Instance-Aware Multi-Camera 3D Object Detection with Structural Priors Mining and Self-Boosting Learning

Yang Jiao[1,2], Zequn Jie[3], Shaoxiang Chen[3], Lechao Cheng[4], Jingjing Chen[1,2†], Lin Ma[3†], Yu-Gang Jiang[1,2]

[1]Shanghai Key Lab of Intell. Info. Processing, School of CS, Fudan University
[2]Shanghai Collaborative Innovation Center on Intelligent Visual Computing
[3]Meituan
[4]Zhejiang Lab

Abstract

Camera-based bird-eye-view (BEV) perception paradigm has made significant progress in the autonomous driving field. Under such a paradigm, accurate BEV representation construction relies on reliable depth estimation for multi-camera images. However, existing approaches exhaustively predict depths for every pixel without prioritizing objects, which are precisely the entities requiring detection in the 3D space. To this end, we propose IA-BEV, which integrates image-plane instance awareness into the depth estimation process within a BEV-based detector. First, a category-specific structural priors mining approach is proposed for enhancing the efficacy of monocular depth generation. Besides, a self-boosting learning strategy is further proposed to encourage the model to place more emphasis on challenging objects in computation-expensive temporal stereo matching. Together they provide advanced depth estimation results for high-quality BEV features construction, benefiting the ultimate 3D detection. The proposed method achieves state-of-the-art performances on the challenging nuScenes benchmark, and extensive experimental results demonstrate the effectiveness of our designs.

Introduction

In recent years, there has been a surge of research interest in multi-camera 3D object detection within the autonomous driving field (Huang et al. 2021; Li et al. 2022c, 2023a; Feng et al. 2023). Compared with LiDAR, the camera excels at capturing object semantics and enjoys the advantage of a lower deployment cost. The recent trend in this field is to transform the multi-view image features to a unified Bird's-Eye-View (BEV) space for the subsequent perception. This paradigm facilitates aligning signals from multiple sensors and timestamps in the BEV space, serving as a generic representation for downstream tasks such as detection (Jiao et al. 2023; Li et al. 2023b), map segmentation (Xie et al. 2022) and question answering (Qian et al. 2023).

Within the BEV-based perception pipeline, depth estimation plays a pivotal role in the perspective projection from the image view to BEV. Pioneering methods estimate depth from monocular images either implicitly (Li et al. 2022c) or explicitly (Li et al. 2023a). Motivated by the success of the multi-view stereo technique (Yao et al. 2018; Wei et al. 2021), follow-up approaches (Li et al. 2022a; Park et al. 2022) leverage consecutive camera frames to construct cost volume for stereo matching. Benefiting from enhanced depth estimation, these methods enjoy high-quality BEV features and thus achieve remarkable detection performances.

Despite significant advancements, existing methods (Li et al. 2023a, 2022a) treat every pixel equally, neglecting the inherent properties encapsulated in foreground objects. In fact, foreground objects can exhibit consistency within a category and heterogeneity among instances, which we

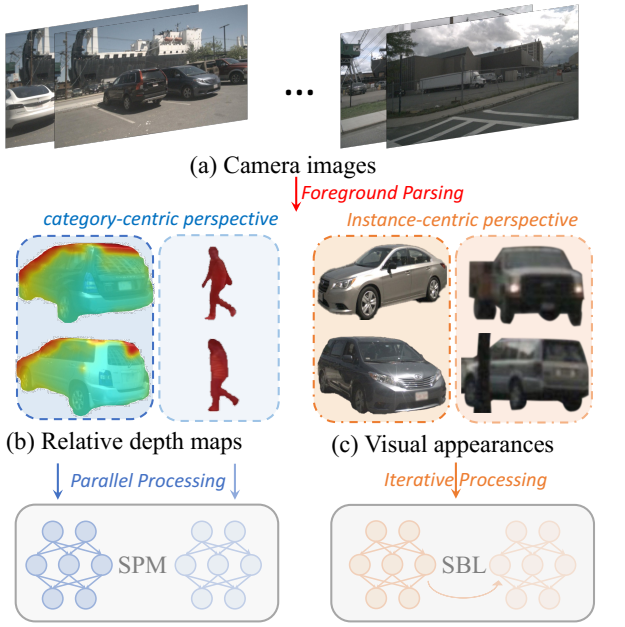

Figure 1: The overall workflow of the proposed IA-BEV. In (b), we first use gaussian kernel to densely fill depth values for each pixel, then calculate relative depths by dividing the maximum depth of the corresponding objects. All objects illustrated in (b) and (c) are extracted from nuScenes images and resized to the same scale for clarity.

†Corresponding authors

find can be utilized to improve depth estimation. On the one hand, objects of the same semantic category share similar structural priors, which manifest in two key aspects. 1) Object scales shown in the image are correlated with their real depths, and this correlation is usually coherent for objects of the same semantic category and varies across categories. For example, the scales of cars in image are inversely proportional to their real depths, but even with the same depth, a car and a pedestrian can exhibit significantly different scales. 2) Objects of the same semantic category have consistent inner-geometric structures. As shown in Fig 1(b), when observed in isolation from the image plane, the objects of the same category (car) have similar distributions of relative depths. On the other hand, for individual objects, their visual appearance can vary dramatically even within the same category due to different resolutions and occlusion statuses. Consequently, the complexity of depth estimation for different object instances also varies. As demonstrated in Fig 1(c), cars on the left column contain more precise textures and shape details versus those on the right column, thus reducing the ambiguity of the challenging depth estimation. Although some approaches (Chu et al. 2023; Wang et al. 2023) have explored 2D object priors for 3D object detection, they primarily leverage detected 2D objects after the perspective projection, thereby ignoring their potential to improve depth estimation for enhanced BEV feature construction.

Motivated by the aforementioned observations and to overcome the limitation of existing methods, we propose IA-BEV, which exploits 2D instance awareness to enhance the depth estimation process in the BEV-based detector. As shown in Fig 1(d), our IA-BEV initially parses a scene into individual objects and then leverages their intrinsic properties to assist both monocular and stereo depth estimation with novelly devised Structural Priors Mining approach (SPM) and Self-Boosting Learning strategy (SBL), respectively. Within SPM, objects belonging to the same or similar semantic categories are grouped and processed by respective lightweight depth decoders to better exploit structural priors. However, expecting these parallel decoders to actively learn category-specific patterns with only grouped input poses significant optimization challenges, resulting in suboptimal performance. To address this, we explicitly encode object scale properties as additional inputs and apply two instance-aware loss functions to supervise the rough instance absolute depth and the fine-grained inner-object relative depth predictions. In contrast to SPM, SBL operates in a class-agnostic manner, which focuses on iteratively distinguishing and emphasizing challenging objects. Within each iteration, objects are first partitioned into two groups according to their stereo-matching uncertainty. Subsequently, the group with higher uncertainty, indicating inaccurate estimation, is further boosted in the subsequent iteration. Thanks to the gradually sparser foreground regions addressed in the later iterations, we can set denser depth hypotheses within the realm of uncertainty for more comprehensive stereo matching on the selected challenging samples. Finally, on the basis of the combined depth estimates from both SPM and SBL, the conventional view transformation process is conducted to construct BEV features for the ultimate detection.

In summary, our contributions are three-fold: (1) We propose IA-BEV, which enhances the depth estimation process within the BEV perception pipeline via exploiting 2D instance awareness. (2) Within IA-BEV, a Structural Priors Mining approach (SPM) and a Self-Boosting Learning strategy (SBL) are introduced to exploit object intrinsic properties to promote monocular and stereo depth estimation, respectively. (3) Our IA-BEV achieves significant improvements over the strong BEVDepth baseline and state-of-the-art performances among all methods that also utilize two keyframes on the challenging nuScenes benchmark.

Related Work

Depth Estimation

Estimating depths from camera images has been a classical topic in computer vision. Contemporary research can be grouped into monocular and stereo depth estimation approaches. Monocular depth estimation aims to predict depths from a single image. Mainstream methods (Bhat, Alhashim, and Wonka 2021; Poggi et al. 2020; Li et al. 2022b) in this line adopt an encoder-decoder pipeline to directly predict the depth values or distribution at the input resolution. However, monocular depth estimation is a longstanding ill-posed problem due to its inherent scale ambiguity. As an alternative, stereo depth estimation is based on multi-view image inputs to construct the cost volume to learn the pixel-to-pixel matching behaviors to meet the epipolar geometry constraints (Yao et al. 2018; Peng et al. 2022; Shen, Dai, and Rao 2021). As a key step to link the 2D and 3D space, depth estimation techniques have been extensively adopted in modern autonomous-driving field 3D detectors (Sun et al. 2020; Li et al. 2023a, 2022a). However, due to the scene layout complexity and supervision signals sparsity in the outdoor scenarios, the depth estimation quality in these detectors remains unsatisfactory.

Multi-camera 3D Object Detection

Modern BEV-based 3D detectors can be categorized into two paradigms. The first paradigm transforms image features to the BEV space via the Lift-Splat-Shoot (Philion and Fidler 2020) technique. The pioneering work, BEVDet (Huang et al. 2021), first implements the complete LSS-based detection pipeline. Follow-up approaches (Li et al. 2023a, 2022a; Park et al. 2022) enhance the depth estimation by introducing depth supervision or leveraging multi-frame information. Another line of work projects 3D object queries to multi-view image planes to collect useful image features. DETR3D (Wang et al. 2022a) and PETR (Liu et al. 2022a) extend DETR (Carion et al. 2020) detector to mutli-camera 3D object detection. BEVFormer-series methods (Li et al. 2022c; Yang et al. 2023) further incorporate both spatial and temporal cues for more robust detection. On the basis of these two fundamental paradigms, some recent studies have explored 2D object priors to assist 3D detection. OA-BEV (Chu et al. 2023) lifts detected regions as foreground objects from the image plane to 3D space, and processes them with a voxel encoder to generate

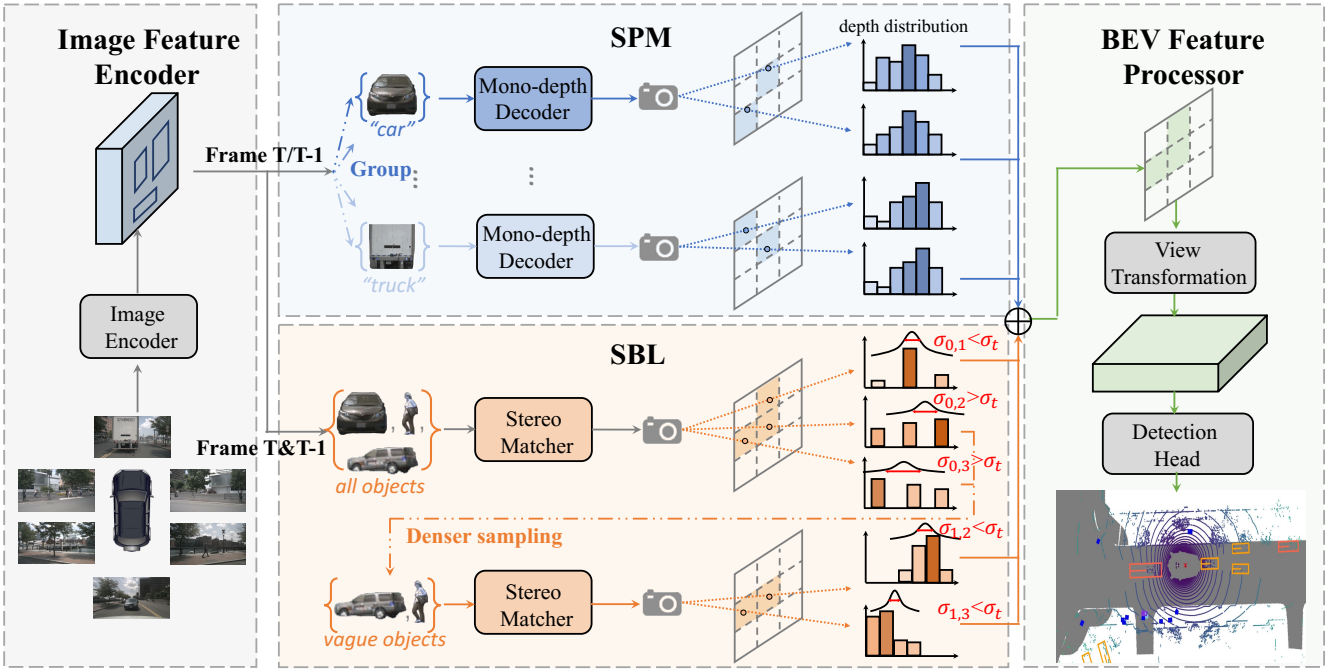

Figure 2: The detailed designs of the proposed IA-BEV. Given images collected from multi-view cameras, foreground objects are first parsed by off-the-shelf 2D scene parsers. Then, these objects, together with image features, are fed into the proposed SPM and SBL for effective depth estimation by exploring object properties from category-centric and instance-centric perspectives, respectively. Finally, the outputs from both SPM and SBL are merged, resulting in the ultimate image depths, which will be used for the conventional view transformation and BEV-based detection. Frame T and $T-1$ are individually fed into SPM, while they are supplied to SBL simultaneously as the stereo matching here requires temporal multi-view information.

object-aware BEV features as the augmentation of the original ones. MV2D (Wang et al. 2023) utilizes 2D object features and detections to directly predict 3D detection results. Instead of relying on 2D object detections for late fusion or final decision, our IA-BEV aims to harness object inherent properties in early depth estimation, enabling the construction of high-quality BEV features.

Method

As shown in Fig 2, the proposed IA-BEV comprises four key components: a feature encoder responsible for extracting image features and parsing foreground objects, the proposed Structural Priors Mining approach (SPM) which enhances the monocular depth estimation by leveraging structural consistency of the same category of objects, the proposed Self-Boosting Learning strategy (SBL) which emphasizes vague objects during stereo depth estimation, and finally, a BEV feature encoder utilized for rendering features and detecting objects in the BEV space. In the following sections, we will elaborate on each of them.

Feature Encoder

Given images collected from multi-view cameras, we extract image features with a prevalent backbone like ResNet-50 (He et al. 2016) or ConvNeXt (Liu et al. 2022c). Meanwhile, we use the off-the-shelf instance segmentor (Zhou, Koltun, and Krähenbühl 2021) to parse foreground objects $\mathcal{O} = \{(F_i, b_i)\}_{i=1}^{N}$, where F_i includes features of all pixels belonging to the current object, b_i refers to box parameters, and N is the number of segemented foreground objects. Note that here we keep all object pixel features rather than pooling them into a single vector because our goal is to densely predict depths for the entire region of objects. On the basis of parsed objects $\mathcal{O}$, we devise a Structural Priors Mining approach (SPM) and a Self-Boosting Learning (SBL) strategy to unleash the potential of objects' inherent properties in the depth estimation from category-centric and instance-centric perspectives, respectively, which will be elaborated in the following parts.

Structural Priors Mining

Category-Specific Depth Decoders. Estimating depths from monocular images is challenging as it requires understanding the relationships between physical scales and depth values of objects with different semantics. Existing BEV-based methods (Li et al. 2023a, 2022a) utilize prevalent pre-trained image backbones (He et al. 2016; Liu et al. 2022c) as feature encoders to endow the model with strong semantic-capturing ability, however, they rely on a single depth decoder to simultaneously learn scale-to-depth mapping patterns of multiple semantic categories, increasing the burden of optimization.

To simplify learning such patterns of different semantic categories, we design multiple parallel lightweight depth decoders, where each of them is responsible for processing objects of the same category as shown in Fig 2. Specifically, we first divide all foreground objects into several non-overlapping semantic groups $\{\mathcal{O}^{(c_i)}\}_{i=1}^{K}$, where K is the number of object categories. Then, taking an object $o_j^{(c_i)} = (F_j^{(c_i)}, b_j^{(c_i)})$ from semantic group $\mathcal{O}^{(c_i)}$ as an example, we feed both object features and box parameters (i.e., normalized box height and width) into a lightweight depth decoder. Within each depth decoder, box parameters are encoded by a linear mapping and then fused with object features using the SE block (Hu, Shen, and Sun 2018). And the outputs will be fed into a convolutional layer to predict the depth logits $D_j^{(c_i)}$ for regions of the current object. The above process can be formulated as:

$$\begin{aligned} \widetilde{F}_j^{(c_i)} &= \texttt{SE}(F_j^{(c_i)}; \texttt{Linear}(b_j^{(c_i)})) \\ D_j^{(c_i)} &= \texttt{Conv}(\widetilde{F}_j^{(c_i)}) \end{aligned} \tag{1}$$

Finally, by merging the predicted depth logits of all instances, the estimated monocular depth can be obtained.

Instance-Aware Supervision. In a typical BEV-based perception pipeline (Li et al. 2023a), the depth prediction is supervised by pixel-wise cross-entropy loss, which fails to capture fine-grained instance-level cues, making it more challenging for the aforementioned category-specific depth decoders to learn semantic structural priors. Therefore, we design two new loss functions to encourage learning the rough instance-level absolute depths and inner-instance relative depths. First, for object $o_j^{(c_i)}$ we convert the discrete depth prediction $D_j^{(c_i)} \in \mathbb{R}^{B\times M}$ into continuous depth values $\tilde{D}_j^{(c_i)} \in \mathbb{R}^{M}$ following MonoDETR (Zhang et al. 2022):

$$\hat{D}_j^{(c_i)} = \sum_{k=1}^{B} (d\,[k] \cdot \texttt{Softmax}(D_j^{(c_i)}, \texttt{dim} = 0)[k]) \tag{2}$$

where $d\,[k]$ represents the depth value of the center of k-th depth bin, B and M represent the number of pre-defined depth bins and object pixels. Then, we project the LiDAR points within ground-truth 3D boxes onto the image plane to obtain ground-truth depth values, and keep those intersecting with foreground objects $\mathcal{O}$ to further construct supervision signals. Here, we denote objects with both predicted and ground-truth depths as $\mathcal{O}' = \{(F_i, b_i, \hat{D}_i, D_i^{gt})\}_{i=1}^{N'}$, where $D_i^{gt} \in \mathbb{R}^{M_i'}$, M_i' is the number of ground-truth depth values. To explicitly supervise the instance-level depth prediction, for each object $o_i' = (F_i, b_i, \hat{D}_i, D_i^{gt})$ in $\mathcal{O}'$, we abstract an absolute depth value d_i^{gt} from D_i^{gt} as the regression target. It is worth noting that there exist some outliers in the ground-truth depths D_i^{gt} due to imperfect sensor calibration (Zhao et al. 2023)[1], which poses great challenges in choosing a proper d_i^{gt}. Therefore, we first scatter all depth values in D_i^{gt} into predefined depth bins, and then only average those in the depth bin with the maximum votes as d_i^{gt}. Afterward, the total absolute depth loss can be calculated as:

$$\mathcal{L}_{depth}^{abs} = \frac{1}{N'} \sum_{i=1}^{N'} \frac{1}{M_i'} \sum_{j=1}^{M_i'} (d_i^{gt} - \hat{D}_i\,[j])^2 \tag{3}$$

On the basis of d_i^{gt}, we also calculate the relative depth loss to encourage the category-specific decoder to learn fine-grained object geometric patterns:

$$\mathcal{L}_{depth}^{rel} = \frac{1}{N'} \sum_{i=1}^{N'} \frac{1}{M_i'} \sum_{j=1}^{M_i'} ((d_i^{gt} - \hat{D}_i\,[j]) - (d_i^{gt} - D_i^{gt}\,[j]))^2 \tag{4}$$

Self-Boosting Learning

The temporal stereo-matching technique relies on geometric consistency through time for depth estimation (Wang et al. 2022b). Concretely, for every pixel in the T-th frame, several depth hypotheses are initially proposed along the depth channel. Then, these hypotheses are warped to the $(T-1)$-th frame to construct cost volume for learning the best match among them. In the above process, the main barrier lies in the large memory cost brought by constructing 3D cost volume for huge amounts of pixels in high-resolution image features and dense hypotheses (Li et al. 2022a). However, the image regions should not be treated equally in our scenario. First, the foreground objects are more important than the background area. Furthermore, it is harder to accurately estimate depth for objects with lower visual clarity and more attention should be paid to them. Therefore, we devise a self-boosting strategy to iteratively focus on harder object regions, which further enables adaptively adjusting the granularity of cost volume construction for different regions and results in a better trade-off between cost and efficacy.

Sparse Cost Volume Construction. In pursuit of enhanced efficiency, we mainly focus on exploring the stereo-matching behaviors of foreground objects at T-th frame, which breaks the conventional dense cost volume construction paradigm. Therefore, we reformulate such a process into a sparse format introduced as follows. Taking a pixel with coordinates (u, v) and depth hypothesis d^h as example, we employ the homography warping between T-th and $(T-1)$-th frames on it to obtain the corresponding projected location (u^{T-1}, v^{T-1}):

$$(u^{T-1}, v^{T-1}) = \texttt{Homo}((u, v, d^h); \mathcal{K}; \mathcal{M}_{T\to T-1}) \tag{5}$$

where $\mathcal{K}$ is the camera intrinsic parameters, and $\mathcal{M}_{T\to T-1}$ is the transformation matrix from T-th to $(T-1)$-th frame. Following the above process, for every object pixel with different depth hypotheses, we establish its correspondence to pixels in $(T-1)$-th frame, and then combine their features to generate the sparse cost volume $V \in \mathbb{R}^{N_p\times N_d\times C_f}$, where N_p and N_d are number of foreground pixels and depth hypotheses, respectively, and C_f is the feature channel dimension. Subsequently, the matching scores are calculated with 3D sparse convolutions (Contributors 2022).

[1]Intuitive illustrations of this phenomenon are included in the supplementary materials.

Method	Input Size	mAP ↑	mATE ↓	mASE ↓	mAOE ↓	mAVE ↓	mAAE ↓	NDS ↑
BEVDet4D-R50 (Huang et al. 2021)	256x704	0.322	0.703	0.278	0.495	0.354	0.206	0.457
PETR-R50 (Liu et al. 2022a)	384x1056	0.313	0.768	0.278	0.495	0.923	0.225	0.381
BEVDepth-R50 (Li et al. 2023a)	256x704	0.351	0.639	0.267	0.479	0.428	0.198	0.475
BEVStereo-R50 (Li et al. 2022a)	256x704	0.372	0.598	**0.270**	**0.438**	0.367	**0.190**	0.500
AeDet-R50 (Feng et al. 2023)	256x704	0.387	0.598	0.276	0.461	0.392	0.196	0.501
FB-BEV-R50 (Li et al. 2023c)	256x704	0.378	0.620	0.273	0.444	0.374	0.200	0.498
IA-BEV-R50 (ours)	256x704	**0.400**	**0.557**	0.275	0.449	**0.347**	0.209	**0.516**
BEVDepth-ConvNeXt-B (Li et al. 2023a)	512x1408	0.462	0.540	0.254	0.355	0.379	0.200	0.558
BEVStereo-ConvNeXt-B (Li et al. 2022a)	512x1408	0.478	-	-	-	-	-	0.575
SA-BEV-ConvNeXt-B (Zhang et al. 2023)	512x1408	0.479	-	-	-	-	-	0.579
IA-BEV-ConvNeXt-B (ours)	512x1408	**0.493**	**0.493**	0.259	0.364	**0.336**	0.207	**0.581**

Table 1: Comparison with state-of-the-art methods on nuScenes val set.

Method	Input Size	mAP ↑	mATE ↓	mASE ↓	mAOE ↓	mAVE ↓	mAAE ↓	NDS ↑
BEVDet4D-Swin-B (Huang et al. 2021)	640x1600	0.451	0.511	0.241	0.386	0.301	0.121	0.569
BEVFormer-Vov99 (Li et al. 2022c)	640x1600	0.481	0.582	0.256	0.375	0.378	0.126	0.569
PETRv2 (Liu et al. 2022b)	640x1600	0.490	0.561	0.243	0.361	0.343	0.120	0.582
BEVDepth-ConvNeXt-B (Li et al. 2023a)	640x1600	0.520	0.445	0.243	0.352	0.347	0.127	0.609
BEVStereo-Vov99 (Li et al. 2022a)	640x1600	0.525	0.431	0.246	0.358	0.357	0.138	0.610
OA-BEV-Vov99 (Chu et al. 2023)	900x1600	0.494	0.574	0.256	0.377	0.385	0.132	0.575
MV2D-Vov99 (Chu et al. 2023)	640x1600	0.511	0.525	0.243	0.357	0.357	0.120	0.596
SOLOFusion-ConvNeXt-B† (Park et al. 2022)	640x1600	0.540	0.453	0.257	0.376	0.276	0.148	0.619
AeDet-ConvNeXt-B (Feng et al. 2023)	640x1600	0.531	0.439	0.247	0.344	0.292	0.130	0.620
CAPE-Vov99 (Xiong et al. 2023)	900x1600	0.525	0.503	0.242	0.361	0.306	**0.114**	0.610
FB-BEV-Vov99 (Li et al. 2023c)	640x1600	0.537	0.439	0.250	0.358	**0.270**	0.128	0.624
SA-BEV-Vov99 (Zhang et al. 2023)	640x1600	0.533	0.430	**0.241**	**0.338**	0.282	0.139	0.624
IA-BEV-ConvNeXt-B (ours)	640x1600	**0.545**	**0.407**	0.248	0.343	0.294	0.133	**0.630**

Table 2: Comparison with state-of-the-art methods on nuScenes test set. † denotes that longer temporal frames (>2) are used.

Method	SILog ↓	AbsRel ↓	SqRel ↓	log10 ↓	RMSE ↓
BEVDepth (Li et al. 2023a)	21.74	0.155	1.223	0.060	5.269
BEVStereo (Li et al. 2022a)	21.74	0.152	1.206	0.059	5.246
IA-BEV (ours)	**17.64**	**0.134**	**1.178**	**0.046**	**3.843**

Table 3: Evaluation of depth prediction performances of different methods.

Iterative Stereo Matching. In the first round, to efficiently recognize objects with rich visual details, we uniformly sample sparse depth hypotheses $H_0 \in \mathbb{R}^{L_0}$ for all pixels $\mathcal{P}_0$ in foreground objects $\mathcal{O}$. Then the sparse cost volume is constructed based on $\mathcal{P}_0$ and H_0 to calculate the matching scores $S_0 \in \mathbb{R}^{N_0 \times L_0}$, where N_0 and L_0 are number of pixels and depth hypotheses, respectively. For pixel $\mathcal{P}_{0,i}$, we calculate the mean $\mu_{0,i}$ and standard deviation $\sigma_{0,i}$ along its depth channel as:

$$\mu_{0,i} = \sum_{j=1}^{L_0} (H_0[j] \cdot S_{0,i}[j]) \tag{6}$$

$$\sigma_{0,i}^2 = \sum_{j=1}^{L_0} ((H_0[j] - \mu_{0,i})^2 \cdot S_{0,i}[j]) \tag{7}$$

The scale of $\sigma_{0,i}$ indicates the uncertainty of the stereo depth estimation. With small $\sigma_{0,i}$, the depth hypotheses have been successfully verified to find the best match. Conversely, large $\sigma_{0,i}$ means that multiple depth hypotheses are preferred, and thus should be further boosted. Therefore, we regard the pixels whose matching scores' standard deviation are smaller than our predefined threshold σ_t as satisfactory results, and filter them in the next iteration. For the remaining pixels, their mean and standard deviation can provide a more accurate search range, which facilitates proposing depth hypotheses more effectively for the next iteration. With μ_0 and σ_0, we update the depth sampling range as:

$$\mathcal{R}_1 = [\mu_0 - 3\sigma_0, \mu_0 + 3\sigma_0] \tag{8}$$

Within $\mathcal{R}_1$, we further uniformly sample L_1 ($L_1 > L_0$) depth hypotheses H_1 for remaining pixels $\mathcal{P}_1$. Both H_1 and $\mathcal{P}_1$ will be utilized for constructing the sparse cost volume and calculating mean and deviation similarly in the next iteration. Since the numbers of depth hypotheses in different iterations are different, we employ an interpolation operation to fill all predefined depth bins for alignment.

BEV Feature Processor

By summing up the predicted monocular and stereo depths from SPM and SBL, the final depth prediction can be obtained for rendering the BEV feature from multi-camera images. Afterward, the BEV feature will be fed into a conventional detection head for the ultimate 3D detection. The total loss functions can be formulated as:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{det} + \lambda_2 \mathcal{L}^{CE}_{depth} + \lambda_3 \mathcal{L}^{abs}_{depth} + \lambda_4 \mathcal{L}^{rel}_{depth} \quad (9)$$

where $\mathcal{L}_{det}$ and $\mathcal{L}^{CE}_{depth}$ are conventional detection loss and pixel-wise cross-entropy depth loss, respectively.

Experiments

Experimental Setup

Dataset and Metrics NuScenes dataset (Caesar et al. 2020) is a large-scale autonomous driving benchmark, encompassing LiDAR, camera and radar data collected from 10,000 unique driving scenarios. These scenarios are systematically grouped into 700 for training, 150 for validation, and 150 for testing. For the purpose of detection, a suite of evaluation metrics are introduced, including the nuScenes Detection Score (NDS), mean Average Precision (mAP), alongside five True Positive (TP) metrics, specifically, mean Average Translation Error (mATE), mean Average Scale Error (mASE), mean Average Orientation Error (mAOE), mean Average Velocity Error (mAVE), and mean Average Attribute Error (mAAE). For depth estimation quality evaluation, the scale-invariant logarithmic error (SILog), mean absolute relative error (AbsRel), mean squared relative error (SqRel), mean log10 error (log10) and root mean squared error (RMSE) are reported.

Implementation Details We use the BEVDepth (Li et al. 2023a) as the baseline method, and follow its official training configurations, including data augmentation, optimizer selection, and basic hyperparameters. Within SPM, objects are divided into 6 semantic groups: *"car", "truck & construction vehicle", "bus & trailer", "barrier", "motorcycle & bicycle", "pedestrian & traffic cone"*. As for SBL, we perform the sparse stereo matching for 2 rounds. The number of sampled depth hypotheses in the first iteration (L_0) and second iteration (L_1) are 12 and 20, respectively. In the loss function, the balance factors λ_1, λ_2, λ_3 and λ_4 are set as 1.0, 3.0, 0.5, 2.0, respectively. We equip our IA-BEV with both ResNet-50 (He et al. 2016) and ConvNeXt-base (Liu et al. 2022c) as image backbones. Input resolutions are rescaled to 256×704, 512×1408 and 640×1600 for comprehensive evaluation.

Comparison with State-of-the-art Methods

In this section, we train our IA-BEV with different configurations for 20 epochs using both CBGS (Zhu et al. 2019) and EMA techniques following prior works (Li et al. 2023a; Feng et al. 2023; Zhang et al. 2023). First, we compare the 3D object detection performances of our IA-BEV with state-of-the-art methods on both nuScenes *val* set and *test* set with different input resolutions and image backbones in Table 1 and Table 2. It can be observed that our IA-BEV consistently surpasses state-of-the-art methods across different configurations on both *val* and *test* sets. Furthermore, compared with OA-BEV (Chu et al. 2023) and MV2D (Wang et al. 2023) methods which also leverage 2D instance priors, our IA-BEV outperforms them by a clear margin. Besides, we also compare the depth estimation quality in Table 3. With the help of instance awareness, our IA-BEV evidently enhances the depth estimation quality, which is key to the effectiveness of our method.

SPM	SBL	mAP ↑	mATE ↓	mAVE ↓	NDS ↑
		0.330	0.700	0.552	0.425
✓		0.345	0.671	0.521	0.443
	✓	0.354	0.667	0.512	0.446
✓	✓	**0.367**	**0.658**	**0.486**	**0.461**

Table 4: Ablation study of the proposed SPM and SBL.

#	PD	CD	$\mathcal{L}^{abs}_{depth}$	$\mathcal{L}^{rel}_{depth}$	mAP	mATE	NDS
1					0.330	0.700	0.425
2	✓				0.329	0.697	0.426
3		✓			0.336	0.689	0.431
4		✓	✓		0.340	0.691	0.439
5		✓	✓	✓	**0.345**	**0.671**	**0.443**

Table 5: Ablation study of designs in SPM. "PD" and "CD" are short for "Parallel Decoders" and "Category-specific decoders". Parallel decoders have the same model structure as category-specific decoders, but take all objects as inputs for each branch.

Iter_num	Memory	mAP ↑	mATE ↓	NDS ↑
0	4.56G	0.330	0.700	0.425
1	4.83G	0.346	0.676	0.439
2	4.84G	0.354	0.667	0.446
3	4.85G	**0.356**	**0.655**	**0.447**

Table 6: Memory cost and performance comparison of iterating different rounds in SBL. The memory costs are based on our reproduction, which are similar to the memory costs measured in SOLOFusion (Park et al. 2022).

Comprehensive Analysis

In this section, all models are trained for 24 epochs without using CBGS or EMA techniques for efficient evaluation. The baseline method is the BEVDepth (Li et al. 2023a) without using its "Depth Refinement" module.

Ablations of main components. We verify the effects of the proposed Structural Priors Mining (SPM) and Self-Boosting Learning (SBL) approaches as shown in Table 4. By introducing SPM or SBL alone can bring 1.8% and 2.1% NDS improvements over the baseline method, respectively. And combining them can further boost NDS from 42.5% to 46.1%. The significant performance improvements demonstrate the effectiveness of our designs.

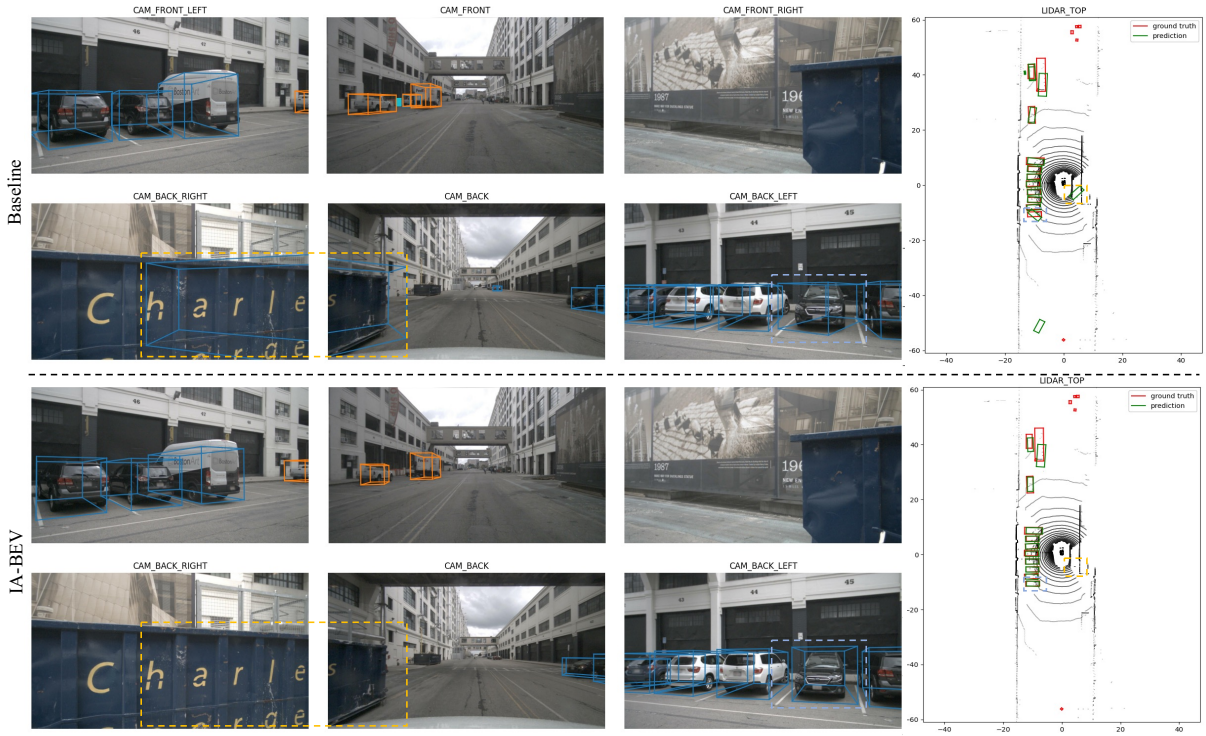

Figure 3: Qualitative results comparison between our IA-BEV and baseline method (i.e., BEVDepth). The dotted rectangles with the same color in image views and BEV planes represent the same regions.

Ablations of SPM designs. We further investigate the effects of each design in SPM. From the Table 5, we have the following two observations. First, simply increasing the number of depth decoders without semantic grouping (#1) can not bring benefits over the baseline (#2), while our category-specific depth decoder design (#3) boosts the performance, showcasing the effectiveness of capturing the instances' structural priors per category. Besides, on the basis of category-specific decoders, adding instance-level absolute depth supervision (#4) can bring 0.4% mAP and 0.8% NDS improvements. And by introducing inner-instance relative depth can achieve further enhancement, indicating that instance-level depth supervision can benefit effective monocular depth estimation.

Effects of SBL iteration rounds. To evaluate the effects of self-boosting learning (SBL) design, we compare results of different number of iterations in Table 6. Iterating 0 rounds stands for the baseline method without using the stereo-matching technique. Compared with only iterating 1 round, boosting unclear regions with the second round can significantly improve the performances while only slightly increase the memory, which demonstrates that our self-boosting learning can efficiently promote comprehensive stereo matching. However, increasing the iteration number to 3 does not bring a significant performance boost, which might be because of the limitation of the current resolution of perception and feature quality. Therefore, we only iterate for 2 rounds in practice.

Visualization

Qualitative results. We illustrate the detection results of the BEVDepth baseline and our IA-BEV in Fig 3. Benefitting from 2D instance awareness, our IA-BEV can predict more accurate results than the baseline. First, as shown in the yellow dotted rectangle, our IA-BEV can predict fewer false positives with 2D semantics as priors. Besides, as shown in the blue dotted rectangle, IA-BEV can also perceive the object's detailed structures more accurately than the baseline.

Conclusion

In this paper, we proposed IA-BEV, which enhances the depth estimation process for the multi-camera BEV-based detector by exploring the inherent properties encapsulated in foreground objects. Within IA-BEV, a Structural Priors Mining approach (SPM) and a Self-Boosting Learning strategy (SBL) are proposed to enhance the monocular and stereo depth estimations, respectively. Equipped with both SPM and SBL, IA-BEV sets new state-of-the-art performances across methods that use short-term (i.e., 2) frames with 54.5% mAP and 63.0% NDS on nuScenes.

Acknowledgements

This work was supported in part by National Natural Science Foundation of China Project (No. 62072116), Shanghai Science and Technology Program Project (No. 21JC1400600) and National Natural Science Foundation of China (Grant

No. 62106235),by the Zhejiang Provincial Natural Science Foundation of China (LQ21F020003).

References

Bhat, S. F.; Alhashim, I.; and Wonka, P. 2021. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4009–4018.

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

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

Chu, X.; Deng, J.; Zhao, Y.; Ji, J.; Zhang, Y.; Li, H.; and Zhang, Y. 2023. OA-BEV: Bringing Object Awareness to Bird's-Eye-View Representation for Multi-Camera 3D Object Detection. *arXiv preprint arXiv:2301.05711*.

Contributors, S. 2022. Spconv: Spatially Sparse Convolution Library. https://github.com/traveller59/spconv.

Feng, C.; Jie, Z.; Zhong, Y.; Chu, X.; and Ma, L. 2023. Aedet: Azimuth-invariant multi-view 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21580–21588.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141.

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

Jiao, Y.; Jie, Z.; Chen, S.; Chen, J.; Ma, L.; and Jiang, Y.-G. 2023. MSMDfusion: Fusing lidar and camera at multiple scales with multi-depth seeds for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21643–21652.

Li, Y.; Bao, H.; Ge, Z.; Yang, J.; Sun, J.; and Li, Z. 2022a. Bevstereo: Enhancing depth estimation in multi-view 3d object detection with dynamic temporal stereo. *arXiv preprint arXiv:2209.10248*.

Li, Y.; Ge, Z.; Yu, G.; Yang, J.; Wang, Z.; Shi, Y.; Sun, J.; and Li, Z. 2023a. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 1477–1485.

Li, Z.; Chen, Z.; Liu, X.; and Jiang, J. 2022b. Depthformer: Exploiting long-range correlation and local information for accurate monocular depth estimation. *arXiv preprint arXiv:2203.14211*.

Li, Z.; Lan, S.; Alvarez, J. M.; and Wu, Z. 2023b. BEVNeXt: Reviving Dense BEV Frameworks for 3D Object Detection. *arXiv preprint arXiv:2312.01696*.

Li, Z.; Wang, W.; Li, H.; Xie, E.; Sima, C.; Lu, T.; Qiao, Y.; and Dai, J. 2022c. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European conference on computer vision*, 1–18. Springer.

Li, Z.; Yu, Z.; Wang, W.; Anandkumar, A.; Lu, T.; and Alvarez, J. M. 2023c. FB-BEV: BEV Representation from Forward-Backward View Transformations. *arXiv preprint arXiv:2308.02236*.

Liu, Y.; Wang, T.; Zhang, X.; and Sun, J. 2022a. Petr: Position embedding transformation for multi-view 3d object detection. In *European Conference on Computer Vision*, 531–548. Springer.

Liu, Y.; Yan, J.; Jia, F.; Li, S.; Gao, A.; Wang, T.; Zhang, X.; and Sun, J. 2022b. Petrv2: A unified framework for 3d perception from multi-camera images. *arXiv preprint arXiv:2206.01256*.

Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; and Xie, S. 2022c. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11976–11986.

Park, J.; Xu, C.; Yang, S.; Keutzer, K.; Kitani, K.; Tomizuka, M.; and Zhan, W. 2022. Time will tell: New outlooks and a baseline for temporal multi-view 3d object detection. *arXiv preprint arXiv:2210.02443*.

Peng, R.; Wang, R.; Wang, Z.; Lai, Y.; and Wang, R. 2022. Rethinking depth estimation for multi-view stereo: A unified representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8645–8654.

Philion, J.; and Fidler, S. 2020. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, 194–210. Springer.

Poggi, M.; Aleotti, F.; Tosi, F.; and Mattoccia, S. 2020. On the uncertainty of self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3227–3237.

Qian, T.; Chen, J.; Zhuo, L.; Jiao, Y.; and Jiang, Y.-G. 2023. NuScenes-QA: A Multi-modal Visual Question Answering Benchmark for Autonomous Driving Scenario. *arXiv preprint arXiv:2305.14836*.

Shen, Z.; Dai, Y.; and Rao, Z. 2021. Cfnet: Cascade and fused cost volume for robust stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13906–13915.

Sun, J.; Chen, L.; Xie, Y.; Zhang, S.; Jiang, Q.; Zhou, X.; and Bao, H. 2020. Disp r-cnn: Stereo 3d object detection via shape prior guided instance disparity estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10548–10557.

Wang, Y.; Guizilini, V. C.; Zhang, T.; Wang, Y.; Zhao, H.; and Solomon, J. 2022a. Detr3d: 3d object detection from

multi-view images via 3d-to-2d queries. In *Conference on Robot Learning*, 180–191. PMLR.

Wang, Z.; Huang, Z.; Fu, J.; Wang, N.; and Liu, S. 2023. Object as query: Equipping any 2d object detector with 3d detection ability. *arXiv preprint arXiv:2301.02364*.

Wang, Z.; Min, C.; Ge, Z.; Li, Y.; Li, Z.; Yang, H.; and Huang, D. 2022b. Sts: Surround-view temporal stereo for multi-view 3d detection. *arXiv preprint arXiv:2208.10145*.

Wei, Z.; Zhu, Q.; Min, C.; Chen, Y.; and Wang, G. 2021. Aarmvsnet: Adaptive aggregation recurrent multi-view stereo network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6187–6196.

Xie, E.; Yu, Z.; Zhou, D.; Philion, J.; Anandkumar, A.; Fidler, S.; Luo, P.; and Alvarez, J. M. 2022. M2BEV: Multi-Camera Joint 3D Detection and Segmentation with Unified Birds-Eye View Representation. *arXiv preprint arXiv:2204.05088*.

Xiong, K.; Gong, S.; Ye, X.; Tan, X.; Wan, J.; Ding, E.; Wang, J.; and Bai, X. 2023. Cape: Camera view position embedding for multi-view 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21570–21579.

Yang, C.; Chen, Y.; Tian, H.; Tao, C.; Zhu, X.; Zhang, Z.; Huang, G.; Li, H.; Qiao, Y.; Lu, L.; et al. 2023. BEVFormer v2: Adapting Modern Image Backbones to Bird's-Eye-View Recognition via Perspective Supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17830–17839.

Yao, Y.; Luo, Z.; Li, S.; Fang, T.; and Quan, L. 2018. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European conference on computer vision (ECCV)*, 767–783.

Zhang, J.; Zhang, Y.; Liu, Q.; and Wang, Y. 2023. SA-BEV: Generating Semantic-Aware Bird's-Eye-View Feature for Multi-view 3D Object Detection. *arXiv preprint arXiv:2307.11477*.

Zhang, R.; Qiu, H.; Wang, T.; Guo, Z.; Xu, X.; Qiao, Y.; Gao, P.; and Li, H. 2022. MonoDETR: depth-guided transformer for monocular 3D object detection. *arXiv preprint arXiv:2203.13310*.

Zhao, L.; Zhou, H.; Zhu, X.; Song, X.; Li, H.; and Tao, W. 2023. Lif-seg: Lidar and camera image fusion for 3d lidar semantic segmentation. *IEEE Transactions on Multimedia*.

Zhou, X.; Koltun, V.; and Krähenbühl, P. 2021. Probabilistic two-stage detection. *arXiv preprint arXiv:2103.07461*.

Zhu, B.; Jiang, Z.; Zhou, X.; Li, Z.; and Yu, G. 2019. Class-balanced grouping and sampling for point cloud 3d object detection. *arXiv preprint arXiv:1908.09492*.