FreqPDE: Rethinking Positional Depth Embedding for Multi-View 3D Object Detection Transformers

Supplementary Material

A. Appendix

A.1. More Related Work

Depth Estimation Depth estimation from 2D camera images is a challenging topic in Computer Vision, categorized into regressing metric depth [2, 3, 20, 27, 61] and relative depth [23, 37, 43]. BinsFormer [27] introduces sufficient interaction between probability distribution and bin predictions to generate proper metric estimation. DPT [43] exploits vision transformers as a backbone for dense relative depth prediction. Recent works [4, 58, 59] attempt to build a foundation model with excellent generalization performance across domains while maintaining metric scale. ZoeDepth [4] uses a lightweight depth head with a novel metric bin design to combine metric and relative depth estimation. DepthAnything [58, 59] introduces the affine-invariant loss to ignore the unknown scale and shift during the training stage, additionally, a data engine has been devised to automatically generate pseudo depth annotations for unlabeled images.

3D Positional Embedding The necessity of the 3D Position Encoder (PE) has been addressed in prior studies [32, 33, 45]. PETR series [32, 33] discretize the camera frustum space into meshgrid coordinates which are transformed to 3D world space with camera parameters, then the 3D coordinates are input to a 3D position encoder with 2D image features to construct the 3D position-aware features. However, leveraging hand-crafted camera-ray depth bins as the channel dimensionality for the point cloud disregards depth variations across different pixels. To ameliorate the aforementioned problem, 3DPPE [45] transforms the pixels to 3D space with camera parameters and predicted pixel-wise depth results, the resulting 3D points are sent to a position encoder to construct the 3D feature with point-level embeddings.

A.2. Implicit Distribution Supervision.

To fully leverage the strengths of the foundation model, we exploit the generated relative depth results as pseudo labels for extra supervision of our depth prediction $D_{i,j}$. To elaborate, the crucial issue is converting metric depth to relative depth, this process can be formulated as:

$$D^{rel} = \frac{1}{\text{scale}}\left(\frac{1}{D^{mtr}} - \text{shift}\right), \tag{13}$$

where D^{rel} is relative depth, D^{mtr} is metric depth, scale and shift are sample-wise parameters for transposition. Noticing the linear relationship between $\frac{1}{D^{mtr}}$ and D^{rel}, we perform mean-variance normalization [1] separately:

$$\begin{aligned}\widehat{\frac{1}{D^{mtr}}} &= \frac{\frac{1}{D^{mtr}} - \mathbb{E}(\frac{1}{D^{mtr}})}{\sqrt{Var(\frac{1}{D^{mtr}})}},\\ \widehat{D^{rel}} &= \frac{\frac{1}{\text{scale}}(\frac{1}{D^{mtr}} - \mathbb{E}(\frac{1}{D^{mtr}}))}{|\frac{1}{\text{scale}}|\sqrt{Var(\frac{1}{D^{mtr}})}},\end{aligned} \tag{14}$$

where $\mathbb{E}$ and Var represent the computation of the mean and variance respectively, $\widehat{\frac{1}{D^{mtr}}}$ and $\widehat{D^{rel}}$ correspond to the normalized outcomes. Given that the coefficient $\frac{1}{\text{scale}}$ is strictly positive, it follows that the two normalization results for each sample are equivalent. Consequently, we take the reciprocal of the predicted depth $D_{i,j}$, normalize the outcome, and employ the normalized pseudo-labels as supervisory signals to facilitate the supervised learning transition from metric depth to relative depth.

A.3. More Ablation Study

Cross View Attention. In our CSDP module, we apply a fixed-ratio mask to the features in order to mitigate the influence of non-overlapping regions. To verify the effectiveness of this masking approach, we conduct ablation studies to evaluate the impact of different mask ratios, as illustrated in Tab. 8. With a mask ratio of 0.2, our method demonstrates improved performance, outperforming the model without masking and other mask ratio.

Table 8. Necessity of cross-view.

Mask Ratio	NDS ↑	mAP ↑	mATE↓
-	58.3	50.3	0.578
0.1	57.3	49.6	0.609
0.2	**58.5**	**50.5**	**0.569**
0.3	58.2	50	0.580

Effect of Positional Depth Encoder This study seeks to provide empirical evidence of that positional encoding, within the multi-level depth maps, enhances the detection capacity of 3D objects by the query. As shown in Tab. 9, wherein multi-level scale-invariant depth prediction results are resized to the same scale and fused together to be fed into a point-wise embedding function, which outperforms the baseline by 1.4% NDS and 2.4% mAP, also exceed the single-level embedding method similar to 3DPPE [45].

Comparison with LSS method To validate the 'plug-and-play' capability of proposed depth predictor, we replace the

Table 9. Ablation for Positional Depth Encoder on nuScenes.

Method	**NDS ↑**	**mAP ↑**	**mATE ↓**
Baseline	57.2	48.2	0.602
Single-level	57.9	49.6	0.587
Multi-levels	**58.6**	**50.6**	**0.576**

Table 10. Comparison with LSS-based method.

Method	Backbone	Input Resolution	mAP	NDS
BEVDepth	R50	256*704	35.1	47.5
BEVDepth-R	R50	256*704	**36.0**	**48.4**

Table 11. Effect of Hybrid Depth Supervision on nuScenes `val` set.

Supervision	**Abs Rel ↓**	**Sq Rel ↓**	**NDS ↑**	**mAP ↑**
LiDAR only	0.17	1.45	58.3	49.9
Pseudo only	0.23	3.71	58.4	50.1
Hybrid	**0.15**	**1.41**	**58.6**	**50.6**

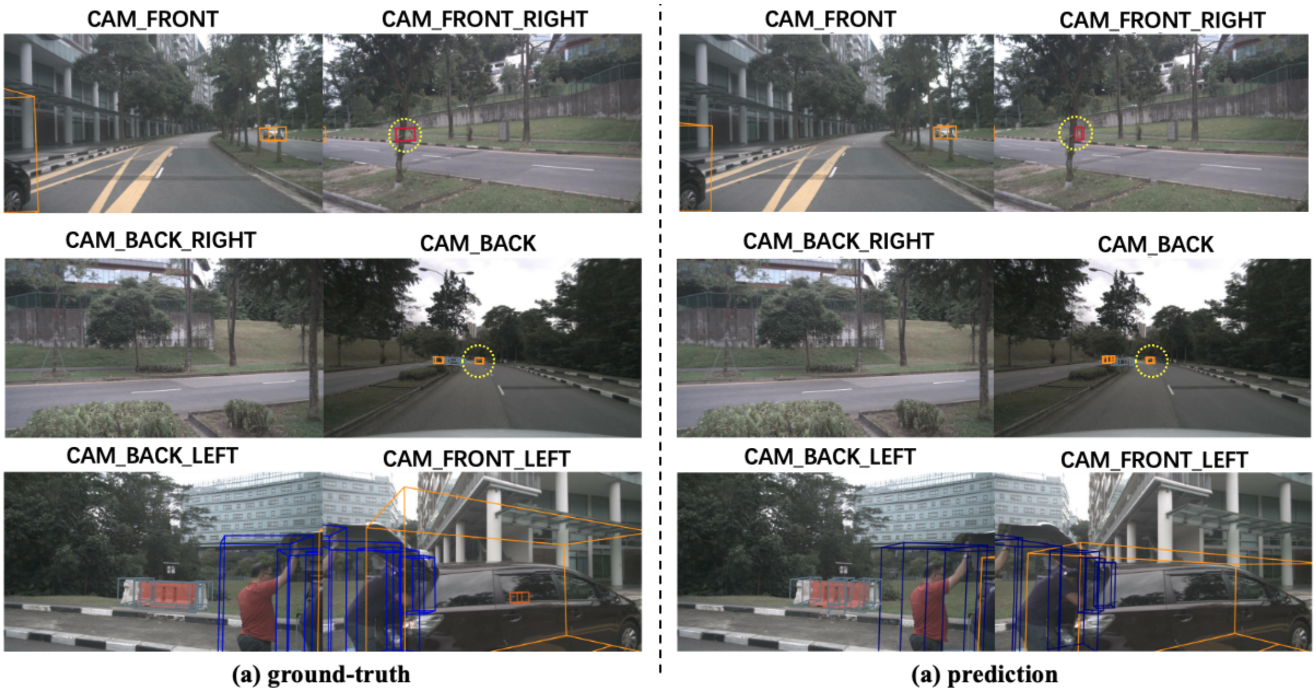
depth predictor in BEVDepth with our FSPE and CSDP modules. The results, presented in Table 10, demonstrate the effectiveness and transferability of our proposed design.

Effect of Hybrid Depth Supervision. To further validate the effectiveness of the hybrid supervision approach for CSDP, we compare the performance of different supervision methodologies. As presented in Tab. 11, employing only pseudo-labels results in an improvement in detection performance; however, it leads to a decrease in depth estimation performance. This indicates that distribution-based supervision provides a more comprehensive supervisory signal for overall depth maps but lacks the precision of absolute depth supervision. Consequently, with hybrid supervision, both the absolute relative error (Abs Rel) and squared relative error (Sq Rel) decrease, while the model achieves a 1.4% increase in mAP and a 0.6% increase in NDS.

A.4. Result Visualization

Qualitative Results. We show the qualitative detection results of FreqPDE in Fig. 6 on multi-view images. The 3D predicted bounding boxes are drawn with different colors for different classes. As illustrated by the highlighted circles, our method accurately detects the category and location of distant targets, while also mitigating the challenges posed by occluded small targets to some extent. This demonstrates an enhancement in the model's detection capability for distant targets following the integration of a more precise depth estimation module.

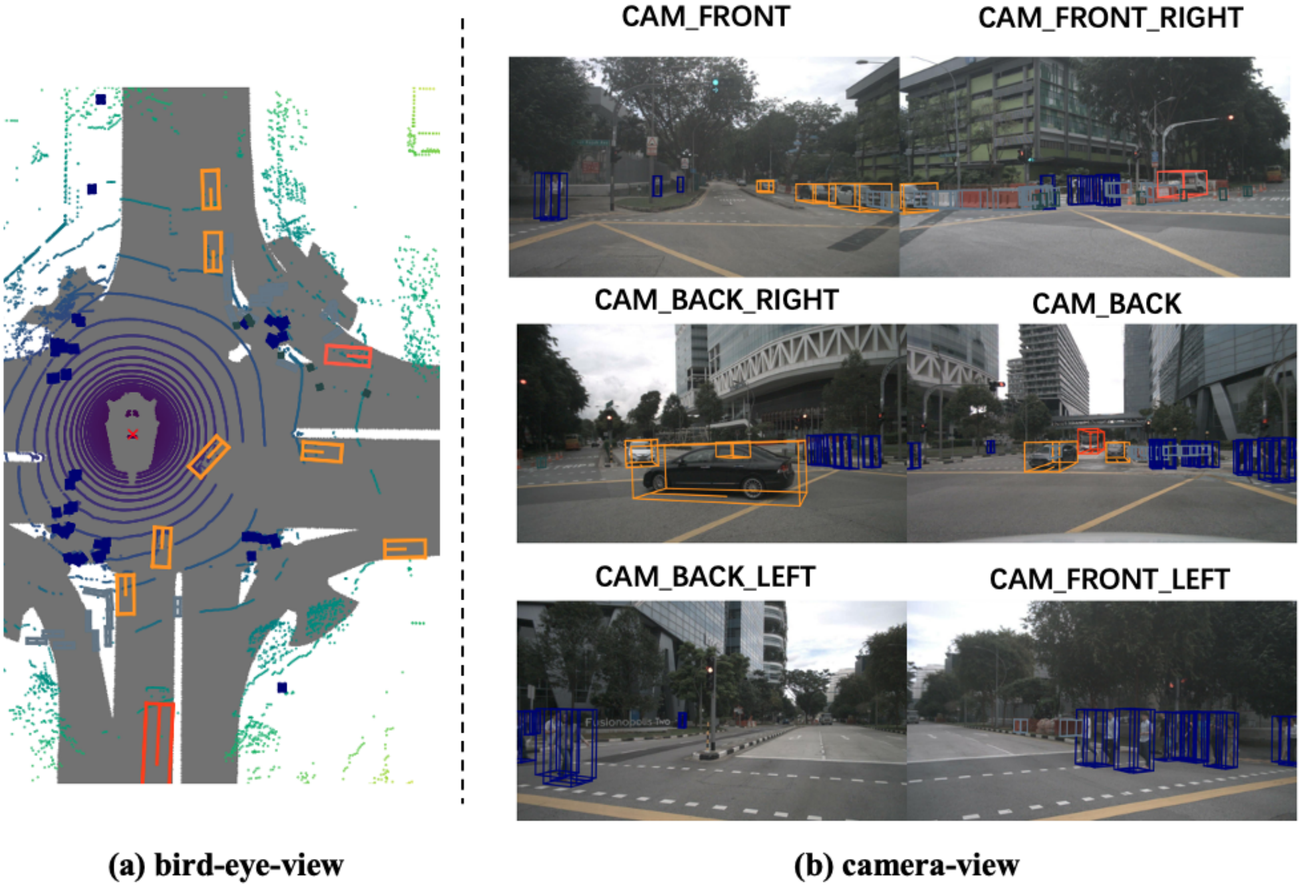
More Visualization. We also show more detection results of some challenging scenes in Fig. 7 and Fig. 8. Our method shows impressive results on crowded and distant objects.



(a) ground-truth

(a) prediction

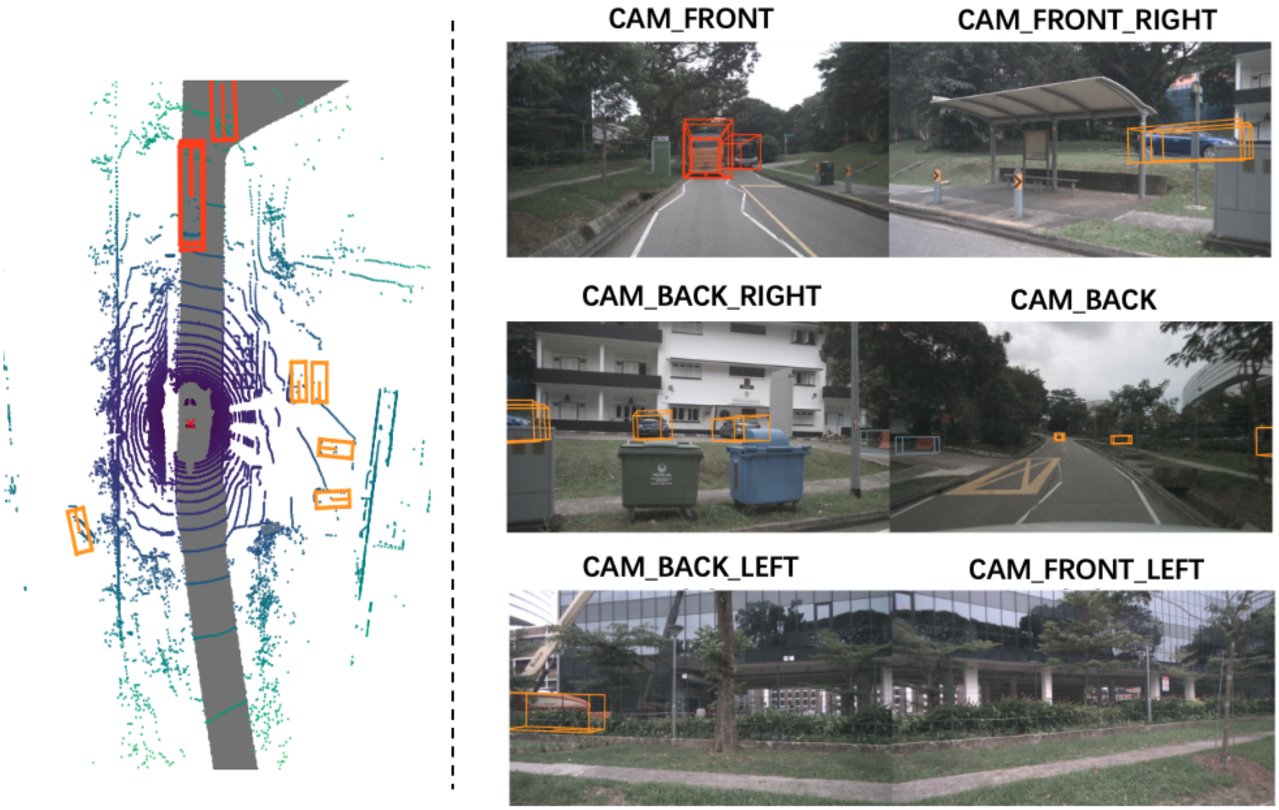
Figure 6. Qualitative detection results on multi-view images on the nuScenes val set. The 3D predicted bounding boxes are drawn with different colors for different classes.



(a) bird-eye-view

(b) camera-view

Figure 7. Qualitative detection results on multi-view images and BEV space on the nuScenes val set.



(a) bird-eye-view **(b) camera-view**

Figure 8. Qualitative detection results on multi-view images and BEV space on the nuScenes val set.