Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Decoupling Geometry and Content: A Reliability-Aware Attention Framework for Robust BEV Perception

Peiyuan Qu
Hainan University
Mengxing Huang
1005467006@qq.com

Hainan University

Research Article

Keywords:

Posted Date: February 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8888255/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Decoupling Geometry and Content: A Reliability-Aware Attention Framework for Robust BEV Perception

Abstract

Multi-camera Bird's-Eye-View (BEV) perception is fundamental to autonomous driving. However, existing query-based methods suffer from a representational coupling bottleneck: a single BEV query is compelled to simultaneously encode spatial invariance for geometric localization and semantic sensitivity for content assessment. Shallow linear projections in standard attention mechanisms struggle to disentangle these conflicting attributes, leading to sampling instability. Furthermore, the conventional unconditional feature fusion paradigm lacks explicit quality assessment, allowing low-quality observations to contaminate temporal features. To address these issues, we propose a Self-Regulating Attention (SRA) framework. A Structurally Decoupled Geometry-Content Attention (SDGCA) replaces the shallow linear projections of standard attention with dedicated sub-networks, enabling it to actively decouple and extract task-specific representations from the BEV query. We posit that a highly reliable feature update should not only stem from a geometrically sound sampling process, but also from a high-confidence content aggregation. Guided by this insight, we introduce the Cooperative Assessment Gating (CAG), which leverages the affine transformation matrix and information entropy from SDGCA to rigorously evaluate the reliability of the proposal along these two orthogonal dimensions. An adaptive gating module then arbitrates by fusing these evaluation signals with the query context to dynamically regulate the feature update intensity, significantly improving update reliability. Experiments on the nuScenes dataset demonstrate that SRA-Former achieves a state-of-the-art mAP of 0.425 and significantly reduces orientation estimation errors compared to strong baselines.

Crucially, these improvements are obtained without relying on explicit depth or height supervision. This research validates that structural decoupling and reliability-aware gating alone can establish a high-performance BEV perception paradigm, offering a robust alternative to prior-dependent methods.

Keywords: BEV Perception, Self-Regulating Attention, Structural Attention Decoupling, Affine Transformation Matrix, Shannon Entropy, Feature Robustness

1. Introduction

Accurate 3D environmental perception is fundamental to autonomous driving [1, 2]. Bird's-Eye-View (BEV) perception, which unifies multi-view data onto a top-down grid, has become a dominant paradigm for temporal fusion [3, 4]. However, reconstructing a dense BEV representation from sparse 2D images remains a formidable challenge [5]. First, the inherent depth ambiguity in 2D-to-3D mapping makes estimating object height and longitudinal position precarious, imposing stringent demands on the geometric precision of sampling points [6, 7]. Second, environmental complexity presents significant hurdles: distant instances exhibit sparse texture information, whereas nearby objects are prone to occlusion and truncation [8]. Addressing these extremes mandates that the model possesses both fine-grained feature extraction and holistic spatial reasoning abilities [9].

Existing methodologies incorporate explicit geometric priors to constrain view transformation or optimize internal attention mechanisms for implicit representation learning. While incorporating explicit depth or height supervision yields immediate gains in localization metrics, they do not resolve the inherent representational conflicts within the feature extraction process itself. We argue that if the core attention operator remains inefficient due to coupled learning objectives, the resulting BEV features will lack discriminability regardless of the geometric constraints applied. Therefore, optimizing this fundamental mechanism to ensure sampling stability and semantic sensitivity is a prerequisite to

advancing the upper limits of BEV perception. Following this trajectory, query-based methods like BEVFormer [10] have established a strong baseline. Despite its success the standard Spatial Cross Attention design suffers from a fundamental bottleneck where a single shared BEV query is compelled to simultaneously encode the conflicting attributes of spatial invariance for geometric localization and semantic sensitivity for content assessment [11, 12]. Relying on shallow linear projections to resolve these fundamentally orthogonal objectives creates an ill-posed optimization problem, leading to confounded gradient signals that destabilize the sampling process. Furthermore, existing paradigms employ an unconditional update mechanism that lacks explicit reliability assessment, allowing noise to contaminate temporal representations [13]. Consequently, robust perception demands structural decoupling to resolve these conflicts and a reliability assessment mechanism to filter invalid updates.

We propose the Self-Regulating Attention (SRA) framework to replace the conventional unconditional attention mechanism. First, the Structurally Decoupled Geometry-Content Attention (SDGCA) module substitutes the standard shallow projections with two dedicated sub-networks: a geometry sub-network is tasked exclusively with predicting sampling points via geometric transformation primitives (affine matrices) to resolve geometric localization, while a content sub-network is responsible for evaluating feature importance. This structural decoupling endows the module with the capability to actively disentangle and extract task-specific representations from the shared query input, significantly enhancing sampling stability and feature extraction precision.

Beyond sampling stability, we argue that a robust fusion paradigm must evaluate the incoming update along two orthogonal axes: the rationality of the geometric sampling and the confidence of the content aggregation. To implement this assessment, we introduce the Cooperative Assessment Gating (CAG) framework. Specifically, the Geometry Assessor utilizes the affine transformation matrix generated by SDGCA to quantify geometric stability (process score S_{geo}). Simultaneously, the Content Assessor calculates the Shannon entropy of the attention weights to measure content certainty (result score S_{con}).

However, a naive entropy-based assessment faces an inherent semantic ambiguity. While high entropy typically signifies uncertainty arising from uninformative background regions, similar flat attentional distributions also characterize valid observations of large-scale objects exhibiting homogeneous textures. To resolve this conflict, we introduce a context-aware MLP as the arbitration mechanism. Distinct from context-free gating strategies, this module explicitly integrates the high-level semantic context embedded in the BEV query $\mathbf{q}$ with the reliability scores, formulated as $g = f(S_{geo}, S_{con}, \mathbf{q})$. This design enables the model to effectively differentiate between uncertainty-induced dispersion and semantically valid uniformity, thereby dynamically modulating the feature update intensity based on context-specific reliability.

In summary, SRA-Former systematically addresses the dual challenges of sampling instability and update unreliability. Without relying on external depth supervision, our approach achieves state-of-the-art detection accuracy (mAP) and significantly reduced orientation and velocity errors compared to baseline methods. The SRA framework propels BEV feature fusion from an unconditional mode to a paradigm of adaptive, regulated decision-making.

2. Related Work

To establish the theoretical groundwork for our reliability mechanism, we examine the transition from geometric projection to query-based attention strategies.

2.1. The Paradigm Evolution of BEV Perception

BEV construction methodologies are primarily categorized into geometry-based and query-based routes [14, 15]. Geometry-based methods, pioneered by LSS [5] and its variants [14, 16, 17, 18], lift 2D features to 3D via explicit depth prediction and aggregate them onto the BEV grid through pooling [19]. However, reliance on explicit depth renders LSS fragile, as estimation errors in complex regions propagate to degrade BEV features [20, 21]. While recent

works attempt to mitigate this by modeling depth uncertainty, the depth dependency limits performance, motivating a shift towards implicit view transformation models [22, 23].

Query-based approaches offer end-to-end transformation and have become mainstream [24, 25]. Such approaches aggregate multi-view features by encoding 3D coordinates as positional priors, as demonstrated in PETR and its subsequent work [26, 27]. BEVFormer evolves this using learnable queries and spatio-temporal attention, circumventing explicit depth reliance to achieve superior performance [10]. However, this implies that the performance ceiling is largely constrained by the core view-transformation module, namely the attention-based sampling mechanism. Consequently, in-depth analysis and refinement of this core sampling mechanism are pivotal to advancing this field.

2.2. Explorations in Decoupling Attention Mechanisms

Deformable Attention has become a standard in vision transformers, overcoming quadratic complexity by attending to sparse points [28, 29, 30]. Although efficient, its core mechanism suffers from inherent *representational coupling*. It relies on a single shared BEV query, using linear projections to simultaneously perform geometric localization and content feature assessment [31]. This compels the query to simultaneously encode conflicting spatial invariance and semantic sensitivity, causing objective conflicts that confound optimization and limit performance [32, 33].

To alleviate this representation bottleneck, improvement strategies have predominantly focused on architectural modifications rather than resolving the core computational conflict within the attention mechanism. At the query input stage, Conditional DETR [34] and DAB-DETR [11] introduce stronger spatial priors via conditional queries or dynamic anchor boxes. Regarding structural design, Decoupled DETR [35] separates prediction heads to mitigate misalignment, while HV-BEV [36] and recent hierarchical frameworks [37] investigate decoupling feature sampling across spatial dimensions or perception tasks; additionally, DINO [38] optimizes convergence strategies. However, these approaches

operate primarily at the macro-architectural level. Inside the attention operation, both sampling offsets and attention weights remain dependent on shallow linear projections from the same shared query. This implies that the core sampling module retains conflicting learning objectives, rendering the process sensitive to initialization and prone to instability [39, 31]. Thus, explicitly decoupling geometric positioning from content evaluation via specialized internal architectures is necessary to fundamentally break this coupling [40, 41].

2.3. Robustness Mechanisms for BEV Feature Updating

Autonomous driving perception systems demand not only high precision but also exceptional robustness in complex environments [6, 42]. In temporal fusion, standard BEV models employ an unconditional residual update, implicitly assuming that all attention-derived vectors are beneficial. However, this assumption fails in occluded or uninformative regions [43], where invalid updates act as noise, contaminating the temporal BEV state [44].

Existing research primarily addresses this via temporal fusion gating and perceptual uncertainty modeling. At the fusion level, learnable gating regulates update strength [45], exemplified by TA-GRU in RecurrentBEV [46] or saliency-guided mechanisms in SP-Det [47], which prioritize informative regions. However, derived from sequence models designed for 1D temporal data, these methods prove inefficient for high-dimensional, sparse spatial updates [48]. Another class focuses on quantifying perceptual uncertainty [19], including explicit modeling or entropy-based methods [19]. Bayesian methods like MCDropout incur prohibitively high computational costs, while ensemble learning increases complexity, rendering both impractical for real-time perception. More critically, they provide *a posteriori* metrics failing to distinguish data from cognitive uncertainty, preventing intervention to avoid temporal contamination [49]. Contrasting with uncertainty quantification, feature recalibration methods like SENet [50] or CBAM [51] amplify useful features rather than suppressing noise, failing to prevent low-quality updates. However, existing methods still lack a mechanism to quantify the reliability of spatial information aggregated by the

attention module itself. To ensure stability, explicit evaluation of attention updates is required prior to fusion, enabling adaptive gating to realize robust feature fusion.

3. Methodology

To address the dual challenges of sampling instability and update unreliability, we propose the SRA framework. Through the integration of dedicated computational structures and explicit evaluation mechanisms, our framework ensures stable and robust BEV perception.

3.1. Overall Framework

As illustrated in Fig. 1, SRA-Former is built upon BEVFormer baseline [10]. While preserving the established encoder-decoder paradigm, our core contribution is replacing the Spatial Cross-Attention (SCA) with the Self-Regulating Attention (SRA), enabling the model to actively regulate feature aggregation reliability without altering the temporal fusion or decoding pipelines. Unlike standard attention mechanisms that blindly accept updates, the SRA framework generates candidate features via the SDGCA module while simultaneously utilizing the CAG framework to evaluate their validity based on geometric and semantic cues prior to fusion. This design enables the model to explicitly weigh the incoming information and filter out unreliable updates derived from unstable sampling or uncertain content.

3.2. Structurally Decoupled Geometry-Content Attention (SDGCA)

In standard attention mechanisms, a single BEV query $\mathbf{q}$ is tasked with simultaneously encoding conflicting geometric information and semantic sensitivity. However, conventional shallow linear projections fail to disentangle this coupled representation, limiting performance in complex scenes. To address this optimization bottleneck, the Structurally Decoupled Geometry-Content Attention (SDGCA), as shown in the left part of Fig. 2, employs structural decoupling. It projects $\mathbf{q}$ into independent subspaces $\mathbf{q}_{geo}$ and $\mathbf{q}_{content}$ via dedicated

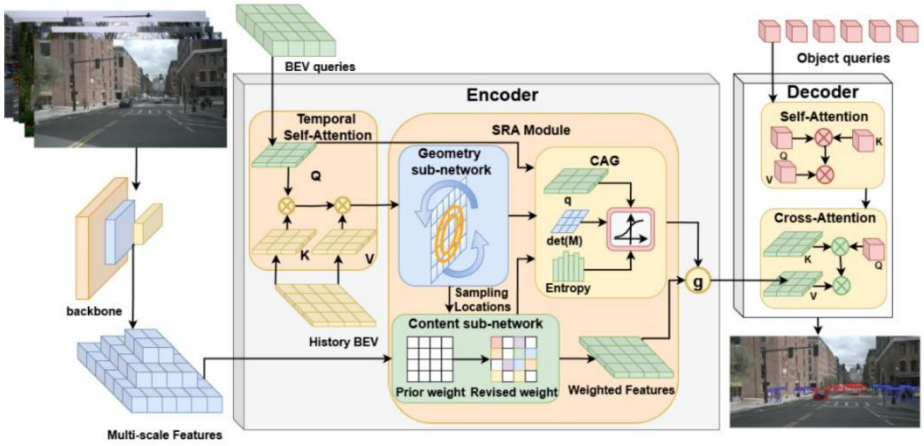


Figure 1: **The overall architecture of SRA-Former.** Adopting a query-based paradigm, the framework utilizes a shared backbone to extract multi-scale features from multi-view inputs. Central to our design is the BEV encoder, where the Self-Regulating Attention (SRA) substitutes standard Spatial Cross-Attention, working in tandem with Temporal Self-Attention to robustly aggregate spatiotemporal features. Within the SRA, the Geometry and Content sub-networks collaboratively generate candidate features that are subjected to dynamic filtration by the Cooperative Assessment Gating (CAG) prior to final integration. The refined BEV queries are finally decoded by the transformer decoder to yield 3D detection predictions.

sub-networks, ensuring that the Geometry Flow and Content Flow traverse along orthogonal manifolds without interference. This effectively extracts task-specific representations to resolve the coupling.

3.2.1. Geometry Sub-network

To address the optimization instability arising from the unconstrained regression of sampling offsets in standard deformable attention, the geometry sub-network takes the query $\mathbf{q}$ to predict optimal sampling points $\mathbf{P}_{\text{final}}$ via a hybrid strategy combining structured priors with residual refinement. Specifically, MLP_{geo} first predicts a set of geometrically meaningful primitives: a rotation angle θ, axial scaling factors s_x, s_y, a shear factor ψ, and a translation

vector $\mathbf{t}$.

$$[\theta, s_x, s_y, \psi, \mathbf{t}, \Delta\mathbf{p}_{\text{residual}}] = \text{MLP}_{\text{geo}}(\mathbf{q}) \tag{1}$$

These primitives are used to construct an affine transformation matrix $\mathbf{M}$:

$$\mathbf{M} = \mathbf{R}(\theta) \cdot \mathbf{S}(s_x, s_y) \cdot \mathbf{H}(\psi) \tag{2}$$

The matrix $\mathbf{M}$ is then applied to a predefined sampling prior $\mathbf{P}_{\text{base}}$ generating a set of Structured Sampling Points $\mathbf{P}_{\text{structured}}$. Specifically, as illustrated in Fig. 2, we employ a base dual-ring template as the initial geometric prior $\mathbf{P}_{\text{base}}$, which allows for simultaneous multi-scale sampling of the object core and periphery:

$$\mathbf{P}_{\text{structured}} = \mathbf{M} \cdot \mathbf{P}_{\text{base}}^{\top} + \mathbf{t} \tag{3}$$

Simultaneously, MLP_{geo} predicts a tanh-constrained residual offset $\Delta\mathbf{p}_{\text{residual}}$, which is added to the structured points for fine-grained local refinement to yield the final sampling points $\mathbf{P}_{\text{final}}$.

$$\mathbf{P}_{\text{final}} = \mathbf{P}_{\text{structured}} + \alpha \cdot \tanh(\Delta\mathbf{p}_{\text{residual}}) \tag{4}$$

This matrix decomposition-based strategy, by imposing constraints on the geometric parameters, such as $s_x, s_y > 0$, structurally helps to ensure the validity of the transformation. It avoids the risk of sampling point collapse or explosion, thereby improving training stability and convergence.

3.2.2. Content Sub-network

Building upon the sampling points $\mathbf{P}_{\text{final}}$, the content sub-network addresses the content-agnostic bottleneck of standard attention by performing feature aggregation via a two-stage strategy. In the first stage, Prior Intent Prediction, the sub-network estimates a set of weights $\mathbf{A}_{\text{prior}}$ solely from $\mathbf{q}$, representing the initial sampling expectation:

$$\mathbf{A}_{\text{prior}} = \text{Linear}_{\text{prior}}(\mathbf{q}). \tag{5}$$

The second stage, Content-Aware Refinement, extracts features $\mathbf{F}_{\text{sampled}}$ through bilinear interpolation at $\mathbf{P}_{\text{final}}$. To balance intent, MLP_{λ} generates

a dynamic factor λ_Q:

$$\lambda_Q = \text{Sigmoid}(\text{MLP}_\lambda(\mathbf{q})) \tag{6}$$

Simultaneously, MLP_{sim} computes matching scores Sim_i between $\mathbf{q}$ and each sampled feature:

$$\text{Sim}_i = \text{MLP}_{\text{sim}}(\mathbf{q}, \mathbf{F}_{\text{sampled},i}) \tag{7}$$

Finally, the prior weights $\mathbf{A}_{\text{prior}}$ and the content-based similarity vector $\mathbf{A}_{\text{sim}}$ (where $\mathbf{A}_{\text{sim}} = [\text{Sim}_1, \ldots, \text{Sim}_N]$) are dynamically balanced by the query-predicted modulation factor λ_Q:

$$\mathbf{A} = \text{Softmax}(\mathbf{A}_{\text{prior}} + \lambda_Q \cdot \mathbf{A}_{\text{sim}}) \tag{8}$$

The weight vector $\mathbf{A}$ (with elements a_i) is then used to perform a weighted sum over $\mathbf{F}_{\text{sampled}}$, producing the feature update vector $\mathbf{u}$:

$$u = \sum_{i=1}^{N} a_i \cdot F_{sampled,i} \tag{9}$$

This design, which combines prior intention with content-aware refinement, resolves the representational coupling of content-agnostic weight prediction in standard deformable attention.

Overall, SDGCA specializes gradient flows, as demonstrated in Equation (10), by decomposing them into two dedicated, deeper computational pathways. Each pathway learns task-specific representations, thereby mitigating the optimization conflicts inherent in shared queries.

$$\frac{\partial L}{\partial q} = \underbrace{\frac{\partial L}{\partial \mathbf{P}_{final}} \cdot \frac{\partial \mathbf{P}_{final}}{\partial q}}_{Geometry} + \underbrace{\frac{\partial L}{\partial \mathbf{A}} \cdot \frac{\partial \mathbf{A}}{\partial q}}_{Content} \tag{10}$$

3.3. Cooperative Assessment Gating (CAG) Framework

SDGCA resolves the sampling instability through structural decoupling. To prevent low-quality updates, we introduce the CAG, as shown in the right part of Fig. 2, which synthesizes reliability signals from both the geometric sampling process and the content aggregation result to dynamically regulate the update intensity, ensuring that only high-confidence features are integrated.

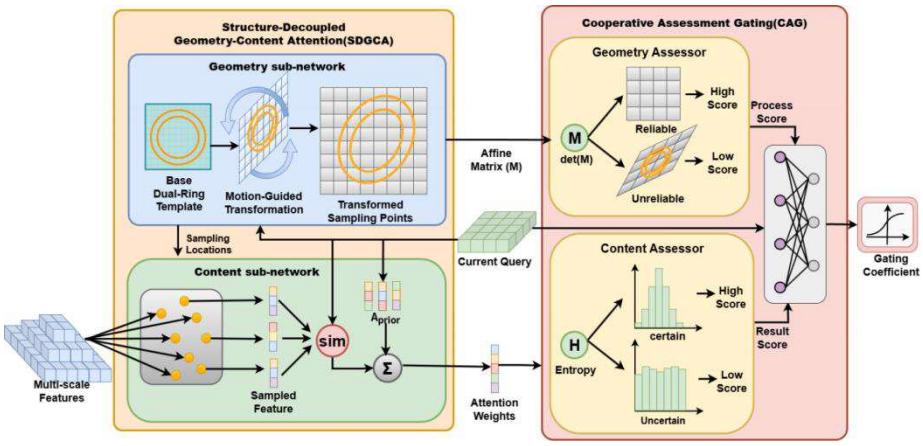


Figure 2: **Detailed structure of the proposed decoupled modules.** (Left) Structurally Decoupled Geometry-Content Attention (SDGCA): This module disentangles the query into two specialized streams. The Geometry sub-network utilizes a base dual-ring template and predicts the Affine Matrix **M** for transformation. The Content Sub-network aggregates features using weights balanced by prior intent and feature similarity. (Right) Cooperative Assessment Gating (CAG): This module evaluates reliability via two pathways. The Geometry Assessor calculates stability using det(**M**), while the Content Assessor measures certainty via Shannon Entropy. The Context-Aware Arbiter finally fuses these signals with the query context.

3.3.1. Geometry Assessor

A highly reliable feature update **u** should stem from a geometrically sound and stable sampling process. To explicitly quantify this geometric soundness, a mechanism is required to evaluate the sampling process prior to the update occurring. Our observations indicate that the key intermediate signal, the affine transformation matrix **M**, generated by the SDGCA geometric subnetwork for constructing sampling points serves as an ideal proxy for this assessment. We therefore introduce the Geometry Assessor, whose core function is to evaluate the rationality of the sampling process using this geometric signal **M**. A high-confidence sampling process corresponds to a geometrically well-posed transformation. Conversely, a highly uncertain sampling process, such as when the model performs exploratory sampling in regions lacking clear geometric cues,

like the sky or distant road surfaces, will primarily manifest as an abnormally large-scale sampling kernel. To extract a concise yet informative metric from **M** that captures this characteristic, we select its determinant det(**M**). Geometrically, the absolute value of the determinant directly quantifies the area or scale covered by the sampling kernel. Thus, an excessively large determinant value corresponds to an overly extensive sampling region, indicating highly uncertain exploratory sampling.

However, using det(**M**) directly presents numerical instability issues, such as values potentially being negative or approaching zero. Therefore, we first compute a stable divergence metric d, defined as:

$$d = \sqrt{|\det(\mathbf{M})| + \epsilon} \tag{11}$$

where $\in$ is a small constant to ensure numerical stability.

While d indicates sampling scale, its relationship with reliability is complex and non-linear, defying simple heuristics. Thus, we pass d to a lightweight $\text{MLP}_{\text{Geometry Assessor}}$ to map this geometric metric onto a normalized score S_{geo}:

$$S_{geo} = \text{Sigmoid}(\text{MLP}_{\text{Geometry Assessor}}(d)) \tag{12}$$

This score S_{geo} represents the quantified assessment of the geometric reasonableness of the current sampling process. It serves as one of the two orthogonal dimensions for evaluating feature reliability and is passed to the subsequent Cooperative Gating Module.

3.3.2. Content Assessor

Beyond geometric stability, the update must also stem from a high-confidence content aggregation. To explicitly quantify this, a mechanism is needed to evaluate the reliability of this aggregation before the update occurs. Our observations indicate that the final attention weights **A** constitute a crucial intermediate signal generated by the SDGCA content subnetwork prior to computing **u**, serving as an ideal proxy for this evaluation. Consequently, we introduce the Content Assessor, whose core function is to assess the quality of aggregated results by

quantifying the distributional properties of the attention weights **A**. Specifically, when the content subnetwork samples effectively target features, the weights a_i become highly concentrated on a few key sampling points, corresponding to low Shannon entropy. This signifies a high-confidence, reliable aggregation result. Conversely, when the sampling region lacks information or features are ambiguous, the attention weights **A** tend towards a uniform distribution, corresponding to high Shannon entropy.

To explicitly quantify this distributional uncertainty, we employ the Shannon Entropy, defined as $H(\mathbf{A})$:

$$H(\mathbf{A}) = -\sum_{i=1}^{N} a_i \log_2(a_i + \epsilon) \tag{13}$$

where N is the total number of sampling points, a_i is the attention weight of the i-th sampling point, and ϵ is a small constant for numerical stability.

To obtain a standardized measure representing confidence, the entropy value $H(\mathbf{A})$ is first normalized by dividing it by its theoretical maximum entropy $\log_2 N$, which corresponds to a perfectly uniform distribution. Subsequently, we convert this measure from an uncertainty score to a confidence score, thereby yielding the final result score S_{con}:

$$S_{con} = 1 - H_{norm} = 1 - \frac{H(\mathbf{A})}{\log_2 N} \tag{14}$$

This S_{con} score quantifies the informational certainty of the aggregated content. It will be sent, along with the S_{geo} score, to the final Context-Aware Arbiter for arbitration.

3.3.3. Context-Aware Arbiter

Geometry Assessor and Content Assessor derive evaluation signals S_{geo} and S_{con} from the geometric process and content aggregation respectively, yet these scores may conflict in specific scenarios. When the SDGCA module correctly samples a large object, such as a nearby lorry, the geometric subnetwork predicts a larger det(**M**) to fully cover the object, potentially yielding a high S_{geo} score.

When sampling large objects, the attention weights **A** tend to distribute uniformly, leading to high Shannon Entropy and a low S_{con} score. To resolve this semantically dependent ambiguity (high S_{geo} vs. low S_{con}), we introduce the Context-Aware Arbiter, employing a lightweight MLP to achieve context-aware arbitration by fusing semantic information from the BEV space. Specifically, the input to this MLP is an enhanced feature vector formed by concatenating the current BEV query **q** with the two scalar scores S_{geo} and S_{con}:

$$\mathbf{X}_{\text{input}} = \text{Concat}(S_{geo}, S_{con}, \mathbf{q}) \tag{15}$$

This concatenation strategy is pivotal to resolving the inherent ambiguity in entropy-based assessment. High entropy signals typically originate from uninformative background dispersion, such as the sky, which warrants suppression. Conversely, similar distributional patterns also characterize the semantic uniformity inherent to large-scale objects, such as buses, which necessitate valid feature updates. To differentiate these conflicting scenarios, the Context-Aware Arbiter explicitly integrates the query context **q** in Eq. (15) to approximate a conditional probability distribution. This formulation enables the network to maintain high gate values even under high entropy conditions by validating the semantic consistency of the target object.

To ensure stability, we initialize the Context-Aware Arbiter output bias to a large negative value, forcing $g \approx 0$ initially to prevent early gradient instability. The generated adaptive coefficient g then modulates the feature update **u** via a gated residual connection:

$$\mathbf{q}' = \mathbf{q} + g \cdot \mathbf{u} \tag{16}$$

4. Experiments

4.1. Experimental Setup

Dataset and Metrics. We conduct experiments on the nuScenes dataset [52]. Focusing on the camera-only perception task, we utilize only surround-view

image data. Following the official protocol, we report NDS, mAP, and five True Positive (TP) metrics (mATE, mASE, mAOE, mAVE, mAAE) [52].

Implementation Details. Following the BEVFormer V1 baseline, we employ a nuImages pre-trained ResNet-101-DCN backbone for all comparisons and ablations. The model is trained for 24 epochs using AdamW [53] with an initial learning rate of 2×10^{-4}, cosine annealing, and a 500-iteration warmup. For generalization analysis, we also adopt the BEVFormer V2 [54] baseline with a ResNet-50 backbone, trained with a learning rate of 4×10^{-4} and a step scheduler.

4.2. State-of-the-Art Comparison

To clearly position the performance of SRA-Former within the domain of vision-only BEV perception, we conducted a comprehensive quantitative comparison with several classic SOTA models on the nuScenes validation set. As shown in Table 1, SRA-Former significantly outperformed the baseline BEVFormer, improving mAP from 0.416 to 0.425, thereby validating the effectiveness of the SRA framework. This success stems directly from the SDGCA module, which, through its dedicated sub-networks, structurally decouples the spatial invariance required for geometric localization from the semantic sensitivity needed for content evaluation. By resolving optimization conflicts inherent in coupled representations, SRA enhances feature discriminability, improving object recognition and false positive suppression. Simultaneously, SRA-Former achieves significant improvements in key localization and motion-related metrics compared to the baseline BEVFormer, further proving the comprehensive advantages of the SRA framework. Without introducing additional geometric priors, the mATE of SRA-Former was reduced from 0.673 to 0.642, mAOE was significantly reduced from 0.372 to 0.341, and mAVE is reduced from 0.394 to 0.375. These reductions stem from precise sampling via SDGCA and noise filtering by CAG, which protect temporal BEV feature integrity to enable reliable estimation of position, orientation, and velocity.

Despite its excellent mAP performance, the NDS score 0.532 of SRA-Former

Table 1: Comparison with State-of-the-Art Methods on the nuScenes Validation Set. Metrics are reported as NDS ↑, mAP ↑, and TP Errors ↓.

Model	Backbone	Image Resolution	NDS ↑	mAP ↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
FCOS3D [25]	R101-DCN	1600×900	0.415	0.343	0.725	0.263	0.422	1.292	0.153
DETR3D [55]	R101-DCN	1600×900	0.434	0.349	0.716	0.268	0.379	0.842	0.200
BEVDet [16]	R101	1054×384	0.396	0.330	0.702	0.272	0.534	0.932	0.251
PETR [26]	R101-DCN	1600×900	0.442	0.370	0.711	0.267	0.383	0.865	0.201
BEVDepth [20]	R101-DCN	1408×512	0.538	0.418	0.565	0.266	0.358	0.331	0.190
PolarFormer [56]	R101-DCN	1600×900	0.470	0.415	0.657	0.263	0.405	0.911	**0.139**
Fast-BEV [57]	R101	1600×900	0.535	0.413	0.584	0.279	0.311	0.329	0.206
BEVHeight [58]	R101	1408×512	**0.554**	0.423	**0.541**	0.262	**0.307**	**0.277**	0.187
BEVFormer [10]	R101-DCN	1600×900	0.517	0.416	0.673	0.274	0.372	0.394	0.198
Ours	R101-DCN	1600×900	0.532	**0.425**	0.642	**0.265**	0.341	0.375	0.187

is lower than that of BEVHeight, which highlights the fundamental difference between our proposed method and existing paths [20]. Methods like BEVHeight and BEVDepth impose strong geometric constraints on the 3D localization problem by introducing explicit depth or height priors during the view transformation stage, thereby achieving a significant advantage in localization-related errors [58, 59]. In contrast, SRA-Former systematically enhances detection accuracy and positioning robustness by optimizing feature extraction stability and temporal fusion reliability, without imposing additional geometric constraints, thereby significantly reducing positioning errors at baseline.

To further validate our proposed SRA framework, we selected the temporal model BEVformerV2 as a baseline with a historical frame of 2 [54]. By directly replacing its original spatial cross-attention mechanism with SRA, the results shown in Table 2 demonstrate a significant improvement in performance. This robustly confirms the generalizability and effectiveness of the SRA framework.

Table 2: Generalization Study of the SRA Module.

Model	Backbone	Image Res.	NDS	mAP
BEVFormerV2-t2	R-50	1600×900	0.518	0.42
BEVFormerV2-t2+SRA	R-50	1600×900	0.535	0.428

4.3. Ablation Study

We conducted an ablation study to systematically validate the effectiveness of each component within the SRA-Former framework.

4.3.1. Incremental Ablation

As shown in Table 3, we first conducted an incremental ablation study on the two core components of the SRA framework: SDGCA and CAG. Replacing the standard attention of the baseline BEVFormer with SDGCA yields a significant improvement of +0.008 NDS and +0.005 mAP, representing the largest single gain within the framework. This confirms that SDGCA mitigates the representational coupling inherent in standard attention, establishing a robust foundation for subsequent gating modules. This result provides compelling evidence for the existence of representational coupling within standard attention mechanisms. SRA-Former mitigates conflict through SDGCA, thereby furnishing a higher-quality foundation for feature extraction in subsequent gating modules.

Table 3: Incremental Ablation Study.

Model	**NDS** ↑	**mAP** ↑
BEVFormer	0.517	0.416
BEVFormer + **SDGCA**	0.525	0.421
BEVFormer + **SDGCA** + **Geometry Assessor**	0.526	0.422
BEVFormer + **SDGCA** + **Content Assessor**	0.527	0.422
Ours	0.532	0.425

Building on SDGCA, integrating the Geometry Assessor alone raises the NDS to 0.526, indicating that Geometry Assessor brings stable and positive performance improvements by suppressing geometrically unstable sampling proposals. Subsequently, integrating the Content Assessor alone on top of SDGCA improves the NDS to 0.527. Compared to the Geometry Assessor, the Content Assessor yields a more significant performance gain, indicating that evaluating the information certainty of the content aggregation result is more critical

than evaluating the geometric rationality of the sampling process. In sparse or occluded scenarios, the attention module is prone to generating high-entropy, uninformative updates. By measuring attention uncertainty via Shannon entropy, the Content Assessor identifies and suppresses these invalid features, thereby safeguarding the integrity of temporal BEV features. The complete SRA-Former achieves 0.532 NDS and 0.425 mAP, with an overall gain significantly exceeding the sum of independent gains from Geometry Assessor and Content Assessor. This non-linear leap demonstrates that Context-Aware Arbiter learns non-linear fusion functions to effectively arbitrate conflicting signals generated by Geometry Assessor and Content Assessor in complex scenarios.

4.3.2. Analysis of Base Vector Selection

Having validated the overall contribution of SDGCA, we further investigated the design of basis vector priors within its geometric subnetwork. As shown in Table 4, we compared several distinct prior structures. Results indicate that the rigid Cartesian grid exhibits the poorest performance, as its view-agnostic structure struggles to effectively match the radial distribution of objects in self-centred panoramic perception. Introducing a single-ring structure with radial sampling yields only marginal improvement, suggesting that while circular structures are superior to Cartesian grids, a single fixed scale is insufficient for capturing complex geometric features. The log-polar structure demonstrates markedly superior performance, underscoring the necessity of multi-scale sampling. Specifically, its dense near-field and sparse far-field patterns align with perspective projection, enabling fine local sampling while extending the receptive field [60, 61, 62]. Ultimately, the dual-ring architecture achieves the best NDS of 0.525. Surpassing simple radial approaches, the inner ring enables the geometry sub-network to learn to focus on the high-response core regions of an object, while the outer ring permits the simultaneous exploration of peripheral contextual features.

In summary, the improvement from Cartesian to log-polar validates multi-scale sampling, while the further gain with dual-ring structures demonstrates

that composite, learnable priors with higher degrees of freedom surpass single patterns. By optimizing two distinct radii simultaneously, the dual-ring structure enables more flexible, data-driven spatial queries.

Table 4: Ablation Study on Base Vector Prior Structures.

Base Vector Structure	**NDS ↑**	**mAP ↑**
Cartesian Grid	0.517	0.416
Single-Ring	0.519	0.416
Log-Polar Grid	0.523	0.418
Dual-Ring	**0.525**	**0.421**

4.3.3. Analysis of Gating Component Design Choices

For the Geometry Assessor, when exploring the geometric properties of the affine transformation **M**, we compared two approaches: the sampling kernel aspect ratio and the versus matrix determinant. As shown in Table 5, employing the determinant yields markedly superior performance, indicating that sampling failures arising from model uncertainty primarily manifest as anomalous variations in the scale of sampled regions rather than excessive stretching of sampled shapes.

For the Content Assessor, we aim to evaluate the distributional uncertainty of the attention weights **A**. We compared two classic metrics from information theory: Gini Impurity and Shannon entropy. As shown in Table 5, Shannon Entropy outperforms Gini Impurity. Mathematically, while Gini is sensitive to peaks, Shannon entropy captures global distributional uncertainty. Since invalid updates manifest as high-entropy distributions rather than low peaks, Shannon entropy proves more effective for identifying such uncertainty.

For the Context-Aware Arbiter, we aim to build an arbitrator that can effectively fuse the S_{geo} and S_{con} signals. A basic solution is a *context-free* MLP, which only learns a non-linear combination of the two scores. To further test our core hypothesis about the importance of context, we designed a context-

aware model, $MLP(q, S_{geo}, S_{con})$, which additionally receives the BEV query **q** as input. As shown in the table, introducing context results in a substantial performance gap of +0.006 NDS. This validates that context-free modules fail to resolve semantic conflicts, confirming that the semantic information carried by BEV query is critical for high-fidelity arbitration.

In summary, the ablation experiments profoundly elucidate the reasons behind the success of SRA-Former: its reliability assessment relies on measurement scales rather than distortion, considers global uncertainty rather than peak probability, and integrates semantic context rather than isolated signals.

Table 5: Ablation Study on Gating Component.

Module	**Core Design**	**NDS ↑**	**mAP ↑**
Geometry Assessor	Condition Number	0.523	0.420
	Determinant	0.526	0.422
Content Assessor	Gini Impurity	0.523	0.421
	Shannon Entropy	0.527	0.422
Context-Aware Arbiter	MLP(S_{geo}, S_{con})	0.526	0.423
	MLP(S_{geo}, S_{con}, **q**)	0.532	0.425

4.4. Qualitative Analysis

To more intuitively demonstrate the performance advantages of SRA-Former shown in the SOTA comparisons and ablation studies, this section provides visual comparison results in challenging scenarios.

4.4.1. Visual Detection Results

As visualized in Fig. 3, we present detection results in complex real-world environments to demonstrate model robustness. While the baseline model fails to detect targets in these extreme conditions, SRA-Former successfully localizes them. Specifically, for distant objects, SRA-Former leverages the structured

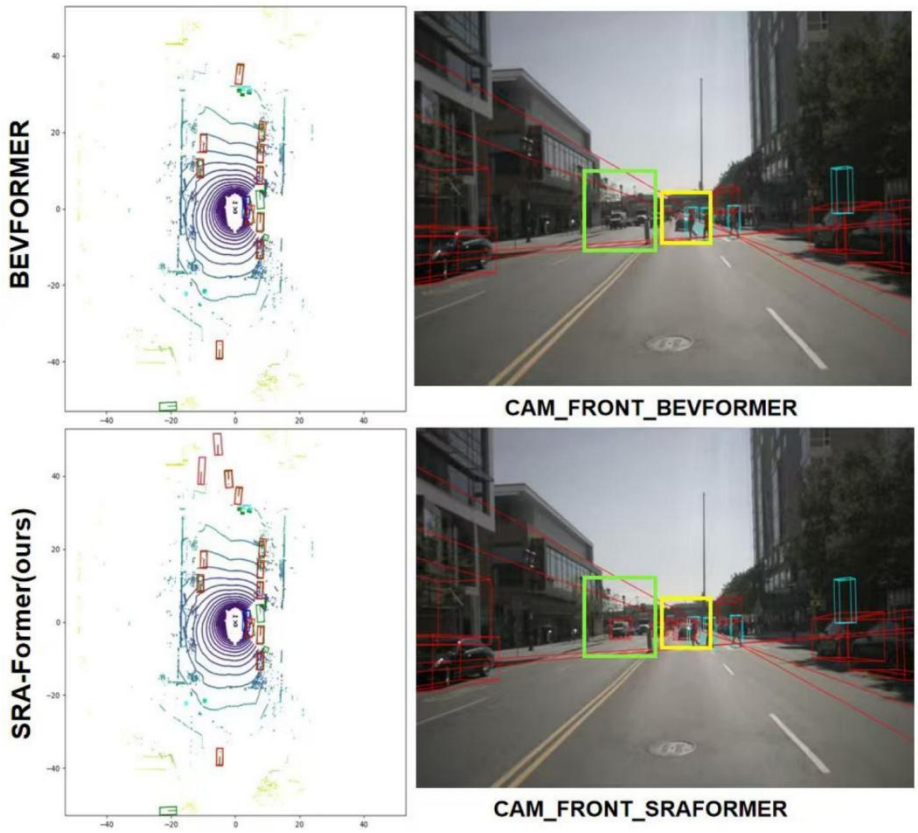


Figure 3: **Visualization in Challenging Scenarios.** The top row shows results from the baseline BEVFormer, while the bottom row shows results from SRA-Former (Ours). Green boxes highlight distant small objects missed by the baseline but detected by ours due to the multi-scale sampling prior of SDGCA. Yellow boxes indicate occluded objects where the baseline fails, whereas SRA-Former successfully recalls them by resolving ambiguous signals via the context-aware gating.

sampling of SDGCA to reliably aggregate sparse features despite background noise. Furthermore, in occlusion scenarios, SRA-Former framework demonstrates superior robustness: the Context-Aware Arbiter leverages semantic context to resolve conflicts between high geometric stability and occlusion-induced semantic ambiguity, thereby integrating critical feature updates to successfully recall the occluded vehicle.

4.4.2. CAG Decision Heatmap

To validate the interpretability of the CAG framework, we visualize the decision surface of the Context-Aware Arbiter in Fig. 4. When the context is background, Context-Aware Arbiter exhibits global suppression with $g \approx 0$, preventing noise contamination regardless of S_{geo} and S_{con}. When the context is vehicles, Context-Aware Arbiter displays a complex non-linear surface where, in the lower-left corner, a low S_{geo} indicates geometrically unreliable samples, while a low S_{con} signifies aggregations with uncertain content. Corresponding to invalid samples from occluded regions, Context-Aware Arbiter outputs $g \approx 0$ to prevent invalid updates from contaminating temporal BEV features. In the bottom-right corner, a high S_{geo} implies reasonable geometry, while a low S_{con} indicates content uncertainty. This confirms that while a context-free module would erroneously suppress updates for large nearby objects due to low S_{con}, Context-Aware Arbiter outputs $g \approx 1$, demonstrating that the semantic context in $\mathbf{q}$ is indispensable for intelligent arbitration. The contrast between vehicle and background decision surfaces underscores the critical role of the semantic context embedded within the query $\mathbf{q}$ in resolving ambiguities associated with large nearby objects.

5. Discussion

The SOTA comparisons and ablation studies of SRA-Former collectively demonstrate that optimizing the internal attention structure and temporal fusion paradigm significantly reduces baseline localization errors without additional geometric priors. Although SRA-Former achieves advancements in mAP and localization robustness, we must acknowledge its limitations.

First, regarding efficiency, SDGCA and CAG indeed introduce additional structures. However, these lightweight operations exhibit linear query complexity, incurring negligible overhead relative to the backbone to efficiently trade marginal FLOPs for significant robustness. Second, there is the gap in the NDS between SRA-Former and methods such as BEVHeight. Methods like

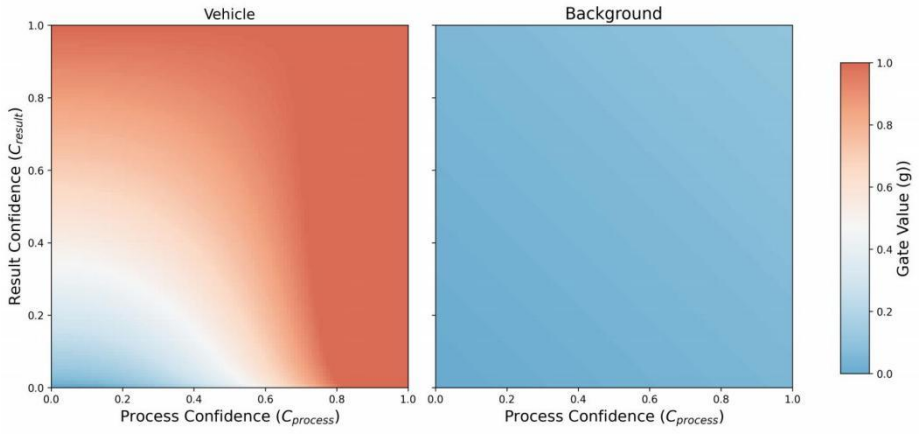


Figure 4: Context-Aware Arbiter Decision Surface. In background contexts Context-Aware Arbiter enforces global suppression whereas in vehicle contexts it learns a non-linear boundary. It suppresses invalid samples in the lower-left region while retaining updates for large objects with conflicting signals in the bottom-right region to demonstrate intelligent context-aware arbitration.

BEVHeight constrain 2D-to-3D projection via external depth priors, naturally minimizing localization errors like mATE. In contrast, SRA-Former prioritizes mechanism optimization to enhance internal stability and fusion robustness, yielding superior mAP. While NDS sensitivity to mATE favors geometric methods, this confirms that the two paradigms are orthogonal and complementary.

Based on above analysis, future work can be pursued in several directions. Given the complementarity of mechanism optimization and geometric priors, integrating SRA-Former with lightweight priors such as utilizing depth uncertainty as a Context-Aware Arbiter input creates a hybrid model combining geometric constraints with robust fusion. Additionally, knowledge distillation techniques could transfer the gating logic of SRA-Former into a lightweight network for efficient deployment. Finally, future work could expand CAG into an intermodal framework to dynamically arbitrate sensor weights, such as suppressing LiDAR in rain or cameras in glare, achieving robust all-weather fusion.

6. Conclusion

This paper introduces SRA-Former, the proposed self-regulating attention framework, to address the optimization conflict in existing BEV perception methods caused by a single shared query vector simultaneously encoding spatial invariance and semantic sensitivity, as well as the problem of unconditionally fusing low-quality features. Experiments on the nuScenes dataset not only validate the significant advantages of SRA-Former over the baseline in key metrics such as mAP and mATE, but also demonstrate the effectiveness of its two core components: the structural decoupling of SDGCA and the reliability gating mechanism of CAG. SRA-Former establishes a reliability-aware decision-making paradigm over unconditional fusion, ensuring the robustness required for safety-critical applications. The structural decoupling and gated evaluation principles proposed possess independent academic merit and offer potential for integration with strong geometric prior models.

References

[1] H. Li, C. Sima, J. Dai, W. Wang, L. Lu, H. Wang, J. Zeng, Z. Li, J. Yang, H. Deng, et al., Delving into the devils of bird's-eye-view perception: A review, evaluation and recipe, IEEE Transactions on Pattern Analysis and Machine Intelligence 46 (4) (2023) 2151–2170.

[2] R. Qian, X. Lai, X. Li, 3D object detection for autonomous driving: A survey, Pattern Recognition 130 (2022) 108796.

[3] Y. Gong, N. Wang, J. Lu, X. Zhang, Y. Gao, J. Zhao, Z. Huang, H. Bai, N. Zeng, N. Su, et al., Progressive Bird's Eye View perception for Safety-Critical autonomous driving: A comprehensive survey, arXiv preprint arXiv:2508.07560.

[4] Q.-V. Lai-Dang, A survey of Vision Transformers in Autonomous Driving: Current Trends and Future Directions, arXiv preprint arXiv:2403.07542.

[5] J. Philion, S. Fidler, Lift, Splat, Shoot: Encoding Images From Arbitrary Camera Rigs by Implicitly Unprojecting to 3D, in: European Conference on Computer Vision (ECCV), 2020, pp. 194–210.

[6] S. Xie, L. Kong, W. Zhang, J. Ren, L. Pan, K. Chen, Z. Liu, Benchmarking and Improving Bird's Eye View Perception Robustness in Autonomous Driving, IEEE Transactions on Pattern Analysis and Machine Intelligence.

[7] Y. Zhang, J. Lu, J. Zhou, Objects are different: Flexible monocular 3D object detection, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3289–3298.

[8] Y. Ma, T. Wang, X. Bai, H. Yang, Y. Hou, Y. Wang, Y. Qiao, R. Yang, X. Zhu, Vision-centric BEV perception: A survey, IEEE Transactions on Pattern Analysis and Machine Intelligence.

[9] K. Wang, T. Zhou, X. Li, F. Ren, Performance and challenges of 3D object detection methods in complex scenes for autonomous driving, IEEE Transactions on Intelligent Vehicles 8 (2) (2022) 1699–1716.

[10] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Q. Yu, J. Dai, BEVFormer: Learning Bird's-Eye-View Representation from Multi-Camera Images via Spatiotemporal Transformers, in: European Conference on Computer Vision (ECCV), 2022, pp. 1–18.

[11] S. Liu, F. Li, H. Zhang, X. Yang, X. Qi, H. Su, J. Zhu, L. Zhang, DAB-DETR: Dynamic anchor boxes are better queries for detr, in: International Conference on Learning Representations (ICLR), 2022.

[12] X. Zhang, K. Bi, S. Chan, S. Lu, X. Zhou, SyNet: A synergistic network for 3D object detection through Geometric-Semantic-Based Multi-Interaction fusion, IEEE Transactions on Multimedia 27 (2025) 4950–4960.

[13] X. Zhao, X. Zhang, G. Chen, X. Wei, Z. Li, Z. Xu, Y. Wang, J. Zhu, P. Zhai, L. Zhang, Denoising Transformer for BEV 3D Object Detection

via Multiview Multiscale Cross-Attention, IEEE Transactions on Intelligent Transportation Systems.

[14] E. Xie, Z. Yu, D. Zhou, J. Philion, A. Anandkumar, S. Fidler, P. Luo, J. M. Alvarez, M^2BEV: Multi-camera joint 3D detection and segmentation with unified birds-eye view representation, arXiv preprint arXiv:2204.05088.

[15] Y. Lee, H.-M. Jeong, Y. Jeon, S. Kim, EVT: Efficient View Transformation for Multi-Modal 3D Object Detection, in: IEEE/CVF International Conference on Computer Vision (ICCV), 2025, pp. 26632–26642.

[16] J. Huang, G. Huang, Z. Zhu, D. Du, BEVDet: High-performance Multi-Camera 3D object detection in Bird-Eye-View, arXiv preprint arXiv:2112.11790.

[17] Y. Li, H. Bao, Z. Ge, J. Yang, J. Sun, Z. Li, BEVStereo: Enhancing depth estimation in Multi-View 3D object detection with temporal stereo, in: AAAI Conference on Artificial Intelligence (AAAI), Vol. 37, 2023, pp. 1486–1494.

[18] A. W. Harley, Z. Fang, J. Li, R. Ambrus, K. Fragkiadaki, Simple-BEV: What really matters for Multi-Sensor BEV perception? (2023) 2759–2765.

[19] N. Durasov, R. Mahmood, J. Choi, M. T. Law, J. Lucas, P. Fua, J. M. Alvarez, Uncertainty estimation for 3D object detection via evidential learning, arXiv preprint arXiv:2410.23910.

[20] Y. Li, Z. Ge, G. Yu, J. Yang, Z. Wang, Y. Shi, J. Sun, Z. Li, BEVDepth: Acquisition of reliable depth for Multi-View 3D object detection, in: AAAI conference on artificial intelligence, Vol. 37, 2023, pp. 1477–1485.

[21] S.-W. Lu, Y.-H. Tsai, Y.-T. Chen, Toward Real-world BEV Perception: Depth Uncertainty Estimation via Gaussian Splatting, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025, pp. 17124–17133.

[22] Z. Li, Z. Yu, W. Wang, A. Anandkumar, T. Lu, J. M. Alvarez, FB-BEV: BEV representation from forward-backward view transformations, in: IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6919–6928.

[23] Z. Li, S. Lan, J. M. Alvarez, Z. Wu, BEVNext: Reviving dense BEV frameworks for 3D object detection, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20113–20123.

[24] T. Yin, X. Zhou, P. Krahenbuhl, Center-based 3D object detection and tracking, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11784–11793.

[25] T. Wang, X. Zhu, J. Pang, D. Lin, FCOS3D: Fully convolutional one-stage monocular 3D object detection, in: IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 913–922.

[26] Y. Liu, T. Wang, X. Zhang, J. Sun, PETR: Position embedding transformation for Multi-View 3D object detection, in: European Conference on Computer Vision (ECCV), Springer, 2022, pp. 531–548.

[27] Y. Zhao, J. Du, Q. Yang, N. Cai, Z. Peng, H. Zhan, H. Cheng, PETR-Fusion: Multi-Sensor Fusion Based BEV Semantic Segmentation Network for Autonomous Driving, in: 2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC), IEEE, 2024, pp. 2966–2971.

[28] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, J. Dai, Deformable DETR: Deformable transformers for End-to-End object detection, in: International Conference on Learning Representations (ICLR), 2021.

[29] X. Lin, Z. Pei, T. Lin, L. Huang, Z. Su, Sparse4D v3: Advancing end-to-end 3D detection and tracking, arXiv preprint arXiv:2311.11722.

[30] H. Liu, Y. Teng, T. Lu, H. Wang, L. Wang, SparseBEV: High-performance sparse 3D object detection from Multi-camera videos, in: IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18580–18590.

[31] Z. Xia, X. Pan, S. Song, L. E. Li, G. Huang, Vision Transformer with Deformable Attention, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4794–4803.

[32] N. Fang, L. Qiu, S. Zhang, Z. Wang, K. Hu, K. Wang, A Cross-Scale hierarchical transformer with Correspondence-Augmented Attention for inferring Bird's-Eye-View semantic segmentation, IEEE Transactions on Intelligent Transportation Systems 25 (7) (2024) 7726–7737.

[33] C. Witte, J. Behley, C. Stachniss, M. Raaijmakers, Epipolar attention field transformers for bird's eye view semantic segmentation, in: IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), IEEE, 2025, pp. 8660–8669.

[34] D. Meng, X. Chen, Z. Fan, G. Zeng, H. Li, Y. Yuan, L. Sun, J. Wang, Conditional DETR for fast training convergence, in: IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3651–3660.

[35] M. Zhang, G. Song, Y. Liu, H. Li, Decoupled DETR: Spatially disentangling localization and classification for improved end-to-end object detection, in: IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6601–6610.

[36] D. Wu, F. Yang, B. Xu, P. Liao, W. Zhao, D. Zhang, HV-BEV: Decoupling horizontal and vertical feature sampling for Multi-View 3D object detection, IEEE Transactions on Intelligent Transportation Systems 26 (11) (2025) 18734–18746.

[37] Y. Dai, J. Sun, S. E. Li, Q. Xu, J. Wang, L. He, K. Li, Hierarchical and decoupled BEV perception learning framework for autonomous driving, arXiv preprint arXiv:2407.12491.

[38] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. M. Ni, H.-Y. Shum, DINO: DETR with improved denoising anchor boxes for end-to-end ob-

ject detection, in: International Conference on Learning Representations (ICLR), 2023.

[39] F. Pu, Y. Wang, J. Deng, W. Yang, MonoDGP: Monocular 3D object detection with decoupled-query and geometry-error priors, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025, pp. 6520–6530.

[40] Y. Cui, Z. Zou, X. Ye, X. Tan, Z. Li, Z. Fang, Coupling and decoupling: Towards temporal feedback for 3D object detection, IEEE Transactions on Multimedia 27 (2025) 7230–7242.

[41] M. Zhang, Y. Zhang, J. Sun, K.-L. Yung, L. Yang, Geometry-guided transformer for monocular 3D object detection, Advanced Intelligent Systems (2025) 2500003.

[42] S. Zeng, W. Zheng, J. Lu, H. Yan, Hardness-Aware scene synthesis for semi-supervised 3D object detection, IEEE Transactions on Multimedia 26 (2024) 9644–9656.

[43] S. Chen, Y. Ma, Y. Qiao, Y. Wang, M-BEV: Masked BEV perception for robust autonomous driving, in: AAAI Conference on Artificial Intelligence, Vol. 38, 2024, pp. 1183–1191.

[44] S. Kumar, H. Truong, S. Sharma, G. Sistu, T. Scanlan, E. Grua, C. Eising, Minimizing occlusion effect on Multi-View camera perception in BEV with Multi-Sensor Fusion, arXiv preprint arXiv:2501.05997.

[45] C. Han, J. Yang, J. Sun, Z. Ge, R. Dong, H. Zhou, W. Mao, Y. Peng, X. Zhang, Exploring recurrent long-term temporal fusion for Multi-View 3D perception, IEEE Robotics and Automation Letters 9 (7) (2024) 6544–6551.

[46] M. Chang, X. Zhang, R. Zhang, Z. Zhao, G. He, S. Liu, RecurrentBEV: A long-term temporal fusion framework for Multi-View 3D detection, in:

European Conference on Computer Vision (ECCV), Springer, 2024, pp. 131–147.

[47] P. An, Y. Duan, Y. Huang, J. Ma, Y. Chen, L. Wang, Y. Yang, Q. Liu, SP-Det: Leveraging saliency prediction for Voxel-Based 3D object detection in sparse point cloud, IEEE Transactions on Multimedia 26 (2023) 2795–2808.

[48] W. Wu, X. Deng, P. Jiang, S. Wan, Y. Guo, Crossfuser: Multi-modal feature fusion for end-to-end autonomous driving under unseen weather conditions, IEEE Transactions on Intelligent Transportation Systems 24 (12) (2023) 14378–14392.

[49] J. Mun, H. Choi, Uncertainty prediction for monocular 3D object detection, Sensors 23 (12) (2023) 5395.

[50] J. Hu, L. Shen, G. Sun, Squeeze-and-Excitation networks, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7132–7141.

[51] S. Woo, J. Park, J.-Y. Lee, I. S. Kweon, CBAM: Convolutional Block Attention Module, in: European Conference on Computer Vision (ECCV), 2018, pp. 3–19.

[52] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, O. Beijbom, nuScenes: A multimodal dataset for autonomous driving, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11621–11631.

[53] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: International Conference on Learning Representations (ICLR), 2019.

[54] 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: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17830–17839.

[55] Y. Wang, V. C. Guizilini, T. Zhang, Y. Wang, H. Zhao, J. Solomon, DETR3D: 3D object detection from Multi-view images via 3D-to-2D queries, in: Conference on Robot Learning (CoRL), PMLR, 2022, pp. 180–191.

[56] Y. Jiang, L. Zhang, Z. Miao, X. Zhu, J. Gao, W. Hu, Y.-G. Jiang, PolarFormer: Multi-camera 3D object detection with polar transformer, in: AAAI Conference on Artificial Intelligence, Vol. 37, 2023, pp. 1042–1050.

[57] Y. Li, B. Huang, Z. Chen, Y. Cui, F. Liang, M. Shen, F. Liu, E. Xie, L. Sheng, W. Ouyang, et al., Fast-BEV: A fast and strong bird's-eye view perception baseline, IEEE Transactions on Pattern Analysis and Machine Intelligence 46 (12) (2024) 8665–8679.

[58] L. Yang, K. Yu, T. Tang, J. Li, K. Yuan, L. Wang, X. Zhang, P. Chen, BEVHeight: A robust framework for vision-based roadside 3D object detection, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21611–21620.

[59] L. Yang, T. Tang, J. Li, K. Yuan, K. Wu, P. Chen, L. Wang, Y. Huang, L. Li, X. Zhang, et al., BEVHeight++: Toward robust visual centric 3D object detection, IEEE Transactions on Pattern Analysis and Machine Intelligence.

[60] C. Wu, J. Li, J. Cao, M. Li, Y. Feng, J. Xu, S. Wu, Z. Gao, S. Du, Y. Li, Omnidirectional depth-aided occupancy prediction based on cylindrical voxel for autonomous driving, arXiv preprint arXiv:2504.01023.

[61] M. Dang, G. Liu, H. Li, D. Wang, R. Pan, Q. Wang, PRA-Det: Anchor-Free oriented object detection with Polar Radius Representation, IEEE Transactions on Multimedia 27 (2025) 145–157.

[62] Z. Yu, Q. Liu, W. Wang, L. Zhang, X. Zhao, PolarBEVDet: Exploring polar representation for Multi-View 3D object detection in bird's-eye-view, arXiv preprint arXiv:2408.16200.