2025 IEEE International Conference on Robotics and Automation (ICRA)
May 19-23, 2025. Atlanta, USA

Stands on Shoulders of Giants: Learning to Lift 2D Detection to 3D with Geometry-Driven Objectives

Jhih-Rong Chen*† Che-Yuan Chang*† Szu-Han Tseng ‡
Chih-Sheng Huang‡ Yong-Sheng Chen† Wei-Chen Chiu†
†National Yang Ming Chiao Tung University, Taiwan ‡ELAN Microelectronics Corp., Taiwan

2025 IEEE International Conference on Robotics and Automation (ICRA) | 979-8-3315-4139-2/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICRA55743.2025.11128173

***Abstract*— 3D detection of vehicles is an essential component for autonomous driving applications. Nevertheless, collecting the supervised training data for learning 3D vehicle detectors would be costly (e.g. utilization of expensive LiDAR sensors) and labor-intensive (for human annotation). In comparison to 3D detection, 2D object detection has achieved a well-developed status, boosting stable and robust performance with widespread application in numerous fields, thanks to the large scale (i.e. amount of samples) of existing training datasets of 2D object detection. Hence, in our work, we propose to realize 3D detection via leveraging the robustness of 2D detectors and developing a network that lifts 2D detections to 3D.**

With the flexibility of building upon various backbone models (e.g. the models which take image regions detected by 2D detector as inputs to predict their corresponding 3D bounding boxes, or the existing monocular 3D detection models which have the intermediate output of 2D bounding boxes), we propose several geometry-driven objectives, including *projection consistency loss*, *geometry depth loss*, and *opposite bin loss*, to improve the training upon 2D-to-3D lifting. Our extensive experimental results demonstrate that our proposed geometry-driven objectives not only contribute to the superior results of 3D detection but also provide better generalizability across datasets.

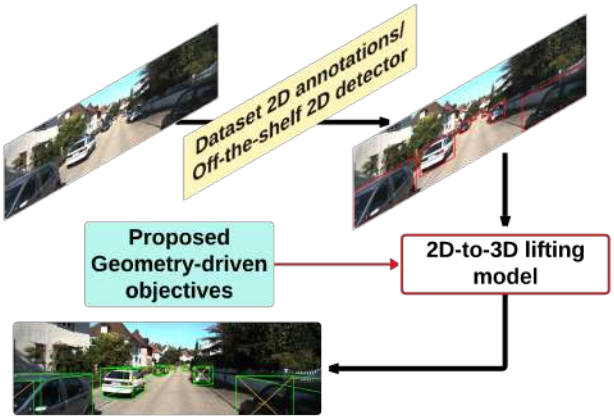

Fig. 1. First, we leverage existing 2D object detectors to obtain the 2D bounding boxes of objects (stands on shoulders of giants). Then, we apply a 2D-to-3D lifting model to transform the 2D detections into corresponding 3D bounding boxes. Our framework is designed to decouple 3D object detection into 2D detection process and the 2D-to-3D transformation, allowing the model to focus solely on mastering the 2D-to-3D lifting without the complexities of detection. The proposed geometry-driven objectives further enhance the model to predict more accurate and robust 3D attributes.

I. INTRODUCTION

Object detection, particularly on vehicles, is a critical component in the application scenarios of autonomous driving. Basically, object detectors have the main objective of identifying and localizing objects within a given scene. The current works of object detection methods can be roughly categorized into two groups, i.e. 2D and 3D object detections, according to the format of output (in which 2D and 3D detections result to 2D and 3D bounding boxes, respectively). In practice, 2D object detectors have been comprehensively developed and widely used across various scenarios, demonstrating their efficiency and applicability, while 3D detectors begin to attract more and more research attention due to their ability to provide richer geometric structure. However, learning monocular 3D object detection is typically quite difficult (as now the output contains only not the 3D locations of the target objects, but also their 3D attributes composed of 3D dimensions and orientations) requires large quantity of training data, where the cost of collecting dataset with groundtruth annotations also becomes extremely expensive. Moreover, the complexity of annotating 3D attributes in turn also limits the overall number of training samples in the 3D detection datasets (in comparison, the scale of existing 2D detection datasets typically is larger than the 3D ones). In results, we can observe from our pilot study that the off-the-shelf 2D detectors (which have larger amount of training data to learn a simpler task) experimentally offers higher recall and more stable performance across diverse scenes in comparison to the 3D ones (which have less quantity of training data and require to learn a harder task).

Motivated by the discussion and the pilot study above, in this work, we would like to explore the following research question: "*in light of the well-established foundation of 2D object detection, are we able to utilize the reliability of 2D detection for enabling/boosting 3D detection?*" Specifically, we propose to establish a **2D-to-3D lifting model** which is trained to predict the attributes of 3D bounding boxes from the image content and the given 2D detections, where such lifting model acts as an efficient bridge to fill the gap between 2D and 3D detection capabilities. Moreover, we further utilize the geometric relationships between 2D and 3D to design novel geometry-driven loss functions, enabling a more effective lifting/transformation.

There are already existing works (e.g. Deep3DBox [1])

∗The authors contributed equally to this work.

of building 2D-to-3D lifting model, or we can leverage the 3D detection models which have the intermediate stage of predicting 2D bounding boxes (e.g. GUPNet [2], where we can replace its 2D bounding boxes with the ones from off-the-shelf 2D detectors, or simply treat its network components prior to this intermediate stage as a 2D detector) and extend them to support the lifting operation. In comparison to them, our proposed method in this paper focuses more on developing the objective functions which can not only boost the training of lifting model (as well as enhance the estimation of 3D attributes for 3D bounding boxes) but also have the flexibility to be integrated with various network architectures of lifting models. Particularly, we propose three geometry-driven objective functions dedicated to utilizing the monocular image content as well as the 2D/3D geometric properties of objects, including:

- **projection consistency loss** which considers the consistency between the given 2D detections and the 2D bounding boxes produced by reprojecting the lifted 3D bounding boxes back onto the image plane;
- **geometric awareness loss** which firstly parameterizes the depth (named as geometric depth) by using the 3D attributes (e.g. 3D dimensions and the orientation of objects), followed by minimizing the errors between the geometric depth derived from groundtruth 3D attributes and the one derived from estimated 3D attributes;
- **opposite bin loss** which specifically aims to improve the estimate of object orientation via minimizing the confusion between front and rear features of vehicles.

We have conducted extensive experiments across multiple datasets, including KITTI [3], nuScenes-mini [4], and a private Taiwan street scene dataset provided by ELAN Microelectronics Corp, to verify the effectiveness of the proposed geometry-driven objectives. Our framework, employing off-the-shelf 2D object detectors and training with the proposed geometry-driven, demonstrates superior performance in constructing and predicting robust 3D bounding boxes across various scenarios.

II. Related Works

A. 2D Object Detection

With the rapid advancement of deep learning and its widespread adoption in 2D object detection, significant progress has been made in developing high-performing detection models. Pioneering frameworks such as the R-CNN family [5], [6], [7], [8] and the YOLO series [9], [10], [11], [12], [13] have set benchmarks in this domain. The R-CNN models, by leveraging region proposal mechanisms, have progressively optimized feature extraction and object localization. YOLO series, with the gradual integration of new technologies in each generation, ultimately achieved excellent detection performance. Building on these foundational advancements, 2D object detection has become highly reliable and is now widely used in various real-world applications. This reliability inspires us to integrate pre-existing 2D detectors into our framework.

B. Monocular 3D Object Detection

Monocular 3D object detection has become a critical research area due to its widespread applications in autonomous driving, robotics, and augmented reality. Despite the absence of explicit depth information in single-image inputs, numerous methods have been developed to address this challenge by incorporating learning strategies, geometric constraints, and prior knowledge. Early works are driven by the use of constraints and adapt 2D object detection frameworks to infer 3D information [1], [14], [15]. More recent approaches have introduced multi-task learning techniques that simultaneously estimate keypoints and 3D attributes [2], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25]. For example, SMOKE [18] completely bypasses 2D detection and directly predicts 3D centers by using keypoint estimation. However, [19] revisits the misalignment between 2D bounding box centers and projected 3D object centers, arguing that 2D detection remains a crucial component. As a result, many recent methods continue to incorporate 2D detection in monocular 3D object detectors [2], [22], [23], [24], [26], [27], [25]. Additionally, several approaches have exploited geometric relationships between objects [28], [29]. Specifically, [28] improves object location precision by enforcing spatial constraints on object pairs, while [29] treats all 3D objects in an image as a unified whole and introduces a novel loss function. These innovations inspire the design of geometry-driven objectives, which can be integrated into monocular 3D detectors.

III. Proposed Method

A. 2D-to-3D lifting model

In this work, we focus on transforming/lifting 2D detections into 3D bounding boxes, as illustrated in Fig. 2. Using GUPNet as an example, we replace its original branch responsible for predicting 2D parameters with an off-the-shelf 2D detector or existing 2D annotations. This modification enables the model to concentrate solely on learning how to transform these known 2D detections into their corresponding 3D bounding boxes, improving its performance in 3D space.

B. Geometry-Driven Objectives

Basically, given an input monocular image $I \in \mathcal{R}^{W \times H \times 3}$ and a 2D bounding box $\mathbf{B}_{2d} = (u, v, w_{2d}, h_{2d})$, where W, H denote the width and height of the image I, (u, v) represents the center of $\mathbf{B}_{2d}$, and (w_{2d}, h_{2d}) indicates the width and height of $\mathbf{B}_{2d}$ in pixels on the image, our objective is to accurately regress the corresponding 3D attributes $(x, y, z, h_{3d}, w_{3d}, l_{3d}, \theta)$ from $\mathbf{B}_{2d}$. Here, (x, y, z) and (h_{3d}, w_{3d}, l_{3d}) represent the 3D center and the 3D dimensions of the object (both measured in meters), respectively, while θ is the yaw angle describing the orientation of the object. Please note that for orientation, both roll and pitch angles are assumed to be zero, following practice as [1]; while both the intrinsic matrix K and the extrinsic matrix E of the imaging camera are assumed to be known as well during both training and inference.

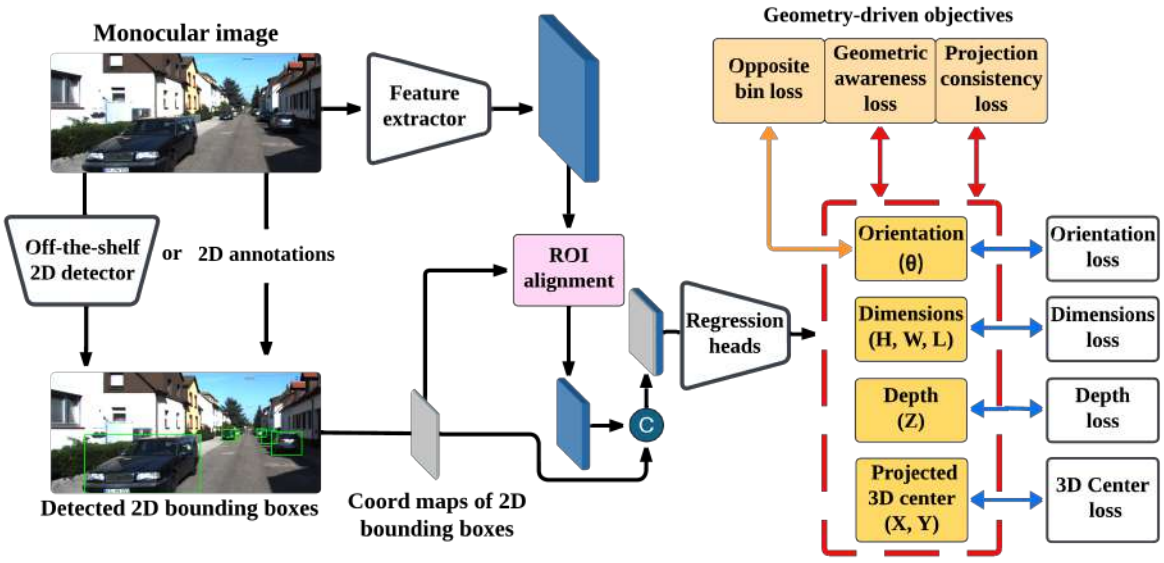

Fig. 2. **Overview of GUPNet with our proposed method for lifting all 2D detection to 3D space.** Given an input monocular image and the 2D bounding boxes of vehicles (which can be obtained using an off-the-shelf 2D object detector), the object-level features are firstly extracted with the help of feature extractor and RoI-align operation. The concatenation between the extracted object-level features and the coordinate maps of their 2D bounding boxes will go through the regression heads to estimate the 3D attributes of the 3D bounding boxes. In addition to the typical objectives (noted with white rounded rectangles) which directly optimize the errors between estimated 3D attributes and their groundtruth. We additionally introduce three geometry-driven objectives (highlighted by orange rectangles). Please refer to Section III-B for details.

Projection Consistency Loss $\mathcal{L}_{proj}$. Given a 3D bounding box $\mathbf{B}_{3d}$ presented by a set of 3D attributes that are predicted/lifted from a 2D bounding box $\mathbf{B}_{2d}^{gt}$ via the lifting model, it can be easily projected onto the image place to obtain the corresponding 2D bounding box $\mathbf{B}_{2d}^{proj}$ (where the projection is based on the general mapping function of a pinhole camera model, driven by both intrinsic and extrinsic camera matrices K and E, we omit its equation here for simplicity), in which such projected 2D bounding box ideally should be accurately aligned with $\mathbf{B}_{2d}^{gt}$, thus leading to our projection consistency loss $\mathcal{L}_{proj}$ defined as the interaction-over-union (IoU) between $\mathbf{B}_{2d}^{proj}$ and $\mathbf{B}_{2d}^{gt}$:

$$\mathcal{L}_{proj} = 1 - \frac{\left|\mathbf{B}_{2d}^{proj} \cap \mathbf{B}_{2d}^{gt}\right|}{\left|\mathbf{B}_{2d}^{proj} \cup \mathbf{B}_{2d}^{gt}\right|} . \tag{1}$$

In practice, as there are 8 corners of a 3D bounding box $\mathbf{B}_{3d}$, where their 3D coordinates can be computed via

$$\begin{bmatrix} \cos(\theta) & 0 & \sin(\theta) \\ 0 & 1 & 0 \\ -\sin(\theta) & 0 & \cos(\theta) \end{bmatrix} \begin{bmatrix} \pm l_{3d}/2 \\ \pm h_{3d}/2 \\ \pm w_{3d}/2 \end{bmatrix} + \begin{bmatrix} x \\ y \\ z \end{bmatrix} , \tag{2}$$

according to the 3D attributes $(x, y, z, h_{3d}, w_{3d}, l_{3d}, \theta)$, we project these corners onto the image space followed by identifying their maximum and minimum values of 2D coordinate to construct the projected 2D bounding box $\mathbf{B}_{2d}^{proj}$.

Intuitively, as the size/area of a projected 2D bounding box $\mathbf{B}_{2d}^{proj}$ is mainly determined by the corresponding 3D dimensions (h_{3d}, w_{3d}, l_{3d}) and yaw angle θ (assuming depth remains unchanged), the employment of our projection consistency loss $\mathcal{L}_{proj}$ contributes to enhance their estimations.

Geometric Awareness Loss $\mathcal{L}_{geo}$. Here, we incorporate additional information by introducing geometric depth $\mathbf{d}$ as a form of supervision. $\mathbf{d}$ is derived from an algorithm that leverages 2D bounding box size and 3D attributes of objects. We provide illustrations from bird-eye-view (BEV) as shown in Fig. 3 and Fig. 4 to explain the connection among depth, 2D bounding box size, and 3D attributes. Firstly, Fig.3(a) visualizes the geometric relationship across geometric depth $\mathbf{d}$, the width $\mathbf{V}_{img}$ of image field-of-view (FoV), and the spread angle Φ of image FoV, where

$$\tan(\frac{\Phi}{2}) = \frac{\mathbf{V}_{img}}{2 \cdot \mathbf{d}} . \tag{3}$$

Moreover, by denoting the width of the object in FoV as $\mathbf{V}_{obj}$ (i.e. observed width along the horizontal direction), we assume (without loss of generality) the ratio between $\mathbf{V}_{obj}$ and $\mathbf{V}_{img}$ equals to the one between their corresponding representations in terms of pixels (i.e. the width w_{2d} of the object 2D bounding box and the image width W, respectively), in which

$$\frac{\mathbf{V}_{obj}}{\mathbf{V}_{img}} = \frac{w_{2d}}{W} , \quad \text{then} \quad \mathbf{V}_{img} = \frac{W}{w_{2d}} \cdot \mathbf{V}_{obj} . \tag{4}$$

With substituting $\mathbf{V}_{img}$ in Eq. 3 by the one in Eq. 4, the geometric depth $\mathbf{d}$ now can be represented as

$$\mathbf{d} = \frac{W}{w_{2d}} \cdot \frac{\mathbf{V}_{obj}}{2 \cdot \tan(\frac{\Phi}{2})} . \tag{5}$$

Furthermore, as illustrated by Fig. 3(b), when the object is positioned right ahead of the camera (i.e. in the center of camera's FoV), its $\mathbf{V}_{obj}$ is fully determined by the object

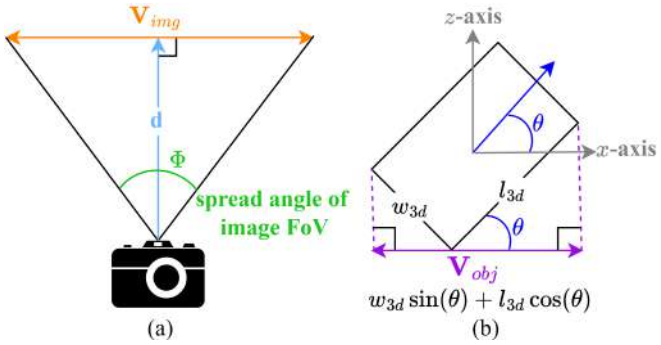

Fig. 3. From the perspective of bird-eye-view, (a) illustrates the geometric connections among geometric depth $\mathbf{d}$, the width of image FoV, $\mathbf{V}_{img}$ and the spread angle of image FoV Φ; while (b) starts from the assumption that object is located in the center of camera's FoV, and links the object's observed width $\mathbf{V}_{obj}$ to its 3D dimensions (i.e. w_{3d} and l_{3d}) and the orientation θ.

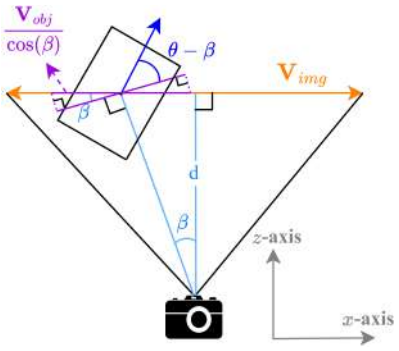

Fig. 4. Extended from Fig. 3(b) with its releasing the assumption upon object's position (i.e. now objects can appear at any position in camera's FoV) and taking the angle β (between camera's optical axis and the ray connecting camera to the object center) into consideration, the **geometric depth** formulated as Eq. 7 enables our geometric awareness loss.

width w_{3d}, object length l_{3d}, and the object orientation θ, that is: $\mathbf{V}_{obj} = w_{3d}\sin(\theta) + l_{3d}\cos(\theta)$.

Nevertheless, as objects could appear at any position in camera's FoV (i.e. not being right ahead of the camera), the formulation among $\mathbf{V}_{obj}$, w_{3d}, l_{3d}, and θ should take the angle β, which is between the optical axis of camera and the ray connecting camera to the object center (as illustrated in Fig. 4), into consideration, where in results we have:

$$\mathbf{V}_{obj} = \frac{w_{3d}\sin(\theta-\beta) + l_{3d}\cos(\theta-\beta)}{\cos(\beta)} \tag{6}$$

With combining Eq. 5 and Eq. 6, we derive geometric depth $\mathbf{d}$ based on the aforementioned geometric relations:

$$\mathbf{d} = \frac{W}{w_{2d}} \cdot \frac{w_{3d}\sin(\theta-\beta) + l_{3d}\cos(\theta-\beta)}{2\cdot\tan(\frac{\Phi}{2})\cdot\cos(\beta)} \tag{7}$$

Finally, given a 3D detection $\mathbf{B}_{3d}$, we denote the geometric depth computed by using its groundtruth 3D attributes as $\mathbf{d}^{gt}$ and the one computed by the predicted 3D attributes (i.e. produced by the lifting model) as $\tilde{\mathbf{d}}$, our geometric depth loss $\mathcal{L}_{geo}$ is then defined as:

$$\mathcal{L}_{geo} = \left|\tilde{\mathbf{d}} - \mathbf{d}^{gt}\right| \tag{8}$$

in which its minimization contributes to the update upon 3D dimensions (i.e. w_{3d} and l_{3d}) and the orientation θ.

Opposite Bin Loss $\mathcal{L}_{oppo}$. The object's orientation is formulated into two sub-tasks, heading classification and residual regression, it is a common practice nowadays to estimate the object's heading via introducing the multi-bin loss [19], [30], where we define N equally split angle bins and the model learns to predict the posterior for each bin to realize the orientation estimation (i.e. the bin with the highest posterior provides a specific range of object's orientation). We then empirically discover that, the confusion in estimation frequently happens between the diametrically opposite bins (i.e. between the bin and its 180°-rotated counterpart), indicating the model's difficulty in distinguishing between vehicle object's front and rear features due to their high similarity. For alleviating such confusion, in addition to the original multi-bin objective applied upon the orientation estimation (which attempts to maximize the posterior of the groundtruth bin), we propose the opposite bin loss $\mathcal{L}_{oppo}$ which explicitly aims to minimize the posterior of the diametrically opposite bins with respect to the groundtruth ones. In details, with denoting the array composed of the posterior of all orientation bins as $\mathbf{P}$, the posterior for the bin corresponding to the groundtruth orientation as $\mathbf{p}^{gt}$, and the posterior for the diametrically opposite bin as $\mathbf{p}^{oppo}$, our opposite bin loss is defined as:

$$\mathcal{L}_{oppo} = (1 - \frac{\mathbf{p}^{gt} - \mathbf{p}^{oppo}}{\max(\mathbf{P}) - \min(\mathbf{P})})^2 \tag{9}$$

where $\max(\mathbf{P}) - \min(\mathbf{P})$ helps to normalize the difference between $\mathbf{p}^{gt}$ and $\mathbf{p}^{oppo}$ thus leading to more numerically stable optimization. $\mathcal{L}_{oppo}$ directly contributes to update θ.

IV. Experiment

To validate the adaptability of the proposed methods, we select two baseline models with distinct architectures, GUPNet [2] and Deep3DBox [1], and incorporate proposed geometry-driven objectives into both frameworks. This allows for a comprehensive evaluation of the methods' robustness across varying network designs. Evaluated across multiple datasets, proposed geometry-driven objectives demonstrate significant effectiveness in enhancing feature extraction and improving regression accuracy for 3D object detection.

A. Baseline Models

GUPNet [2] processes full images at a resolution of 1280×384 and employs DLA-34 [31] as the feature extractor. After generating 2D bounding boxes in intermediate stage, the model performs ROI alignment on the feature maps, followed by multi-head regression to estimate 3D attributes.

Deep3DBox [1] accepts cropped images based on provided 2D bounding box coordinates and resizes them to 224×224 for input. The model employs VGG [32] backbone as the feature extractor, followed by fully connected layers to predict the orientation and dimensions of objects. Subsequently, 3D center locations are determined by geometric constraints between the 2D and 3D bounding boxes.

B. Implementation Details

Our experiments are performed on two public datasets, KITTI and nuScenes-mini, and one in-house dataset provided by ELAN Microelectronics Corporation. ELAN dataset consists of 4182 RCCC images, which are sensitive to red lights, and the color representation significantly deviates from traditional RGB images.

We incorporate the proposed geometry-driven objectives into GUPNet [2] and Deep3DBox [1], training both models on the KITTI [3] *train* set and evaluating them on the KITTI *val.* set. Additionally, to assess the adaptability of the proposed methods across different scenarios, we further evaluate the models trained on KITTI directly on the nuScenes-mini [4] dataset. This cross-dataset evaluation provides insights into the generalization capability of our approach in diverse environments.

For the ELAN dataset, we partition the data into 80% for training and 20% for validation and then retrain both baseline models from scratch on the RCCC images to evaluate the effectiveness of the geometry-driven objectives across various images.

C. Enhancement on baseline models

Our work focuses on transforming each 2D bounding box into a corresponding 3D bounding box, evaluating the precision of key 3D object attributes such as orientation, depth, and dimensions. To assess the yaw angle estimation, we compute the cosine distance, defined as one minus the cosine similarity. For depth and dimensions, we calculate the mean absolute differences between the predicted values and the ground truth for the car category, providing a detailed evaluation of the model's accuracy in predicting 3D attributes.

Tab. I highlights the performance enhancements of GUPNet and Deep3DBox across different datasets. For the car category on the KITTI *val.* set, improvements are observed across all metrics, particularly in depth and orientation estimation, illustrating the superior capability of our method in transforming/lifting 2D detection into accurate 3D bounding boxes. On the nuScenes-mini dataset, our method further demonstrates its robustness in lifting 2D objects to 3D bounding boxes across diverse scenes. Notably, these models are trained on the KITTI dataset and evaluated directly on nuScenes-mini without any fine-tuning. For the ELAN dataset of RCCC images, the results demonstrate that our proposed geometry-driven objectives are closely tied to the inherent geometric properties of objects, independent of the image format in which they were captured.

D. Ablation Study

To evaluate how the proposed geometric objectives enhance regression performance, we conducted ablation studies on GUPNet by incrementally adding the three proposed objectives. (We keep the original losses of GUPNet during training.) The results, presented in Tab. II, highlight the impact of each objective on improving the model's performance.

TABLE I

PERFORMANCE ENHANCEMENT FOR GUPNET AND DEEP3DBOX ACROSS DIFFERENT DATASETS, AND IMPROVED VALUES ARE IN **BOLD**.

Dataset	Model	$\lvert\Delta z\rvert$ (↓)	$\lvert\cos(\Delta_\theta)\rvert$ (↓)	$\lvert\Delta h_{3d}\rvert$ (↓)	$\lvert\Delta w_{3d}\rvert$ (↓)	$\lvert\Delta l_{3d}\rvert$ (↓)
KITTI	GUPNet	0.973	0.1229	0.0756	0.0698	0.3129
	GUPNet+ours	**0.856**	**0.0883**	**0.0675**	**0.0695**	**0.3078**
	Deep3DBox	2.518	0.0635	0.0962	0.0733	0.2934
	Deep3DBox+ours	**2.330**	**0.0462**	**0.0894**	**0.0722**	**0.2874**
nuScenes-mini	GUPNet	1.591	0.1423	0.138	0.1117	0.325
	GUPNet+ours	**1.510**	**0.0791**	**0.1373**	**0.1084**	**0.2746**
	Deep3DBox	6.618	0.8321	0.1282	0.1840	0.2858
	Deep3DBox+ours	**6.541**	**0.7858**	**0.1276**	0.1844	**0.2844**
ELAN	GUPNet	1.1166	0.0440	0.0586	0.0968	0.2728
	GUPNet+ours	**0.7765**	**0.0255**	**0.0385**	**0.0836**	**0.2561**
	Deep3DBox	3.195	0.1110	0.1134	0.1651	0.3938
	Deep3DBox+ours	**2.899**	**0.0752**	**0.1095**	**0.1622**	**0.3728**

In Tab. II (b), the loss term $\mathcal{L}geo$ significantly improves all metrics by enabling the model to better capture geometric attributes. Although both $\mathcal{L}proj$ and $\mathcal{L}geo$ involve the complete set of 3D object attributes, potentially complicating model convergence during training, $\mathcal{L}proj$ still enhances depth and object height estimation. Additionally, the third loss, $\mathcal{L}_{oppo}$, contributes to refining the orientation and works synergistically with the other two geometry-driven objectives. As a result, the model achieves top performance across four metrics and secures second place in object length estimation, as shown in Tab. II (d).

TABLE II

ABLATION STUDY OF THE PROPOSED OBJECTIVES ON GUPNET.

	$\mathcal{L}_{geo}$	$\mathcal{L}_{proj}$	$\mathcal{L}_{oppo}$	$\lvert\Delta z\rvert$ (↓)	$\lvert\cos(\Delta_\theta)\rvert$ (↓)	$\lvert\Delta h_{3d}\rvert$ (↓)	$\lvert\Delta w_{3d}\rvert$ (↓)	$\lvert\Delta l_{3d}\rvert$ (↓)
(a)	×	×	×	1.0121	0.1942	0.8230	0.0743	0.3250
(b)	✓	×	×	0.9913	<u>0.1757</u>	0.0717	<u>0.0735</u>	**0.3112**
(c)	✓	✓	×	<u>0.9850</u>	0.1766	<u>0.0701</u>	0.0739	0.3180
(d)	✓	✓	✓	**0.9849**	**0.1602**	**0.0684**	**0.0706**	<u>0.3122</u>

E. Quantitative Results

Besides the enhancement of lifting ability, we evaluate the performance in AP_{3D}, AP_{BEV} metrics on the KITTI *val.* set car category. Here, we integrate our proposed geometry-driven objectives on GUPNet [2], DID-M3D [25], and MonoDistill [39]. As shown in Table III, the improvement on each model states the precision of constructed 3D bounding boxes is better. In conclusion, our proposed geometry-driven objectives not only benefit the 2D-to-3D lifting models but also enhance the performance of monocular 3D object detectors.

F. Qualitative Results

Fig. 5 shows the qualitative results of our methods. The left column illustrates the results detected by the baseline GUPNet model, whereas the right column presents the results of our proposed method, lifting 2D detection into 3D. The blue boxes represent the groundtruth, the orange boxes represent the results predicted by the model, and the red circles highlight the objects that are not detected by the baseline model. Because the 3D monocular object detection models may cause more miss detection on objects than 2D, our proposed framework leverages mature 2D object detectors and learns to transform 2D detection into 3D attributes. Aided by geometry-driven objectives, we can obtain more accurate results for 3D object bounding boxes.

TABLE III

QUANTITATIVE RESULTS ON KITTI *val. Car*. THE RED INDICATES THE IMPROVEMENT, AND * MEANS REPRODUCED MODELS.

Method	Venues	AP_{3D} @ IoU=0.7			AP_{BEV} @ IoU=0.7		
		Easy	**Moderate**	**Hard**	**Easy**	**Moderate**	**Hard**
M3D-RPN [33]	ICCV2019	14.53	11.07	8.65	20.85	15.62	11.88
SMOKE [18]	CVPRW2020	14.76	12.85	11.50	19.99	15.61	15.28
MonoPair [28]	CVPR2020	16.28	12.30	10.42	24.12	18.17	15.76
MonoDLE [19]	CVPR2021	17.45	13.66	11.68	24.97	19.33	17.01
MonoRUn [34]	CVPR2021	17.26	12.27	10.41	-	-	-
GrooMeD-NMS [35]	CVPR2021	19.67	14.32	11.27	27.38	19.75	15.92
MonoFlex [21]	CVPR2021	23.64	17.51	14.83	28.23	19.75	16.89
DCD [36]	ECCV2022	23.81	15.90	13.21	32.55	21.50	18.25
MoGDE [26]	NeurIPS2022	23.35	20.35	17.71	-	-	-
MonoCon [22]	AAAI2022	26.33	19.01	15.98	-	-	-
MonoDDE [24]	CVPR2022	26.66	19.75	16.72	35.51	26.48	23.07
MonoDTR [37]	CVPR2022	24.52	18.57	15.51	33.33	25.35	21.68
MonoGround [23]	CVPR2022	25.24	18.69	15.58	32.68	24.79	20.56
MonoNeRD[38]	ICCV2023	22.75	17.13	15.63	31.13	23.46	20.97
MonoPGC [27]	ICRA2023	25.67	18.63	15.65	34.06	24.26	20.78
DID-M3D [25]	ECCV2022	29.16	21.92	18.57	37.22	26.25	24.37
DID-M3D+ours	-	29.71	22.28	18.83	37.13	28.86	24.63
Improvement	-	+0.55	+0.36	+0.26	-0.09	+2.61	+0.26
MonoDistill [39]	ICLR2022	25.74	21.70	20.11	36.87	29.66	25.59
MonoDistill+ours	-	26.78	21.76	20.57	37.16	29.99	25.98
Improvement	-	+1.04	+0.06	+0.46	+0.29	+0.33	+0.39
GUPNet* [2]	ICCV2021	27.31	21.63	18.53	34.30	25.48	24.10
GUPNet*+ours	-	27.98	22.11	18.96	36.27	26.52	25.21
Improvement	-	+0.67	+0.48	+0.43	+1.93	+1.04	+1.11

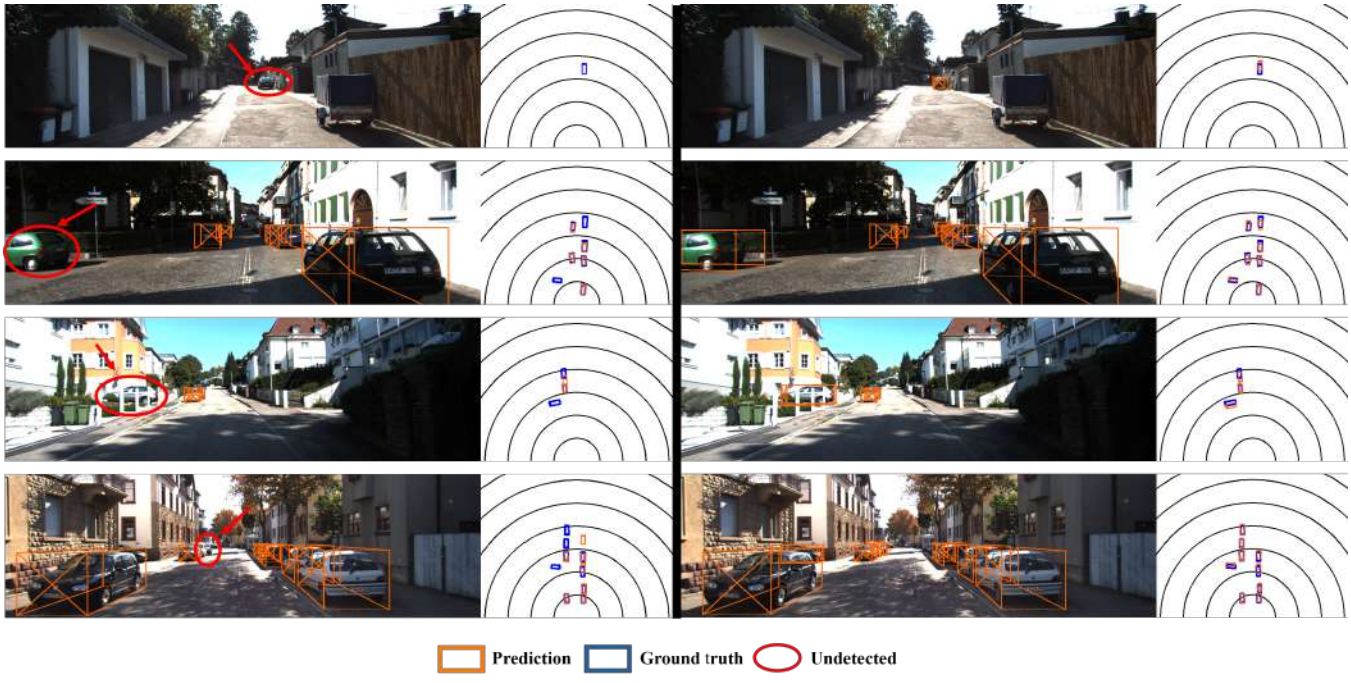

Fig. 5. **Qualitative results of comparison between GUPNet and GUPNet+ours method on KITTI *val* set of Car category.** The left column shows the baseline model results with its corresponding BEV results. The right column presents the results of our proposed method. Blue boxes represent the groundtruths; orange boxes indicate the predicted results, and the red circles highlight the objects not detected by the baseline model.

V. CONCLUSION

In this work, we proposed a framework to transform/lift the bounding boxes from 2D object detection into their 3D counterparts. Specifically, we leverage the geometric relationships of objects and introduce geometry-driven objectives to significantly improve the estimation of 3D object attributes. Our experiments demonstrate that the proposed method can be seamlessly integrated with various model architectures, exhibiting strong generalizability across diverse scenes and datasets.

REFERENCES

[1] A. Mousavian, D. Anguelov, J. Flynn, and J. Kosecka, "3d bounding box estimation using deep learning and geometry," in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2017, pp. 7074–7082.

[2] Y. Lu, X. Ma, L. Yang, T. Zhang, Y. Liu, Q. Chu, J. Yan, and W. Ouyang, "Geometry uncertainty projection network for monocular 3d object detection," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 3111–3121.

[3] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," The International Journal of Robotics Research, vol. 32, no. 11, pp. 1231–1237, 2013.

[4] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 11 621–11 631.

[5] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," 2014. [Online]. Available: https://arxiv.org/abs/1311.2524

[6] R. Girshick, "Fast r-cnn," in Proceedings of the IEEE international conference on computer vision, 2015, pp. 1440–1448.

[7] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," 2016.

[8] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2961–2969.

[9] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," 2020.

[10] C. Li, L. Li, H. Jiang, K. Weng, Y. Geng, L. Li, Z. Ke, Q. Li, M. Cheng, W. Nie, Y. Li, B. Zhang, Y. Liang, L. Zhou, X. Xu, X. Chu, X. Wei, and X. Wei, "Yolov6: A single-stage object detection framework for industrial applications," 2022.

[11] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," pp. 7464–7475, 2023.

[12] D. Reis, J. Kupec, J. Hong, and A. Daoudi, "Real-time flying object detection with yolov8," 2023.

[13] J. Terven, D.-M. Córdova-Esparza, and J.-A. Romero-González, "A comprehensive review of yolo architectures in computer vision: From yolov1 to yolov8 and yolo-nas," Machine Learning and Knowledge Extraction, vol. 5, no. 4, pp. 1680–1716, 2023.

[14] W. Murray, "The sensitivity of the higgs boson branching ratios to the w boson width," Physics Letters B, vol. 758, p. 98–100, July 2016. [Online]. Available: http://dx.doi.org/10.1016/j.physletb.2016.04.056

[15] F. Chabot, M. Chaouch, J. Rabarisoa, C. Teuliere, and T. 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, 2017, pp. 2040–2049.

[16] P. Li, H. Zhao, P. Liu, and F. Cao, "Rtm3d: Real-time monocular 3d detection from object keypoints for autonomous driving," in European Conference on Computer Vision. Springer, 2020, pp. 644–660.

[17] G. Wang, J. Wu, B. Tian, S. Teng, L. Chen, and D. Cao, "Centernet3d: An anchor free object detector for point cloud," IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 8, pp. 12 953–12 965, 2021.

[18] Z. Liu, Z. Wu, and R. 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, 2020, pp. 996–997.

[19] 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 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 4721–4730.

[20] A. Simonelli, S. R. R. Bulò, L. Porzi, M. López-Antequera, and P. Kontschieder, "Disentangling monocular 3d object detection," 2019. [Online]. Available: https://arxiv.org/abs/1905.12365

[21] Y. Zhang, J. Lu, and J. Zhou, "Objects are different: Flexible monocular 3d object detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 3289–3298.

[22] X. Liu, N. Xue, and T. Wu, "Learning auxiliary monocular contexts helps monocular 3d object detection," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, no. 2, 2022, pp. 1810–1818.

[23] Z. Qin and X. Li, "Monoground: Detecting monocular 3d objects from the ground," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 3793–3802.

[24] Z. Li, Z. Qu, Y. Zhou, J. Liu, H. Wang, and L. Jiang, "Diversity matters: Fully exploiting depth clues for reliable monocular 3d object detection," 2022. [Online]. Available: https://arxiv.org/abs/2205.09373

[25] L. Peng, X. Wu, Z. Yang, H. Liu, and D. Cai, "Did-m3d: Decoupling instance depth for monocular 3d object detection," in European Conference on Computer Vision. Springer, 2022, pp. 71–88.

[26] Y. Zhou, Q. Liu, H. Zhu, Y. Li, S. Chang, and M. Guo, "Mogde: Boosting mobile monocular 3d object detection with ground depth estimation," Advances in Neural Information Processing Systems, vol. 35, pp. 2033–2045, 2022.

[27] Z. Wu, Y. Gan, L. Wang, G. Chen, and J. Pu, "Monopgc: Monocular 3d object detection with pixel geometry contexts," 2023. [Online]. Available: https://arxiv.org/abs/2302.10549

[28] Y. Chen, L. Tai, K. Sun, and M. Li, "Monopair: Monocular 3d object detection using pairwise spatial relationships," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 12 093–12 102.

[29] J. Gu, B. Wu, L. Fan, J. Huang, S. Cao, Z. Xiang, and X.-S. Hua, "Homography loss for monocular 3d object detection," 2022. [Online]. Available: https://arxiv.org/abs/2204.00754

[30] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, "Frustum pointnets for 3d object detection from rgb-d data," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 918–927.

[31] F. Yu, D. Wang, E. Shelhamer, and T. Darrell, "Deep layer aggregation," 2019.

[32] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015.

[33] G. Brazil and X. Liu, "M3d-rpn: Monocular 3d region proposal network for object detection," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 9287–9296.

[34] H. Chen, Y. Huang, W. Tian, Z. Gao, and L. Xiong, "Monorun: Monocular 3d object detection by reconstruction and uncertainty propagation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 10 379–10 388.

[35] A. Kumar, G. Brazil, and X. Liu, "Groomed-nms: Grouped mathematically differentiable nms for monocular 3d object detection," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 8973–8983.

[36] Y. Li, Y. Chen, J. He, and Z. Zhang, "Densely constrained depth estimator for monocular 3d object detection," in European Conference on Computer Vision. Springer, 2022, pp. 718–734.

[37] K.-C. Huang, T.-H. Wu, H.-T. Su, and W. H. Hsu, "Monodtr: Monocular 3d object detection with depth-aware transformer," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 4012–4021.

[38] J. Xu, L. Peng, H. Cheng, H. Li, W. Qian, K. Li, W. Wang, and D. Cai, "Mononerd: Nerf-like representations for monocular 3d object detection," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 6814–6824.

[39] Z. Chong, X. Ma, H. Zhang, Y. Yue, H. Li, Z. Wang, and W. Ouyang, "Monodistill: Learning spatial features for monocular 3d object detection," arXiv preprint arXiv:2201.10830, 2022.