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

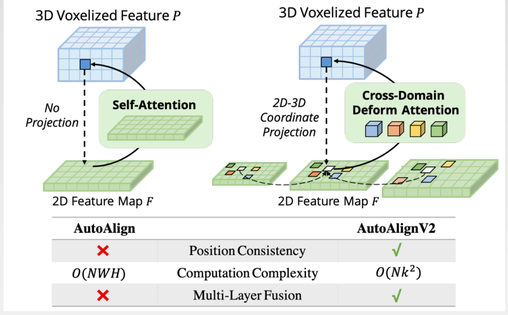
Zehui Chen, Zhenyu Li, Shiquan Zhang, Liangji Fang, Qinhong Jiang, Feng Zhao*

USTC, HIT, SenseTime

ABSTRACT

Point clouds and RGB images are two general perceptional sources in autonomous driving. The former can provide accurate localization of objects, and the latter is denser and richer in semantic information. Recently, AutoAlign 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 crossmodal 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.

A Quick Comparision with AutoAlign & AutoAlignV2



AutoAlign		AutoAlignV2
✗	Position Consistency	√
$O(NWH)$	Computation Complexity	$O(Nk^2)$
✗	Multi-Layer Fusion	√

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.

Revisit CAFA in AutoAlign

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 $F = \{f_1, f_2, \ldots, f_{hw}\}$, f_i indicates the image feature of the i th spatial position) and voxel features $P = \{p_1, p_2, \ldots, p_j\}$ (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.

METHOD

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.

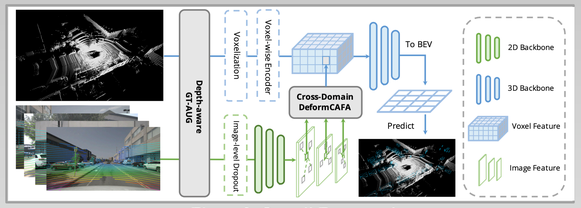


Figure 2. Overall Framework

Deformable Feature Aggregation

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 , 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 $F \in \boldsymbol{R}^{h \times w \times d}$ extracted from the image backbone (e.g., ResNet, CSPNet) and non-empty voxel features $P \in \boldsymbol{R}^{N \times c}$, we first compute the reference points $R_i = \left(r_x^i, r_y^i\right)$ in the image plane from each voxel feature center $V_i = \left(v_x^i, v_y^i, v_z^i\right)$ with the camera projection matrix. 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. The final deformable cross-attention feature aggregation is

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

Cross-Domain Token Generation

Motivated by DeFCN, we hypothesize that the representation of each object can be explicitly disentangled into two components: the domain-specific and the instance-specific information. Given the corresponding paired image feature F_i and voxel feature P_j , we have, $F_i = D_i^{2D} \cdot M_{obj}^i, P_j = D_j^{3D} \cdot M_{obj}^j$, where D_i^{2D} and D_j^{3D} are domain-related features in the image and point domains, while M_{obj} denotes the object-specific representations. Based on this, we can implicitly interact features of different domain knowledges with,

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

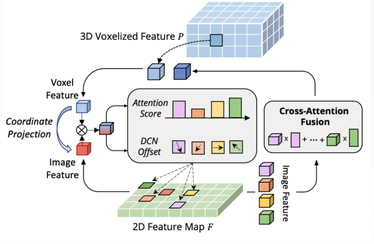


Figure 3. Illustration of the *Cross-Domain DeformCAFA* module. It first combines coordinate-corresponding voxel and image features to generate cross-domain tokens, which are then used to guide the aggregated positions in 2D feature map through learnable convolutional offset.

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 PointAugmenting, 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, we incorporate the depth information from 3D object annotations to mix up the image regions.

Algorithm 1: Depth-Aware GT-AUG

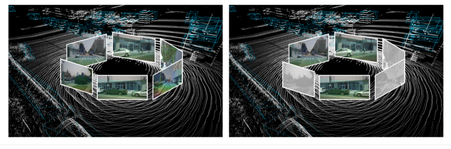
Input: Object Points Set $\mathbf{P^{3D}}$, Object Image Patches Set $\mathbf{P^{2D}}$, Object Depths Set $\mathbf{D}$, Points $\mathbf{P}$, Image $\mathbf{I}$.

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

Output: $\mathbf{P}, \mathbf{I}$

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. 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.



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

Figure 4. 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 while-black (b) denote the dropout RGB images where we pad them with zeros for fusion.

Figure 5. Visualization of the augmented images with the proposed Depth-Aware GT-AUG. The samples are randomly selected from nuScenes dataset.

RESULTS

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**

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

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 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.

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

LINK

WeChat Code Paper

CONCLUSIONS

In this paper, we develop a dynamic and fast multimodal 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.

ACKNOWLEDGEMENT

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.