agile training. The validation split contains 8500 frames from 41 scenes.

A. Implementation Details

We use a VovNet-99 backbone to extract image features. The transformer decoder consists of 6 layers and 900 queries. We train the model for 24 epochs using the AdamW optimizer [30]. The initial learning rate is set to 2×10^{-4} and is decayed using a cosine annealing schedule. Our model is trained on 2 NVIDIA A5000 GPUs with a batch size of 8 and the learning rate is scaled linearly with the batch size. All models are implemented within the MMDetection3D [31] framework. For data preprocessing, we crop and resize the fisheye image to the same size as the pinhole image of 1408x376.

For the other models reported in Table I: Polar-BEVDet is adapted from PolarBEVDet without the temporal modeling and auxiliary 2d supervision. BEVFormer(F2BEV) is the small static version of BEVFormer without temporal modeling, and the VTM is replaced with the fisheye projection from F2BEV. For PETR(CAMConv), we append three further channels: (ρ, ϕ, z_s) from the unprojected 3D rays on the first convolution layer.

PolarPETR is our re-implementation of PETR in polar coordinates [29]. 3D PE and object queries of PETR are converted from Cartesian (x, y, z) to cylindrical coordinates. For each 3D point generated by PETR's frustum sampling, we compute (ρ, θ, z) and normalize with the maximum detection range $\rho_{\max}$ and angular range 2π, which are fed into the 3D position encoder. Object queries are initialized uniformly in polar space and passed directly to the transformer decoder, allowing the model to learn polar representations and predict offsets relative to polar reference points. For loss computation and the final box regression in the detection head, predicted (ρ, θ, z) and offsets are converted back to Cartesian coordinates to maintain compatibility with standard 3D detection annotations and evaluation protocols.

B. Main Results

We benchmark several state-of-the-art BEV detection frameworks on our converted KITTI-360 dataset, with results summarized in Table I.

First, we evaluate representative projection-based models. Both forward-projection (BEVDet) and backward-projection (BEVFormer) methods struggle with the native fisheye imagery, achieving only 0.121 and 0.184 mAP, respectively. While their distortion-aware (F2BEV) or polar variants (Polar-BEVDet) bring modest gains, they remain significantly outperformed by projection-free approaches, validating our choice to build upon PETR.

Our main analysis focuses on PETR variants. The standard PETR baseline on native fisheye images achieves a strong 0.272 mAP. We evaluate existing camera-aware adaptations like CAM-Conv, which provides a slight improvement. The explicit geometric approach, PolarPETR, also boosts performance, confirming the benefits of a polar representation.

Finally, our proposed model, DAPETR, which integrates our learned distortion-aware modules, achieves a new state-of-the-art on the native mixed-camera setup. It reaches 0.286 mAP and 0.306 NDS, demonstrating performance surpassing baseline. The performance is significant, as our method handles severe fisheye distortion natively without rectification, preserving the full field of view and image information.

Per-class AP analysis reveals a large performance disparity across classes in KITTI-360 dataset. While all models perform well on common classes like cars, they struggle on less frequent classes like buses, due to the extremely limited samples (cars: 430K, buses: 1K). Although the number of pedestrian samples is comparable to nuScenes dataset, the performance gap is still large, which reveals the challenge of detecting small objects with fisheye images.

Runtime and Efficiency Analysis: To validate our claims of efficiency, we compare the computational costs of DAPETR against the baseline models in Table II. Inference speeds (FPS) and detection head GFLOPs are measured on a single GPU with a batch size of 1. As shown, DAPETR introduces nominal overhead: bidirectional feature modulation adds merely 0.57M parameters and 4.67 GFLOPs to the detection head compared to the baseline. Meanwhile, it maintains a highly competitive inference speed of 6.8 FPS (compared to 7.0 FPS for PETR). This demonstrates that our distortion-aware mechanism preserves the inherent efficiency of projection-free architectures while significantly boosting detection performance.

TABLE II
COMPUTATIONAL COST COMPARISON

Method	Params (M)	Head GFLOPs	FPS
PETR	81.80	51.27	7.0
PolarPETR	82.06	53.39	7.7
DAPETR (Ours)	82.37	55.94	6.8

C. Ablation Studies

We perform a series of ablation studies to analyze the contribution of each component, with results shown in Table III. Our baseline is PETR applied to native fisheye images, where only the fundamental distortion model (DM) for ray unprojection is used. This performs worse than the baseline on rectified images (23.3 vs 25.0 mAP), highlighting the limit of only distortion modeling. With further enhancements, we all observed consistent transcendence over the rectified baseline on the small-scale dataset.

Effect of Polar Representation. Introducing a polar representation for BEV queries and 3D coordinates improves performance significantly over the baseline (+1.6 mAP). This confirms that explicitly aligning the query space with the radial distortion nature of fisheye geometry provides a good inductive bias.

Effect of Co-Modulation. Independently, spatial modulation (SM) and feature-guided modulation (FgM) also improve performance. For FgM, using FiLM is notably more effective than a simpler SE layer (+0.8 mAP), demonstrating the benefit of its affine transformation for fusing geometric and appearance information.

TABLE III
ABLATION STUDY ON MODEL COMPONENTS. DM: DISTORTION MODELLING, PR: POLAR REPRESENTATION, SM: SPATIAL MODULATION, FGM: FEATURE-GUIDED MODULATION, PPE: POLAR POSITIONAL ENCODING. THE FIRST ROW INDICATES THE PERFORMANCE OF BASELINE ON RECTIFIED IMAGES OF KITTI-360.

DM	PR	SM	FgM	PPE	mAP↑	NDS↑
					25.0	25.3
✓					23.3	25.7
✓	✓				24.9	26.8
✓		✓			25.2	26.8
✓		✓	SE		25.4	26.0
✓		✓	FiLM		26.2	27.8
✓		✓	FiLM	✓	**26.3**	**28.5**
✓	✓	✓	FiLM	✓	25.5	27.5

Spatial FiLM Position Ablation: We further study where to place the spatial FiLM module that uses a distortion map to modulate image features for distortion awareness. Intuitively, conditioning early in the network should allow the model to normalize distortion from the start and propagate geometry-aware features through all subsequent stages. However, our results show the opposite trend: applying spatial FiLM at the detection head yields the best performance, while inserting it in the backbone or neck underperforms (Table IV). We hypothesize that early modulation may overfit low-level textures to camera-specific distortion and harm feature generality across views; the neck mixes multi-scale features and may dilute the signal. In contrast, head-level modulation acts closest to the cross-attention and box decoding, aligning geometry and appearance precisely where queries meet features. This also keeps the shared backbone features largely camera-agnostic, improving cross-camera consistency.

TABLE IV
SPATIAL FILM PLACEMENT ABLATION ON KITTI-360. HEAD-LEVEL PLACEMENT PERFORMS BEST, CONTRARY TO THE HYPOTHESIS THAT "EARLIER IS BETTER."

Placement	mAP ↑	NDS ↑
Backbone	23.5	25.7
Neck	23.3	25.9
Head	**25.4**	**26.0**

Effect of Polar Positional Encoding. Building on the best co-modulation model, adding the polar positional encoding (PPE) for 2D image tokens further improves performance, achieving our best result of 26.3 mAP and 28.5 NDS. This shows that providing the 2D features with explicit, token-wise information about the fisheye geometry is crucial for robust cross-attention.

Conflicting Interaction with Polar Representation. However, when we combine the explicit polar representation (PR) with our best-performing model (SM + FgM + PPE), performance unexpectedly decreases by 0.8 mAP. This suggests a conflict between the two strategies. The learned co-modulation and positional encoding likely create their own implicit correction for the fisheye geometry, which becomes redundant or even counter-productive when applied to a BEV space that is already explicitly warped into polar coordinates. This finding indicates that explicit geometric re-parameterization and learned feature adaptation should be treated as competing, rather than complementary, design choices.

D. Range and Angular Stratified Results

We evaluate the detection performance tiered by distance and angle. As shown in Table V, both PolarPETR and our DAPETR outperform the baseline PETR across all distances. Notably, DAPETR shows the most significant gains in the 10-30m range, improving mAP by over 1.4% compared to PolarPETR and 5.5% compared to the baseline PETR. This suggests that our learned distortion-aware modules are particularly effective at handling the smaller object sizes at mid-range distances.

Similarly, the angular-stratified results in Table VI demonstrate the effectiveness of our approach in full 360° perception. DAPETR achieves the highest mAP in the front and

TABLE V

MAP (%) ACROSS DISTANCE RANGES ON KITTI-360 MIXED PINHOLE-FISHEYE IMAGES.

Method	0–10 m	10–20 m	20–30 m	30–40 m	40–50 m
PETR	54.42	30.75	12.01	4.21	1.01
PolarPETR	56.57	35.52	12.49	4.50	1.44
DAPETR	57.29	36.27	13.94	4.85	1.36

TABLE VI

ANGULAR-STRATIFIED MAP (%) ON KITTI-360 MIXED PINHOLE-FISHEYE IMAGES.

Method	Front 120°	Back 120°	Sides 120°
PETR	29.84	21.87	28.77
PolarPETR	31.35	22.68	29.84
DAPETR	32.19	23.86	28.67

back sectors, with particularly notable gains in the front sector (+2.35% over baseline PETR). In the back sector, which is primarily covered by fisheye cameras, DAPETR improves mAP by 1.99% over the baseline, confirming that our learned distortion-aware modules effectively mitigate fisheye distortions. For the side views, PolarPETR achieves the best performance at 29.84%, outperforming DAPETR by 1.17%. This suggests that explicit polar representation may offer advantages for certain viewing angles, while our learned approach excels in the more challenging front and back regions where accurate depth estimation and feature-geometry alignment are critical for detection accuracy.

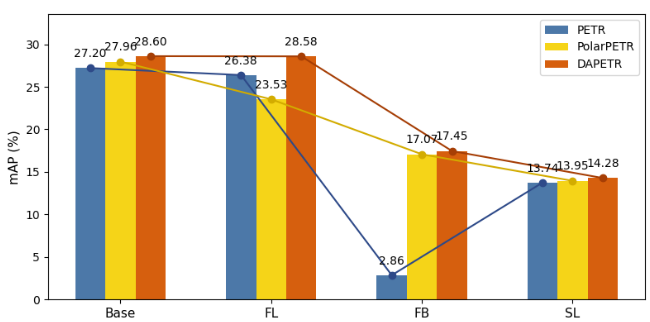

Fig. 4. Model robustness under camera failure scenarios. FL, FB and SL denote the front-left, front-both and side-left.

E. Camera Loss Robustness

To evaluate the fault tolerance of our approach, we test model robustness under camera failure scenarios by randomly mocking cameras during inference. As shown in Fig. 4, for front-left camera failure (FL), all methods show relatively modest degradation, with DAPETR maintaining 28.58% mAP compared to 26.38% for PETR. This suggests that both front cameras can partially compensate for the missing view. Although the FOVs of both front cameras overlap significantly, the BEV object detection does not benefit much from the redundancy.

The most significant degradation occurs when both front cameras are lost (FB), leaving only the side fisheye cameras for perception. Here, DAPETR demonstrates superior resilience with 17.45% mAP, substantially outperforming both PETR (3.86%). Without specifically designed distortion-aware modules, the baseline PETR nearly fails completely, as it cannot effectively leverage the fisheye views alone.

For side-left camera failure (SL), all methods show graceful degradation, with DAPETR still achieving the highest mAP. The relatively smaller performance drop compared to front camera failures indicates that the remaining cameras can provide sufficient coverage for reasonable detection performance. Overall, these results confirm that DAPETR's learned distortion adaptation creates more robust feature representations when parts of the sensor suite fail.

F. Qualitative Results

Fig. 5. Qualitative comparison between the baseline PETR (left) and our DAPETR (right). Our model demonstrates improved localization and classification accuracy. From top to bottom: better rotation estimation for cars, tighter scale for a motorcycle, and the correct classification of a truck, which the baseline misses.

As illustrated in Fig. 5, our DAPETR model produces significantly improved qualitative results compared to the baseline PETR. The left column shows the baseline's predictions, while the right column shows our model's predictions. In the top row, DAPETR demonstrates more accurate orientation estimation for cars in the near distance. The middle row highlights better scale prediction, with our model producing a much tighter bounding box for the motorcycle. Finally, the bottom row shows a clear case of improved classification and detection; our model correctly identifies the large vehicle as a truck, whereas the baseline fails to detect it, showcasing the benefits of our distortion-aware approach in challenging fisheye views.

V. CONCLUSION

In this paper, we addressed the challenge of 3D object detection in mixed pinhole-fisheye camera systems by proposing Distortion-Aware PETR (DAPETR). Instead of relying on explicit geometric transformations or rectification, DAPETR incorporates two novel learned-adaptive modules: a unified distortion-aware positional embedding using

MEI for both 2D image features and 3D position coordinates, and a bidirectional co-modulation module to mutually refine image features and 3D positional embeddings. Through extensive experiments on the converted KITTI-360 dataset, we found that projection-free methods (PETR) prove most adaptable, achieving the highest mAP compared to projection-based methods (variants of BEVDet and BEVFormer). Our robustness analysis confirms that DAPETR maintains superior performance even under camera failure scenarios, retaining viability with fisheye-only perception.

Although polar re-parameterization has been a common strategy for BEV object detection, even we proved its effectiveness in fisheye settings, we discovered a negative interaction when combining it with our feature modulation approach. This suggests that learned feature adaptation and explicit polar transformation are competing, rather than complementary choices. Our DAPETR model, relying on learned adaptation, achieves the best performance in all our experiments, particularly in challenging mid-range distances and fisheye-covered views. This work not only delivers a fisheye 3D object detector without rectification but also provides practical guidance for future research in distortion-aware perception. Future work will focus on addressing the remaining challenges in detecting small objects and assessing cross-dataset generalizability on other fisheye datasets.

REFERENCES

[1] V. R. Kumar, C. Eising, C. Witt, and S. K. Yogamani, "Surround-view fisheye camera perception for automated driving: Overview, survey & challenges," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 4, pp. 3638–3659, 2023.

[2] H. Li, C. Sima, J. Dai, W. Wang, L. Lu, H. Wang, J. Zeng, Z. Li, J. Yang, H. Deng, H. Tian, E. Xie, J. Xie, L. Chen, T. Li, Y. Li, Y. Gao, X. Jia, S. Liu, J. Shi, D. Lin, and Y. Qiao, "Delving into the devils of bird's-eye-view perception: A review, evaluation and recipe," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–20, 2023.

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

[4] Y. Liu, T. Wang, X. Zhang, and J. Sun, "Petr: Position embedding transformation for multi-view 3d object detection," in *European conference on computer vision*, pp. 531–548, Springer, 2022.

[5] Y. Jiang, L. Zhang, Z. Miao, X. Zhu, J. Gao, W. Hu, and Y.-G. Jiang, "Polarformer: Multi-camera 3d object detection with polar transformer," in *Proceedings of the AAAI conference on Artificial Intelligence*, vol. 37, pp. 1042–1050, 2023.

[6] Z. Yu, Q. Liu, W. Wang, L. Zhang, and X. Zhao, "Polarbevdet: Exploring polar representation for multi-view 3d object detection in bird's-eye-view," *arXiv preprint arXiv:2408.16200*, 2024.

[7] S. Chen, X. Wang, T. Cheng, Q. Zhang, C. Huang, and W. Liu, "Polardetr: Polar parametrization for vision-based surround-view 3d detection," *Image and Vision Computing*, vol. 156, p. 105438, 2025.

[8] Y. Liao, J. Xie, and A. Geiger, "Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 3292–3310, 2022.

[9] C. Mei and P. Rives, "Single view point omnidirectional camera calibration from planar grids," in *Proceedings 2007 IEEE International Conference on Robotics and Automation*, pp. 3945–3950, IEEE, 2007.

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

[11] J. Huang, G. Huang, Z. Zhu, Y. Ye, and D. Du, "Bevdet: High-performance multi-camera 3d object detection in bird-eye-view," *arXiv preprint arXiv:2112.11790*, 2021.

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

[13] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Q. Yu, and J. Dai, "Bevformer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[14] Y. Wang, V. C. Guizilini, T. Zhang, Y. Wang, H. Zhao, and J. Solomon, "Detr3d: 3d object detection from multi-view images via 3d-to-2d queries," in *Conference on robot learning*, pp. 180–191, PMLR, 2022.

[15] E. U. Samani, F. Tao, H. R. Dasari, S. Ding, and A. G. Banerjee, "F2bev: Bird's eye view generation from surround-view fisheye camera images for automated driving," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 9367–9374, IEEE, 2023.

[16] S. Yogamani, D. Unger, V. Narayanan, and V. R. Kumar, "Fisheyebevseg: Surround view fisheye cameras based bird's-eye view segmentation for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1331–1334, 2024.

[17] S. Yogamani, C. Hughes, J. Horgan, G. Sistu, P. Varley, D. O'Dea, M. Uricár, S. Milz, M. Simon, K. Amende, *et al.*, "Woodscape: A multi-task, multi-camera fisheye dataset for autonomous driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9308–9318, 2019.

[18] G. Sistu and S. Yogamani, "Fisheyedetnet: 360 {\deg} surround view fisheye camera based object detection system for autonomous driving," *arXiv preprint arXiv:2404.13443*, 2024.

[19] O. Carlsson, J. E. Gerken, H. Linander, H. Spieß, F. Ohlsson, C. Petersson, and D. Persson, "Heal-swin: A vision transformer on the sphere," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6067–6077, 2024.

[20] A. Athwale, A. Afrasiyabi, J. Lagüe, I. Shili, O. Ahmad, and J.-F. Lalonde, "Darswin: Distortion aware radial swin transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5929–5938, 2023.

[21] J. M. Facil, B. Ummenhofer, H. Zhou, L. Montesano, T. Brox, and J. Civera, "Cam-convs: Camera-aware multi-scale convolutions for single-view depth," in *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pp. 11826–11835, 2019.

[22] H. Reichert, M. Hetzel, A. Hubert, K. Doll, and B. Sick, "Sensor equivariance: A framework for semantic segmentation with diverse camera models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1254–1261, 2024.

[23] R. Griffiths and D. G. Dansereau, "Adapting cnns for fisheye cameras without retraining," *arXiv preprint arXiv:2404.08187*, 2024.

[24] B. Berenguel-Baeta, M. Santos-Villafranca, J. Bermudez-Cameo, A. P. Yus, and J. Guerrero, "Convolution kernel adaptation to calibrated fisheye," in *34th British Machine Vision Conference 2023, BMVC 2023, Aberdeen, UK, November 20-24, 2023*, BMVA, 2023.

[25] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, 2018.

[26] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, 2018.

[27] Y. Liu, J. Yan, F. Jia, S. Li, A. Gao, T. Wang, and X. Zhang, "Petrv2: A unified framework for 3d perception from multi-camera images," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3262–3272, 2023.

[28] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*, pp. 213–229, Springer, 2020.

[29] X. Liu and H. Shen, "Benchmarking multi-view bev object detection with mixed pinhole and fisheye cameras," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2026. (to appear).

[30] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[31] M. Contributors, "MMDetection3D: OpenMMLab next-generation platform for general 3D object detection." `https://github.com/open-mmlab/mmdetection3d`, 2020.