

AutoAlignV2: Deformable Feature Aggregation for Dynamic Multi-Modal 3D Object Detection

Zehui Chen¹ Zhenyu Li² Shiquan Zhang³ Liangji Fang³ Qinhong Jiang³ Feng Zhao¹

¹University of Science and Technology of China

²Harbin Institute of Technology ³SenseTime Research

Abstract

Point clouds and RGB images are two general perceptual sources in autonomous driving. The former can provide accurate localization of objects, and the latter is denser and richer in semantic information. Recently, AutoAlign [6] presents a learnable paradigm in combining these two modalities for 3D object detection. However, it suffers from high computational cost introduced by the global-wise attention. To solve the problem, we propose Cross-Domain DeformCAFA module in this work. It attends to sparse learnable sampling points for cross-modal relational modeling, which enhances the tolerance to calibration error and greatly speeds up the feature aggregation across different modalities. To overcome the complex GT-AUG under multi-modal settings, we design a simple yet effective cross-modal augmentation strategy on convex combination of image patches given their depth information. Moreover, by carrying out a novel image-level dropout training scheme, our model is able to infer in a dynamic manner. To this end, we propose AutoAlignV2, a faster and stronger multi-modal 3D detection framework, built on top of AutoAlign. Extensive experiments on nuScenes benchmark demonstrate the effectiveness and efficiency of AutoAlignV2. Notably, our best model reaches 72.4 NDS on nuScenes test leaderboard, achieving new state-of-the-art results among all published multi-modal 3D object detectors. Code will be available at <https://github.com/zehuichen123/AutoAlignV2>.

1. Introduction

3D object detection serves as a fundamental computer vision task in autonomous driving. Modern 3D object detectors [13, 20, 24, 33] have demonstrated promising performance on competitive benchmarks including KITTI [10], Waymo [28], and nuScenes [2] datasets. Despite the rapid progress in detection accuracy, the room for further improvement is still large. Recently, an upsurging stream in combining RGB images and LiDAR points for accurate de-

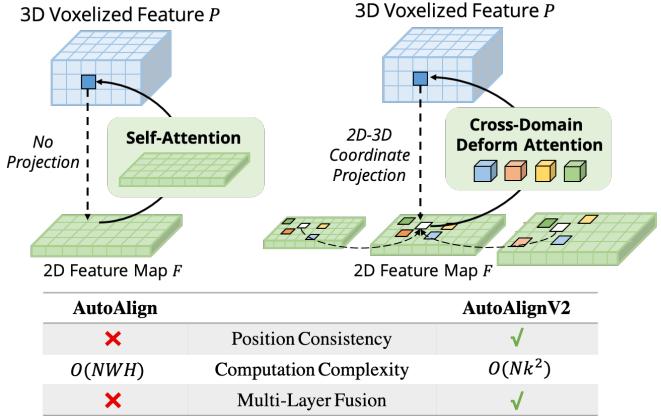


Figure 1. The comparison between AutoAlignV2 and AutoAlign. AutoAlignV2 hints at the alignment module with general mapping relationship guaranteed by deterministic projection matrix, and simultaneously reserves the ability to automatically adjust the positions of feature aggregation. Due to the lightweight computational cost, AutoAlignV2 is able to aggregate multi-layer features for hierarchical imagery information.

tection has drawn many attentions [1, 14, 16, 19, 30, 39]. Different from the point clouds which are beneficial for spatial localization, imagery data are more superior in providing semantic and textural information, *i.e.*, more suitable for classification. Therefore, it is believed that these two modalities are complementary to each other and can further promote the detection accuracy.

However, how to effectively combine these heterogeneous representations for 3D object detection has not been fully explored. In this work, we mainly attribute the current difficulties of training cross-modal detectors to two aspects. On one hand, the fusion strategy in combining imagery and spatial information remains sub-optimal. Due to the heterogeneous representations between RGB images and point clouds, features need to be carefully aligned before being aggregated together. This is often achieved by establishing deterministic correspondence between the point and the image pixel through LiDAR-camera projec-

tion matrix [27, 30, 39]. AutoAlign [6] proposes a learnable global-wise alignment module for automatic registration and achieves good performance. However, it has to be trained with the help of CSFI module to acquire the inner positional matching relationship between points and image pixels. Besides, the complexity of attention-style operation is quadratic to the image size, making it impractical to apply queries on high-resolution feature maps (*e.g.*, C_2 , C_3). Such a limitation can lead to coarse and inaccurate image information and the loss of hierarchical representations brought by FPN (See Figure 1). On the other hand, data augmentation, especially GT-AUG [33], is a crucial step for 3D detectors to achieve competitive results. In terms of multi-modal methods, an important problem is how to keep synchronization between images and point clouds when conducting cut and paste operations. MoCa [39] uses labor-intensive mask annotations in 2D domain for accurate image features. Box-level annotations are also applicable but delicate and complex points filtering is required [30].

In this work, we propose AutoAlignV2 to mitigate the aforementioned issues in a much simpler and more effective way. It hints at the alignment module with the general mapping relationship guaranteed by deterministic projection matrix and simultaneously reserves the ability to automatically adjust the positions of feature aggregation. As for the synchronization issue in 2D-3D joint augmentation, a novel depth-aware GT-AUG algorithm is introduced to cope with object occlusion in the image domain, getting rid of the complex point cloud filtering or the need for delicate mask annotations. We also present a new training scheme named image-level dropout strategy, which enables the model to infer results dynamically even without images. Through extensive experiments, we validate the effectiveness of AutoAlignV2 on two representative 3D detectors: Object DGCNN [32] and CenterPoint [36], and achieve new state-of-the-art performance on the competitive nuScenes benchmark.

2. Related Work

2.1. Object Detection with Point Cloud

Existing 3D object detectors can be broadly categorized as point-based and voxel-based approaches. Point-based methods directly predict the regression boxes from points [26, 35]. For example, Point R-CNN [25] adopts a semantic network to segment the point clouds and then generates the proposals at each foreground point. 3DSSD [34] fully applies point-level predictions on the one-stage architecture, where an anchor-free head is designed after the PointNet-like feature extraction. Although the accurate 3D localization information is maintained, these algorithms often suffer from high computational cost [24]. Different from the point-wise detection, voxel-based approaches transform

sets of unordered points into 2D feature map through voxelization, which can be directly applied with convolutional neural networks [8, 22, 41]. For instance, VoxelNet [41] is a widely-used paradigm where a VFE layer is proposed to extract unified features for each 3D voxel. Based on this, CenterPoint [36] presents a center-based label assignment strategy, achieving competitive performance in 3D object detection.

2.2. Multi-Modal 3D Object Detection

Recently, there has been an increasing attention on multi-modal data for 3D object detection [17, 21]. AVOD [12] and MV3D [5] are two pioneer works in this field, where 2D and 3D ROI are directly concatenated before box prediction. Qi *et al.* [23] utilized images to generate 2D proposals and then lifted them up to 3D space (frustum), which narrows the searching space in point clouds. 3D-CVF [37] and EPNet [11] explore the fusion strategy on feature maps across different modalities with a learned calibration matrix. Though easy-to-implement, they are likely to suffer from coarse feature aggregation. To mitigate this issue, various approaches [27, 29, 39] fetch pixel-wise image features with camera-LiDAR projection matrix given by 3D coordinates. As an example, MVX-Net [27] provides an easy-to-extend framework for cross-modal 3D object detection with joint optimization in 2D and 3D branches. AutoAlign [6] formulates the projection relationship as an attention map and automates the learning of such an alignment through the network. In this work, we explore a faster and more efficient alignment strategy to further boost the performance of point-wise feature aggregation.

3. AutoAlignV2

The aim of AutoAlignV2 is to effectively aggregate image features for further performance enhancement of 3D object detectors. We start with the basic architecture of AutoAlign: the paired images are input into a light-weight backbone, ResNet [31], followed by FPN [18] to get the feature maps. Then, relevant imagery information is aggregated through a learnable alignment map to enrich the 3D representations of non-empty voxels during the voxelization phase. Finally, the enhanced features will be fed into the subsequent 3D detection pipeline to generate the instance predictions.

Such a paradigm could aggregate heterogeneous features in a data-driven way. However, there are two main bottlenecks that still hinder the performance. The first one is inefficient feature aggregation. Although global-wise attention map automates the feature alignment between RGB images and LiDAR points, the computational cost is high: given the voxel number N and the size of image feature $W \times H$, the complexity is $O(NWH)$. Due to the large value of WH , AutoAlign discards other layers except C_5 to reduce the

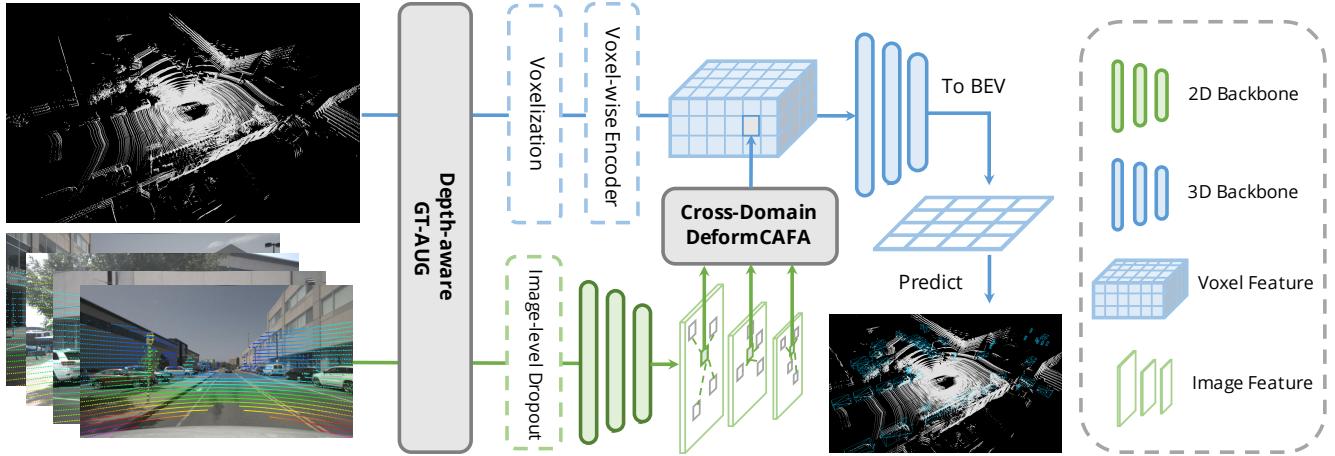


Figure 2. The overall framework of AutoAlignV2. It differs from AutoAlign in three aspects: (i) the proposed Cross-Domain DeformCAFA module enhances the representations with better imagery features and improves the efficiency of the fusion process, (ii) the Depth-Aware GT-AUG algorithm greatly simplifies the synchronization issue among 2D-3D joint augmentations, and (iii) the adoption of image-level dropout training strategy enables our model to infer in a dynamic fusion manner.

cost. The second one is complex data augmentation synchronization between image and points. GT-AUG is an essential step for high-performance 3D object detectors, but how to keep the semantic consistency between the points and the image during training remains a complicated problem.

In this section, we show that the aforementioned challenges can be effectively resolved through the proposed AutoAlignV2, which consists of two parts: Cross-Domain DeformCAFA module and Depth-Aware GT-AUG data augmentation strategy (see Figure 2). We also present a novel image-level dropout training strategy, which enables our model to infer in a more dynamic manner.

3.1. Deformable Feature Aggregation

3.1.1 Revisiting to CAFA

We first revisit the Cross-Attention Feature Alignment module proposed in AutoAlign. Instead of establishing deterministic correspondence with the camera-LiDAR projection matrix, it models the mapping relationship with a learnable alignment map, which enables the network to automate the alignment of non-homogenous features in a dynamic and data-driven manner. Specifically, given the feature map $F = \{f_1, f_2, \dots, f_{hw}\}$ (f_i indicates the image feature of the i^{th} spatial position) and voxel features $P = \{p_1, p_2, \dots, p_J\}$ (p_j indicates each non-empty voxel feature) extracted from raw point clouds, each voxel feature p_j will query the whole image pixels and generate the attention weights based on the dot-product similarity between the voxel feature and the pixel feature. The final output of each voxel feature is the linear combination of values on all the pixel features according to the attention weights. Such a paradigm

enables the model to aggregate semantically relevant spatial pixels to update p_j and demonstrates superior performance compared to bilinear interpolation of features. However, the huge computational cost limits the query candidate to C_5 only, losing the fine-grained information from high-resolution feature maps.

3.1.2 Cross-Domain DeformCAFA

The bottleneck of CAFA is that it takes all the pixels as possible spatial positions. Based on the attributes of 2D images, the most relative information is mainly located at *geometrically-nearby* locations. Therefore, it is unnecessary to consider all the positions but only several key-point regions. Inspired by this, we introduce a novel *Cross-Domain DeformCAFA* operation (see Figure 3), which greatly reduces the sampling candidates and dynamically decides the key-point regions on the image plane for each voxel query feature.

More formally, given the feature map $\mathbf{F} \in \mathbb{R}^{h \times w \times d}$ extracted from the image backbone (e.g., ResNet, CSPNet) and non-empty voxel features $\mathbf{P} \in \mathbb{R}^{N \times c}$, we first compute the reference points $R_i = (r_x^i, r_y^i)$ in the image plane from each voxel feature center $V_i = (v_x^i, v_y^i, v_z^i)$ with the camera projection matrix $T_{cam-lidar}$,

$$R_i = \mathbf{RC} \cdot T_{cam-lidar} \cdot V_i, \quad (1)$$

where \mathbf{RC} is the combination of the rectifying rotation matrix and calibration matrix of the camera. After obtaining the reference point R_i , we adopt bilinear interpolation to get the feature F_i in the image domain. The query feature Q_i is derived as the element-wise product of the image feature F_i and the corresponding voxel feature P_j (to be dis-

cussed later). The final deformable cross-attention feature aggregation is calculated by,

$$\text{DeformCAFA}(Q_i, R_i, \mathbf{F}) = \sum_{m=1}^M \mathbf{W}_m \left[\sum_{k=1}^K A_{mqk}(Q_i) \cdot \mathbf{W}'_m \mathbf{F}(R_i + \Delta R_{mqk}) \right], \quad (2)$$

where \mathbf{W}_m and \mathbf{W}'_m are learnable weights, and A_{mqk} is a MLP to generate attention scores on the aggregated image features. Following the design of self-attention mechanism, we adopt M attention split heads. Here, K is the number of sampling positions ($K^2 \ll HW$, e.g., $K = 4$). With the help of dynamically generated sampling offset ΔR_{mqk} , DeformCAFA is able to conduct cross-domain relational modeling much faster than vanilla operation. The complexity is reduced from $O(NWH)$ to $O(NK^2)$, enabling us to perform multi-layer feature aggregation, *i.e.*, to fully utilize the hierarchical information provided by FPN layers. Another advantage of DeformCAFA is that it explicitly maintains the positional consistency with the camera projection matrix to obtain the reference points. Hence, even without adopting the CFSI module proposed in AutoAlign, our DeformCAFA can yield a semantically and positionally consistent alignment.

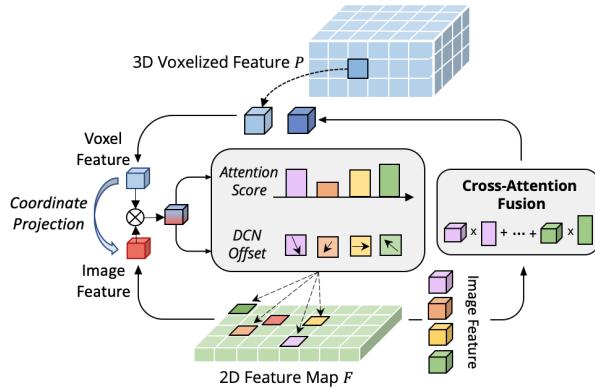


Figure 3. Illustration of the Cross-Domain DeformCAFA module. It first combines coordinate-corresponding voxel and image features to generate cross-domain tokens, which are then used to guide the aggregated positions in 2D feature map through learnable convolutional offset. The final fused feature is obtained by the cross-attention fusion of aggregated image feature and original voxel feature.

3.1.3 Cross-Domain Token Generation

The sparse-style DeformCAFA greatly improves the efficiency compared to vanilla non-local operation. However, when directly applying voxel features as the token to generate attention weights and deformable offsets, the detection



Figure 4. Visualization of the augmented images with the proposed Depth-Aware GT-AUG. The samples are randomly selected from nuScenes dataset.

performance is barely comparable to or even worse than its bilinear-interpolation counterparts. After careful analysis, we find a cross-domain knowledge translation issue in the token generation process. Different from the original deformable operation, which is usually performed under the unimodal setting [3, 43], cross-domain attention requires information from both modalities. However, the voxel features that only consist of spatial representations, can hardly perceive information in the image domain. Therefore, allowing interaction between different modalities is of great importance.

Motivated by [15], we hypothesize that the representation of each object can be explicitly disentangled into two components: the domain-specific and the instance-specific information. The former refers to the data related to the representation itself, including the built-in attributes of domain features, while the latter represents the identity information about the object, regardless of the domain it is encoded in. Concretely, given the corresponding paired image feature F_i and voxel feature P_j , we have,

$$F_i = D_i^{2D} \cdot M_{obj}^i, \quad P_j = D_j^{3D} \cdot M_{obj}^j, \quad (3)$$

where D_i^{2D} and D_j^{3D} are domain-related features in the image and point domains, while M_{obj}^i and M_{obj}^j are the object-specific representations, respectively. Since F_i and P_j are the geometrically-paired features, M_{obj}^i and M_{obj}^j can be close in the instance-specific representation space (*i.e.*, $M_{obj}^i \approx M_{obj}^j \approx M_{obj}^j$). Based on this, we can implicitly interact features of different domain knowledges with,

$$Token = f(F_i \cdot P_j) = f(D_i^{2D} \cdot D_j^{3D} \cdot (M_{obj})^2), \quad (4)$$

where f is one fully connected (FC) layer to aggregate cross-domain information and improve the flexibility of token generation.

3.2. Depth-Aware GT-AUG

Data augmentation is a crucial part of achieving competitive results for most deep learning models. However, in terms of multi-modal 3D object detection, it is hard to keep synchronization between point clouds and images when combining them together in data augmentation, mainly due to object occlusions or changes in the viewpoints. To solve the problem, we design a simple yet effective cross-modal data augmentation named *Depth-Aware GT-AUG*. Different from the methods described in [30, 39], our approach abandons the complex point cloud filtering process or the requirement of delicate mask annotation in the image domain. Instead, inspired by the MixUp proposed in [38], we incorporate the depth information from 3D object annotations to mix up the image regions.

Specifically, given the virtual objects P to paste, we follow the same 3D implementation in GT-AUG [33]. As for the image domain, we first sort them in a far-to-near order. For each to-paste object, we crop the same region from the original image and combine them with a mix-up ratio of α on the target image. The detailed implementation is shown in Algorithm 1.

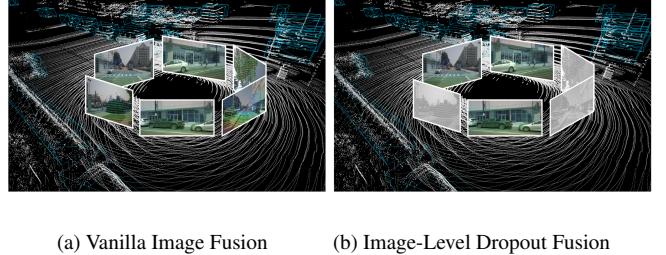
Algorithm 1: Depth-Aware GT-AUG

Input: Object Points Set P^{3D} , Object Image Patches Set P^{2D} , Object Depths Set D , Points P , Image I .

- 1: $\text{ObjectInds} \leftarrow \text{AscendingSort}(D);$
- 2: **for all** i such that $i \in \text{ObjectInds}$ **do**
- 3: // point augmentation
- 4: $P \leftarrow P + P_i^{3D};$
- 5: // image augmentation
- 6: $P_{\text{origin}} = \text{CROP}(I, \text{Coord}(P_i^{2D}));$
- 7: $P_{\text{new}} = \alpha P_{\text{origin}} + (1 - \alpha) P_i^{2D};$
- 8: $I \leftarrow \text{PASTE}(I, P_{\text{new}})$
- 9: **end for**

Output: P, I

Depth-Aware GT-AUG simply follows the augmentation strategy in the 3D domain, but at the same time, keeps the synchronization in the image plane through MixUp-based cut-and-paste. The key intuition is that the MixUp technique does not fully remove the corresponding information after pasting augmented patches on top of the original 2D image. On the contrary, it decays the compactness of such information with respect to the depth to guarantee the existence of the feature from the corresponding points. Concretely, if one object is occluded by other instances n times, the transparency of this object region will be decayed by a factor of $(1 - \alpha)^n$ according to its depth order.



(a) Vanilla Image Fusion (b) Image-Level Dropout Fusion

Figure 5. Visualization of our proposed image-level dropout training strategy compared to the vanilla fusion method. We enable the model to acquire ad-hoc inference by randomly blinding several cameras during training. The images in white-black (b) denote the dropout RGB images where we pad them with zeros for fusion.

3.3. Image-Level Dropout Training Strategy

Actually, image is usually an optional input and may not be supported in all 3D detection systems. Therefore, a more realistic and applicable solution to multi-modal detection should be in a dynamic fusion manner: when images are unavailable, the model detects objects based on raw point clouds; when images are available, the model conducts feature fusion and yields better prediction. To achieve this goal, we propose an image-level dropout training strategy by randomly dropping the aggregated image features at the image level and padding them with zeros during training, as shown in Figure 5. Since the imagery information is intermittently missed, the model should gradually learn to utilize 2D features as one alternative input. Later, we will show that such a strategy not only speeds up the training speed greatly (with fewer images to process per batch) but also improves the final performance.

4. Experiments

4.1. Dataset and Experimental Setup

Dataset. The nuScenes dataset [2] is one of the most popular datasets for 3D object detection, consisting of 700 scenes for training, 150 scenes for validation, and 150 scenes for testing. For each scene, it includes 6 camera images to cover the whole viewpoint. In terms of the overlapping regions between images, we predefine the image fetching priority sequence to avoid the ambiguous problem.

Experimental Setup. We select Object DGCNN [32] and CenterPoint [36] as 3D base detectors for the nuScenes dataset. For the image branch, we adopt a light-weight backbone CSPNet [31], the same one used in YOLOX-Tiny [9], as the feature extractor, followed by PAFFPN [18]. We also pretrain the image branch with 2D detection supervision on nuImages by adding an extra head [9]. The voxel size is set to $(0.1m, 0.1m, 0.1m)$ if not specified. To avoid the redundant computational cost, we adopt dynamic vox-

elization [40] to reduce the number of voxel features. As for the DeformCAFA module, the head number is set to 4 and the deformable point is set to 8. All the feature pyramid layers share the same weight for the feature aggregation operation. All runtimes are measured on a NVIDIA V100 GPU. The whole framework is optimized with hybrid optimizers in an end-to-end manner. The 3D branch is optimized with AdamW and the 2D branch is optimized with SGD. We use MMDetection3D [7] as our codebase and apply the default settings, if not specified.

4.2. Main Results

4.2.1 Results on 3D Object Detectors

We first implement AutoAlignV2 on two representative 3D detector baselines: CenterPoint (anchor/center-based) and Object DGCNN (transformer-based) on nuScenes validation subset. The final performance is reported in Table 1. Our AutoAlignV2 greatly boosts its vanilla 3D baselines by 3.7/4.5 on mAP and 2.4/2.4 on NDS score, respectively. This validates the effectiveness and generalization of the proposed method under different 3D detection frameworks.

Table 1. Comparison of detection results based on Object DGCNN and CenterPoint with and without AutoAlignV2 on nuScenes validation subset.

Method	AutoAlignV2	mAP	NDS
Object DGCNN [32]		60.73	67.14
Object DGCNN [32]	✓	64.42	69.52
CenterPoint [36]		62.56	68.84
CenterPoint [36]	✓	67.05	71.23

4.2.2 Comparison with State-of-the-Arts

In addition to offline results, we also report the detection performance on nuScenes test leaderboard compared to various detection approaches. The results are shown in Table 2. Our final model is based on CenterPoint with a voxel size of $(0.075m, 0.075m, 0.2m)$. It surpasses all the other counterparts including the recently developed MoCa [39] and PointAugmenting [30] by roughly 2.0 mAP, achieving new state-of-the-arts on this competitive benchmark. When observing the results in detail, we can find that the construction vehicle, motorcycle, and bicycle are separately improved by 13.1, 13.4, and 17.4 mAP. Such huge enhancements manifest the superiority of our proposed AutoAlignV2 to deal with hard-to-detect examples.

4.3. Ablation Studies

In this section, we provide extensive ablations to gain a deeper understanding of AutoAlignV2. For efficiency, 1/8

nuScenes training set is used.

4.3.1 Ablation Studies on AutoAlignV2

To understand how each component in AutoAlignV2 facilitates the detection performance, we test each module independently on the baseline detector: CenterPoint and report its performance in Table 3. The overall mAP score starts from 50.3. When we add the Cross-Domain DeformCAFA module together with the image branch, the mAP score is raised by 6.7%. Such a significant improvement validates the correctness of the incorporation of image features and the effectiveness of the proposed deformable feature alignment module. Then, we adopt the image-level dropout strategy to improve the training speed. The performance does not drop and is even slightly improved by another 0.1 mAP. When the depth-aware GT-AUG is added, the accuracy is further promoted by 1.4 mAP. Although the improvement is not remarkable, depth-aware GT-AUG greatly simplifies the synchronization process in the joint image-point augmentation.

4.3.2 Ablation Studies on Cross-Domain DeformCAFA

Comparison with other fusion mechanisms. In this experiment, we keep all settings the same except for the cross-modal feature fusion method for a fair comparison. We consider the following strategies used in PointPainting [27], MoCa [39], AutoAlign [6], and PointAugmenting [30], and compare them with Cross-Domain DeformCAFA in Table 4. It can be found that AutoAlignV2 outperforms all the other fusion mechanisms by a large margin, verifying the effectiveness of our proposed approach. The enhancement mainly stems from two aspects: (i) multi-level features are fully utilized thanks to the optimization of computational complexity and (ii) superiority of relational modeling on cross-domain features across different modalities.

Strategies on token generation. To validate the necessity of the cross-domain token generation, we compare our method with various policies: generated from voxel features only, image features only, and their combinations including concatenation, addition, and multiplication. As given in Table 5, utilizing the voxel features as the query token cannot guarantee satisfying results, since 3D features can hardly perceive information in the interaction between cross-modal features. The result produced by the image features is also limited, possibly due to the lack of information from 3D points. The performance of simply concatenating or adding them together remains poor. We infer the reason that though both features contain the same identity information, it is still hard for the model to figure them out when blending with the domain-specific representation. Finally, we obtain

Table 2. Comparison with previous methods on nuScenes test leaderboard. ‘‘C.V.’’ and ‘‘Ped.’’ are the abbreviations of construction vehicle and pedestrian, respectively. NDS score, mAP, and APs of each category are reported. The single class AP not reported in the paper is marked by ‘‘-’’. The best results are highlighted in bold.

Method	NDS	mAP	Car	Truck	Bus	Trailer	C.V.	Ped.	Motor	Bicycle
3D-CVF [37]	49.8	42.2	79.7	37.9	55.0	36.3	-	71.3	37.2	-
PointPainting [29]	58.1	46.4	77.9	35.8	36.1	37.3	15.8	73.3	41.5	24.1
CVCNet [4]	66.6	58.2	82.6	49.5	59.4	51.1	16.2	83.0	61.8	38.8
AFDetV2 [42]	68.5	62.4	86.3	54.2	62.5	58.9	26.7	85.8	63.8	34.3
MVP [36]	70.5	66.4	86.8	58.5	67.4	57.3	26.1	89.1	70.0	49.3
MoCa [39]	70.9	66.6	86.7	58.6	67.2	60.3	32.6	87.1	67.8	52.0
AutoAlign [6]	70.9	65.8	85.9	55.3	67.7	55.6	29.6	86.4	71.5	51.5
PointAugmenting [30]	71.1	66.8	87.5	57.3	65.2	60.7	28.0	87.9	74.3	50.9
CenterPoint [36]	67.3	60.3	85.2	53.5	63.6	56.0	20.0	84.6	59.5	30.7
AutoAlignV2 (Ours)	72.4	68.4	87.0	59.0	69.3	59.3	33.1	87.6	72.9	52.1

Table 3. Effect of each component in our AutoAlignV2. Results are reported on nuScenes validation set with CenterPoint.

DeformCAFA	Image-level Dropout	Depth-aware GT-AUG	mAP	NDS
✓			50.28	58.71
✓	✓		56.96	62.54
✓	✓	✓	57.03	62.52
✓	✓	✓	58.45	63.16

Table 4. Comparison with different feature fusion strategies adopted in current multi-modal detectors. Methods with * indicate our own implementation.

Fusion Strategy	mAP	NDS
Baseline w/o Img	50.28	58.71
PointPainting* [27]	55.45	61.44
MoCa [39]	55.91	61.54
AutoAlign [6]	56.69	61.93
PointAugmenting* [30]	56.75	62.11
Cross-Domain DeformCAFA	58.45	63.16

the best performance with the multiplication version, which proves the assumption in Section 3.1.3.

4.3.3 Ablation Studies on Depth-Aware GT-AUG

Comparison with other cross-modal GT-AUG. We compare depth-aware GT-AUG together with other cross-modal data augmentation approaches proposed in MoCa [39] and PointAugmenting [30]. As shown in Table 6, the depth-aware GT-AUG slightly surpasses all the other strategies even without point filtering or 2D occlusion checking, which greatly overcomes the difficulty in cross-domain synchronization. Moreover, we can see from Figure 4 that the

Table 5. Ablations on different strategies in query token generation for Cross-Domain DeformCAFA module. ‘‘Operation’’ denotes the interact operation between the points and image features to generate tokens.

Pts Feat	Img Feat	Operation	mAP	NDS
✓		-	57.10	61.77
	✓	-	57.77	62.08
✓	✓	Concat	58.01	62.32
✓	✓	Add	57.94	62.13
✓	✓	Multiply	58.45	63.16

depth-aware GT-AUG is able to produce smoother images for image fusion, which enhances the quality of 2D features during the cross-modal fusion process.

GT-AUG Mix-up Ratio. In Figure 6, we investigate how the mix-up ratio α in the depth-aware GT-AUG affects the model performance. It can be seen that the detection result is not sensitive to the mix-up ratio ranging from 0.5 to 0.8, where the NDS only fluctuates within 0.1%. However, the score drops about 0.7 mAP with $\alpha = 1.0$, where the depth-aware GT-AUG degenerates to the original GT-AUG implementation in MoCa [39]. Since no occlusion checking or point filtering is performed, points may be fused with other imagery information, leading to the ambiguous learn-

Table 6. Comparison with various cross-modal GT-AUG strategies. “Occ Check”: abandoning the instance paste if it has certain overlap with the original instances in the images; “Pts Filter”: filtering the points to guarantee that points of one instance will not aggregate the image features from another occluded one.

Method	Occ Check	Pts Filter	mAP	NDS
w/o Aug			40.12	45.39
MoCa [39]	✓		53.08	56.54
Wang et.al [30]		✓	53.16	56.91
DA-GTAUG			53.48	57.16

ing issue.

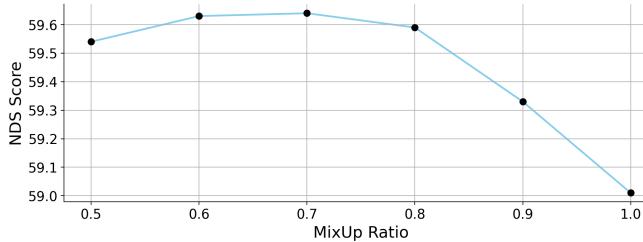


Figure 6. Ablation study on the mix-up ratio α introduced in depth-aware GT-AUG.

4.3.4 Ablation Studies on Image-level Dropout.

Considering that AutoAlignV2 can be dynamically trained with or without images, namely dynamic image fusion, we study such an attribute and how it contributes to the final performance. Concretely, we vary the number of images for training in our image-level dropout strategy and report the detection accuracy as well as the training time in Table 7. From the table, we can find that reducing the number of training images from 6 to 3 has little effect on the performance of the model but greatly reduces the training time by $1.5\times$. However, if continuously reducing this number to 1, the performance incurs an evident decline. We infer the reason that single image training is not enough for fully cross-modal fusion learning. Therefore, we adopt 3 images per scene in our experiments.

Table 7. Ablation studies on the number of images for fusion during the training process with our proposed image-level dropout strategy.

# Images	Training Time	mAP	NDS
0	7.6h	50.28	58.71
1	8.5h	57.93	62.84
3	9.7h	58.45	63.16
6	14.1h	58.51	63.11

4.4. Dynamic Inference and Runtime

Autonomous driving is a direct application of multi-modal 3D object detection. Therefore, the practicality and inclusiveness of the model are also vital. As mentioned in Section 3.1.3, AutoAlignV2 fits to different inference modes, no matter the images are available or not. We carefully measure the inference performance of AutoAlignV2 under different settings and report its runtime per frame in Table 8. Compared with the LiDAR-only detector: CenterPoint, our AutoAlignV2 takes only 123 ms for the extra 2D image branch, thanks to its light-weight backbone: CSPNet. We resize all the images to 640×1280 for efficient fusion. In addition to fully surrounding images for cross-modal fusion, our method is also qualified for the LiDAR-only scenarios without any extra computational cost compared to vanilla CenterPoint, but still maintains the detection accuracy.

Table 8. Inference time of AutoAlignV2 on nuScenes dataset. “# Images” means the number of images to load during inference.

Method	# Images	Inference Time	mAP	NDS
CenterPoint	-	85ms	50.28	58.71
AutoAlignV2	6	208ms	58.45	63.16
AutoAlignV2	3	181ms	54.32	60.84
AutoAlignV2	0	87ms	50.29	58.67

5. Conclusion

In this paper, we develop a dynamic and fast multi-modal 3D object detection framework, AutoAlignV2. It greatly speeds up the fusion process by utilizing multi-layer deformable cross-attention networks to extract and aggregate features from different modalities. We also design the depth-aware GT-AUG strategy to simplify the synchronization between 2D and 3D domains during the multi-modal data augmentation process. Interestingly, our AutoAlignV2 is much more flexible and can infer with and without images in an ad-hoc manner, which is more suitable for the real-world systems. We hope AutoAlignV2 can serve as a simple yet strong paradigm in multi-modal 3D object detection.

Acknowledgments: This work was supported by the USTC-NIO Joint Research Funds KD2111180313. We acknowledge the support of GPU cluster built by MCC Lab of Information Science and Technology Institution, USTC.

References

- [1] Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan 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*, pages 1090–1099, 2022. 1
- [2] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giacarlo Baldan, and Oscar Beijbom. nuScenes: A multi-modal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 1, 5
- [3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 4
- [4] Qi Chen, Lin Sun, Ernest Cheung, and Alan L Yuille. Every view counts: Cross-view consistency in 3D object detection with hybrid-cylindrical-spherical voxelization. *Advances in Neural Information Processing Systems*, pages 21224–21235, 2020. 7
- [5] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3D object detection network for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1907–1915, 2017. 2
- [6] Zehui Chen, Zhenyu Li, Shiquan Zhang, Liangji Fang, Qinghong Jiang, Feng Zhao, Bolei Zhou, and Hang Zhao. AutoAlign: Pixel-instance feature aggregation for multi-modal 3D object detection. *IJCAI*, pages 1–7, 2022. 1, 2, 6, 7
- [7] MMDetection3D Contributors. MMDetection3D: Open-MMLab next-generation platform for general 3D object detection. <https://github.com/open-mmlab/mmdetection3D>, 2020. 6
- [8] Jiajun Deng, Shaoshuai Shi, Peiwei Li, Wengang Zhou, Yanyong Zhang, and Houqiang Li. Voxel R-CNN: Towards high performance voxel-based 3D object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1201–1209, 2021. 2
- [9] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, pages 1–6, 2021. 5
- [10] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3354–3361. IEEE, 2012. 1
- [11] Tengteng Huang, Zhe Liu, Xiwu Chen, and Xiang Bai. EP-Net: Enhancing point features with image semantics for 3D object detection. In *Proceedings of the European Conference on Computer Vision*, pages 35–52. Springer, 2020. 2
- [12] Jason Ku, Melissa Mozifian, Jungwook Lee, Ali Harakeh, and Steven L Waslander. Joint 3D proposal generation and object detection from view aggregation. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1–8. IEEE, 2018. 2
- [13] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. PointPillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12697–12705, 2019. 1
- [14] Yingwei Li, Adams Wei Yu, Tianjian Meng, Ben Caine, Ji-quan Ngiam, Daiyi Peng, Junyang Shen, Yifeng Lu, Denny Zhou, Quoc V Le, et al. DeepFusion: LiDAR-camera deep fusion for multi-modal 3D object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17182–17191, 2022. 1
- [15] Yanwei Li, Hengshuang Zhao, Xiaojuan Qi, Liwei Wang, Zeming Li, Jian Sun, and Jiaya Jia. Fully convolutional networks for panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 214–223, 2021. 4
- [16] Zhenyu Li, Zehui Chen, Ang Li, Liangji Fang, Qinrong Jiang, Xianming Liu, Junjun Jiang, Bolei Zhou, and Hang Zhao. SimIPU: Simple 2d image and 3d point cloud unsupervised pre-training for spatial-aware visual representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1500–1508, 2022. 1
- [17] Ming Liang, Bin Yang, Shenlong Wang, and Raquel Urtasun. Deep continuous fusion for multi-sensor 3D object detection. In *Proceedings of the European Conference on Computer Vision*, pages 641–656, 2018. 2
- [18] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8759–8768, 2018. 2, 5
- [19] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela Rus, and Song Han. BEVFusion: Multi-task multi-sensor fusion with unified bird’s-eye view representation. *arXiv preprint arXiv:2205.13542*, 2022. 1
- [20] Jiageng Mao, Yujing Xue, Minzhe Niu, Haoyue Bai, Jiashi Feng, Xiaodan Liang, Hang Xu, and Chunjing Xu. Voxel transformer for 3D object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3164–3173, 2021. 1
- [21] Su Pang, Daniel Morris, and Hayder Radha. CLOCs: Camera-LiDAR object candidates fusion for 3D object detection. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1–10. IEEE, 2020. 2
- [22] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3D object detection in point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9277–9286, 2019. 2
- [23] Charles R Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J Guibas. Frustum pointnets for 3D object detection from RGB-D data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 918–927, 2018. 2
- [24] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. PV-RCNN: Point-voxel feature set abstraction for 3D object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10529–10538, 2020. 1, 2
- [25] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Point RCNN: 3D object proposal generation and detection from

- point cloud. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–779, 2019. 2
- [26] Shaoshuai Shi, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. From points to parts: 3D object detection from point cloud with part-aware and part-aggregation network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–14, 2020. 2
- [27] Vishwanath A Sindagi, Yin Zhou, and Oncel Tuzel. MVX-Net: Multimodal voxelnet for 3D object detection. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 7276–7282. IEEE, 2019. 2, 6, 7
- [28] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2446–2454, 2020. 1
- [29] Sourabh Vora, Alex H Lang, Bassam Helou, and Oscar Beijbom. PointPainting: Sequential fusion for 3D object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4604–4612, 2020. 2, 7
- [30] Chunwei Wang, Chao Ma, Ming Zhu, and Xiaokang Yang. PointAugmenting: Cross-modal augmentation for 3D object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11794–11803, 2021. 1, 2, 5, 6, 7, 8
- [31] Chien-Yao Wang, Hong-Yuan Mark Liao, Yueh-Hua Wu, Ping-Yang Chen, Jun-Wei Hsieh, and I-Hau Yeh. CSPNet: A new backbone that can enhance learning capability of cnn. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 390–391, 2020. 2, 5
- [32] Yue Wang and Justin M Solomon. Object DGCNN: 3D object detection using dynamic graphs. *Advances in Neural Information Processing Systems*, 2021. 2, 5, 6
- [33] Yan Yan, Yuxing Mao, and Bo Li. SECOND: Sparsely embedded convolutional detection. *Sensors*, pages 3337–2247, 2018. 1, 2, 5
- [34] Zetong Yang, Yanan Sun, Shu Liu, and Jiaya Jia. 3DSSD: Point-based 3D single stage object detector. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11040–11048, 2020. 2
- [35] Zetong Yang, Yanan Sun, Shu Liu, Xiaoyong Shen, and Jiaya Jia. STD: Sparse-to-dense 3D object detector for point cloud. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1951–1960, 2019. 2
- [36] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3D object detection and tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11784–11793, 2021. 2, 5, 6, 7
- [37] Jin Hyeok Yoo, Yecheol Kim, Jisong Kim, and Jun Won Choi. 3D-CVF: Generating joint camera and LiDAR features using cross-view spatial feature fusion for 3D object detection. In *Proceedings of the European Conference on Computer Vision*, pages 720–736. Springer, 2020. 2, 7
- [38] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. MixUp: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, pages 1–12, 2017. 5
- [39] Wenwei Zhang, Zhe Wang, Chen Change Loy, and Wenwei Zhang. Improving data augmentation for multi-modality 3D object detection. *arXiv preprint arXiv:2012.12741*, pages 1–10, 2021. 1, 2, 5, 6, 7, 8
- [40] Yin Zhou, Pei Sun, Yu Zhang, Dragomir Anguelov, Jiyang Gao, Tom Ouyang, James Guo, Jiquan Ngiam, and Vijay Vasudevan. End-to-end multi-view fusion for 3D object detection in LiDAR point clouds. In *Conference on Robot Learning*, pages 923–932. PMLR, 2020. 6
- [41] Yin Zhou and Oncel 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*, pages 4490–4499, 2018. 2
- [42] Benjin Zhu, Zhengkai Jiang, Xiangxin Zhou, Zeming Li, and Gang Yu. Class-balanced grouping and sampling for point cloud 3D object detection. *arXiv preprint arXiv:1908.09492*, pages 1–10, 2019. 7
- [43] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, pages 1–12, 2020. 4