OnlineBEV: Recurrent Temporal Fusion in Bird's Eye View Representations for Multi-Camera 3D Perception

Junho Koh, Youngwoo Lee, Jungho Kim, Dongyoung Lee, and Jun Won Choi, *Member, IEEE*

***Abstract*— Multi-view camera-based 3D perception can be conducted using bird's eye view (BEV) features obtained through perspective view-to-BEV transformations. Several studies have shown that the performance of these 3D perception methods can be further enhanced by combining sequential BEV features obtained from multiple camera frames. However, even after compensating for the ego-motion of an autonomous agent, the performance gain from temporal aggregation is limited when combining a large number of image frames. This limitation arises due to dynamic changes in BEV features over time caused by object motion. In this paper, we introduce a novel temporal 3D perception method called OnlineBEV, which combines BEV features over time using a recurrent structure. This structure increases the effective number of combined features with minimal memory usage. However, it is critical to spatially align the features over time to maintain strong performance. OnlineBEV employs the Motion-guided BEV Fusion Network (MBFNet) to achieve temporal feature alignment. MBFNet extracts motion features from consecutive BEV frames and dynamically aligns historical BEV features with current ones using these motion features. To enforce temporal feature alignment explicitly, we use Temporal Consistency Learning Loss, which captures discrepancies between historical and target BEV features. Experiments conducted on the nuScenes benchmark demonstrate that OnlineBEV achieves significant performance gains over the current best method, SOLOFusion. OnlineBEV achieves 63.9% NDS on the nuScenes test set, recording state-of-the-art performance in the camera-only 3D object detection task.**

***Index Terms*— 3D perception, camera-based perception, BEV perception, 3D object detection, recurrent temporal fusion.**

Received 4 September 2024; revised 5 May 2025; accepted 4 July 2025. Date of publication 28 August 2025; date of current version 3 November 2025. This work was supported in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) Grant funded by Korea Government through Ministry of Science and ICT (MSIT) [Artificial Intelligence Graduate School Program (Seoul National University)] under Grant RS-2021-II211343 and in part by the National Research Foundation (NRF) funded by Korean Government through MSIT under Grant RS-2024-00421129. The Associate Editor for this article was K. Wang. *(Junho Koh and Youngwoo Lee contributed equally to this work.) (Corresponding author: Jun Won Choi.)*

Junho Koh is with the Autonomous Driving Development Center, Hyundai Motors, Seongnam-si 13529, Republic of Korea (e-mail: junhkoh@hyundai.com).

Youngwoo Lee is with the Department of Electrical Engineering, Hanyang University, Seoul 04763, Republic of Korea (e-mail: youngwoolee@spa.hanyang.ac.kr).

Jungho Kim is with the Interdisciplinary Program in Artificial Intelligence, Seoul National University, Seoul 08826, Republic of Korea (e-mail: jhkim@spa.snu.ac.kr).

Dongyoung Lee and Jun Won Choi are with the Department of Electrical and Computer Engineering, Seoul National University, Seoul 08826, Republic of Korea (e-mail: dylee@spa.snu.ac.kr; junwchoi@snu.ac.kr).

Digital Object Identifier 10.1109/TITS.2025.3589152

I. Introduction

3D PERCEPTION plays a crucial role in autonomous driving and robot navigation by gathering comprehensive information about the surrounding 3D environment through sensor data. It encompasses various tasks such as 3D object detection, bird's eye view (BEV) segmentation, and 3D occupancy grid prediction.

Multi-view cameras covering a 360-degree perspective have enabled the identification of a 3D environment around the ego-vehicle. Recent research efforts have been dedicated to a task of transforming multiple 2D images into a unified 3D representation for 3D perception. Existing architectures for generating 3D representations from multi-view camera images can be categorized into two strategies: dense BEV-based methods [1], [2], [3], [4], [5] and sparse query-based methods [6], [7], [8]. BEV-based methods utilize the lift-Splat-Shoot (LSS) mechanism [9] to convert features extracted from 2D images into unified representations within the BEV domain. On the other hand, the query-based methods leverage attention mechanisms to decode object query features by utilizing multi-view 2D features.

Baseline models for 3D perception process single-frame images captured by multi-view cameras at each time step. This approach may underperform when the current frame experiences occlusion or motion blur. To compensate for this degradation, it is beneficial to use multiple consecutive frames for 3D perception through a technique known as temporal fusion. In this technique, features from historical frames are utilized to enhance robustness in perception tasks.

There are two main strategies for temporal fusion: (1) parallel temporal fusion and (2) recurrent temporal fusion, as illustrated in Fig. 1 (a) and Fig. 1 (b). In parallel temporal fusion, information from the most recent K frames is aggregated simultaneously, requiring the memory bank that stores features from all K frames. Furthermore, the computational complexity of this approach increases with larger K, which limits the number of frames that can be combined. In contrast, recurrent temporal fusion maintains and updates a single feature that encapsulates historical information in a recurrent manner. This design supports long-term temporal fusion while keeping the computational complexity low.

Parallel temporal fusion has been employed in 3D perception frameworks such as [2], [10], [11], [12], [13], and

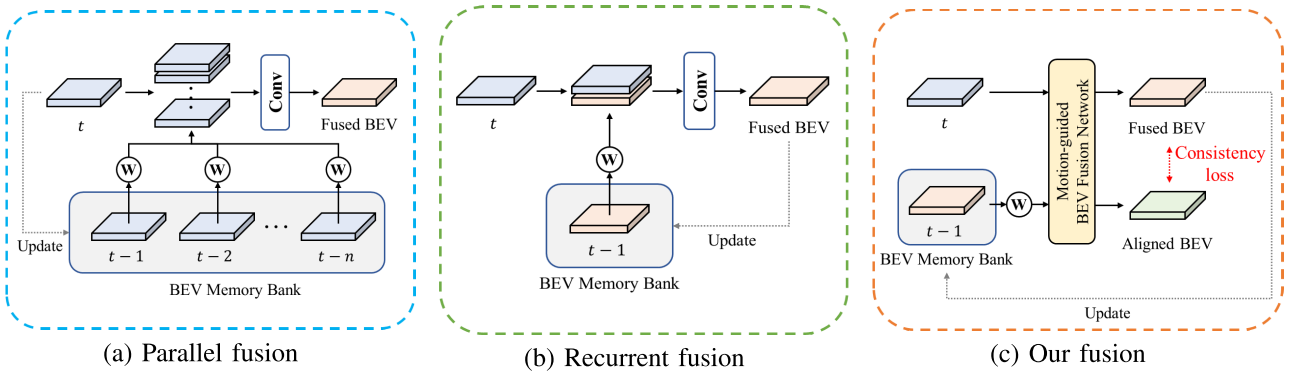


Fig. 1. Different temporal fusion strategies. (a) Parallel temporal fusion aggregates historical BEV features within a fixed-length window at each time step. (b) Recurrent temporal fusion progressively updates historical BEV features over time. (c) Our OnlineBEV approach aligns historical BEV features with current BEV features to enable effective recurrent temporal fusion.

[14]. Notably, SOLOFusion achieved significant computational complexity reduction by storing previous BEV features in memory rather than the raw sensor data. Recurrent temporal fusion has also been adopted in StreamPETR [15], which integrates features from historical frames using a recurrent architecture. While StreamPETR achieved high computational efficiency through its sparse, object-centric design, it lacks a unified BEV representation. This limitation constrains its applicability to broader 3D scene understanding tasks—such as BEV segmentation and 3D occupancy prediction—that require dense, spatially aligned features. In this paper, we aim to design recurrent temporal fusion architecture tailored for BEV representations.

One of the key challenges in temporal fusion lies in handling significant spatial changes within dynamic scenes, which can cause substantial feature misalignments across frames. If not properly addressed during feature aggregation, these misalignments may propagate over time, leading to a noticeable degradation in performance. Moreover, spatial changes induced by object motion are often more pronounced in the BEV domain than in the perspective view. Therefore, it is essential to both align and appropriately weight features from different frames to fully leverage the advantages of temporal fusion.

In this paper, we propose a novel recurrent temporal fusion framework, OnlineBEV, for multi-camera 3D perception. The objective of OnlineBEV is to effectively aggregate BEV features across frames using a recurrent design, as illustrated in Fig. 1 (c). While enabling long-term temporal fusion, OnlineBEV maintains high performance by compensating for spatial misalignment over time through the Motion-Guided BEV Fusion Network (MBFNet).

OnlineBEV introduces two key innovations. First, it utilizes MBFNet to align BEV feature maps across time using a spatio-temporal deformable attention mechanism. This alignment is guided by two core modules: the Motion Feature Extractor (MFE) and the Motion-Guided BEV Warping Attention (MGWA). MFE captures spatial changes between adjacent BEV features and encodes them into motion features. MGWA then leverages deformable attention [16] to spatially align the features based on these motion cues. Second, we introduce a novel temporal consistency learning strategy that explicitly guides the alignment of BEV features. By enforcing consistency between the heatmap representations of historical and current BEV feature maps, the proposed consistency loss significantly improves the temporal alignment capability of MBFNet. Experimental results confirm that this approach enhances both the accuracy and robustness of feature fusion over time.

The proposed OnlineBEV framework is evaluated on the public nuScenes benchmark [17]. Our evaluation shows that OnlineBEV achieves significant performance gains over baseline models across all tasks, including 3D object detection, BEV segmentation, and 3D occupancy prediction. Additionally, OnlineBEV outperforms other existing methods in the official nuScenes 3D object detection benchmark.

The key contributions of our work are summarized as follows:

- We propose OnlineBEV, a novel temporal fusion framework that efficiently aligns and aggregates sequential BEV features for multi-view 3D perception.
- OnlineBEV leverages the advantages of the recurrent fusion strategy while achieving strong performance through a robust feature alignment mechanism enabled by the proposed MBFNet.
- We present MBFNet that spatially aligns BEV features by utilizing motion features capturing the spatial changes between previous and current BEV feature maps.
- We further introduce a temporal consistency learning strategy that minimizes the difference between historical and current BEV features, thereby enhancing alignment quality and temporal coherence.

II. Related Work

A. Multi-View 3D Perception

Multi-view 3D perception has rapidly advanced by transforming feature maps from multiple viewpoints into a unified 3D representation, enabling more efficient processing of multi-camera inputs compared to single-view perception [18], [19], [20]. Based on their transformation strategies, multi-view 3D perception frameworks can be broadly categorized into dense BEV-based methods [1], [2], [3], [4], [5] and sparse query-based methods [6], [7].

Dense BEV-based approaches adopt the LSS paradigm [9] to project 2D perspective image feature maps into a unified BEV representation. In contrast, sparse query-based methods [6], [7], [8] employ learnable object queries and 3D position-aware features to directly aggregate multi-view image features using attention mechanisms. Despite their effectiveness, these frameworks are limited by their reliance on single-frame inputs, lacking temporal context from image sequences.

Recently, both paradigms have been extended to incorporate temporal modeling. BEVFormer [21] pioneered temporal modeling for multi-view 3D object detection by introducing a temporal attention mechanism. Dense BEV-based methods have since achieved significant performance gains by fusing adjacent BEV feature maps across time. For instance, BEVStereo [11] improved depth estimation by applying multi-view stereo (MVS) [22] to consecutive keyframes, while SOLOFusion [13] enhanced long-term temporal fusion by storing historical BEV features in a memory bank and integrating them with current features through lightweight processing.

Sparse query-based approaches have similarly evolved to exploit temporal information. DETR4D [23] and Sparse4D [24] introduced temporal attention to enable interactions between past and current object queries. StreamPETR [15] propagated long-term historical object proposals as queries, encoding ego-motion and surrounding object dynamics for temporal reasoning. SparseBEV [14] defined object queries in BEV space and utilized adaptive spatio-temporal sampling and mixing to interact with multi-view image features for 3D object prediction.

B. Consistency Learning

Consistency learning has been widely explored as a strategy to enhance feature representations by leveraging both labeled and unlabeled datasets. For example, CSD [25] introduced a consistency loss by enforcing constraints between foreground proposals predicted from the original input image and its horizontally flipped counterpart. Similarly, CCT [26] proposed a robust and efficient consistency learning approach that exploits the invariance of predictions to various perturbations at the feature level.

Recently, consistency learning has been extended to videos to exploit temporal correspondences across frames. To stabilize predictions for video inputs, the method in [27] utilized temporal consistency by incorporating optical flow between consecutive frames. Furthermore, consistency learning has also been applied to exploit temporal information without relying exclusively on unsupervised datasets. MaskFreeVIS [28], for instance, applied a temporal consistency loss between one-to-many matched patches across consecutive frames, improving instance segmentation by capturing complex object shapes through temporal information.

Building on these advances, we propose a novel temporal consistency learning strategy designed specifically for feature alignment. This approach guides the alignment of features from the previous to the current time step, thereby enhancing 3D perception performance.

III. OnlineBEV

The overall architecture of the proposed OnlineBEV is illustrated in Fig. 2. An image backbone network with shared weights processes multi-camera images at time t to produce multi-view feature maps. To generate the BEV feature F_t, we adopt the LSS method [9], which efficiently transforms perspective features into the BEV space using depth predictions. Notably, our framework can also incorporate alternative BEV generation methods, such as inverse perspective mapping (IPM) or BEVFormer [21]. OnlineBEV maintains historical features H_{t-1} in memory, which are updated in a recurrent manner. To conduct temporal fusion, we devise a target query $q_t^{(l)}$ and a historical query $q_{t-1}^{(l)}$, where l denotes the transformer layer. The target query is initialized with F_t while the historical query is initialized with H_{t-1}.

In each transformer layer, MGWA aligns the historical query $q_{t-1}^{(l)}$ with the target query $q_t^{(l)}$, guided by the motion features produced by MFE. To generate these motion features, MFE leverages both the target and historical queries. The aligned historical query $\hat{q}_{t-1}^{(l)}$ is then updated as the historical query $q_{t-1}^{(l+1)}$ for the subsequent layer. Simultaneously, the target query $q_t^{(l+1)}$ for the next layer is obtained by fusing $\hat{q}_{t-1}^{(l)}$ with $q_t^{(l)}$. Through this iterative process, the historical query progressively converges toward the target query, ultimately producing enhanced BEV features.

After passing through L transformer layers, the queries are transformed into the final target query $q_t^{(L)}$ and the final historical query $q_{t-1}^{(L)}$. We let $H_t = q_t^{(L)}$ and $\hat{H}_t = q_{t-1}^{(L)}$, where $\hat{H}_t$ represents the features temporally aligned with H_t. To further enhance alignment, we introduce the Heatmap-based Temporal Consistency Loss (HTC-loss), which encourages greater similarity between the heatmaps derived from H_t and $\hat{H}_t$. Finally, a detection head is applied to H_t to produce the final output and H_t is stored in the memory bank for use in the next time step.

A. Motion-Guided BEV Fusion Network

Fig. 3 presents the structure of the MBFNet. MBFNet consists of MFE and MGWA. MGWA spatially aligns the historical BEV features with the current BEV features using motion context information produced by MFE. Then, MGWA aggregates the aligned features with the current BEV features to yield the enhanced BEV representation.

1) Motion Feature Extractor: MFE performs differential encoding between the historical query $q_{t-1}^{(l)}$ and the target query $q_t^{(l)}$ to extract motion features. It generates a motion map by computing the difference between $q_{t-1}^{(l)}$ and $q_t^{(l)}$. Static objects are expected to exhibit minimal changes across frames, whereas dynamic objects produce significant feature differences over time. By encoding these variations, the network effectively captures object motion within the BEV space. The resulting difference is further processed through fully connected (FC) layers followed by Channel-Wise Attention (CWA) [29], yielding the motion context feature $M_t^{(l)}$, i.e.,

$$M_t^{(l)} = \mathrm{CWA}(\mathrm{FC}(q_t^{(l)} - q_{t-1}^{(l)})). \tag{1}$$

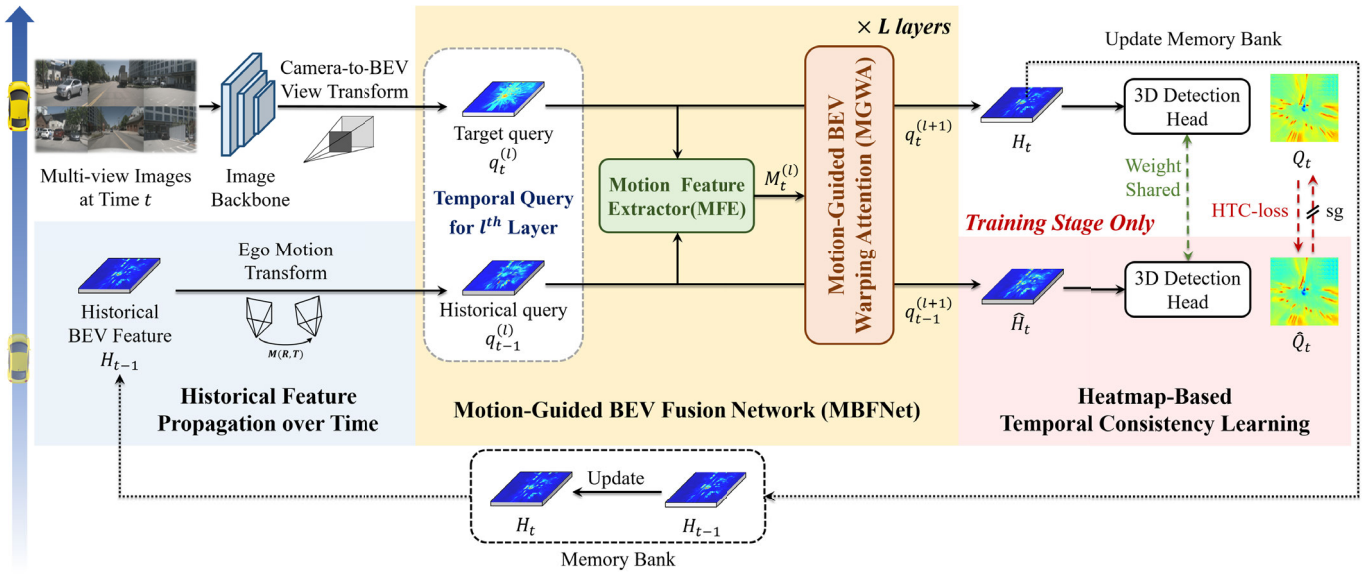


Fig. 2. The overall architecture of OnlineBEV. OnlineBEV aggregates historical BEV features with current BEV features using a recurrent structure. Before aggregation, MGWA aligns the historical BEV features to the current ones, guided by the motion features produced by MFE. During training, HTC-loss further facilitates the feature alignment process. Here, 'sg' denotes stop-gradient.

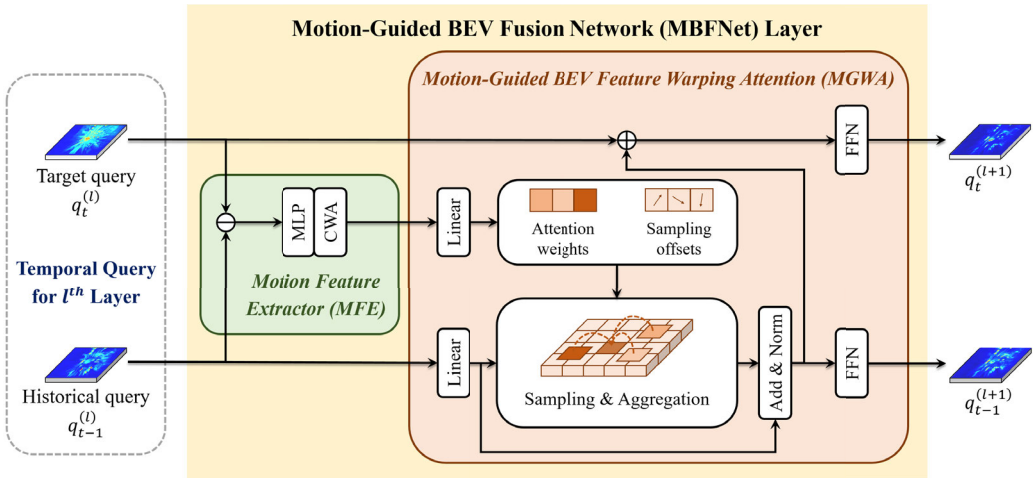


Fig. 3. Structure of Motion-Guided BEV Fusion Network (MBFNet). MBFNet comprises MFE and MGWA, where MFE generates motion features and MGWA applies deformable attention for feature alignment and fusion.

2) Motion-Guided BEV Warping Attention: Using the motion features $M_t^{(l)}$ extracted by MFE, MGWA aligns the historical query $q_{t-1}^{(l)}$ with the target query $q_t^{(l)}$. The feature alignment is achieved by a deformable attention mechanism originally proposed in [16]. The motion features $M_t^{(l)}$ is used to determine the offsets and attention weights of a mask, which is applied to $q_{t-1}^{(l)}$. At the reference point p, the aligned features $Z_{t-1}^{(l)}(p)$ are given by

$$Z_{t-1}^{(l)}(p) = \text{DeformAttn}(M_{t-1}^{(l)}(p),\, p,\, q_{t-1}^{(l)}), \tag{2}$$

where $\text{DeformAttn}(Q, p, V)$ is Deformable Cross Attention module [16], and Q and V denote query and value, respectively. The historical query $q_{t-1}^{(l+1)}$ for the next iteration is finally updated as

$$q_{t-1}^{(l+1)} = \text{FFN}_1(\hat{q}_{t-1}^{(l)}), \tag{3}$$

$$\hat{q}_{t-1}^{(l)} = \text{LN}(\text{Dropout}(Z_{t-1}^{(l)}) + q_{t-1}^{(l)}), \tag{4}$$

where $\text{FFN}_1(\cdot)$ is the feed-forward network [16], $\text{Dropout}(\cdot)$ is the Dropout operation [30], and $\text{LN}(\cdot)$ is the Layer Normalization [31]. The target query $q_t^{(l+1)}$ for the next iteration is also updated by combining the aligned historical query $\hat{q}_{t-1}^{(l)}$ with the target query $q_t^{(l)}$, i.e.,

$$q_t^{(l+1)} = \text{FFN}_2(q_t^{(l)} \oplus \hat{q}_{t-1}^{(l)}), \tag{5}$$

where $\oplus$ denotes channel-wise concatenation and $\text{FFN}_2(\cdot)$ is the feed-forward network composed of one fully-connected layer with ReLU activation.

After L transformer layers, the final aggregated BEV features $q_t^{(L)}$ and the aligned historical BEV features $q_{t-1}^{(L)}$ become H_t and $\hat{H}_t$, respectively. The final object detection results are obtained by applying the 3D detection head to H_t [32]. Note that H_t is returned to the memory queue for BEV detection in the subsequent time step.

TABLE I

PERFORMANCE COMPARISON ON nuScenes *Valid* SET

Methods	Backbone	Input Size	# Frames	mAP ↑	NDS ↑	mATE ↓	mASE ↓	mAOE ↓	mAVE ↓	mAAE ↓
PETRv2 [34]	ResNet50	320 × 800	2	35.0	45.6	0.726	0.277	0.505	0.503	0.181
BEVDepth [2]	ResNet50	256 × 704	2	33.3	44.1	0.683	0.276	0.545	0.526	0.226
BEVStereo [11]	ResNet50	256 × 704	2	34.4	44.9	0.659	0.276	0.579	0.503	0.216
AEDet [3]	ResNet50	256 × 704	2	35.9	47.3	0.649	0.277	0.496	0.432	0.216
FB-BEV [4] †	ResNet50	256 × 704	2	37.8	49.8	0.620	0.273	0.444	0.374	0.200
P2D [12]	ResNet50	256 × 704	4	37.4	48.6	0.631	0.272	0.508	0.384	0.212
SOLOFusion [13]	ResNet50	256 × 704	17	42.7	53.4	**0.567**	0.274	0.511	0.252	0.188
StreamPETR [15]	ResNet50	256 × 704	*rnt*	43.2	54.0	0.581	**0.272**	0.413	0.295	0.195
SparseBEV [14]	ResNet50	256 × 704	8	43.2	54.5	0.606	0.274	**0.387**	0.251	**0.186**
OnlineBEV	ResNet50	256 × 704	*rnt*	**44.4**	**54.5**	0.590	0.275	0.466	**0.244**	0.197
StreamPETR [15] †	ResNet50	256 × 704	*rnt*	45.0	55.0	0.613	**0.267**	0.413	0.265	0.196
SparseBEV [14] †	ResNet50	256 × 704	8	44.8	55.8	0.581	0.271	**0.373**	0.247	**0.190**
OnlineBEV †	ResNet50	256 × 704	*rnt*	**46.3**	**56.0**	**0.552**	0.271	0.450	**0.245**	0.197
SOLOFusion [13]	ResNet101	512 × 1408	17	48.3	58.2	0.503	0.264	0.381	0.246	0.207
StreamPETR [15] †	ResNet101	512 × 1408	*rnt*	50.4	59.2	0.569	0.262	**0.315**	0.257	0.199
SparseBEV [14] †	ResNet101	512 × 1408	8	50.1	59.2	0.562	0.265	0.321	0.243	**0.195**
OnlineBEV †	ResNet101	512 × 1408	*rnt*	**50.8**	**59.9**	**0.508**	**0.262**	0.353	**0.223**	0.205

TABLE II

PERFORMANCE COMPARISON ON nuScenes *Test* SET

Methods	Backbone	Input Size	# Frames	mAP ↑	NDS ↑	mATE ↓	mASE ↓	mAOE ↓	mAVE ↓	mAAE ↓
BEVDepth [2]	V2-99	640 × 1600	2	50.3	60.0	0.445	0.245	0.378	0.320	0.126
BEVStereo [11]	V2-99	640 × 1600	2	52.5	61.0	0.431	0.246	0.358	0.357	0.137
AEDet [3]	ConvNeXt-B	640 × 1600	2	53.1	62.0	0.439	0.247	0.344	0.292	0.130
FB-BEV [4]	V2-99	640 × 1600	2	53.7	62.4	0.439	0.250	0.358	0.270	0.128
SOLOFusion [13]	ConvNeXt-B	640 × 1600	17	54.0	61.9	0.453	0.257	0.376	0.276	0.148
StreamPETR [15]	V2-99	640 × 1600	*rnt*	55.0	63.6	0.479	**0.239**	**0.317**	0.241	0.119
SparseBEV [14]	V2-99	640 × 1600	8	55.6	63.6	0.485	0.244	0.332	0.246	**0.117**
OnlineBEV	V2-99	640 × 1600	*rnt*	**55.8**	**63.9**	**0.424**	0.253	0.362	**0.233**	0.133

TABLE III

PERFORMANCE COMPARISON FOR THE OTHER 3D PERCEPTION TASKS ON THE nuScenes *Valid* SET

3D Perception Task	*Method*	# Frames	Performance	
			mIOU-seg ↑	mIOU-occ ↑
BEV Segmentation	BEVDet-Depth [1]	1	43.6	-
	BEVDepth [2]	2	44.7	-
	SOLOFusion [13]	17	49.5	-
	OnlineBEV	*rnt*	52.4	-
3D Occupancy Prediction	BEVDet-Depth [1]	1	-	40.8
	BEVDepth [2]	2	-	42.6
	SOLOFusion [13]	17	-	44.7
	OnlineBEV	*rnt*	-	45.9

B. Heatmap-Based Temporal Consistency Loss

HTC-loss provides explicit supervision for feature alignment during training by enforcing consistency between the current BEV feature H_t and the temporally aligned historical BEV feature $\hat{H}_t$. This promotes stability and coherence in BEV representations across consecutive frames, enabling more effective long-term temporal fusion.

To compute HTC-loss, two prediction heads with shared weights [32] are applied to H_t and $\hat{H}_t$, producing heatmaps Q_t and $\hat{Q}_t$, respectively. The HTC-loss is then defined as

$$\mathcal{L}_{\text{cons}} = \| Q_t - \hat{Q}_t \|_2^2. \tag{6}$$

Here, gradient flow is blocked for the branch generating Q_t to ensure that $\hat{H}_t$ is aligned to the target BEV features H_t.

C. Overall Training Loss Function

The proposed OnlineBEV is an end-to-end trainable with a total loss

$$\mathcal{L} = \omega_{cls}\mathcal{L}_{cls} + \omega_{reg}\mathcal{L}_{reg} + \omega_{cons}\mathcal{L}_{cons}, \tag{7}$$

where w_{cls}, w_{reg}, and w_{cons} are the regularization parameters, and $\mathcal{L}_{cls}$, $\mathcal{L}_{reg}$, and $\mathcal{L}_{cons}$ are the focal loss [33], the L1 loss for 3D box regression, and the HTC-loss, respectively. We set w_{cls} and w_{reg} to 1 and 0.25, respectively, following the BEVDepth [2]. Meanwhile, w_{cons} is set to 2 based on our experiments. More details of training procedures are provided in Section IV-B.

IV. EXPERIMENTS

A. Datasets and Performance Metrics

The nuScenes dataset [17] is a challenging large-scale autonomous driving dataset comprising 1,000 scenes, each with a duration of about 20 seconds. These 1000 driving scenes are split into 700 scenes for training (*train*), 150 validation (*val*), and 150 for testing (*test*). Six cameras provide perspective-view images, covering the entire 360° field of view (FOV). Moreover, 3D bounding boxes from 10 categories are annotated at 2Hz. The proposed method was evaluated in regarding mean average precision (mAP) and nuScenes detection score (NDS), which are official 3D object detection benchmark metrics of nuScenes. mAP is computed based on the 2D center distance between the ground truth data and the predictions at the BEV. NDS is a weighted sum of mAP and 5 kinds of true positive (TP) metrics including average

TABLE IV
ABLATION STUDY TO EVALUATE THE CONTRIBUTIONS OF THE INDIVIDUAL COMPONENTS

Method	# Frames	*BEV Temporal Fusion Strategy*				Performance			GFLOPs	FPS
		Parallel Temporal Fusion	Recurrent Temporal Fusion	Motion-Guided BEV Fusion	Heatmap-based Temporal Consistency	mAP ↑	NDS ↑	mAVE ↓		
Baseline-S	1					35.4	41.9	0.758	192.2	13.8
Baseline-M	2	✓				37.0	46.9	0.430	194.2	13.7
	5	✓				38.9	49.2	0.310	194.8	13.6
	17	✓				40.8	50.7	0.287	198.7	13.4
	2	✓		✓		37.9	47.3	0.401	205.7	12.6
	5	✓		✓		40.9	50.1	0.301	249.0	7.7
Method (a)	*rnt*		✓			40.8	50.4	0.296	192.6	13.8
Method (b)	*rnt*		✓	✓		42.1	51.4	0.287	205.7	12.6
Method (c)	*rnt*		✓	✓	✓	42.5	51.9	0.280	205.7	12.6

TABLE V
ABLATION STUDY FOR EVALUATING THE EFFECT OF MOTION GUIDANCE OF MBFNET

MFE Block	Performance			
	mAP ↑	NDS ↑	mATE ↓	mAVE ↓
Baseline	41.4	50.9	0.614	0.295
w diff.	41.7	51.1	0.613	0.289
w diff. + CWA	42.1	51.4	0.611	0.287

TABLE VI
ABLATION STUDY CONDUCTED ON AGOVERSE 2 DATASET

Method	mAP ↑	CDS ↑	mATE ↓	mASE ↓	mAOE ↓
Baseline-S	13.6	8.2	0.967	0.452	0.854
Baseline-M (17)	14.7	10.4	0.938	0.431	0.782
OnlineBEV	15.6	12.3	0.899	0.411	0.746

TABLE VII
PERFORMANCE COMPARISON ON THE CORRUPTED nuScenes TEST SET

Test Input	SOLOFusion		OnlineBEV	
	mAP	NDS	mAP	NDS
Original	42.7	53.4	44.4	54.5
Motion Blur	32.2 (↓ 10.5)	46.9 (↓ 6.5)	39.6 (↓ 4.8)	51.5 (↓ 3.0)
Occlusion	29.7 (↓ 13.0)	42.3 (↓ 11.1)	35.6 (↓ 8.8)	48.8 (↓ 5.7)

translation error (mATE), average scale error (mASE), average orientation error (mAOE), average velocity error (mAVE), and average attribute error (mAAE).

The Argoverse 2 dataset [35] comprises 1,000 driving scenes, each lasting 15 seconds, with annotations provided at 10 Hz. The dataset is divided into 700 scenes for training, 150 for validation, and 150 for testing. It features seven high-resolution ring cameras providing a complete 360° field of view. Annotations cover 26 object categories within a sensing range of 150 meters. We evaluate our method using the official Argoverse 2 3D detection metrics. In addition to mAP, the primary evaluation metric is the Composite Detection Score (CDS), defined as a weighted sum of mAP and three TP metrics: mATE, mASE, and mAOE.

B. Implementation Details

We conducted experiments on the nuScenes dataset. We employed ResNet50, ResNet101 [36], and V2-99 [37] as image backbones. When we employed ResNet50, the input resolution of multi-view images was set to 256×704, and the size of BEV feature was set to 128×128. For larger backbones such as ResNet101 and V2-99, the input image sizes were set to 512×1408 and 640×1600, respectively, with the BEV resolution configured to 256×256.

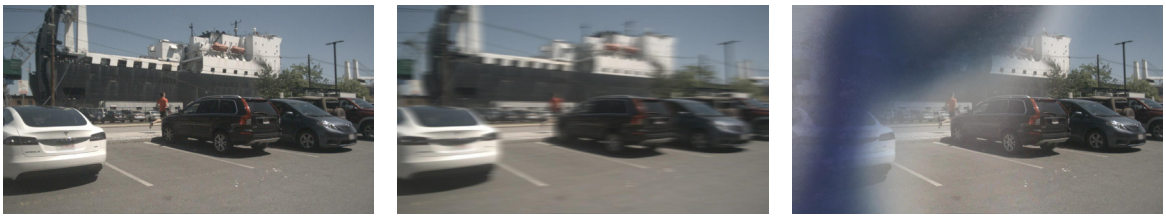
OnlineBEV was trained in two phases. In the first phase, a single-frame model without MBFNet and HTC-loss was trained for 6 epochs. In the second phase, the full OnlineBEV model was trained with the recurrent structure for a total of 90 epochs using the AdamW optimizer [38] with a learning rate of 2e-4. We used three transformer layers with a dropout rate of 0.1 during training, and dropout was disabled during inference. A batch size of 32 was used with the ResNet50 backbone on 4 NVIDIA RTX 3090 GPUs. For larger backbones (ResNet101 and V2-99), the batch size was reduced to 16, and training was conducted on 8 NVIDIA Tesla V100 GPUs. We adopted data augmentation used in [2].

To train the recurrent structure, we utilized a sequential dataloader that preserves the temporal order of frames within each sequence. Frames were processed sequentially, and all historical frames were ego-motion compensated prior to temporal fusion. To avoid feature contamination across scenes, the memory was reset to zero at the start of each new scene. OnlineBEV and other competitive methods were trained and evaluated without access to future frames.

When we trained the OnlineBEV on the Argoverse 2 dataset, we used the V2-99 backbone with an input resolution of 640×960 and a BEV resolution of 256×256. We used the same optimization settings and training strategies as used in the nuScenes experiments. The model was trained for 6 epochs with a batch size of 8 on 8 NVIDIA Tesla V100 GPUs.

C. Performance Comparison

1) Performance on nuScenes Valid Set: Table I summarizes the performance of *OnlineBEV* compared to existing multi-view 3D object detectors on the nuScenes *validation* set. Here, # Frames denotes the number of frames used during training, † indicates methods leveraging perspective-view pre-training [39], and *rnt* denotes a recurrent structure. *OnlineBEV* achieves substantial performance gains over other temporal fusion approaches. With a ResNet-50 backbone pretrained on ImageNet-1k [40], *OnlineBEV* delivers a 1.7% improvement



(a) (b) (c)

Fig. 4. Examples of corrupted input images on nuScenes dataset. (a), (b), and (c) show the original input, motion blur, and occlusion, respectively.

TABLE VIII
COMPUTATIONAL COMPLEXITY AND MEMORY REQUIREMENTS

Method		mAP	NDS	Memory (GB)	Inference Time (ms)	GFLOPs	Params (M)
Query-based Method	SparseBEV	43.2	54.5	4.1	43.2	192.0	44.34
	StreamPETR	43.2	54.0	2.3	37.5	145.7	37.21
BEV-based Method	SOLOFusion	42.7	53.7	3.9	72.8	198.7	64.99
	OnlineBEV	44.4	54.5	3.4	79.3	205.7	65.03

in mAP and a 1.1% increase in NDS compared to SOLOFusion [13], which uses 16 historical frames. In contrast, *OnlineBEV* maintains only a single historical feature map. When ResNet-50 is pretrained on nuImages [17], *OnlineBEV* achieves state-of-the-art results with 46.3% mAP and 56.0% NDS. Furthermore, with a ResNet-101 backbone and an input resolution of 512×1408, *OnlineBEV* surpasses the previous best, StreamPETR [15], by 0.4% in mAP and 0.7% in NDS.

2) Performance on nuScenes Test Set: Table II provides the performance of OnlineBEV evaluated on the nuScenes test set. Note that ConvNeXt-B [41] backbone is pretrained on ImageNet-22K [40], and V2-99 [37] backbone is initialized with weights from DD3D [39]. Using the V2-99 [37] backbone, OnlineBEV outperforms other temporal fusion methods. The performance of OnlineBEV surpasses that of the latest state-of-the-art method, SparseBEV [14], by 0.2% in mAP and 0.3% in NDS. Additionally, our OnlineBEV outperforms SOLOFusion [13] configured with 16 historical frames, by 1.8% in mAP and 2.0% in NDS.

3) Performance on Other 3D Perception Tasks: Table III demonstrates the effectiveness of *OnlineBEV* on additional 3D perception tasks, including BEV segmentation and 3D occupancy prediction. Although numerous sophisticated methods exist, we compare *OnlineBEV* with representative temporal fusion approaches, BEVDepth [2] and SOLOFusion [13]. For the BEV segmentation task, *OnlineBEV* achieves a 7.7% mean Intersection-over-Union (mIoU) improvement over BEVDepth and a 2.9% gain over SOLOFusion. In the 3D occupancy grid prediction task, *OnlineBEV* outperforms BEVDepth by 3.3% mIoU and SOLOFusion by 1.2%.

D. Ablation Studies

We performed a series of ablation studies on the nuScenes *validation* set to evaluate the contributions of each submodule. Two baseline models were considered: Baseline-S, which processes a single frame without temporal fusion, and Baseline-M, which fuses 2, 5, and 17 frames using a parallel temporal fusion strategy. For clarity, we denote these configurations as Baseline-M (2), Baseline-M (5), and Baseline-M (17). Furthermore, we applied MGWA to Baseline-M (2) and Baseline-M (5). However, applying MGWA to Baseline-M (17) was not feasible due to memory limitations.

1) Contributions of Main Ideas: Table IV highlights the contributions of three key concepts: 1) recurrent temporal fusion, 2) motion-guided BEV fusion, and 3) heatmap-based consistency loss to overall performance. First, Method (a) is obtained by applying recurrent temporal fusion to Baseline-S. This offers an 8.5% NDS gain over Baseline-S. Adding motion-guided BEV fusion to Method (a) results in Method (b), which further improves the NDS by 1%. Finally, enabling heatmap-based consistency loss in Method (c) leads to an additional 0.5% performance gain. Altogether, these three approaches yield a total NDS improvement of 10%. Notice that the complexity overhead incurred by our method over Baseline-S is only $(205.7 - 192.2)/205.7 = 6.56\%$. Since Heatmap-based Temporal Consistency operates only during the training phase, it does not increase the complexity in run-time.

We also compare our method with the parallel fusion approach, **Baseline-M**. The performance of parallel fusion becomes comparable to that of recurrent fusion only when 17 frames are aggregated. In all other cases, recurrent fusion significantly outperforms parallel fusion.

2) Effect of Motion-Guidance in MBFNet: Table V presents the impact of incorporating motion features for BEV feature alignment on overall performance. The baseline model is derived by disabling MFE in Method (b) of Table III. Here, "w diff." refers to differential encoding without channel-wise attention, while "w diff. + CWA" indicates the use of both differential encoding and channel-wise attention. Incorporating differential encoding alone improves mAP by 0.3% and NDS

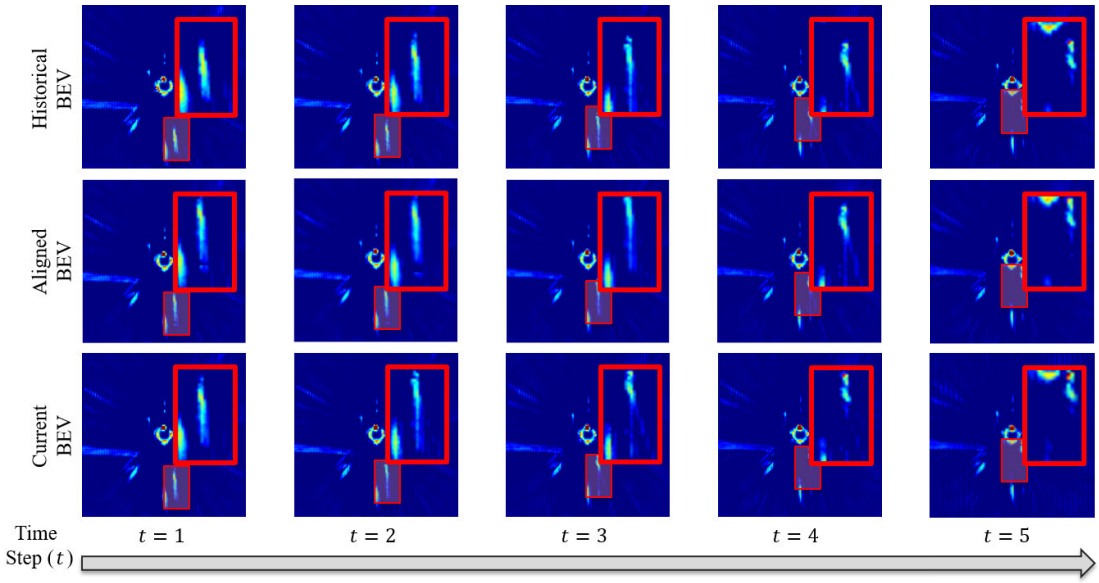


Fig. 5. Visualization of BEV features over time. Historical, aligned, and current BEV features are visualized over five consecutive time steps ($t = 1$ to $t = 5$) to show the temporal evolution and the effectiveness of the alignment process.

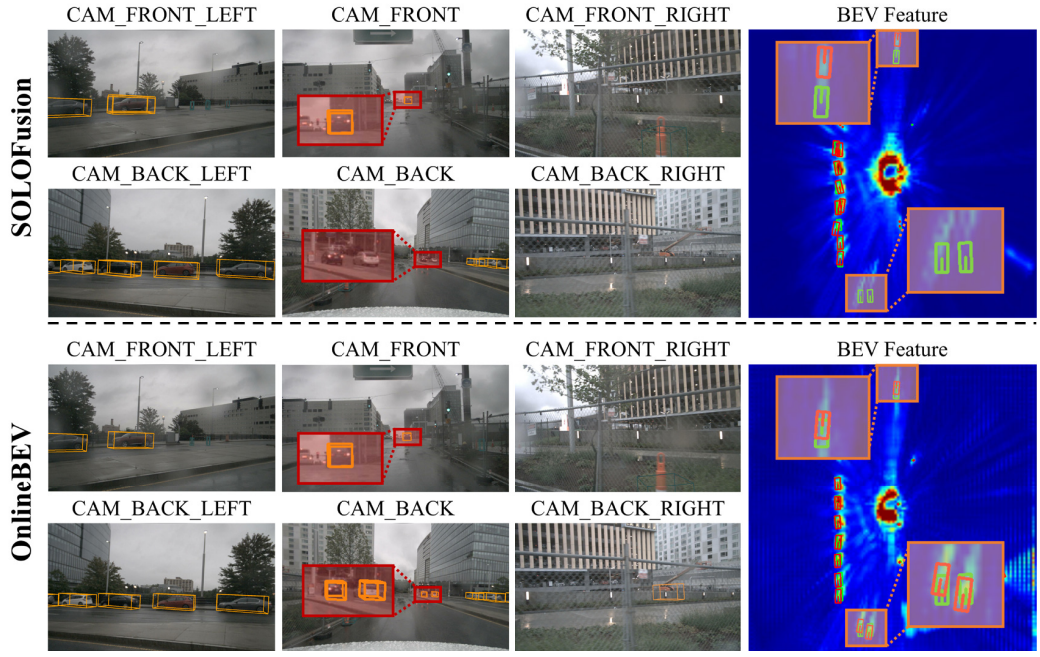


Fig. 6. Qualitative results of SOLOFusion and OnlineBEV. Predicted 3D boxes are shown in red, while ground-truth 3D boxes are shown in green.

by 0.2% over the baseline, and adding CWA provides an additional gain of 0.4% in mAP and 0.3% in NDS. These results demonstrate that motion features significantly enhance feature alignment.

3) Ablation Study Conducted on Argoverse 2 Valid Set: Table VI compares the performance of OnlineBEV against Baseline-S and Baseline-M (17) on the Argoverse 2 *validation* set, evaluating the generalization capability of our method beyond the nuScenes dataset. OnlineBEV demonstrates superior performance, surpassing Baseline-M (17) by 0.9% in mAP and 1.9% in CDS.

4) Robustness to Corrupted Image Input: Table VII highlights the robustness of OnlineBEV under adverse real-world conditions. As illustrated in Fig. 4, we applied synthetic input corruptions to simulate motion blur and occlusion scenarios, following the protocols of nuScenes-C [42] and nuScenes-R [43]. Across all corruption types, OnlineBEV consistently outperforms SOLOFusion [13]. Under the clean setting, OnlineBEV achieves improvements of 1.7% in mAP and 1.1% in NDS over SOLOFusion. Notably, the gaps further widen under adverse conditions, reaching 7.4% mAP and 4.6% NDS under motion blur, and 5.9% in mAP and 6.5% in NDS under occlusion. These results clearly demonstrate

the superior robustness of OnlineBEV in handling challenging visual conditions.

5) Complexity Analysis: Table VIII summarizes the computational cost and memory usage of OnlineBEV compared to other methods, including SOLOFusion [13], StreamPETR [15], and SparseBEV [14]. All measurements were conducted on an NVIDIA RTX 3090 with a ResNet-50 backbone. BEV-based approaches such as OnlineBEV and SOLOFusion incur higher computational costs than sparse query-based methods like StreamPETR and SparseBEV, as they operate on dense BEV representations. However, they offer global scene understanding and support a wider range of 3D perception tasks, including segmentation and occupancy prediction (see Section IV-C.3). Notably, OnlineBEV achieves the highest accuracy (44.4% mAP and 54.5% NDS) while requiring 205.7 GFLOPs and 79.3 ms of inference time. Despite its dense structure, it consumes only 3.4 GB of memory—lower than the 3.9 GB used by SOLOFusion.

6) Visualization of Feature Alignment: We illustrate the effectiveness of our feature alignment strategy by visualizing BEV features over time. Fig. 5 shows the historical, aligned, and current BEV features at each time step t, where t denotes the sequential frame index. To highlight the temporal evolution of alignment, five consecutive frames are visualized. As depicted, moving objects in the BEV space exhibit noticeable spatial misalignment between historical and current features. After applying MBFNet, the aligned features demonstrate substantially improved spatial correspondence with the current BEV features, highlighting the effectiveness of our alignment strategy.

7) Qualitative Results: Fig. 6 presents qualitative comparisons between OnlineBEV and SOLOFusion [13]. OnlineBEV achieves higher accuracy in localizing distant objects, benefiting from its temporal alignment strategy. Moreover, while SOLOFusion fails to detect objects on the rear side, OnlineBEV successfully identifies them. These results highlight OnlineBEV's ability to leverage long-term spatio-temporal information more effectively through temporal feature alignment.

V. Conclusion

In this paper, we proposed a novel temporal fusion framework for multi-camera 3D perception. Conventional temporal fusion methods often suffer from misaligned BEV features across consecutive frames due to the motion of dynamic objects, which limits their performance. To address this challenge, we introduced OnlineBEV, an effective feature alignment technique tailored for recurrent temporal fusion. OnlineBEV continuously updates historical BEV features by spatially aligning them and fusing them with current BEV features. Our approach leverages spatio-temporal deformable attention to adaptively align BEV features across frames, guided by motion context features extracted from two adjacent frames. To further enhance alignment, we incorporated temporal consistency learning, providing explicit supervision for BEV feature alignment.

Experiments on the nuScenes public dataset demonstrated that OnlineBEV achieves significant performance gains over existing temporal fusion methods. Moreover, when combined with recurrent temporal fusion, our method offers greater computational efficiency compared to parallel fusion approaches.

In this study, we utilized implicit motion features to predict alignment offsets. As a future direction, explicit motion information such as velocity vectors could be extracted from temporal features and integrated into our alignment process. This could be supervised using ground-truth motion data obtained from auxiliary sensors (e.g., LiDAR or radar). We plan to investigate this strategy to further enhance the performance of temporal fusion.

References

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

[2] Y. Li et al., "BEVDepth: Acquisition of reliable depth for multi-view 3D object detection," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, Jun. 2023, pp. 1477–1485.

[3] C. Feng, Z. Jie, Y. Zhong, X. Chu, and L. Ma, "AeDet: Azimuth-invariant multi-view 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 21580–21588.

[4] Z. Li, Z. Yu, W. Wang, A. Anandkumar, T. Lu, and J. M. Alvarez, "FB-BEV: BEV representation from forward-backward view transformations," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 6919–6928.

[5] J. Zhang, Y. Zhang, Q. Liu, and Y. Wang, "SA-BEV: Generating semantic-aware bird's-eye-view feature for multi-view 3D object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 3348–3357.

[6] Y. Wang, V. Guizilini, T. Zhang, Y. Wang, H. Zhao, and J. Solomon, "DETR3D: 3D object detection from multi-view images via 3D-to-2D queries," in *Proc. Conf. Robot Learn.*, Jan. 2022, pp. 180–191.

[7] Y. Liu, T. Wang, X. Zhang, and J. Sun, "PETR: Position embedding transformation for multi-view 3D object detection," in *Proc. 17th Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 531–548.

[8] D. Chen, J. Li, V. Guizilini, R. Ambrus, and A. Gaidon, "Viewpoint equivariance for multi-view 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 9213–9222.

[9] J. Philion and S. Fidler, "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3D," in *Proc. 16th Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Aug. 2020, pp. 194–210.

[10] J. Huang and G. Huang, "BEVDet4D: Exploit temporal cues in multi-camera 3D object detection," 2022, *arXiv:2203.17054*.

[11] Y. Li, H. Bao, Z. Ge, J. Yang, J. Sun, and Z. Li, "BEVStereo: Enhancing depth estimation in multi-view 3D object detection with temporal stereo," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2023, vol. 37, no. 2, pp. 1486–1494.

[12] S. Kim, Y. Kim, I.-J. Lee, and D. Kum, "Predict to detect: Prediction-guided 3D object detection using sequential images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 18011–18020.

[13] J. Park et al., "Time will tell: New outlooks and a baseline for temporal multi-view 3D object detection," in *Proc. Int. Conf. Learn. Represent.*, Jan. 2022, pp. 1–16.

[14] H. Liu, Y. Teng, T. Lu, H. Wang, and L. Wang, "SparseBEV: High-performance sparse 3D object detection from multi-camera videos," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 18534–18544.

[15] S. Wang, Y. Liu, T. Wang, Y. Li, and X. Zhang, "Exploring object-centric temporal modeling for efficient multi-view 3D object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 3598–3608.

[16] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," 2020, *arXiv:2010.04159*.

[17] H. Caesar et al., "nuScenes: A multimodal dataset for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11621–11631.

[18] M. A. Haq, S.-J. Ruan, M.-E. Shao, Q. M. U. Haq, P.-J. Liang, and D.-Q. Gao, "One stage monocular 3D object detection utilizing discrete depth and orientation representation," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 11, pp. 21630–21640, Nov. 2022.

[19] H. Yao et al., "Occlusion-aware plane-constraints for monocular 3D object detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 5, pp. 4593–4605, May 2024.

[20] W. Chen, J. Zhao, W.-L. Zhao, and S.-Y. Wu, "Shape-aware monocular 3D object detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 6, pp. 6416–6424, Jun. 2023.

[21] Z. Li et al., "Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 1–18.

[22] T. Kanade and M. Okutomi, "A stereo matching algorithm with an adaptive window: Theory and experiment," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, no. 9, pp. 920–932, Sep. 1994.

[23] Z. Luo, C. Zhou, G. Zhang, and S. Lu, "DETR4D: Direct multi-view 3D object detection with sparse attention," 2022, *arXiv:2212.07849*.

[24] X. Lin, T. Lin, Z. Pei, L. Huang, and Z. Su, "Sparse4D: Multi-view 3D object detection with sparse spatial–temporal fusion," 2022, *arXiv:2211.10581*.

[25] J. Jeong, S. Lee, J. Kim, and N. Kwak, "Consistency-based semi-supervised learning for object detection," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 10759–10768.

[26] Y. Ouali, C. Hudelot, and M. Tami, "Semi-supervised semantic segmentation with cross-consistency training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12674–12684.

[27] S. Varghese et al., "An unsupervised temporal consistency (TC) loss to improve the performance of semantic segmentation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 12–20.

[28] L. Ke, M. Danelljan, H. Ding, Y.-W. Tai, C.-K. Tang, and F. Yu, "Mask-free video instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 22857–22866.

[29] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2018, pp. 7132–7141.

[30] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, Jan. 2014.

[31] J. Lei Ba, J. Ryan Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.

[32] T. Yin, X. Zhou, and P. Krahenbuhl, "Center-based 3D object detection and tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11784–11793.

[33] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[34] Y. Liu et al., "PETRv2: A unified framework for 3D perception from multi-camera images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 3239–3249.

[35] B. Wilson et al., "Argoverse 2: Next generation datasets for self-driving perception and forecasting," 2023, *arXiv:2301.00493*.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[37] Y. Lee, J.-W. Hwang, S. Lee, Y. Bae, and J. Park, "An energy and GPU-computation efficient backbone network for real-time object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 752–760.

[38] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2019, *arXiv:1711.05101*.

[39] D. Park, R. Ambrus, V. Guizilini, J. Li, and A. Gaidon, "Is pseudo-LiDAR needed for monocular 3D object detection?" in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3122–3132.

[40] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[41] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 11976–11986.

[42] Y. Dong et al., "Benchmarking robustness of 3D object detection to common corruptions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 1022–1032.

[43] K. Yu et al., "Benchmarking the robustness of LiDAR-camera fusion for 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2023, pp. 3188–3198.

Junho Koh received the B.S., M.S., and Ph.D. degrees in electrical engineering from Hanyang University, Seoul, South Korea. In 2024, he joined Hyundai Motor Company, where he has been developing deployable 3D perception systems for end-to-end autonomous driving. His research interests include machine learning, robot vision, and 3D perception for autonomous vehicles.

Youngwoo Lee received the B.S. and M.S. degrees in electrical engineering from Hanyang University, Seoul, South Korea. In 2024, he joined Samsung Electronics Company, where he has been developing a multi-image fusion model for high-resolution image generation. His research interests include super-resolution, denoising, and tone mapping for software-based image signal processing.

Jungho Kim received the B.S. degree in automotive engineering from Kookmin University, Seoul, South Korea, in 2022, and the M.S. degree in artificial intelligence from Hanyang University, Seoul, in 2025. He is currently pursuing the Ph.D. degree with the Interdisciplinary Program in Artificial Intelligence, Seoul National University. His research interests include 3D perception, trajectory prediction, and planning in autonomous driving.

Dongyoung Lee received the B.S. degree in electronic engineering from Kookmin University, Seoul, South Korea, in 2021, and the M.S. degree in artificial intelligence from Hanyang University, Seoul, in 2025. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, Seoul National University. His research interests include deep learning for perception, 3D computer vision, and sensor fusion in autonomous driving.

Jun Won Choi (Member, IEEE) received the B.S. and M.S. degrees from Seoul National University and the Ph.D. degree from the University of Illinois at Urbana–Champaign. Following his studies, he joined Qualcomm, San Diego, USA, in 2010. From 2013 to 2024, he was a Faculty Member at the Department of Electrical Engineering, Hanyang University. Since 2024, he has been a Faculty Member at the Department of Electrical and Computer Engineering, Seoul National University. His research interests include signal processing, machine learning, robot perception, autonomous driving, and intelligent vehicles. He serves as an Associate Editor for IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, and *International Journal of Automotive Technology*.