MonoCD: Monocular 3D Object Detection with Complementary Depths

Longfei Yan[1] Pei Yan[1] Shengzhou Xiong[1] Xuanyu Xiang[1] Yihua Tan[1*]
[1]Hubei Engineering Research Center of Machine Vision and Intelligent Systems,
School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, China
{longfeiyan, yanpei}@hust.edu.cn, xiongshengzhou@126.com, {xuanyuxiang, yhtan}@hust.edu.cn

Abstract

Monocular 3D object detection has attracted widespread attention due to its potential to accurately obtain object 3D localization from a single image at a low cost. Depth estimation is an essential but challenging subtask of monocular 3D object detection due to the ill-posedness of 2D to 3D mapping. Many methods explore multiple local depth clues such as object heights and keypoints and then formulate the object depth estimation as an ensemble of multiple depth predictions to mitigate the insufficiency of single-depth information. However, the errors of existing multiple depths tend to have the same sign, which hinders them from neutralizing each other and limits the overall accuracy of combined depth. To alleviate this problem, we propose to increase ***the complementarity*** *of depths with two novel designs. First, we add a new depth prediction branch named complementary depth that utilizes global and efficient depth clues from the entire image rather than the local clues to reduce the similarity of depth predictions. Second, we propose to fully exploit the geometric relations between multiple depth clues to achieve complementarity in form. Benefiting from these designs, our method achieves higher complementarity. Experiments on the KITTI benchmark demonstrate that our method achieves state-of-the-art performance without introducing extra data. In addition, complementary depth can also be a lightweight and plug-and-play module to boost multiple existing monocular 3d object detectors. Code is available at* https://github.com/elvintanhust/MonoCD.

1. Introduction

As a significant research topic in both academia and industry, 3D object detection can empower non-human intelligences to perceive the 3D world. Compared with LiDAR-based [11, 27, 28, 34] and stereo-based [12, 13, 23, 30] approaches, monocular 3D object detection has attracted widespread attention due to its lower price and simpler configuration [15, 24]. However, its 3D localization accuracy is significantly lower than those based on LiDAR and stereo. To advance and promote automation technologies, such as autonomous driving and robotics, it is essential to enhance the 3D localization precision of monocular 3D object detection.

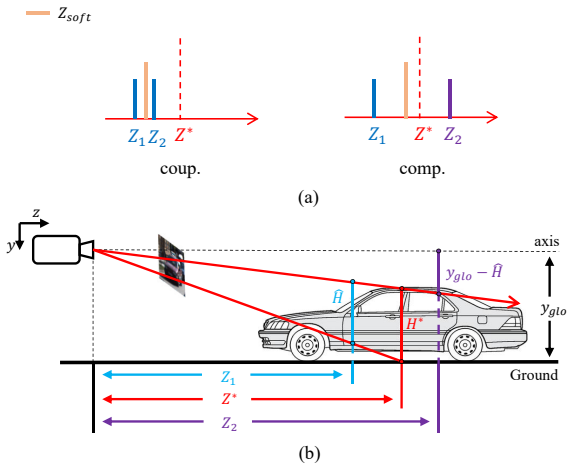

Figure 1. (a) Comparison of coupling(coup) and complementary(comp) multi-depth with two depth branches Z_1 and Z_2, where Z^* and Z_{soft} represents the ground truth of the depth and the final combined depth respectively. (b) A complementary demonstration of the two depth branches with the help of geometrical relations when considering only the inaccurate estimation of the object 3D height H. Both Z_1 generated by the widely used local height clue and Z_2 generated by our newly introduced global clue y_{glo} are related to H. H^* and $\hat{H}$ denote the ground truth of H and the underestimated H respectively.

Recently, many monocular 3D object detection algorithms have realized that the main reason limiting the 3D localization precision of monocular 3D object detection is inaccurate depth estimation [15, 22, 25, 40, 45]. Following mainstream CenterNet paradigm [42], they explore multiple local depth clues and formulate depth estimation as an ensemble of multiple depth predictions to mitigate the insufficiency of single-depth information. For instance, MonoFlex [40] explores local depth clues from direct es-

*Corresponding author.

timate and object heights, and subsequently combines them into one depth by weighted averaging. MonoDDE [15] further reveals clues from the object perspective point on top of that.

However, experiments on KITTI dataset [7] show that 95% of the existing multi-depth prediction ensembles have the same error sign, *i.e.*, multiple predicted depths are usually distributed on the same side of the ground truth as shown by the coupling in Fig. 1(a), which leads to depth errors that cannot be neutralized with each other, hindering the improvement of combined depth accuracy. We attribute this coupling phenomenon to the fact that the local depth clues they used are all derived from the same local features around the object in the CenterNet paradigm.

In this paper, we propose to increase **the complementarity** of depths to alleviate the problem. Complementarity here refers that these predictions not only aim for high accuracy but also have different error signs. To this end, we propose two novel designs. First, considering the aforementioned coupling phenomenon, we add a new depth prediction branch that utilizes global and efficient depth clues from the entire image rather than the local clues to reduce the similarity of depth predictions. It relies on the global information that all objects in one image approximately lie on the same plane. Second, to further improve complementarity, we propose to fully exploit the geometric relations between multiple depth clues to achieve complementarity in form, which utilizes the fact that errors in the same geometric quantity may have opposite effects on different branches. For example, in Fig. 1(b), Z_1 has a negative error because the related clue 3D height H is underestimated, whereas in this case, Z_2 has a positive error because the effect of H on Z_2 combined with new clues y_{glo} is opposite to Z_1. Therefore, the geometric relation based on H provides complementarity to Z_1 and Z_2 in form.

Incorporating all the designs, we propose a novel monocular 3D detector with complementary depths, named MonoCD, which compensates for the complementarity neglected in previous multi-depth predictions. The main contributions of this paper are summarized as follows:

- We point out the coupling of existing monocular object depth predictions, which limits the accuracy of the combined depths. Therefore we propose to improve the depths complementarity to alleviate this problem.
- We propose to add a new depth prediction branch named complementary depth that utilizes global and efficient depth clue and fully exploit the geometric relations between multiple depth clues to achieve complementarity in form.
- Evaluated on KITTI benchmark, our method achieves state-of-the-art performance without introducing extra data. Moreover, complementary depth can be a lightweight plug-and-play module to boost multiple existing detectors.

2. Related work

2.1. Center-based Monocular 3D Detector

Many recent works [5, 16, 20, 36, 41, 43] are extended from the popular center-based paradigm CenterNet [42], which is an anchor-free method initially applied to 2D object detection. It makes the detection process simpler and more efficient due to converting all attributes of a 3D bounding box into a center to estimate. SMOKE [18] inherits the center-based framework and proposes that the estimation of the 2D bounding box can be omitted. MonoDLE [21] finds that the estimation of the 2D bounding box contributes to the prediction of 3D attributes and demonstrates that depth error is the main reason limiting the accuracy of monocular 3D object detection. MonoCon [17] finds that adding auxiliary learning tasks around the center can improve the generalization performance. Although there are many benefits in the center-based framework, it makes the prediction of all 3D attributes highly correlated with the local center. It ignores the exploitation of global information, leading to the coupling of predicted 3D attributes.

2.2. Transformer-based Monocular 3D Detector

Benefiting from the non-local encoding of attention mechanism [32] and its development in object detection [2], multiple Transformer-based monocular 3D detectors have recently been proposed to enhance the global perception capability. MonoDTR [8] proposes to perform depth position encoding to inject global depth information into Transformer to guide the detection, which requires LIDAR for auxiliary supervision. Different from it, MonoDETR [39] uses foreground object labels to predict foreground Depth Maps to achieve depth guidance. In order to improve the inference efficiency, MonoATT [44] proposes an adaptive token Transformer and makes it possible for finer tokens to be assigned to more significant regions in images. Although the above methods perform well, the drawbacks of high computational complexity and slow inference of Transformer-based monocular 3D detectors are still apparent. Thus there is currently a lack of a method that has both the capability of synthesizing global information and low latency in real-world autonomous driving scenarios.

2.3. Estimation of Multi-Depth

In addition to directly estimating object depth using deep neural networks, many recent works have broadened the depth estimation branch by mediately predicting geometric clues associated with depth. [20, 29] utilizes mathematical priors and uncertainty modeling to restore depth information through the ratio of 3D to 2D height. Based on them, MonoFlex [40] further extends the geometric depths to three

sets by other supporting lines of the 3D bounding box and proposes to use uncertainties as weights to combine multiple depths into a final depth. MonoGround [25] introduces a local ground plane prior and enriches the depth supervision sources using randomly sampling dense points in the bottom plane of each object. MonoDDE [15] utilizes keypoint information to expand the number of depth prediction branches to 20, highlighting the importance of depth diversity. However, the complementarity between multiple depths is hardly explored. Errors in geometric clues (such as 2D/3D height) accumulate into the corresponding depth errors. Without effective complementarity, existing depth errors cannot be neutralized.

3. Approach

3.1. Problem Definition

The task of monocular 3D object detection is to recognize objects of interest from a 2D image only and predict their corresponding 3D attributes including 3D location (x, y, z), dimension (h, w, l), and orientation θ. The 3D location (x, y, z) is usually transformed into 2.5D information (u_c, v_c, z) for prediction. The recovery process of x and y can be formulated as:

$$x = \frac{(u_c - c_u)z}{f_x}, \quad y = \frac{(v_c - c_v)z}{f_y} \tag{1}$$

where (u_c, v_c) is the projected 3D center in the image and (c_u, c_v) is the camera optical center. f_x and f_y denote the horizontal and vertical focal lengths respectively.

As described in Sec. 1, many methods [15, 25, 40] have realized that depth z is the main reason limiting the performance of monocular 3D detector and utilize multi-depth to improve the accuracy of depth prediction via:

$$z_{soft} = \sum_{i=1}^{n} w_i z_i \tag{2}$$

where $\{z_i\}_{i=1}^n$ represents n predicted depths and $\{w_i\}_{i=1}^n$ represents their weights determined by the predicted uncertainty [9, 10]. z_{soft} is used as the final depth of the output.

3.2. The Effect of Complementary Depths

To demonstrate the effectiveness of complementary depths, we present its superiority from a mathematical perspective. Define two different depth prediction branches $\hat{z}_1$ and $\hat{z}_2$ as follows:

$$\hat{z}_1 = z^* + e_1, \quad \hat{z}_2 = z^* + e_2 \tag{3}$$

where z^* represents the ground truth of depth. e_1 and e_2 are the errors of the two depth branches in a single prediction, respectively. Note that the positive and negative of e_1 and e_2 correspond to the sign of error. We define $e_1 e_2 > 0$ to simulate the case of multiple depth coupling, as shown in Fig. 1(a). We term the final combination error of multiple coupling depths as coupling depth error. Hence, referring to Eq. (2), **the coupling depth error** E_1 of $\hat{z}_1$ and $\hat{z}_2$ can be formulated as:

$$\begin{aligned} E_1 &= |w_1 \hat{z}_1 + w_2 \hat{z}_2 - z^*| \\ &= |w_1 e_1 + w_2 e_2| \end{aligned} \tag{4}$$

where w_1 and w_2 satisfy $w_1, w_2 > 0$ and $w_1 + w_2 = 1$. We then flip $\hat{z}_1$ symmetrically along z^* without changing the accuracy of the prediction through:

$$\begin{aligned} \hat{z}_1' &= z^* - (\hat{z}_1 - z^*) \\ &= z^* - e_1 \end{aligned} \tag{5}$$

After flipping, the error sign in $\hat{z}_1'$ and $\hat{z}_2$ are opposite and higher complementarity between them is artificially achieved. We term the final combination error of multiple complementary depths as complementary depth error. Similarly, **the complementary depth error** E_2 of $\hat{z}_1'$ and $\hat{z}_2$ can be formulated as:

$$\begin{aligned} E_2 &= |w_1 \hat{z}_1' + w_2 \hat{z}_2 - z^*| \\ &= |w_1 e_1 - w_2 e_2| \end{aligned} \tag{6}$$

By mathematical transformations we further express Eqs. (4) and (6) as:

$$\begin{aligned} E_1 &= \sqrt{(w_1 e_1 + w_2 e_2)^2} \\ &= \sqrt{(w_1 e_1)^2 + 2 w_1 w_2 e_1 e_2 + (w_2 e_2)^2} \end{aligned} \tag{7}$$

$$\begin{aligned} E_2 &= \sqrt{(w_1 e_1 - w_2 e_2)^2} \\ &= \sqrt{(w_1 e_1)^2 - 2 w_1 w_2 e_1 e_2 + (w_2 e_2)^2} \end{aligned} \tag{8}$$

It is obvious that the complementary depth error E_2 is consistently less than the coupling depth error E_1 due to the condition $e_1 e_2 > 0$. Regardless of variations in weight or error magnitude, this relationship remains constant. Similarly, the conclusion is equivalent by maintaining z_1 unchanged during the flip of z_2. Therefore we can draw **the conclusion**: realizing the complementary relationship between two depth branches contributes to reducing the overall depth error, even without improving the accuracy of individual branches.

To demonstrate the effectiveness of complementary depths in practice, we select a classical multi-depth prediction baseline [40] for evaluation in KITTI val set. It contains 4 depth prediction branches (1 directly estimated depth and 3 geometric depths) and the coupling rate of any two branches is around 95% after testing. As shown on the left in Fig. 3, we flip the direct depth estimation branch among them symmetrically along the ground truth based on

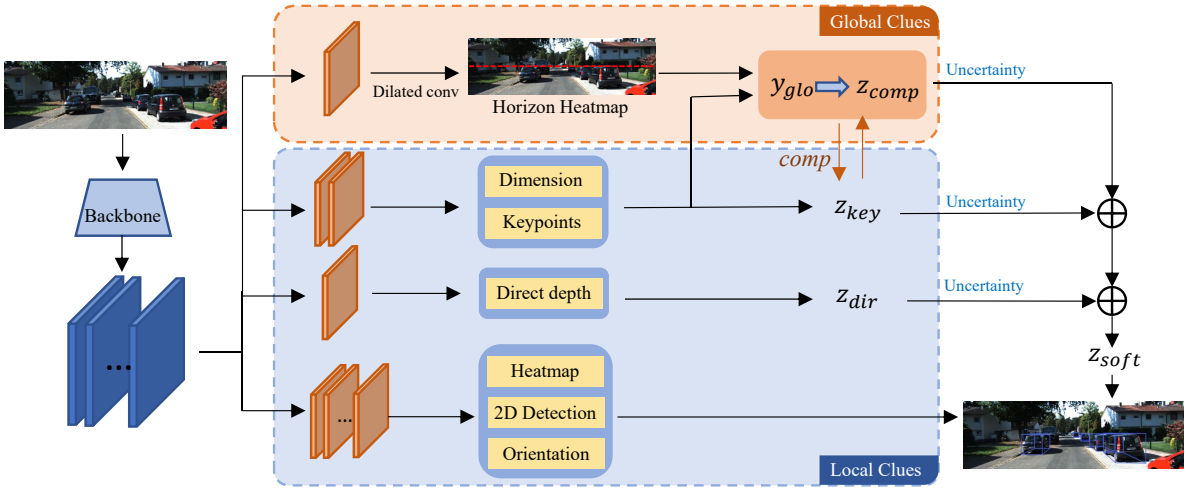

Figure 2. Overview of the approach. The input image is first subjected to processing by a feature extraction network and subsequently directed into multiple prediction heads. The prediction heads are divided into two parts. The upper orange section is used to predict the global horizon heatmap of the image, serving as a global clue to generate the prediction of complementary depths (z_{comp}). The lower blue section, after predicting local information for each point of interest, further generates keypoint depths (z_{key}) and direct depth (z_{dir}). Finally, the three depth prediction branches are weighted and combined using simultaneously predicted uncertainties to obtain the final depth estimation.

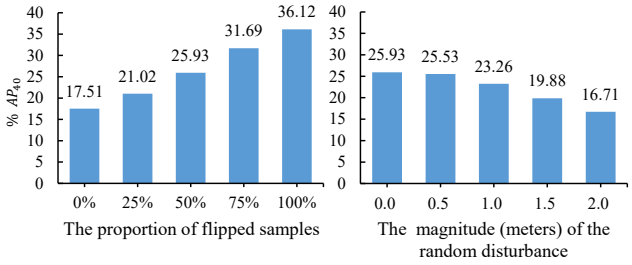

Figure 3. Evaluation of complementary effect on the KITTI *validation* set. The metric is AP_{40} for the moderate Car category at 0.7 IoU threshold. *Left*: Different proportions of flipped samples achieve different levels of complementarity. *Right*: Fixing the proportion of flipped samples to 50% and applying random disturbances of different magnitudes to the flipped depth branch.

Eq. (5) across a 0% to 100% sample scale to achieve depths complementary at different levels. Additionally, considering the difficulty of obtaining depth predictions with opposite error signs while maintaining the same accuracy in practice, we conduct another experiment by flipping the depth branch while applying random disturbances of different magnitudes on top of it. The results are presented on the right of Fig. 3. Similar results are observed in other branches by performing the same operation as above. Based on this, we have the following three observations:

Observation 1: On the left of Fig. 3, the detection accuracy increases as the proportion of flipped samples rises. It demonstrates that increasing complementarity between multiple depth prediction branches can improve detection accuracy continuously.

Observation 2: For two independent depth prediction branches, ideally, the proportion of their predictions with opposite signs in all samples should be 50%. The situation is similar to the 50% flipped proportion on the left of Fig. 3 due to the coupling of multiple branches in the baseline. Therefore reducing the similarity of multiple depth prediction branches can also increase their complementarity.

Observation 3: In the case where the flipped proportion is fixed at 50%, as shown in the right of Fig. 3, it is not until the application of random disturbance with an amplitude of 2 meters (which is quite significant [21] for Car in KITTI) that the complementary effect disappeared. This indicates that complementary effect can still contribute to overall performance even if losing some depth estimation accuracy and ultimately whether the overall performance can be improved depends on both the proportion of opposite signs and the depth estimation accuracy.

Additionally, we select models with different total numbers of depth prediction branches to perform flipping and evaluation. We find that as the number of flipped branches approaches the number of unflipped branches, the overall performance improves accordingly. For more experiments and details, please refer to the supplementary materials.

3.3. 3D Detector with Complementary Depths

Framework Overview. As shown in Fig. 2, the network we design extends from CenterNet [42]. The regression heads are divided into two parts: local clues and global clues, where DLA-34 [38] is chosen as the backbone of the

network. The branch of local clues is designed with reference to MonoFlex [40], which estimates dimension, keypoints, direct depth, orientation, and 2D detection for each local peak point based on the predicted Heatmap. Since the prediction of these geometric quantities is highly correlated with the position of the local peak point in the image, they are referred to as local clues. Both z_{dir} and z_{key} are derived from them. The branch of global clues predicts the Horizon Heatmap of the entire image based on all extracted pixel features, which is used to obtain the trend of y_{glo} in scenes, and then outputs the complementary depth z_{comp} embedding the global clues. How to construct a depth prediction branch with the global clues and further achieve complementarity in form will be elaborated below. Following [9, 10], we model uncertainty for all seven depth predictions (1 direct depth, 3 keypoint depths, and 3 complementary depths augmented by diagonal columns as [40]). The final depth is obtained according to Eq. (2), with $w_i = \frac{1}{\sigma_i}$.

Depth Prediction with Global Clues. Inspired by [6], the neural network sees depth from a single image through:

$$z = \frac{f_y y}{v_b - c_v} \tag{9}$$

where y denotes the y-axis coordinates of the object in the camera coordinate system, and v_b denotes the vertical coordinate of the projected bottom center in the pixel coordinate system. Considering that y also represents the elevation of the plane in which the objects are located and that all objects lie approximately in one plane, y contains such a global characteristic and can be distinguished from other depth clues. Unlike previous neural networks that implicitly utilize Eq. (9), we propose to predict y explicitly.

To avoid falling into the coupling, we do not utilize the center-based approach discussed in Sec. 2.1 to predict y. We propose to first obtain the sloping trend of y in the scene by the ground plane equation. The prediction of the ground plane equation is based on the Horizon Heatmap branch, similar to [35], but we omit the edge prediction and obtain prediction results as:

$$\begin{aligned} Ax + By + Cz + 1.65 = 0 \\ s.t. \quad A^2 + B^2 + C^2 = 1 \end{aligned} \tag{10}$$

where $A = F\frac{k_h f_x}{f_y}$, $B = -F$ and $C = F\frac{k_h c_u + b_h - c_v}{f_y}$. k_h and b_h represent the slope and intercept of the horizon fitted by Horizon Heatmap. After it, then considering Eq. (1) and the projected bottom center (u_b, v_b) of the object, y with global information can be derived as:

$$y_{glo} = -\frac{1.65}{An + Cm + B} \tag{11}$$

where $n = \frac{f_y(u_b - c_u)}{f_x(v_b - c_v)}$, $m = \frac{f_y}{v_b - c_v}$.

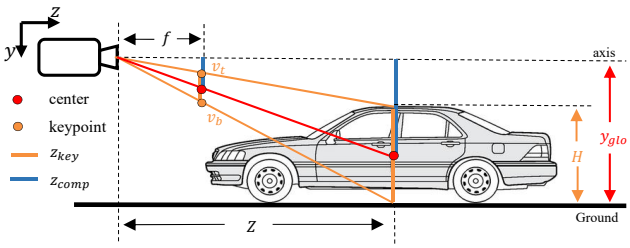

Figure 4. Geometric correspondence of different depths. To avoid overlap, the geometric correspondences of z_{key} and z_{comp} are marked with orange and blue lines, respectively.

Inserting Eq. (11) into Eq. (9), a new depth prediction branch with the global clue is obtained:

$$z_{glo} = \frac{f_y y_{glo}}{v_b - c_v} \tag{12}$$

In addition, to better utilize the global features as well as to expand the receptive field, we use dilated convolution [37] to predict the Horizon Heatmap.

Complementary Form in Solving. Simply achieving more independent depth prediction is not enough, we hope to fully exploit the geometric relations between multiple depth prediction branches to improve complementarity further. Considering the projected bottom center (u_b, v_b) and top center (u_t, v_t), as shown in the orange part of Fig. 4, the depth derived from keypoint and height in [29] can be rewritten as:

$$z_{key} = \frac{f_y H}{v_b - v_t} \tag{13}$$

where H represents the 3D height of the object. Combining the global y_{glo} information obtained by Eq. (11) and the geometric quantities used in Eq. (13), we further propose a depth prediction that is complementary to z_{key} in form:

$$z_{comp} = \frac{f_y(y_{glo} - \frac{1}{2}H)}{\frac{1}{2}(v_b + v_t) - c_v} \tag{14}$$

The geometric correspondence is shown in the blue part of Fig. 4. It can be observed that the signs of H and v_t in the designed Eq. (14) are exactly opposite to those in Eq. (13). This means that the errors of H and v_t have opposite effects on z_{key} and z_{comp} during the prediction of 3D information for each object. Although Eq. (13) and Eq. (14) are not strictly symmetrical, this further increases the probability that the errors e_{key} and e_{comp} of z_{key} and z_{comp} satisfy the condition of $e_{key}e_{comp} < 0$. As proved by Sec. 3.2, eventually a part of the depth error is neutralized in the weighted averaging of Eq. (2).

Methods, Venues	Extra data	Test, AP_{3D}			Test, AP_{BEV}			Time(ms)
		Eazy	Mod.	Hard	Eazy	Mod.	Hard	
DDMP-3D [33], CVPR2021	Depth	19.71	12.78	9.80	28.08	17.89	13.44	180
Kinematic3D [1], ECCV2020	Video	19.07	12.72	9.17	26.69	17.52	13.10	120
AutoShape [19], ICCV2021	CAD	22.47	14.17	11.36	30.66	20.08	15.59	50
DCD [14], ECCV2022		23.81	15.90	13.21	32.55	21.50	18.25	-
MonoRUn [3], CVPR2021	LiDAR	19.65	12.30	10.58	27.94	17.34	15.24	70
CaDDN [26], CVPR2021		19.17	13.41	11.46	27.94	18.91	17.19	630
MonoDTR [8], CVPR2022		21.99	15.39	12.73	28.59	20.38	17.14	37
SMOKE [18], CVPRW2020	None	14.03	9.76	7.84	20.83	14.49	12.75	30
MonoDLE [21], CVPR21		17.23	12.26	10.29	24.79	18.89	16.00	40
MonoRCNN [29], ICCV2021		18.36	12.65	10.03	25.48	18.11	14.10	70
MonoFlex [40], CVPR2021		19.94	13.89	12.07	28.23	19.75	16.89	35
MonoGround [25], CVPR2022		21.37	14.36	12.62	30.07	20.47	17.74	30
GPENet [35], -		22.41	15.44	12.84	30.31	20.79	18.21	-
MonoJSG [16], CVPR2022		24.69	16.14	13.64	32.59	21.26	18.18	42
MonoCon [17], AAAI2022		22.50	16.46	<u>13.95</u>	31.12	22.10	<u>19.00</u>	25.8
MonoDETR [39], ICCV2023		<u>25.00</u>	<u>16.47</u>	13.58	**33.60**	<u>22.11</u>	18.60	43
MonoCD(Ours)	None	**25.53**	**16.59**	**14.53**	<u>33.41</u>	**22.81**	**19.57**	36
Improvement	*v.s. second-best*	+0.53	+0.12	+0.58	-0.19	+0.70	+0.57	-

Table 1. Comparison with current state-of-the-art methods on Car category on the KITTI test set. Methods are grouped according to extra data. Follow [7], the methods in each group are sorted by AP_{3D} performance in Moderate difficulty setting. We **bold** the best results and <u>underline</u> the second results.

4. Experiments

4.1. Dataset

Our experiments are conducted on the widely-adopted KITTI 3D Object [7] dataset, which contains 7481 training images and 7518 test images. Since the annotations of the test images are not publicly accessible, we follow [4] and further divide the 7481 training images into 3712 and 3769 as the training and validation sets, respectively. Each category is further refined into three difficulties: Easy, Moderate, and Hard based on 2D height, truncation, and occlusion.

4.2. Evaluation Metrics

As in previous methods, we use Average Precision AP_{3D} and AP_{BEV} as the overall evaluation metrics. Following [31], 40 recall positions are used for the above AP calculations. The IoU threshold is 0.7 for Car.

In the ablation study of Sec. 4.5, the mean absolute error (MAE) of y is introduced as a metric to evaluate the accuracy of the different y sources. In addition, to better measure the complementarity between different designs, we quantify the magnitude of complementarity as the Complementarity Score. As discussed in Sec. 3.2, both the error sign opposite proportion and depth estimation accuracy are crucial in achieving enhanced performance. Thus we formulate the **Complementarity Score**(CS) as:

$$CS = \frac{ESOP_z}{MAE_z} \tag{15}$$

where $ESOP_z$ represents depths **Error Sign Opposite Proportion** (ESOP) between global and local clue branches, and MAE_z represents the Mean Absolute Error of z_{comp}. For a baseline without z_{comp}, ESOP counts the proportion between z_{key} and z_{dir}.

4.3. Implementation Details

In order to demonstrate the effectiveness of the proposed framework, we choose three recent center-based methods with excellent performance as the baseline model, MonoFlex [40], MonoDLE [21], and MonoCon [17]. All experiments are performed on a single RTX 2080Ti GPU. The aforementioned baseline models all employ DLA-34 [38] as the feature extraction network. In the Global Clues branch, the prediction head of Horizon Heatmap contains two 3×3 conv layers with BN and ReLU (where the dilation rate is set to 2) and an output conv layer. The horizon equation is obtained by taking out all the largest elements in each column of the Horizon heatmap and fitting them. The ground truth of Horizon Heatmap is generated by fitting the scene ground plane through the bottom coordinate annotation of each object and then projecting to the 2D image plane [35], so only RGB image data and camera annotations are used throughout the training process. The radius

Method	Val, AP_{BEV}			Val, AP_{3D}		
	Eazy	Mod.	Hard	Eazy	Mod.	Hard
MonoDLE [21]	24.97	19.33	17.01	17.45	13.66	11.68
+ Ours	**26.84**	**20.86**	**17.89**	**18.60**	**15.09**	**12.86**
Improvement	+1.87	+1.53	+0.88	+1.15	+1.43	+1.18
MonoFlex [40]	30.51	23.16	19.87	23.64	17.51	15.14
+ Ours	**31.49**	**23.56**	**20.12**	**24.22**	**18.27**	**15.42**
Improvement	+0.98	+0.40	+0.25	+0.58	+0.76	+0.28
MonoCon [17]	33.36	24.39	21.03	26.33	19.01	15.98
+ Ours	**34.60**	**24.96**	**21.51**	**26.45**	**19.37**	**16.38**
Improvement	+1.24	+0.57	+0.48	+0.12	+0.36	+0.40

Table 2. In order to fully demonstrate the effectiveness of the proposed method, we extend complementary depth to three center-based monocular 3D detectors. Evaluation is performed on the KITTI val set. The increased performance is highlighted in blue.

of the Gaussian kernel used for each pixel is 2 when mapping the horizon equation into Heatmap. The z_{direct}, z_{key} and z_{comp} loss weight proportions are set to $1:0.2:0.1$. The remaining settings such as optimizer, batch sizes, image padding size, *etc*. remain consistent with the baseline.

4.4. Quantitative Results

To demonstrate the effectiveness of the proposed method, we conduct quantitative experiments on test and val sets of KITTI [7].

As shown in Tab. 1, the proposed method is compared with the state-of-the-art methods in recent years on the widely used KITTI test set. Our method achieves the best performance in the majority of metrics without using any additional data. Compared with the previous multi-depth solving method MonoFlex [40], our performance for AP_{3D}/AP_{BEV} improves by 19.44%/15.49%, respectively. The performance for AP_{3D}/AP_{BEV} improves from 15.44/20.79 to 16.59/22.81 compared to the method GPENet [35], which also incorporated the ground plane equation solution. Even when compared to the latest Transformer-based detector MonoDETR [39], we outperform it in most metrics while ensuring real-time operation.

As shown in Tab. 2, we extend the complementary depth branch to three competitive center-based monocular 3d detectors. The results of the KITTI val set demonstrate that the proposed complementary depth is flexible and achieves stable increments across multiple frameworks and metrics. It is worth noting that the boost of our design performs better on AP_{BEV} than AP_{3D} in general. We attribute this to the focus of our method on improvements in depth estimation, since AP_{BEV} is more emphasis on the accuracy of localization along the Z-axis compared to AP_{3D} [7].

4.5. Ablation Study

In this section, we select MonoFlex [40] as the baseline to discuss the impact of different designs.

Source of Depth Clue. To demonstrate the effectiveness of introducing global depth clue, we adopt different approaches to obtain depth clue y, and the results are presented in rows 2, 3, 4, and 5 of Tab. 3. By comparing the ESOP metric, it can be observed that the ESOP of 3rd, 4th, and 5th in Tab. 3 with global characteristic (*i.e*., not determined by a single object) are significantly higher than that of the baseline and using local clue branch, which demonstrates the necessity of introducing global clues and the coupling of multi-depth prediction is alleviated. In addition, it can be found that the accuracy of z_{comp} is largely related to the accuracy of y.

Setting	Val, AP_{3D}			y	z_{comp}	ESOP	CS↑
	Eazy	Mod.	Hard	MAE	MAE	(%)	
Baseline	23.64	17.51	15.14	-	-	4.08	-
Baseline+lo.	18.41	13.49	10.90	0.127	4.03	18.63	4.62
Baseline+fi.	21.93	15.86	13.22	0.250	8.47	45.72	5.40
Baseline+gl.	22.97	17.85	15.11	0.139	3.29	36.91	11.22
Baseline+gt.	26.21	19.43	16.50	0.097	3.23	59.08	18.29
Baseline+gl.+ed.	21.85	15.97	13.26	0.242	6.72	42.51	6.33
Baseline+gl.+di.	24.22	18.27	15.42	0.131	3.09	38.19	12.36

Table 3. Ablation study of y sources on KITTI val set. "lo." means using the local clues branch to predict y for each object. "fi." means using fixed 1.65 meters as the y source. "gl." means using the global clue branch to predict. "gt." means directly using the ground plane equation generated by the ground truth of val set. "ed." means using edge detection to obtain the horizon slope in the global clues branch. "di." means using dilated convolution.

Depth Form	Val, AP_{3D}			z	ESOP	CS↑
	Eazy	Mod.	Hard	MAE	(%)	
Baseline	23.64	17.51	15.14	-	4.08	-
Eq. (12)	23.16	17.62	14.73	2.27	25.69	11.32
Eq. (16)	21.83	15.97	13.19	8.65	45.40	5.25
Eq. (14)	24.22	18.27	15.42	3.09	38.19	12.36

Table 4. Ablation Study of complementary forms in KITTI val set. z MAE reflects the depth estimation accuracy in each form

By comparing the results of z_{comp} MAE and ESOP pairs under different settings, it can be found that determining whether complementary depth can lead to overall performance enhancement often requires evaluation from two perspectives: depth estimation accuracy and ESOP. This trend can be effectively quantified by complementary scores.

The results in the 6th to 7th rows of Tab. 3 justify the removal of edge detection and the use of dilated convolution when predicting the ground plane equation.

Complementary Form. To validate the effectiveness of achieving complementary form in enhancing detection accuracy, we present the results of different depth forms in Tab. 4. According to the results of the 2nd and 4th row in Tab. 4, the ESOP and CS of Eq. (14) are further enhanced after considering the complementary form compared to Eq. (12). Although a part of the depth estimation accuracy is sacrificed, the complementarity and overall per-

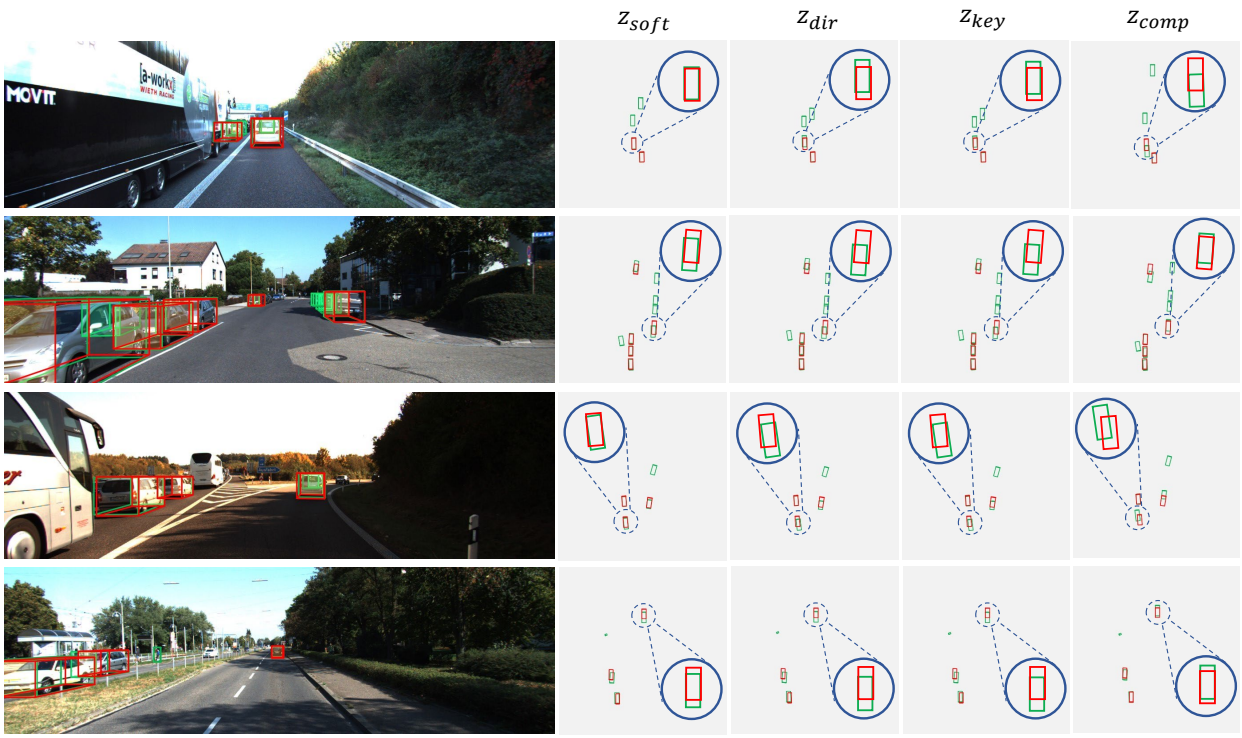

Figure 5. Qualitative examples on KITTI validation set. In each row, we provide one final front view (left) and four bird's-eye view (right) visualizations. The detection results for the various bird's-eye views vary only in terms of the depth output, progressing from z_{soft} to z_{dir}, z_{key}, and z_{comp} from left to right. Red represents the ground truth of boxes, while Green represents the predictions. We circle some objects to highlight the differences across multiple depth prediction branches.

formance are eventually improved, which is consistent with observation 3 in Sec. 3.2.

In addition to Eqs. (12) and (14) mentioned in Sec. 3.3, we also consider the following complementary form:

$$z = \frac{f_y(y_{glo} - H)}{v_t - c_v} \tag{16}$$

Although it appears that Eq. (16) is more symmetrical and complementary to z_{key} in form, its depth estimation error is significantly higher than that of Eq. (14). This is due to the fact that v_t and c_v in the denominator are relatively close, as well as the y_{glo} and H in the numerator, which causes an unstable depth estimation. This is also why Eq. (16) has a higher ESOP because the instability of the estimate mitigates the prediction tendency, but it does not contribute to the overall performance. It demonstrates the importance of an appropriate form of complementary depth.

4.6. Qualitative Results

Based on the qualitative results shown in Fig. 5, it can be observed that z_{comp} from the global clue branch is significantly different from z_{dir} and z_{key} from the local clue branch and has the opposite error sign. After combining z_{comp}, the predicted box is closer to the ground truth. This visualizes the process of error neutralization.

5. Conclusion

In this paper, we point out the coupling phenomenon that the existing multi-depth predictions tend to have the same sign, which limits the accuracy of combined depth. We analyze how complementary depth fixes it by mathematical derivation and find that the complementarity needs to be considered both from depth estimation accuracy and error sign opposite proportion. To improve depth complementarity, we propose to add a new depth prediction branch with the global clue and achieve complementarity in form through geometric relations. Extensive experiments demonstrate the effectiveness of our method. **Limitations.** The performance of our framework is limited by the accuracy of the vertical position of objects and the complementary effect may be lost when the ground plane is undulating. Future work could involve improving the understanding and prediction of global road scenarios.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China (No.62371201), by the Basic Research Surpport Plan of HUST (No.6142113-JCKY2022003), and by the China Scholarship Council for funding visiting Ph.D. student (No.202106160054).

References

[1] Garrick Brazil, Gerard Pons-Moll, Xiaoming Liu, and Bernt Schiele. Kinematic 3d object detection in monocular video. In *ECCV*, pages 135–152. Springer, 2020. 6

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229. Springer, 2020. 2

[3] Hansheng Chen, Yuyao Huang, Wei Tian, Zhong Gao, and Lu Xiong. Monorun: Monocular 3d object detection by reconstruction and uncertainty propagation. In *CVPR*, pages 10379–10388, 2021. 6

[4] Xiaozhi Chen, Kaustav Kundu, Yukun Zhu, Andrew G Berneshawi, Huimin Ma, Sanja Fidler, and Raquel Urtasun. 3d object proposals for accurate object class detection. *NeurIPS*, 28, 2015. 6

[5] Yongjian Chen, Lei Tai, Kai Sun, and Mingyang Li. Monopair: Monocular 3d object detection using pairwise spatial relationships. In *CVPR*, pages 12093–12102, 2020. 2

[6] Tom van Dijk and Guido de Croon. How do neural networks see depth in single images? In *ICCV*, pages 2183–2191, 2019. 5

[7] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, pages 3354–3361. IEEE, 2012. 2, 6, 7

[8] Kuan-Chih Huang, Tsung-Han Wu, Hung-Ting Su, and Winston H Hsu. Monodtr: Monocular 3d object detection with depth-aware transformer. In *CVPR*, pages 4012–4021, 2022. 2, 6

[9] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *NeurIPS*, 30, 2017. 3, 5

[10] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *CVPR*, pages 7482–7491, 2018. 3, 5

[11] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *CVPR*, pages 12697–12705, 2019. 1

[12] Peiliang Li, Xiaozhi Chen, and Shaojie Shen. Stereo r-cnn based 3d object detection for autonomous driving. In *CVPR*, pages 7644–7652, 2019. 1

[13] Peixuan Li, Shun Su, and Huaici Zhao. Rts3d: Real-time stereo 3d detection from 4d feature-consistency embedding space for autonomous driving. In *AAAI*, pages 1930–1939, 2021. 1

[14] Yingyan Li, Yuntao Chen, Jiawei He, and Zhaoxiang Zhang. Densely constrained depth estimator for monocular 3d object detection. In *ECCV*, pages 718–734. Springer, 2022. 6

[15] Zhuoling Li, Zhan Qu, Yang Zhou, Jianzhuang Liu, Haoqian Wang, and Lihui Jiang. Diversity matters: Fully exploiting depth clues for reliable monocular 3d object detection. In *CVPR*, pages 2791–2800, 2022. 1, 2, 3

[16] Qing Lian, Peiliang Li, and Xiaozhi Chen. Monojsg: Joint semantic and geometric cost volume for monocular 3d object detection. In *CVPR*, pages 1070–1079, 2022. 2, 6

[17] Xianpeng Liu, Nan Xue, and Tianfu Wu. Learning auxiliary monocular contexts helps monocular 3d object detection. In *AAAI*, pages 1810–1818, 2022. 2, 6, 7

[18] Zechen Liu, Zizhang Wu, and Roland Tóth. Smoke: Single-stage monocular 3d object detection via keypoint estimation. In *CVPRW*, pages 996–997, 2020. 2, 6

[19] Zongdai Liu, Dingfu Zhou, Feixiang Lu, Jin Fang, and Liangjun Zhang. Autoshape: Real-time shape-aware monocular 3d object detection. In *ICCV*, pages 15641–15650, 2021. 6

[20] 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 *ICCV*, pages 3111–3121, 2021. 2

[21] Xinzhu Ma, Yinmin Zhang, Dan Xu, Dongzhan Zhou, Shuai Yi, Haojie Li, and Wanli Ouyang. Delving into localization errors for monocular 3d object detection. In *CVPR*, pages 4721–4730, 2021. 2, 4, 6, 7

[22] Liang Peng, Xiaopei Wu, Zheng Yang, Haifeng Liu, and Deng Cai. Did-m3d: Decoupling instance depth for monocular 3d object detection. In *ECCV*, pages 71–88. Springer, 2022. 1

[23] Xidong Peng, Xinge Zhu, Tai Wang, and Yuexin Ma. Side: center-based stereo 3d detector with structure-aware instance depth estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 119–128, 2022. 1

[24] Rui Qian, Xin Lai, and Xirong Li. 3d object detection for autonomous driving: A survey. *Pattern Recognition*, 130: 108796, 2022. 1

[25] Zequn Qin and Xi Li. Monoground: Detecting monocular 3d objects from the ground. In *CVPR*, pages 3793–3802, 2022. 1, 3, 6

[26] Cody Reading, Ali Harakeh, Julia Chae, and Steven L Waslander. Categorical depth distribution network for monocular 3d object detection. In *CVPR*, pages 8555–8564, 2021. 6

[27] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In *CVPR*, pages 10529–10538, 2020. 1

[28] Shaoshuai Shi, Li Jiang, Jiajun Deng, Zhe Wang, Chaoxu Guo, Jianping Shi, Xiaogang Wang, and Hongsheng Li. Pv-rcnn++: Point-voxel feature set abstraction with local vector representation for 3d object detection. *International Journal of Computer Vision*, 131(2):531–551, 2023. 1

[29] Xuepeng Shi, Qi Ye, Xiaozhi Chen, Chuangrong Chen, Zhixiang Chen, and Tae-Kyun Kim. Geometry-based distance decomposition for monocular 3d object detection. In *ICCV*, pages 15172–15181, 2021. 2, 5, 6

[30] Yuguang Shi, Yu Guo, Zhenqiang Mi, and Xinjie Li. Stereo centernet-based 3d object detection for autonomous driving. *Neurocomputing*, 471:219–229, 2022. 1

[31] Andrea Simonelli, Samuel Rota Bulo, Lorenzo Porzi, Manuel López-Antequera, and Peter Kontschieder. Disentangling monocular 3d object detection. In *ICCV*, pages 1991–1999, 2019. 6

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017. 2

[33] Li Wang, Liang Du, Xiaoqing Ye, Yanwei Fu, Guodong Guo, Xiangyang Xue, Jianfeng Feng, and Li Zhang. Depth-conditioned dynamic message propagation for monocular 3d object detection. In *CVPR*, pages 454–463, 2021. 6

[34] Qiangeng Xu, Yiqi Zhong, and Ulrich Neumann. Behind the curtain: Learning occluded shapes for 3d object detection. In *AAAI*, pages 2893–2901, 2022. 1

[35] Fan Yang, Xinhao Xu, Hui Chen, Yuchen Guo, Jungong Han, Kai Ni, and Guiguang Ding. Ground plane matters: Picking up ground plane prior in monocular 3d object detection. *arXiv preprint arXiv:2211.01556*, 2022. 5, 6, 7

[36] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *CVPR*, pages 11784–11793, 2021. 2

[37] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015. 5

[38] Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. Deep layer aggregation. In *CVPR*, pages 2403–2412, 2018. 4, 6

[39] Renrui Zhang, Han Qiu, Tai Wang, Ziyu Guo, Ziteng Cui, Yu Qiao, Hongsheng Li, and Peng Gao. Monodetr: Depth-guided transformer for monocular 3d object detection. In *ICCV*, pages 9155–9166, 2023. 2, 6, 7

[40] Yunpeng Zhang, Jiwen Lu, and Jie Zhou. Objects are different: Flexible monocular 3d object detection. In *CVPR*, pages 3289–3298, 2021. 1, 2, 3, 5, 6, 7

[41] Yunpeng Zhang, Wenzhao Zheng, Zheng Zhu, Guan Huang, Dalong Du, Jie Zhou, and Jiwen Lu. Dimension embeddings for monocular 3d object detection. In *CVPR*, pages 1589–1598, 2022. 2

[42] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 1, 2, 4

[43] 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*, 44(12): 10114–10128, 2021. 2

[44] Yunsong Zhou, Hongzi Zhu, Quan Liu, Shan Chang, and Minyi Guo. Monoatt: Online monocular 3d object detection with adaptive token transformer. In *CVPR*, pages 17493–17503, 2023. 2

[45] Minghan Zhu, Lingting Ge, Panqu Wang, and Huei Peng. Monoedge: Monocular 3d object detection using local perspectives. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 643–652, 2023. 1