MonoATT: Online Monocular 3D Object Detection with Adaptive Token Transformer

Yunsong Zhou [1] Hongzi Zhu [1] * Quan Liu [1] Shan Chang [2] Minyi Guo [1]
[1]Shanghai Jiao Tong University [2]Donghua University
{zhouyunsong,hongzi,liuquan2017,guo-my}@sjtu.edu.cn changshan@dhu.edu.cn

Abstract

Mobile monocular 3D object detection (Mono3D) (e.g., on a vehicle, a drone, or a robot) is an important yet challenging task. Existing transformer-based offline Mono3D models adopt grid-based vision tokens, which is suboptimal when using coarse tokens due to the limited available computational power. In this paper, we propose an online Mono3D framework, called MonoATT*, which leverages a novel vision transformer with heterogeneous tokens of varying shapes and sizes to facilitate mobile Mono3D. The core idea of MonoATT is to adaptively assign finer tokens to areas of more significance before utilizing a transformer to enhance Mono3D. To this end, we first use prior knowledge to design a scoring network for selecting the most important areas of the image, and then propose a token clustering and merging network with an attention mechanism to gradually merge tokens around the selected areas in multiple stages. Finally, a pixel-level feature map is reconstructed from heterogeneous tokens before employing a SOTA Mono3D detector as the underlying detection core. Experiment results on the real-world KITTI dataset demonstrate that MonoATT can effectively improve the Mono3D accuracy for both near and far objects and guarantee low latency. MonoATT yields the best performance compared with the state-of-the-art methods by a large margin and is ranked number one on the KITTI 3D benchmark.*

1. Introduction

Three-dimensional (3D) object detection has long been a fundamental problem in both industry and academia and enables various applications, ranging from autonomous vehicles [17] and drones, to robotic manipulation and augmented reality applications. Previous methods have achieved superior performance based on the accurate depth information from multiple sensors, such as LiDAR signal [11,23,35,43,44,69] or stereo matching [9,10,21,34,37,57]. In order to lower the sensor requirements, a much cheaper, more energy-efficient, and easier-to-deploy alternative, *i.e.*, monocular 3D object detection (Mono3D) has been proposed and made impressive progress. A practical online Mono3D detector for autonomous driving should meet the following two requirements: 1) given the constrained computational resource on a mobile platform, the 3D bounding boxes produced by the Mono3D detector should be accurate enough, not only for near objects but also for far ones, to ensure, *e.g.*, high-priority driving safety applications; 2) the response time of the Mono3D detector should be as low as possible to ensure that objects of interest can be instantly detected in mobile settings.

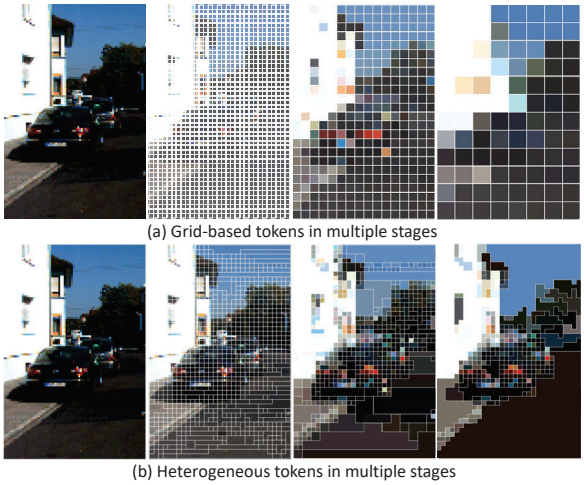

Figure 1. Illustration of (a) grid-based tokens used in traditional vision transformers and (b) heterogeneous tokens used in our adaptive token transformer (ATT). Instead of equally treating all image regions, our ATT distributes dense and fine tokens to meaningful image regions (*i.e.*, distant cars and lane lines) yet coarse tokens to regions with less information such as the background.

Current Mono3D methods, such as depth map based [15, 29, 36], pseudo-LiDAR based [15, 29–31, 36, 54, 57], and image-only based [2, 3, 12, 22, 26, 28, 42, 48, 51, 64–67], mostly follow the pipelines of traditional 2D object detectors [41, 42, 48, 66] to first localize object centers from heatmaps and then aggregate visual features around each object center to predict the object's 3D properties, *e.g.*, location, depth, 3D sizes, and orientation. Although it is con-

*Corresponding authors

ceptually straightforward and has low computational overhead, merely using local features around the predicted object centers is insufficient to understand the scene-level geometric cues for accurately estimating the depth of objects, making existing Mono3D methods far from satisfactory. Recently, inspired by the success of transformers in natural language processing, visual transformers with long-range attention between image patches have recently been developed to solve Mono3D tasks and achieve state-of-the-art (SOTA) performance [19, 64]. As illustrated in Figure 1 (a), most existing vision transformers follow the *grid-based* token generation method, where an input image is divided into a grid of equal image patches, known as tokens. However, using grid-based tokens is sub-optimal for Mono3D applications such as autonomous driving because of the following two reasons: 1) far objects have smaller size and less image information, which makes them hard to detect with coarse grid-based tokens; 2) using fine grid-based tokens is prohibitive due to the limited computational power and the stringent latency requirement.

In this paper, we propose an online Mono3D framework, called *MonoATT*, which leverages a novel vision transformer with *heterogeneous* tokens of varying sizes and shapes to boost mobile Mono3D. We have one key observation that not all image pixels of an object have equivalent significance with respect to Mono3D. For instance, pixels on the outline of a vehicle are more important than those on the body; pixels on far objects are more sensitive than those on near objects. The core idea of MonoATT is to automatically assign fine tokens to pixels of more significance and coarse tokens to pixels of less significance before utilizing a transformer to enhance Mono3D detection. To this end, as illustrated in Figure 1 (b), we apply a *similarity compatibility principle* to dynamically cluster and aggregate image patches with similar features into heterogeneous tokens in multiple stages. In this way, MonoATT neatly distributes computational power among image parts of different importance, satisfying both the high accuracy and low response time requirements posed by mobile Mono3D applications.

There are three main challenges in designing MonoATT. First, it is essential yet non-trivial to determine keypoints on the feature map which can represent the most relevant information for Mono3D detection. Such keypoints also serve as cluster centers to group tokens with similar features. To tackle this challenge, we score image features based on prior knowledge in mobile Mono3D scenarios. Specifically, features of targets (*e.g.*, vehicles, cyclists, and pedestrians) are more important than features of the background. Moreover, more attention is paid to features of distant targets and the outline of targets. Then, a predefined number of keypoints with the highest scores are selected as cluster centers to guide the token clustering in each stage. As a result, an image region with dense keypoints will eventually be assigned with fine tokens while a region with sparse keypoints will be assigned with coarse tokens.

Second, given the established cluster centers in each stage, how to group similar tokens into clusters and effectively aggregate token features within a cluster is non-intuitive. Due to the local correlation of 2D convolution, using naive minimal feature distance for token clustering would make the model insensitive to object outlines. Furthermore, a straightforward feature averaging scheme would be greatly affected by noise introduced by outlier tokens. To deal with these issues, we devise a token clustering and merging network. It groups tokens into clusters, taking both the feature similarity and image distance between tokens into account, so that far tokens with similar features are more likely to be designated into one cluster. Then, it merges all tokens in a cluster into one combined token and aggregates their features with an attention mechanism.

Third, recovering multi-stage vision tokens to a pixel-level feature map is proved to be beneficial for vision transformers [46, 62]. However, how to restore a regular image feature map from heterogeneous tokens of irregular shapes and various sizes is challenging. To transform adaptive tokens of each stage into feature maps, we propose an efficient multi-stage feature reconstruction network. Specifically, the feature reconstruction network starts from the last stage of clustering, gradually upsamples the tokens, and aggregates the token features of the previous stage. The aggregated tokens correspond to the pixels in the feature map one by one and are reshaped into a feature map. As a result, accurate 3D detection results can be obtained via a conventional Mono3D detector using the enhanced feature map.

Experiments on KITTI dataset [17] demonstrate that our method outperforms the SOTA methods by a large margin. Such a framework can be applied to existing Mono3D detectors and is practical for industrial applications. The proposed MonoATT is ranked *number one* on the KITTI 3D benchmark by submission. The whole suite of the code base will be released and the experimental results will be posted to the public leaderboard. We highlight the main contributions made in this paper as follows: 1) a novel online Mono3D framework is introduced, leveraging an adaptive token transformer to improve the detection accuracy and guarantee a low latency; 2) a scoring network is proposed, which integrates prior knowledge to estimate keypoints for progressive adaptive token generation; 3) a feature reconstruction network is designed to reconstruct a detailed image feature map from adaptive tokens efficiently.

2. Related Work

Standard Monocular 3D object detection. The monocular 3D object detection aims to predict 3D bounding boxes from a single given image. Except for methods assisted by additional inputs, such as depth maps [15, 29, 36], CAD models [6, 8, 27, 33, 56], and LiDAR [7, 30, 40, 54],

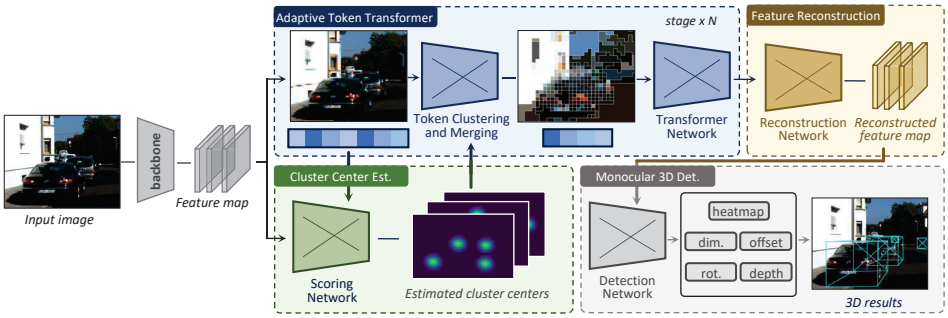

Figure 2. MonoATT consists of four main components, *i.e.*, *cluster center estimation* (CCE), *adaptive token transformer* (ATT), *multi-stage feature reconstruction* (MFR), and *monocular 3D detection*. CCE involves a scoring network to predict the most essential image areas which serve as cluster centers in each heterogeneous token generation stage. Given the initial fine grid-based tokens obtained by slicing the feature map, ATT first generates heterogeneous tokens adaptive to the significance of image areas by grouping and merging tokens in multiple stages; then it leverages the long-range self-attention mechanism provided by a transformer network to associate features on heterogeneous tokens. MFR reconstructs an enhanced pixel-level feature map from all irregular tokens for the ease of Mono3D. Finally, a standard Mono3D detector is employed as the underlying detection core.

standard monocular detectors [18, 20, 24, 45, 68] take as input only a single image and mostly adopt center-guided pipelines following conventional 2D detectors [42, 48, 66]. M3D-RPN [2] designs a depth-aware convolution along with 3D anchors to generate better 3D region proposals. With very few handcrafted modules, SMOKE [26] and FCOS3D [51] propose concise architectures for one-stage monocular detection built on CenterNet [66] and FCOS [48], respectively. Many methods turn to geometric constraints for improving performance. MonoPair [12] considers adjacent object pairs and parses their spatial relations with uncertainty. MonoEF [67] first proposes a novel method to capture the camera pose in order to formulate detectors that are not subject to camera extrinsic perturbations. MonoFlex [65] conducts an uncertainty-guided depth ensemble and categorizes different objects for distinctive processing. GUPNet [28] solves the error amplification problem by geometry-guided depth uncertainty and collocates a hierarchical learning strategy to reduce the training instability. To further strengthen the detection accuracy, recent methods have introduced more effective but complicated vision transformers into the networks. MonoDTR [19] proposes to globally integrate context- and depth-aware features with transformers and inject depth hints into the transformer for better 3D reasoning. MonoDETR [64] adopts a depth-guided feature aggregation scheme via a depth-guided transformer and discards the dependency for center detection. The above geometrically dependent designs largely promote the overall performance of image-only methods, but the underlying problem still exists, namely, the detection accuracy for distant objects is still not satisfactory.

Object detection via the transformer. Transformer [50] is first introduced in sequential modeling in natural language processing tasks, and it has been successfully leveraged in DETR [5] which improves the detection performance in the computer vision field by using the long-range attention mechanism. Several methods [13, 58, 59] have made a demonstration of how to apply a vision transformer to a monocular camera model. Transformerfusion [1] leverages the transformer architecture so that the network learns to focus on the most relevant image frames for each 3D location in the scene, supervised only by the scene reconstruction task. MT-SfMLearner [49] first demonstrates how to adapt vision transformers for self-supervised monocular depth estimation focusing on improving the robustness of natural corruptions. Some recent works, MonoDTR [19] and MonoDETR [64], have tried to apply transformers to monocular 3D detection tasks. However, the token splitting of these models is still based on a grid of regular shapes and sizes. None of these methods consider how to merge unnecessary tokens to reduce the computational complexity of the high-resolution feature maps, which will not be available in a typical Mono3D task in autonomous driving scenarios.

There exist some schemes working on improving the efficiency of transformers. Yu *et al.* [61] propose to reformulate the cross-attention learning as a clustering process. Some approaches [16, 39, 47] study the efficiency of ViTs and propose a dynamic token sparsification framework to prune redundant tokens progressively. Wang *et al.* [55] propose to automatically configure a proper number of tokens for each input image. The methods mentioned above are all variants of grid-based token generation, which modify the resolution, centers of grids, and number specifically. In contrast, the token regions of MonoATT are not restricted by grid structure and are more flexible in three aspects, *i.e.* location, shape, and size.

Our MonoATT inherits DETR's superiority for non-local encoding and long-range attention. Inspired by transformers based on variant-scaled and sparse tokens [39, 55, 63], we use dynamically generated adaptive tokens to obtain high accuracy for both near and far targets with low compu-

tational overhead.

3. Design of MonoATT

The guiding philosophy of MonoATT is to utilize adaptive tokens with irregular shapes and various sizes to enhance the representation of image features for transformer-based Mono3D so that two goals can be achieved: 1) superior image features are obtained from coarse to fine to increase Mono3D accuracy for both near and far objects; 2) irrelevant information (*e.g.*, background) is cut to reduce the number of tokens to improve the timeliness of the vision transformer. Figure 2 depicts the architecture of our framework. Specifically, MonoATT first adopts the DLA-34 [60] as its backbone, which takes a monocular image of size $(W \times H \times 3)$ as input and outputs a feature map of size $(W_s \times H_s \times C)$ after down-sampling with an s-factor. Then, the feature map is fed into four components as follows:

Cluster Center Estimation (CCE). CCE leverages a scoring network to pick out the most crucial coordinate point locations from monocular images that are worthy of being used as cluster centers based on the ranking of scores and quantitative requirements in each stage.

Adaptive Token Transformer (ATT). Starting from the initial fine grid-based tokens obtained by slicing the feature map and the selected cluster centers, ATT groups tokens into clusters and merges all tokens within each cluster into one single token in each stage. After that, a transformer network is utilized to establish a long-range attention relationship between adaptive tokens to enhance image features for Mono3D. The ATT process is composed of N stages.

Multi-stage Feature Reconstruction (MFR). MFR restores and aggregates all N stages of irregularly shaped and differently sized tokens into an enhanced feature map of size $(W_s \times H_s \times C^{'})$.

Monocular 3D Detection. MonoATT employs GUPNet [28], a SOTA monocular 3D object detector as its underlying detection core.

3.1. Cluster Center Estimation

In order to generate adaptive tokens, it is key to be aware of the significance of each image region with respect to the Mono3D task. We have the following two observations:

Observation 1: *As a depth knowledge, distant objects are more difficult to detect and should be paid more attention to.*

Observation 2: *As a semantic knowledge, features of targets (e.g., vehicles, pedestrians, and cyclists) are more valuable than those of backgrounds, and outline features (e.g., lanes, boundaries, corner points) are more crucial than inner features of a target.*

Therefore, we propose to design two scoring functions to measure the depth and semantic information, respectively. For the depth scoring function, it is straightforward to estimate the depth information using a monocular depth estimation network but it would greatly increase the computational overhead and training burden. In addition, pixel-level depth labels are required, which is not allowed in a standard Mono3D task (*e.g.*, pixel-level depth labels are not available in the KITTI 3D detection dataset [17]).

Instead, we take an effective depth estimation scheme based on the camera's pinhole imaging principle and have the following proposition:

Proposition 1: *Given the camera coordinate system* $\mathbf{P}$*, the virtual horizontal plane can be projected on the image plane of the camera according to the ideal pinhole camera model and the depth corresponding to each pixel on the image is determined by the camera intrinsic parameter* $\mathbf{K}$.

Particularly, we envision a virtual scene to quickly estimate the depth of each pixel in the scene, where there is a vast and infinite horizontal plane in the camera coordinate system $\mathbf{P}$. Specifically, for each pixel locating at (u, v) with an assumed depth $\hat{z}$, it can be back-projected to a point $(x_{3d}, y_{3d}, \hat{z})$ in the 3D scene:

$$x_{3d} = \frac{u - c_x}{f_x}\hat{z} \quad y_{3d} = \frac{v - c_y}{f_y}\hat{z}, \tag{1}$$

where f_x and f_y are the focal lengths expressed in pixels along the $x-$ and $y-$ axes of the image plane and c_x and c_y are the possible displacements between the image center and the foot point. These are referred to as the camera intrinsic parameters $\mathbf{K}$.

Assume that the elevation of the camera from the ground, denoted as H, is known (for instance, the mean height of all vehicles in the KITTI dataset, including ego vehicles, is 1.65m [17]), the depth of a point on the depth feature map (u, v) can be calculated as:

$$z = \frac{f_y \cdot H}{v - c_y}. \tag{2}$$

Note that (2) is not continuous when the point is near the vanishing point, *i.e.*, $v = c_y$, and does not physically hold when $v \leq c_y$. To address this issue, we use the reciprocal to score the depth as follows:

$$\mathbf{S}_d = -\text{ReLU}(B\frac{\mathbf{v} - c_y}{f_y \cdot H}), \tag{3}$$

where $\mathbf{v}$ is the vector for $y-$ axis, B is a constant, the ReLU activation is applied to suppress virtual depth values smaller than zero, which is not physically feasible for monocular cameras.

For the semantic scoring function, we introduce the subsequent neural network to detect the possible key points from images. Specifically, in addition to the regular regression tasks in CenterNet [66] based network, we introduce a regression branch for semantic scoring:

$$\mathbf{S}_s = \mathbf{f}(\mathbf{H}), \tag{4}$$

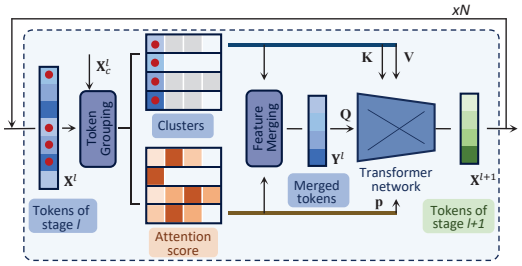

Figure 3. A schematic diagram of the ATT module. The pixels in the feature map are regarded as initial vision tokens. In each stage l, ATT assigns n_{l-1} input tokens to the selected n_l cluster center tokens δ (*i.e.*, denoted as those red dots) and calculates the attention score $\mathbf{p}$ in the respective cluster. Then, tokens in one cluster are merged by feature $\mathbf{x}$ with score $\mathbf{p}$ to obtain a unified token $\mathbf{y}$. Finally, adaptive tokens are associated using a transformer and are used as the input tokens in the next stage $l+1$.

where $\mathbf{H}$ is the input image feature and $\mathbf{f}$ is the CNN architecture. We represent the loss of point detection task as:

$$\mathcal{L}_{\text{CCE}} = \text{FL}(\mathbf{g}^m(\mathbf{u_t}, \mathbf{v_t}), \mathbf{S}), \tag{5}$$

where FL is the Focal Loss used to deal with sample imbalance for key point labels; $(\mathbf{u_t}, \mathbf{v_t})$ is the ground truth key point coordinate; $\mathbf{g}^m$ is the mapping function $\mathbf{g}^m : (\mathbb{R}^m, \mathbb{R}^m) \mapsto \mathbb{R}^{W_s \times H_s}$ which turns m point coordinates into heatmap; $\mathbf{S} = \mathbf{S}_d + \alpha \mathbf{S}_s$ is the score matrix for the image feature map with a size of $(W_s \times H_s)$; α is a hyperparameter. The matrix addition method expands the dimensions and adds content when necessary. The detection network is supervised by $\mathcal{L}_{\text{CCE}}$ and can be trained jointly with other Mono3D branches.

After scoring the whole feature map, CCE calculates the mean value of the pixel scores within each token to assess the importance of that token. We define the *cluster center token* as a token that has the highest average score and serves as the starting center for token clustering. As the number of cluster centers required for different stages is inconsistent, for stage l, we rank and pick out the set of cluster center tokens with number n_l from n_{l-1} original tokens:

$$\mathbf{X}_c^l = \mathbf{g}^r(\mathbf{X}^l, n_l, \mathbf{S}), \tag{6}$$

where $\mathbf{X}_c^l \in \mathbb{R}^{n_l \times C'}$ is the cluster center token features; $\mathbf{g}^r$ is the ranking and picking function selects the highest ranked n_l token features from the input tokens $\mathbf{X}^l \in \mathbb{R}^{n_{l-1} \times C'}$ which is consistent with the output of previous stage $l-1$.

3.2. Adaptive Token Transformer

To enhance the image features for Mono3D, inspired by [63], we leverage an ATT to exploit the long-range self-attention mechanism in an efficient way. As shown in Figure 3, our AAT loops through N stages, where each stage goes through two consecutive processes: *i.e.*, *outline-preferred token grouping*, and *attention-based feature merging*.

3.2.1 Outline-preferred Token Grouping

It is infeasible to cluster tokens based on the straightforward spatial distance of features as it fails to identify outlines of objects due to the local feature correlation brought by 2D convolution [38]. We utilize a variant of the nearest-neighbor clustering algorithm which considers both the feature similarity and image distance between tokens [63].

Specifically, given a set of tokens $\mathbf{X}$ and cluster center tokens $\mathbf{X}_c$, for each token, we compute the indicator δ_i as the minimal feature distance minus average pixel distance between it and any other cluster center token:

$$\delta_i = \min_{j:\mathbf{x}_j \in \mathbf{X}_c} (||\mathbf{x}_i - \mathbf{x}_j||_2^2 - \beta\, ||\mathbf{g}^l(\mathbf{x}_i) - \mathbf{g}^l(\mathbf{x}_j)||_2^2), \tag{7}$$

where δ_i is the indicator that represents which cluster token i should be subordinated to, $\mathbf{x}_i$ and $\mathbf{x}_j$ are feature vectors of token i and j. $\mathbf{g}^l$ is the look-up function that can find the mean position on the feature map corresponding to each token. β is a hyperparameter. The distance constraint requires that two close tokens in the image space have to have extremely similar features in order to be clustered into the same cluster. In this way, we assign all tokens to their corresponding clusters.

3.2.2 Attention-based Feature Merging

To merge token features, an intuitive scheme is to directly average the token features in each cluster. However, such a scheme would be greatly affected by outlier tokens. Inspired by the attention mechanism [39], we attach each token with the attention score $\mathbf{p}$ to explicitly represent the importance, which is estimated from the token features. The token features are averaged with the guidance of attention scores as

$$\mathbf{y}_i = \frac{\sum_{j \in C_i} e^{p_j} \mathbf{x}_j}{\sum_{j \in C_i} e^{p_j}}, \tag{8}$$

where $\mathbf{y_i}$ is the merged token feature; C_i indicates the set of i-th cluster; $\mathbf{x_j}$ and p_j are the original token features and the corresponding attention score. The region of the merged token is the union of the original cluster.

For associating adaptive tokens via the long-range self-attention mechanism, as shown in Figure 3, merged tokens are fed into a transformer network as queries $\mathbf{Q}$, and the original tokens are used as keys $\mathbf{K}$ and values $\mathbf{V}$. In order to differentially allow more important tokens to contribute more to the output and Reduce the impact of outliers, the attention score $\mathbf{p}$ is involved in the calculation of the attention

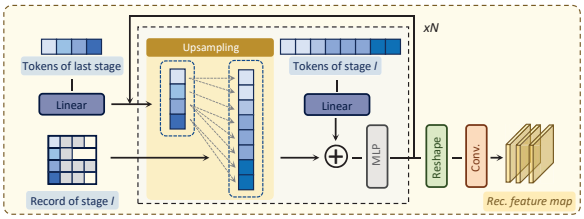

Figure 4. A illustration diagram of the MFR module. MFR starts from the last stage N and progressively aggregates features by stacked upsampling processes and MLP blocks. In the token upsampling process, we use the recorded token relationship to copy the merged token features to the corresponding upsampled tokens. The final tokens are in one-to-one correspondence with the pixels in feature maps and reshaped to the feature maps for Mono3D.

matrix of the transformer:

$$\text{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}} + \mathbf{p})\mathbf{V}, \quad (9)$$

where d_k is the channel number of the queries. When the dimensions of the matrix addition are inconsistent, the matrix performs expansion of the data to the appropriate dimension. Introducing the token attention score $\mathbf{p}$ equips our ATT with the capability to focus on the critical image features when merging vision tokens.

3.3. Multi-stage Feature Reconstruction

Prior work [46,62] has proved the benefits of multi-stage stacking and aggregation of feature maps of different scales for detection tasks. In order to reconstruct the feature map from irregular tokens for feature enhancement, we propose the Multi-stage Feature Reconstruction (MFR), which is able to upsample the tokens by history record and restore the feature maps.

Figure 4 shows the proposed token upsampling process. During the token clustering and feature merging process in Section 3.2, each token is assigned to a cluster and then each cluster is represented by a single merged token. We record the positional correspondence between the original tokens and the merged tokens. In the upsampling process, we use the record to copy the merged token features to the corresponding upsampled tokens. To aggregate detailed features in multiple stages, MFR adds the token features in the previous stage to the upsampled vision tokens. The tokens are then processed by a multi-layer processing (MLP) block. Such processing is executed progressively in N stages until all tokens are aggregated. The lowest level of tokens can be reshaped to feature maps and are processed by 2D convolution for further Mono3D detection.

Some DETR-based Mono3D detectors, such as MonoDETR [64] and MonoDTR [19], which use the Hungarian algorithm to detect the 3D properties of the objects directly from the tokens.

4. Performance Evaluation

We conduct experiments on the widely-adopted KITTI 3D dataset [17]. We report the detection results with three-level difficulties, *i.e.* easy, moderate, and hard, in which the moderate scores are normally for ranking and the hard category is generally distant objects that are difficult to distinguish.

4.1. Quantitative and Qualitative Results

We first show the performance of our proposed MonoATT on KITTI 3D object detection benchmark [1] for the car category. Comparison results with other SOTA Mono3D detectors are shown in Table 1. For the official *test* set, it achieves the highest score for all kinds of samples and is ranked No.1 with no additional data inputs on all metrics. Compared to the second-best models, MonoATT surpasses them under easy, moderate, and hard levels respectively by +1.07, +1.45, and +2.01 in AP_{3D}, especially achieving a significant increase (15%) in the hard level. The comparison fully proves the effectiveness of the proposed adaptive tokens for letting the model spend more effort on the more crucial parts of the image. The first two columns of Figure 5 show the qualitative results on the KITTI dataset. Compared with the baseline model without the aid of adaptive tokens, the predictions from MonoATT are much closer to the ground truth, especially for distinct objects. It shows that using image patches with irregular shapes and various sizes indeed helps locate the object precisely.

4.2. Ablation Study

Effectiveness of each proposed component. In Table 2, we conduct an ablation study to analyze the effectiveness of the proposed components: (a) the baseline which only uses image features for Mono3D based on GUPNet [28]; (b) an improved version of (a) which uses a 3-stage DETR for enhancing image features with regular tokens; (c) grouping tokens based on minimal feature distance and the token features within one cluster are averaged. (d) the proposed *outline-preferred token grouping* is exploited and token features are averaged; (e) *attention-based feature merging* is also used for token aggregation within the cluster. All of (c), (d), and (e) do not consider the issue of how to select cluster centers, they determine them using random sampling in each stage. Based on (e), (f) and (g) consider cluster center selection based on scores. The difference is that (f) only uses the depth knowledge while (g) considers both the depth knowledge and the semantic knowledge. (h) is the final version of our proposed MonoATT which additively takes into account the reconstruction of adaptive tokens into a feature map.

From a $\rightarrow$ b, it can be seen that the effectiveness of the transformer on overall performance, which helps the model to understand the long-range attention relationship between

[1] https://www.cvlibs.net/datasets/kitti/eval_object.php?obj_benchmark=3d

Method	Extra data	Time (ms)	Test, AP_{3D}			Test, AP_{BEV}		
			Easy	Mod.	Hard	Easy	Mod.	Hard
PatchNet [29]	Depth	400	15.68	11.12	10.17	22.97	16.86	14.97
D4LCN [15]		200	16.65	11.72	9.51	22.51	16.02	12.55
Kinematic3D [3]	Multi-frames	120	19.07	12.72	9.17	26.69	17.52	13.10
MonoRUn [7]	Lidar	70	19.65	12.30	10.58	27.94	17.34	15.24
CaDDN [40]		630	19.17	13.41	11.46	27.94	18.91	17.19
AutoShape [27]	CAD	-	22.47	14.17	11.36	30.66	20.08	15.59
SMOKE [26]	None	30	14.03	9.76	7.84	20.83	14.49	12.75
MonoFlex [65]		30	19.94	13.89	12.07	28.23	19.75	16.89
GUPNet [28]		40	20.11	14.20	11.77	-	-	-
MonoDTR [19]		37	21.99	15.39	12.73	28.59	20.38	17.14
MonoDETR [64]		43	23.65	15.92	12.99	32.08	21.44	17.85
MonoATT (Ours)	None	56	**24.72**	**17.37**	**15.00**	**36.87**	**24.42**	**21.88**
Improvement	*v.s. second-best*	-	+1.07	+1.45	+2.01	+4.79	+2.98	+4.03

Table 1. AP_{40} scores(%) of the car category on KITTI *test* set at 0.7 IoU threshold referred from the KITTI benchmark website. We utilize bold to highlight the best results, and color the second-best ones and our performance gain over them in blue. Our model is ranked NO. 1 on the benchmark.

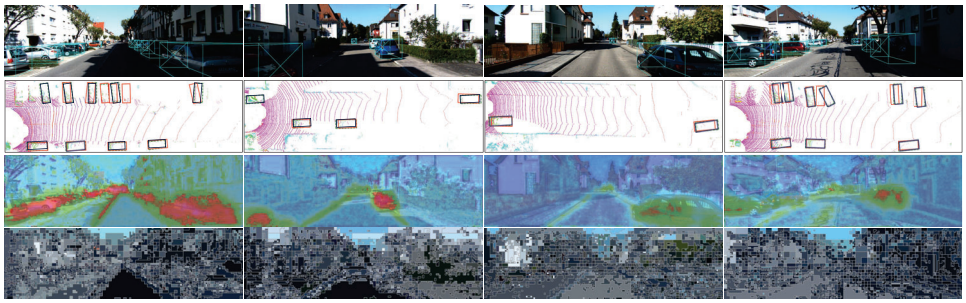

Figure 5. Qualitative results on KITTI dataset. The predicted 3D bounding boxes of our proposed MonoATT are shown in the first row. The second row shows the detection results in the bird's eye view (z-direction from right to left). The green dashed boxes are the ground truth, and the blue and red solid boxes are the prediction results of our MonoATT and the comparison baseline (GUPNet [28]), respectively. The third row and fourth rows visualize the heatmap for estimating cluster centers and adaptive tokens in the last stage.

pixels. However, the grid-based token makes the model unable to cope with small-sized targets and shows no significant change in accuracy at a distance (*i.e.*, usually the hard case). From b $\rightarrow$ c, we can observe that the use of adaptive tokens has a certain enhancement effect compared to the grid-based tokens, especially in the hard case. Because of averaging merged token features and random sampling to select cluster centers, the performance improvement of (c) is not obvious. From c $\rightarrow$ d, it shows that the introduced distance constraint plays a role in improving the detection accuracy. From d $\rightarrow$ e, it can be seen that the performance gain of treating tokens in a cluster differently and using the attention mechanism to achieve feature merging. Both (f) and (g) demonstrate that appropriate cluster centers are crucial for the generation of adaptive tokens. In addition, both the depth knowledge and the semantic knowledge are indispensable in determining cluster centers. Group (h) indicates that for the CenterNet-based Mono3D model, transforming the token into a feature map by MFR for subsequent processing seems to be a more efficient way compared to the Hungarian algorithm in [64].

Response time analysis. Other Mono3D models may require some additional operations to assist the prediction during inference. Compared to these methods, MonoATT is based on CenterNet [66] and we adopt an efficient GPU implementation of ATT module, which only costs 9.4% of the forward time. We can see from Table 1 that our method also has a great advantage in terms of response time.

Visualization of the adaptive tokens. To facilitate the understanding of our ATT, we visualize the cluster center scoring heatmap and the corresponding adaptive tokens in Figure 5. In the heatmap, a warmer color means a higher score and a higher likelihood of becoming a cluster center. As shown in the third row of the figure, the heat is concentrated on the outlines of vehicles, lane lines, and distant targets in the image. It is worth mentioning that the outlines in the scene are of great help for depth discrimination, so even though there are no associated labels, the model can still learn that outlines are crucial for the Mono3D task based on semantics alone. From the visualization, we can see that the model can adaptively divide the images into tokens of different sizes and shapes. This enables the model

	Abla.	ATT	CCE	MFR	Easy	Mod.	Hard
(a)	base.	-	-	-	22.76	16.46	13.72
(b)	+T.				23.18	17.68	13.95
(c)	$+\delta_i$	✓	-	-	24.45	19.32	16.23
(d)	$+\mathbf{g}^l$				24.81	19.68	16.57
(e)	$+\mathbf{p}$				25.62	20.67	17.76
(f)	$+\mathbf{S}_d$	✓	✓	-	27.72	21.19	18.15
(g)	$+\mathbf{S}_s$				28.36	21.78	18.87
(h)	-	✓	✓	✓	**29.01**	**23.49**	**19.60**

Table 2. Effectiveness of different components of our approach on the KITTI *val* set for car category. The Ablation column indicates which new variables and modules we have added to the previous experimental group compared to the previous one. **base.** is the GUPNet [28] baseline. **+T.** stands for the addition of a DETR-based 3-stage transformer.

Method	Val, AP_{3D}			Val, AP_{BEV}		
	Easy	Mod.	Hard	Easy	Mod.	Hard
MonoDTR [19]	24.52	18.57	15.51	33.33	25.35	21.68
+ **Ours**	**26.98**	**21.46**	**18.41**	**35.49**	**27.76**	**24.34**
Imp.	+2.46	+2.89	+2.90	+2.16	+2.41	+2.66
MonoDETR [64]	28.84	20.61	16.38	37.86	26.95	22.80
+ **Ours**	**29.56**	**22.47**	**18.65**	**38.93**	**29.76**	**25.73**
Imp.	+0.72	+2.13	+2.27	+1.07	+2.81	+2.93

Table 3. Extension of MonoATT to existing transformer-based monocular 3D object detectors. We show the AP_{40} scores(%) evaluated on KITTI3D *val* set. **+Ours** indicates that we apply the ATT and CCE modules to the original methods. All models benefit from the MonoATT design.

to use fine tokens for image parts that contain more details (*e.g.*, small targets and boundary lines) and coarse tokens for image parts that do not matter (*e.g.*, the sky and the ground). This implies that ATT is able to put more experience on more important tokens, leading to more accurate predictions.

Plugging into existing transformer-based methods. Our proposed approach is flexible to extend to existing transformer-based Mono3D detectors. We respectively plug the CCE and the ATT components into MonoDTR and MonoDETR, the results of which are shown in Table 3. It can be seen that, with the aid of our CCE and ATT, these detectors can achieve further improvements on KITTI 3D *val* set, demonstrating the effectiveness and flexibility of our proposed adaptive token generation. Particularly, MonoATT enables models to achieve more performance gains in the hard category. For example, for MonoDETR, the AP_{3D}/AP_{BEV} gain is +0.72/+1.07 in the easy category and +2.27/+2.93 in the hard category.

Efficacy for detecting objects at different distances. In Table 4, we compare the accuracy gain of our model for detecting objects at different distances. We present the accuracy (%) of ours with Kinemantic3D, MonoDTR, and MonoDETR as the baselines in the KITTI *val* set for the following three distance ranges: near (5m-10m), middle (20-25m), and far (40m-45m). It can be seen that MonoDTR and MonoDETR using a transformer with grid-based tokens do not perform well in the far case, although they outperform Kinemantic3D in terms of overall accuracy. For MonoDTR, the AP_{3D}/AP_{BEV} gain is +8.79/+8.36 on the far case. From this, we can see that our method has a significant improvement in detecting distant objects.

Method	Val, AP_{3D}			Val, AP_{BEV}		
	Near	Mid.	Far	Near	Mid.	Far
Kinematic3D [3]	34.52	16.50	5.82	246.86	22.81	7.35
+ **Ours**	**35.70**	**21.86**	**10.48**	**48.23**	**27.84**	**12.34**
Imp.	+1.18	+5.36	+4.66	+1.37	+5.03	+4.99
MonoDTR [19]	48.51	17.87	2.16	61.25	25.03	3.29
+ **Ours**	**49.69**	**22.49**	**10.95**	**63.24**	**29.81**	**11.65**
Imp.	+1.18	+4.62	+8.79	+1.99	+4.78	+8.36
MonoDETR [64]	48.66	17.91	2.35	61.21	25.25	3.38
+ **Ours**	**49.92**	**22.48**	**11.07**	**63.29**	**30.00**	**11.91**
Imp.	+1.26	+4.57	+8.72	+2.08	+4.75	+8.53

Table 4. The comparison of the performance gain of our MonoATT over existing models at different distances. We show the AP_{40} scores(%) evaluated on KITTI3D *val* set. **+Ours** indicates that we apply our modules to the original methods. Near (5m-10m), middle (20m-25m), and far (40m-45m) are three different distance intervals. All models benefit from the MonoATT design, especially for far objects.

5. Conclusion

In this paper, we have proposed a Mono3D framework, called *MonoATT*, which can effectively utilize the generated adaptive vision tokens to improve online Mono3D. The advantages of MonoATT are two-fold: 1) it can greatly improve the Mono3D accuracy, especially for far objects, which is an open issue for Mono3D; 2) it can guarantee low latency of Mono3D detectors by omitting backgrounds suitable for appealing mobile applications. Nevertheless, MonoATT still has two main limitations as follows: 1) the computational complexity of the nearest neighbor algorithm in stage 1 is still linear with respect to the token number, which limits the speed of MonoATT for a large initial token input; 2) it heavily counts on the scoring network in CCE, which may sometimes be disturbed by semantic noise, such as complex vegetation areas. These limitations also direct our future work. We have implemented Mono3D and conducted extensive experiments on the real-world KITTI dataset. MonoATT yields the best performance compared with the SOTA methods by a large margin and is ranked number one on the KITTI 3D benchmark.

Acknowledgement

This research was supported in part by National Natural Science Foundation of China (Grant No. 61972081) and the Natural Science Foundation of Shanghai (Grant No. 22ZR1400200).

References

[1] Aljaz Bozic, Pablo Palafox, Justus Thies, Angela Dai, and Matthias Nießner. TransformerFusion: Monocular RGB Scene Reconstruction using Transformers. *Advances in Neural Information Processing Systems*, 34, 2021. 3

[2] Garrick Brazil and Xiaoming Liu. M3D-RPN: Monocular 3D Region Proposal Network for Object Detection. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 1, 3, 13

[3] Garrick Brazil, Gerard Pons-Moll, Xiaoming Liu, and Bernt Schiele. Kinematic 3D Object Detection in Monocular Video. *arXiv preprint arXiv:2007.09548*, 2020. 1, 7, 8, 13

[4] 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 Multimodal Dataset for Autonomous Driving. *arXiv preprint arXiv:1903.11027*, 2019. 15

[5] 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. 3

[6] Florian Chabot, Mohamed Chaouch, Jaonary Rabarisoa, Céline Teulière, and Thierry Chateau. Deep MANTA: A Coarse-to-Fine Many-Task Network for Joint 2D and 3D Vehicle Analysis from Monocular Image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2040–2049, 2017. 2

[7] Hansheng Chen, Yuyao Huang, Wei Tian, Zhong Gao, and Lu Xiong. MonoRUn: Monocular 3D Object Detection by Reconstruction and Uncertainty Propagation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10379–10388, 2021. 2, 7, 13

[8] Xiaozhi Chen, Kaustav Kundu, Ziyu Zhang, Huimin Ma, Sanja Fidler, and Raquel Urtasun. Monocular 3D Object Detection for Autonomous Driving. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2147–2156, Las Vegas, NV, USA, June 2016. IEEE. 2

[9] Xiaozhi Chen, Kaustav Kundu, Yukun Zhu, Andrew G Berneshawi, Huimin Ma, Sanja Fidler, and Raquel Urtasun. 3D Object Proposals for Accurate Object Class Detection. In *Advances in Neural Information Processing Systems*, pages 424–432, 2015. 1

[10] Xiaozhi Chen, Kaustav Kundu, Yukun Zhu, Huimin Ma, Sanja Fidler, and Raquel Urtasun. 3D Object Proposals Using Stereo Imagery for Accurate Object Class Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(5):1259–1272, May 2018. 1

[11] 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. 1

[12] Yongjian Chen, Lei Tai, Kai Sun, and Mingyang Li. MonoPair: Monocular 3D Object Detection Using Pairwise Spatial Relationships. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12093–12102, 2020. 1, 3, 12, 13

[13] Zeyu Cheng, Yi Zhang, and Chengkai Tang. Swin-Depth: Using Transformers and Multi-Scale Fusion for Monocular-Based Depth Estimation. *IEEE Sensors Journal*, 21(23):26912–26920, 2021. 3

[14] MMDetection3D Contributors. MMDetection3D: OpenMMLab next-generation platform for general 3D object detection. https://github.com/open-mmlab/mmdetection3d, 2020. 13

[15] Mingyu Ding, Yuqi Huo, Hongwei Yi, Zhe Wang, Jianping Shi, Zhiwu Lu, and Ping Luo. Learning Depth-Guided Convolutions for Monocular 3D Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1000–1001, 2020. 1, 2, 7

[16] Mohsen Fayyaz, Soroush Abbasi Koohpayegani, Farnoush Rezaei Jafari, Sunando Sengupta, Hamid Reza Vaezi Joze, Eric Sommerlade, Hamed Pirsiavash, and Jürgen Gall. Adaptive token sampling for efficient vision transformers. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XI*, pages 396–414. Springer, 2022. 3

[17] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 1, 2, 4, 6

[18] Tong He and Stefano Soatto. Mono3D++: Monocular 3D Vehicle Detection with Two-Scale 3D Hypotheses and Task Priors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8409–8416, 2019. 3

[19] Kuan-Chih Huang, Tsung-Han Wu, Hung-Ting Su, and Winston H. Hsu. Monodtr: Monocular 3d object detection with depth-aware transformer. In *CVPR*, 2022. 2, 3, 6, 7, 8, 12

[20] Abhijit Kundu, Yin Li, and James M Rehg. 3D-RCNN: Instance-Level 3D Object Reconstruction via Render-and-Compare. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3559–3568, 2018. 3

[21] Peiliang Li, Xiaozhi Chen, and Shaojie Shen. Stereo R-CNN Based 3D Object Detection for Autonomous Driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7644–7652, 2019. 1

[22] Peixuan Li, Huaici Zhao, Pengfei Liu, and Feidao Cao. RTM3D: Real-time Monocular 3D Detection from Object Keypoints for Autonomous Driving. In *European Conference on Computer Vision*, pages 644–660. Springer, 2020. 1

[23] Ming Liang, Bin Yang, Shenlong Wang, and Raquel Urtasun. Deep Continuous Fusion for Multi-sensor 3D Object Detection. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, Lecture Notes in Computer Science, pages 663–678. Springer International Publishing, 2018. 1

[24] Lijie Liu, Jiwen Lu, Chunjing Xu, Qi Tian, and Jie Zhou. Deep Fitting Degree Scoring Network for Monocular 3D Object Detection. In *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition, pages 1057–1066, 2019. 3

[25] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 12

[26] Zechen Liu, Zizhang Wu, and Roland Tóth. SMOKE: Single-Stage Monocular 3D Object Detection via Keypoint Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 996–997, 2020. 1, 3, 7

[27] Zongdai Liu, Dingfu Zhou, Feixiang Lu, Jin Fang, and Liangjun Zhang. AutoShape: Real-Time Shape-Aware Monocular 3D Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15641–15650, 2021. 2, 7

[28] Yan Lu, Xinzhu Ma, Lei Yang, Tianzhu Zhang, Yating Liu, Qi Chu, Junjie Yan, and Wanli Ouyang. Geometry Uncertainty Projection Network for Monocular 3D Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3111–3121, 2021. 1, 3, 4, 6, 7, 8, 12, 13

[29] Xinzhu Ma, Shinan Liu, Zhiyi Xia, Hongwen Zhang, Xingyu Zeng, and Wanli Ouyang. Rethinking Pseudo-Lidar Representation. *arXiv preprint arXiv:2008.04582*, 2020. 1, 2, 7

[30] Xinzhu Ma, Zhihui Wang, Haojie Li, Pengbo Zhang, Wanli Ouyang, and Xin Fan. Accurate Monocular 3D Object Detection via Color-Embedded 3D Reconstruction for Autonomous Driving. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 1, 2

[31] Fabian Manhardt, Wadim Kehl, and Adrien Gaidon. ROI-10D: Monocular Lifting of 2D Detection to 6D Pose and Metric Shape. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1

[32] Arsalan Mousavian, Dragomir Anguelov, John Flynn, and Jana Kosecka. 3D Bounding Box Estimation Using Deep Learning and Geometry. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7074–7082, 2017. 12

[33] J Krishna Murthy, GV Sai Krishna, Falak Chhaya, and K Madhava Krishna. Reconstructing Vehicles from a Single Image: Shape Priors for Road Scene Understanding. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 724–731. IEEE, 2017. 2

[34] Cuong Cao Pham and Jae Wook Jeon. Robust Object Proposals Re-ranking for Object Detection in Autonomous Driving using Convolutional Neural Networks. *Signal Processing: Image Communication*, 53:110–122, Apr. 2017. 1

[35] 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. 1

[36] Zengyi Qin, Jinglu Wang, and Yan Lu. MonoGRNet: A Geometric Reasoning Network for Monocular 3D Object Localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8851–8858, 2019. 1, 2

[37] Zengyi Qin, Jinglu Wang, and Yan Lu. Triangulation Learning Network: From Monocular to Stereo 3D Object Detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1

[38] Michael Ramamonjisoa, Yuming Du, and Vincent Lepetit. Predicting sharp and accurate occlusion boundaries in monocular depth estimation using displacement fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14648–14657, 2020. 5

[39] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances in neural information processing systems*, 34:13937–13949, 2021. 3, 5

[40] Cody Reading, Ali Harakeh, Julia Chae, and Steven L Waslander. Categorical Depth Distribution Network for Monocular 3D Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8555–8564, 2021. 2, 7

[41] Joseph Redmon and Ali Farhadi. Yolov3: An Incremental Improvement. *arXiv preprint arXiv:1804.02767*, 2018. 1

[42] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *Advances in neural information processing systems*, 28, 2015. 1, 3

[43] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. PointRCNN: 3D Object Proposal Generation and Detection from Point Cloud. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–779, 2019. 1

[44] Kiwoo Shin, Youngwook Paul Kwon, and Masayoshi Tomizuka. RoarNet: A Robust 3D Object Detection Based on Region Approximation Refinement. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, pages 2510–2515. IEEE, 2019. 1

[45] Andrea Simonelli, Samuel Rota Bulo, Lorenzo Porzi, Manuel López-Antequera, and Peter Kontschieder. Disentangling Monocular 3D Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1991–1999, 2019. 3

[46] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2, 6, 12

[47] Yehui Tang, Kai Han, Yunhe Wang, Chang Xu, Jianyuan Guo, Chao Xu, and Dacheng Tao. Patch slimming for efficient vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12165–12174, 2022. 3

[48] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: Fully Convolutional One-Stage Object Detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9627–9636, 2019. 1, 3

[49] Arnav Varma, Hemang Chawla, Bahram Zonooz, and Elahe Arani. Transformers in Self-Supervised Monocular Depth Estimation with Unknown Camera Intrinsics. *arXiv preprint arXiv:2202.03131*, 2022. 3

[50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. *Advances in neural information processing systems*, 30, 2017. 3

[51] Tai Wang, Xinge Zhu, Jiangmiao Pang, and Dahua Lin. FCOS3D: Fully Convolutional One-Stage Monocular 3D Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 913–922, 2021. 1, 3, 15

[52] Tai Wang, Xinge Zhu, Jiangmiao Pang, and Dahua Lin. Probabilistic and Geometric Depth: Detecting objects in perspective. In *Conference on Robot Learning (CoRL) 2021*, 2021. 15

[53] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 568–578, October 2021. 12

[54] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-LiDAR From Visual Depth Estimation: Bridging the Gap in 3D Object Detection for Autonomous Driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8445–8453, 2019. 1, 2

[55] Yulin Wang, Rui Huang, Shiji Song, Zeyi Huang, and Gao Huang. Not all images are worth 16x16 words: Dynamic transformers for efficient image recognition. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 3

[56] Yu Xiang, Wongun Choi, Yuanqing Lin, and Silvio Savarese. Subcategory-Aware Convolutional Neural Networks for Object Proposals and Detection. *arXiv:1604.04693 [cs]*, Mar. 2017. arXiv: 1604.04693. 2

[57] Bin Xu and Zhenzhong Chen. Multi-level Fusion Based 3D Object Detection from Monocular Images. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2345–2353, Salt Lake City, UT, USA, June 2018. IEEE. 1

[58] Guanglei Yang, Hao Tang, Mingli Ding, Nicu Sebe, and Elisa Ricci. Transformer-Based Attention Networks for Continuous Pixel-Wise Prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16269–16279, 2021. 3

[59] Guanglei Yang, Hao Tang, Mingli Ding, Nicu Sebe, and Elisa Ricci. Transformers Solve Limited Receptive Field for Monocular Depth Prediction. *arXiv e-prints*, pages arXiv–2103, 2021. 3

[60] Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. Deep Layer Aggregation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2403–2412, 2018. 4, 12

[61] Qihang Yu, Huiyu Wang, Siyuan Qiao, Maxwell Collins, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. k-means mask transformer. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIX*, pages 288–307. Springer, 2022. 3

[62] Yuhui Yuan, Rao Fu, Lang Huang, Weihong Lin, Chao Zhang, Xilin Chen, and Jingdong Wang. Hrformer: High-resolution transformer for dense prediction. 2021. 2, 6, 12

[63] Wang Zeng, Sheng Jin, Wentao Liu, Chen Qian, Ping Luo, Wanli Ouyang, and Xiaogang Wang. Not all tokens are equal: Human-centric visual analysis via token clustering transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11101–11111, 2022. 3, 5

[64] Renrui Zhang, Han Qiu, Tai Wang, Xuanzhuo Xu, Ziyu Guo, Yu Qiao, Peng Gao, and Hongsheng Li. MonoDETR: Depth-Aware Transformer for Monocular 3D Object Detection. *arXiv preprint arXiv:2203.13310*, 2022. 1, 2, 3, 6, 7, 8, 12

[65] Yunpeng Zhang, Jiwen Lu, and Jie Zhou. Objects Are Different: Flexible Monocular 3D Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3289–3298, 2021. 1, 3, 7, 13

[66] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects As Points. *arXiv:1904.07850 [cs]*, Apr. 2019. arXiv: 1904.07850. 1, 3, 4, 7, 12

[67] Yunsong Zhou, Yuan He, Hongzi Zhu, Cheng Wang, Hongyang Li, and Qinhong Jiang. Monocular 3D Object Detection: An Extrinsic Parameter Free Approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7556–7566, 2021. 1, 3

[68] Yunsong Zhou, Yuan He, Hongzi Zhu, Cheng Wang, Hongyang Li, and Qinhong Jiang. Monoef: Extrinsic parameter free monocular 3d object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 3

[69] 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. 1