2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Robust Multi-Camera BEV Perception: An Image-Perceptive Approach to Counter Imprecise Camera Calibration

Rundong Sun[1,2], Mengyin Fu[1,2,3], Hao Liang[1,2], Chunhui Zhu[1,2], Zhipeng Dong[1,2] and Yi Yang*[,1,2]

***Abstract*— Recently, Bird's Eye View (BEV) detection methodologies that utilize surround-view cameras have seen significant advancements in autonomous driving systems. Traditional methods, however, are constrained by their reliance on specific camera parameters, which poses challenges in generalizing across different vehicle-mounted cameras with varying poses and under adverse conditions. To address these challenges, we propose a robust BEV representation network that integrates Dual-Space Positional Encoding (DSPE) and image perception. This network is designed to enhance resilience to calibration errors and pose fluctuations, resulting in reliable detection performance on the Nuscenes dataset, even with imprecise extrinsic inputs. Our approach demonstrates competitive accuracy when compared to other methods that do not rely on temporal data, highlighting the effectiveness of our DSPE strategy in improving the robustness and accuracy of BEV detection in dynamic and challenging environments.**

I. Introduction

Detecting 3D objects within three-dimensional spaces is a crucial aspect of scene perception in autonomous driving systems. While LiDAR-based methods [1], [2] and multi-modal approaches [3]–[7] have their merits, camera-based techniques have increasingly attracted researchers due to their cost-effectiveness and high data density. Early methods largely employed monocular camera architectures [8]–[11], using post-processing in the global 3D space for basic fusion. Although straightforward, these methods often failed to effectively merge information from multiple views, leading to inferior performance. Perception in Birds-Eye-View (BEV) has garnered increasing interest due to its unified representation of 3D position and scale, as well as its suitability for downstream tasks such as motion planning. Recent transformer-based techniques [12] that comprehensively assimilate multi-view data and consistently represent it from a BEV have shown marked superiority over traditional methods.

Camera-based BEV perception methods primarily predict 3D geometric targets using the features extracted from camera-captured images, with the key challenge lying in learning the transformation between 2D views and 3D space. Existing methods can be broadly categorized into two types:

This work was partly supported by the National Natural Science Foundation of China under Grant 62233002, the National Key Research and Development Program of China under Grant 2022YFC2603602, and the Fundamental Research Funds for the Central Universities.

*Corresponding author: Yi Yang (email: yang_yi@bit.edu.cn)
[1]School of Automation, Beijing Institute of Technology, Beijing, China
[2]National Key Lab of Autonomous Intelligent Unmanned Systems, Beijing Institute of Technology, Beijing, China
[3]School of Automation, Nanjing University of Science and Technology, Nanjing, China

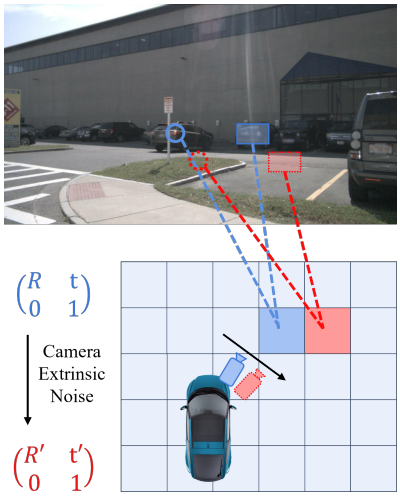

Fig. 1: Illustration of BEV perception under noisy camera extrinsics. Arrows and different color are employed to visualize the effects of noise in camera extrinsic parameters. The figure contrasts two scenarios: blue represents the perception under noise-free camera extrinsics, resulting in an accurate BEV representation, while red signifies the perception with distortions introduced by inaccurate camera extrinsics.

explicit BEV representation methods and implicit BEV representation methods. The former constructs an explicit BEV feature map by lifting 2D features into 3D space using camera parameters [13]–[15]. The latter implicitly associates 2D features with 3D space in an end-to-end manner by treating camera parameters as positional embeddings [16], [17].

However, these techniques are greatly reliant on precise camera calibration, encompassing both intrinsic and extrinsic parameters. Calibration noise or camera pose changes, as a vehicle operates, often diminishes the accuracy of the extrinsic parameters. This reduced precision can introduce errors in the BEV features, affecting overall performance. Unfortunately, this problem is quite common. Prior research has highlighted challenges such as imprecise camera parameters [14] and camera dropouts [17]. Given the significance of camera calibration in these systems, even minor inconsistencies can lead to major performance degradation. After outlining the importance of BEV perception and the challenges posed by noisy camera extrinsics, we provide a visual representation in Fig. 1. The figure distinctly illustrates the influence of camera parameters on BEV positional encoding. It further showcases how different strategies, namely using

reference points and employing local attention, are affected by the camera extrinsic noise.

In this study, we introduce a BEV representation network that maintains robustness against inaccuracies in extrinsic camera parameters. To achieve this, we have developed Dual-Space Positional Encoding (DSPE), a strategy that integrates global and local positional encodings to effectively decouple the network from its dependency on precise camera extrinsic attributes. By doing so, we enable the network to generalize better in the presence of varying camera setups. Additionally, we harness image perception to further bolster the network's resilience, enhancing its tolerance to noise in these parameters and ensuring that our BEV representations remain reliable even under challenging conditions.

To summarize, our primary contributions are:

- We introduce an image-perception-based positional encoding mechanism for multi-camera BEV representations to address calibration errors during camera setup and pose fluctuations during operation.
- We propose a dual-space positional encoding strategy that combines local and global encodings to mitigate the network's strong coupling with camera extrinsic parameters, enhancing its robustness to varying conditions.
- We achieve notable detection accuracy on the Nuscenes dataset, comparable to existing techniques without utilizing temporal data, and maintain consistent performance even with imprecise extrinsic inputs.

II. Related Work

A. Monocular Perception Based on Transformer

Object perception from two-dimensional imagery poses a significant challenge. DETR [18] pioneered the application of the transformer framework to this domain, eschewing post-processing and achieving end-to-end detection. Later approaches have sought to overcome the extended training time of DETR, with the deformable DETR [19] being a notable example.

A pivotal component of monocular 3D detection is the retrieval of depth information from 2D images. Various strategies have been adopted to augment 2D detection results with corresponding 3D positions by leveraging depth data. Some methods derive depth from pre-training [20]–[22], whereas others embed depth estimation within the network to facilitate simultaneous training [15], [23], [24]. Furthermore, certain approaches exploit geometric principles, gleaning depth information straight from 2D images via perspective transformations [25]–[27].

B. Multi-Camera BEV Perception

Multi-camera BEV perception predominantly harnesses multiple cameras situated around a vehicle to comprehensively capture its surroundings. This collected data is then transformed into a BEV representation. Initial approaches relied on Inverse Perspective Mapping (IPM) to directly convert Perceptual View (PV) images into BEV imagery [28]. Building upon this, later methodologies, inspired by LiDAR systems, harnessed depth estimation to craft BEV features analogous to point clouds [8], [22], [29], [30]. The Lift-Splat-Shoot (LSS) technique [15] used depth distribution to craft a streamlined voxel-based BEV representation, while PYVA [27] implemented a Multi-Layer Perceptron (MLP) to implicitly deduce the external camera parameters needed for PV-to-BEV conversions. Recent research endeavors have identified the transformer as a potent tool for multi-view camera BEV perception [16], [31], [32]. For instance, CVT [17] adeptly merges overlapping camera views by employing cross-view attention to fuse local features, and BEVFormer [14] innovatively integrates temporal self-attention to embed temporal dynamics within BEV features. Predominantly, these methods either demand substantial computational commitment for global attention or exhibit a strong dependence on accurate external camera parameters for pivotal transformations. In stark contrast, our proposed methodology innovatively employs image perceptual positional encoding with localized attention. This not only trims computational demands but also bolsters resilience against inaccuracies in external camera parameters.

III. Method

A. Overall Architecture

As depicted in Fig. 2, our primary framework is an evolution of the CVT [17] model, with a significant improvement in the positional encoding strategy. We have replaced the traditional camera positional encoding with our innovative image perception-based positional encoding, further augmented by our proposed Local Positional Encoding technique. For images captured from n cameras, represented as $\mathbf{I} \in \mathbb{R}^{n\times h_i\times w_i\times 3}$, they are first processed by the backbone network, extracting multi-scale features denoted as $\mathbf{F} \in \mathbb{R}^{n\times h\times w\times c}$.

In contrast to earlier methods that directly fuse results in the global 3D space, our approach applies extrinsic parameters, given by $\mathbf{T} \in \mathbb{R}^{n\times 4\times 4}$, and camera intrinsic, denoted as $\mathbf{K} \in \mathbb{R}^{n\times 3\times 3}$, more effectively by leveraging Local Positional Encoding. This technique operates by independently transforming 2D features into 3D space within the coordinate system of each camera. By decoupling the positional relationships between cameras from the 2D-to-3D transformation, our framework enhances the network's ability to generalize across various environments and improves its robustness to changes in camera extrinsics, which are common in challenging working conditions such as vehicle operation.

The integration of Local Positional Encoding into our framework aims to provide a more accurate and stable representation of the multi-camera data, which is crucial for applications requiring precise spatial understanding and reconstruction.

Initial BEV features, represented as $\mathbf{Q} \in \mathbb{R}^{H\times W\times C}$, are then channeled into the encoder. Each encoder layer processes these features through both a self-attention module and a spatial cross-attention module, refining the BEV feature set for subsequent layers. The self-attention module is particularly important as it ensures the robustness of the image

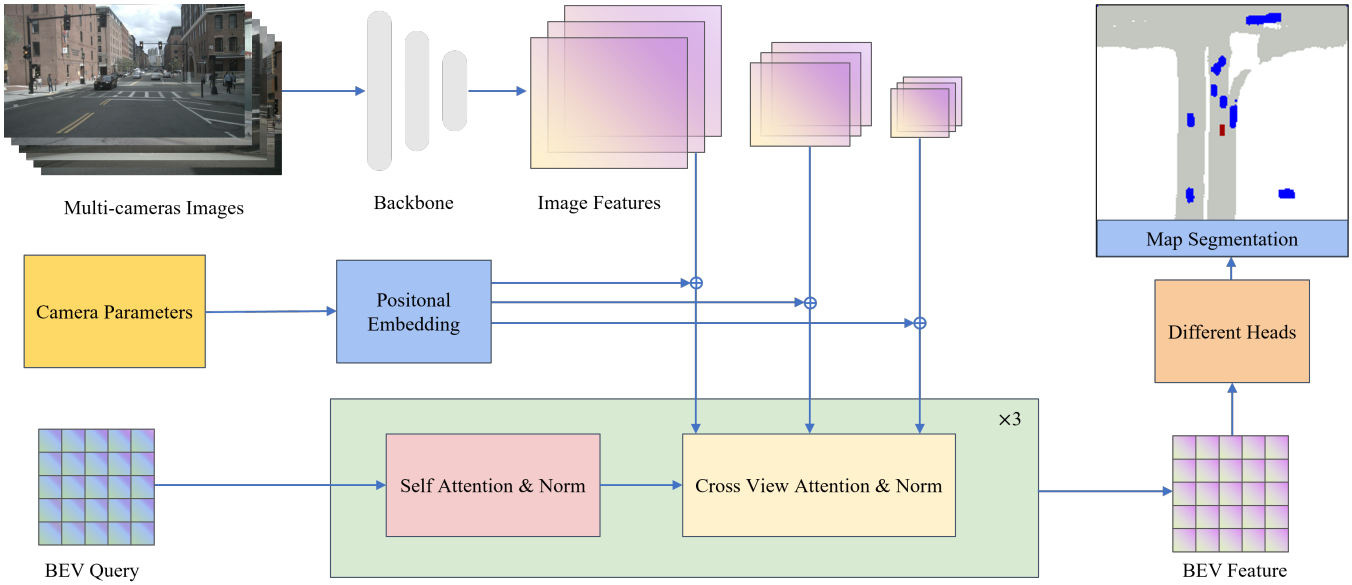

Fig. 2: An overview of our proposed network showcasing the integration of image perception and BEV encoding modules.

features as they transition to BEV features, mitigating the impact of potential inaccuracies in the external parameters.

Upon passing through all three encoder layers, the resulting BEV features are interpreted, depending on the specific detection or segmentation header being employed, to yield a binary mask output, $y \in \{0,1\}^{h \times w}$.

B. Dual-Space Positional Encoding

It is known that early methods derived results directly from monocular cameras and fused them in the global 3D space through post-processing. This included the transformation of 2D features into 3D space and the subsequent fusion of results from multiple cameras, with camera intrinsic and extrinsic parameters applied in these two stages, respectively. Positional encoding plays a pivotal role in neural network architectures by capturing spatial hierarchies and relationships. Recent methods opt to embed 2D image features from multiple cameras directly into the global 3D space through positional encoding. This approach is straightforward, but positioning all cameras in the global space through direct encoding entangles the positional relationships between cameras with the transformation from 2D views to 3D space. This not only hinders the network's ability to generalize universally but also, in harsh working environments, the changes in camera extrinsics during vehicle operation could lead to a significant degradation in network performance.

To address the issue at hand, we introduce Dual-Space Positional Encoding (DSPE), which encompasses both local and global positional encoding strategies. DSPE operates by independently converting 2D views to 3D space within the coordinate system of each camera. To elucidate the workings of our encoding, consider a singular camera viewpoint. Each 2D coordinate in the image gets mapped to its respective set of D 3D coordinates in the real world, denoted as $\mathbf{c}_d(d = 1,2,...,D)$. These 3D coordinates are initially transformed to align with the camera's coordinate system as:

$$\mathbf{c}_d^{cam} = \mathbf{K}_i^{-1}\mathbf{c}_d \tag{1}$$

where $\mathbf{K}_i$ signifies the i-th set of intrinsic parameters of the camera. And then we adapt Local Positional Encoding as:

$$\mathbf{p}^{local} = (\mathrm{MLP}(\mathrm{LN}(\mathbf{c}_d^{cam}))) \tag{2}$$

To fully leverage the advantages of multi-camera view perception, we also employ Global Positional Encoding. A subsequent transformation aligns these coordinates with the reference coordinate system (specifically, the LiDAR system):

$$\mathbf{c}_i^{ref} = \mathbf{T}_i\mathbf{c}_d^{cam} \tag{3}$$

with $\mathbf{T}_i$ denoting the i-th set of extrinsic camera parameters. Subsequently, the 3D coordinates are transformed into 3D positional embeddings as:

$$\mathbf{p}^{global} = (\mathrm{MLP}(\mathrm{LN}(\mathbf{c}_i^{ref}))) \tag{4}$$

In this equation, MLP stands for Multi-Layer Perceptron, LN is Layer Normalization. Finally, we concatenate the local positional encoding with the global positional encoding to obtain the 3D positional encoding:

$$\mathbf{p} = \mathrm{Concat}(\mathbf{p}^{local}, \mathbf{p}^{global}) \tag{5}$$

C. Image Perception Positional Encoding

Positional encoding plays a pivotal role in neural network architectures by capturing spatial hierarchies and relationships. In the case of the original CVT network's cross-view attention module, a camera-aware positional encoding is used. This traditional method leans on geometric priors to link points in the world coordinate system with their

counterparts in the image coordinate system. Although effective, it hinges heavily on precise camera parameters. In our approach, we introduce IPPE, an image perception positional encoding that relies directly on transforming the image based on these parameters, ensuring enhanced resilience against potential inaccuracies

To ensure that camera images from different viewpoints can be correlated in the global coordinate system and to provide additional information for 3D positional embeddings, we extract features from the images for image perception. These features offer guidance on the content information for how the transformed 3D points should be embedded with positional encoding. This contextual guidance is adopted by the image-perceived positional encoding in the form of weights and is parameterizable for learning. Therefore, we update the formula (2) as:

$$\mathbf{p}^{local} = (\mathrm{MLP}(\mathrm{LN}(c_d^{cam}))) \circ (\mathrm{MLP}(\mathrm{LN}(\mathbf{F}_i))) \qquad (6)$$

where $\circ$ symbolizes Hadamard multiplication.

IV. Experiments

A. Dataset and Metrics

We conducted our experiments on the Nuscenes dataset [33]. The Nuscenes dataset contains 1,000 distinct scenes, with each scene spanning 20 seconds. These scenes capture a variety of weather, lighting, and traffic conditions. All scenes were recorded using six RGB cameras, offering a comprehensive 360° view around the vehicle. Neither LiDAR nor radar data was utilized during the training or evaluation phases. The dataset provides calibrated intrinsic K and extrinsic T parameters for the cameras at each timestamp.

To ensure a fair evaluation, we assessed the predictions against the ground truth from CVT [17]. The evaluation region was set to a 100m$\times$100m area centered around the vehicle, with the BEV grid resolution fixed at 0.5m. Our primary performance metric is the Intersection over Union (IoU) between the model predictions and the ground truth labels from the map.

B. Implementation Details

For the sake of a fair comparison, we replicated the experimental settings of the image encoder from CVT [17]. We employed the EfficientNet B4 [34] for image feature extraction, in line with the approaches used in CVT and Fiery [23]. Features were extracted from the original image dimensions of 224$\times$480 across three scales, yielding resolutions of (56, 120), (28, 60), and (14, 30).

We set the default BEV query size to 200$\times$200, encompassing a perceptual range of [-50m, 50m] along both the X and Y axes at a resolution of 0.5m. Our architecture includes three encoder layers for BEV encoding. Training of the model was conducted using the focal loss function with a batch size of 4 per GPU spanning 50 epochs. We used the AdamW optimizer in conjunction with a single-cycle learning rate scheduler. The entire training duration amounted to roughly 5 hours using 8 RTX2080 Ti GPUs.

TABLE I: Vehicle map-view segmentation on nuScenes. We evaluate using the Intersection over Union (IoU) metric and report the frames per second (FPS) to indicate computational speed.

	IoU	#Params(M)	FPS
LSS [15]	32.1	14	25
FIERY† [23]	35.8	7	8
CVT [17]	36.0	5	35
Ours	**37.6**	8	25

† We use FIERY static to compare.

C. Main Results

The comparative outcomes between our model and various leading-edge approaches that eschew temporal models—including CVT [17], FIERY [23], and LSS [15]—are showcased in Table I. To maintain the integrity of comparison, every model was assessed solely based on single-time-step data. As evidenced by the table, our technique registered a superior IOU relative to other methodologies, notwithstanding similar parameter dimensions.

To assess the robustness of our model in the presence of external camera noise, we conducted a comparative evaluation with the CVT [17] model under 6 Degree-of-Freedom (6DoF) camera disturbances, which include unintended rotations and translations along the X, Y, and Z axes of the camera coordinate system. Perturbations were induced by adjusting camera poses with zero-mean Gaussian noise, with the severity governed by the standard deviation of the applied Gaussian noise. We analyzed six distinct types of disturbances independently. The results are presented in Fig. 3, where IoU curves comparing the performance of our method against the CVT method are provided. Upon comparison, it was evident that our method generally outperformed CVT across the various types of disturbances. Our method demonstrated remarkable robustness to translations along the X and Z axes, whereas CVT displayed heightened sensitivity to translations along the Z axis. Although our method exhibited somewhat diminished performance against Y-axis translations, it still outperformed CVT. For rotational disturbances around the X and Z axes, our method maintained a superior performance overall. Nonetheless, it is important to note that both our method and CVT showed reduced robustness against rotations around the Y axis. The IoU deteriorated sharply as the noise intensity increased. The primary reason for this vulnerability is the significant disruption of the predefined multi-view layout [33] caused by rotations around the Y axis.

D. Ablation Study

To delve deeper into the contributions of distinct components under varying noise environments, we conducted ablation studies. This process was instrumental in ascertaining the significance of each module, particularly in the face of diverse noise conditions. As elucidated in Fig 4, the findings underscore the indispensable contribution of every component in enhancing the model's resilience to varying

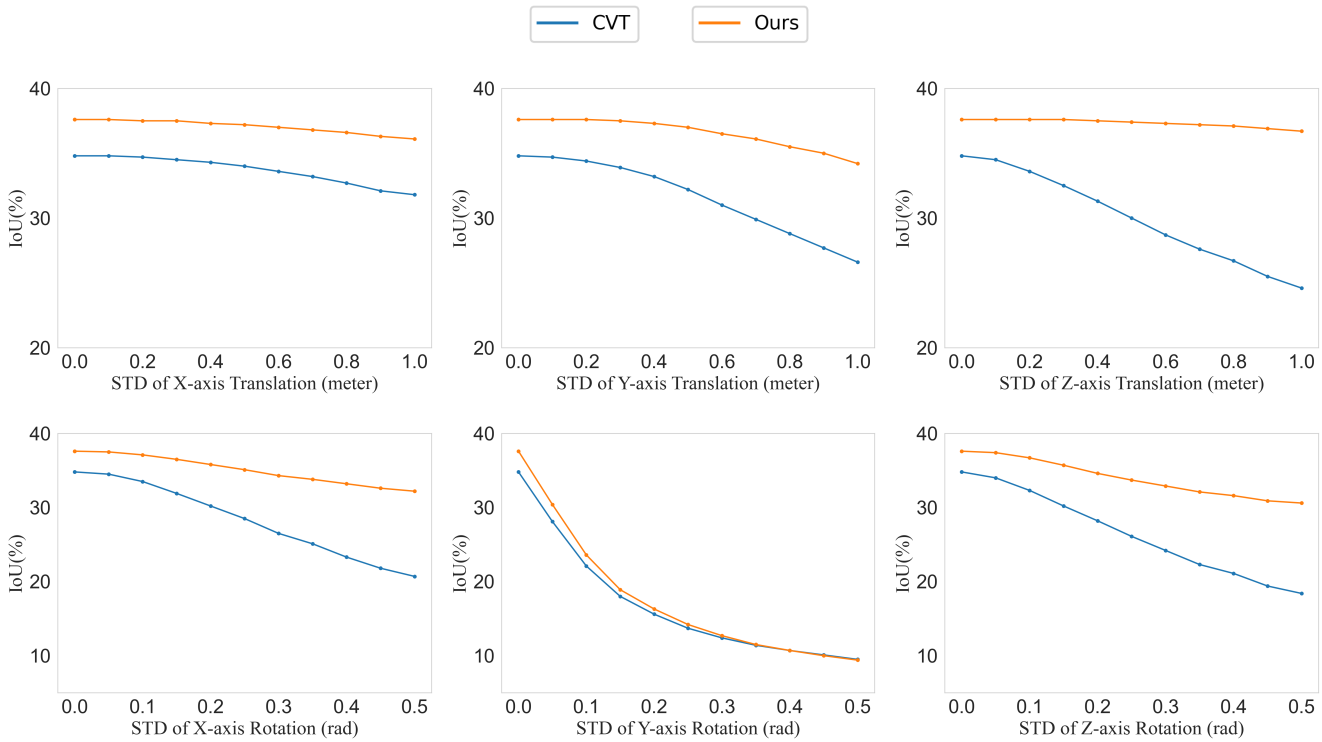

Fig. 3: Comparison between CVT and our method under 6DoF camera perturbations. The perturbations are induced by adjusting camera poses with zero-mean Gaussian noise, where the standard deviation of the noise determines the magnitude of the perturbations. Each camera's pose is perturbed independently within its own coordinate system.

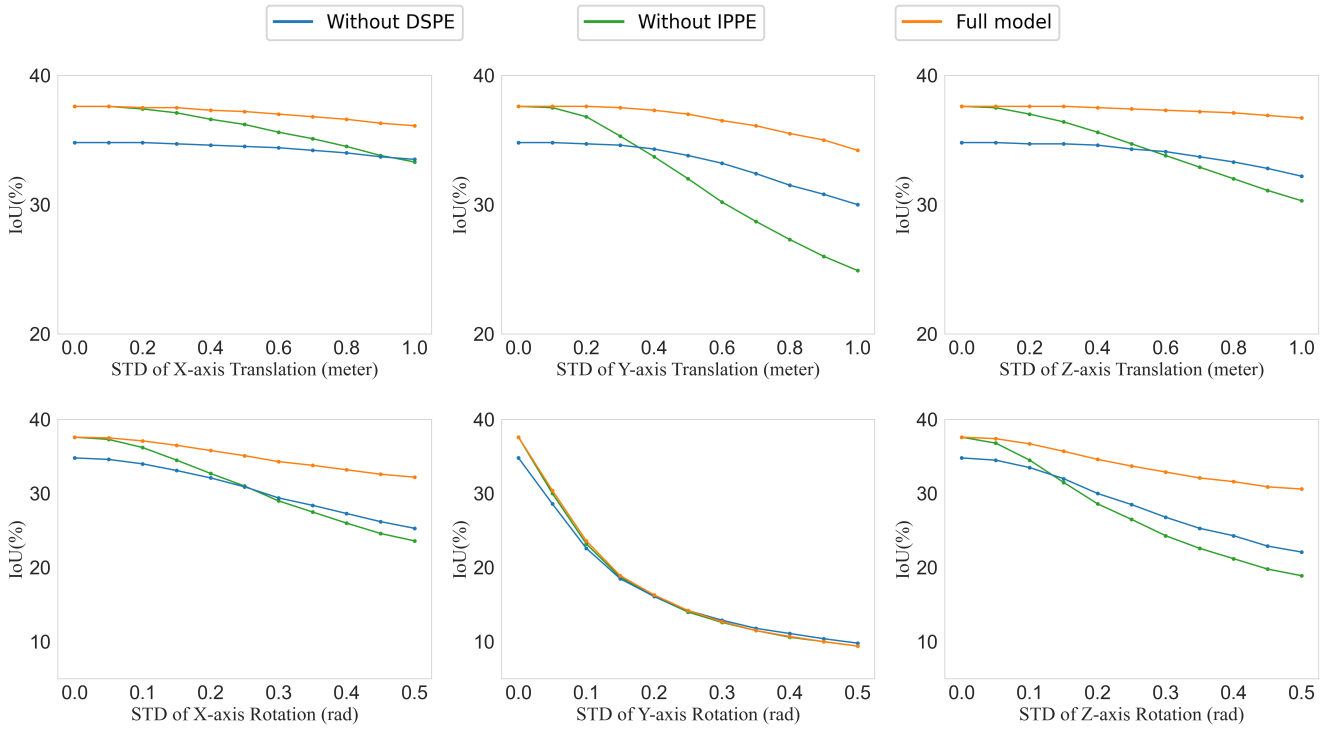

Fig. 4: Ablations of positional embeddings under 6DoF camera perturbations. This figure illustrates the effects of ablating either the Dual-Space Positional Encoding (DSPE) or the Image Perception Positional Encoding (IPPE) under 6 Degree-of-Freedom (6DoF) camera perturbations.

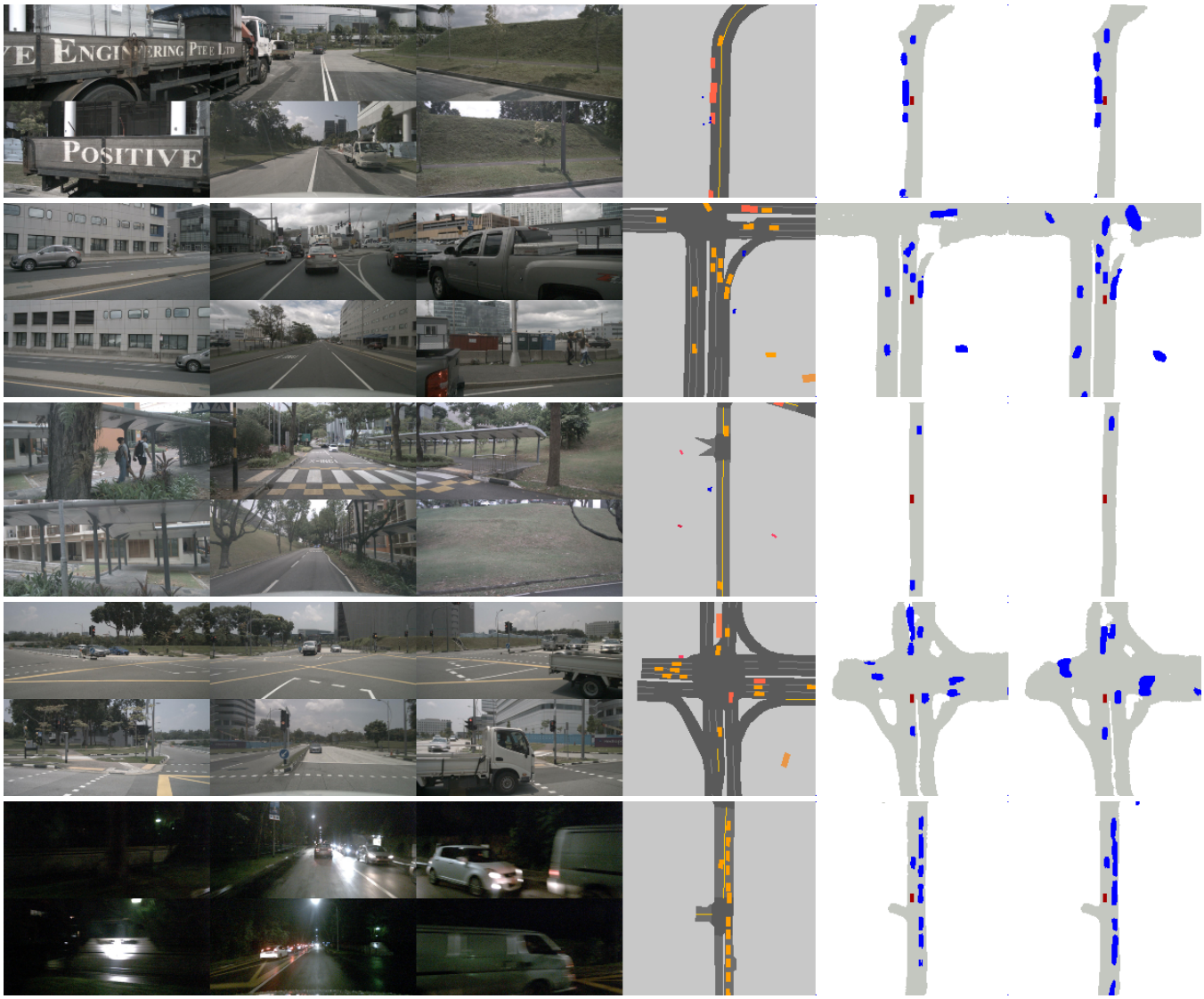

Fig. 5: Qualitative results on scenes with subtle camera noise, including translation noise with a standard deviation of 0.1m and rotation noise with a standard deviation of 0.05rad. Each row contains, from left to right: six camera views surrounding the vehicle (top 3 are front-facing, bottom 3 are back-facing), the ground truth BEV segmentation, our predicted BEV segmentation, and CVT [17]'s predicted BEV segmentation. The ego-vehicle is centrally positioned on the BEV segmentations.

magnitudes of camera extrinsic noise. Our proposed IPPE method has effectively enabled the model to learn a variety of features within images, thereby improving the model's capability and robustness. On the other hand, the DSPE method has played a significant role in greatly enhancing the model's robustness, allowing it to maintain relatively slower degradation in accuracy even when confronted with drastic changes in inaccurate extrinsic parameters. The final model performs best with all its positional embedding components.

E. Visualization

Fig. 5 showcases a selection of qualitative detection outcomes. In the BEV space, the results of vehicle segmentation are vividly displayed. The representations within this domain reveal that, when juxtaposed with other techniques, our model consistently generates predictions more congruent with the ground truth, even amidst camera extrinsic noise disturbances. Such visual evidence reiterates the superior detection efficacy of our proposed technique.

V. CONCLUSIONS

This research presents an innovative image-perceptive multi-camera BEV perception approach that mitigates the prevalent dependency on camera parameters exhibited by prior surround-view camera-based BEV detection methodologies. Significantly, our approach delivers real-time, state-of-the-art performance, distinguishing itself by circumventing the need for temporal models, which is a common requisite in similar studies.

As we move forward, our aspirations are clear: to delve deeper into the realm of real-time robust BEV perception by seamlessly incorporating temporal information.

ACKNOWLEDGMENT

We would like to express our sincere gratitude to Yufeng Yue for his invaluable review and suggestions on the manuscript. Additionally, we extend our appreciation to the other members of the ININ Lab of Beijing Institute of Technology for their assistance and contributions during this research.

REFERENCES

[1] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 12 697–12 705.

[2] Y. Zhou and O. Tuzel, "Voxelnet: End-to-end learning for point cloud based 3d object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4490–4499.

[3] X. Chen, T. Zhang, Y. Wang, Y. Wang, and H. Zhao, "Futr3d: A unified sensor fusion framework for 3d detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 172–181.

[4] X. Bai, Z. Hu, X. Zhu, Q. Huang, Y. Chen, H. Fu, and C.-L. Tai, "Transfusion: Robust lidar-camera fusion for 3d object detection with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 1090–1099.

[5] Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. L. Rus, and S. Han, "Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 2774–2781.

[6] Y. Li, Y. Chen, X. Qi, Z. Li, J. Sun, and J. Jia, "Unifying voxel-based representation with transformer for 3d object detection," *Advances in Neural Information Processing Systems*, vol. 35, pp. 18 442–18 455, 2022.

[7] T. Yin, X. Zhou, and P. Krähenbühl, "Multimodal virtual point 3d detection," *Advances in Neural Information Processing Systems*, vol. 34, pp. 16 494–16 507, 2021.

[8] D. Park, R. Ambrus, V. Guizilini, J. Li, and A. Gaidon, "Is pseudo-lidar needed for monocular 3d object detection?" in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3142–3152.

[9] L. Reiher, B. Lampe, and L. Eckstein, "A sim2real deep learning approach for the transformation of images from multiple vehicle-mounted cameras to a semantically segmented image in bird's eye view," in *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2020, pp. 1–7.

[10] T. Wang, Z. Xinge, J. Pang, and D. Lin, "Probabilistic and geometric depth: Detecting objects in perspective," in *Conference on Robot Learning*. PMLR, 2022, pp. 1475–1485.

[11] T. Wang, X. Zhu, J. Pang, and D. Lin, "Fcos3d: Fully convolutional one-stage monocular 3d object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 913–922.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

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

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

[15] J. Philion and S. Fidler, "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*. Springer, 2020, pp. 194–210.

[16] Y. Liu, T. Wang, X. Zhang, and J. Sun, "Petr: Position embedding transformation for multi-view 3d object detection," in *European Conference on Computer Vision*. Springer, 2022, pp. 531–548.

[17] B. Zhou and P. Krähenbühl, "Cross-view transformers for real-time map-view semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 13 760–13 769.

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

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

[20] S. Schulter, M. Zhai, N. Jacobs, and M. Chandraker, "Learning to look around objects for top-view representations of outdoor scenes," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 787–802.

[21] B. Liu, B. Zhuang, S. Schulter, P. Ji, and M. Chandraker, "Understanding road layout from videos as a whole," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4414–4423.

[22] C. Reading, A. Harakeh, J. Chae, and S. L. Waslander, "Categorical depth distribution network for monocular 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8555–8564.

[23] A. Hu, Z. Murez, N. Mohan, S. Dudas, J. Hawke, V. Badrinarayanan, R. Cipolla, and A. Kendall, "Fiery: Future instance prediction in bird's-eye view from surround monocular cameras," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 273–15 282.

[24] H. Wang, P. Cai, Y. Sun, L. Wang, and M. Liu, "Learning interpretable end-to-end vision-based motion planning for autonomous driving with optical flow distillation," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 13 731–13 737.

[25] T. Yin, X. Zhou, and P. Krahenbuhl, "Center-based 3d object detection and tracking," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 11 784–11 793.

[26] V. Sitzmann, J. Thies, F. Heide, M. Nießner, G. Wetzstein, and M. Zollhofer, "Deepvoxels: Learning persistent 3d feature embeddings," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2437–2446.

[27] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 15 536–15 545.

[28] S. Ammar Abbas and A. Zisserman, "A geometric approach to obtain a bird's eye view from an image," in *Proceedings of the IEEE/CVF international conference on computer vision workshops*, 2019, pp. 4095–4104.

[29] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8445–8453.

[30] R. Qian, D. Garg, Y. Wang, Y. You, S. Belongie, B. Hariharan, M. Campbell, K. Q. Weinberger, and W.-L. Chao, "End-to-end pseudo-lidar for image-based 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5881–5890.

[31] Y. Jiang, L. Zhang, Z. Miao, X. Zhu, J. Gao, W. Hu, and Y.-G. Jiang, "Polarformer: Multi-camera 3d object detection with polar transformer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 1, 2023, pp. 1042–1050.

[32] Y. Wang, V. C. Guizilini, T. Zhang, Y. Wang, H. Zhao, and J. Solomon, "Detr3d: 3d object detection from multi-view images via 3d-to-2d queries," in *Conference on Robot Learning*. PMLR, 2022, pp. 180–191.

[33] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.

[34] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.