LET-3D-AP: Longitudinal Error Tolerant 3D Average Precision for Camera-Only 3D Detection

Wei-Chih Hung  Vincent Casser  Henrik Kretzschmar  Jyh-Jing Hwang  Dragomir Anguelov
Waymo LLC

Abstract—The 3D Average Precision (3D AP) relies on the intersection over union between predictions and ground truth objects. However, camera-only detectors have limited depth accuracy, which may cause otherwise reasonable predictions that suffer from such longitudinal localization errors to be treated as false positives. We therefore propose variants of the 3D AP metric to be more permissive with respect to depth estimation errors. Specifically, our novel longitudinal error tolerant metrics, LET-3D-AP and LET-3D-APL, allow longitudinal localization errors of the prediction boxes up to a given tolerance. To evaluate the proposed metrics, we also construct a new test set for the Waymo Open Dataset, tailored to camera-only 3D detection methods. Surprisingly, we find that state-of-the-art camera-based detectors can outperform popular LiDAR-based detectors with our new metrics past 10% depth error tolerance, suggesting that existing camera-based detectors already have the potential to surpass LiDAR-based detectors in downstream applications. We believe the proposed metrics and the new benchmark dataset will facilitate advances in the field of camera-only 3D detection by providing more informative signals that can better indicate the system-level performance.

I. INTRODUCTION

Detecting objects in 3D space is a fundamental task in many robotics applications, including autonomous driving, unmanned aerial vehicles, robot navigation, and augmented reality. While LiDAR-based object detection [1]–[7] has been studied extensively in recent years with impressive results, reliable camera-based 3D object detection remains a challenging and active area of research [8]–[20].

According to common metrics, LiDAR-based detectors outperform camera-based detectors by a large margin, suggesting that camera-based object detectors fail to detect many objects. However, a careful evaluation of the failure cases reveals that many camera-based detectors identify objects reasonably well. The issue is rooted in these detectors often poorly estimating depth. Common metrics, such as the 3D Average Precision (3D AP), rely on the intersection over union (IoU) to associate the prediction boxes with the ground truth boxes. As a result, they may treat reasonable predictions that suffer from these longitudinal localization errors as false positives, leading to lower and less informative scores.

We therefore propose a variant of the 3D AP metric that is more permissive to depth errors. The proposed longitudinal error tolerant metrics, LET-3D-AP and LET-3D-APL, allow longitudinal localization errors up to a given tolerance. Specifically, we define the longitudinal error to be the object localization error along the line of sight between the camera and the ground truth object. The maximum longitudinal error that our metric tolerates is an adjustable percentage of the depth estimation error. Commonly, such errors are penalized by scaling the precision. However, under certain conditions, this metric does not penalize any corrected errors and is therefore comparable to the original 3D AP metric, but with more tolerant matching. In contrast to this, the metric LET-3D-APL penalizes longitudinal localization errors by scaling the precision.

Fig. 1. Evaluating camera-only 3D detections when using the 3D Average Precision (3D AP) metric (left) and when using the proposed longitudinal error tolerant, LET-3D-AP(L) metric (right). The figures depict the bipartite matching between the detections (green) and the ground truth objects (black). Regular 3D AP matching (left) is based on the intersection over union (IoU) values and cannot match the detections that suffer from longitudinal localization errors, though the detection is reasonable and can provide useful signals to downstream modules. In contrast to this, the proposed LET-3D-AP(L), shown on the right, is more permissive by shifting the predictions to mitigate the longitudinal localization errors. We show the shifted predictions in blue, which are used for computing the longitudinal error tolerant intersection over union (LET-IoU). To account for the used longitudinal tolerance, we propose longitudinal affinity (LA) as a measure of how close the original prediction is to the ground truth in the longitudinal direction.

* Work done while at Waymo.
To evaluate the proposed metrics, we extended the existing Waymo Open Dataset [21] by (1) creating a new dedicated test set of 80 segments (80,000 images) with LiDAR data redacted, (2) providing camera-synchronized 3D box labels for the entire dataset, eliminating the camera/label synchronization gap. Waymo included these extensions as part of the Waymo Open Dataset 3D Camera-Only Detection Challenge. The challenge remains open for submissions, and can be used to benchmark camera-only detection methods.

II. RELATED WORK

Following popular 2D detection benchmarks, such as PASCAL VOC [22] and COCO [23], most 3D object detection benchmarks rely on the metric 3D AP to evaluate detections based on the intersection over union (IoU), either in 3D or in a bird’s eye view (BEV), with predefined IoU thresholds. For example, the KITTI dataset [24] and the Waymo Open Dataset [21] adopt 3D IoU as the main matching function. The Waymo Open Dataset proposes to use a heading accuracy weighted AP, referred to as aP.H, as the primary metric to penalize incorrect heading prediction. NuScenes [25] uses the center distance between predicted objects and ground truth objects as the true positive matching criterion and proposes a set of true positive metrics to quantify other errors, including localization, scale, and orientation.

While LiDAR-based detectors [1]–[7] remain the most popular methods for 3D object detection for autonomous driving, monocular camera-based 3D object detection has been gaining traction in recent years. Methods can be roughly categorized into two groups: 1) The first camp leverages mature perspective 2D detectors and equips the networks with 3D box attributes for 3D detection [8]–[14], [26]–[30]. An emerging direction is to leverage Transformer models for implicit depth encoding [17], [31]–[34]. 2) The second camp leverages mature LiDAR detectors and projects 2D perspective features into 3D in the form of point clouds [16], [20], [35]–[40], feature maps in bird’s eye view (BEV) [15]–[17], [41]–[52], or in voxel space [18], [19], [53], [54]. These methods then predict the detections by applying 3D detection heads to the 3D/BEV feature maps. Accurate monocular depth estimation is essential for all methods. Since monocular depth estimation is intrinsically an ill-posed problem, however, it is difficult for camera-based methods to accurately estimate the depth that corresponds to the objects, resulting in longitudinal localization errors that lead to reduced 3D AP scores when compared to LiDAR-based methods.

As pointed out by Ma et al. [55], the performance of monocular camera 3D detection methods can be greatly improved when their longitudinal localization errors are mitigated by using ground truth depth or localization. This is why we are proposing longitudinal error tolerant (LET) metrics for evaluating camera-only 3D detection methods. While generalized intersection over union (GIoU) [56] and its 3D variant [57] can also be used to match non-overlapping bounding boxes, they do not target errors in a specific direction, and the shape and heading mismatch is not factored into the non-overlapping cases.

III. LONGITUDINAL ERROR TOLERANT 3D AP

The 3D AP metric relies on the IoU to match prediction boxes with ground truth boxes. Therefore, prediction boxes that have little or no overlap with the corresponding ground truth boxes will be treated as false positives. However, these predictions may still contribute valuable information to the decision making of an autonomous driving system. We therefore propose a metric that rewards such detections.

Our proposed metrics are inspired by the following objectives: First, given a model of an assumed localization error distribution, design a new matching criterion so that a predicted box may still match with a target ground truth box even when they do not match in terms of IoU. Second, design a new bipartite matching cost function that takes the localization, shape, and heading errors into account so that frame-level matching can be properly calculated. Last, design a penalty term to penalize detections that can only be matched with the ground truth when using the longitudinal error tolerant matching.

A. Localization Errors in Camera-Based 3D Detection

We assume a ground truth bounding box $G$ with center $\vec{g} \in \mathbb{R}^3$ and a predicted bounding box $P$ with center $\vec{p} \in \mathbb{R}^3$ such that the origin $[0, 0, 0]$ is the camera location, or a mean position in the case of multi-camera setup. That is, all 3D location vectors are in the camera frame. We define the localization error as the vector

$$\vec{e}_{\text{loc}} = \vec{p} - \vec{g}. \quad (1)$$

We decompose the localization error into two components: noitemsep

- The longitudinal error $\vec{e}_{\text{lon}}$ is the error along the line of sight from the center of the prediction box to the center of the ground truth box, giving $\vec{e}_{\text{lon}} = (\vec{e}_{\text{loc}} \cdot \vec{u}_G) \cdot \vec{u}_G$, where $\vec{u}_G = \vec{g}/|\vec{g}|$ is the unit vector of $\vec{g}$.
- The lateral error $\vec{e}_{\text{lat}}$ is the distance between the predicted box $P$ and the line of sight to the ground truth box $G$. It is defined as the shortest distance from the predicted box to any point on the line of sight, leading to $\vec{e}_{\text{lat}} = \vec{e}_{\text{loc}} - \vec{e}_{\text{lon}}$.

Figure 2 illustrates the error terms. We observe that localization errors tend to have the following attributes for...
camera-based 3D detectors. Localization errors tend to be the most pronounced along the line of sight because of imperfect depth estimation. We assume that the standard deviation of the longitudinal error, $e_{\text{lon}}$, is proportional to the distance between the sensor and the center of the ground truth bounding box, $\vec{g}$. The lateral error, $e_{\text{lat}}$, is the result of an imperfect estimation of the object center on the camera image plane. Since the size of the bounding box projected onto the camera plane is inversely proportional to the range to a given ground truth, $G$, the standard deviation of the center estimation error in pixels, $\sigma(|\vec{r}_{\text{cam}}|)$, is also inversely proportional to the range of a given ground truth $\vec{g}$, that is, $\sigma(|\vec{r}_{\text{cam}}|) \propto 1/|\vec{g}|$. The lateral error in 3D space, however, is scaled by the range projection, that is, $e_{\text{lat}} = |\vec{g}| \cdot \sigma(|\vec{r}_{\text{cam}}|)$. Therefore, the standard deviation of the lateral error is independent of the range of the ground truth object and can be set as a constant. We therefore only introduce a tolerance for the longitudinal localization errors.

B. Longitudinal Affinity

We propose a scalar value, longitudinal affinity $a_l(\vec{p}, \vec{g})$, to determine the scores for matching prediction boxes with ground truth boxes given a tolerance for the longitudinal error. Specifically, the longitudinal affinity, whose value is in $[0, 1.0]$, estimates how well the centers of a prediction box and the ground truth box align. Given the longitudinal error, $e_{\text{lon}}$, between a pair of prediction and ground truth, we define the longitudinal affinity based on the following hyperparameters: noiteseq

- longitudinal_tolerance_percentage $T_l^p$: The maximum longitudinal error $e_{\text{lon}}$ is expressed as a percentage $T_l^p$ of the range to the ground truth $G$. For example, $T_l^p = 0.1$ provides a 10% tolerance, and for a ground truth object that is 50 meters away ($|\vec{g}| = 50$), the longitudinal tolerance is 5 meters.
- min_longitudinal_tolerance_meter $T_l^m$: When a ground truth object is close to the sensor origin, the percentage-based tolerance can result in an small matching region. This parameter controls the minimum absolute tolerance, thus mainly affecting near range objects.

Finally, we define the longitudinal affinity $a_l(\vec{p}, \vec{g})$ between a prediction box with center $\vec{p}$ and a ground truth box with center $\vec{g}$ as:

$$a_l(\vec{p}, \vec{g}) = 1 - \min\left(\frac{|\vec{e}_{\text{lon}}(\vec{p}, \vec{g})|}{T_l}, 1.0\right),$$  

where $T_l = \max(T_l^p \cdot |\vec{g}|, T_l^m)$.

C. LET-IoU: Longitudinal Error Tolerant IoU

The longitudinal affinity captures the longitudinal error of a prediction based on the center of the predicted bounding box and the center of the ground truth bounding box. To determine whether a prediction can be associated with a ground truth, we also want to take into account the shape, size, and heading error. When relying on the regular 3D AP, this is only captured by 3D IoU. We propose LET-IoU, where IoU is calculated between the ground truth bounding box and the prediction bounding box after compensating for the longitudinal error. Specifically, we mitigate the longitudinal error of the prediction by aligning its center along the line of sight with the ground truth.

Given a ground truth with center $\vec{g}$ and a prediction with center $\vec{p}$, the objective is to move the center location of the prediction box along the line of sight so that the IoU between the moved prediction box and the ground truth box is maximized:

$$\vec{p}_{\text{aligned}} = \arg\max_{\vec{p}_{\text{aligned}}} \text{3D-IoU}(P_{\text{aligned}}, G),$$  

where $P_{\text{aligned}}$ is the prediction box $P$ with updated center $\vec{p}_{\text{aligned}}$. However, there is no closed-form solution to (3), and exhaustive search along the line of sight is computationally expensive. We therefore approximate the objective by minimizing the distance between the moved prediction box center and the ground truth center, where the longitudinal error is compensated. In other words, the center of the aligned prediction is the projection of the ground truth center $\vec{g}$ onto the line of sight from the sensor to the prediction object, leading to

$$\vec{p}_{\text{aligned}} = (\vec{g} \cdot \vec{u}_P) \cdot \vec{u}_P,$$

where $\vec{u}_P = \vec{p}/|\vec{p}|$ is the unit vector along the line of sight to prediction center $\vec{p}$. A detailed analysis on the quality of approximation (4) can be found in the supplementary material.
Finally, we obtain the LET-IoU by calculating the typical 3D IoU between the aligned prediction box $P_{aligned}$ and the target ground truth box $G$, giving

$$\text{LET-IoU}(P, G) = 3\text{D-IoU}(P_{aligned}, G).$$

(5)

See Figure 3 for an illustration. Note that for simplicity, this function does not assume that the longitudinal error is within a given tolerance. In the final metrics computation, LET-IoU will only be calculated on pairs with non-negative longitudinal affinity, where $a_l > 0$.

Figure 4 shows that the prediction (green) is aligned to the target ground truth (red), where the 3D IoU is computed between the aligned prediction (yellow) and the ground truth, resulting in 0.7 LET-IoU.

D. Bipartite Matching with Longitudinal Error Tolerance

To calculate the precision and recall values, detection metrics need to perform a bipartite matching between the prediction set and the ground truth set. Most bipartite matching algorithms involve computing an association weight matrix $W \in \mathbb{R}^{N_P \times N_G}$, where $N_P$ and $N_G$ are the numbers of detections and ground truth objects. The objective is to maximize the summed weights among all the matched pairs. We may take into consideration the longitudinal error, the shape error, and potentially the heading error.

The typical matching weight function in 3D AP computes the IoU between a prediction box and a ground truth box with an IoU threshold $T_{iou}$:

$$W(i, j) = \begin{cases} 
\text{IoU}(P(i), G(j)), & \text{if } \text{IoU}(P(i), G(j)) > T_{iou} \\
0, & \text{otherwise}
\end{cases},$$

(6)

where $P$ and $G$ are the prediction box set and ground truth box set for a single frame, respectively. However, in our setting, there are two affinity terms to be considered: longitudinal affinity $a_l$ and LET-IoU. As a result, we propose to calculate the bipartite matching weight function where both terms are taken into account:

$$W(i, j) = \begin{cases} 
a_l \cdot \text{LET-IoU}, & \text{if } a_l > 0 \\
& \text{and } \text{LET-IoU} > T_{iou}, \\
0, & \text{otherwise}
\end{cases},$$

(7)

where we omit the function input terms $(P(i), G(j))$ of $a_l$ and LET-IoU for simplicity. This allows us to take the shape and heading error into account while prioritizing the detections that have higher longitudinal affinity.

After computing the matching weight matrix, we run a bipartite matching method (Greedy or Hungarian) to compute the matching results, including matched pairs (true positives / TP), non-matched predictions (false positives / FP), and non-matched ground truths (false negatives / FN).

E. LET-3D-AP and LET-3D-APL

a) LET-3D-AP: Average Precision with Longitudinal Error Tolerance.: Once the matching results are finalized, each matched prediction will be counted as a true positive (TP). A prediction without matching ground truth is counted as a false positive (FP). If a ground truth is not matched with any prediction, the ground truth is counted as a false negative (FN). Then the precision $p = |\text{TP}|/(|\text{TP} + |\text{FP}|)$ and recall $r = |\text{TP}|/(|\text{TP} + |\text{FN}|)$ can be calculated. After computing the precision and recall values for detection subsets with different score cutoffs, a PR curve can be obtained. Finally, the LET-3D-AP is calculated from the average precision of the PR curve: LET-3D-AP = $\int_0^1 p(r)dr$, where $p(r)$ is the precision value at recall $r$.

Here, we do not penalize depth errors when computing the PR curve, even though we may have had to adjust some predictions to compensate for their depth errors. This provides a number comparable to 3D AP but with a more tolerant matching criterion.

b) LET-3D-APL: Longitudinal Affinity Weighted LET-3D-AP.: Here, we penalize those predictions that do not overlap with any ground truth, that is, those which only match a ground truth bounding box owing to the above-mentioned center alignment. We penalize these predictions by using the longitudinal affinity $a_l(\bar{p}, \bar{g})$ proposed above. To this end, we propose a weighted variant of precision values.

Though the number of true positives $|\text{TP}|$ is used both in the precision and recall computation, they represent different ideas. In precision calculation, $|\text{TP}_P|$ means the number of matched predictions, while in the recall calculation, $|\text{TP}_G|$ means the number of matched ground truths. As a result, we rewrite the precision and recall as

$$\text{Precision} = \frac{|\text{TP}_P|}{|\text{TP}_P| + |\text{FP}|}$$

(8)

$$\text{Recall} = \frac{|\text{TP}_G|}{|\text{TP}_G| + |\text{FN}|}$$

(9)

to emphasize the difference between the prediction accumulator and the ground truth accumulator.

In the precision computation, we traverse through the predictions to accumulate the FP's and TP's. For any unmatched prediction, it only contributes to the FP-accumulator with 1.0. However, for a matched prediction P with respect to a ground truth $\bar{g}$, it contributes to TP-accumulator with the quantity of $a_l(\bar{p}, \bar{g})$ and also contributes to the FP-accumulator with the quantity of $1 - a_l(\bar{p}, \bar{g})$. Essentially, we distribute part of the TP-accumulator to the FP-accumulator based on the longitudinal affinity. Then, the soft TP- and FP-accumulators are:

$$|\text{TP}_P| = \sum_{\text{matched } P} a_l(\bar{p}, \bar{g})$$

(10)

$$|\text{FP}| = \sum_{\text{matched } P} (1 - a_l(\bar{p}, \bar{g})) + \sum_{\text{unmatched } P} 1.0$$

(11)

Based on the definition of soft TP and FP, the soft precision can then be computed as:

$$\text{Prec}_L = \frac{|\text{TP}_P|}{|\text{TP}_P| + |\text{FP}|} = a_l \cdot \text{Precision}$$

(12)

which results in a mean longitudinal affinity weighted precision point.

For the recall computation, we propose to not weight the $TP_G$ since the ground truth does get matched, and there is no reason to penalize the metric twice. As a result, the
precision values are discounted by the multiplier $\pi_l$, which is the average longitudinal affinity of all the matched predictions that was treated as TP.

Finally, the LET-3D-APL can be calculated with the longitudinal affinity weighted PR curve:

$$\text{LET-3D-APL} = \int_0^1 \text{Prec}_L(r) dr = \int_0^1 \pi_l \cdot \text{Prec}(r) dr,$$

where $\text{Prec}_L(r)$ is the longitudinal affinity weighted precision, and $\text{Prec}(r)$ is the precision at recall $r$. Note that the weight $a_l(\vec{p}, \vec{q})$ depends on the specific score cutoff at the PR point and therefore cannot be taken out of the integral.

IV. CAMERA-PRIMARY 3D DETECTION DATASET

To evaluate the proposed LET metrics, we extend the existing Waymo Open Dataset [21]. Our contributions are two-fold: (1) We provide a dedicated test set of 80 segments with LiDAR sensor data redacted, (2) We significantly improve camera/label synchronization throughout the dataset. Waymo conducted the Waymo Open Dataset 3D Camera-Only Detection Challenge* between April and May 2022, with LET-3D-APL being the primary metric for determining the winning submissions.

Our dedicated test set contains 80 segments in the same format as the original dataset release (20s, 10Hz, 5 cameras), for a total of 1,000 camera images per segment and 80,000 images overall. LiDAR sensor data was used during labeling to ensure high-quality 3D boxes, but redacted before the release to ensure compliance with the dataset challenge rule with regards to basing predictions solely on cameras.

The original Waymo Open Dataset has a synchronization gap of $[-6, 7]$ms between camera and LiDAR sensor data [21]. Since the 3D object labels are based on LiDAR sensor data, this gap translates into a corresponding synchronization gap between camera and LiDAR-based labels. Motion of the ego-vehicle and independent object motion can lead to misaligned labels with respect to the camera point of view.

To address this issue, we add a new field `CAMERA_SYNCHED_BOX`, a variant of `BOX`, which is adjusted to eliminate the synchronization gap to the camera that perceives it. The adjustment is based on solving an optimization problem that takes into account the camera rolling shutter, the motion of the autonomous vehicle and the motion of the object of interest.

V. EXPERIMENTAL RESULTS

In this section, we analyze the proposed LET metrics against both LiDAR-based and camera-only 3D detectors.

A. Evaluation Dataset and Baseline Methods

To verify the proposed LET metrics, we evaluate two LiDAR-based detectors and two camera-only 3D detectors trained on the Waymo Open Dataset [21] (WOD) and report metrics on the newly proposed test set.

**SWFormer** [58]: A state-of-the-art LiDAR-based 3D detector that builds upon the idea of Swin Transformer [59]

and operates on sparse voxels. It achieves 73.36 L2 vehicle mAPH on the WOD original test set.

**PointPillars** [3]: A well-established LiDAR-based 3D detector baseline that achieves 60.05 L2 vehicle mAPH on the WOD original test set.

**BEVFormer** [41]: A camera-only 3D detector that learns unified BEV representations with spatio-temporal transformers to tackle both 3D detection and segmentation tasks. The method achieves state-of-the-art performance on nuScenes [25] and WOD. We evaluate a specific variant of BEVFormer that won the 2022 Waymo Open Dataset 3D Camera-Only Detection Challenge.

**MV-FCOS3D++** [30]: A camera-only 3D detector that adds 2D supervision and temporal information onto the BEV style detector and achieves 2nd place in the 2022 Waymo Open Dataset 3D Camera-Only Detection Challenge.

B. Comparisons of Longitudinal Tolerance Values

The main hyperparameter of the proposed metrics is `longitudinal_tolerance_percentage`. We show the performance of the LiDAR-based detectors (SWFormer, PointPillars) and camera-only detectors (BEVFormer, MV-FCOS3D++) with different tolerance values in Figure 5 for the vehicle class.

As shown in Figure 5, higher longitudinal error tolerance leads to higher LET-3D-AP and LET-3D-APL because more ground truth objects are matched with detections. Note that for the LiDAR-based detectors, LET-AP remains similar to the AP values for all longitudinal tolerance values since the longitudinal errors of LiDAR-based detections are already small. The results also suggest that the proposed LET metrics

---

*https://waymo.com/open/challenges/2022/3d-camera-only-detection
TABLE I. LET metrics of different breakdowns. The methods are evaluated on the new test set with 80 multi-camera sequences. We report 3D AP, LET-3D-AP, LET-3D-APL, and mean longitudinal affinity (mLA) for all the class and range breakdowns, with 10% of longitudinal tolerance. For the LiDAR-based detector, the proposed metrics show similar performance since most longitudinal tolerance is not used for matching predictions and ground truths. For camera-based detectors, the results show that proposed LET metrics are more suitable for evaluation since predictions can be better matched with ground truths, especially for small objects like pedestrian and long-range detections as indicated by the difference between 3D AP and LET-3D-AP(L).

| Method    | Metrics          | All | Vehicle | Pedestrian | Cyclist | Range (Vehicle) | [0, 30) | [30, 50) | [50, ∞) |
|-----------|------------------|-----|---------|------------|---------|----------------|--------|----------|---------|
| SWFormer  | 3D AP (%)        | N/A | 89.0    | 86.7       | N/A     | 96.0           | 89.5   | 78.7     |
|           | LET-3D-AP (%)    | N/A | 89.7    | 87.3       | N/A     | 96.2           | 90.2   | 79.7     |
|           | LET-3D-APL (%)   | N/A | 86.3    | 85.4       | N/A     | 91.6           | 87.9   | 78.1     |
|           | mLA              | N/A | 0.962   | 0.978      | N/A     | 0.952          | 0.975  | 0.980    |
| PointPillars | 3D AP (%)    | N/A | 78.6    | 58.3       | N/A     | 88.2           | 76.4   | 58.4     |
|           | LET-3D-AP (%)    | N/A | 80.7    | 62.7       | N/A     | 88.5           | 78.5   | 66.2     |
|           | LET-3D-APL (%)   | N/A | 76.7    | 61.0       | N/A     | 83.0           | 75.9   | 54.1     |
|           | mLA              | N/A | 0.950   | 0.973      | N/A     | 0.938          | 0.967  | 0.817    |
| BEVFormer | 3D AP (%)        | 35.3| 50.7    | 22.3       | 33.0    | 79.7           | 47.2   | 20.5     |
|           | LET-3D-AP (%)    | 70.7| 82.9    | 71.1       | 58.1    | 89.7           | 80.6   | 50.1     |
|           | LET-3D-APL (%)   | 56.2| 68.8    | 53.2       | 46.5    | 75.3           | 68.5   | 38.9     |
|           | mLA              | 0.795| 0.830  | 0.748      | 0.800   | 0.839          | 0.850  | 0.776    |
| MV-FCOS3D++ | 3D AP (%)    | 30.3| 45.6    | 16.7       | 28.5    | 79.2           | 37.5   | 13.0     |
|           | LET-3D-AP (%)    | 66.0| 82.1    | 63.0       | 53.0    | 89.3           | 78.0   | 66.1     |
|           | LET-3D-APL (%)   | 51.1| 66.9    | 45.0       | 41.3    | 74.5           | 63.6   | 52.0     |
|           | mLA              | 0.774| 0.815  | 0.714      | 0.779   | 0.834          | 0.815  | 0.787    |

C. LET-AP vs. AP

We compare the proposed LET-AP and LET-APL metrics with AP on the selected baselines and show results in Figure 5. For the LiDAR-based detectors (SWFormer, PointPillars), we show that LET-AP is mostly consistent with AP, which suggests that the matching results are similar in the presence of longitudinal tolerance. For higher tolerance values, LET-3D-AP is slightly higher than 3D AP due to the small amount of additional detections being matched with ground truth objects with more error tolerance. Also, due to the accuracy of LiDAR range estimates, the boxes are fairly accurate, resulting in minimal difference between LET-AP and LET-APL since the allowed longitudinal tolerance is not used by most matching pairs.

The LET-AP metrics results of the camera-only detectors in Figure 5 exhibit a much larger gap to the AP baseline. This shows that many detections can be properly matched to ground truth objects with a certain amount of longitudinal tolerance, making the resulting LET-AP(L) metrics better suited for evaluating these methods. Under 10% tolerance, BEVFormer achieves higher metrics than PointPillars, indicating BEVFormer may actually have better downstream performance provided sufficient robustness towards localization errors. Compared to the LiDAR-based detectors, the drop off between LET-AP and LET-APL is larger, demonstrating higher longitudinal error among the true positive examples.

D. Metrics of Different Breakdowns

We show detailed metrics on the proposed new test set in Table I. We report 3D AP, LET-3D-AP, LET-3D-APL, and mean longitudinal affinity (mLA) for all class and range breakdowns, with 10% of longitudinal tolerance. Note that the 3D AP metrics are computed with $\theta = 0.5, 0.3, 0.3$ IoU thresholds for vehicles, pedestrians and cyclists. As shown in Table I, LET-3D-AP values are consistently higher than 3D AP values since more detections are matched with ground truth objects with additional longitudinal tolerance. The difference is especially pronounced for small objects like pedestrians, as well as far-away objects. We also report the mean longitudinal affinity values to show the difference between LET-3D-AP and LET-3D-APL. We observe that methods with higher metrics not only have more valid detections but also have better localization as indicated by higher mLA. We can also verify the assumption of depth errors being roughly proportional to range by checking the consistent mLA values with different range breakdowns. Our results suggest that, under the proposed metrics, some camera-only 3D detectors can already surpass the LiDAR-based detectors.

VI. Conclusion

We proposed LET-3D-AP(L) metrics for evaluating camera-only 3D detectors. The tolerance along the longitudinal axis allows detections to be associated with the ground truth objects despite depth estimation errors. The results show that state-of-the-art camera-based detectors can outperform popular LiDAR-based detectors with our metrics, suggesting that existing camera-based methods have the potential in real-world applications. We also construct a new test set for the Waymo Open Dataset, tailored to camera-only 3D detection methods. We hope the proposed metrics and dataset will help advance the field of camera-only 3D detection by providing a more meaningful indication for method performance.

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
[52] W. Roh, G. Chang, S. Moon, G. Nam, C. Kim, Y. Kim, S. Kim, and J. Kim, “Ora3d: Overlap region aware multi-view 3d object detection,” arXiv preprint arXiv:2207.00865, 2022.
[53] T. Wang, J. Pang, and D. Lin, “Monocular 3d object detection with depth from motion,” in ECCV, 2022.
[54] J. Lu, Z. Zhou, X. Zhu, H. Xu, and L. Zhang, “Learning ego 3d representation as ray tracing,” arXiv preprint arXiv:2206.04042, 2022.
[55] X. Ma, Y. Zhang, D. Xu, D. Zhou, S. Yi, H. Li, and W. Ouyang, “Delving into localization errors for monocular 3d object detection,” in CVPR, 2021.
[56] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, “Generalized intersection over union,” in CVPR, 2019.
[57] J. Xu, Y. Ma, S. He, and J. Zhu, “3d-giou: 3d generalized intersection over union for object detection in point cloud,” Sensors, 2019.
[58] P. Sun, M. Tan, W. Wang, C. Liu, F. Xia, Z. Leng, and D. Anguelov, “Swformer: Sparse window transformer for 3d object detection in point clouds,” in ECCV, 2022.
[59] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” in CVPR, 2021.