Exploring Geometric Consistency for Monocular 3D Object Detection

Qing Lian [1], Botao Ye[2,3], Ruijia Xu[1], Weilong Yao[3], Tong Zhang[1]
[1]The Hong Kong University of Science and Technology,
[2]Institute of Computing Technology, Chinese Academy of Sciences, China [3] Autowise.AI
qlianab@connect.ust.hk, botao.ye@vipl.ict.ac.cn, rxuaq@connect.ust.hk,
yaoweilong@autowise.ai, tongzhang@ust.hk

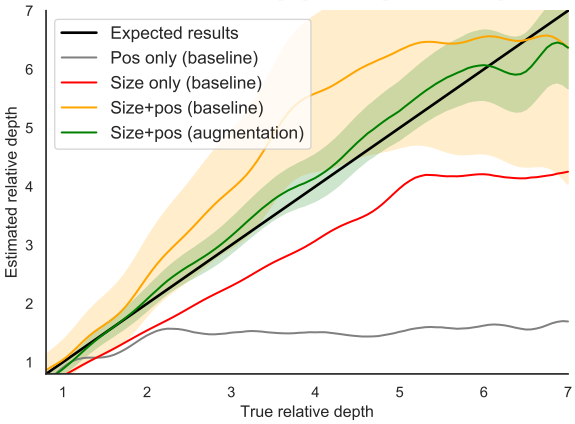
Abstract

This paper investigates the geometric consistency for monocular 3D object detection, which suffers from the ill-posed depth estimation. We first conduct a thorough analysis to reveal how existing methods fail to consistently localize objects when different geometric shifts occur. In particular, we design a series of geometric manipulations to diagnose existing detectors and then illustrate their vulnerability to consistently associate the depth with object apparent sizes and positions. To alleviate this issue, we propose four geometry-aware data augmentation approaches to enhance the geometric consistency of the detectors. We first modify some commonly used data augmentation methods for 2D images so that they can maintain geometric consistency in 3D spaces. We demonstrate such modifications are important. In addition, we propose a 3D-specific image perturbation method that employs the camera movement. During the augmentation process, the camera system with the corresponding image is manipulated, while the geometric visual cues for depth recovery are preserved. We show that by using the geometric consistency constraints, the proposed augmentation techniques lead to improvements on the KITTI and nuScenes monocular 3D detection benchmarks with state-of-the-art results. In addition, we demonstrate that the augmentation methods are well suited for semi-supervised training and cross-dataset generalization.

1. Introduction

Given an input image, the objective of monocular 3D object detection is to detect objects of interest and recover their position in 3D space. Recently, it has received increasing attention due to its importance in many downstream tasks, such as autonomous driving, robot navigation, *etc.* Different from stereo or lidar sensors, a monocular camera requires a lower cost to perceive the surrounding environments. However, it suffers from unreliable depth recovery, leading to unsatisfied performance for deployment.

(a) Visualization of different copy-paste manipulation techniques.



(b) Visualization of the estimated depth from the baseline and augmentation-enhanced detectors under the copy-paste manipulation (see the details in Sec 4.2).

Figure 1. We select one of the proposed manipulation techniques (copy-paste) to illustrate the instability of object localization under distortion of objects' apparent size and vertical position. "Size+pos" denotes geometry-consistent manipulation that shifts the two visual cues with satisfying geometric constraints, "Size only" and "Pos only" denote geometry-inconsistent manipulation that only shifts the vertical position or apparent size. The shaded region indicates the std of the depth in the "Size + pos" manipulation.

To alleviate the ambiguity in depth estimation, recent approaches [1, 3, 22, 24] leverage deep neural networks to model the semantic and geometric information for depth

reasoning. However, what geometric features existing detectors use and if they are robust when the used features are perturbed are still under-explored. As a result, this work conducts a comprehensive study on the geometry robustness of existing detectors and proposes several augmentation techniques to enhance their geometric consistency under geometric shifts. Different from 2D detection, the geometric visual cues for depth recovery are supposed to be preserved when the objects' coordinates are manipulated, which is not straightforward.

It is demonstrated in [10] that neural networks might rely on the features of appearance size and vertical position to estimate object depth. As visualized in Figure 1a, objects farther away from the camera have smaller apparent sizes and their vertical position is closer to the vanishing points. To study if detectors utilize these two pictorial visual cues in localizing objects, we conduct controlled experiments that shift one of the visual cues during manipulating. As the results of "Size + pos", "Size only" and "Pos only" shown in Figure 1, the estimated depth changes as the shift of pictorial visual cues, especially for the objects' apparent size. We further evaluate the robustness of the detectors in utilizing them to estimate depth by manually distorting the visual cues (*i.e.,* shifting the objects' apparent size or vertical position) with the proposed manipulations (visualized in Figure 2 and 1a). Through the evaluation, we observe that detectors cannot capture consistent relationships between depth with the two pictorial visual cues, even they can identify the variation of them. As shown in Figure 1b and 3, the estimated depth from the baseline detectors has a strong deviation when the images are manipulated.

Inspired by the above analysis, we convert the manipulations into several geometry-aware data augmentation techniques to improve the geometric consistency of existing detectors. The awareness means that the pictorial visual cues for estimating object depth are preserved during manipulating. At the image level, we lift random scale and random crop, the commonly used 2D augmentation to 3D space by connecting the image manipulation with the shift of camera focal lengths and receptive field. With the help of a dense depth estimation network, we provide a new 3D augmentation method that models the shifts of the camera's 3D location. At the instance level, we propose a geometry-aware copy-paste that leverage the guidance of geometric hints to guide the pasting procedure. Through modeling the geometric constraints, the objects are pasted to novel scenes, while their pictorial visual cues are still preserved.

By enhancing the geometric consistency, the proposed augmentation techniques yield significant performance boost in both state-of-the-art anchor-free and anchor-based detectors. Compared with the baseline in Figure 1b, the estimated depth from the enhanced detectors with the designed geometric augmentation methods has less deviation under manipulation. With regularizing the geometric consistency, the trained detectors also show strong robustness in the cross-domain scenario. Furthermore, the consistency regularization techniques also can be applied in the semi-supervised setting, which boosts the performance by regularizing the output consistency under different levels of manipulations. Our contributions are summarized as follows:

- Through a study of how monocular detectors estimate depth, we identified an instability problem of depth recovery under the changes of the object's apparent size and position.
- We provide four geometry-aware augmentation techniques at the image-level and instance-level to address this problem. With the proposed augmentation techniques, we achieve state-of-the-art results on both the KITTI and nuScenes monocular 3D object detection benchmarks.
- We extend the geometry augmentation techniques into semi-supervised training and cross-domain evaluation, showing the effectiveness of improving performance by regularizing the geometric consistency.

2. Related work

In this section, we present the review on monocular 3D object detection and the data augmentation techniques used in object detection.

2.1. Monocular 3D detection

Current monocular 3D object detectors can be split into two categories: image-based and pseudo-lidar based.

Image-based approaches estimate the 3D information by lifting 2D detectors [36, 49] to the 3D space. Traditional approaches [1, 38, 49] infer the 3D bounding boxes by additionally estimating location, dimension, and orientation based on 2D detectors [36,49]. M3D-RPN [1] redesigns the anchor proposal module to better extract 3D information. MonoDis [38] and MonoFlex [48] address the multi-task learning by disentangling the loss functions and neural network architectures. Shi et al. [37] and Yan et al. [30] decompose the depth into two easier estimated metrics: 2D and 3D height. To alleviate the label noise in object location, multiple approaches [30–32, 37, 50] model the aleatoric uncertainty in both the training and inference stages. In addition, several methods take external information [3, 11, 33, 34] (*e.g.,* semantic segmentation, CAD model, the ground surface) to enrich the contextual information for localization.

Except for directly regressing depth, several approaches design 2D and 3D geometry constraints for object depth recovery. RTM3D [24], KM3D-Net [23], and MonoPair [5] propose to use the geometric constraints to recovery depth from the constraints in single instance [23, 24] or pairwise

instances [5]. Similar to MonoPair [5], RAR-Net [26] proposes a reinforcement learning based post-processing strategy to refine the 3D information. To alleviate the sparse constraints, AutoShape [28] utilizes CAD models to learn dense keypoints to label the semantic keypoints. MonoRun enriches the sparse keypoint constraint to a self-supervised dense constraint, where a modified PnP algorithm is proposed to solve the designed constraint.

In addition to directly taking the monocular image as input, pseudo-lidar based approaches [35, 41–43, 46] adopt a depth estimation network [14] to convert the 2D images into 3D point cloud and then apply a point cloud detector on them. Although they achieve superior performance, the input transformation requires an extra depth estimation module during inference, leading to high latency.

2.2. Data augmentation in object detection

Data augmentation is an effective technique to boost the performance of object detection [25, 36, 51]. Both geometry-based (*e.g.,* random scale, random crop, and *etc.*) and color-based (*e.g.,* color distortion) augmentation techniques have been widely adopted in 2D detection models [25, 36, 49, 51]. In addition, copy-paste augmentation has also proven to be an effective technique to improve the generalization in detection and segmentation. Dvornik et al and Zuo et al [12, 39] propose to guide the object pasting by aligning the visual context before and after the augmentation. InstaBoost [13] proposes a probability heatmap to learn where to paste. In the 3D space, Moca [47] proposes an occlusion-aware copy-paste approach for multi-modality 3D detection. In lidar-based detection, data augmentation is also widely adopted [6, 21, 40]. Besides the common schemes used in object detection, there are several special augmentation methods tailored to point cloud data, such as the random erasing in SECOND [45], part-aware data augmentation method in [7].

While these aggressive data augmentation methods have yielded impressive gains for either 2D cases or some specific 3D data representation, however, they are hardly leveraged in current monocular 3D detection frameworks due to the violation of geometric constraints, where horizontal flip and color distortion are the only two methods used in this field for a long time. To this end, we hope to reshape this embarrassing situation by offering more diverse geometry-consistent data augmentation techniques to enhance the baseline monocular 3D detectors.

3. Preliminaries

3.1. Baselines

In this section, we first introduce the basic setup of the monocular detectors. We use lower-case and upper-case letters to represent the 2D and 3D coordinates, respectively. Monocular 3D detectors are required to recover the following 3D information: (1) 3D bounding box dimension $[W, H, L]^T$, (2) 3D bounding box center location $P = [X, Y, Z, 1]^T$ (3) object yaw angle θ. On the KITTI dataset [15], the following coordinate conversion is adopted to connect the 2D and 3D coordinate:

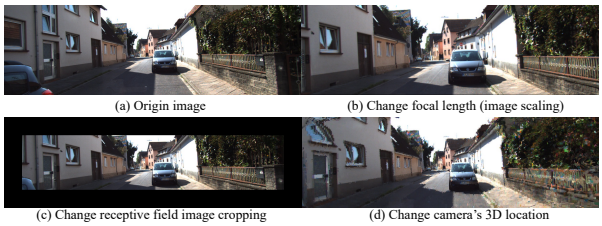


Figure 2. Visualization of the image-level manipulation.

$$p = \frac{1}{Z} K P, \tag{1}$$

where $p = [u, v, 1]^T$ is the 2D location of the 3D center projected in the image and the transformation matrix K is formulated as:

$$K = \begin{pmatrix} f & 0 & c_u & 0 \\ 0 & f & c_v & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}. \tag{2}$$

In this work, we adopt one anchor-free (CenterNet [49]) and one anchor-based (M3D-RPN [1]) detectors as our baselines and lift them to state-of-the-art results by several recently proposed techniques. (1) For depth estimation, we follow [5, 32, 48] and model the regression uncertainty with laplacian distribution during training and inference. (2) We add an integral corner loss as in [38, 48] to directly supervise the estimated bounding box coordinates with ground-truth. (3) Following [24, 27, 32, 48], we replace the objective of the classification heatmap in CenterNet from the 2D bounding box center to the projected 3D bounding box center.

3.2. Pictorial visual cues

In human and machine perception, researchers [10, 17] provide several pictorial visual cues that might be used for 3D reconstruction, including object apparent size, vertical position, occlusion, shading, and *etc.* As part of the objective in 3D object detection, the object's apparent size and vertical position are the two most relevant cues for object depth recovery. We visualize the relationships between them with depth in Figure 4. As shown in Figure 4, the orange triangle displays the relationship between 2D bounding box height h and 3D bounding box height H with depth Z. Given the camera focal length f, we can infer the depth with the following equation:

$$Z = f \frac{H}{h}. \tag{3}$$

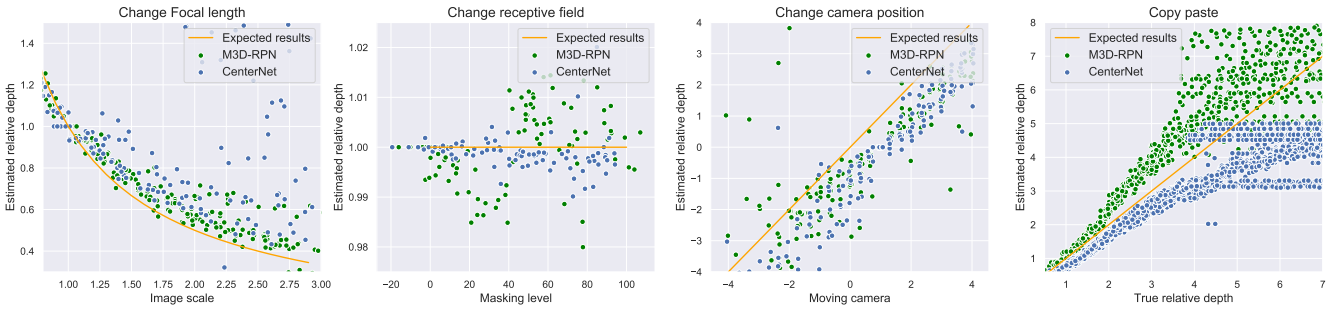


Figure 3. Empirical analysis of anchor-based (M3D-RPN) and anchor-free (CenterNet) detectors under geometric manipulations. As displayed, their object depth estimation modules are not robust under different geometric manipulations.

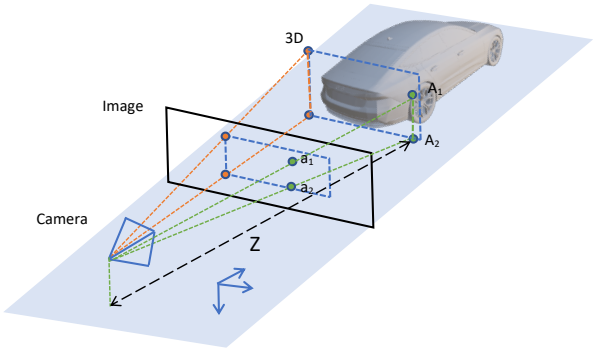


Figure 4. Visualization of the geometric relationships between depth with objects' apparent size and position.

The intuition behind this visual cue is that objects that are farther away from the camera tend to have smaller apparent sizes.

Except for the apparent size, depth also can be recovered by localizing the vertical position of the object's ground contact points. Given the camera height Y_{cam} relative to the ground and the height of the horizon line v_h in the image, depth can be obtained by:

$$Z = f\frac{Y_{cam}}{v - v_h}. \tag{4}$$

In Figure 4, we visualize the relationship of vertical position with depth in the green triangle, where point A_1 represents one of the horizon lines projected in the object, point A_2 represents one of the object's ground contact points. The points that A_1 and A_2 projected in image coordinate are a_1 and a_2, whose vertical positions are v and v_h, respectively. The intuition behind this visual cue is that an object closer to the camera would have a lower vertical position in the image. Although the two geometric relationships require several assumptions, most of them are satisfied in autonomous driving environments. We refer readers to [10] for a more thorough review of the pictorial cues.

4. Analysis based on Geometric manipulations

In this section, we first present three image-level and one instance-level geometric manipulation techniques to disturb the aforementioned visual cues in the image. Then we introduce the robustness analysis based on the presented manipulation techniques. KITTI validation set [4] is adopted to conduct the empirical analysis.

4.1. Image-level

Random Scale. Random scale resizes the image with a specific scale, which corresponds to shifting the camera focal length in the imaging process. Under the same camera intrinsic in the pinhole camera, image scaling also can be treated as moving all the objects towards a relative scale. For a scaling factor s, the location change in 3D space is formulated as:

$$P_{new} = \begin{pmatrix} 1 & 0 & (1-s)\frac{c_u}{f} & 0 \\ 0 & 1 & (1-s)\frac{c_v}{f} & 0 \\ 0 & 0 & s & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} P. \tag{5}$$

We evaluate if the detector can identify this location change when the objects are scaled with different sizes.

Random Crop. The second manipulation is randomly cropping the image, which corresponds to changing the camera receptive field. To preserve the pictorial visual cue during manipulating, we pad the cropped region to keep the objects' vertical position in random scale. As demonstrated by Md et al [19], neural networks would utilize the padding region to extract the position information. We evaluate if the detectors are robust under this manipulation technique by checking if they can estimate consistent depth after cropping and padding.

Moving Camera The third manipulation is moving the camera's location, which equals to taking images from a different location. In this manipulation, we change the camera's location in the Z coordinate, where the object-to-

Table 1. Experimental results of anchor-based (M3D-RPN) and anchor-free (CenterNet) detectors under different manipulation techniques. Except the baseline, we replace the ground-truth with estimated results. For example, "Depth*" denotes replacing the ground truth depth with the estimation and setting all other components with ground truth. (Results of $AP|_{40}$ with IoU$\geq$0.5 on car (easy) are reported.)

Network	Method	Base	Depth*	Dim*	Pos*
M3D-RPN	Origin	54.3	55.6	99.1	98.9
	Random scale	31.3	34.8	98.2	98.4
	Random crop	40.2	42.3	95.6	96.7
	Moving cam	25.6	29.4	91.0	89.3
	Copy-paste	35.2	43.3	83.4	97.3
CenterNet	Origin	49.9	50.6	98.9	99.0
	Random scale	23.3	27.3	97.8	97.9
	Random crop	38.8	41.0	94.7	94.2
	Moving cam	25.9	28.8	91.7	88.6
	Copy-paste	36.2	42.3	82.0	97.0

camera distance should be shifted with an offset d:

$$P_{new} = P + [0, 0, d, 0]^T. \tag{6}$$

To generate corresponding images, we adopt a depth estimation network: DORN [14] to regress the location of each pixel. With the manipulated images, we evaluate if the detectors cannot identify the offset in the generated image.

4.2. Instance-level: Copy-paste

In addition to the image-level manipulation, we further provide an instance-level manipulation: copy-paste. Copy-paste is widely used in 2D instance segmentation, where several approaches are proposed to preserve the semantic context during pasting. However, most of the approaches [12, 13, 39] ignore the geometric relationships, destroying the pictorial visual cues during manipulation. We first provide a geometric consistent copy-paste to study the robustness of the detectors and then introduce two geometric violated copy-paste to study how neural networks estimate depth.

Geometric consistent copy-paste This manipulation is split into two stages: (1) what to copy and (2) how to paste.

What to Copy. In this stage, we first collect an instance database from the training data. Specifically, we crop the objects of interest in the training images by a pre-trained instance segmentation model [44]. To filter out outliers, we remove the instances that are truncated or have low visibility. Since the two pictorial visual cues we studied assume the ground is flat, we further remove the unqualified objects by comparing their corresponding vanish points as in [10].

How to Paste. In the pasting stage, we sample depth in a valid region (*i.e.,* [0m, 60m]) and then calculate the corresponding bounding box size and the pasting location based on Equation 3 and 4. The whole pipeline of pasting is described in Algorithm 1.

Algorithm 1 Procedure of copy-paste augmentation.

1: **Input**:Original object with ground truth: $[(u_1, v_1, u_2, v_2), (X, Y, Z), (W, H, L), \theta]$.
2: Sample a new scene for pasting.
3: Sample new depth $\hat{Z}$.
4: Set the orientation $\hat{\theta} = \theta$.
5: Set the location of $\hat{X} = X\frac{\hat{Z}}{Z}$.
6: Compute the location of $\hat{Y}$ based on Eq 4.
7: Set the dimension as $\hat{W} = W, \hat{H} = H, \hat{L} = L$.
8: Generate a 2D bounding box $(\hat{u_1}, \hat{v_1}, \hat{u_2}, \hat{v_2})$ by projecting the corner points in 3D boxes to the image.
9: **if** the new instances does not satisfy the Eq 3. **then**
10: Go back to Step 2.
11: **end if**
12: **Output**: the new instances with ground truth: $[(\hat{u_1}, \hat{v_1}, \hat{u_2}, \hat{v_2}), (\hat{X}, \hat{Y}, \hat{Z}), (\hat{W}, \hat{H}, \hat{L}), \hat{\theta}]$.

Note that to simplify the generation process, we fix the object yaw and alpha angle during pasting. Step 4 and Step 5 display how we use the geometric relationship to determine the objects' apparent size and vertical position. For the geometry violated manipulation, the value in step 3 and step 5 are randomly sampled. The if statement in step 9 would be false when the height of the ground plan in the origin and pasted scenes are different. Figure 5 visualizes the difference between geometry consistent and geometry violated copy-paste.

4.3. Stability under different manipulations

In Figure 3, we plot the estimated depth of the detectors for manipulated images and compare it with the expected depth to measure whether the detectors are robust against the four above-mentioned manipulations. As illustrated, while the estimated depth in anchor-based and anchor-free detectors is approximately correlated with the expected result, however, both of them suffer from a large deviation, especially for the anchor-free detector. To further evaluate if the detectors can capture the variation of each visual cue and learn consistent geometric relationships, we report the mAP with the prediction of depth, 3D dimension and position in Table 1. As illustrated, the *base* version denotes the overall mAP with the estimation results. The versions of *depth**, *dim** and *pos** mark the mAP with the estimated depth, dimension and position offset respectively, while leaving the other components the same as the ground truth. We draw the following observations: 1) In the origin setting, the performance drop in *depth** is larger than *dim** and *pos**, showing that the depth recovery is more challenging; 2) Both detectors suffer from a significant performance drop under the four kinds of manipulations, especially for the anchor-free

detector; 3) For the results of *dim** and *pos**, they almost achieve 100% mAP, showing that the detectors accurately estimate the dimensions and positions of the objects, even in the manipulated image. However, the accuracy in *depth** is much less than 100%, indicating that the detectors cannot capture consistent geometry relationships under the manipulations; 4) Unlike the phenomenon in the image-level manipulation, detectors are unable to accurately regress the objects' dimensions for the inserted objects.

5. Geometry-aware data augmentation

After diagnosing the geometric inconsistency of the detectors, we convert the manipulations into geometric consistent augmentation approaches to enhance this consistency.

Random Scale As aforementioned in Section 4.1, we distort the camera focal length to generate the image with scales from 0.8 to 1.2. Although images with different scales are generated, cameras' intrinsic may be inconsistent with the testing data, which would be harmful to the testing performance. To customize the detectors with this augmentation method, we disentangle the training objective of depth from Z to a camera intrinsic irrelevant $\frac{Z}{f}$. During inference, we recover the depth by timing $\frac{Z}{f}$ with the corresponding camera focal length. For the other 3D metrics, we fix them as the original value, because they are consistent under different image scales.

Random Crop As discussed in Section 4.1, we adopt a crop-then-pad operation to make sure the geometric cue is consistent during training and inference. We randomly cropped out 25% of the region and adopt a zero-padding to fill the image in the vertical direction.

Moving camera Regarding moving the camera, we randomly move the camera in the Z direction with a range from -5m to 5m. For the coordinate conversion in the 2D and 3D coordinates, we adopt the same operation as in Section 4.1. To simplify the augmentation process, we do not adopt sophisticated novel view synthesis models, while leveraging neural networks to convert the pixel in the origin view to the target view. For the pixels that cannot find the corresponding pixel in the source view, we fill them by the nearest neighbor pixel.

Copy-Paste For the copy-paste augmentation technique, we adopt the geometric-consistent version as discussed in Section 4.2. As visualized in Figure 5, the apparent size and vertical position are matched with ground truth depth after considering the geometry relationships.

6. Experiments

We first introduce the experimental setup, including evaluation benchmarks, metrics, and our implementation details. Then, we present and analyse the results of our experiments. In addition, we verify the effectiveness of the proposed augmentation techniques in label-efficient settings.

6.1. Experimental setup

We evaluate the effectiveness of the proposed data augmentation approaches on the KITTI [16] and nuScenes [2] 3D object detection benchmarks.

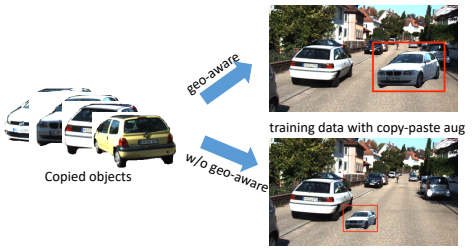


Figure 5. Visualization of copy-paste data augmentation with and without geometry-aware consideration.

KITTI [16] consists of 7,481 training frames and 7,518 test frames with 80,256 annotated 3D bounding boxes. For fair comparisons, we follow prior work [3, 4] and split the training data into training and validation subsets. We evaluate the effectiveness of the proposed components on the validation set and evaluate the final model on the test set.

nuScenes [2] is a recently released autonomous driving dataset. It contains up to 40K annotated key frames from 6 cameras with 4 different scene locations. Compared with the KITTI dataset, it has 7x as many annotations with 23 different object classes. The dataset is split into 700 video sequences for training, 150 for validation, and 150 for testing. Due to the limited computation resources, we train the detectors on the training subset and evaluate the performance on the official validation subset.

Evaluation metrics In the KITTI dataset, we follow the official protocol [16] and adopt the $AP|_{40}$ evaluation metrics on both bird-eye view (BEV) and 3D bounding box estimation tasks. The evaluation is conducted separately based on the difficulty levels (Easy, Moderate, and Hard) and object categories (Car, Pedestrian, and Cyclist). In the nuScenes dataset, we adopt the provided [2] evaluation metrics from the perspective of entire boxes (mAP), translation (mATE), size (mASE), *etc*.

Implementation details As described in Section 3.1, our experiments are conducted based on CenterNet [49] and M3D-RPN [1]. We use the modified DLA-34 [49] (CenterNet) and DesNet-141 [18] (M3D-RPN) as detectors' backbone and initialize the parameters with ImageNet [8] pre-trained weights. Before applying the proposed augmentation techniques, we first pad the images in KITTI to the size of 1280×384 and downsample the images in nuScenes to half of the resolution. Regarding optimization, we train the two detectors with 90 epochs in the KITTI dataset and 12 epochs in the nuScenes dataset. We adopt the AdamW [29] optimizer for training and set the initial learning rate as 3e-

4. The detailed descriptions of the experimental setup are provided in the supplementary material.

6.2. Individual and composite effect of the proposed augmentation methods

To evaluate the effectiveness of our geometry-aware strategy, we first conduct experiments with different augmentation strategies for comparison. In the vanilla strategy, we adopt the horizontal flip augmentation in both 2D and 3D tasks. For the other augmentation techniques (random scale, random crop, and copy-paste), we only adopt them in the 2D task, because the vanilla operations violate the geometric constraints and cannot directly get the corresponding 3D ground-truth. In our geometry-aware scheme, we add the coordinate-based augmentation to 3D task with the proposed geometric-preserving operations, where the 3D related ground-truth are calculated as in Section 5. Table 2 displays the comparison results with anchor-based (M3D-RPN) and anchor-free (CenterNet) detectors. As illustrated, the geometry-aware scheme consistently improves the vanilla strategy and the combination of four augmentation techniques yields consistently performance boosting with 5.99%/4.79%, 4.96%/3.75%, and 3.85%/2.35% of the three settings on the two detectors, respectively. We also observe that the improvement of "vanilla aug" over "w/o aug" is limited. The potential reason is that the performance of monocular 3D detection heavily relies on the accuracy of depth recovery, while vanilla augmentation destroys the pictorial visual cue for recovery.

Table 2. Comparison among different augmentation strategies on the KITTI validation dataset. $AP|_{40}$ of 3d bounding box on the Car category are reported.

Method	Setting	Easy	Mod	Hard
M3D-RPN	W/o aug	17.45	10.03	9.42
	Vanilla aug	18.21	11.28	9.56
	+ Random scale	22.06	15.43	12.04
	+ Random crop	20.91	14.42	11.60
	+ Moving cam	21.73	14.56	11.37
	+ Copy-paste	22.63	15.94	12.61
	All aug	**23.42**	**16.24**	**13.41**
CenterNet	W/o aug	18.74	13.21	10.80
	Vanilla aug	20.16	13.49	11.95
	+ Random scale	22.46	15.60	13.57
	+ Random crop	22.63	16.02	13.21
	+ Moving cam	21.34	15.10	12.92
	+ Copy-paste	22.23	15.47	13.24
	All aug	**24.53**	**17.23**	**14.32**

6.3. Results on the KITTI test set

In Table 3, we present the comparison of the proposed augmentation enhanced detectors with state-of-the-art methods on the KITTI test set. Quantitatively, the two baseline approaches with vanilla augmentation already achieve comparable results in each setting. Powered by the proposed geometry-aware augmentation, we outperform the baseline with 3.89%/4.00%, 3.03%/2.05%, and 2.00%/1.76% of the three different difficulties in the 3D task. For the anchor-based detectors, we outperform the state-of-the-art approach DDMP-3D [41] a large margin while keeping a low running time. For the anchor-free detector, we achieve almost 2% improvement over the state-of-the-art method MonoEF [50].

6.4. Results on the nuScenes dataset

Except for the KITTI dataset, we also evaluate the proposed augmentation techniques on the nuScenes dataset. Table 4 presents the experimental results of the modified CenterNet on nuScenes validation set. Although nuScenes contains more training instances, the proposed geometry-aware augmentation strategies still improve the vanilla setting in different evaluation metrics. Typically, regarding the most important mAP metric, the geometry-aware strategy outperforms the vanilla version over 3.89%.

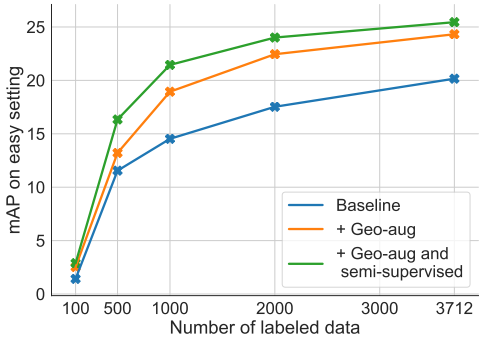


Figure 6. Experimental results of our geometric data augmentation on the semi-supervised learning setting.

6.5. On the benefit of the proposed augmentation methods to label-efficient settings

It is worth mentioning that our proposed augmentation techniques are orthogonal to which setting it is conducted. In this part, besides supervised 3D detection, we also investigate the effectiveness of our proposed augmentation in label-efficient settings that include semi-supervised and cross-domain scenarios.

Semi-supervised training. In semi-supervised learning, one of the common practices [9, 20] is to regularize the output consistency of the unlabeled data under image manipulations. As for monocular 3D detection, we utilize our proposed augmentation to generate different views of unlabeled data and then feed them into mean-teacher architecture [9, 20] to regularize the geometric consistency of their outputs. In terms of the different levels of manipulation,

Table 3. Experimental results of the "Car" class on the KITTI Test set. The best results are marked with **bold**.

	Setting	3D (Test)			BEV (Test)			Running time (ms)
		Easy	Mod	Hard	Easy	Mod	Hard	
Anchor-based	M3DSSD	17.51	11.46	8.98	24.15	15.93	12.11	-
	Mono R-CNN	18.36	12.65	10.03	25.48	18.11	14.10	70
	GrooMed-NMS	18.10	12.32	9.65	26.19	18.27	14.05	-
	Kinemantic	19.07	12.72	9.17	26.69	17.52	13.10	-
	MonoRun	19.65	12.30	10.58	27.94	17.34	15.24	70
	DDMP-3D	19.71	12.78	9.80	28.08	17.89	13.44	-
	CaDDN	19.17	13.41	11.46	27.94	18.01	17.19	630
	M3D-RPN (vanilla aug)	16.45	11.24	10.02	26.53	17.78	12.11	40
	M3D-RPN (geo aug)	**20.34**	**14.27**	**12.02**	**28.15**	**19.67**	**16.73**	40
Anchor-free	MonoFlex	19.94	13.89	12.07	28.23	19.75	16.89	30
	MonoEF	21.29	13.87	11.71	29.03	19.70	17.26	30
	AutoShape	22.47	14.17	11.36	30.66	20.08	15.59	50
	Monodle	17.23	12.26	10.29	27.94	17.34	15.24	40
	CenterNet (vanilla aug)	19.41	13.21	11.04	27.89	19.24	15.53	30
	CenterNet (geo aug)	**23.41**	**15.26**	**12.80**	**31.58**	**20.75**	**17.66**	30

Table 4. Experimental results of the anchor-free detector on the nuScenes validation set.

Setting	mAP↑	mATE↓	mASE↓	NDS↑
Vanilla aug	33.2	0.69	0.28	38.4
Geo aug	34.5	0.68	0.27	39.4

Table 5. Cross-domain evaluation between different augmentation methods with the anchor-free detector. Results of car on the KITTI (easy with 3D mAP) and nuScenes datasets (mAP) are reported.

Training data	Setting	KITTI	nuScenes
KITTI	Vanilla aug	20.16	10.23
	Geo aug	24.53	19.40

the regularization requires detectors to estimate consistent object dimension and yaw angle and predict depth that satisfied the geometric relationships.

We conduct this case study on the KITTI dataset by using the "Eigen-clean" split [33] with 14,940 images as the unlabeled subset and the training split as the labeled subset. We provide the detailed setup of the mean-teacher framework on the supplementary material. Figure 6 shows the detection performance with different numbers of labeled data. Compared with the "baseline" that adopts the vanilla augmentation, the version with geometry-aware data augmentation obtains significant improvements when 500∼1500 labeled data are sampled. Furthermore, when semi-supervised training is conducted with the unlabeled data, it achieves higher performance over the baseline version. This superior results demonstrate the potential of our augmentation techniques to reduce the labeling budget.

Cross-domain evaluation. As stated in Section 5, the geometric manipulations correspond to the shift of the camera configurations. We adopt a cross-domain evaluation to evaluate if the proposed augmentation techniques can enhance the detectors' robustness in real-scenario camera configuration shifts. Specifically, we conduct a KITTI to nuScenes evaluation, where the models are trained on the source domain (KITTI) and tested in the unseen target domain (nuScenes). On the KITTI and nuScenes datasets, the cameras' focal length and their receptive field are different. As shown in Table 5, the augmentation enhanced detector not only outperforms baseline in the in-domain scenario but also shows better robustness in the cross-domain situation.

7. Conclusion and Discussion

In this work, we diagnosed the instability issues of monocular detectors under geometric shifts. To alleviate the geometric inconsistency issues observed in the diagnosis, we proposed diverse augmentation techniques for regularizing the monocular object detectors. Our work provides a new way to improve the 3D detection performance by generating more training data with preserving the geometric properties. With more diverse training data, the augmentation methods yield consistently improvement over state-of-the-art approaches on the KITTI and nuScenes datasets.

Except for the simple image perturbations, sophisticated augmentation techniques have already emerged in 2D object detection and 3D scene understanding for improving model robustness (*e.g.,* mixup, novel view synthesis, sim-to-real, adversarial example, *etc.*). On the other hand, monocular 3D object detection also has its robustness issues (*e.g.,* the perturbation of camera pitch and roll angle, occlusion, *etc.*), which could be alleviated by customized data augmentation methods. We hope this paper will provide a baseline setup for future work in leveraging augmentation methods to enhance monocular 3D object detection.

References

[1] Garrick Brazil and Xiaoming Liu. M3d-rpn: Monocular 3d region proposal network for object detection. In *ICCV*, 2019. 1, 2, 3, 6

[2] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020. 6

[3] Xiaozhi Chen, Kaustav Kundu, Ziyu Zhang, Huimin Ma, Sanja Fidler, and Raquel Urtasun. Monocular 3d object detection for autonomous driving. In *CVPR*, 2016. 1, 2, 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. In *NeurIPS*, 2015. 4, 6

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

[6] Shuyang Cheng, Zhaoqi Leng, Ekin Dogus Cubuk, Barret Zoph, Chunyan Bai, Jiquan Ngiam, Yang Song, Benjamin Caine, Vijay Vasudevan, Congcong Li, Quoc V. Le, Jonathon Shlens, and Dragomir Anguelov. Improving 3d object detection through progressive population based augmentation. In *ECCV*, 2020. 3

[7] Jaeseok Choi, Yeji Song, and Nojun Kwak. Part-aware data augmentation for 3d object detection in point cloud. *arXiv preprint arXiv:2007.13373*, 2020. 3

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 6

[9] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *CVPR*, 2021. 7

[10] Tom van Dijk and Guido de Croon. How do neural networks see depth in single images? In *ICCV*, 2019. 2, 3, 4, 5

[11] Mingyu Ding, Yuqi Huo, Hongwei Yi, Zhe Wang, Jianping Shi, Zhiwu Lu, and Ping Luo. Learning depth-guided convolutions for monocular 3d object detection. In *CVPR*, 2020. 2

[12] Nikita Dvornik, Julien Mairal, and Cordelia Schmid. Modeling visual context is key to augmenting object detection datasets. In *ECCV*, 2018. 3, 5

[13] Hao-Shu Fang, Jianhua Sun, Runzhong Wang, Minghao Gou, Yong-Lu Li, and Cewu Lu. Instaboost: Boosting instance segmentation via probability map guided copy-pasting. In *ICCV*, 2019. 3, 5

[14] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *CVPR*, 2018. 3, 5

[15] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013. 3

[16] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 6

[17] James J. Gibson. *The perception of the visual world*. Hooughton Mifflin Boston, 1950. 3

[18] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, 2017. 6

[19] Md Amirul Islam, Sen Jia, and Neil D. B. Bruce. How much position information do convolutional neural networks encode? In *ICLR*, 2020. 4

[20] Jisoo Jeong, Seungeui Lee, Jeesoo Kim, and Nojun Kwak. Consistency-based semi-supervised learning for object detection. In *NeurIPS*, 2019. 7

[21] 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*, 2019. 3

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

[23] Peixuan Li and Huaici Zhao. Monocular 3d detection with geometric constraint embedding and semi-supervised training. *IEEE Robotics and Automation Letters*, 6:5565–5572, 2021. 2

[24] Peixuan Li, Huaici Zhao, Pengfei Liu, and Feidao Cao. Rtm3d: Real-time monocular 3d detection from object keypoints for autonomous driving. In *ECCV*, 2020. 1, 2, 3

[25] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 3

[26] Lijie Liu, Chufan Wu, Jiwen Lu, Lingxi Xie, Jie Zhou, and Qi Tian. Reinforced axial refinement network for monocular 3d object detection. In *ECCV*, 2020. 3

[27] Zechen Liu, Zizhang Wu, and Roland Tóth. Smoke: Single-stage monocular 3d object detection via keypoint estimation. In *CVPRW*, 2020. 3

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

[29] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 6

[30] 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*, 2021. 2

[31] Shujie Luo, Hang Dai, Ling Shao, and Yong Ding. M3dssd: Monocular 3d single stage object detector. In *CVPR*, 2021. 2

[32] 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*, 2021. 2, 3

[33] Dennis Park, Rares Ambrus, Vitor Guizilini, Jie Li, and Adrien Gaidon. Is pseudo-lidar needed for monocular 3d object detection? In *ICCV*, 2021. 2, 8

[34] Zengyi Qin, Jinglu Wang, and Yan Lu. Monogrnet: A geometric reasoning network for monocular 3d object localization. In *AAAI*, 2019. 2

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

[36] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 2, 3

[37] 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*, 2021. 2

[38] Andrea Simonelli, Samuel Rota Rota Bulò, Lorenzo Porzi, Manuel López-Antequera, and Peter Kontschieder. Disentangling monocular 3d object detection. In *ICCV*, 2019. 2, 3

[39] Hao Wang, Qilong Wang, Fan Yang, Weiqi Zhang, and Wangmeng Zuo. Data augmentation for object detection via progressive and selective instance-switching. *arXiv preprint arXiv:1906.00358*, 2019. 3, 5

[40] Kaixuan Wang and Shaojie Shen. MVDepthNet: real-time multiview depth estimation neural network. In *3DV*, 2018. 3

[41] 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*, 2021. 3, 7

[42] Li Wang, Li Zhang, Yi Zhu, Zhi Zhang, Tong He, Mu Li, and Xiangyang Xue. Progressive coordinate transforms for monocular 3d object detection. In *NeurIPS*, 2021. 3

[43] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q. Weinberger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *CVPR*, 2019. 3

[44] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 5

[45] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018. 3

[46] Yurong You, Yan Wang, Wei-Lun Chao, Divyansh Garg, Geoff Pleiss, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar++: Accurate depth for 3d object detection in autonomous driving. *arXiv preprint arXiv:1906.06310*, 2019. 3

[47] Wenwei Zhang, Zhe Wang, and Chen Change Loy. Multi-modality cut and paste for 3d object detection. *arXiv preprint arXiv:2012.12741*, 2020. 3

[48] Yunpeng Zhang, Jiwen Lu, and Jie Zhou. Objects are different: Flexible monocular 3d object detection. In *CVPR*, 2021. 2, 3

[49] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 2, 3, 6

[50] Yunsong Zhou, Yuan He, Hongzi Zhu, Cheng Wang, Hongyang Li, and Qinhong Jiang. Monocular 3d object detection: An extrinsic parameter free approach. In *CVPR*, 2021. 2, 7

[51] Barret Zoph, Ekin D. Cubuk, Golnaz Ghiasi, Tsung-Yi Lin, Jonathon Shlens, and Quoc V. Le. Learning data augmentation strategies for object detection. In *ECCV*, 2020. 3