Front. Comput. Sci., 2024, 18(5): 185706
https://doi.org/10.1007/s11704-023-3242-2

RESEARCH ARTICLE

# GRAMO: geometric resampling augmentation for monocular 3D object detection

**He GUAN**[1,2], **Chunfeng SONG**[1,2], **Zhaoxiang ZHANG** (✉)[1,2]

1 School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China
2 Center for Research on Intelligent Perception and Computing, State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation Chinese Academy of Sciences, Beijing 100190, China



**Abstract** Data augmentation is widely recognized as an effective means of bolstering model robustness. However, when applied to monocular 3D object detection, non-geometric image augmentation neglects the critical link between the image and physical space, resulting in the semantic collapse of the extended scene. To address this issue, we propose two geometric-level data augmentation operators named Geometric-Copy-Paste (Geo-CP) and Geometric-Crop-Shrink (Geo-CS). Both operators introduce geometric consistency based on the principle of perspective projection, complementing the options available for data augmentation in monocular 3D. Specifically, Geo-CP replicates local patches by reordering object depths to mitigate perspective occlusion conflicts, and Geo-CS re-crops local patches for simultaneous scaling of distance and scale to unify appearance and annotation. These operations ameliorate the problem of class imbalance in the monocular paradigm by increasing the quantity and distribution of geometrically consistent samples. Experiments demonstrate that our geometric-level augmentation operators effectively improve robustness and performance in the KITTI and Waymo monocular 3D detection benchmarks.

**Keywords** 3D detection, monocular, augmentation, geometry

## 1 Introduction

3D object detection is a hot topic in computer vision, with widespread applications in autonomous driving, robot navigation, and virtual reality. A key requirement is the ability to accurately classify and localize objects in physical space. Currently, there are several alternatives for in-vehicle sensing devices, such as LiDAR sensors, which provide accurate depth but have high hardware costs, and stereo cameras, which are much cheaper but require strict calibration and sufficient texture. Recently, multi-cameras have gained attention due to the dual advantages of comprehensive view coverage and low equipment cost. As the cornerstone of multi-camera systems, monocular systems tend to be ubiquitous and scalable for improvement, providing a wider scope for exploration.

Data augmentation improves the generality and robustness of the model by increasing the number of samples. While many options have been validated in 2D tasks, extending augmentation to monocular systems requires a rethinking of geometric consistency constraints. Objects of the same scale in the physical world exhibit the visual rule of ‘bigger near, smaller far’ in a perspective environment, resulting in an uneven scale distribution. At the same time, occlusion and truncation of objects within the field of view, as well as differences in data collection scenarios, lead to an uneven distribution of samples. The unreliability and ambiguity of monocular depth estimation makes it difficult to observe distant objects in detail. Data augmentation promises to improve the degree of distributional homogeneity, but there are few augmentation operators that are geometrically consistent, and the discussion of whether geometric features remain robust after perturbation is not exhaustive.

Compared to image-level data augmentation, geometry-level data augmentation prioritizes the relationship between visual content and camera parameters, preserving more geometric cues for depth estimation. Objects further away from the camera have smaller aspect ratios and are closer to the vanishing point in the vertical direction [1]. Additionally, objects are typically placed on the ground and assumed to be perpendicular to the image plane, especially in driving scenes. The model can estimate object depth using both appearance and vertical height relative to the image. Inspired by the above analysis, recent work has developed geometric representations, such as leveraging the triangular proportion theorem for estimating object distances [2,3].

Considering the coherence of the visual-geometric space, we propose two novel data augmentation operations based on geometric resampling: Geometric-Copy-Paste and Geometric-Crop-Shrink. The former uses depth reordering and vertical view de-overlap to guide the paste process, while the latter simultaneously modifies the visual scale and visual depth of objects based on the principles of perspective and mitigates depth ambiguity by locally scaling patches. Geometric

Received March 24, 2023; accepted October 9, 2023

E-mail: zhaoxiang.zhang@ia.ac.cn

constraints are introduced into the data processing phase to increase sample diversity while preserving the original landmarks in the visual image. By increasing the geometric coherence, the proposed augmentation operations yield significant performance gains over state-of-the-art detectors, as shown in Figure 1. The detector also exhibits strong robustness to inter-domains with geometric consistency regularization. Our contributions are summarized below:

- In this paper, we provide two geometry-level data augmentation operators to improve the class imbalance and depth ambiguity problems in monocular 3D object detection while maintaining the geometric consistency of objects and scenes.
- Utilizing the proposed data augmentation operators, we achieve significant improvements in monocular 3D object detectors in both the KITTI and Waymo benchmarks.

## 2 Related work

**Monocular 3D object detection** Several researchers have focused on monocular 3D detection and explored different approaches. Estimating 3D objects from RGB inputs is inherently challenging due to the lack of depth information. The intuitive idea is to use representation transformation, where depth-guided approaches employ a depth estimation network to support 3D detection [4−6] or reuse depth information in an embedded form [7]. On the other hand, the pseudo-LiDAR based counterparts boost uplift 2D image to 3D point clouds and then apply standard lidar-based detectors [8–11], providing a powerful but high latency inference option. The construction of the bird's-eye-view (BEV) representation [12] also allows the recovery of detection results from the BEV projection space. A second line of work attempts 3D extensions of 2D detectors, covering both anchor-based [13,14] and anchor-free [15,16] series.

In recent work, researchers have either explicitly embedded geometric priors in the network [3,17,18] or implicitly learned

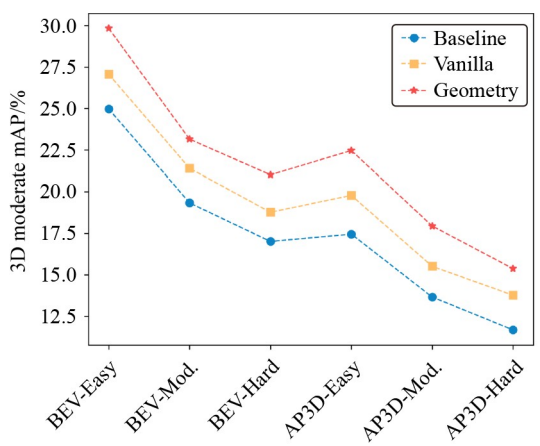


**Fig. 1** Below is a comparison of the performance of the automotive category in the KITTI 3D object detection benchmark. The results show a sustained performance improvement with the proposed geometric augmentation operators. In contrast, the "Vanilla" approach represents a clean baseline achieved by removing the existing enhancement operators

them through 2D-3D coherence constraints [19–21]. In addition to common 3D bounding boxes, some more efficient ways of inscribing object models are also attractive, such as wireframe models [22] or CAD models [23]. Keypoint-based methods either directly estimate the projected 3D centre of the object [19,24] or model the uncertainty in the relevant attributes [2,25] to reduce dense constraints, achieving a satisfactory compromise between performance and speed. 2D-3D keypoint correlation can provide advanced geometric constraints, a typical DCD [26] is to merge a large number of depth candidates generated from the associated points with a graph matching weighting module. MonoRUn [27] proposes an improved PnP algorithm to solve dense constraints in a self-supervised manner. In contrast, EPro-PnP [28] uses a derivable probability density to solve for the optimal poses.

**Data augmentation in detection** Data augmentation plays a significant role in improving the performance of the model. Common image augmentations include but are not restricted to colour fields (i.e., photometric distortion, etc.) as well as geometric fields (i.e., random shift, random flip, multi-scale training, etc.). Cut and paste is a representational operation for perceptual tasks involving object regions. Recent work explores location probability maps [29], semantic and depth information [30], and visual context modelling [31] as qualitative alternatives to the random paste principle [32]. Notice that the augmentation operator covers more than just 2D, with 3D extensions tailored to point clouds [33,34] such as rotation, translation, GT sampling [35] and point cloud mixing [36]. In the multimodal domain, multimodal copy-paste with perceptible occlusion [37] and cross-modal simultaneous data augmentation [38] are worth mentioning.

Although significant benefits have been achieved with augmentation in many 2D tasks [39], the lack of stable constraints on geometric consistency has meant that these strategies are rarely used in monocular systems [40]. It is difficult to synchronise the numerical values of the 3D attributes (position, dimension, orientation etc.) and the corresponding 2D appearance for editing the same object. The most relevant works [41,42] attempt to break this awkwardness. The former establishes a correlation between perspective distance and visible scale and modifies objects according to visual cues, while the latter distinguishes attribute depth from visual depth and proposes an instance depth segmentation strategy for data augmentation.

## 3 Overview and framework

### 3.1 Preliminaries

Given an RGB image with corresponding camera parameters, the monocular detector is primarily responsible for classifying the object of interest and determining its precise location with 3D bounding boxes. The bounding box is determined by the object's central position represented by $[X,Y,Z]^{\mathrm{T}}$, its dimensions represented by $[H,W,L]^{\mathrm{T}}$, and its orientation represented by the angle $\theta$. In autonomous driving scenarios, the ground plane is typically assumed to be horizontal, meaning that the orientation refers to the yaw angle, while the roll and pitch angles are typically set to zero by default. To

achieve geometric coherence, the 3D spatial state of an object must be synchronised with its 2D projection and the camera parameters. Considering the camera intrinsic $\mathbf{K} \in \mathbb{R}^{3\times3}$, the 2D-3D projection transformation from the coordination $(x,y,z)^{\mathrm{T}}$ of the image to the coordination $(X,Y,Z)^{\mathrm{T}}$ of camera is formulated as:

$$\begin{bmatrix} f_x & 0 & p_x \\ 0 & f_y & p_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = z \cdot \begin{bmatrix} u \\ v \\ 1 \end{bmatrix}, \tag{1}$$

where $f$ and $p$ are the focal length and optical centre, and $x$ and $y$ indicate the image coordinate axes of the image, respectively.

We subjectively select the simple and versatile centre-guided detector MonoDLE as a baseline and sequentially equip it with several data augmentation combinations to obtain state-of-the-art results. The effectiveness and stability of the proposed geometric operators are then verified on the depth-guided detector MonoDETR.

### 3.2 Non-geometric operation and geometric repackaging

Current monocular 3D object detectors have not reaped the research dividends of augmentation techniques, and the available operations remain limited, including flipping, color jittering, and affine resize. We believe that filling this omission could improve the overall benchmarks.

Depending on whether the perturbation involves camera parameters, we simply divide the available data augmentation in monocular systems into non-geometric and geometric levels. For non-geometric sets, color jittering is the only operation that does not involve geometric consistency. By varying the exposure, saturation, and hue of the image, the model can be adapted to scenes with different lighting conditions.

In contrast to the 2D detection scenario, we consider the synchronization issue between the visual content and camera parameters in a monocular system. First, we repackage the common image-level transformations, which include, but are not limited to operations such as cropping, shifting, rotation, and scaling. The flipping operator mirrors both the content and the optical centre of an image, without the need for vertical flipping for the driving scenario. The rotation operation is not necessary due to the assumption that the ground is flat and the camera plane is perpendicular to the ground. Conversely, altering the camera's position by means of shifting and cropping provides the flexibility to adjust the location of the optical centre, enabling the capture of varied image content. For instance, a top crop can be employed to lower the position of the optical centre along the $Y$-axis while simultaneously capturing the desired image content. Scaling or resizing an image with a fixed camera intrinsic, which has the same effect as pushing/pulling all objects in the scene proportionally relative to the camera.

Neglecting to synchronise camera parameters during data augmentation can directly affect 2D-3D projection consistency, leading to misalignment between features and annotations, and even destroying potentially common visual cues across images collected from the same device. Therefore, we incorporate the affine transformation into the camera intrinsic matrix during the perspective projection transformation process and formalise it as follows:

$$P_{new} = \begin{pmatrix} 1 & 0 & \frac{s_x(w-c_x)+d_x}{f_x} \\ 0 & 1 & \frac{s_y(w-c_y)+d_y}{f_y} \\ 0 & 0 & 1 \end{pmatrix} \cdot Pori, \tag{2}$$

where $s_x$ and $s_y$ are the ratios of shape scaling along $X$-axis and $Y$-axis, while $d_x$ and $d_y$ are the values of pixel shifting along $X$-axis and $Y$-axis respectively. $w$ denotes the long side of the image (usually the width). In the implementation we use the reversible order of flipping-scaling-shifting and stow the cropping into the shifting operator.

### 3.3 Geometric copy-paste in monocular system

Copy-paste is a content-driven data augmentation operator where a local patch is randomly selected and pasted to other locations. This scheme has been extended to LiDAR-based 3D object detection with a GT sampling strategy [35]. Since objects are collision-free with each other and naturally separated from the background in physical space, it is easy to collect objects of interest from other frames and insert them into the current frame.

However, a direct extension of the above operations to a monocular system is clearly not feasible. Severe object occlusion can greatly increase the ambiguity of 2D semantics, even if it appears to fit the scene in 3D space. Patches that are visible in the bird's eye view (BEV) are not necessarily guaranteed to remain distinguishable in the front view. Randomly pasting instances into a new scene may blur key visual cues, including geometric consistency and physical plausibility. Therefore, it is crucial to ensure that the pasted objects not only avoid overlapping with the original objects in the BEV, but also maintain a reasonable perspective-projection relationship between the objects.

**Copy-paste principle** The process of copying and pasting involves two fundamental principles: what to copy and where to paste. During the copying phase, we initially gather a database of box-level instances from the training data. We employ both offline and online preprocessing manners. The former involves offline traversal to store all instance samples that meet the criteria, while the latter filters and pastes across samples within the current batch. The online approach is suitable for large datasets and cases where pre-traversal is not feasible, but there is no guarantee that sufficient unduplicated samples will be collected for pasting in any batch. Any anomalous cases exceeding the truncation and occlusion thresholds are removed by filtering.

The plausibility of the object patch occlusion is the primary issue in the pasting session. To limit the number of pasted objects in the scene, we adjust the intersection over union (IoU) of the bounding boxes in 2D between occlusion samples, given the set of pre-collected box-level samples during the copying process. Objects with an IoU larger than a given threshold are discarded. Each object is then pasted to a specific location within the 2D bounding box based on its

inverse depth order. In other words, the further away an object is, the earlier it is pasted. The part obscured by a foreground object is represented by a block of overwritten frames. Additionally, both the number of single-category pastes and the proportion of multi-category pastes affect model performance. If the number or proportion of pastes is too small, the existing distribution will not be improved by augmentation. Conversely, if the number or proportion of pastes is too high, the pasted samples will tip over each other and there will be too few insufficient complete instances instead. The visualization of the Geometric-Copy-Paste operator is shown in Fig. 2.

### 3.4 Geometric crop-shrink in monocular system

The LiDAR-based approach uses a centred object scaling strategy that maintains the the bird's-eye view (BEV) position at the centre of the object. This assumption is reasonable because the point cloud is in a discrete distribution, and the points reflected from the object are centered on the 3D center.However, when object distance is changed by scaling the object, using a centered scaling pattern with the current monocular system results in the visual size and visual depth not being decoupled. To address this issue, it would be more reasonable to use the bottom center of the projected 2D bounding boxes as the base point for scaling, as it follows the triangular isometric relationship between object height and distance.

Considering the continuity of the pixels, we keep the visible edges of the BEV in place and smooth out the jumpy paste edges with a width of 3 pixels. In addition, directly shrinking the cropped object patch will result in a blank edge on the original image. To avoid giving away information about the object, assuming that the object is shrunk at scale $s$, we first scale up 2D crop by a factor of $1/s$ and then scale it back down by a factor of $s$. As a result, the black edges are filled in by the background on the fly and the object is correctly reduced. Based on the principle of similar triangles mapping [2,18], we synchronize the corresponding depth properties of the object to update them and achieve instance-level geometric augmentation in depth. The visualization of the Geometric-Crop-Shrink operator is shown in Fig. 3.

**Fig. 2** Visual illustration of the Geometric-Copy-Paste operation

## 4 Experiments

In this section, we first introduce the experimental setup, including evaluation benchmarks, metrics, and details of our implementation, then present and analyze the experimental results. We also verify the effectiveness of the proposed geometric augmentation technique in an common setting.

### 4.1 Experimental setup

**Datasets** We evaluate the effectiveness of the presented data augmentation approaches on the KITTI and Waymo 3D object detection benchmarks. The KITTI dataset comprises of 7,481 training images and 7,518 test images with 80,256 annotated 3D instances. To ensure fair comparisons, we followed the methodology of previous work [43] and divided the training data into a training set consisting of 3,712 samples and a validation set comprising of 3,769 samples. We evaluated the effectiveness of the proposed components on the validation set and then assessed the final model on the test set. In the Waymo Open dataset, there are 798 training sequences and 202 validation sequences. To ensure consistency, we collected one sample every 5 frames for both the training and validation sets.

**Evaluation metrics** KITTI assesses on 3 categories of items: Easy, Moderate, and Hard. The algorithm assigns each object to a category according to its occlusion, truncation and height in the image space. KITTI uses $AP_{3D|R_{40}}$ percentage of moderate difficulty as the benchmark indicator. Waymo evaluates on two object levels: Level_1 and Level_2. It assigns each object to a level based on the number of LiDAR

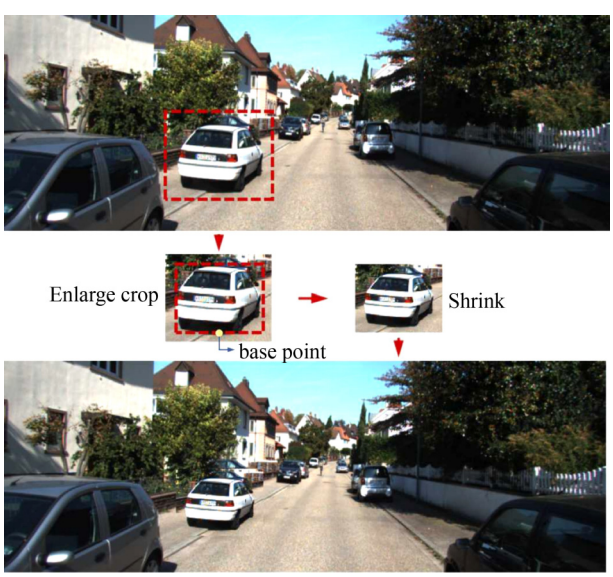


**Fig. 3** Visual illustration of the Geometric-Crop-Shrink operation

points included in its 3D box. Waymo uses the APH-3D percent metric which is the incorporation of heading information into AP3D to benchmark the models. In addition, it provides assessment at three distances [0, 30), [30, 50), and [50, $\infty$) metres. The primary metric used in Waymo is the recently proposed Longitudinal Error Tolerant 3D Average Precision (LET-3D-AP) which allows for longitudinal location errors of bounding boxes predicted up to a set tolerance.

**Implementation details** As stated in Section 1, the experiments are conducted with MonoDLE [24] as the baseline and DLA-34 [44] as the backbone, with parameters initialized with pre-trained ImageNet model weights. For feature map regularity, we pad the images in KITTI to the size of $1280 \times 384$ while the front-view images in Waymo to the size of $960 \times 640$ (one-half of the original resolution). The optimisation network costs 140 epochs on the KITTI dataset and only 12 epochs on the Waymo dataset. We adopt the Adam optimizer for training process and set the initial learning rate to 0.00025. All experiments are conducted with 2 NVIDIA 3090 GPUs with the total batch size to 16.

### 4.2 Ablation experiments on individual and compound effects of components.

We first remove the existing data augmentation operators to obtain a clean baseline for fair comparison. General-purpose operators for generating variants such as random flipping, colour dithering and affine resizing (using built-in random scaling, random cropping and random shifting) are added incrementally on top of it, and finally equipped with the two proposed geometric resampling operators. The image and camera intrinsic are modified synchronously throughout to maintain the consistency of the 2D-3D projections.

Table 1 shows the results of ablation for the amplification operations based on the MonoDLE baseline. As illustrated, the conventional operations steadily improve the performance of the vanilla model as expected, obtaining 9.56%, 6.78% and 5.46% on the three settings on the baseline, respectively. In contrast, our customised geometric manipulations gained a further 2.77%, 1.75%, and 2.24% with our resampling balance and depth perturbations, respectively. In addition, we observed that the improvement of geometry-free augmentation over geometry-aware augmentation tricks is limited. The potential reason for this is that monocular 3D detection relies heavily on geometric cues represented by depth.

We also conduct ablation experiments on some core sensitive parameters of the operation. Table 2 presents an inverse relationship between the intersection ratio and gain for different threshold settings. Interestingly, there is also a slight benefit of setting the threshold to zero. It means that newly inserted instances should preferably avoid visual occlusion with the original object, but there is no guarantee that non-overlapping sample candidates are sufficient at a threshold of zero. Table 3 verifies the effect of resampling ratio on multiple categories. The limited visual range of the pivot view restricts the amount or proportion of resampling from being too high. Sampling instances at a reasonable proportion helps to address the long-tailed distribution of raw samples.

To verify the generality and stability, we additionally employ a depth-guided paradigm detector called MonoDETR [45]. As shown in Table 4, there is still a significant improvement after equipping proposed augmentation operators, which suggests that our method is independent of detector choice and yields stable gains without additional time-consuming inference. Note that † indicates that the performance is reproduced by ourselves.

### 4.3 Results on the KITTI test set

In Table 5, we present a comparison of the proposed augmented detector with the state-of-the-art method on the KITTI test set. Quantitatively, the baseline approach with the fictitious augmentation has achieved comparable results in each setting. Supported by the proposed geometric resampling augmentation strategy, we outperform the baseline with 5.11%, 3.41%, and 2.83% in three different difficulty levels of the 3D task, respectively. It is worth noting that the runtime metric is measured in milliseconds and our operation without any extra inference latency cost. Figure 4 displays the qualitative results on the KITTI dataset generated by the

**Table 1** Ablations on the KITTI *val* set for amplification operations

| | $AP_{BEV/3D}\|R_{40}$(IoU = 0.7, Car) | | |
|---|---|---|---|
| | Easy | Mod. | Hard |
| Vanilla | 17.50/11.47 | 14.62/10.04 | 12.36/8.30 |
| + Random Flipping | 22.00/15.39 | 17.64/12.68 | 15.93/10.67 |
| + Color Jittering | 23.64/16.84 | 18.84/13.83 | 16.25/11.65 |
| + Affine Resize | 27.06/19.78 | 21.40/15.50 | 18.77/13.77 |
| + Geo-CP | 28.86/20.67 | 22.83/16.99 | 20.80/14.65 |
| + Geo-CS | 29.83/22.47 | 23.15/17.94 | 21.01/15.37 |

**Table 2** Ablations on the KITTI *val* set for different IoU thresholds

| | $AP_{BEV/3D}\|R_{40}$(IoU = 0.7, Car) | | |
|---|---|---|---|
| | Easy | Mod. | Hard |
| 0.0 | 30.26/21.13 | 22.57/15.66 | 19.45/13.07 |
| 0.05 | 29.83/22.47 | 23.15/17.94 | 21.01/15.37 |
| 0.1 | 28.25/22.13 | 22.62/17.57 | 19.68/14.99 |
| 0.3 | 29.40/21.15 | 23.26/17.16 | 20.96/14.73 |
| 0.5 | 28.05/20.88 | 22.44/17.03 | 20.46/14.84 |
| 0.7 | 27.88/21.45 | 22.51/16.75 | 19.84/14.91 |

**Table 3** Ablations on the KITTI *val* set for different sampling proportions

| | $AP_{BEV/3D}\|R_{40}$(IoU = 0.7, Car : Pedestrian : Cyclist) | | |
|---|---|---|---|
| | Easy | Mod. | Hard |
| 5:0:0 | 26.57/19.95 | 21.27/15.40 | 18.79/13.80 |
| 5:2:2 | 28.06/21.24 | 21.90/15.98 | 19.27/14.31 |
| 10:0:0 | 25.20/18.34 | 19.95/14.76 | 17.91/12.43 |
| 10:1:1 | 26.85/19.02 | 21.98/15.62 | 19.27/14.03 |
| 10:3:3 | 29.83/22.47 | 23.15/17.94 | 21.01/15.37 |
| 10:5:5 | 28.13/20.10 | 22.65/16.32 | 19.62/14.34 |
| 15:2:2 | 26.60/19.98 | 21.67/15.98 | 19.14/14.24 |
| 15:5:5 | 26.38/18.99 | 21.23/15.87 | 19.30/14.25 |

**Table 4** Ablations on the KITTI *val* set for another detector

| Method | $AP_{BEV/3D}\|R_{40}$(IoU = 0.7, Car) | | |
|---|---|---|---|
| | Easy | Mod. | Hard |
| MonoDETR† | 35.88/26.26 | 24.78/18.59 | 20.92/15.34 |
| MonoDETR (ours) | 38.26/27.04 | 27.15/19.93 | 23.04/16.10 |
| Improvement | +2.38/+0.78 | +2.37/+1.34 | +2.12/+0.76 |

**Table 5** Comparative analyses on the KITTI *test* set. We highlight the best results in **bold** and the second place in underlined

| Methods | Runtime/ms | $AP_{BEV}\|R_{40}$(IoU = 0.7) | | | $AP_{3D}\|R_{40}$(IoU = 0.7) | | |
|---|---|---|---|---|---|---|---|
| | | Easy | Moderate | Hard | Easy | Moderate | Hard |
| MonoGRNet [5] | 400 | 18.19 | 11.17 | 8.73 | 15.74 | 9.61 | 4.25 |
| M3D-RPN [14] | 160 | 21.02 | 13.67 | 10.23 | 14.76 | 9.71 | 7.42 |
| MonoPair [20] | 60 | 19.28 | 14.83 | 12.89 | 13.04 | 9.99 | 8.65 |
| PatchNet [11] | 400 | 22.97 | 16.86 | 14.97 | 15.68 | 11.12 | 10.17 |
| D4LCN [4] | 200 | 22.51 | 16.02 | 12.55 | 16.65 | 11.72 | 9.51 |
| GrooMeD-NMS [46] | 120 | 26.19 | 18.27 | 14.05 | 18.10 | 12.32 | 9.65 |
| MonoRCNN [13] | 70 | 25.48 | 18.11 | 14.10 | 18.36 | 12.65 | 10.03 |
| CaDDN [12] | 630 | 27.94 | 18.91 | 17.19 | 19.17 | 13.41 | 11.46 |
| MonoFlex [25] | 35 | 28.23 | 19.75 | 16.89 | 19.94 | 13.89 | 12.07 |
| AutoShape [23] | 40 | 30.66 | 20.08 | 15.95 | 22.47 | 14.17 | 11.36 |
| GUPNet [2] | 34 | 30.29 | 21.19 | 18.20 | 22.26 | 15.02 | 13.12 |
| MonoCon [19] | 26 | 31.12 | **22.10** | **19.00** | **22.50** | **16.46** | **13.95** |
| MonoDLE [24] (baseline) | 40 | 24.79 | 18.89 | 16.00 | 17.23 | 12.26 | 10.29 |
| MonoDLE (ours) | 40 | **32.44** | 21.74 | 18.38 | 22.34 | 15.67 | 13.12 |
| Improvements | | +7.65 | +2.85 | +2.38 | +5.11 | +3.41 | +2.83 |

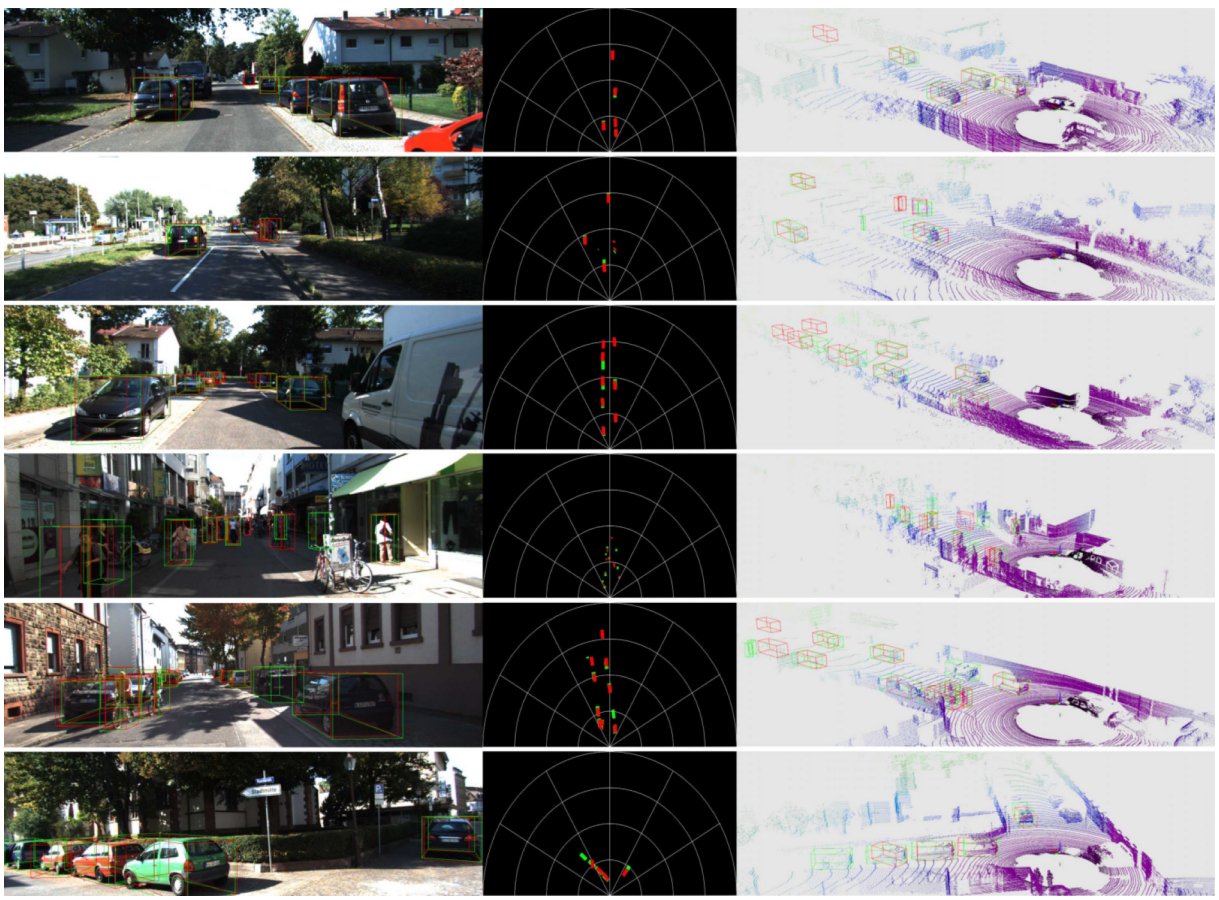

**Fig. 4** Qualitative results on the KITTI dataset. The prediction results are shown from left to right in the following order: image, BEV and Lidar. We color the predictions red and the ground truths green, including Car, Pedestrian, and Cyclist. LiDAR signals are shown for visualization only. Best viewed in color with zoom-in

MonoDLE applied with proposed Geo-CP and Geo-CS augmentation operations.

### 4.4 Results on the Waymo *val* set

With the exception of the KITTI dataset, we also evaluate the proposed augmentation techniques on the Waymo dataset on a large scale. Table 6 presents the experimental results of the modified MonoDLE on Waymo validation set. Although Waymo contains more richer set of training instances, the proposed geometry-aware augmentation techniques still improve the vanilla setting in different evaluation metrics. In general, with respect to the most important mAP metric, the geometry aware strategy outperforms the baseline over 1.90% at LEVEL_1 and 2.21% at LEVEL_2. This further validates the efficacy of the proposed method.

**Table 6** Results on Waymo *val* set. We highlight the best results in **bold** and the second place in underlined

| Methods | LEVEL_1 | | | | LEVEL_2 | | | |
|---|---|---|---|---|---|---|---|---|
| | $3DAP_{70}$ | $3DAPH_{70}$ | $3DAP_{50}$ | $3DAPH_{50}$ | $3DAP_{70}$ | $3DAPH_{70}$ | $3DAP_{50}$ | $3DAPH_{50}$ |
| M3D-RPN [14] | 0.35 | 0.34 | 3.79 | 3.63 | 0.33 | 0.33 | 3.61 | 3.46 |
| PatchNet [11] | 0.39 | 0.37 | 2.92 | 2.74 | 0.38 | 0.36 | 2.42 | 2.28 |
| GUPNet [2] | 2.28 | 2.27 | 10.02 | 9.94 | 2.14 | 2.12 | 9.39 | 9.31 |
| CaDDN [12] | 5.03 | 4.99 | 17.54 | 17.31 | 4.49 | 4.45 | 16.51 | 16.28 |
| DID-M3D [42] | – | – | 20.66 | 20.47 | – | – | 19.37 | 19.19 |
| BEVFormer [47] | – | 7.70 | – | 30.80 | – | 6.90 | – | 27.70 |
| MonoFlex [25] | 11.70 | 11.64 | 32.26 | 32.06 | 10.96 | 10.90 | 30.31 | 30.12 |
| DCD [26] | 12.57 | 12.50 | **33.44** | **33.24** | 11.78 | 11.72 | 31.43 | 31.25 |
| MonoDLE (baseline) | 10.93 | 9.93 | 28.35 | 27.93 | 9.66 | 9.58 | 27.90 | 27.55 |
| MonoDLE (ours) | **12.83** | **12.74** | 33.01 | 32.38 | **11.87** | **12.04** | **32.13** | **31.58** |
| Improvements | +1.90 | +2.81 | +4.67 | +4.45 | +2.21 | +2.46 | +4.23 | +4.03 |

## 5 Discussion

According to Table 1, Geo-CP and Geo-CS exhibit improvements for both near and distant objects. The former having high generalization to the full depth range, while the latter still falls short in perceiving far objects. Among them, Geo-CP improves the occurrence rate with the original distribution, which is equivalent to enlarging the set of distance-independent instances. Since the projections of distant objects are naturally focused on a few pixels locally at the top side of the image, while the projections of near objects are scattered over a large area at the bottom of the image, it is difficult to improve the detection quality of distant objects by Geo-CS by shrinking a few pixels precisely.

## 6 Conclusion

When extending image-based augmentation techniques to monocular systems, the geometric consistency issues caused by the perspective principle are a major challenge and become a bottleneck for improving the performance of 3D detectors. To alleviate this dilemma, we propose two geometric operators for data augmentation, namely Geometric Copy-Paste and Geometric Crop-Shrink, to generate more diverse samples without breaking the geometric consistency principle. Experiments on KITTI and Waymo datasets have confirmed that these geometric augmentations provides stable improvements over state-of-the-art methods. Although the above discussion is limited to perspective occlusion and depth, we will continue to explore the stability under the camera perturbations environment in order to provide a wider range of augmentation techniques for future monocular 3D detection.

**Acknowledgements** This work was supported in part by the National Key R&D Program of China (No. 2022ZD0160102), and the National Natural Science Foundation of China (Grant Nos. 61836014, U21B2042, 62072457, 62006231).

**Competing interests** The authors declare that they have no competing interests or financial conflicts to disclose.



## References

1. Dijk T V, Croon G D. How do neural networks see depth in single images? In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019, 2183−2191
2. Lu Y, Ma X, Yang L, Zhang T, Liu Y, Chu Q, Yan J, Ouyang W. Geometry uncertainty projection network for monocular 3D object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021, 3111−3121
3. Qin Z, Li X. MonoGround: detecting monocular 3D objects from the ground. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022, 3793−3802
4. Ding M, Huo Y, Yi H, Wang Z, Shi J, Lu Z, Luo P. Learning depth-guided convolutions for monocular 3D object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020, 1000−1001
5. Qin Z, Wang J, Lu Y. MonoGRNet: a geometric reasoning network for monocular 3D object localization. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2019, 8851−8858
6. Wang L, Du L, Ye X, Fu Y, Guo G, Xue X, Feng J, Zhang L. Depth-conditioned dynamic message propagation for monocular 3D object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021, 454−463
7. Park D, Ambrus R, Guizilini V, Li J, Gaidon A. Is pseudo-lidar needed for monocular 3D object detection? In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021, 3142−3152
8. Wang Y, Chao W, Garg D, Hariharan B, Campbell M, Weinberger K. Pseudo-LiDAR From Visual Depth Estimation: Bridging the Gap in 3D Object Detection for Autonomous Driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019, 8445−8453
9. Qian R, Garg D, Wang Y, You Y, Belongie S, Hariharan B, Campbell M, Weinberger K, Chao W. End-to-end Pseudo-LiDAR for image-based 3D object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020, 5881−5890
10. Chen Y, Dai H, Ding Y. Pseudo-Stereo for monocular 3D object detection in autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022, 887−897
11. Ma X, Liu S, Xia Z, Zhang H, Zeng X, Ouyang W. Rethinking Pseudo-LiDAR representation. In: Proceedings of European Conference on Computer Vision. 2020, 311−327
12. Reading C, Harakeh A, Chae J, Waslander S. Categorical depth

distribution network for monocular 3D object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021, 8555−8564
13. Shi X, Ye Q, Chen X, Chen C, Chen Z, Kim T. Geometry-based distance decomposition for monocular 3D object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021, 15172−15181
14. Brazil G, Liu X. M3D-RPN: monocular 3D region proposal network for object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019, 9287−9296
15. Luo S, Dai H, Shao L, Ding Y. M3DSSD: monocular 3D single stage object detector. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021, 6145−6154
16. Wang T, Zhu X, Pang J, Lin D. FCOS3D: fully convolutional one-stage monocular 3D object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021, 913−922
17. Mousavian A, Anguelov D, Flynn J, Kosecka J. 3D bounding box estimation using deep learning and geometry. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2017, 7074−7082
18. Shi X, Chen Z, Kim T. Distance-normalized unified representation for monocular 3D object detection. In: Proceedings of European Conference on Computer Vision. 2020, 91−107
19. Liu X, Xue N, Wu T. Learning auxiliary monocular contexts helps monocular 3D object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2022, 1810−1818
20. Chen Y, Tai L, Sun K, Li M. MonoPair: monocular 3D object detection using pairwise spatial relationships. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020, 12093−12102
21. Gu J, Wu B, Fan L, Huang J, Cao S, Xiang Z, Hua X. Homography loss for monocular 3D object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022, 1080−1089
22. Chabot F, Chaouch M, Rabarisoa J, Teuliere C, Chateau T. 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, 2040−2049
23. Liu Z, Zhou D, Lu F, Fang J, Zhang L. AutoShape: Real-Time Shape-Aware Monocular 3D Object Detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021, 15641−15650
24. Ma X, Zhang Y, Xu D, Zhou D, Yi S, Li H, Ouyang W. Delving into localization errors for monocular 3D object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021, 4721−4730
25. Zhang Y, Lu J, Zhou J. Objects are different: flexible monocular 3D object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021, 3289−3298
26. Li Y, Chen Y, He J, Zhang Z. Densely constrained depth estimator for monocular 3D object detection. In: European Conference on Computer Vision. 2022, 718−734
27. Chen H, Huang Y, Tian W, Gao Z, Xiong L. MonoRUn: monocular 3D object detection by reconstruction and uncertainty propagation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021, 10379−10388
28. Chen H, Wang P, Wang F, Tian W, Xiong L, Li H. EPro-PnP: generalized end-to-end probabilistic perspective-n-points for monocular object pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022, 2781−2790
29. Fang H, Sun J, Wang R, Gou M, Li Y, Lu C. InstaBoost: boosting instance segmentation via probability map guided copy-pasting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019, 682−691
30. Georgakis G, Mousavian A, Berg A, Kosecka J. Synthesizing training data for object detection in indoor scenes. 2017, arXiv preprint arXiv: 1702.07836
31. Dvornik N, Mairal J, Schmid C. Modeling visual context is key to augmenting object detection datasets. In: Proceedings of the European Conference on Computer Vision. 2018, 364−380
32. Dwibedi D, Misra I, Hebert M. Cut, paste and learn: surprisingly easy synthesis for instance detection. In: Proceedings of the IEEE International Conference on Computer Vision. 2017, 1301−1310
33. Wang H, Huang D, Wang Y. GridNet: efficiently learning deep hierarchical representation for 3D point cloud understanding. Frontiers of Computer Science, 2022, 16(1): 161301.
34. Xian Y, Xiao J, Wang Y. A fast registration algorithm of rock point cloud based on spherical projection and feature extraction. Frontiers of Computer Science, 2019, 13(1): 170−182
35. Yan Y, Mao Y, Li B. SECOND: sparsely embedded convolutional detection. Sensors, 2018, 18(10): 3337
36. Xiao A, Huang J, Guan D, Cui K, Lu S, Shao L. PolarMix: a general data augmentation technique for LiDAR point clouds. In: Proceedings of Advances in Neural Information Processing Systems. 2022, 11035−11048
37. Zhang W, Wang Z, Loy C. Exploring data augmentation for multi-modality 3D object detection. 2021, arXiv preprint arXiv: 2012.12741.
38. Wang C, Ma C, Zhu M, Yang X. Point augmenting: cross-modal augmentation for 3D object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021, 11794−11803
39. Jiang H, Cheng M, Li S, Borji A, Wang J. Joint salient object detection and existence prediction. Frontiers of Computer Science, 2019, 13(1): 778−788
40. Yang X, Xue T, Luo H, Guo J. Fast and accurate visual odometry from a monocular camera. Frontiers of Computer Science, 2019, 13(1): 1326−1336
41. Lian Q, Ye B, Xu R, Yao W, Zhang T. Exploring geometric consistency for monocular 3D object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022, 1685−1694
42. Peng L, Wu X, Yang Z, Liu H, Cai D. DID-M3D: decoupling instance depth for monocular 3D object detection. In: Proceedings of European Conference on Computer Vision. 2022, 71−88
43. Chen X, Kundu K, Zhang Z, Ma H, Fidler S, Urtasun R. Monocular 3D object detection for autonomous driving. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016, 2147−2156
44. Yu F, Wang D, Shelhamer E, Darrell T. Deep layer aggregation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018, 2403−2412
45. Zhang R, Qiu H, Wang T, Guo Z, Qiao Y, Li H, Gao P. MonoDETR: Depth-guided transformer for monocular 3D object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023, 9155−9166
46. Kumar A, Brazil G, Liu X. 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, 8973−8983
47. Li Z, Wang W, Li H, Xie E, Sima C, Lu T, Yu Q, Dai J. BEVFormer: learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In: Proceedings of European Conference on Computer Vision. 2022, 1−18

He Guan is now a PhD candidate with the University of Chinese Academy of Sciences, China. He received the bachelor degree from Harbin Institute of Technology, China in 2015, and the master's degree from the Institute of Automation, Chinese Academy of Sciences, China in 2018. His research interests include computer graphics and computer vision.

Chunfeng Song received the PhD degree from University of Chinese Academy of Sciences, China in 2020. He is now working at the Center for Research on Intelligent Perception and Computing, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences as an Assistant Professor. He has published more than 20 conference and journal papers such as IEEE TPAMI, TIP, IJCV, CVPR, ECCV, and AAAI. His current research focuses on person identification, image segmentation, and unsupervised learning.

Zhaoxiang Zhang received his bachelor degree in Circuits and Systems from the University of Science and Technology of China, China in 2004, and he received his PhD degree in 2009. He is now a full Professor in the Center for Research on Intelligent Perception and Computing and the State Key Laboratory of Multimodal Artificial Intelligence Systems, China. His research interests include computer vision, pattern recognition, and machine learning. Specifically, he recently focuses on biologically inspired intelligent computing and its applications in human analysis and scene understanding. He has published more than 150 papers in international journals and conferences, such as IEEE TPAMI, TIP, TIFS, IJCV, CVPR, ICCV, ECCV, and NeurIPS.