Calibration-free BEV Representation for Infrastructure Perception

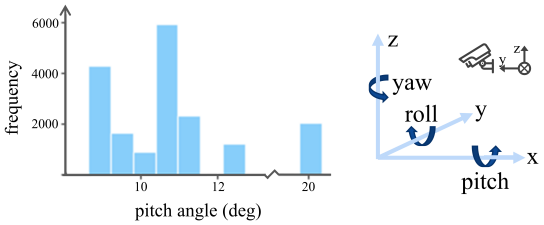
Siqi Fan[1] Zhe Wang[1] Xiaoliang Huo[1,2] Yan Wang*[1] Jingjing Liu[1]

***Abstract*— Effective BEV object detection on infrastructure can greatly improve traffic scenes understanding and vehicle-to-infrastructure (V2I) cooperative perception. However, cameras installed on infrastructure have various postures, and previous BEV detection methods rely on accurate calibration, which is difficult for practical applications due to inevitable natural factors (e.g., wind and snow). In this paper, we propose a Calibration-free BEV Representation (CBR) network, which achieves 3D detection based on BEV representation without calibration parameters and additional depth supervision. Specifically, we utilize two multi-layer perceptrons for decoupling the features from perspective view to front view and bird-eye view under boxes-induced foreground supervision. Then, a cross-view feature fusion module matches features from orthogonal views according to similarity and conducts BEV feature enhancement with front view features. Experimental results on DAIR-V2X demonstrate that CBR achieves acceptable performance without any camera parameters and is naturally not affected by calibration noises. We hope CBR can serve as a baseline for future research addressing practical challenges of infrastructure perception.**

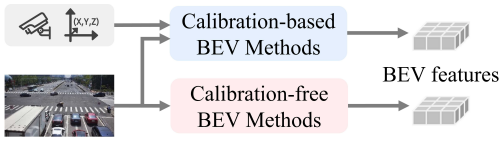
I. Introduction

3D object detection is one of the key enabling technologies for environment perception. Compared with LiDAR-based methods, vision-based methods are cost-effective and easy to implement. However, it is an ill-posed problem to recover 3D information from 2D image. Existing methods can be grouped into three categories according to when the 2D information is lifted to 3D, including data lifting-based methods, feature lifting-based methods, and result lifting-based methods [1]. Most of them are not designed specifically for infrastructure side. Different from typical applications, such as vehicle-side environment perception, object detection on infrastructure side has two main challenges: *1) computing resource is limited, 2) cameras are installed in various postures (Figure 1.a), and accurate calibration parameters are hard to obtain or dynamically correct due to natural factors, like wind and snow*.

Among the aforementioned categories, 2D images are directly transformed into pseudo 3D data (e.g., point cloud) and processed via LiDAR-based pipeline in data lifting-based methods [2], [3], [4], which are computational expensive for infrastructure. Result lifting-based methods [5], [6], [7], [8], [9] recover 3D information, including 3D locations and dimensions, based on 2D perspective view features and fully leverage the advantages of 2D detection pipelines. Although they can basically address the aforementioned challenges, the features in perspective view hinder the further development



(a) The distribution of pitch angles in DAIR-V2X dataset.



(b) Calibration-based and calibration-free BEV methods.

Fig. 1. Cameras are installed in various postures, especially various pitch angles, on infrastructures (data is from DAIR-V2X dataset [18]). Compared with previous calibration-based methods, our approach is calibration-free, which is naturally not constrained by calibration accuracy.

of V2I cooperative perception researches, which only enables result-level fusion. Considering feature-level fusion, feature lifting-based methods [10], [11], [12], [13], [14], [15] first transform 2D image feature into 3D voxel feature and collapse them to generate BEV features. BEV features from different agents, time series, and modalities can be fused in a physics-interpretable manner [16], and 3D detection based on BEV representations has attracted immense attention in recent years. However, these methods rely on accurate camera calibration (i.e. intrinsic and extrinsic parameters) and/or additional depth supervision to assist cross-view feature projection, which are not suitable for infrastructure-side perception because of unavoidable calibration noise, and their performance would be significantly degraded if using noisy parameters. PYVA [17] generated BEV representations via multi-layer perceptrons (MLPs) without camera parameters for road scene layout estimation on vehicle-side, but the performance is far from satisfaction when migrated to detection task on infrastructure side.

To address the practical challenges of infrastructure-side perception, we propose a Calibration-free BEV Representation (CBR) network, which is naturally not constrained by calibration accuracy (Figure 1.b). Specifically, we utilize light image backbone ResNet-18 to extract perspective view feature. In the face of various camera postures, two MLPs are used for view decoupling from perspective view to front and bird-eye view. The view transformation is supervised by boxes-induced foreground segmentation labels generated

* Corresponding author: wangyan@air.tsinghua.edu.cn;
[1] Institute for AI Industry Research, Tsinghua University;
[2] Beihang University;

from 3D bounding box labels, without additional labeling cost. To compensate information loss of BEV features along height dimension (z-axis), a cross-view feature fusion module is proposed for BEV feature enhancement using front view features. Assuming that same object should have similar features under different views, features from orthogonal views are fused according to similarity distribution.

Our contributions are summarized as follows:

- We point out the practical challenges of infrastructure-side perception, and propose the Calibration-free BEV Representation network, CBR, to address various install postures and calibration noise.
- The perspective view features are decoupled to front view and bird-eye view via MLPs without any calibration parameters, and orthogonal feature fusion is similarity-based without additional depth supervision.
- Experimental results demonstrate that CBR achieves acceptable detection performance based on BEV representation on large-scale real-world dataset DAIR-V2X and can output BEV foreground segmentation predictions at the meantime.

II. Related Work

In this section, we briefly review two related topics, BEV object detection and BEV representation generation.

A. Image-based BEV Object Detection

BEV object detection methods have attracted more attention recently, and have made great progress in performance. However, most of them are designed for typical vehicle-side perception and not suitable for infrastructure side. Depth-based methods [10], [11], [19] infer depth to recover 3D information along y-axis of BEV coordinate system, but the depth in image view on infrastructure is the compound information along y-axis and z-axis due to the pitch angle, which cannot be directly utilized for BEV detection. Projection-based methods [14], [20] are not affected by camera postures, since the features are projected to 3D according to calibration parameters before fed into detection heads, nevertheless, their performance highly relies on calibration accuracy. Transformer-based methods [21], [13], [12], [22] achieve better performance with higher computational cost, and calibration parameters may also be needed for attention guidance. To address the practical challenges of infrastructure-side perception and get rid of the dependence on calibration accuracy, we propose CBR to achieve 3D perception in a calibration-free manner.

B. BEV Representation Generation

With the advantage of succinct and physics-interpretable, BEV representations are deployed in more and more downstream real-world applications, especially for traffic scenes. Despite the aforementioned approaches adopted in detection tasks, how to generate BEV representation from image is also well-studied in segmentation researches, consisting of geometry-based (homograph or depth) and learning-based (MLP or transformer) approaches [16]. Homograph-based methods [23] realize view projection relying on physical mapping under horizontal plane constraint. Depth-based methods [20], [19], [24], [11], [10] explicitly leverage depth distribution to lift 2D features to 3D space (e.g. voxel and points cloud), and depth supervision is an essential cue to them. Learning-based approaches ignore the geometric priors from calibrations. MLP-based methods [25], [26], [17] model the transformation via the global mapping capability of MLP. On account of strong modeling ability, transformer-based methods [13], [12], [22] are further developed recently. It would be a considerable option for devices with sufficient computing resources. In this paper, MLP is used for view decoupling, and similarity-based cross-view fusion is proposed taking inspiration from depth-based methods.

III. CBR Framework

In this section, we describe the proposed calibration-free BEV representation network, which mainly consists of feature view decoupling module and similarity-based cross-view fusion module.

A. Overall Architecture

Addressing practical challenges, CBR achieves feature view standardization via decoupled feature reconstruction (Figure 2). Images captured from infrastructure are fed into an image backbone to extract perspective view features. With the consideration of limited computing resources, ResNet-18 is employed. The feature maps are further processed by a convolution operation with a filter in the size of 3×3 and a mean pooling operation to save the computational cost of the view decoupling. Image scale is gradually decreased from $H \times W$ to $\frac{H}{64} \times \frac{W}{64}$, while the channel size is increased to 1024. Taking the advantage of global receptive field, the perspective view feature f_{pv} is spatially decoupled to two orthogonal views via FVD (feature view decoupling) module. Next, a SCF (similarity-based cross-view fusion) module is used to match the features from different views and generate enhanced BEV features f_e leveraging front view features f_{fv} and bird-eye view features f_{bev}. Finally, f_e is fed to 4 detection heads for classification and regression tasks. Each detection head is composed of a basic convolution block, including convolution, batch normalization, and RELU.

B. Feature View Decoupling

In real-world scenes, cameras installed on infrastructure side usually have various postures, including x-y-z location and pitch-yaw-roll orientation. Compared with location, orientation, especially pitch, will directly affect the perspective view features. To generate unified representations despite of various orientation, we propose the FVD module for feature view decoupling. Since the features of different views are not spatially aligned, MLP can better facilitate the view decoupling compared with convolution operation. The MLP structure consisted of two fully connected layers is deployed following the practice of previous works [27], [17]. The decoupled features are fed to four consecutive decoder layers, and we utilize nearest interpolation for upsampling from

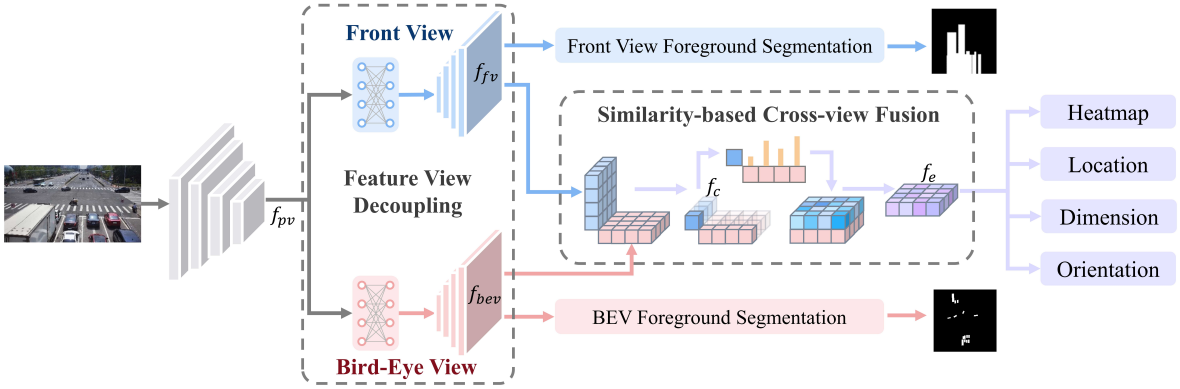


Fig. 2. Framework of the Calibration-free BEV Representation (CBR) network.

$\frac{H}{64} \times \frac{W}{64}$ to $\frac{H}{4} \times \frac{W}{4}$. After that, the front view features f_{fv} and BEV features f_{bev} are obtained:

$$f_{fv} = \Phi_{fv}(MLP(f_{pv})) \tag{1}$$

$$f_{bev} = \Phi_{bev}(MLP(f_{pv})) \tag{2}$$

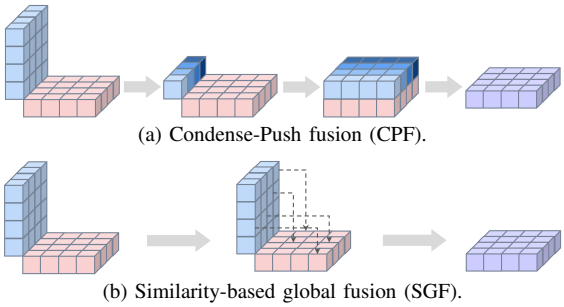
where '$\Phi(\cdot)$' denotes the feature decoding operation, and f_{pv} is the extracted features in perspective view.

To guide the view decoupling without using calibration parameters, f_{fv} and f_{bev} are further input to the corresponding foreground segmentation heads (composed of a basic convolution block), respectively. The segmentation prediction is under the boxes-induced foreground supervision, which is generated by projecting the 3D bounding boxes to front/bird-eye view, without additional labeling cost. The benefits of the foreground segmentation supervision are two-fold. On the one hand, the pixel-level supervision can effectively guide the view transformation and encourage the module focus on foreground objects (e.g., cars). On the other hand, the BEV foreground segmentation predictions are output as by-product, which indicates the dynamic foreground layout of the traffic scenes.

C. Similarity-based Cross-view Fusion

BEV features can effectively represent the foreground layout in bird-eye view. However, 3D detection performance based on that will naturally be influenced by the information loss along z-axis, especially when the view projection is not accurately guided with calibration parameters. Therefore, it is necessary to enhance the BEV representations with features in the front view, and the main difficulty is matching the corresponding features across orthogonal views.

There are two heuristic options, as shown in Figure 3. Assuming that the feature of the same object in different views should be similar, SGF (similarity-based global fusion) can match features globally according to similarity, but it is computational expensive. To reduce the searching space of feature matching between two views, CPF (condense-push fusion) first condenses f_{fv} along z-axis and then pushes the obtained f_c along y-axis making use of geometric constrains.



(a) Condense-Push fusion (CPF).

(b) Similarity-based global fusion (SGF).

Fig. 3. Illustration of condense-push fusion (CPF) and similarity-based global fusion (SGF).

To embrace the advantages of both CPF and SGF, we design SCF (similarity-based cross-view fusion) module, which matches the features based on similarity with the geometric constrains (Figure 2). Specifically, we only take the similarity among the features with the same x-axis value into consideration. To reduce computational cost, we utilize the condensed feature $f_c = Avg(f_{fv})$ for feature fusion, where '$Avg(\cdot)$' denotes mean pooling operation along z-axis. The similarity s_{ij} is measured by the inner-product:

$$s_{ij} = \langle f_{c_i}, f_{bev_{ij}} \rangle \tag{3}$$

where '$\langle \cdot, \cdot \rangle$' denotes inner-product operation, and i is the index of x-axis, j is the index along y-axis. The calculated similarity is used as fusion weights to enhance BEV feature f_{bev} with condensed front view feature f_c:

$$f_e = Conv(Concat(f_{bev}, s \cdot f_c)) \tag{4}$$

where '$Conv(\cdot)$' and '$Concat(\cdot)$' denote convolution and concatenation operations.

It can be presumed that the feature similarity distribution across orthogonal views along y-axis is implicit depth distribution, since the closer to the real depth, the more similar the features across views are. SCF bridges the cross-view features without additional depth supervision. Moreover, the similarity-based fusion indirectly facilitates the spatial-wise alignment across different views, as the corresponding features are encouraged to be in the same x column.

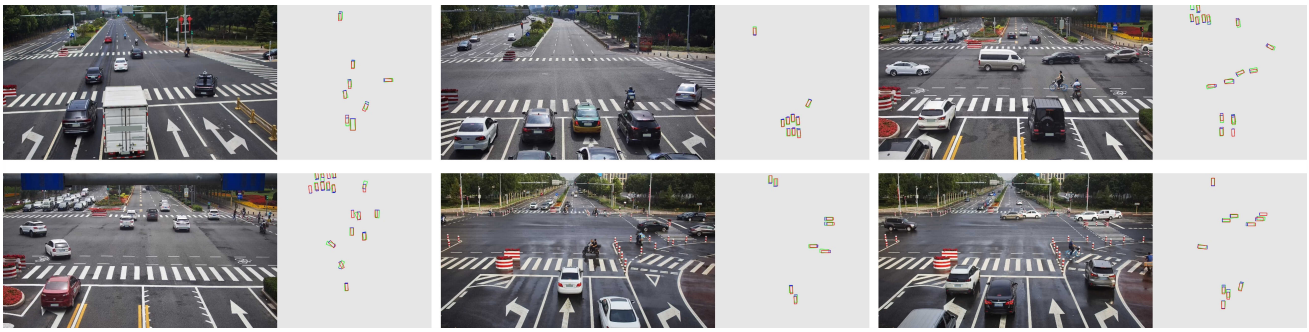


Fig. 4. Visualization examples. Red: groundtruth. Green: predictions of CBR. Blue line indicates the head of vehicle.

TABLE I

QUANTITATIVE EVALUATION ON DAIR-V2X DATASET WITH CALIBRATION NOISE ON ROTATION ANGLES. THE PERFORMANCE IS SIGNIFICANTLY DEGRADED WITH NOISY CALIBRATION PARAMETERS, WHILE OUR APPROACH IS NOT INFLUENCED. ALL SCORES ARE IN %.

Methods	**Calib. Noise (deg)**	$AP_{3D\|R40}$ (IoU=0.5)			$AP_{3D\|R40}$ (IoU=0.7)			$AP_{BEV\|R40}$ (IoU=0.5)			$AP_{BEV\|R40}$ (IoU=0.7)		
		easy	**mod.**	**hard**	**easy**	**mod.**	**hard**	**easy**	**mod.**	**hard**	**easy**	**mod.**	**hard**
ImVoxelNet [14]	/	47.6	29.2	27.1	27.1	16.2	14.8	51.9	32.7	30.4	35.4	20.5	20.1
	0.1	44.5	26.5	26.2	24.2	13.9	12.5	50.9	32.0	29.9	32.1	19.0	17.5
	0.2	38.6	23.1	22.6	21.0	11.5	11.2	45.1	26.8	26.4	27.6	16.4	15.0
	0.5	29.3	16.9	15.4	12.9	6.8	6.5	35.0	20.1	19.7	19.1	10.9	9.9
	1.0	19.7	11.4	10.2	5.0	2.4	2.4	25.5	14.7	14.3	9.6	5.3	4.7
	2.0	8.2	4.4	4.3	1.0	0.5	0.5	13.6	7.2	7.0	2.2	1.0	1.0
	5.0	0.6	0.3	0.3	0.0	0.0	0.0	1.4	0.7	0.7	0.1	0.1	0.1
PYVA-det	calibration-free	12.6	7.3	7.1	0.9	0.6	0.5	23.3	14.0	13.6	5.5	2.9	2.9
CBR (Ours)	calibration-free	24.7	15.7	14.7	1.3	0.8	0.8	40.0	24.9	24.5	4.9	3.2	3.2

IV. EXPERIMENTS

This section describes experiments on real-world infrastructure detection dataset. We compare our model with other typical BEV detection methods in noisy calibration setting, and provide detailed ablation study and error analysis. Further evaluation on BEV foreground segmentation also validates the scene layout understanding capacity of CBR.

A. Experimental Setting

Datasets We evaluate the proposed CBR model on the large-scale real-world cooperative perception dataset DAIR-V2X [18]. It provides 12,424 images captured from diverse infrastructure-side cameras with 3D annotations, which comprises 8800 images for training and 3624 images for validation. We follow the official split scheme and report experimental results on validation set. All of the objects inside the camera view are labeled, and the perception range of our method is set as $90m \times 90m \times 5m$. The input images are resized to a fixed size of 1024×1024.

Foreground Segmentation Label Generation To generate foreground segmentation labels in orthogonal views, each bounding box of labeled objects in perception range is projected to bird-eye and front view. The generated pixel-level groundtruth is with the size of 256×256.

Calibration Noise To simulate the natural calibration noise in practical environments, we introduce several levels of Gaussian noise to rotation angles

$$\theta_n = x_n * n_{range} \tag{5}$$

where $x_n \sim N(\mu, \sigma^2)$, $\mu = 0, \sigma = \frac{1}{3}$, and $n_{range} \in \{0.1, 0.2, 0.5, 1.0, 2.0, 5.0\}$ denotes the noise level in degree.

Baselines We compare CBR with both calibration-based and calibration-free BEV methods. ImVoxelNet [14] is a typical projection-based detection method, which projects features from perspective view to BEV with the guidance of calibration parameters. PYVA [17] is originally proposed for calibration-free segmentation. We develop PYVA-det based on that by adopting additional detection heads.

Implementation Details The image backbone, ResNet 18, is initialized with pre-trained weights provided by PyTorch. The initial learning rate is set to 2×10^{-4}. We use adam optimizer for training with batch size of 6, and the network is trained until convergence (200 epochs). Random flipping and random color jitter are applied while training. Experiments are implemented in PyTorch on a server with NVIDIA A30.

B. Main Results

Table I reports the comparison results between CBR and baselines with calibration noise. The performance with the IoU threshold of 0.5 is regarded as major concern. CBR achieves 15.7% and 24.9% mAP on 3D detection and BEV detection tasks for moderate difficulty. Some visualization examples are shown in Figure 4.

It can be seen that ImVoxelNet [14] performs better leveraging accurate calibration parameters. However, the performance is significantly degraded while noise raises, almost cut in half if small rotation noise (within 0.5 degree) is introduced randomly. Different from that, calibration-free methods are naturally not affected by noisy calibration parameters, and show the superiority in noisy cases, visually shown in Figure 5. Our CBR has performance advantage

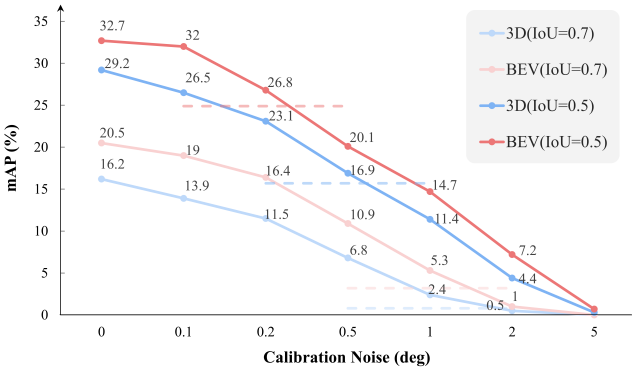


Fig. 5. Comparison between calibration-based method [14] (solid line) and our CBR (dotted line) with calibration noise under moderate difficulty. CBR is naturally not affected by noisy rotation angles.

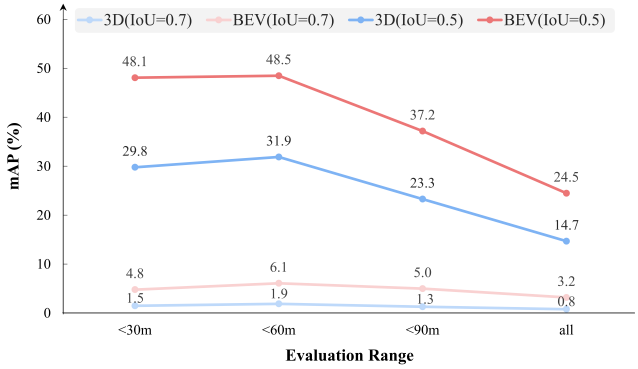


Fig. 6. Error analysis: evaluation with distance.

on BEV detection, when n_{range} is greater than 0.2 degree. The watershed on 3D detection task is 0.5 degree. We also report experimental results with a more strict IoU threshold, $IoU = 0.7$, and performance trend in noisy case is in line with that of $IoU = 0.5$. Since calibration noise on infrastructure side is unavoidable due to the complex natural factors (e.g. wind and snow), calibration-free methods are more robust under various scenes.

Although PYVA-det is not limited by calibration accuracy either, the performances on 3D/BEV tasks are far from satisfaction. Our approach achieves better accuracy-robustness balance for infrastructure perception.

C. Ablation Study

We conduct the following experiments to study the impact of different cross-view feature fusion methods. The performance comparisons are summarized in Table II, and the baseline is vanilla BEV representation, which is directly obtained via feature view decoupling without cross-view feature fusion. Benefiting from similarity-based fusion, SGF effectively leads to an improvement of 0.9% on BEV detection, but the performance growth on 3D detection is limited. CPF performs better than SGF leveraging the front view features with geometric constrains, and the performance on 3D task is lifted to 13.6%. Embracing advantages of geometry and similarity, SCF further boosts the performances on both tasks, and achieves 15.7% and 24.9% respectively. The performance advantage of SCF is obvious compared with SGF and CPF.

TABLE II
ABLATION STUDY ON CROSS-VIEW FEATURE FUSION.

Ablated	$AP_{3D\|R40}$ (IoU 0.5)			$AP_{BEV\|R40}$ (IoU 0.5)		
	easy	**mod.**	**hard**	**easy**	**mod.**	**hard**
Vanilla-BEV	20.2	12.6	11.7	36.8	22.2	20.7
SGF	21.8	13.0	12.8	37.3	23.1	22.7
CPF	21.8	13.6	13.3	38.9	23.5	23.1
SCF	24.7	15.7	14.7	40.0	24.9	24.5

D. Error Analysis

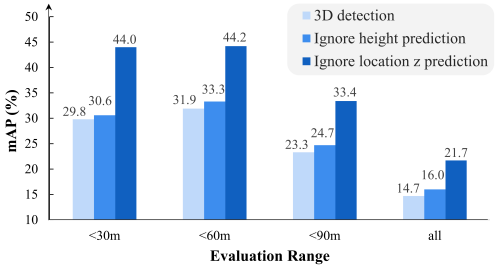
To better understand CBR, we further conduct several groups of experiments for error analysis, and the limitations of calibration-free BEV representation are discussed.

Evaluation with distance. The following experiments are conducted to analysis the performance under different perception range (Figure 6). The infrastructure-side perception region is split to four parts with three thresholds of 30m, 60m, and 90m. It can be seen that the performance degradation is inevitable along with distance increasement. The performance is almost doubled if we only take the objects within the range of 60m into consideration. Note that the performance within 60m is slightly better than that within 30m. We think this increment is caused by the amount of information of vehicles in different views. Specifically, the object within 30m captured from infrastructure is almost in a top view, while it tends to be in side view when extended to 60m. Intuitively, the side view is more informative than top view. In addition, the decline is obvious out of the range of 90m, which is the designed representation range of our BEV feature. Therefore, detection capability is limited by the the manually set perception range, and the objects lie out of that are theoretically ignored.

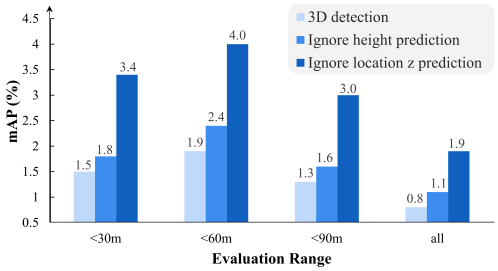
Error sources of 3D detection. Compared with BEV detection, performance on 3D detection is worse since the additional prediction along z-axis, including location-z and height. To analysis the major source of error, we evaluated the ablated predictions by ignoring either location-z or height predictions. As shown in Figure 7, the score increment is more obvious if ignoring predictions of location-z rather than height predictions, regardless of the IoU threshold (0.5 or 0.7), which indicates location-z prediction is the major error source on 3D detection. It is difficult to estimate the location along z axis for infrastructure perception due to the various installation height of cameras, especially without the reference of calibration parameters.

E. Evaluation on BEV Foreground Segmentation

We further evaluate CBR on BEV foreground segmentation, which is a by-product of feature view decoupling (Table III). The performance of PYVA [17] is declined when



(a) 3D detection performance (IoU 0.5).



(b) 3D detection performance (IoU 0.7).

Fig. 7. Error analysis: major error source of 3D detection.

additional detection heads are introduced. Although CBR is slightly worse than PYVA, it is still better than PYVA-det.

TABLE III

QUANTITATIVE EVALUATION ON BEV FOREGROUND SEGMENTATION.

Methods	mIoU (%)	mAP (%)
PYVA [17]	42.3	56.0
PYVA-det	33.9	43.5
CBR (Ours)	40.3	50.4

V. CONCLUSION

Addressing the practical challenges of various installation postures and calibration noises, we point out the significant performance degradation of calibration-based BEV detection approach under calibration noise, and propose a calibration-free BEV representation for infrastructure perception in this paper. The extracted image features are decoupled to two orthogonal views, and BEV representations are enhanced via similarity-based cross-view fusion. Extensive experiments on real-world dataset demonstrate that CBR achieves a better accuracy-robustness balance. In addition, error analysis are reported, and limitations of the proposed calibration-free BEV representations are further discussed. In future work, the way to utilize partial stable calibration parameters to improve perception performance deserves to be studied, and how to leverage multi-view images for adaptive camera re-calibration is also worth to be further explored.

APPENDIX: MORE EXPERIMENTAL RESULTS

CBR is proposed for practical infrastructure perception to facilitate the development of V2I cooperative perception. The experiments in Sec.IV are conducted on DAIR-V2X-C to provide an infrastructure-side baseline for VIC3D task, and we further report more experimental results on DAIR-V2X-I to make comparison with more recent approaches.

Datasets Different from the main dataset DAIR-V2X-C for V2I cooperative perception, DAIR-V2X-I only contains infrastructure-side data. It includes around 10 thousand frames and is divided into train/val/test (50/20/30) subsets. We evaluate CBR on validation set following [28].

Baselines We compare our CBR with other SOTA camera-based methods like ImVoxelNet [14], BEVFormer [12], BEVDepth [11], and BEVHeight [28]. In addition, some LiDAR-based and multimodal-based methods are also reported for reference, including PointPillars [29], SECOND [30], and MVXNet [31].

Experimental results on DAIR-V2X-I We report both 3D and BEV perception performance under two IoU threshold in Table IV. CBR achieves 60.1% $AP_{3D|R40}$ and 64.5% $AP_{BEV|R40}$ for moderate difficulty, which demonstrates the effectiveness of our method in practical application. Besides, it also achieves 57.9% mIoU and 60.2% mAP on BEV foreground segmentation. Some visualization examples are shown in Figure 8.

TABLE IV

EXPERIMENTAL RESULTS ON DAIR-V2X-I.

Task	IoU=0.5			IoU=0.7		
	easy	mod.	hard	easy	mod.	hard
$AP_{3D\|R40}$	72.0	60.1	60.1	38.5	31.6	31.7
$AP_{BEV\|R40}$	78.7	64.5	64.6	56.5	46.1	46.2

Comparison on DAIR-V2X-I benchmark We compare CBR with recent methods on the original benchmark, as shown in Table V. The experimental results of others are from [28], and all of them are calibration-based. It can be seen that CBR outperforms five of them without the limitation of accurate extrinsic parameters.

TABLE V

COMPARISON ON DAIR-V2X-I BENCHMARK. THE RESULTS OF OTHERS ARE FROM [28]. 'L' AND 'C' DENOTE LIDAR AND CAMERA.

Method	Modality	$AP_{3D\|R40}$ (IoU=0.5)		
		easy	mod.	hard
PointPillars [29]	L	63.1	54.0	54.0
SECOND [30]	L	71.5	54.0	54.0
MVXNet [31]	LC	71.0	53.7	53.8
ImVoxelNet [14]	C	44.8	37.6	37.6
BEVFormer [12]	C	61.4	50.7	50.7
BEVDepth [11]	C	75.5	63.6	63.7
BEVHeight [28]	C	77.8	65.8	65.9
CBR (Ours)	C	72.0	60.1	60.1

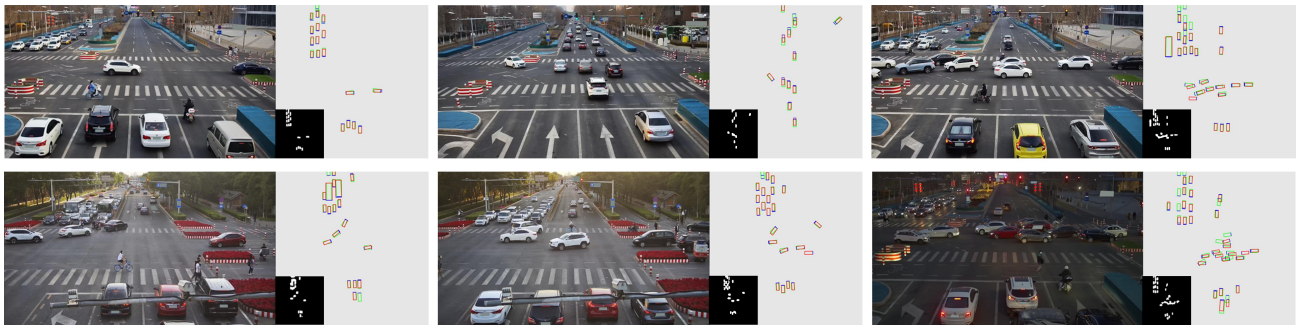


Fig. 8. Visualization examples on DAIR-V2X-I. Red: groundtruth. Green: predictions of CBR. Blue line indicates the head of vehicle.

REFERENCES

[1] X. Ma, W. Ouyang, A. Simonelli, and E. Ricci, "3D object detection from images for autonomous driving: a survey," *arXiv:2202.02980*, 2022.

[2] X. Ma, Z. Wang, H. Li, P. Zhang, W. Ouyang, and X. Fan, "Accurate monocular 3d object detection via color-embedded 3d reconstruction for autonomous driving," in *ICCV*, 2019, pp. 6851–6860.

[3] Y. Wang, W.-L. Chao, D. Garg, B. Hariharan, M. Campbell, and K. Q. Weinberger, "Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving," in *CVPR*, 2019, pp. 8445–8453.

[4] X. Weng and K. Kitani, "Monocular 3d object detection with pseudo-lidar point cloud," in *ICCVW*, 2019, pp. 857–866.

[5] R. Zhang, H. Qiu, T. Wang, X. Xu, Z. Guo, Y. Qiao, P. Gao, and H. Li, "MonoDETR: Depth-aware transformer for monocular 3d object detection," *arXiv:2203.13310*, 2022.

[6] Y. Zhang, J. Lu, and J. Zhou, "Objects are different: Flexible monocular 3d object detection," in *CVPR*, 2021, pp. 3289–3298.

[7] Y. Chen, L. Tai, K. Sun, and M. Li, "Monopair: Monocular 3d object detection using pairwise spatial relationships," in *CVPR*, 2020, pp. 12 093–12 102.

[8] Z. Liu, Z. Wu, and R. Tóth, "Smoke: Single-stage monocular 3d object detection via keypoint estimation," in *CVPRW*, 2020, pp. 996–997.

[9] X. Chen, K. Kundu, Y. Zhu, A. G. Berneshawi, H. Ma, S. Fidler, and R. Urtasun, "3d object proposals for accurate object class detection," *NIPS*, vol. 424-432, 2015.

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

[11] Y. Li, Z. Ge, G. Yu, J. Yang, Z. Wang, Y. Shi, J. Sun, and Z. Li, "BEVDepth: Acquisition of reliable depth for multi-view 3d object detection," *arXiv preprint arXiv:2206.10092*, 2022.

[12] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Q. Yu, and J. Dai, "BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," *arXiv preprint arXiv:2203.17270*, 2022.

[13] Y. Liu, T. Wang, X. Zhang, and J. Sun, "PETR: Position embedding transformation for multi-view 3d object detection," *arXiv preprint arXiv:2203.05625*, 2022.

[14] D. Rukhovich, A. Vorontsova, and A. Konushin, "ImVoxelNet: Image to voxels projection for monocular and multi-view general-purpose 3d object detection," in *WACV*, 2022, pp. 2397–2406.

[15] C. Reading, A. Harakeh, J. Chae, and S. L. Waslander, "Categorical depth distribution network for monocular 3d object detection," in *CVPR*, 2021, pp. 8555–8564.

[16] Y. Ma, T. Wang, X. Bai, H. Yang, Y. Hou, Y. Wang, Y. Qiao, R. Yang, D. Manocha, and X. Zhu, "Vision-centric bev perception: A survey," *arXiv:2208.02797*, 2022.

[17] W. Yang, Q. Li, W. Liu, Y. Yu, Y. Ma, S. He, and J. Pan, "Projecting your view attentively: Monocular road scene layout estimation via cross-view transformation," in *CVPR*, 2021, pp. 15 536–15 545.

[18] H. Yu, Y. Luo, M. Shu, Y. Huo, Z. Yang, Y. Shi, Z. Guo, H. Li, X. Hu, J. Yuan, and Z. Nie, "DAIR-V2X: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection," in *CVPR*, 2022, pp. 21 361–21 370.

[19] C. Reading, A. Harakeh, J. Chae, and S. L. Waslander, "Categorical depth distribution network for monocular 3d object detection," in *CVPR*, 2021, pp. 8555–8564.

[20] T. Roddick, A. Kendall, and R. Cipolla, "Orthographic feature transform for monocular 3d object detection," in *BMVC*, 2020, pp. 1–1.

[21] A. Saha, O. Mendez, C. Russell, and R. Bowden, "Translating images into maps," in *ICRA*, 2022, pp. 9200–9206.

[22] C. Yang, Y. Chen, H. Tian, C. Tao, X. Zhu, Z. Zhang, G. Huang, H. Li, Y. Qiao, L. Lu, J. Zhou, and J. Dai, "BEVFormer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision," *arXiv preprint arXiv:2211.10439*, 2022.

[23] H. A. Mallot, H. H. Bülthoff, J. Little, and S. Bohrer, "Inverse perspective mapping simplifies optical flow computation and obstacle detection," *Biological cybernetics*, vol. 64, no. 3, pp. 177–185, 1991.

[24] J. Philion and S. Fidler, "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d," in *ECCV*, 2020, pp. 194–210.

[25] B. Pan, J. Sun, H. Y. T. Leung, A. Andonian, and B. Zhou, "Cross-view semantic segmentation for sensing surroundings," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 4867–4873, 2020.

[26] T. Roddick and R. Cipolla, "Predicting semantic map representations from images using pyramid occupancy networks," in *CVPR*, 2020, pp. 11 138–11 147.

[27] B. Pan, J. Sun, H. Y. T. Leung, A. Andonian, and B. Zhou, "Cross-view semantic segmentation for sensing surroundings," *IEEE RAL*, vol. 5, no. 3, pp. 4867–4873, 2020.

[28] L. Yang, K. Yu, T. Tang, J. Li, K. Yuan, L. Wang, X. Zhang, and P. Chen, "BEVHeight: A robust framework for vision-based roadside 3d object detection," *arXiv preprint arXiv:2303.08498*, 2023.

[29] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *CVPR*, 2019, pp. 12 697–12 705.

[30] Y. Yan, Y. Mao, and B. Li, "Second: Sparsely embedded convolutional detection," *Sensors*, vol. 18, no. 10, p. 3337, 2018.

[31] V. A. Sindagi, Y. Zhou, and O. Tuzel, "Mvxnet: Multimodal voxelnet for 3d object detection," in *ICRA*, 2019, pp. 7276–7282.