

DAMap: Distance-aware MapNet for High Quality HD Map Construction

Jinpeng Dong, Chen Li, Yutong Lin, Jingwen Fu, Sanping Zhou, Nanning Zheng*

State Key Laboratory of Human-Machine Hybrid Augmented Intelligence,
National Engineering Research Center for Visual Information and Applications,
Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University

{djp1235a, edward82, yutonglin, fu1371252069}@stu.xjtu.edu.cn

{spzhou, nnzheng}@xjtu.edu.cn

Abstract

High-definition (HD) map is an important component to support navigation and planning for autonomous driving vehicles. Predicting map elements with high quality (high classification and localization scores) is crucial to the safety of autonomous driving vehicles. However, current methods perform poorly in high quality predictions due to inherent task misalignment. Two main factors are responsible for misalignment: 1) inappropriate task labels due to one-to-many matching queries sharing the same labels, and 2) sub-optimal task features due to task-shared sampling mechanism. In this paper, we reveal two inherent defects in current methods and develop a novel HD map construction method named DAMap to address these problems. Specifically, DAMap consists of three components: Distance-aware Focal Loss (DAFL), Hybrid Loss Scheme (HLS), and Task Modulated Deformable Attention (TMDA). The DAFL is introduced to assign appropriate classification labels for one-to-many matching samples. The TMDA is proposed to obtain discriminative task-specific features. Furthermore, the HLS is proposed to better utilize the advantages of the DAFL. We perform extensive experiments and consistently achieve performance improvement on the NuScenes and Argoverse2 benchmarks under different metrics, baselines, splits, backbones, and schedules.

1. Introduction

High-definition (HD) map is an important component developed specifically for autonomous driving vehicles. Traditional offline HD maps are constructed using SLAM-based methods [27, 35] that require manual annotation on the Lidar point clouds. The offline pipeline is complex, high-cost, and difficult to keep up-to-date frequently. To address these challenges, there has been a growing interest

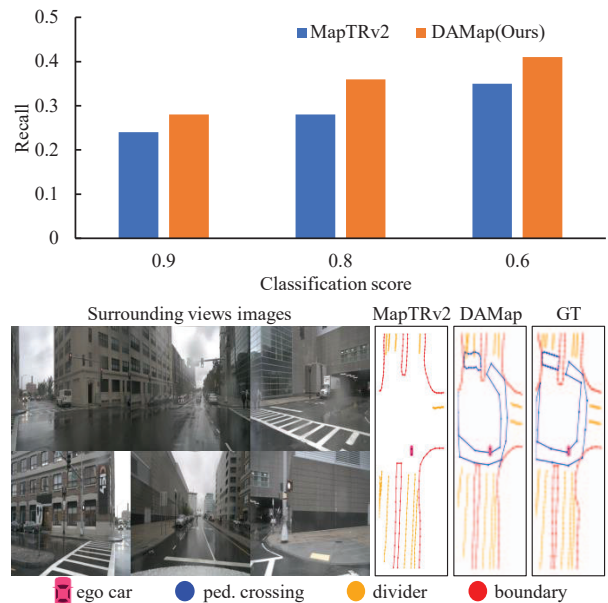


Figure 1. Top: The analysis about recall at 0.5m threshold (high localization). Even with the 0.6 classification threshold, the recall is only 35%. This indicates that the baseline performs poorly in high quality predictions. Our method improves the recall of high quality predictions at each score threshold. Bottom: A visualization example with a 0.4 classification threshold. At this threshold, the baseline model fails to predict pedestrian crossings, which is dangerous for autonomous driving vehicles.

in the development of online HD map. Inspired by DETR [3, 7, 18, 34, 38] and BEV perception [10, 14, 24], MapTR [15] designs a transformer-based pipeline with permutation-equivalent modeling and hierarchical query embedding for more efficient HD map construction. MapTRv2 [16] builds upon the MapTR framework by incorporating decoupled self-attention to reduce memory consumption and the one-to-many matching to speed up convergence.

Predicting map elements with high classification and lo-

*Corresponding author.

calization quality is crucial to the safety of autonomous driving vehicles. Although MapTRv2 [16] yields promising performance gains, the performance of high quality predictions is still poor, as analyzed in Figure 1. We argue that two inherent defects in current methods limit high quality predictions as follows:

Inappropriate task labels is that one-to-many matching samples share the same labels easily leads to the misalignment results of high classification and low localization. When one-to-many matching branches are introduced, each GT instance corresponds to several different candidate samples. For example, each GT map element instance in the MapTRv2 [16] corresponds to 7 candidate samples. When using loss backpropagation training, different candidate samples are given the same classification labels, ignoring the differences in localization of different samples. This way allows low localization samples to output high classification scores. Many results with high classification but low localization will result in poor performance. Based on the above observations, we propose to assess the localization quality of these candidate samples and assign proper labels based on this assessment. In this way, the potential of one-to-many branching can be better exploited, and low quality predictions can be suppressed.

Sub-optimal task features are that task-shared features may cause inherent conflicts between tasks, leading to sub-optimal task outputs. Cross-attention layer (deformable attention) is used to extract the instance features through the interaction between queries and image features. In the current map construction framework, the classification and localization branches use the same instance features. However, classification and localization tasks usually have different feature preferences. Conventional deformable attention does not extract task-specific features. This is not conducive to obtaining more accurate results for both tasks. Typically, utilizing task modulated methods tends to obtain more flexible feature representations. So, we argue that deformable attention in the decoder can be modulated to extract discriminative task-specific features.

In this paper, we propose the DAMap, a simple yet effective HD map construction method that uses three components to obtain high quality predictions. Specifically, Distance-aware Focal Loss (DAFL) is proposed to assign appropriate classification labels for one-to-many matching samples by introducing the localization quality of candidate samples into the classification loss. The DAFL utilizes likelihood estimation to convert localization loss into localization quality. Task Modulated Deformable Attention (TMDA) is introduced to obtain discriminative task-specific sampling features by learning task-specific sampling weights in the cross-attention layer. Furthermore, to better utilize the advantages of our DAFL, we propose a Hybrid Loss Scheme (HLS) to use Focal Loss [17] and our

DAFL in different decoders during the training stage. We perform extensive experiments to validate the effectiveness of DAMap on the NuScenes [1] and Argoverse2 [30] benchmark. Our method consistently achieves performance improvement. We summarize our contributions as follows:

- We analyze two inherent defects in the current methods, inappropriate classification labels and sub-optimal task features, which limit the predictions with high classification and localization quality.
- A new online vectorized HD map construction method named DAMap is proposed to address these defects with Distance-aware Focal Loss, Task Modulated Deformable Attention, and Hybrid Loss Scheme.
- We validate the proposed DAMap on the NuScenes and Argoverse2 benchmarks under different metrics, baselines, splits, backbones, and schedules, and it consistently yields performance improvement.

2. Related Work

2.1. HD Map Construction

HDMaNet [12] proposes to group pixel-level segmentation results into vectorized representations. VectorMapNet [21] further proposes a two-stage coarse-to-fine framework with the auto-regressive decoder to sequentially generate points. Inspired by DETR[3], MapTR [15] designs an end-to-end DETR-like pipeline with permutation-equivalent point set modeling and hierarchical query embedding. MapTRv2 [16] builds upon the MapTR by incorporating decoupled self-attention to reduce memory consumption and the one-to-many matching branch to speed up the convergence. Piece-wise Bézier curve modeling is introduced into BeMapNet [26], and PivotNet [5] utilizes dynamic pivot-based points to represent map elements. MGMap [20] utilizes the guidance of learned masks to enhance element features in detail. HIMap [37] proposes the hybrid representation learning with a point-element interaction module and a point-element consistency constraint scheme. MapQR [23] replaces the point-element query with the proposed scatter-and-gather query to address content conflicts that points of the same map element may have different content features. In addition, StreamMapNet [32] proposes a long-sequence temporal modeling pipeline for complex scenarios and broader perception ranges. In contrast to these works, our focus lies on task labels and features for high quality predictions. In addition, we conducted experiments using these SOTA methods as the baseline models, and our method is complementary to theirs.

2.2. High Quality Learning

Although high quality learning is less focused on HD map construction, there exists some works [2, 8, 11, 13, 25, 33] to improve high quality predictions for other tasks. For

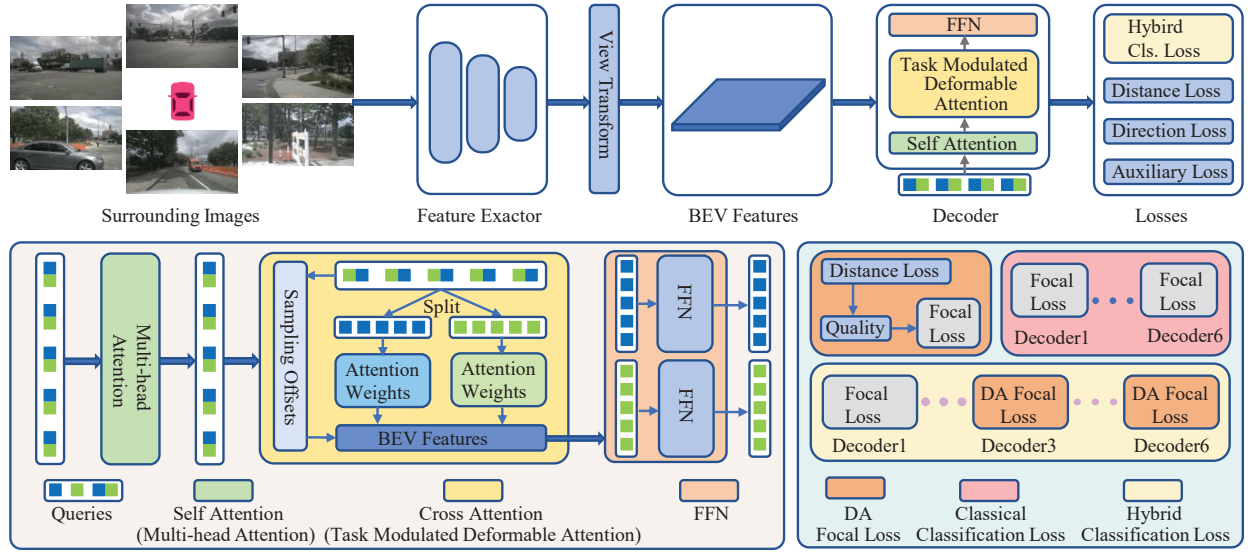


Figure 2. The pipeline of the proposed DAMap. Taking surrounding images as inputs, a shared backbone as the feature extractor first extracts image features; The extracted image features are then transformed to BEV space as BEV features; The transform decoder finally achieves the interaction between queries and BEV features to predict instances of map elements. These predictions are used to calculate the loss for model optimization in the training phase.

example, IoU-Net [11] introduces an auxiliary Intersection over Union (IoU) head to output IoU scores for post-processing. GFL [13] designs a quality focal loss to obtain the joint representation of classification scores and IoU scores. VFNet [33] proposes a novel IoU-based classification loss and a refinement scheme to obtain more accurate results. Since these methods achieved success in the convolutional era, some recent works like Stable-DINO [19], Align-DETR [2], and Rank-DETR [25] introduce these ideas into the DETR framework. In addition to assigning an appropriate label, there are some works [6, 9, 28, 31, 36] that explore task feature enhancement. Double-Head RCNN [31] uses two separate branches to decouple the sibling task head. TSD [28] proposes a spatial disentanglement module to generate different proposals for different tasks. For the DETR [3] framework, DEST [9] and Decoupled DETR [36] utilize two cross-attention (multi-head attention) modules for different tasks. Although these methods have been successful, their efficacy in the HD map construction requires further validation. Furthermore, our main contribution, assigning an appropriate label without explicit IoU representation, designing the task modulated deformable attention, and hybrid using loss in different decoders, has not yet been explored in previous works.

3. Method

3.1. Framework Overview

Taking multi-view RGB images as inputs, our framework is to predict instances of map elements. The overall pipeline

of our DAMap is shown in Figure 2. Giving multi-view RGB images as input, a shared backbone as features extractor is first used to extract image features. Then, these extracted multi-view image features are transformed to BEV space with a pre-defined size $H \times W$ through the view transform module. The generated BEV features are denoted as $\mathbf{F} \in \mathbb{R}^{C \times H \times W}$, where C represents the number of channels. Finally, the DETR-like transform decoder is used to achieve the interaction between queries and BEV features and to generate instances of map elements.

3.2. Preliminary

Focal Loss (FL) [17] is a classical classification loss designed to address extreme imbalance between positive and negative samples during the object detector training stage. It is also used as the classification loss in the HD map construction task. It can be simply formulated as:

$$\text{FL}(p, y) = \begin{cases} -(1-p)^\gamma \log(p), & y = 1 \\ -p^\gamma \log(1-p), & y = 0 \end{cases} \quad (1)$$

where $y \in \{1, 0\}$ indicates the ground truth class and $p \in [0, 1]$ denotes the predicted classification probability score for the foreground class. γ is the hyper-parameter to adjust the effect of the modulation factor. As shown in Equation 1, focal loss consists of a modulation scaling factor part $(1-p)^\gamma$ and a conventional cross entropy part $-\log(p)$. Specifically, the modulation scaling factor $(1-p)^\gamma$ and p^γ correspond to the foreground and background classes, respectively.

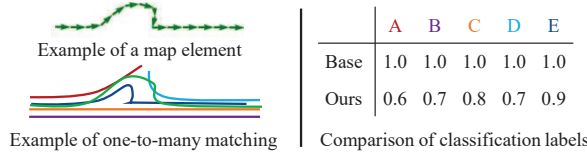


Figure 3. A simple example of inappropriate classification labels for one-to-many matching queries. The green line represents GT, and the other colored lines represent queries.

Decoder is used to extract the features of map instances through interaction between queries and image features. The decoder consists of a self-attention layer, a cross-attention layer, and an FFN layer. Let $\mathbf{Q}$ denote the pre-defined queries and $\mathbf{F}$ denote the generated BEV features. The pipeline of the decoder can be simply formulated as:

$$\hat{\mathbf{Q}} = \{\text{SA}(\mathbf{Q}), \text{CA}(\mathbf{Q}, \mathbf{F}), \text{FFN}(\mathbf{Q})\}, \quad (2)$$

where $\text{SA}(\cdot)$ is the self-attention layer that performs interactions between different queries. $\text{CA}(\cdot)$ is the cross-attention layer that extracts the instance features from the image features. In MapTRv2 [16], Deformable Attention is used to construct a cross-attention layer. $\text{FFN}(\cdot)$ is the feed-forward network layer that enhances the instance features. The normalization layers are omitted in the pipeline.

After obtaining the instance features extracted by the decoder, the instance features are fed into the classification and localization task heads, consisting of linear layers, to output the predictions. The process can be formulated as:

$$\mathbf{O}_{\text{cls}} = H_{\text{cls}}(\hat{\mathbf{Q}}), \mathbf{O}_{\text{loc}} = H_{\text{loc}}(\hat{\mathbf{Q}}), \quad (3)$$

where H_{cls} indicates the classification head, H_{loc} indicates the localization head. The final predictions are usually obtained by cascading multiple decoders and task heads.

3.3. Distance-aware Focal Loss

Although the integration of Focal Loss can speed up convergence and yield promising performance gains, its modulation factor, considering only the classification score of samples, is not optimal. When one-to-many matching branches are introduced, each GT instance corresponds to several different candidate samples. However, these different candidate samples are given the same label of 1, which is not conducive to obtaining a more discriminative model, as shown in Figure 3. We need to assess the localization quality of these candidate samples and assign appropriate labels based on this localization assessment.

The localization loss can be used to indicate the localization quality, so we first introduce the localization loss (distance loss) as follows:

$$\mathcal{L}_{\text{dist}} = \sum_{i=0}^{N-1} \mathbb{1}_{\{c_i \neq \emptyset\}} \sum_{j=0}^{N_v-1} D_{\text{L1}}(\hat{v}_{\hat{\pi}(i),j}, v_{i,\hat{\gamma}_i(j)}), \quad (4)$$

where i is the index of GT instance, j is the index of point in GT instance, $\hat{\gamma}_i$ denotes the point-level matching result, $\hat{\pi}$ denotes instance-level matching result. $\hat{v}_{\hat{\pi}(i),j}$ represents each predicted point, $v_{i,\hat{\gamma}_i(j)}$ represents the assigned a GT point. The localization loss is obtained by computing the L1 distance between predicted points and assigned GT points.

The large range of values for the distance loss ($\mathcal{L}_{\text{dist}} \in [0, +\infty]$) causes it not to be used as directly as the classification score ($p \in [0, 1]$). To align the classification scores, we need to represent the distance loss as a probability to indicate the localization quality. The distance loss can be converted into a likelihood probability using maximum likelihood estimation theory, as follows:

$$P_{\text{dist}}^i = e^{-\lambda \mathcal{L}_{\text{dist}}^i}, \quad (5)$$

where P_{dist}^i denotes the instance localization confidence to indicate the localization quality. When $\mathcal{L}_{\text{dist}}^i = 0$, P_{dist}^i is equal to 1. When $\mathcal{L}_{\text{dist}}^i$ tends to infinity, P_{dist}^i tends to 0. λ is a hyper-parameter to adjust the effect of loss.

After obtaining the localization probability, we can use the localization probability as continuous labels to assign more reasonable labels to different samples. Specifically, we replace the binary labels in Focal Loss with continuous labels and thus obtain the proposed Distance-aware Focal Loss (DAFL). The DAFL is defined as:

$$\text{DAFL}(p, y) = -(y - p)^\gamma (y \log(p) + (1 - y) \log(1 - p)), \quad (6)$$

where $y \in [0, 1]$ denotes the localization confidence. As shown in Equation 6, when the sample is a negative sample ($y = 0$), the DAFL is the same as the raw Focal Loss. When the samples are positive, the localization confidence will be used as the classification label as a supervised signal so that the classification score can indicate the localization quality.

Our proposed DAFL can mitigate the issue of the classification task not perceiving the localization quality, thus avoiding the output of predictions with high classification and low localization quality.

3.4. Hybrid Loss Scheme

MapTRv2[16] typically stacks 6 decoder layers to output predictions using a cascade style. All decoder layers are supervised by classification and localization loss. The total classification loss can be simply formulated as:

$$\mathcal{L}_{\text{cls}} = \sum_{l=1}^L \text{FL}(p, y), \quad (7)$$

where L is the number of decoder layers.

To introduce our proposed DAFL into the classification loss, a straightforward way is to replace the Focal Loss in all decoders with our proposed DAFL. The total classification

loss is converted as:

$$\mathcal{L}_{\text{cls}} = \sum_{l=1}^L \text{DAFL}(p, y). \quad (8)$$

The random initialization of the query in MapTRv2 [16] leads to poor localization results at the early stage of training, which may affect our proposed DAFL.

To take full advantage of our proposed DAFL, we propose the hybrid loss scheme to combine the Focal Loss [17] with the proposed DAFL. Specifically, we use Focal Loss as the classification loss in the earlier decoder layer and DAFL as the classification loss in the later decoder layer. Focal Loss is independent of the localization quality of samples. Predictions of the later decoders tend to have a higher localization quality than the earlier ones due to the cascade property of the decoders. The total classification loss can be simply formulated as:

$$\mathcal{L}_{\text{cls}} = \sum_{l=1}^{L1} \text{FL}(p, y) + \sum_{l=1}^{L2} \text{DAFL}(p, y), \quad (9)$$

where $L1$ is the number of decoder layers with Focal Loss, $L2$ is the number of decoder layers with DAFL. The sum of $L1$ and $L2$ is equal to L .

3.5. Task Modulated Deformable Attention

Sharing the same instance features may cause inherent conflicts across tasks, leading to sub-optimal task output results. For example, features in salient regions within instances may have rich semantic information for classification tasks, while features at instance boundaries may be more conducive to localization tasks. The shared paradigm in current decoder designs prevents decoders from obtaining more discriminative task-specific features to achieve further performance gains.

To deal with the above potential issue, we propose the Task Modulated Deformable Attention to obtain discriminative task-specific features. Specifically, we first double the channels of the initialized query $\mathbf{Q}$, with half of the channels responsible for classification and the other half for localization. Then, as with the conventional decoder, the queries utilize the self-attention layer for interactions between different queries. Next, we split the queries obtained from the self-attention layer to generate task-aware queries $\mathbf{Q}_{\text{cls}}$ and $\mathbf{Q}_{\text{loc}}$ as follows:

$$\mathbf{Q}_{\text{cls}}, \mathbf{Q}_{\text{loc}} = \text{Split}(\text{SA}(\mathbf{Q})), \quad (10)$$

where $\mathbf{Q} \in \mathbb{R}^{2C}$, $\mathbf{Q}_{\text{cls}} \in \mathbb{R}^C$, $\mathbf{Q}_{\text{loc}} \in \mathbb{R}^C$. $\text{Split}(\cdot)$ represents the split operation for the channels.

After obtaining the task-aware queries, we feed them to the cross-attention layer to obtain the task-specific features.

It is worth noticing that we do not simply use two cross-attention layers (Deformable Attention module [38]) to process task-aware queries separately, but utilize our proposed Task Decoupled Deformable Attention module to process them. The standard deformable attention module can be simply formulated as:

$$\mathbf{A} = W_a \mathbf{Q}, \Delta r = W_p \mathbf{Q}, \mathbf{V} = \text{Samp}(W_v \mathbf{F}, r + \Delta r), \\ \text{DeformableAttention}(\mathbf{Q}, \mathbf{V}) = \text{Softmax}(\mathbf{A})\mathbf{V}, \quad (11)$$

where Δr are the offsets of reference points, r are the reference points, W_a , W_p and W_v represent the linear projection. $\mathbf{A}$ are the attention weights of the sample features, Samp is a function that extracts the features corresponding to the location $(r + \Delta r)$ by bilinear interpolation.

Our proposed Task Decoupled Deformable Attention module utilizes task-aware query to learn task-specific attention weights, while offset learning remains shared. The reason for this design is that we argue that 1) it is difficult to simultaneously learn task-specific weights and offsets using the task-aware query due to the large number of variables to be optimized, and 2) offset learning itself is more difficult to optimize because its optimization objective is unbounded. Our proposed Task Decoupled Deformable Attention module can be simply formulated as:

$$\mathbf{A}_{\text{cls}} = W_a \mathbf{Q}_{\text{cls}}, \mathbf{A}_{\text{loc}} = W'_a \mathbf{Q}_{\text{loc}} \\ \Delta r = W_p \text{Cat}(\mathbf{Q}_{\text{cls}}, \mathbf{Q}_{\text{loc}}), \mathbf{V} = \text{Samp}(W_v \mathbf{F}, r + \Delta r), \\ \hat{\mathbf{Q}}_{\text{cls}}, \hat{\mathbf{Q}}_{\text{loc}} = \text{Softmax}(\mathbf{A}_{\text{cls}})\mathbf{V}, \text{Softmax}(\mathbf{A}_{\text{loc}})\mathbf{V}, \quad (12)$$

where Cat represents the concatenation operation.

Next, we feed the task-specific features extracted by the cross-attention layer into different FFN networks, respectively. Finally, task-specific features are fed into the two task heads to output the predictions of the map element.

4. Experiments

4.1. Datasets

NuScenes. NuScenes [1] dataset includes 1000 driving scenes recorded in Boston and Singapore. Each scene lasts about 20 seconds and is sampled at 2Hz to get 40 key-frame samples. For each sample, there are 6 RGB images from surrounding cameras that provide the full 360-degree field of view. Following previous methods, 700/150/150 split scenes are utilized for training/validation/testing, respectively. We report the performance of three categories, including the road boundary, the lane divider, and the pedestrian crossing, on the validation set. The ground truth in the NuScenes is annotated as the 2D vectorized map elements. **Argoverse2.** Argoverse2 [30] dataset includes 1000 driving scenes recorded in six cities. Each scene lasts about 15 seconds and is sampled at 10Hz to get 150 key-frame samples.

Methods	Reference	Epoch	AP _{ped.}	AP _{div.}	AP _{bou.}	mAP	AP _{ped.}	AP _{div.}	AP _{bou.}	mAP
			hard : {0.2, 0.5, 1.0}m				easy : {0.5, 1.0, 1.5}m			
HMapNet [12]	ICRA22	30	7.1	28.3	32.6	22.7	24.1	23.6	43.5	31.4
PivotNet [5]	CVPR23	30	34.8	42.9	39.3	39.0	53.8	58.8	59.6	57.4
BeMapNet [26]	CVPR23	30	39.0	46.9	37.8	41.3	57.7	62.3	59.4	59.8
MapTR [15]	ICLR23	24	23.2	30.7	28.2	27.3	46.3	51.5	53.1	50.3
StreamMapNet [32]	WACV24	30	-	-	-	-	61.7	66.3	62.1	63.4
MapTRv2 [16]	IJCV24	24	35.4	40.0	36.3	37.2	59.8	62.4	62.4	61.5
MapTRv2+MGMap [20]	CVPR24	24	-	-	-	-	61.8	65.0	67.5	64.8
HIMap [37]	CVPR24	30	37.2	49.2	42.8	43.1	62.6	68.4	69.1	66.7
SQD-MapNet [29]	ECCV24	24	-	-	-	-	63.6	66.6	64.8	65.0
MapTRv2 [†] [16]	IJCV24	24	34.2	39.3	36.2	36.6	58.1	60.8	62.3	60.4
MapQR [23]	ECCV24	24	38.6	49.9	41.5	43.3	63.4	68.0	67.7	66.4
Mask2Map [4]	ECCV24	24	-	-	-	-	70.6	71.3	72.9	71.6
MapTRv2+Ours	-	24	34.7	43.2	39.1	39.0(+2.4)	58.5	64.7	65.1	62.8(+2.4)
MapQR+Ours	-	24	40.3	52.4	45.3	46.0(+2.6)	65.2	70.8	70.3	68.8(+2.4)
Mask2Map+Ours	-	24	55.6	62.2	59.4	51.2	70.7	72.3	74.9	72.6(+1.0)
VectorMapNet [21]	ICML23	110	18.2	27.2	18.4	21.3	36.1	47.3	39.3	40.9
MapTR [‡] [15]	ICLR23	110	31.4	40.5	35.5	35.8	55.8	60.9	61.1	59.3
BeMapNet [26]	CVPR23	110	44.5	52.7	44.2	47.1	62.6	66.7	65.1	64.8
MapTRv2 [16]	IJCV24	110	-	-	-	-	68.1	68.3	69.7	68.7
MGMap [20]	CVPR24	110	54.5	42.1	37.4	44.7	64.4	67.6	67.7	66.5
MapTRv2+SGQ [23]	ECCV24	110	44.0	52.9	45.3	47.4	67.7	69.8	70.4	69.3
MapTRv2 [†] [16]	IJCV24	110	43.4	48.2	42.9	44.9	67.5	68.3	69.1	68.3
MapTRv2+HLS (Ours)	-	110	44.8	50.0	46.3	47.0(+2.1)	68.2	69.7	70.5	69.4(+1.1)
MapTRv2+Ours	-	110	44.0	51.6	46.6	47.4(+2.5)	68.6	71.3	71.2	70.4(+2.1)

Table 1. **Comparison to the state-of-the-art on nuScenes val set with ResNet-50.** The [†] is the reproduced result from public code. Our reproduced models trained on 4 NVIDIA GeForce RTX 3090 GPUs with batch size 16 and 3×10^{-4} with cosine decay, which is half the learning rate at total batch size 32. So some reproduced results are low than the results with batch size 32 from paper. The numbers in brackets are corresponding mAP gains compared with reproduced baseline models under the same settings. The [‡] means the results from released model. “-” means that the corresponding results are not available. The metric of easy setting and hard setting are follow [5].

For each sample, there are 7 RGB images from surrounding cameras. Following previous methods, 700/150 split scenes with 110K/24K frames are used for training/validation. We report the performance of the same categories as NuScenes on the validation set. The ground truth in Argoverse2 is annotated as the 3D vectorized map elements.

4.2. Metric

Following previous methods, the mean Average Precision (mAP) metric based on Chamfer Distance is chosen for evaluation. A prediction that satisfies a specific chamfer distance threshold is defined as a True-Positive (TP). Two different threshold sets (hard and easy) are adopted for a more comprehensive evaluation. The hard and easy settings correspond to {0.2, 0.5, 1.0}m and {0.5, 1.0, 1.5}m, respectively. The prediction map range is $[-15.0m, 15.0m]$ for the X -axis and $[-30.0m, 30.0m]$ for the Y -axis.

4.3. Implementation Details

We adopt the ResNet50 as the backbone to conduct experiments unless otherwise specified. The total batch size is 16

and the learning rate is 3×10^{-4} . All models are trained on 4 NVIDIA GeForce RTX 3090 GPUs unless otherwise specified. The training schedule is 24 epochs for NuScenes and 6 epochs for Argoverse2 unless otherwise specified. Other settings are the same as the baseline MapTRv2 [16].

4.4. Main Results

Results on nuScenes. For nuScenes [1], different methods output 2D vectorized HD map elements for evaluation. Table 1 shows the comparison of the experimental results against the SOTA methods. Combining our components with MapTRv2, our proposed DAMap achieves 62.8 and 39.0 mAP, which are 2.4 and 2.4 mAP higher than our reimplemented MapTRv2 baseline with the 24 epochs setting. Since recent work MapQR [23] and Mask2Map [4] have shown good performance, we also chose them as the baselines to validate the generality of our method. Combining our components with MapQR, our proposed DAMap achieves 68.8 and 46.0 mAP, which are 2.4 and 2.6 mAP higher than our reimplemented MapQR baseline with the 24 epochs. Combining our components with Mask2Map, our

Methods	Backbone	Epoch	AP _{ped.}	AP _{div.}	AP _{bou.}	mAP	AP _{ped.}	AP _{div.}	AP _{bou.}	mAP
			<i>hard</i> : {0.2, 0.5, 1.0}m				<i>easy</i> : {0.5, 1.0, 1.5}m			
MapQR [23]	SwinB	24	41.4	52.3	43.2	45.6	67.5	72.1	70.5	70.1
MapQR [†] [23]	SwinB	24	41.6	53.2	46.3	47.0	66.9	71.9	71.0	70.0
MapQR+Ours	SwinB	24	43.2	54.6	47.9	48.6(+1.6)	68.7	73.9	73.2	71.9(+1.9)

Table 2. **Comparison to the state-of-the-art on nuScenes val set with SwinB.** The † is the reproduced result from public code.

Methods	Dim	Epoch	AP _{ped.}	AP _{div.}	AP _{bou.}	mAP	AP _{ped.}	AP _{div.}	AP _{bou.}	mAP
			<i>hard</i> : {0.2, 0.5, 1.0}m				<i>easy</i> : {0.5, 1.0, 1.5}m			
MapTRv2 [16]		6	30.5	48.1	36.7	38.4	60.7	68.9	64.5	64.7
MapTRv2 [†] [16]		6	30.1	47.9	36.4	38.1	58.4	69.0	63.4	63.6
MapTRv2+Ours		6	32.5	49.7	40.5	40.9(+2.8)	61.0	70.3	67.2	66.2(+2.6)
MapQR [23]		6	30.6	49.4	38.9	39.6	60.1	71.2	66.2	65.9
MapQR [†] [23]	3	6	33.9	50.1	39.5	41.1	61.3	68.9	66.1	65.4
MapQR+Ours		6	35.3	52.4	43.3	43.6(+2.5)	62.9	70.9	68.4	67.4(+2.0)
MapTRv2 [†] [16]		24	36.3	52.3	41.8	43.5	65.1	72.2	68.1	68.5
MapTRv2+Ours		24	36.0	54.2	45.0	45.0(+1.5)	64.7	73.2	70.2	69.4(+0.9)
MapQR [†] [23]		24	36.7	54.5	44.0	45.1	63.3	71.0	69.3	67.9
MapQR+Ours		24	37.6	55.6	46.2	46.5(+1.4)	64.4	72.1	69.8	68.8(+0.9)

Table 3. **Comparison to the state-of-the-art on Argoverse2 val set with ResNet-50.** The † is the reproduced result from public code. The reproduced setting is same as the nuScenes. The numbers in brackets are corresponding mAP gains.

proposed DAPMap achieves 72.6 mAP, which is 1.0 mAP higher than Mask2Map baseline with the 24 epochs.

To evaluate the effectiveness of our method under a longer training schedule, we perform experiments under the 110 epochs training schedule. Our method achieves 70.4 and 47.4 mAP, which are 2.1 and 2.5 mAP higher than the baseline. Besides, we conduct the experiment of introducing only our HLS with DAFL into the baseline MapTRv2 [16]. It is worth noting that our HLS with DAFL as the loss function does not introduce parameters and computational effort in the inference stage. Our proposed HLS with DAFL achieves 69.4 and 47.0 mAP, which are 1.1 and 2.1 mAP higher than the baseline MapTRv2.

In addition, our DAPMap can consistently obtain performance gains even with a stronger Swin-Transformer [22] backbone. As shown in Table 2, when using SwinB as the backbone, our DAPMap achieves 71.9 and 48.6 mAP, which are 1.9 and 1.6 mAP higher than the baseline. These results validate the effectiveness and generalizability of our method on different metrics, baselines, schedules, and backbones.

Results on Argoverse2. For Argoverse2 [30], different methods output 3D vectorized HD map elements for evaluation. We also evaluate our proposed DAPMap on the Argoverse2 val dataset under different metrics (hard and easy), baselines (MapTRv2 [16] and MapQR [23]), and schedules (6 and 24 epochs). Table 3 shows the comparison of the experimental results. Our proposed DAPMap consistently obtains performance improvement under all settings. Specifically, under the 6 epochs schedule, our method outperforms MapTRv2 by 2.6 and 2.8 mAP and MapQR by 2.0

Methods	AP _{ped.}	AP _{div.}	AP _{bou.}	mAP
MapTRv2 [†] [16]	12.1	26.3	42.2	26.9
MapTRv2+Ours	12.3	27.6	45.8	28.6(+1.7)
StreamMapNet [32]	32.2	29.3	40.8	34.1
StreamMapNet+Ours	35.1	27.8	43.8	35.6(+1.5)

Table 4. **Comparison to the baseline methods on new nuScenes split val set under the ResNet50 backbone.** The † is the reproduced result from public code.

and 2.5 mAP, respectively. Under the 24 epochs schedule, our method outperforms MapTRv2 by 0.9 and 1.5 mAP and MapQR by 0.9 and 1.4 mAP, respectively. These results validate the effectiveness and generalizability of our method on the different datasets.

Results on New nuScenes Split. To ensure a fair comparison, we also perform experiments on the new nuScenes split proposed by StreamMapNet [32]. As shown in Table 4, our proposed DAPMap demonstrates superior performance over baseline methods across all categories in the new nuScenes split. The results show the effectiveness of our method on different dataset splits.

Results with Centerline Category. To investigate the effectiveness for more categories, we conduct experiment of our method with the Centerline category. As shown in Table 5, our DAPMap can obtain performance gains even with the Centerline category. Specifically, our DAPMap achieves 58.3 mAP and 62.5 mAP in nuScenes and Argoverse2, respectively, which is 5.1 mAP and 4.0 mAP higher than the reproduced baseline. These results further validate the ef-

Methods	Data.	$AP_{ped.}$	$AP_{div.}$	$AP_{bou.}$	$AP_{cen.}$	mAP
MapTRv2[16]	nuS	50.1	53.9	58.8	53.1	54.0
MapTRv2 [†] [16]	nuS	47.7	52.8	59.2	53.2	53.2
MapTRv2+Ours	nuS	51.0	60.6	64.5	57.2	58.3
MapTRv2 [†] [16]	Arg2	51.6	62.1	61.3	58.9	58.5
MapTRv2+Ours	Arg2	53.5	68.6	64.7	63.3	62.5

Table 5. **Comparison to the the baseline methods on nuScenes and Argoverse2 val set with Centerline category under the ResNet50.** The [†] is the reproduced result from public code.

DAFL	HLS	TMDA	$AP_{ped.}$	$AP_{div.}$	$AP_{bou.}$	mAP
-	-	-	58.1	60.8	62.3	60.4
✓			58.6	61.1	63.1	60.9
✓	✓		58.5	62.9	63.4	61.6
		✓	58.5	63.1	63.2	61.6
✓	✓	✓	58.5	64.7	65.1	62.8

Table 6. **Ablations on each component of our DAMap.** "DAFL", "HLS" and "TMDA" represent Distance-aware Focal Loss, Hybrid Loss Scheme and Task Modulated Deformable Attention, respectively.

fectiveness and generalizability of our method.

4.5. Ablation Study

Ablation on Each Component. To analyze the effect of each component in our DAMap, we gradually apply Distance-aware Focal Loss (DAFL), Hybrid Loss Scheme (HLS), and Task Modulated Deformable Attention (TMDA) to the baseline. As shown in Table 6, the DAFL produces 0.5 mAP gains compared to the baseline. The result shows that introducing the localization quality into the Focal Loss is beneficial to the model. The HLS with DAFL improves the performance from 60.4 mAP to 61.6 mAP. This result suggests that our HLS can take greater advantage of the DAFL. The TMDA brings 1.2 mAP gains compared to the baseline. When all components are introduced into the baseline, it achieves 62.8 mAP with 2.4 mAP gains. Hyperparameter experiments on these components and runtime analysis can be found in the Supplementary Material. The results show that these components are complementary and address different issues in the HD map task.

Ablation on TMDA Design. To validate the effectiveness of our TMDA design, we conduct experiments on other designs with the same complexity. As shown in Table 7, other designs can obtain a slight gain. For example, the design (Setting 1) of task-specific sampling offsets and attention weights generated by the task-aware query can bring 0.3 mAP gains. This design is equivalent to having a separate cross-attention layer for each of the classification and localization tasks. Our design (Setting 4) of task-agnostic sampling offsets generated by task-agnostic query and the task-specific attention weights generated by task-aware query can achieve 1.2 mAP gains. The results show that our method outperforms the other designs. In addition, we also

Setting	$AP_{ped.}$	$AP_{div.}$	$AP_{bou.}$	mAP	Para.
-	58.1	60.8	62.3	60.4	40M
1	57.5	61.8	62.9	60.7	52M
2	57.4	63.0	62.3	60.9	52M
3	57.0	62.6	61.7	60.4	52M
4 (Ours)	58.5	63.1	63.2	61.6	52M

Table 7. **Ablations on our TMDA with other designs.** 1: task-specific sampling offsets and attention weights are generated by task-aware query. 2: task-specific sampling offsets and attention weights are learned by task-agnostic query (concatenated task-aware query). 3: task-specific sampling offsets and task-agnostic attention weights are learned by task-agnostic query.

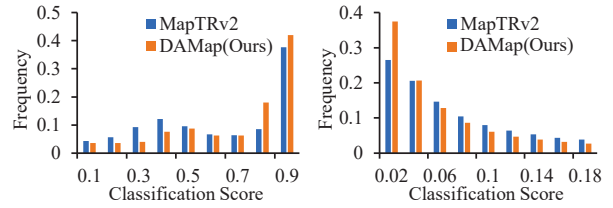


Figure 4. Left: Frequency of the classification scores on the True Positive with or without our method. Right: Frequency of the classification scores on the False Positive with or without our method.

show the attention weights visualization of TMDA in the Supplementary Material.

Quantitative Analysis of Classification Score. To further illustrate that our method can produce high quality predictions, we count the classification scores of the predictions. As shown in Figure 4, our method improves the percentage of predictions with high classification score for True Positive (Left) and the percentage of predictions with low classification score for False Positive (Right). The results show that our method can encourage positive samples and suppress negative samples. This phenomenon is further evidence that our method can produce high quality predictions. Quantitative analysis of localization quality can be found in the Supplementary Material.

5. Conclusion

In this paper, we analyze two inherent defects in current methods that limit the high quality predictions. To address these problems, we propose a novel map construction method named DAMap, which consists of three components: Distance-aware Focal Loss (DAFL), Hybrid Loss Scheme (HLS) and Task Modulated Deformable Attention (TMDA). The DAFL is introduced to assign appropriate labels for one-to-many matching samples. The TMDA is proposed to obtain discriminative task-specific features. Furthermore, the HLS is proposed to better utilize the advantages of the DAFL. Extensive experiments conducted on the nuScenes and Argoverse2 benchmarks under different metrics, baselines, splits, backbones, and schedules validate the effectiveness of our method.

Acknowledgements

This work is supported by National Natural Science Foundation of China (NSFC) under Grant 62088102. We also appreciate Yufeng Hu for his suggestions. We thank all the anonymous reviewers and Area Chairs for their constructive and helpful comments, which have significantly improved the quality of the paper.

References

- [1] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multi-modal dataset for autonomous driving. In *CVPR*, pages 11621–11631, 2020. 2, 5, 6
- [2] Zhi Cai, Songtao Liu, Guodong Wang, Zheng Ge, Xiangyu Zhang, and Di Huang. Align-detr: Improving detr with simple iou-aware bce loss. *arXiv preprint arXiv:2304.07527*, 2023. 2, 3
- [3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229, 2020. 1, 2, 3
- [4] Sehwan Choi, Jungho Kim, Hongjae Shin, and Jun Won Choi. Mask2map: Vectorized hd map construction using bird’s eye view segmentation masks. In *ECCV*, pages 19–36, 2024. 6
- [5] Wenjie Ding, Limeng Qiao, Xi Qiu, and Chi Zhang. Pivotnet: Vectorized pivot learning for end-to-end hd map construction. In *ICCV*, pages 3672–3682, 2023. 2, 6
- [6] Jinpeng Dong, Yuhao Huang, Songyi Zhang, Shitao Chen, and Nanning Zheng. Construct effective geometry aware feature pyramid network for multi-scale object detection. In *AAAI*, pages 534–541, 2022. 3
- [7] Jinpeng Dong, Yutong Lin, Chen Li, Sanping Zhou, and Nanning Zheng. Augdetr: Improving multi-scale learning for detection transformer. In *ECCV*, pages 238–255, 2024. 1
- [8] Jinpeng Dong, Dingyi Yao, Yufeng Hu, Sanping Zhou, and Nanning Zheng. A novel dense object detector with scale balanced sample assignment and refinement. *TCSVT*, 2025. 2
- [9] Liqiang He and Sinisa Todorovic. Destr: Object detection with split transformer. In *CVPR*, pages 9377–9386, 2022. 3
- [10] Junjie Huang, Guan Huang, Zheng Zhu, Yun Ye, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021. 1
- [11] Borui Jiang, Ruixuan Luo, Jiayuan Mao, Tete Xiao, and Yunying Jiang. Acquisition of localization confidence for accurate object detection. In *ECCV*, pages 784–799, 2018. 2, 3
- [12] Qi Li, Yue Wang, Yilun Wang, and Hang Zhao. Hdmapnet: An online hd map construction and evaluation framework. In *ICRA*, pages 4628–4634, 2022. 2, 6
- [13] Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. In *NeurIPS*, pages 21002–21012, 2020. 2, 3
- [14] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird’s-eye-view representation from multi-camera images via spatiotemporal transformers. In *ECCV*, pages 1–18, 2022. 1
- [15] Bencheng Liao, Shaoyu Chen, Xinggang Wang, Tianheng Cheng, Qian Zhang, Wenyu Liu, and Chang Huang. Maptr: Structured modeling and learning for online vectorized hd map construction. In *ICLR*, 2023. 1, 2, 6
- [16] Bencheng Liao, Shaoyu Chen, Yunchi Zhang, Bo Jiang, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. Maptrv2: An end-to-end framework for online vectorized hd map construction. *IJCV*, pages 1–23, 2024. 1, 2, 4, 5, 6, 7, 8
- [17] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal Loss for Dense Object Detection. In *ICCV*, pages 2999–3007, 2017. 2, 3, 5
- [18] Yutong Lin, Yuhui Yuan, Zheng Zhang, Chen Li, Nanning Zheng, and Han Hu. Detr does not need multi-scale or locality design. In *CVPR*, pages 6545–6554, 2023. 1
- [19] Shilong Liu, Tianhe Ren, Jiayu Chen, Zhaoyang Zeng, Hao Zhang, Feng Li, Hongyang Li, Jun Huang, Hang Su, Jun Zhu, et al. Detection transformer with stable matching. In *ICCV*, pages 6491–6500, 2023. 3
- [20] Xiaolu Liu, Song Wang, Wentong Li, Ruizi Yang, Junbo Chen, and Jianke Zhu. Mgmmap: Mask-guided learning for online vectorized hd map construction. In *CVPR*, pages 14812–14821, 2024. 2, 6
- [21] Yicheng Liu, Tianyuan Yuan, Yue Wang, Yilun Wang, and Hang Zhao. Vectormapnet: End-to-end vectorized hd map learning. In *ICML*, pages 22352–22369, 2023. 2, 6
- [22] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021. 7
- [23] Zihao Liu, Xiaoyu Zhang, Guangwei Liu, Ji Zhao, and Ningyi Xu. Leveraging enhanced queries of point sets for vectorized map construction. In *ECCV*, pages 461–477, 2024. 2, 6, 7
- [24] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *ECCV*, pages 194–210, 2020. 1
- [25] Yifan Pu, Weicong Liang, Yiduo Hao, Yuhui Yuan, Yukang Yang, Chao Zhang, Han Hu, and Gao Huang. Rank-detr for high quality object detection. In *NeurIPS*, 2024. 2, 3
- [26] Limeng Qiao, Wenjie Ding, Xi Qiu, and Chi Zhang. End-to-end vectorized hd-map construction with piecewise bezier curve. In *CVPR*, pages 13218–13228, 2023. 2, 6
- [27] Tixiao Shan and Brendan Englot. Lego-loam: Lightweight and ground-optimized lidar odometry and mapping on variable terrain. In *IROS*, pages 4758–4765, 2018. 1
- [28] Guanglu Song, Yu Liu, and Xiaogang Wang. Revisiting the sibling head in object detector. In *CVPR*, pages 11563–11572, 2020. 3
- [29] Shuo Wang, Fan Jia, Weixin Mao, Yingfei Liu, Yucheng Zhao, Zehui Chen, Tiancai Wang, Chi Zhang, Xiangyu

- Zhang, and Feng Zhao. Stream query denoising for vectorized hd-map construction. In *ECCV*, pages 203–220, 2024. [6](#)
- [30] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaese model Pontes, et al. Argoverse 2: Next generation datasets for self-driving perception and forecasting. In *NeurIPS Datasets and Benchmarks Track*, 2021. [2](#), [5](#), [7](#)
 - [31] Yue Wu, Yinpeng Chen, Lu Yuan, Zicheng Liu, Lijuan Wang, Hongzhi Li, and Yun Fu. Rethinking classification and localization for object detection. In *CVPR*, pages 10186–10195, 2020. [3](#)
 - [32] Tianyuan Yuan, Yicheng Liu, Yue Wang, Yilun Wang, and Hang Zhao. Streammapnet: Streaming mapping network for vectorized online hd map construction. In *WACV*, pages 7356–7365, 2024. [2](#), [6](#), [7](#)
 - [33] Haoyang Zhang, Ying Wang, Feras Dayoub, and Niko Sunderhauf. Varifocalnet: An iou-aware dense object detector. In *CVPR*, pages 8514–8523, 2021. [2](#), [3](#)
 - [34] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In *ICLR*, 2023. [1](#)
 - [35] Ji Zhang, Sanjiv Singh, et al. Loam: Lidar odometry and mapping in real-time. In *RSS*, pages 1–9, 2014. [1](#)
 - [36] Manyuan Zhang, Guanglu Song, Yu Liu, and Hongsheng Li. Decoupled detr: Spatially disentangling localization and classification for improved end-to-end object detection. In *ICCV*, pages 6601–6610, 2023. [3](#)
 - [37] Yi Zhou, Hui Zhang, Jiaqian Yu, Yifan Yang, Sangil Jung, Seung-In Park, and ByungIn Yoo. Himap: Hybrid representation learning for end-to-end vectorized hd map construction. In *CVPR*, pages 15396–15406, 2024. [2](#), [6](#)
 - [38] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR*, 2021. [1](#), [5](#)