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

Benchmarking Multi-View BEV Object Detection with Mixed Pinhole and Fisheye Cameras

Xiangzhong Liu[1] and Hao Shen[2]

***Abstract*— Modern autonomous driving systems increasingly rely on mixed camera configurations with pinhole and fisheye cameras for full view perception. However, Bird's-Eye View (BEV) 3D object detection models are predominantly designed for pinhole cameras, leading to performance degradation under fisheye distortion. To bridge this gap, we introduce a multi-view BEV detection benchmark with mixed cameras by converting KITTI-360 into nuScenes format. Our study encompasses three adaptations: rectification for zero-shot evaluation and fine-tuning of nuScenes-trained models, distortion-aware view transformation modules (VTMs) via the MEI camera model, and polar coordinate representations to better align with radial distortion. We systematically evaluate three representative BEV architectures, BEVFormer, BEVDet and PETR, across these strategies. We demonstrate that projection-free architectures are inherently more robust and effective against fisheye distortion than other VTMs. This work establishes the first real-data 3D detection benchmark with fisheye and pinhole images and provides systematic adaptation and practical guidelines for designing robust and cost-effective 3D perception systems. The code is available at `https://github.com/CesarLiu/FishBEVOD.git`.**

I. INTRODUCTION

Current BEV multi-view 3D object detection (3DOD) systems have achieved remarkable success on standardized datasets like nuScenes [1], where uniform pinhole camera configurations provide consistent geometric properties for all views. However, this setup diverges from real-world autonomous driving systems that adopt mixed camera configurations for cost-effectiveness. Serial production vehicles combine pinhole cameras for forward-facing perception with fisheye cameras for surround-view visualization, near-field parking assistance, and 2D object recognition rather than 3D perception tasks [2], achieving 360° coverage at a fraction of the cost of an all-pinhole system.

While BEV methods excel on pinhole cameras [3], their performance remains largely underexplored on mixed configurations, especially regarding fisheye cameras. Fisheye images present several fundamental challenges:

1) Severe radial distortion causes non-uniform spatial sampling and non-linear transformations, leading to inconsistent object scales and feature representations.
2) The wide field of view compresses distant objects, introducing ambiguity and information loss in BEV projections.
3) Depth estimation is unreliable in highly distorted regions, hindering accurate 3D lifting.

[1]Xiangzhong Liu and [2]Hao Shen are with Machine Learning Group, fortiss GmbH, Guerickestraße 25, 80805 Munich, Germany `xiangzhong.liu@tum.de`, `shen@fortiss.org`

Fig. 1: Qualitative detection results on KITTI-360 mixed camera configuration. 3D bounding box and point cloud rendering overlaid on pinhole and fisheye images. The visualization is created with a customized nuScenes devkit. Our method successfully handles the severe radial distortion in fisheye cameras while maintaining consistent detection accuracy across different camera types.

These issues fundamentally disrupt the spatial consistency assumptions underlying BEV models, leading to significant performance degradation of existing BEV architectures in 3D perception tasks [4]. Due to the lack of large-scale datasets with fisheye imagery and reliable 3D annotations, this issue has remained largely unaddressed in current literature.

The KITTI-360 dataset [5] presents a unique opportunity to address this gap, with a comprehensive sensor suite including two 180° fisheye cameras alongside forward-facing stereo cameras, creating a simplified mixed configuration that resembles real-world deployments. While KITTI-360 has been utilized for 3D semantic segmentation and monocular object detection, we identify our work as the first to systematically evaluate multi-view BEV-based 3DOD on a real-world dataset with mixed imagery.

While rectification offers conceptual simplicity, it introduces computational overhead and information loss through cropping. Direct fisheye modeling preserves the full FoV advantages while eliminating rectification overhead, allowing models to adapt to non-linear projections by learning [4]. We

investigate three representative BEV architectures: PETR [6] (projection-free), BEVFormer [7] (backward projection), and BEVDet [8] (forward projection). We propose polar coordinate transformation to align the BEV representation with the fisheye distortion, as polar grids naturally capture the angular consistency and radial compression inherent to wide-angle lenses [9], [10]. Our contributions are summarized as follows:

- **First Multi-View Benchmark on KITTI-360:** We establish KITTI-360 as a comprehensive benchmark for mixed camera BEV object detection, developing conversion pipelines to enable standardized evaluation with the nuScenes devkit.
- **Systematic Architecture Adaptation:** We implement both geometric rectification and distortion modeling with MEI camera model across three VTMs, providing the first systematic comparison of these paradigms.
- **Polar Representation tailored for Fisheye Geometry:** We introduce polar coordinate transformations for fisheye camera MEI model, better reflecting the fisheye radial structure and enabling geometry-aware feature alignment in the BEV space.

While recent advances such as temporal modeling [11], [12], [13] and 2D auxiliary detection heads [14] have proven effective for general BEV pipelines, our study focuses specifically on the geometric challenges posed by distortion.

II. RELATED WORK

A. BEV Multi-View 3D Detection

BEV representation has emerged as the dominant paradigm for 3DOD in autonomous driving [3], transforming multiple perspective views into a unified top-down coordinate system. Based on their view transformation modules, current approaches are categorized into three families [15]:

Forward Projection. LSS [16] pioneered this paradigm by predicting per-pixel depth distributions, then lifting 2D features into 3D space via geometric projection. BEVDet [8] and BEVFusion [17] extend this approach but remain vulnerable to depth estimation errors—particularly problematic for geometrically distorted fisheye imagery.

Backward Projection. BEVFormer [7] predefines 3D coordinates and projects reference points back to 2D images through deformable cross-attention, enabling denser BEV representations. BEVFormerv2 [13] improved efficiency through perspective supervision and temporal modeling.

Projection-Free. PETR [6] bypasses explicit geometric projection by enriching 2D features with 3D positional embeddings derived from camera parameters. StreamPETR [12] extended this with temporal cues. While computationally efficient, these methods rely on learned geometric priors potentially sensitive to camera configuration changes.

Polar BEV Representation. Traditional multi-view 3DOD methods employ Cartesian BEV representations with uniform grid rasterization, creating a structural mismatch with the natural radial distribution of camera frustum information. PolarFormer [9] and PolarBEVDet [10] replace rectangular representations with polar coordinates (r, θ) for more coherent multi-view feature aggregation.

These methods excel on pinhole-camera datasets like nuScenes but inherently assume linear projection models, making their performance on real-world fisheye images with severe distortions largely unexplored and potentially problematic.

B. Fisheye Camera Modeling in 3D Perception

The integration of fisheye cameras into 3D perception pipelines presents critical challenges due to severe radial distortion that violates the linear perspective assumptions underlying standard computer vision algorithms [2]. Fisheye lenses achieve ultra-wide fields of view (typically 180° or greater) through intentional geometric distortion [18], creating non-uniform spatial sampling, which violates the translational equivariance of standard CNNs.

Traditional approaches address fisheye distortion via rectification, mapping distorted images to approximate pinhole views [2]. This preprocessing enables direct application of existing BEV models without architectural modifications, but introduces computational overhead, inevitable information loss through cropping and non-uniform resolution distribution that affect feature quality for downstream tasks. Recent methods bypass rectification by directly integrating fisheye camera models in learning-based approaches. F2BEV [4] pioneered this approach by replacing BEVFormer's pinhole projection with the MEI camera model [19]. Beyond that, spherical transformer architectures [20] extend Vision Transformers to spherical projections, while DarSwin [21] presents a distortion-aware encoder-only architecture, offering generalized distortion modeling. RectConv [22] introduces distortion-aware rectified convolution that can handle camera distortion effects directly within the convolution operation. These models eliminate rectification loss while preserving the full field of view, but require geometry-aware coordinate and sampling adaptations.

C. Benchmarks with Fisheye Imagery

Current real-world datasets primarily target 2D perception tasks, reflecting the historical limitations of fisheye cameras for situational awareness. WoodScape [23] provides 40K fisheye images but lacks public 3D annotations. SynWoodScape [24] addresses 3D annotation absence through synthetic data generation, though its real-world transferability remains questionable, with only 500 samples publicly available. FB-SSEM [4] provides a Unity-generated synthetic dataset featuring 20,000 samples of fisheye SVC images with corresponding BEV maps for simulated parking lot scenarios. Major autonomous driving benchmarks (nuScenes [1], Waymo [25]) exclusively use pinhole cameras, leaving mixed-camera evaluation underexplored.

KITTI-360 [5] uniquely provides both forward-facing stereo pinhole cameras and dual 180° fisheye cameras with high-quality LiDAR-derived 3D annotations. However, it has been primarily used for 3D semantic segmentation with

limited BEV detection exploration due to the format incompatibilities with nuScenes framework.

This work addresses this by developing the first systematic KITTI-360 to nuScenes conversion pipeline, establishing a standardized mixed-camera benchmark for BEV 3DOD. Our contribution enables the direct application of nuScenes-compatible models to fisheye images, fostering research into practical perception systems with mixed or pure fisheye cameras.

III. METHODOLOGY

Our core methodological contribution is a framework for the adaptation of existing BEV-based 3DOD models to mixed pinhole and fisheye camera configurations and evaluation on a real-world dataset. Our approach is three-pronged. First, to enable rigorous and standardized evaluation, we establish a new multi-view benchmark by converting the KITTI-360 dataset into the nuScenes format. Second, we perform a systematic adaptation of three representative VTMs with MEI for unified camera modeling. Additionally, we introduce polar coordinate transformations that naturally align with fisheye geometry to improve spatial feature aggregation.

A. Dataset Conversion and Rectification

While KITTI-360 provides rich urban scene annotations with comprehensive 360° sensor coverage for both cameras and LiDARs [5], existing BEV detection methods are predominantly designed around nuScenes data structure. The format conversion enables the first multi-view evaluation on European urban driving scenarios, complementing existing nuScenes evaluations from America and Asia.

a) Conversion Pipeline: The XML/pose annotations differ significantly from the tokenized JSON structure of nuScenes and static objects are only labeled once globally without frame assignments. We convert KITTI-360 to nuScenes format in three key steps:

1) **Scene partitioning and split definitions** - We adopt the official train/val splits and partition long sequences into scene windows (200m) aligned with nuScenes.
2) **Sample identification** - Keyframes are defined by matching accurate georegistered vehicle poses to synchronized data frames, excluding non-keyframes to maintain dense annotations.
3) **Annotation conversion** - Static objects are assigned to frames via distance and LiDAR-visibility checks, and objects are transformed into nuScenes' center–size–quaternion format. CityScapes' label definitions are mapped to nuScenes' 10 detection classes, with extra KITTI-360 categories preserved as extensions (e.g. pole and traffic signs).

b) Geometric Rectification: For comparison, we establish a rectification baseline by transforming fisheye cameras into virtual pinhole views. Each fisheye generates two cameras with the same focal length as the front ones: forward (30° rotation) and backward-facing (−46° rotation), with −4° downward pitch. Together with the stereo pinhole cameras, this creates a 6-camera setup comparable to nuScenes as shown in Figure 2. We evaluate this rectified dataset in two ways: zero-shot inference with nuScenes-trained models to measure transferability, and fine-tuning to identify the domain gap and establish performance upper bounds. While rectification enables compatibility with existing BEV detectors, it also suffers from reduced FoV, resolution distortion and computational overhead, motivating the direct fisheye-aware approaches.

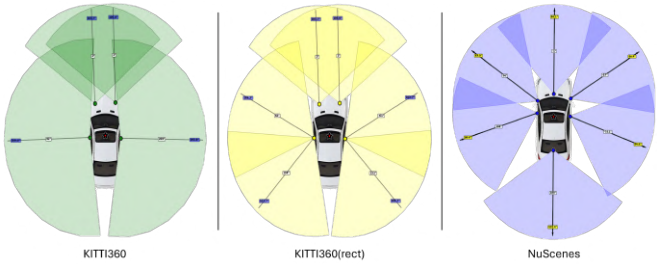

Fig. 2: Camera configuration and full field-of-view coverage comparison. Left: Original KITTI-360 with stereo cameras and fisheye cameras. Middle: Rectified KITTI-360 with 6 pinhole cameras. Right: NuScenes configuration with 6 pinhole cameras.

B. Distortion-Aware View Transformation

We adopt the unified MEI camera model [19] to handle mixed pinhole and fisheye lens distortion in the KITTI-360 dataset. 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, fisheye, and catadioptric cameras through a single parameter set, enabling seamless integration across different camera types within the same multi-view system [4].

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.

a) 3D Position Encoding for PETR: PETR employs 3D position encoding to establish spatial correspondences between image features and 3D queries. We replace the reference points generation with our distortion-aware ray generation process in the position encoding module to account for non-linear distortion characteristics:

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

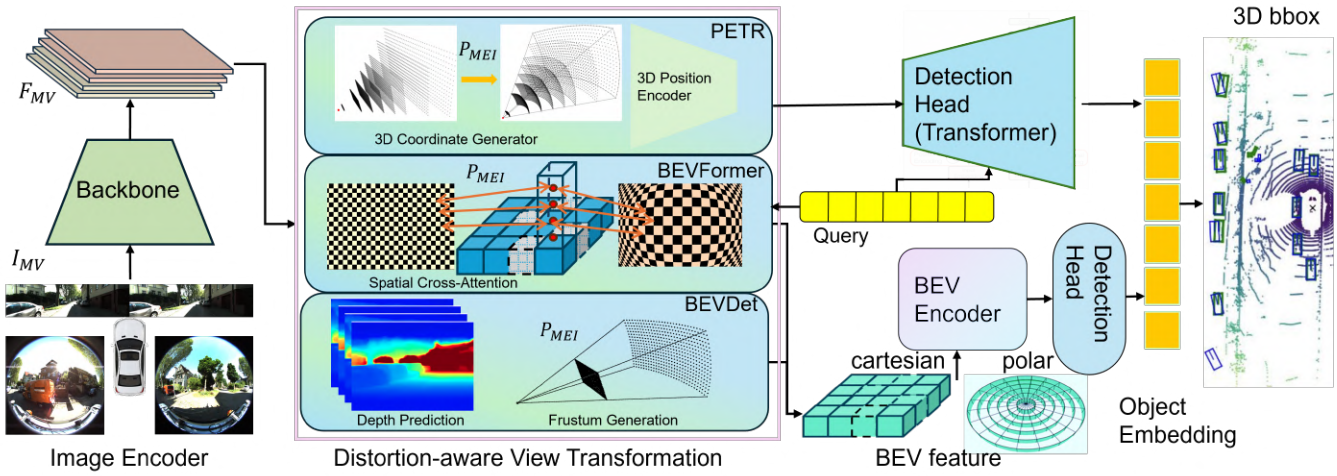

Fig. 3: Overview of distortion-aware BEV 3D object detection framework. Multi-view images (pinhole+fisheye) are processed by a shared backbone encoder, then fed into three distortion-aware view transformation modules via MEI camera model. The resulting BEV features can be represented in either Cartesian or polar coordinates to better align with fisheye geometry. A detection head(Transformer or CNN) processes BEV features to produce final 3D object detection outputs.

where UnprojectMEI$(\cdot)$ implements the inverse unified camera model transformation along uniformly distributed depth $\mathbf{D}_{\text{u}}$. The 3D position encoding queries $\mathbf{Q}_{\text{pos}}$ are subsequently computed as:

$$\mathbf{Q}_{\text{pos}} = \text{PE}_{\text{fpe}}(\mathbf{R}_{\text{f}}(u,v,d)) \tag{3}$$

where PE_{fpe} denotes the feature-guided positional encoding (FPE) function [26]. The FPE mechanism makes PE data-dependent by leveraging features learned from raw images to adaptively modulate the positional embedding. Specifically, a small MLP ϕ processes the projected 2D features from the distortion-aware ray $\mathbf{R}_{\text{f}}(u,v,d)$, yielding attention weights that adapt according to the underlying distorted image content. This feature-dependent encoding helps the model implicitly consider depth cues and geometric relationships that are particularly challenging in distorted image regions. By incorporating fisheye-specific geometric priors through the MEI unprojection, the downstream cross-attention layers attend along geometrically accurate fisheye rays rather than erroneous straight rays ignoring radial distortion, while the FPE mechanism ensures that positional embeddings correctly capture the non-uniform spatial sampling and varying information density characteristic of distorted imagery.

b) Spatial Cross-Attention in BEVFormer: BEVFormer builds BEV representations by applying spatial cross-attention between 3D BEV queries and 2D image features. We introduce a distortion-aware spatial cross-attention module that explicitly models non-linear projection effects of distortion in the reference point sampling (RPS) module for better spatial alignment.

The RPS is reformulated as:

$$\mathbf{p}_{\text{ref}} = \text{ProjectMEI}(\mathbf{q}_{\text{bev}} + \Delta\mathbf{p}, \mathbf{K}_{\text{fisheye}}, \xi, k_1, k_2) \tag{4}$$

where BEV queries $\mathbf{q}_{\text{bev}}$ are first lifted to 3D space and ProjectMEI$(\cdot)$ applies the unified camera model to project 3D reference points to 2D image feature space as a primary alignment. The sampling offset $\Delta\mathbf{p}$ is predicted by the deformable attention mechanism accounting for the non-uniform information density in fisheye images, particularly in edge regions where distortion is strong:

$$\text{Attn}(\mathbf{q}_{\text{bev}}, \mathbf{f}_{\text{img}}) = \sum_{k=1}^{K} W_k \cdot \mathbf{f}_{\text{img}}(\mathbf{p}_{\text{ref}_k}) \cdot \gamma(\mathbf{p}_{\text{ref}_k}) \tag{5}$$

where W_k are the learned attention weights, K is the number of sampling points, and $\gamma(\mathbf{p}_{\text{ref}_k})$ represents distortion-aware geometric weighting.

c) Depth-Based Lifting of BEVDet: Unlike PETR's projection-free approach that relies on learned positional embeddings, BEVDet performs view transformation through explicit depth estimation. This architectural difference necessitates MEI integration directly into the geometric transformation pipeline rather than the positional encoding stage.

We modify BEVDet's VTM to handle fisheye distortion by incorporating the MEI camera model into the depth-based lifting operation as PETR to generate 3D frustums $\mathbf{P}_{3D}$, but with $\mathbf{D}_{\text{pred}}$, which is predicted with the depth estimation head along 3D rays that account for radial distortion. The subsequent BEV projection maintains the original formulation:

$$\mathbf{P}_{\text{BEV}} = \text{ProjectBEV}(\mathbf{P}_{3D}, \mathbf{T}_{\text{cam2ego}}) \tag{6}$$

This modification preserves BEVDet's explicit depth supervision and LSS design philosophy while ensuring geometrically accurate feature lifting under distortion. By addressing the fundamental challenge of unreliable depth cues in heavily distorted regions through proper geometric modeling, this approach enables BEVDet to maintain its depth-based advantages.

C. Polar Coordinate Transformation

Recognizing that the uniform-sampling assumption in Cartesian representations is invalid for fisheye imagery, we

redesign the BEV parameterization in cylindrical coordinates (ρ, θ, z). In an equidistant fisheye model [18], pixels sample equal azimuth increments, yet those increments cover large ground-plane spans near the image center and small spans near the rim. A polar grid representation aligns with fisheye geometry by:

- Preserving angular consistency across camera frustums,
- Enforcing uniform sampling density in radial and angular dimensions, and
- Enabling more natural feature aggregation for wide-angle perception

a) Polar-PETR: We re-parameterize PETR's 3D position encoding and object queries from Cartesian (x, y, z) to cylindrical coordinates to better align with fisheye geometry. For each 3D point generated by PETR's frustum sampling, we compute its polar position (ρ, θ, z) and normalize with the maximum detection range $\rho_{\max}$ and angular range (2π), which are further fed into the 3D position encoder. To handle the angular wrap-around at $\theta = 0/2\pi$, we deploy sinusoidal positional encoding, $[\sin\theta, \cos\theta, \rho_{\text{norm}}, z_{\text{norm}}]$ ensuring continuous embeddings across the angular boundary.

Object queries (anchors) are initialized uniformly in polar space and directly fed to the transformer decoder without conversion, as we expect the transformer to learn polar representations and predict spatial offsets to the polar reference points. This approach places queries on radial beams at fixed angular increments, naturally exploiting multi-view symmetry where objects at the same radius appear similarly across different camera views. However, for loss calculation and final box regression in the detection head, predicted (ρ, θ, z) coordinates and offsets are converted back to Cartesian coordinates, to ensure compatibility with standard 3D detection evaluation protocols and ground truth annotations.

b) Polar-BEVFormer: Within BEVFormer's spatial cross-attention, reference points sampled from the BEV grid are likewise re-parameterized from Cartesian to polar coordinates.

The reference point generator creates normalized polar coordinates that serve as 3D anchors for deformable attention sampling. For camera projection, polar coordinates are converted to Cartesian form before applying standard camera intrinsics and extrinsics. This conversion occurs within the spatial cross-attention module in the Transformer encoder, allowing existing projection code to remain unchanged while enabling polar BEV queries to attend to appropriate image regions.

BEV query embeddings are enhanced with polar-aware positional encoding. Instead of Cartesian coordinates, we encode each query's (ρ, θ) position using sinusoidal functions applied to normalized polar coordinates, ensuring that the transformer captures radial symmetry and angular relationships inherent to fisheye geometry while maintaining spatial consistency across different radial distances and angular orientations.

c) PolarBEVDet: We build upon the existing PolarBEVDet implementation [10], which adapts the Lift-Splat-Shoot pipeline into a polar grid. Frustum points (u, v, d) are back-projected into polar coordinates with modeling the distortion via MEI model, then "splatted" into $N_\theta \times N_\rho$ angular-radial bins via simple indexing and sum-pooling. Features are aggregated via sum pooling into a regular feature map $F \in \mathbb{R}^{C \times N_\theta \times N_r}$, naturally aligning with fisheye geometry while preserving full field-of-view coverage. This plug-and-play implementation provides a polar BEV map with minimal changes to the original depth-lifting implementation.

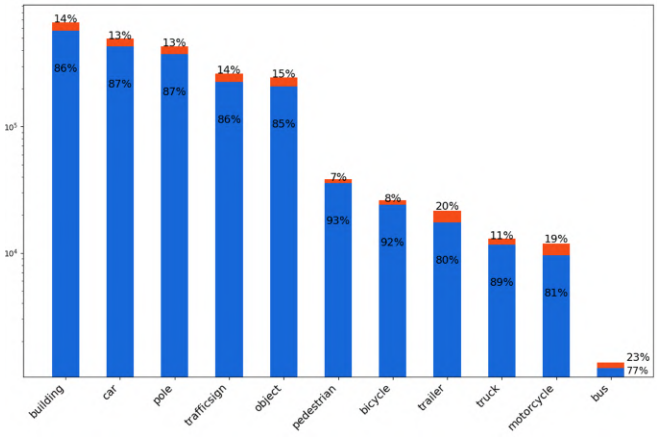

Fig. 4: Class distribution in KITTI-360 dataset (log scale). The annotation distribution exhibits significant class imbalance, heavily skewed toward static infrastructure objects, while dynamic objects relevant to autonomous driving other than car represent a smaller portion. Blue bars indicate training samples, orange bars for validation, with consistent ratios maintained across splits.

IV. EXPERIMENTS

To validate our proposed distortion-aware adaptation framework, we conduct a series of experiments designed to answer three key questions:

- What is the performance degradation when SotA BEV models are naively applied to fisheye data?
- How effectively does our direct adaptation approach, using the MEI camera model, improve performance for different model paradigms?
- Does our proposed polar coordinate positional encoding provide additional benefits for the most suitable architecture?

A. Experimental Setup

a) Dataset and Preprocessing: The converted KITTI-360 dataset comprises 55,526 training samples from 258 distinct scenes, and we validate on 8,554 samples from 41 scenes that are geographically separate from the training scenes to ensure fair evaluation of generalization. The dataset exhibits significant class imbalance, as shown in Figure 4. While the training and validation splits maintain consistent class ratios, the annotation distribution is heavily skewed toward static objects (buildings, poles and traffic signs) compared to movable objects for autonomous driving.

TABLE I: Comprehensive evaluation on KITTI-360 fisheye benchmark. Results show zero-shot performance of nuScenes-trained models, baseline training ignoring distortion, our distortion-aware adaptations with MEI camera model, polar coordinate enhancements, and upper-bound performance on rectified data.

Model	Backbone	Method	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓	AP_{car} ↑	AP_{ped} ↑	AP_{bus} ↑
BEVDet	ResNet-50	Zero-Shot	0.008	0.030	1.036	0.830	1.139	0.957	0.055	0.000	0.000
		Baseline	0.121	0.159	0.736	0.481	1.031	1.017	0.462	0.096	0.000
		DA-VTM	0.150	0.212	0.686	0.400	1.089	**0.816**	0.513	0.138	**0.015**
		+ Polar	**0.153**	**0.219**	**0.655**	**0.384**	**0.979**	0.846	**0.517**	**0.153**	0.002
		Rectification	0.169	0.232	0.673	0.388	1.004	0.760	0.538	0.189	0.024
BEVFormer	ResNet-101	Zero-Shot	0.034	0.083	0.998	0.469	1.031	1.770	0.195	0.009	0.000
		Baseline	0.167	0.216	0.716	0.378	0.847	1.216	0.468	**0.131**	**0.103**
		DA-VTM	0.167	**0.230**	0.731	0.361	**0.736**	**1.027**	0.467	0.113	0.081
		+ Polar	**0.175**	0.221	**0.715**	**0.357**	0.858	1.032	**0.494**	0.120	0.066
		Rectification	0.218	0.275	0.660	0.357	0.792	0.862	0.566	0.187	0.098
PETR	VovNet-99	Zero-Shot	0.001	0.039	1.119	0.698	1.157	1.125	0.010	0.000	0.000
		Baseline	0.266	0.280	0.668	0.340	0.885	**0.932**	0.597	0.260	0.046
		DA-VTM	0.269	0.283	**0.596**	**0.334**	0.884	1.057	**0.606**	0.230	**0.137**
		+ Polar	**0.280**	**0.288**	0.598	0.347	**0.875**	0.993	0.602	**0.296**	0.104
		Rectification	0.295	0.337	0.629	0.337	0.839	0.677	0.610	0.320	0.146

b) Implementation Details: All models are implemented within the MMDetection3D [27] framework. We maintain the original backbone configurations for each model: VovNet-99 for PETR, ResNet-101 for BEVFormer (small, static version), and ResNet-50 for BEVDet, all pre-trained on ImageNet [28]. The fisheye images are cropped to 1408×376 resolution to maintain a unified image tensor size. For BEVDet, we preserve the original input height and downscale images to 960×256 for a fair comparison with baselines. All models are trained in a distributed manner for 24 epochs with a batch size of 8, using the AdamW optimizer [29] with an initial learning rate scaled according to the number of used NVIDIA A5000 GPUs, decayed using a cosine annealing schedule. Beyond the core architectural changes described in Section III, all other hyperparameters follow the default configurations for the respective models in the MMDetection3D library to isolate the impact of the introduced adaptations. 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.

B. Main Results and Analysis

a) Zero-Shot Performance and Rectification: The zero-shot results in Table I reveal the severe impact of geometric and sensor setup mismatch when applying nuScenes-trained models directly to fisheye data. All models exhibit catastrophic performance degradation, with PETR and BEVDet achieving near-zero mAP and BEVFormer slightly better at 0.034 mAP. This confirms that models trained exclusively on pinhole camera images suffer from severe data domain gap.

The rectification results establish performance upper bounds. PETR achieves the highest overall performance (0.295 mAP, 0.337 NDS), followed by BEVFormer and BEVDet. Per-class AP reveals strong car detection (0.538–0.610) but poor results for underrepresented classes like buses (<0.1) and small-scale objects like pedestrians (0.2–0.3), especially for projection-based methods. PETR shows a more balanced performance, suggesting projection-free architectures handle class imbalance better. The dataset's extreme skew toward static infrastructure (buildings: 572K, poles: 375K) versus dynamic objects (cars: 430K, buses: 1K) biases projection-based models toward dominant classes, reducing overall performance.

b) Distortion-Aware View Transformation: When trained directly on fisheye data (Baseline) while ignoring distortion modeling, performance varies significantly across architectures. PETR achieves the highest baseline performance (0.266 mAP, 0.280 NDS), followed by BEVFormer (0.167 mAP, 0.216 NDS), while BEVDet shows more modest results (0.121 mAP, 0.159 NDS) with slightly lower input resolution. This ranking suggests that projection-free architectures (PETR) are inherently more robust to geometric inconsistency than other methods, as they can partially compensate for distortion through learned representations rather than explicit geometric transformations. These results confirm that the pinhole assumption embedded in these architectures fundamentally breaks down when confronted with fisheye distortion, leading to incorrect feature projection and poor spatial reasoning.

Integrating the MEI camera model (DA-VTM) produces varied results across architectures. PETR shows modest improvement (0.269 mAP, 0.283 NDS), benefiting from geometrically accurate 3D positional encoding. BEVFormer maintains similar mAP (0.167) but achieves better NDS (0.230) through improved True Positive metrics. Most notably, BEVDet demonstrates significant gains (0.150 mAP, 0.212 NDS), a 24% relative mAP improvement over baseline, suggesting that explicit distortion modeling particularly benefits depth-based approaches. Examining error metrics, models show reduced translation error (mATE) with DA-VTM, confirming better spatial localization after proper distortion handling.

c) Polar Coordinate Enhancement: Adding polar coordinate representations (+ Polar) consistently improves all models, with PETR achieving the best overall performance (0.280 mAP, 0.288 NDS), while BEVFormer (0.175 mAP, 0.221 NDS) and BEVDet (0.153 mAP, 0.219 NDS) also show clear benefits. Per-class analysis reveals that polar coordinates particularly enhance pedestrian detection, with PETR achieving 0.296 AP (vs. 0.230) and BEVDet reaching 0.153 AP (vs. 0.138). This confirms our hypothesis that polar coordinates naturally align with fisheye geometry, providing more intuitive spatial representations for transformer-based reasoning about radially distorted images, especially for smaller objects that suffer more from distortion effects.

C. Robustness Analysis

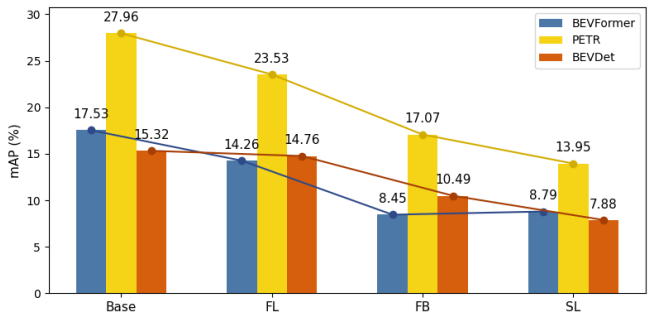

Fig. 5: Model robustness under camera failure scenarios. FL, FB and SL denote the front-left, front-both and side-left.

a) Robustness under camera failure: Figure 5 shows the mAP performance under different camera loss scenarios. PETR consistently outperforms other models across all configurations, with BEVFormer and BEVDet showing different robustness patterns. When losing the front-left camera (FL), BEVDet shows remarkable resilience with only a 3.7% drop, while BEVFormer and PETR experience larger decreases. With both front cameras disabled (FB), PETR maintains 61.1% of its baseline performance, demonstrating stronger fisheye-only perception capabilities than others. Side-left camera failure (SL) significantly impacts all models with similar severity, causing approximately 50% performance drops across architectures. This suggests side cameras provide critical information that cannot be fully compensated by remaining views. PETR's consistent superiority confirms that projection-free architectures better handle geometric uncertainty in mixed camera configurations.

b) Performance along distance ranges: Table II reports mAP across distance bins for three BEV architectures. For all models, accuracy declines rapidly with range, highlighting the challenge of long-range perception in fisheye views. Training with rectified images excels at nearly all distances over distortion-aware models, as rectification restores pinhole-like geometry that these architectures were originally designed for. However, distortion-aware PETR achieves the best near-field score (56.57% 0–10 m), slightly higher than on rectified images, while both BEVFormer and BEVDet show slightly lower performance with distortion modeling than with rectification (-2.16% and -0.95% respectively). Beyond 10m, rectified models consistently outperform their distortion-aware counterparts across all architectures, with the performance gap widening at longer ranges. These results suggest that distortion-aware modeling preserves near-field detail effectively but struggles with long-range depth estimation and spatial reasoning.

TABLE II: mAP (%) across distance ranges on KITTI-360. (rect) for models on rectified images and (ours) represent distortion-aware models on mixed inputs.

Method	0–10 m	10–20 m	20–30 m	30–40 m	40–50 m
BEVFormer(rect)	49.54	26.08	7.87	2.17	0.63
BEVFormer (ours)	47.38	20.17	5.43	1.83	0.40
PETR(rect)	55.87	37.70	14.95	4.83	1.45
PETR (ours)	56.57	35.52	12.49	4.50	1.44
BEVDet(rect)	35.17	19.90	6.41	2.35	0.57
BEVDet (ours)	34.22	18.90	5.64	1.65	0.36

TABLE III: Angular-stratified mAP (%) on KITTI-360. Baseline for standard training on mixed images and (ours) = distortion-aware models.

Method	Front 120°	Back 120°	Sides 120°
BEVFormer baseline	**19.42**	12.56	18.08
BEVFormer (ours)	19.36	**14.06**	**21.44**
PETR baseline	29.84	21.87	28.77
PETR (ours)	**31.35**	**22.68**	**29.84**
BEVDet baseline	14.43	7.81	13.67
BEVDet (ours)	**17.58**	**12.05**	**17.30**

c) Angular-stratified performance: Table III compares baseline and distortion-aware models across three azimuthal sectors: front (120°, primarily covered by pinhole cameras), back (120°, severe fisheye distortion), and sides (120°, moderate distortion). Distortion-aware adaptations yield clear improvements in regions with higher distortion, like side and back areas. Interestingly, only BEVFormer shows a slight decrease in front-facing mAP (-0.06), while all other models demonstrate improvements across all sectors, confirming that unified MEI camera modeling does not degrade performance on pinhole views. BEVDet demonstrates the most substantial relative improvements, with gains of +3.15 mAP (+21.8%) in front, +4.24 mAP (+54.3%) in back, and +3.63 mAP (+26.6%) in side regions. PETR, despite starting from a higher baseline, still achieves meaningful gains, particularly in front regions (+1.51). The consistent improvements in the side and back regions across all architectures confirm that proper distortion modeling is particularly critical for areas where fisheye distortion is most pronounced, enabling more accurate spatial reasoning for 3D object detection.

V. CONCLUSIONS

This work addresses the critical gap between mixed pinhole-fisheye camera configurations and BEV 3D object detection with distortion-aware adaptations and enhance-

ments. Through systematic evaluation on our converted KITTI-360 benchmark for multi-view 3DOD, we demonstrate that directly applying pinhole-trained models to an out-of-distribution dataset results in catastrophic failure with near-zero mAP. Our analysis reveals architectural differences in handling fisheye distortion. Projection-free methods (PETR) prove most adaptable, achieving highest mAP with distortion-aware 3D positional encoding along the polar coordinate enhancement. Backward-projection approaches (BEVFormer) show moderate gains. While starting from a lower baseline, forward-projection methods (BEVDet) show the most dramatic relative gains from our adaptations. These trends suggest that decoupling feature learning from explicit geometry improves adaptability to sensor variation. Introducing polar positional embeddings consistently benefits all methods, confirming that aligning BEV grids with the radial nature of fisheye distortion is a powerful geometric prior. Our robustness analysis confirms that PETR maintains superior performance even under camera failure scenarios, retaining viability with fisheye-only perception. These findings establish architectural guidelines for fisheye-based 3D detection: projection-free designs are preferable to depth-dependent approaches, and geometric priors should align with sensor characteristics. While our work demonstrates the feasibility of distortion-aware BEV detection, significant challenges remain, particularly for small object detection in highly distorted image regions and handling severely imbalanced sample distributions that bias models toward dominant classes.

REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

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

[12] S. Wang, Y. Liu, T. Wang, Y. Li, and X. Zhang, "Exploring object-centric temporal modeling for efficient multi-view 3d object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3621–3631, 2023.

[13] C. Yang, Y. Chen, H. Tian, C. Tao, X. Zhu, Z. Zhang, G. Huang, H. Li, Y. Qiao, L. Lu, *et al.*, "Bevformer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 17830–17839, 2023.

[14] S. Wang, X. Jiang, and Y. Li, "Focal-petr: Embracing foreground for efficient multi-camera 3d object detection," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 1, pp. 1481–1489, 2023.

[15] Z. Li, Z. Yu, W. Wang, A. Anandkumar, T. Lu, and J. M. Alvarez, "Fb-bev: Bev representation from forward-backward view transformations," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6919–6928, 2023.

[16] 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.

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

[18] J. Kannala and S. S. Brandt, "A generic camera model and calibration method for conventional, wide-angle, and fish-eye lenses," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 8, pp. 1335–1340, 2006.

[19] 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.

[20] 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.

[21] 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.

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

[23] 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.

[24] A. R. Sekkat, Y. Dupuis, V. R. Kumar, H. Rashed, S. Yogamani, P. Vasseur, and P. Honeine, "Synwoodscape: Synthetic surround-view fisheye camera dataset for autonomous driving," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 8502–8509, 2022.

[25] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, *et al.*, "Scalability in perception for autonomous driving: Waymo open dataset," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2446–2454, 2020.

[26] 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.

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

[28] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009.

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