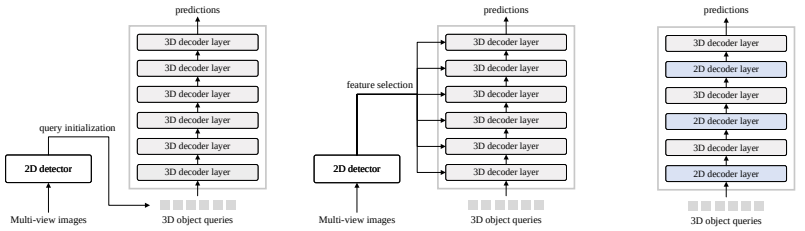
# Appendix

This supplementary material contains additional details of the main manuscript and provides more experiment analysis. In Appendix A, we present the difference between SimPB and other previous works that utilize 2D results as priors. Next, we elaborate on the complete architecture and give more implementation details in Appendix B. Then, we provide more experiment analysis about runtime, encoder ablation study, and association accuracy in Appendix C. Finally, more visualization results are illustrated in D.

## A Utilizing 2D Results as Priors




**(a)** query initilization **(b)** feature selection **(c)** cyclic 2D&3D layers

**Fig. S1:** Comparison of SimPB with previous approaches utilizing 2D results as priors. We roughly categorize these previous methods into two categories, (a) query initialization and (b) feature selection. Instead, SimPB introduces a unified paradigm using novel cyclic 2D&3D layers as in (c).

We highlight the difference between SimPB and previous approaches that use 2D results as priors in two aspects: architecture and association.

**Architecture.** In Fig. S1, we categorize the previous methods into two groups: query initialization and feature selection, as summarized below.

- Query Initialization: In this category, 3D queries are typically initialized from 2D boxes that are detected by a 2D detector (as shown in Fig. S1 (a)).
- Feature Selection: The methods focus on foreground tokens through 2D supervision and then select them for interaction with 3D queries (as shown in Fig. S1 (b)).

All these methods employ a 2D detector (or utilize a 2D head) to predict 2D results as a preliminary step before applying a 3D detector. In contrast, SimPB takes a distinct approach. It performs simultaneous multi-view 2D and 3D detection within a single model using cyclic 2D & 3D decoder layers (as illustrated in

Fig. S1 (c)). SimPB is a one-stage method that does not rely on an off-the-shelf 2D detector.

**Association.** For association, we refer to the connection between 2D and 3D results for the same target. A summary of the association of previous methods is listed as follows.

- Query Initialization: This method employs a heuristic default association, where a 3D query is linked to a 2D box for its initialization. This association is referred to as a 2D-to-3D association.
- Feature Selection: In this approach, the association between 3D queries and selected 2D image tokens, supervised by a 2D detector, is learned through the transformer. However, it does not explicitly establish a direct association between 2D and 3D results.

In contrast, our method determines the association by projecting 3D anchors and matching them with the corresponding 2D results. In this way, our approach establishes a 3D-to-2D association between 2D and 3D results. The 3D-to-2D association has the advantage of aggregating 2D information more efficiently and avoiding the generation of redundant 3D results. We give a detailed analysis in Appendix C.3 and Appendix D.1.

# B More Implementation Details

## B.1 Architecture Details

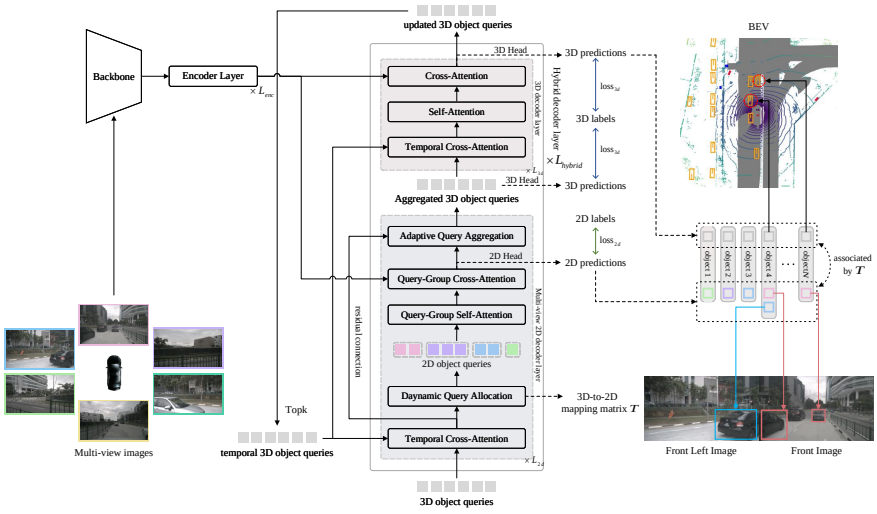



**Fig. S2:** Comprehensive architecture of SimPB.

To maintain clarity, some minor components in Fig. 2 of the main manuscript have been omitted. We provide the complete architecture of SimPB in Fig. S2 for a comprehensive view. Specifically, we include arrow lines to illustrate the connections between self-attention and cross-attention in both the multi-view 2D decoder layer and the 3D decoder layer. Additionally, we visualize the residual connection from the output of temporal cross-attention to the Adaptive Query Aggregation module. The aggregated 3D queries are separately shown as the output of a multi-view 2D decoder layer, which is used as input for the 3D head for deep supervision. Furthermore, to display the temporal propagation, we add an arrow line to indicate the updated object queries linking to the temporal 3D object queries.

### B.2 More Allocation Details

In the Dynamic Query Allocation module, a 3D query is allocated to a maximum of one object center and multiple projection centers across different camera views by projecting it using camera parameters. The projection center typically represents a truncated portion of a cross-view target. The total number of 2D object queries is equal to the combined count of object centers and projection centers.

During the early stages of training, the presence of inaccurate anchors can lead to a rapid increase in the number of projection centers, resulting in convergence challenges. To address it, we introduce two constraint strategies to optimize the allocation during training.

- The number of projection centers is limited to a maximum of 100 for each camera group. Consequently, the total number of 2D queries is restricted to a maximum of $N + 100 \times V$, where $N$ represents the number of 3D queries and $V$ denotes the number of cameras.
- To mitigate the impact of incorrectly projected anchors, we limit the maximum size $\{l, w, h\}$ of the anchors to $\{35, 35, 10\}$, which is computed from the training split of Nuscenes dataset.

In our implementation, the number of 3D queries is fixed at $N = 900$, while the number of 2D queries $M$ dynamically adjusts based on anchor projection. During inference, the number of 2D queries $M$ varies around an average of 1100, which is approximately 200 more than the original number of 3D queries $N = 900$. Nevertheless, this increase in the number of queries introduces only a negligible rise in computational overhead.

## C More Experimental Analysis

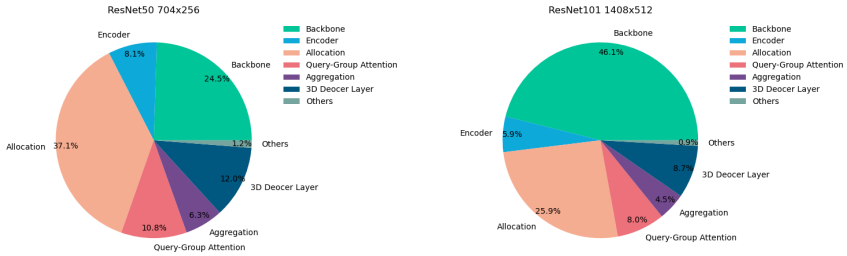
### C.1 Runtime Analysis

We compare the inference speeds of SimPB with state-of-the-art methods using two different backbones and model input resolutions. The evaluation is conducted

**Table S1:** Comparison of inference speeds.

| Method | Backbone | Resolution | FPS↑ |
|---|---|---|---|
| MV2D [6] | ResNet50 | 704 × 256 | 9.5 |
| StreamPETR [4] | ResNet50 | 704 × 256 | 27.1 |
| Sparse3Dv3 [2] | ResNet50 | 704 × 256 | 19.8 |
| SimPB | ResNet50 | 704 × 256 | 10.9 |
| MV2D [6] | ResNet101 | 1408 × 512 | 3.9 |
| StreamPETR [4] | ResNet101 | 1408 × 512 | 6.4 |
| Sparse3Dv3 [2] | ResNet101 | 1408 × 512 | 8.2 |
| SimPB | ResNet101 | 1408 × 512 | 7.1 |

on an NVIDIA 3090 GPU. As shown in Tab. S1, SimPB provides inferior performance compared to StreamPETR [4] and Sparse4Dv3 [2] at a resolution of 704 × 256. However, at higher resolutions, SimPB achieves comparable inference speeds to these methods. Notably, SimPB consistently outperforms MV2D [6] in terms of inference speed.




**(a)** ResNet50 & 704 × 256 **(b)** ResNet101 & 1408 × 512

**Fig. S3:** Run time decomposition on two configurations.

To gain a better understanding of computational complexity, we provide an analysis of the runtime distribution for each module of SimPB under these two settings. The percentage of runtime for each module is illustrated in Fig. S3. The allocation process in SimPB is responsible for a significant portion of the runtime for the ResNet50 backbone, making it a major bottleneck that affects the overall inference speed. In the ResNet101 setting, where both the model capacity and the model input increase, the backbone itself takes up more time and becomes a significant bottleneck. However, the processing time for the allocation step does not vary with changes in model size or input. As a result, the model experiences a relatively smaller negative impact when utilizing a larger backbone and higher resolution. We plan to optimize the inference speed of the network in our future work.

### C.2 Impact of Encoder

**Table S2:** Impact of encoder layer of SimPB.

| Encoder layers | mAP↑ | NDS↑ | $AP_{2d}$ ↑ | FPS↑ | Memory(G)↓ |
|---|---|---|---|---|---|
| - | 0.412 | 0.519 | 0.211 | 12.9 | 10.64 |
| 1 | 0.421 | 0.527 | 0.217 | 10.9 | 14.46 |
| 2 | 0.425 | 0.529 | 0.220 | 9.0 | 17.89 |
| 3 | 0.432 | 0.536 | 0.222 | 7.9 | 21.07 |
| 4 | 0.439 | 0.545 | 0.224 | 6.1 | 23.94 |

Our objective is to develop a unified architecture capable of simultaneously output both 2D and 3D results from multiple cameras. To achieve this, we adopt an encoder-decoder structure following the design of the DETR-like scheme [7]. In contrast to previous sparse query-based methods such as DETR3D [5], PETR [3], and Sparse4D [1], which often exclude the encoder for 3D detection, we recognize the potential benefits of incorporating an encoder.

To explore its effectiveness, we conduct an additional ablation study, as listed in Tab. S2. The experimental settings align with Sec. 4.4 of the main manuscript. By incorporating the encoder, the model boosts the performance in mAP, NDS, and $AP_{2d}$. Increasing the number of encoder layers further enhances performance but at the cost of increased computation time and memory consumption. To achieve a balance between efficiency and effectiveness, we employ a single encoder layer in SimPB.

### C.3 Effect of Association

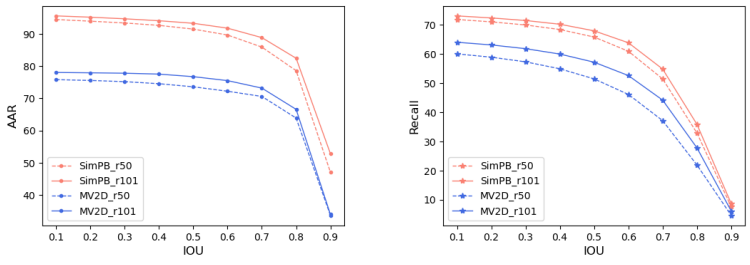



**Fig. S4:** AAR (Association Accuracy Rate) and Recall curves.

Two-stage methods usually only provide a default association between 3D and 2D results during query initialization. In contrast, SimPB explicitly constructs

the association of 2D and 3D detection results. To quantitatively evaluate the association, we design a metric termed Association Accuracy Rate (AAR) to measure the accuracy rate of association and also apply Recall as an evaluation metric as well.

Suppose there are $N_{3\mathrm{d}}$ 3D grountruth boxes $\{G^i_{3d}\}^{N_{3\mathrm{d}}}_{i=1}$ and $N_{2\mathrm{d}}$ projected 2D boxes $\{G^i_{2d}\}^{N_{2\mathrm{d}}}_{i=1}$ on the image views as 2D boxes label in the validation dataset. We can obtain $M_{3\mathrm{d}}$ 3D box prediction $\{P^i_{3d}\}^{M_{3\mathrm{d}}}_{i=1}$ and $M_{2\mathrm{d}}$ 2D prediction $\{P^i_{2d}\}^{M_{2\mathrm{d}}}_{i=1}$ from the network. Typically, a candidate match is established between a 3D prediction and a 2D groundtruth box. The total number of candidates matching is denoted as $\#\mathrm{Matching}_{\{\text{3D-pred, 2D-gt}\}}$. From these candidate matches, we select valid associations between 3D predictions and 2D predictions generated by the network. The valid matching number is $\#\mathrm{ValidMatching}_{\{\text{3D-pred, 2D-pred}\}}$. Therefore, we define the association evaluation metric AAR as follows:

$$\mathrm{AAR} = \frac{\#\mathrm{ValidMatching}_{\{\text{3D-pred, 2D-pred}\}}}{\#\mathrm{Matching}_{\{\text{3D-pred, 2D-gt}\}}} \times 100\%, \tag{1}$$

$$\mathrm{Recall} = \frac{\#\mathrm{Matching}_{\{\text{3D-pred, 2D-gt}\}}}{N_{2d}} \times 100\%. \tag{2}$$

For a given 3D prediction $P^i_{3d}$ and $j$-th 2D groundtruth $G^j_{2d}$ on $v$-th image view. And we denote the associated 3D grountruth of $G^j_{2d}$ as $G^j_{3d}$ for simplicity. $P^i_{3d\to 2d}$ is the bounding rectangle on $v$-th view projected from $P^i_{3d}$. The connection between $P^i_{3d}$ and $G^j_{2d}$ is a candidate matching if the following conditions are met:

$$\Phi(P^i_{3d}, G^j_{2d}) = \begin{cases} 1 & \text{if } \mathrm{Dist}(P^i_{3d}, G^j_{3d}) \le \tau_{dis} \ \&\ \mathrm{IoU}(P^i_{3d\to 2d}, G^j_{2d}) \ge \tau_{iou} \ \&\ \mathrm{Cls}(P^i_{3d}, G^j_{3d}) = 1 \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

where Dist represents the Euclidean distance between the centers of two 3D boxes, and Cls indicates whether the labels of the two boxes are the same or not. Similarly, we denote a valid candidate between the $i$-th 3D prediction $P^i_{3d}$ and the $k$-th 2D prediction $P^k_{2d}$ when the following conditions are met:

$$\Psi(P^i_{3d}, P^k_{2d}) = \begin{cases} 1 & \text{if } \Phi(P^i_{3d}, G^j_{2d}) = 1 \ \&\ \mathrm{IoU}(P^k_{2d}, G^j_{2d}) \ge \tau_{iou} \ \&\ \mathrm{Cls}(P^k_{2d}, G^j_{2d}) = 1 \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Therefore, AAR can be rewritten as:

$$\mathrm{AAR} = \frac{\sum_{i=1}^{M_{3d}} \sum_{k=1}^{M_{2d}} \Psi(P^i_{3d}, P^k_{2d})}{\sum_{i=1}^{M_{3d}} \sum_{j=1}^{N_{2d}} \Phi(P^i_{3d}, G^j_{2d})} \times 100\% \tag{5}$$

We fix the $\tau_{dis} = 2$ and adjust $\tau_{iou}$ from 0.1 to 0.9 to draw the AAR and Recall curves. As shown in Fig. S4, the accuracy and recall decrease as the IOU threshold $\tau_{iou}$ increases. However, SimPB constantly gains higher AAR and Recall metrics with a large margin on both settings. The utilization of the 3D-to-2D association in SimPB demonstrates higher accuracy compared to the 2D-to-3D association used in MV2D. This approach not only maintains a larger

number of matched predictions but also ensures better alignment of 2D and 3D features. As a result, the 2D information to the same target is effectively leveraged for 3D tasks, leading to improved performance.

## D Qualitative Evaluation

**(a)** The detection results of MV2D. 2D-to-3D association (red arrow) may produce duplicate 3D results or unrelated results from 2D priors for a cross-camera target.

**(b)** The detection results of SimPB. The process of 3D-to-2D association (green arrow) effectively yields accurate 3D results along with their corresponding 2D boxes for cross-camera targets.

**Fig. S5:** Illustration of association between 2D and 3D results by MV2D and SimPB.

### D.1 Qualitative Comparison on Association

To conduct a qualitative analysis of the association establishment between 2D-to-3D and 3D-to-2D, we compare the detection results of MV2D and SimPB in the same keyframe. The detection results of several cross-camera targets are shown in Fig. S5, where the yellow boxes represent the 2D detection results, and the blue boxes represent the 3D detection results

In the case of a cross-camera target $O$, MV2D initially employs a 2D detector to generate multiple 2D bounding boxes (using two as an example). These 2D results are used to initialize 3D queries through a 2D-to-3D association method. However, multiple 3D queries are associated with the target $O$. Only one of these 3D queries accurately predicts the target, while the other may produce a duplicated nearby object (circle 1 in Fig. S5 (a)) or even an unrelated result (circle 2 in Fig. S5 (a)). This discrepancy arises during the Hungarian matching step, where only the best candidate query is optimized as the positive sample, resulting in the suppression of the remaining 3D queries. Consequently, the 2D information from a specific view of the suppressed 3D query is discarded, despite it can provide relevant information about the same target.

To address this issue, SimPB adopts a novel approach to establish the association between 2D and 3D results using a 3D-to-2D method. For a cross-camera target, we distribute its 3D queries to different views for 2D detection tasks and subsequently aggregate the results to form a single 3D query. To this end, for a cross-camera object, SimPB only produces one 3D detection result along with its corresponding 2D detection in each relevant camera. Also, our cyclic 3D-2D-3D interaction ensures there is a single coherent representation of the target across different views, eliminating redundancy outputs and enhancing the accuracy of the results (in Fig. S5 (b)).

### D.2 Qualitative Comparison with State-of-the-Art Methods

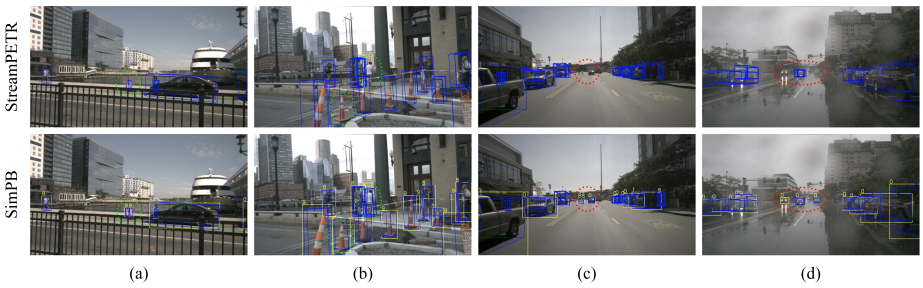



**Fig. S6:** Visualization results of StreamPETR and SimPB.

SimPB provides improved accuracy in detecting crowd objects such as traffic cones and pedestrians compared to StreamPETR. For instance, while StreamPETR incorrectly identifies two traffic cones as a single entity, SimPB accurately detects them as separate objects (green circle in Fig. S6 (a)). In Figure S6 (b), StreamPETR also provides an inaccurate estimation of the locations of pedestrians and traffic cones, showing that crowd objects tend to cluster around their neighboring objects. In contrast, SimPB provides more precise results and successfully distinguishes crowded and small objects. This improvement can be

attributed to the novel cyclic 3D-2D-3D scheme of SimPB, where the iterative and interactive process of 2D and 3D information enhances the refinement of queries, resulting in more accurate detection results.

SimPB also demonstrates its advantage in detecting distant targets and performs well even in challenging scenarios. For example, SimPB successfully detects pedestrians and cars at far distances, whereas StreamPETR fails to do so (red circle in Fig. S6 (c)). Furthermore, despite encountering difficulties in predicting small and distant targets within complex environments, such as rain (as shown in Fig. S6 (d)), SimPB can still provide reliable 2D detections. These 2D detections can be utilized in subsequent post-processing steps within a practical autonomous driving perception system.

### D.3 More Visualization Results

We present the visualization of the 2D and 3D detection results of SimPB using the ResNet101 backbone and a model input resolution of $1408 \times 512$. The visualizations are shown in Fig. S7 and Fig. S8. The number on the detected box represents its predicted category.

**Fig. S7:** Detection results on Nuscenes validation dataset during the daytime. 2D predict results are visualized in yellow and 3D results are visualized in bule.

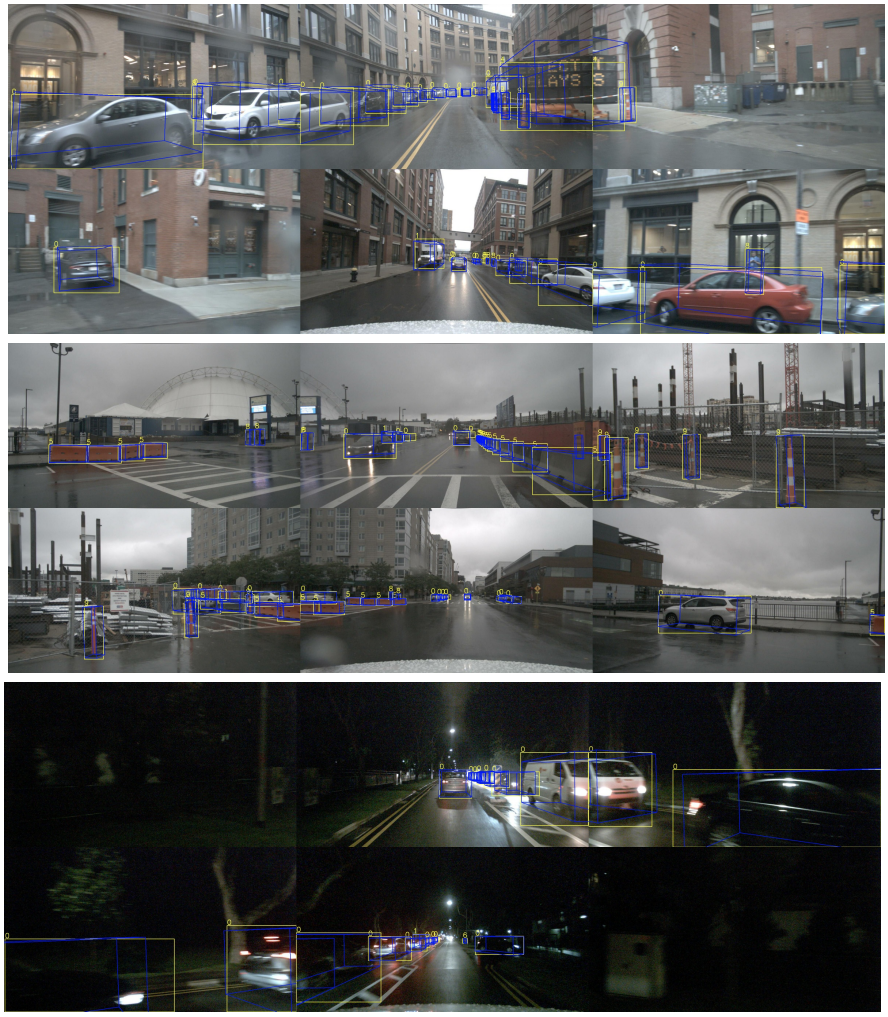


**Fig. S8:** Detection results on Nuscenes validation dataset during the rain and night. 2D predict results are visualized in yellow and 3D results are visualized in bule.

## References

1. Lin, X., Lin, T., Pei, Z., Huang, L., Su, Z.: Sparse4d: Multi-view 3d object detection with sparse spatial-temporal fusion. arXiv preprint arXiv:2211.10581 (2022)
2. Lin, X., Pei, Z., Lin, T., Huang, L., Su, Z.: Sparse4d v3: Advancing end-to-end 3d detection and tracking. arXiv preprint arXiv:2311.11722 (2023)
3. Liu, Y., Wang, T., Zhang, X., Sun, J.: Petr: Position embedding transformation for multi-view 3d object detection. In: ECCV. pp. 531–548 (2022)
4. Wang, S., Liu, Y., Wang, T., Li, Y., Zhang, X.: Exploring object-centric temporal modeling for efficient multi-view 3d object detection. In: ICCV. pp. 3621–3631 (2023)
5. Wang, Y., Guizilini, V., Zhang, T., Wang, Y., Zhao, H., Solomon, J.: Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In: COLR. pp. 180–191 (2021)
6. Wang, Z., Huang, Z., Fu, J., Wang, N., Liu, S.: Object as query: Equipping any 2d object detector with 3d detection ability. In: ICCV. pp. 3791–3800 (2023)
7. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. In: ICLR (2020)