# BEVNeXt: Reviving Dense BEV Frameworks for 3D Object Detection

Zhenxin Li[1, 2] Shiyi Lan[3] Jose M. Alvarez[3] Zuxuan Wu[1, 2†]
[1]Shanghai Key Lab of Intell. Info. Processing, School of CS, Fudan University
[2]Shanghai Collaborative Innovation Center of Intelligent Visual Computing [3]NVIDIA

## Abstract

*Recently, the rise of query-based Transformer decoders is reshaping camera-based 3D object detection. These query-based decoders are surpassing the traditional dense BEV (Bird's Eye View)-based methods. However, we argue that dense BEV frameworks remain important due to their outstanding abilities in depth estimation and object localization, depicting 3D scenes accurately and comprehensively. This paper aims to address the drawbacks of the existing dense BEV-based 3D object detectors by introducing our proposed enhanced components, including a CRF-modulated depth estimation module enforcing object-level consistencies, a long-term temporal aggregation module with extended receptive fields, and a two-stage object decoder combining perspective techniques with CRF-modulated depth embedding. These enhancements lead to a "modernized" dense BEV framework dubbed BEVNeXt. On the nuScenes benchmark, BEVNeXt outperforms both BEV-based and query-based frameworks under various settings, achieving a state-of-the-art result of 64.2 NDS on the nuScenes test set.*

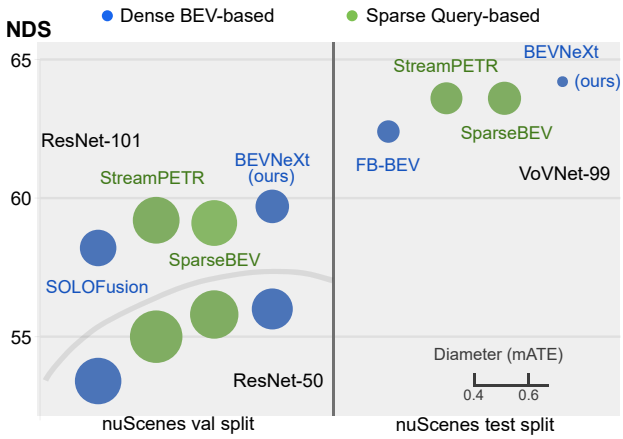


Figure 1. **Previous SOTAs vs. BEVNeXt on the nuScenes 3D Object Detection Benchmark.** On the nuScenes val split and test split, we compare BEVNeXt with previous SOTAs using (ResNet-50, bottom in the left panel), (ResNet-101, top in the left panel), and (VoVNet-99, right panel) as the backbone. BEVNeXt outperforms all previous sparse query-based ones in terms of comprehensive performance, meanwhile generating much fewer localization errors. The diameter of each bubble represents the mean Average Translation Error (mATE) each model produces. Higher and smaller bubbles are better. Best viewed in color.

## 1. Introduction

Visual-based 3D object detection [19, 30, 40, 63] is a critical component of autonomous driving and intelligent transportation systems. Unlike LiDAR-based systems with access to depth data, the primary challenge in visual-based 3D object detection is accurately perceiving depth, a task largely reliant on empirical knowledge of images. As an important part of detection, object localization depends heavily on the accuracy of depth [49]. Precise and robust object localization is the cornerstone of 3D perception, as it helps identify obstacles [1], lays a foundation for scene forecasting [46, 71], and leads to reassuring planning [16, 55].

To detect 3D objects with visual information, two research directions have prevailed: *dense BEV (Bird's Eye View)-based methods* and *sparse query-based methods*. BEV-based methods transform image feature maps into one unified dense bird's-eye-view feature map and thus apply detection decoders on it. In contrast, sparse query-based methods learn a set of object queries, which focus on sparse foreground objects rather than background details, and then predict 3D objects by leveraging multiple stages of cross-attention between object queries and image features and self-attention among object queries.

Despite the superior performance of recent query-based methods over dense BEV-based approaches, we maintain that retaining the dense feature map is advantageous for a complete environmental understanding, regardless of background or foreground elements. This trait makes BEV-based frameworks suitable for dense prediction tasks such as occupancy prediction [56, 57]. Further, the dense processing equips them with robustness in object localization, which

†Corresponding author.

accounts for their fewer localization errors compared with sparse counterparts as shown in Fig. 1. We argue that BEV-based detectors lag behind query-based ones due to less advanced network designs and training techniques. Building on this, we summarize the shortcomings of classic dense BEV-based approaches as follows:

- **Insufficient 2D Modeling.** Recent sparse query-based methods have demonstrated that improved 2D modeling can significantly enhance detection accuracy [21, 59]. In dense BEV-based approaches, efforts to boost 2D modeling include an auxiliary depth estimation task supervised by LiDAR inputs [29]. Yet, the impact is constrained by the low resolution of LiDAR points [3], leading to imprecise depth perception and suboptimal model performance.
- **Inadequate Temporal Modeling.** BEV frameworks often suffer from limited temporal modeling capabilities, which are critical in visual-based 3D detectors [36, 38, 49, 60]. Establishing a large receptive field in a dynamic 3D space during temporal fusion is crucial, especially when the ego vehicle and surrounding objects are in motion. Query-based methods [38, 60] can easily achieve this through the global attention mechanism [58], while BEV-based ones [13, 49] are bounded by the locality of convolutions.
- **Feature Distortion in Uplifting.** In dense BEV-based methods, feature distortion is a natural consequence of transforming feature maps across different coordinate systems and resolutions. On the other hand, sparse query-based approaches are unaffected since they attend to image feature maps in the 2D space rather than transformed features, thus avoiding feature distortion.

We introduce BEVNeXt, a modern dense BEV framework for 3D object detection comprising three main components. First, we employ a Conditional Random Field (CRF) to enhance depth accuracy and address depth supervision challenges, integrating depth probabilities with color information without extra supervision or significant computational cost. Second, the Res2Fusion module, inspired by Res2Net convolution blocks, expands the receptive field in dynamic 3D settings. Third, leveraging the predicted depth information, we have developed a two-stage object decoder. This decoder blends the spirit of sparse query-based techniques with CRF-enhanced depth embedding to improve instance-level BEV features using depth-focused 2D semantics. Together, these elements make BEVNeXt a stronger framework for object detection and localization.

We conduct in-depth experiments on the nuScenes dataset. As shown in Fig. 1, BEVNeXt achieves the highest 56.0% NDS and 64.2% NDS on the val split and test split, respectively, as well as the lowest mATE compared with all prior methods, demonstrating its outstanding comprehensive performance and preciseness in 3D object localization. More specifically, BEVNeXt outperforms previous state-of-the-art BEV-based SOLOFusion [49] by 2.6% and 2.3% on val split and test split, respectively.

## 2. Related Work

### 2.1. Dense BEV-based 3D Object Detection

Ever since the pioneering work of LSS [50], which introduces the concept of lifting 2D image features to the BEV space by predicting pixel-level depth probabilities, a significant research direction [13, 17–19, 28, 29, 49] has emerged, dedicated to improving the quality and efficiency of constructing the BEV space for 3D object detection and other perception tasks (*e.g.* map segmentation [15, 45, 50], occupancy prediction [56, 57]). The lifting procedure is also known as forward projection [32]. In particular, the BEVDet series [17–19] proposed an efficient pipeline to perform 3D object detection in the BEV space, as well as the short-term temporal modeling for velocity estimation [17]. BEVDepth [29] and BEVStereo [28] have respectively advanced the critical depth estimation process by leveraging explicit supervision from LiDAR point clouds [29] and stereo matching [28]. To extend their capabilities with long-term temporal information, SOLOFusion [49] adopts a straightforward concatenation technique across historical BEV representations, demonstrating exceptional performance, while VideoBEV [13] alleviates the heavy computation budget of SOLOFusion with recurrent modelling. However, their long-term fusion strategies suffer from an insufficient receptive field and rely on ego-motion transformation to discriminate stationary objects from moving ones, which can lead to motion misalignment of dynamic objects [20]. We argue that expanding the receptive field allows the model to distinguish different objects automatically.

Backward projection [30, 65] is the inverse operation of forward projection, a technique that samples multi-view 2D features and populates the BEV space with them. In recent advancements presented in FB-BEV [32] and FB-OCC [31], these two projection techniques are unified to obtain a stronger BEV representation, benefiting 3D object detection and occupancy prediction. This technique is used in our object decoder. However, unlike prior work, we utilize backward projection only to refine object-level BEV features, rather than the entire BEV representation. Further, this process is boosted with CRF-modulated depth embedding, which proves conducive to attribute prediction.

### 2.2. Sparse Query-based 3D Object Detection

Following query-based 2D object detectors [5, 73], a parallel avenue of research [35, 40, 63] has emerged. This alternative approach performs 3D object detection by directly querying 2D features, sidestepping the need for explicit 3D space construction. The querying procedure is typically carried out using the conventional attention mechanism [58],

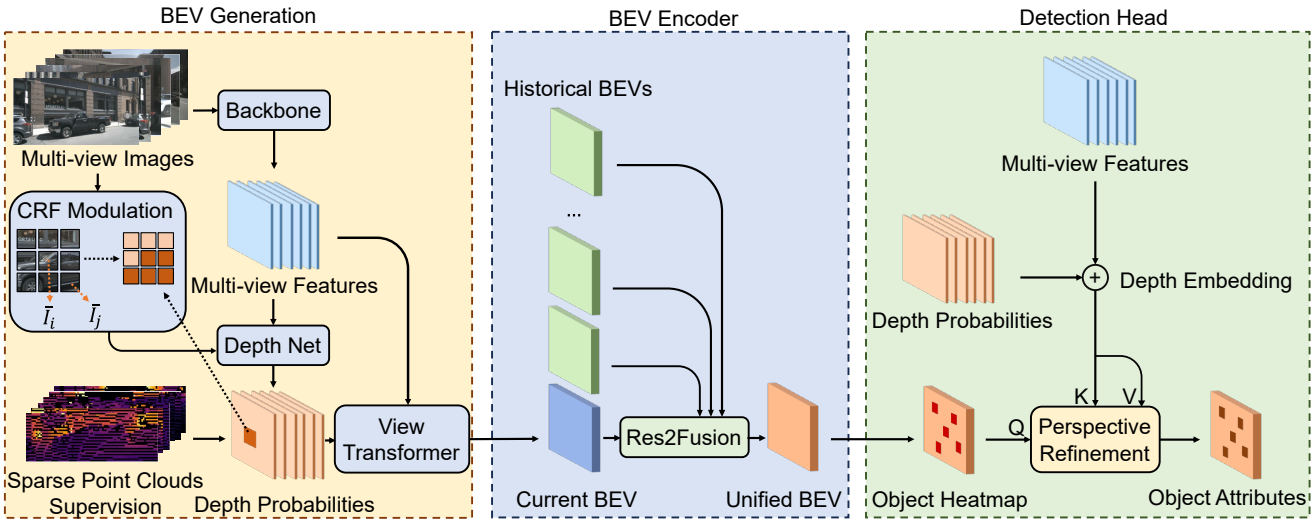


Figure 2. **Overall Architecture of BEVNeXt.** The backbone first extracts multi-view image features, which are converted into depth distributions with a depth network and CRF modulation. The BEV feature at the current frame is fused with previous ones through a Res2Fusion module. Finally, a CenterPoint detection head, coupled with perspective refinement, generates object heatmaps and attributes.

as seen in the PETR series [40, 41, 59, 60] or the sparse deformable attention mechanism [73], as seen in the Sparse4D series [35, 36]. After the emergence of SOLOFusion [49], the PETR series [40, 41, 59] embraces the concept of long-term temporal fusion and integrate it into the query-based framework. Through a well-designed propagation algorithm in the query space, StreamPETR [60] achieves substantial improvements compared to its static [40, 59] or short-term counterparts [41]. In addition, the recent breakthroughs made by SparseBEV [38] have demonstrated that these object queries can be explicitly defined within the BEV space, while Far3D [21] constructs 3D queries through the employment of a 2D object detector and a depth network, significantly expanding the range of 3D object detection. These detectors tend to produce more localization errors as they locate objects through the cross-attention mechanism instead of depth information. Unlike these methods, our work fully builds upon dense BEV frameworks, which are generally more robust in object localization.

### 2.3. CRF for Dense Predictions

Conditional Random Fields (CRF) have long been a fundamental tool for addressing dense prediction tasks such as semantic segmentation [23] and depth estimation [39], predating the widespread adoption of CNNs. With the emergence of CNNs, CRF as RNN [69] first shows that CRF can evolve as a seamless part of a CNN, whose job is to modulate pixel-level probabilities generated by the last CNN layer based on observed image features. For depth estimation, [4] treats the task as a pixel-wise classification problem, making it suitable for the application of CRF, while [37] calculates the CRF energy from continuous depth values. Additionally, in the domain of weakly-supervised instance segmentation (WSIS), where pixel-level mask annotations are not available, [24, 25] leverages CRF to enforce prediction consistency among pixels with similar color characteristics.

### 2.4. 3D Object Detection with LiDAR sensors

LiDAR sensors are widely used in 3D object detection [7, 11, 53, 54, 68, 70] since they generate reliable range information. These detectors typically decode objects from densely encoded BEV features [11, 53, 68, 70] or from sparse voxels [7, 54]. Furthermore, LiDAR sensors are jointly used with camera sensors in multi-modal 3D object detectors [2, 22, 43, 64, 67] due to their complementary nature. The advancements in this field are often directly applied to the pseudo point clouds generated by dense BEV frameworks [17, 19] for object decoding. Similarly, our object decoder derives inspiration from CenterFormer [70] and TransFusion [2], two 2-stage center-based detectors. Instead of densely attending to image features [2] or BEV features [70], our decoder employs depth-guided perspective refinement, an enhanced spatial cross-attention mechanism based on BEVFormer [30].

## 3. Method

We propose BEVNeXt, an enhanced dense BEV framework building upon existing LSS-based methods. BEVNeXt consists of three key components as shown in Fig. 2:

- **BEV Generation:** Given multi-view images $\{\boldsymbol{I^i}\}_{i=1}^{6}$, the backbone extracts multi-scale image features denoted as $\{F^i_{1/4}, F^i_{1/8}, F^i_{1/16}, F^i_{1/32}\}_{i=1}^{6}$, which are processed by a depth network for depth probabilities $\{\boldsymbol{d^i}\}_{i=1}^{6}$. For the spatial accuracy of the BEV feature, the CRF layer is employed to modulate $\{\boldsymbol{d^i}\}_{i=1}^{6}$ with image color information

$\{\boldsymbol{I^i}\}_{i=1}^6$, producing depth probabilities $\{\tilde{\boldsymbol{d}}^{\boldsymbol{i}}\}_{i=1}^6$ that are consistent on the object level. Then, the View Transformer computes the BEV feature $B_t$ at the current timestamp $t$ using features and modulated depth probabilities.

- **BEV Encoder:** The BEV Encoder is designed to fuse historical BEV features $\{B_{t-k+1}, ..., B_t\}$ across $k$ frames into a unified BEV representation $\tilde{B}$. The aggregation process demands a sufficient receptive field over the dynamic 3D environment, which is fulfilled in Res2Fusion.
- **Detection Head:** Finally, a center-based 3D object detection head [68] processes the output $\tilde{B}$ of the BEV Encoder, decoding the BEV representation into 3D objects. CRF-modulated depth probabilities $\{\tilde{\boldsymbol{d}}^{\boldsymbol{i}}\}_{i=1}^6$ are used as an embedding to help the object decoder attend to discriminative 2D features.

For the rest of this section, we elaborate on the specific enhancements of these components following the order of the detection pipeline.

### 3.1. CRF-modulated Depth Estimation

In dense BEV-based methods, depth estimation acts as a 2D auxiliary task, improving 2D modeling and potentially helping prevent feature distortion during uplifting. Therefore, obtaining accurate and higher-resolution depth prediction is beneficial. Considering depth estimation as a segmentation task, with each class representing a specific depth range, we can use Conditional Random Fields (CRF) to enhance the depth estimation quality. Following the utilization of CRFs in sparsely supervised prediction tasks [24, 25, 37], we aim at using CRF-modulated depth estimation to mitigate insufficient depth supervision by imposing a color smoothness prior [23], which enforces depth consistency at the pixel level. Let $\{X_1, ..., X_N\}$ represent the $N$ pixels in the downsampled feature map $F_{1/n}^i$, acquired by a stride of $n$, and $\{D_1, ..., D_k\}$ be $k$ discrete depth bins. The depth network's responsibility is to assign each pixel to various depth bins, represented as $\boldsymbol{d} = \{x_1, ..., x_N | x_i \in \{D_1, ..., D_k\}\}$. The camera index is discarded for convenience. Given this assignment $\boldsymbol{d}$, our objective is to minimize its corresponding energy cost $E(\boldsymbol{d}|\boldsymbol{I})$, as defined following [23]:

$$E(\boldsymbol{d}|\boldsymbol{I}) = \sum_i \psi_u(x_i) + \sum_{i \neq j} \psi_p(x_i, x_j), \qquad (1)$$

where $\sum_i \psi_u(x_i)$ are the unary potentials, measuring the cost associated with the initial output from the depth network. Building upon prior research [4, 69], we define the pairwise potential as:

$$\psi_p(x_i, x_j) = \sum_w w \exp\left(-\frac{|\bar{\boldsymbol{I}}_i - \bar{\boldsymbol{I}}_j|^2}{2\theta^2}\right)|x_i - x_j|, \qquad (2)$$

where $\bar{\boldsymbol{I}}_i$ and $\bar{\boldsymbol{I}}_j$ represent average RGB color values of image patches with dimensions matching the downsampling stride. Furthermore, $|x_i - x_j|$ is the label compatibility between two depth bins, which measures the distance of their centers in the real-world scale. The CRF is attached to the depth network as an extra layer. We denote the CRF-modulated depth probabilities as $\tilde{\boldsymbol{d}}$.

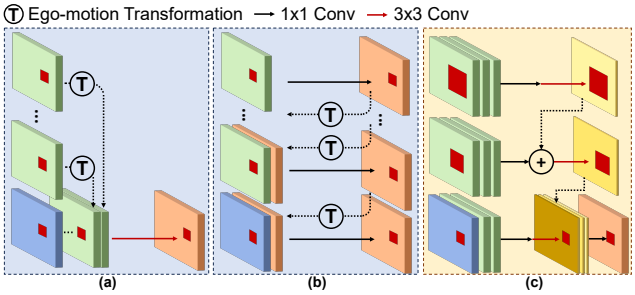


Figure 3. **Overview of Res2Fusion.** We list three major types of BEV temporal fusion techniques: (a) parallel fusion, (b) recurrent fusion, and (c) Res2Fusion (used in BEVNeXt).

In existing BEV-based solutions that rely on explicit depth supervision, such as BEVDepth [29] and SOLOFusion [49], the depth network typically processes the feature map at a small scale (*i.e.* $F_{1/16}^i$). Given a low-resolution input, there is a dense depth label coverage, which comes at the price of discarding too many labels with aggressive downsampling. This, in turn, compromises the effectiveness of our CRF-modulated depth estimation, as indicated in Tab. 4. To demonstrate the effectiveness of our approach under this resolution constraint, our depth network operates on a larger feature map (*i.e.* $F_{1/8}^i$) while halving the channel size. Nevertheless, the advantages of CRF modulation becomes increasingly noticeable as the input resolution scales up, as shown in Tab. 4.

### 3.2. Res2Fusion

The dense BEV-based approaches fuse the current BEV representation $B_t$ with past representations over an extended period, which is vital for perceiving dynamic 3D scenes, especially over long temporal windows where object locations significantly change. However, expanding the receptive field in BEV space is challenging; simply increasing the kernel size causes excessive computation and risks overfitting in uniform 3D environments [6].

To address these issues, we develop a temporal aggregation technique named Res2Fusion as shown in Fig. 3, which enlarges the receptive field by incorporating multi-scale convolution blocks from the Res2Net architecture [12]. Given $k$ historical BEV features $\{B_{t-k+1}, ..., B_t\}$ in which $B_t$ represents the BEV feature at the current frame, we first partition adjacent BEV features into $g = \frac{k}{w}$ groups with a fixed window size $w$, in which zero padding is used on the least recent group if $k \bmod w \neq 0$. The window size $w$ determines how much short-term locality the aggregation enjoys. After win-

dow partitioning, $1 \times 1$ convolutions $\{K_i^{1\times1}\}_{i=1}^{g}$ are used on these groups individually to reduce the channel size, which can be expressed as:

$$B'_i = K_i^{1\times1}([B_{t-(i+1)\times w}; ...; B_{t-i\times w}])(i = 0, ..., g), \quad (3)$$

where $[\cdot;\cdot]$ represents the concatenation operation. Next, the multi-scale convolutions proposed by [12] is used as:

$$B''_i = \begin{cases} K_i^{3\times3}(B'_i) & \text{if } i = g; \\ K_i^{3\times3}(B'_i + B'_{i+1}) & \text{if } 0 < i < g; \\ B'_i & \text{if } i = 0. \end{cases} \quad (4)$$

The increased receptive field, in turn, allows us to skip ego-motion transformation across historical BEVs, which avoids motion misalignment issues [20] in previous techniques [13, 49]. Finally, the output of the Res2Fusion module $\tilde{B}$ can be written as:

$$\tilde{B} = K_{final}^{1\times1}([B''_g; ...; B''_0]), \quad (5)$$

which is further processed with strided layers with a similar structure and an FPN [34] for multi-scale information following BEVDet [19].

### 3.3. Object Decoder with Perspective Refinement

With the unified BEV representation $\tilde{B}$ available, we apply a LiDAR-based 3D object detection head (*e.g.* CenterPoint Head [68]) to $\tilde{B}$ for final detection. Nevertheless, forward projection distorts 2D features, leading to a discrete and sparse BEV representation, as observed in FB-BEV [32]. Thus, we aim to compensate for the distortion using perspective refinement before regressing BEV features of ROI (Regions of Interest) to object attributes.

In the object decoder, we follow CenterPoint [68] to calculate the object heatmap $H$ by applying a $3 \times 3$ convolution and a sigmoid function to the output of the BEV Encoder $\tilde{B}$. The heatmap $H$ contains $K$ channels, corresponding to the $K$ object classes. For attribute regression, we first sample the features $B^{center} = \{\tilde{B}_{x,y} | H_{x,y} > \tau\}$ from object centers in $\tilde{B}$ by employing a heatmap threshold $\tau = 0.1$. As a regression head in CenterPoint typically consists of three convolution layers (one is shared among all heads), we expand $B^{center}$ to $B^{roi}$ by taking into account a $7 \times 7$ neighboring region of each element in $B^{center}$. Along with a set of learnable queries $\{Q_{x,y}\}$, the feature set $B^{roi}$ then goes through the perspective refinement process through a spatial cross-attention layer following [30]:

$$SCA(B^{roi}_{x,y}, F_{1/n}) = \sum_{i=1}^{N}\sum_{j=1}^{N_{ref}} \mathcal{F}_d(B^{roi}_{x,y} + Q_{x,y}, \mathcal{P}_i(x, y, z_j), F^i_{1/n}), \quad (6)$$

where $\mathcal{F}_d$ is the deformable attention function [73] and $\mathcal{P}_i(x, y, z_j)$ is a reference point lifted to height $z_j$. To introduce depth guidance, we embed the 2D features with CRF-modulated depth probabilities $\tilde{\boldsymbol{d}}^{\boldsymbol{i}}$, which are object-consistent after exploiting color information:

$$SCA(B^{roi}_{x,y}, F_{1/n}) = \sum_{i=1}^{N}\sum_{j=1}^{N_{ref}} \mathcal{F}_d(B^{roi}_{x,y} + Q_{x,y}, \mathcal{P}_i(x, y, z_j), F^i_{1/n} + Mlp(\tilde{\boldsymbol{d}}^{\boldsymbol{i}})). \quad (7)$$

Finally, regression heads from CenterPoint regress the refined feature set $B^{roi}$ to the final object attributes.

## 4. Experiments

### 4.1. Implementation Details

Our BEVNeXt builds upon BEVPoolv2 [18], a dense BEV-based framework featuring an efficient forward projection technique and the camera-aware depth network from BEVDepth [29]. BEVPoolv2 is also the baseline model in the following experiments. As the default configuration, we employ ResNet50 [14] as the image backbone, an input resolution of $256 \times 704$ for multi-view images, and a grid size of $128 \times 128$ for the BEV space. Only when we use larger backbones (*i.e.* ResNet101 [14], Vit-Adapter-L [8], V2-99 [26]), the BEV resolution is increased to $256 \times 256$. To maximize CRF-modulation's effect, as mentioned in Sec. 3.1, the depth network operates on $F_{1/8}$ to produce more fine-grained depth probabilities given an input resolution of $256 \times 704$, while on $F_{1/16}$ under other circumstances. When considering the incorporation of historical temporal information, we calculate BEV features from the past 8 frames in addition to the current frame, organized into 3 BEV groups with a window size of 3 for Res2Fusion.

For the training setup, our models are trained with data augmentations for both BEV and images [19]. The CBGS strategy [72] is adopted for a duration of 12 epochs, with the first 2 epochs without temporal information [49]. When employing ViT-Adapter-L and V2-99 as the image backbone, a shorter duration of 6 epochs is adopted to prevent over-fitting. We use the AdamW optimizer [44] and a total batch size of 64. The training process is consistent with BEVDepth [29], where the depth network and the detection head are optimized under supervision simultaneously.

### 4.2. Datasets and Metrics

We conduct extensive evaluations of our method on nuScenes [3], a multi-modal dataset encompassing 1000 distinct scenarios, which are recorded with a 32-beam LiDAR, six surround-view cameras, and five radars, annotated at a 2Hz frequency. Our assessment is based on the metrics of nuScenes, including mean Average Precision (mAP), mean Average Translation Error (mATE), mean Average Scale Error (mASE), mean Average Orientation Error (mAOE), mean Average Velocity Error (mAVE), mean Average Attribute Error (mAAE), and the final nuScenes Detection Score (NDS).

| Method | Backbone | Input Size | NDS↑ | mAP↑ | mATE↓ | mASE↓ | mAOE↓ | mAVE↓ | mAAE↓ |
|---|---|---|---|---|---|---|---|---|---|
| BEVDepth [29] | ResNet50 | 256 × 704 | 0.475 | 0.351 | 0.639 | 0.267 | 0.479 | 0.428 | 0.198 |
| BEVPoolv2 [18] | ResNet50 | 256 × 704 | 0.526 | 0.406 | 0.572 | 0.275 | 0.463 | 0.275 | 0.188 |
| SOLOFusion [49] | ResNet50 | 256 × 704 | 0.534 | 0.427 | 0.567 | 0.274 | 0.511 | 0.252 | 0.181 |
| VideoBEV [13] | ResNet50 | 256 × 704 | 0.535 | 0.422 | 0.564 | 0.276 | 0.440 | 0.286 | 0.198 |
| Sparse4Dv2 [36] | ResNet50 | 256 × 704 | 0.539 | **0.439** | 0.598 | 0.270 | 0.475 | 0.282 | **0.179** |
| StreamPETR [60] | ResNet50 | 256 × 704 | 0.540 | 0.432 | 0.581 | 0.272 | 0.413 | 0.295 | 0.195 |
| SparseBEV [38] | ResNet50 | 256 × 704 | 0.545 | 0.432 | 0.606 | 0.274 | **0.387** | **0.251** | 0.186 |
| **BEVNeXt** | ResNet50 | 256 × 704 | **0.548** | 0.437 | **0.550** | **0.265** | 0.427 | 0.260 | 0.208 |
| StreamPETR* [60] | ResNet50 | 256 × 704 | 0.550 | 0.450 | 0.613 | 0.267 | 0.413 | 0.265 | 0.196 |
| SparseBEV* [38] | ResNet50 | 256 × 704 | 0.558 | 0.448 | 0.581 | 0.271 | **0.373** | **0.247** | **0.190** |
| **BEVNeXt*** | ResNet50 | 256 × 704 | **0.560** | **0.456** | **0.530** | **0.264** | 0.424 | 0.252 | 0.206 |
| BEVDepth [29] | ResNet101 | 512 × 1408 | 0.535 | 0.412 | 0.565 | 0.266 | 0.358 | 0.331 | 0.190 |
| SOLOFusion [49] | ResNet101 | 512 × 1408 | 0.582 | 0.483 | 0.503 | 0.264 | 0.381 | 0.246 | 0.207 |
| SparseBEV* [38] | ResNet101 | 512 × 1408 | 0.592 | 0.501 | 0.562 | 0.265 | 0.321 | 0.243 | 0.195 |
| StreamPETR* [60] | ResNet101 | 512 × 1408 | 0.592 | 0.504 | 0.569 | 0.262 | **0.315** | 0.257 | 0.199 |
| Sparse4Dv2* [36] | ResNet101 | 512 × 1408 | 0.594 | 0.505 | 0.548 | 0.268 | 0.348 | 0.239 | **0.184** |
| Far3D* [21] | ResNet101 | 512 × 1408 | 0.594 | **0.510** | 0.551 | **0.258** | 0.372 | **0.238** | 0.195 |
| **BEVNeXt*** | ResNet101 | 512 × 1408 | **0.597** | 0.500 | **0.487** | 0.260 | 0.343 | 0.245 | 0.197 |
| StreamPETR [60] | ViT-L | 320 × 800 | 0.609 | 0.530 | 0.564 | **0.255** | **0.302** | 0.240 | 0.207 |
| **BEVNeXt** | ViT-Adapter-L [8] | 320 × 800 | **0.622** | **0.535** | **0.467** | 0.260 | 0.309 | **0.227** | **0.195** |

Table 1. **Comparison on the nuScenes val set.** ViT-L [10] is pretrained on COCO [33] and Objects365 [52], while ViT-Adapter-L [8] is pretrained on DINOv2 [47]. * The backbone benefits from perspective pretraining [61].

| Method | Backbone | Input Size | NDS↑ | mAP↑ | mATE↓ | mASE↓ | mAOE↓ | mAVE↓ | mAAE↓ |
|---|---|---|---|---|---|---|---|---|---|
| BEVDepth [29] | V2-99 | 640 × 1600 | 0.600 | 0.503 | 0.445 | 0.245 | 0.378 | 0.320 | 0.126 |
| BEVStereo [28] | V2-99 | 640 × 1600 | 0.610 | 0.525 | 0.431 | 0.246 | 0.358 | 0.357 | 0.138 |
| SOLOFusion [49] | ConvNeXt-B | 640 × 1600 | 0.619 | 0.540 | 0.453 | 0.257 | 0.376 | 0.276 | 0.148 |
| FB-BEV [38] | V2-99 | 640 × 1600 | 0.624 | 0.537 | 0.439 | 0.250 | 0.358 | 0.270 | 0.128 |
| VideoBEV [13] | ConvNeXt-B | 640 × 1600 | 0.629 | 0.554 | 0.457 | 0.249 | 0.381 | 0.266 | 0.132 |
| BEVFormerv2 [65] | InternImage-XL [62] | 640 × 1600 | 0.634 | 0.556 | 0.456 | 0.248 | **0.317** | 0.293 | 0.123 |
| StreamPETR [60] | V2-99 | 640 × 1600 | 0.636 | 0.550 | 0.479 | 0.239 | **0.317** | 0.241 | 0.119 |
| SparseBEV [38] | V2-99 | 640 × 1600 | 0.636 | 0.556 | 0.485 | 0.244 | 0.332 | 0.246 | 0.117 |
| Sparse4Dv2 [36] | V2-99 | 640 × 1600 | 0.638 | 0.556 | 0.462 | **0.238** | 0.328 | 0.264 | **0.115** |
| **BEVNeXt** | V2-99 | 640 × 1600 | **0.642** | **0.557** | **0.409** | 0.241 | 0.352 | **0.233** | 0.129 |

Table 2. **Comparison on the nuScenes test set.** ConvNeXt-B [42] is pretrained on ImageNet-22K [9], while V2-99 is initialized from a DD3D [48] backbone. The listed methods do not use future frames during training or testing.

| Method | AMOTA↑ | AMOTP↓ | RECALL↑ | IDS↓ |
|---|---|---|---|---|
| SPTR-QTrack [60, 66] | 0.566 | 0.975 | 0.650 | 784 |
| DORT [51] | 0.576 | 0.951 | 0.634 | 774 |
| **BEVNeXt-PolyMot [27]** | **0.578** | **0.917** | **0.720** | **519** |

Table 3. **3D multi-object tracking on the nuScenes test set.** Ours uses V2-99 as the backbone while the others use ConvNeXt-B.

3D Multi-Object Tracking is evaluated on Average Multi-Object Tracking Accuracy (AMOTA), Average Multi-Object Tracking Precision (AMOTP), Recall, and ID Switch (IDS).

### 4.3. Main Results

Tab. 1 shows detailed detection performance on the validation split of nuScenes. Specifically, using a ResNet50 [14] backbone and an input resolution of 256 × 704, BEVNeXt outperforms BEV-based SOLOFusion [49] by a clear margin of 2.6% NDS and 2.9% mAP, exceeding recently proposed query-based SparseBEV [38] with and without perspective pretraining [61]. This demonstrates the detection abilities of BEVNeXt in a lightweight setting. Furthermore, equipped with ResNet101, BEVNeXt outperforms the strong query-based framework Far3D [21] by 0.3% NDS, while using ViT-Adapter-L [8], BEVNeXt yields a new state-of-the-art result of 62.2% NDS on the val split. These experiments verify the scalability of our proposed modules, which allows BEVNeXt to thrive with a larger and more modern backbone. On the test split of nuScenes, Tab. 2 shows that

| Backbone | Input Size | LiDAR Coverage | CRF | NDS↑ | mAP↑ | mATE↓ | mAOE↓ | mAVE↓ |
|---|---|---|---|---|---|---|---|---|
| ResNet50 | 256 × 704 | ≈ 85% | | 0.490 | 0.368 | 0.628 | **0.485** | **0.355** |
| | | | ✓ | **0.492** | **0.372** | **0.614** | 0.490 | 0.369 |
| | 384 × 1056 | ≈ 65% | | 0.502 | 0.390 | 0.599 | 0.456 | 0.395 |
| | | | ✓ | **0.521** | **0.406** | **0.581** | **0.419** | **0.343** |
| ResNet101 | 512 × 1408 | ≈ 50% | | 0.535 | 0.412 | 0.565 | 0.358 | 0.331 |
| | | | ✓ | **0.553** | **0.433** | **0.544** | **0.332** | **0.308** |

Table 4. **Ablation of CRF modulation with different backbones and input resolutions.** All depth networks operate on $F_{1/16}$. Only 1 history frame is used. The effect of CRF modulation is minor given dense point clouds supervision.

BEVNeXt outperforms all prior methods with a moderate-sized V2-99 [26] backbone under the same input resolution of 640 × 1600. Specifically, BEVNeXt surpasses the BEV-based SOLOFusion [49] by a 2.3% NDS and the query-based Sparse4Dv2 [36] by a 0.4% NDS. Meanwhile, BEVNeXt produces fewer translation errors (*i.e.* mATE) than prior detectors consistently, demonstrating its superiority in object localization. In particular, BEVNeXt achieves 5.3% less mATE compared with Sparse4Dv2 [36]. Integrating the PolyMot [27] tracker with the detections of BEVNeXt-V2-99, BEVNeXt surpasses the BEV-based specialized tracker DORT by 0.2% AMOTA on the nuScenes test set (Tab. 3).

### 4.4. Ablation Studies

| Res2Fusion | $F_{1/n}$ | CRF | Refinement | NDS↑ | mAP↑ |
|---|---|---|---|---|---|
| | 16 | | | 0.526 | 0.406 |
| ✓ | 16 | | | 0.537 | 0.420 |
| ✓ | 8 | | | 0.540 | 0.430 |
| ✓ | 8 | ✓ | | 0.542 | 0.434 |
| ✓ | 8 | ✓ | ✓ | **0.548** | **0.437** |

Table 5. **Ablation of BEVNeXt Components**. The baseline is BEVPoolv2 with an input resolution of 256 × 704 , ResNet50 as the backbone, and a long-term history of 8 frames.

**Ablation on Different Components.** To verify the effectiveness of our proposed components, we gradually remove the components of BEVNeXt. As depicted in Tab. 5, the absence of each component causes a distinct degradation of NDS and mAP. Though the increase of the feature map scale aims to accommodate CRF modulation, we observe this modification brings performance improvements (+0.3% NDS) itself, which is also confirmed in the localization potential analysis made by SOLOFusion [49]. The integrated version of BEVNeXt brings an overall performance gain of 2.2% NDS and 3.1% mAP compared with the baseline.

**CRF Modulation.** In Tab. 4, the role of CRF modulation under sparse supervision is studied. The impact of CRF modulation is insignificant (+0.2% NDS) when supervision is dense already, but becomes much more notable (+1.9% NDS) under sparse supervision. It can be shown that the object-level consistency provided by CRF modulation comprehensively enhances the detection performance of BEV frameworks, which also demonstrates scalability (+1.8% NDS) using a larger backbone (*i.e.* ResNet101). Yet, in long-term settings, the improvements become moderate due to saturating localization potential [49], as displayed in Tab. 5.

| Ego-motion Trans. | Window Size $w$ | NDS↑ | mAP↑ |
|---|---|---|---|
| - | 2 | 0.531 | 0.417 |
| ✓ | 3 | 0.535 | 0.416 |
| - | 3 | **0.537** | **0.420** |
| - | 4 | 0.529 | 0.418 |

Table 6. **Ablation of Res2Fusion.** We compare different window sizes $w$ and the effect of ego-motion transformation over 8 historical frames (9 frames in total). Zero padding is used if the number of frames cannot be divided evenly by $w$.

| Depth Embedding | Depth Source | NDS↑ | mAP↑ |
|---|---|---|---|
| - | - | 0.497 | 0.379 |
| ✓ | Depth Network | 0.501 | 0.379 |
| ✓ | CRF | **0.505** | **0.382** |

Table 7. **Ablation of Depth Embedding in Perspective Refinement.** All depth networks operate on $F_{1/8}$, as the input resolution is 256 × 704. Only 1 history frame is used.

**Design of Res2Fusion.** Tab. 6 shows how Res2Fusion is affected by the window size and ego-motion transformation. The window size determines the number of adjacent BEVs to process in parallel, which only demands smaller receptive fields than the long-term scenario. Our experiment shows that a window size of 3 maximizes the effect of Res2Fusion by reaching a balance between the short-term locality and the long-term receptive field. In addition, forcibly warping previous BEV features to the current timestamp causes a performance degradation, which is caused by the misalignment of dynamic objects [20].

**Effect of Depth Embedding.** In Tab. 7, we ablate the effect of depth embedding in the perspective refinement module. The purpose of depth embedding is to help the 3D object

| Component | Enhancements | Param. (M) | GFLOPs |
|---|---|---|---|
| Depth Network | w/o CRF | 27.3 | 3864.4 |
| | w/ CRF | 27.3 | 3873.9 |
| BEV Encoder | Parallel Fusion | 54.6 | 172.8 |
| | Res2Fusion | 6.9 | 31.4 |
| Detection Head | w/o Refinement | 1.6 | 99.7 |
| | w/ Refinement | 2.3 | 125.4 |

| Method | Speed (FPS) |
|---|---|
| StreamPETR-R101 [60] | 6.4 |
| SOLOFusion-R101 [49] | 1.5 |
| BEVNeXt-R101 | 4.4 |

Table 8. **Analysis of Runtime Efficiency**. The listed methods use ResNet101 as the image backbone. Both SOLOFusion-R101 and BEVNeXt-R101 utilize a BEV resolution of $256 \times 256$.

decoder attend to discriminative features by forming object-level consistencies in the 2D space. When the depth is produced by the depth network, there exists positive but minor influences. However, when the CRF-modulated depth information is adopted, we witness an 0.8% increase in NDS, which is a boost in the prediction of object attributes. This phenomenon also verifies the effect of CRF modulation.

### 4.5. Visualization and Efficiency Analysis


Figure 4. **Comparison of Depth Estimation with and without CRF modulation on the nuScenes val split.** We visualize depth ranges using an argmax operation on various depth bins. The CRF-modulated depth probabilities can distinguish objects from the background better.

**Visualization.** We first visualize CRF-modulated depth estimation in Fig. 4. The modulated depth probabilities are more boundary-sticky and achieve higher object-level consistencies. Besides, they contain fewer artifacts, which interfere with the spatial accuracy of BEV features. For perspective refinement, both large and small objects can benefit from this process as shown in Fig. 5. Compared with a coarse object decoder, refined objects are more accurate in orientation.

**Efficiency Analysis.** As shown in Tab. 8, our proposed modules require negligible or even less computation, making them suitable for deployment. With a PyTorch fp32 backend and an RTX 3090, though falling behind query-based StreamPETR [60], BEVNeXt is faster than BEV-based SOLOFusion [49]. This can be attributed to the absence of a temporal stereo, which is computationally expensive.

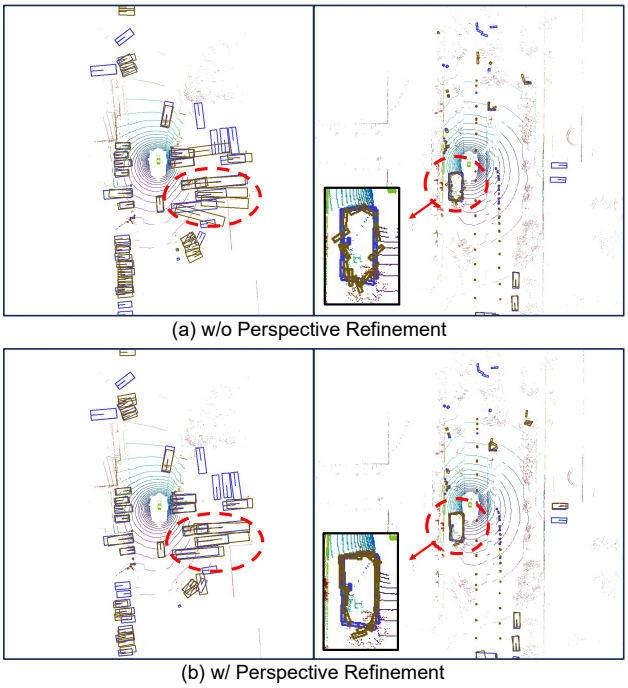


Figure 5. **Comparison of Detection Results with and without Perspective Refinement on the nuScenes val split.** Compared with the coarse predictions of a CenterPoint head [68], our refined objects are more aligned with the ground truths.

## 5. Conclusion

In this work, we proposed a fully enhanced dense BEV framework for multi-view 3D object detection dubbed BEVNeXt. We first identify three shortcomings of classic dense BEV-based frameworks: (1) insufficient 2D modeling, (2) inadequate temporal modeling, and (3) feature distortion in uplifting. To address these inherent issues, we propose three corresponding components: (1) CRF-modulated depth estimation, (2) Res2Fusion for long-term temporal aggregation, and (3) an object decoder with perspective refinement. Extensive experiments are carried out, showing that BEVNeXt excels in object localization and supersedes sparse query paradigms and dense BEV frameworks on the nuScenes benchmark. Specifically, BEVNeXt achieves a new state-of-the-art of 56.0% NDS and 64.2% NDS on the nuScenes val split and test split, respectively.

**Limitations.** Though BEVNeXt demonstrates stronger performance than existing sparse query paradigms, it still falls behind in terms of efficiency. Integrating BEV frameworks into long-range settings also present challenges. We expect these issues to be addressed in future research.

**Acknowledgement** This project was supported by NSFC under Grant No. 62102092.

## References


[1] Samira Badrloo, Masood Varshosaz, Saied Pirasteh, and Jonathan Li. Image-based obstacle detection methods for the safe navigation of unmanned vehicles: A review. *Remote Sens.*, 14(15):3824, 2022. 1

[2] Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *CVPR*, pages 1090–1099, 2022. 3

[3] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, pages 11621–11631, 2020. 2, 5

[4] Yuanzhouhan Cao, Zifeng Wu, and Chunhua Shen. Estimating depth from monocular images as classification using deep fully convolutional residual networks. *IEEE Trans. Circuits Syst. Video Technol.*, 28(11):3174–3182, 2017. 3, 4

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229. Springer, 2020. 2

[6] Yukang Chen, Jianhui Liu, Xiangyu Zhang, Xiaojuan Qi, and Jiaya Jia. Largekernel3d: Scaling up kernels in 3d sparse cnns. In *CVPR*, pages 13488–13498, 2023. 4

[7] Yukang Chen, Jianhui Liu, Xiangyu Zhang, Xiaojuan Qi, and Jiaya Jia. Voxelnext: Fully sparse voxelnet for 3d object detection and tracking. In *CVPR*, pages 21674–21683, 2023. 3

[8] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. In *ICLR*, 2022. 5, 6

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009. 6

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 6

[11] Lue Fan, Ziqi Pang, Tianyuan Zhang, Yu-Xiong Wang, Hang Zhao, Feng Wang, Naiyan Wang, and Zhaoxiang Zhang. Embracing single stride 3d object detector with sparse transformer. In *CVPR*, pages 8458–8468, 2022. 3

[12] Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. Res2net: A new multi-scale backbone architecture. *IEEE TPAMI*, 43(2):652–662, 2019. 4, 5

[13] Chunrui Han, Jianjian Sun, Zheng Ge, Jinrong Yang, Runpei Dong, Hongyu Zhou, Weixin Mao, Yuang Peng, and Xiangyu Zhang. Exploring recurrent long-term temporal fusion for multi-view 3d perception. *arXiv preprint arXiv:2303.05970*, 2023. 2, 5, 6

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 5, 6

[15] Anthony Hu, Zak Murez, Nikhil Mohan, Sofía Dudas, Jeffrey Hawke, Vijay Badrinarayanan, Roberto Cipolla, and Alex Kendall. Fiery: Future instance prediction in bird's-eye view from surround monocular cameras. In *ICCV*, pages 15273–15282, 2021. 2

[16] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *CVPR*, pages 17853–17862, 2023. 1

[17] Junjie Huang and Guan Huang. Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. *arXiv preprint arXiv:2203.17054*, 2022. 2, 3

[18] Junjie Huang and Guan Huang. Bevpoolv2: A cutting-edge implementation of bevdet toward deployment. *arXiv preprint arXiv:2211.17111*, 2022. 5, 6

[19] Junjie Huang, Guan Huang, Zheng Zhu, Yun Ye, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021. 1, 2, 3, 5

[20] Linyan Huang, Zhiqi Li, Chonghao Sima, Wenhai Wang, Jingdong Wang, Yu Qiao, and Hongyang Li. Leveraging vision-centric multi-modal expertise for 3d object detection. *arXiv preprint arXiv:2310.15670*, 2023. 2, 5, 7

[21] Xiaohui Jiang, Shuailin Li, Yingfei Liu, Shihao Wang, Fan Jia, Tiancai Wang, Lijin Han, and Xiangyu Zhang. Far3d: Expanding the horizon for surround-view 3d object detection. *arXiv preprint arXiv:2308.09616*, 2023. 2, 3, 6

[22] Yang Jiao, Zequn Jie, Shaoxiang Chen, Jingjing Chen, Lin Ma, and Yu-Gang Jiang. Msmdfusion: Fusing lidar and camera at multiple scales with multi-depth seeds for 3d object detection. In *CVPR*, pages 21643–21652, 2023. 3

[23] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *NeurIPS*, 24, 2011. 3, 4

[24] Shiyi Lan, Zhiding Yu, Christopher Choy, Subhashree Radhakrishnan, Guilin Liu, Yuke Zhu, Larry S Davis, and Anima Anandkumar. Discobox: Weakly supervised instance segmentation and semantic correspondence from box supervision. In *ICCV*, pages 3406–3416, 2021. 3, 4

[25] Shiyi Lan, Xitong Yang, Zhiding Yu, Zuxuan Wu, Jose M Alvarez, and Anima Anandkumar. Vision transformers are good mask auto-labelers. In *CVPR*, pages 23745–23755, 2023. 3, 4

[26] Youngwan Lee, Joong-won Hwang, Sangrok Lee, Yuseok Bae, and Jongyoul Park. An energy and gpu-computation efficient backbone network for real-time object detection. In *CVPRW*, pages 0–0, 2019. 5, 7

[27] Xiaoyu Li, Tao Xie, Dedong Liu, Jinghan Gao, Kun Dai, Zhiqiang Jiang, Lijun Zhao, and Ke Wang. Poly-mot: A polyhedral framework for 3d multi-object tracking. In *IROS*, pages 9391–9398. IEEE, 2023. 6, 7

[28] Yinhao Li, Han Bao, Zheng Ge, Jinrong Yang, Jianjian Sun, and Zeming Li. Bevstereo: Enhancing depth estimation in multi-view 3d object detection with dynamic temporal stereo. *AAAI*, 2022. 2, 6

[29] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth:

Acquisition of reliable depth for multi-view 3d object detection. In *AAAI*, pages 1477–1485, 2023. 2, 4, 5, 6

[30] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *ECCV*, pages 1–18. Springer, 2022. 1, 2, 3, 5

[31] Zhiqi Li, Zhiding Yu, David Austin, Mingsheng Fang, Shiyi Lan, Jan Kautz, and Jose M Alvarez. Fb-occ: 3d occupancy prediction based on forward-backward view transformation. *arXiv preprint arXiv:2307.01492*, 2023. 2

[32] Zhiqi Li, Zhiding Yu, Wenhai Wang, Anima Anandkumar, Tong Lu, and Jose M Alvarez. Fb-bev: Bev representation from forward-backward view transformations. In *ICCV*, pages 6919–6928, 2023. 2, 5

[33] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014. 6

[34] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017. 5

[35] Xuewu Lin, Tianwei Lin, Zixiang Pei, Lichao Huang, and Zhizhong Su. Sparse4d: Multi-view 3d object detection with sparse spatial-temporal fusion. *arXiv preprint arXiv:2211.10581*, 2022. 2, 3

[36] Xuewu Lin, Tianwei Lin, Zixiang Pei, Lichao Huang, and Zhizhong Su. Sparse4d v2: Recurrent temporal fusion with sparse model. *arXiv preprint arXiv:2305.14018*, 2023. 2, 3, 6, 7

[37] Fayao Liu, Chunhua Shen, and Guosheng Lin. Deep convolutional neural fields for depth estimation from a single image. In *CVPR*, pages 5162–5170, 2015. 3, 4

[38] Haisong Liu, Yao Teng, Tao Lu, Haiguang Wang, and Limin Wang. Sparsebev: High-performance sparse 3d object detection from multi-camera videos. In *ICCV*, pages 18580–18590, 2023. 2, 3, 6

[39] Miaomiao Liu, Mathieu Salzmann, and Xuming He. Discrete-continuous depth estimation from a single image. In *CVPR*, pages 716–723, 2014. 3

[40] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petr: Position embedding transformation for multi-view 3d object detection. In *ECCV*, pages 531–548. Springer, 2022. 1, 2, 3

[41] Yingfei Liu, Junjie Yan, Fan Jia, Shuailin Li, Aqi Gao, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petrv2: A unified framework for 3d perception from multi-camera images. In *ICCV*, pages 3262–3272, 2022. 3

[42] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, pages 11976–11986, 2022. 6

[43] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela L Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *ICRA*, pages 2774–2781. IEEE, 2023. 3

[44] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2017. 5

[45] Mong H Ng, Kaahan Radia, Jianfei Chen, Dequan Wang, Ionel Gog, and Joseph E Gonzalez. Bev-seg: Bird's eye view semantic segmentation using geometry and semantic point cloud. *arXiv preprint arXiv:2006.11436*, 2020. 2

[46] Jiquan Ngiam, Benjamin Caine, Vijay Vasudevan, Zhengdong Zhang, Hao-Tien Lewis Chiang, Jeffrey Ling, Rebecca Roelofs, Alex Bewley, Chenxi Liu, Ashish Venugopal, et al. Scene transformer: A unified architecture for predicting multiple agent trajectories. In *ICLR*, 2021. 1

[47] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 6

[48] Dennis Park, Rares Ambrus, Vitor Guizilini, Jie Li, and Adrien Gaidon. Is pseudo-lidar needed for monocular 3d object detection? In *CVPR*, pages 3142–3152, 2021. 6

[49] Jinhyung Park, Chenfeng Xu, Shijia Yang, Kurt Keutzer, Kris Kitani, Masayoshi Tomizuka, and Wei Zhan. Time will tell: New outlooks and a baseline for temporal multi-view 3d object detection. In *ICLR*, 2022. 1, 2, 3, 4, 5, 6, 7, 8

[50] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *ECCV*, pages 194–210. Springer, 2020. 2

[51] LIAN Qing, Tai Wang, Dahua Lin, and Jiangmiao Pang. Dort: Modeling dynamic objects in recurrent for multi-camera 3d object detection and tracking. In *CoRL*, pages 3749–3765. PMLR, 2023. 6

[52] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *ICCV*, pages 8430–8439, 2019. 6

[53] Guangsheng Shi, Ruifeng Li, and Chao Ma. Pillarnet: High-performance pillar-based 3d object detection. *ECCV*, 2022. 3

[54] Pei Sun, Mingxing Tan, Weiyue Wang, Chenxi Liu, Fei Xia, Zhaoqi Leng, and Dragomir Anguelov. Swformer: Sparse window transformer for 3d object detection in point clouds. In *ECCV*, pages 426–442. Springer, 2022. 3

[55] Siyu Teng, Xuemin Hu, Peng Deng, Bai Li, Yuchen Li, Yunfeng Ai, Dongsheng Yang, Lingxi Li, Zhe Xuanyuan, Fenghua Zhu, et al. Motion planning for autonomous driving: The state of the art and future perspectives. *IEEE Trans. Intell. Veh.*, 2023. 1

[56] Xiaoyu Tian, Tao Jiang, Longfei Yun, Yue Wang, Yilun Wang, and Hang Zhao. Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. *NeurIPS*, 2023. 1, 2

[57] Wenwen Tong, Chonghao Sima, Tai Wang, Silei Wu, Hanming Deng, Li Chen, Yi Gu, Lewei Lu, Ping Luo, Dahua Lin, et al. Scene as occupancy. In *ICCV*, pages 8406–8415, 2023. 1, 2

[58] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017. 2

[59] Shihao Wang, Xiaohui Jiang, and Ying Li. Focal-petr: Embracing foreground for efficient multi-camera 3d object detection. *arXiv preprint arXiv:2212.05505*, 2022. 2, 3

[60] Shihao Wang, Yingfei Liu, Tiancai Wang, Ying Li, and Xiangyu Zhang. Exploring object-centric temporal modeling for efficient multi-view 3d object detection. In *ICCV*, pages 3621–3631, 2023. 2, 3, 6, 8

[61] Tai Wang, Xinge Zhu, Jiangmiao Pang, and Dahua Lin. Fcos3d: Fully convolutional one-stage monocular 3d object detection. In *ICCV*, pages 913–922, 2021. 6

[62] Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, et al. Internimage: Exploring large-scale vision foundation models with deformable convolutions. In *CVPR*, pages 14408–14419, 2023. 6

[63] Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *CoRL*, pages 180–191. PMLR, 2022. 1, 2

[64] Junjie Yan, Yingfei Liu, Jianjian Sun, Fan Jia, Shuailin Li, Tiancai Wang, and Xiangyu Zhang. Cross modal transformer: Towards fast and robust 3d object detection. In *ICCV*, pages 18268–18278, 2023. 3

[65] Chenyu Yang, Yuntao Chen, Hao Tian, Chenxin Tao, Xizhou Zhu, Zhaoxiang Zhang, Gao Huang, Hongyang Li, Yu Qiao, Lewei Lu, et al. Bevformer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision. In *CVPR*, pages 17830–17839, 2023. 2, 6

[66] Jinrong Yang, En Yu, Zeming Li, Xiaoping Li, and Wenbing Tao. Quality matters: Embracing quality clues for robust 3d multi-object tracking. *arXiv preprint arXiv:2208.10976*, 2022. 6

[67] Zeyu Yang, Jiaqi Chen, Zhenwei Miao, Wei Li, Xiatian Zhu, and Li Zhang. Deepinteraction: 3d object detection via modality interaction. *NeurIPS*, 35:1992–2005, 2022. 3

[68] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *CVPR*, pages 11784–11793, 2021. 3, 4, 5, 8

[69] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip HS Torr. Conditional random fields as recurrent neural networks. In *ICCV*, pages 1529–1537, 2015. 3, 4

[70] Zixiang Zhou, Xiangchen Zhao, Yu Wang, Panqu Wang, and Hassan Foroosh. Centerformer: Center-based transformer for 3d object detection. In *ECCV*, pages 496–513. Springer, 2022. 3

[71] Zikang Zhou, Jianping Wang, Yung-Hui Li, and Yu-Kai Huang. Query-centric trajectory prediction. In *CVPR*, pages 17863–17873, 2023. 1

[72] Benjin Zhu, Zhengkai Jiang, Xiangxin Zhou, Zeming Li, and Gang Yu. Class-balanced grouping and sampling for point cloud 3d object detection. *arXiv preprint arXiv:1908.09492*, 2019. 5

[73] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR*, 2020. 2, 3, 5