Correlation Field for Boosting 3D Object Detection in Structured Scenes

Jianhua Sun, Hao-Shu Fang, Xianghui Zhu, Jiefeng Li, Cewu Lu*

Shanghai Jiao Tong University
{gothic, hg1216, ljf_likit, lucewu}@sjtu.edu.cn, fhaoshu@gmail.com

Abstract

Data augmentation is an efficient way to elevate 3D object detection performance. In this paper, we propose a simple but effective online crop-and-paste data augmentation pipeline for structured 3D point cloud scenes, named **CorrelaBoost**. Observing that 3D objects should have reasonable relative positions in a structured scene because of the objects' functionalities and natural relationships, we express this correlation as a kind of interactive force. An energy field called Correlation Field can be calculated correspondingly across the whole 3D space. According to the Correlation Field, we propose two data augmentation strategies to explore highly congruent positions that a designated object may be pasted to: 1) Category Consistent Exchanging and 2) Energy Optimized Transformation. We conduct exhaustive experiments on various popular benchmarks with different detection frameworks and the results illustrate that our method brings huge free-lunch improvement and significantly outperforms state-of-the-art approaches in terms of data augmentation. It is worth noting that the performance of VoteNet with mAP@0.5 is improved by **7.7** on ScanNetV2 dataset and **5.0** on SUN RGB-D dataset. Our method is simple to implement and increases few computational overhead.

Introduction

3D object detection (Shi, Wang, and Li 2019; Qi et al. 2019; Lang et al. 2019; Shi et al. 2020; Chen et al. 2020; You et al. 2020) exploits the position and recognizes the category of objects in a 3D scene, and has numerous applications in downstream tasks (Li et al. 2021b,a; Fang et al. 2020; Mahler et al. 2017). Although they show promising detection results, a large amount of training data is needed to cover volatile cases in the test set and guarantee the performance. However, high-quality 3D data is difficult to obtain.

To tackle this problem, traditional methods such as global transformation and noise addition are widely used by researchers for better training performance (Yan, Mao, and Li 2018; Shi, Wang, and Li 2019; Lang et al. 2019). In recent years, instance-level crop-and-paste data augmentation methods (Fang et al. 2019; Dwibedi, Misra, and Hebert 2017) are adopted for 3D object detection gradually (Yan, Mao, and Li 2018; Lang et al. 2019; Shi, Wang, and Li 2019). They first crop the objects according to ground truth bounding boxes and then paste them into a 3D scene after a collision test. However, such process does not take the functional relationships between objects into consideration when finding a position to paste. This omission may often result in functional inconsistency in a highly structured scene, between objects that are newly pasted in and other objects that originally exist in the scene. (e.g. a bathtub may be pasted next to a bed according to (Yan, Mao, and Li 2018).)

In this paper, we aim to explore highly probable locations where a given object will naturally appear in a structured 3D scene by investigating the functional relationship between objects. We look back to the area of scene interactions modeling, from which we get inspiration for a better-refined positioning strategy. Previous works (Helbing and Molnar 1998; Dan, Todorovic, and Zhu 2013) abstract a force to model the degree of attraction or repulsion between functional objects in order to calculate reasonable relative positions for them. Similarly, we assume the functional relationship between a pair of objects can be represented by a kind of force, and thus an energy field will be implied between an object pair, which we call Correlation Field. Further, these fields can be superimposed together to reflect relationships between multiple objects (see Fig. 1). In this aspect, a functionally consistent scene is equivalent to the situation that all the objects achieve a stable state in the Correlation Field and the energy of the field is minimal.

With the guidance of Correlation Field, we propose a crop-and-paste augmentation pipeline called CorrelaBoost to explore positions that a designated object may be pasted to coherently, including two strategies. **1) Category Consistent Exchanging:** Observing that objects of the same category often share similar functional relationships and interactions with others, we generate various training samples by exchanging objects in the same category without compromising the stability of the Correlation Field. **2) Energy Optimized Transformation:** we define a probability map for the cropped object on a given 3D scene and link this to its Correlation Field, where a lower energy refers to a higher probability to paste. To this end, by sampling locations with high probabilities, we can find proper pasting positions through the whole scene with little time consumption.

To sum up, the contributions of this paper are as follows:

*Cewu Lu is the corresponding author.

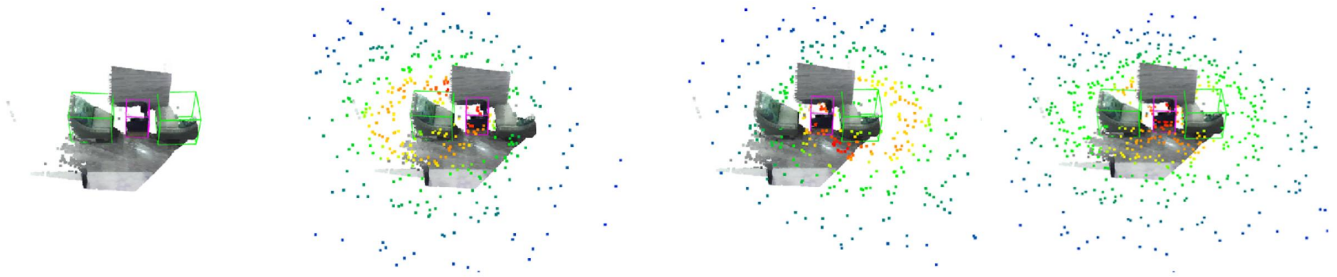


Figure 1: Overview of Correlation Field. Figures from left to right are: a given scene where objects are annotated by bounding boxes, Correlation Field from left sofa to the table, Correlation Field from right sofa to the table, and the superimposed field. The target object is marked by purple. Colored dots refer to the strength of Correlation Field, where red denotes low and blue denotes high.

i) we first introduce functional relationships of objects to explore proper positions for crop-and-paste 3D data augmentation, ii) we design a novel physical model named Correlation Field to model functional relationships between objects, and iii) we propose an efficient data augmentation pipeline including two strategies, interpreting data augmentation to an optimization problem under the Correlation Field framework. Comprehensive experiments are conducted on multiple 3D detection benchmarks with different typical frameworks to show that our approach can deliver improvement across different datasets and frameworks and outperforms state-of-the-art approaches in terms of data augmentation. Our online data augmentation is easy to implement and increases few computational overhead without using heavy deep neural networks (Zhou, While, and Kalogerakis 2019).

Related Work

3D object detection. 3D object detection is a task to locate and recognize objects in 3D scenes. Numerous studies have been carried out in this field. Some researches (Chabot et al. 2017; Chen et al. 2016, 2015; Mousavian et al. 2017) try to use 2D images to explore positions of 3D objects. These methods only need easily accessible 2D data to work. However, because of the lack of depth information, the results are easily affected by different forms of noise.

To this end, more detection approaches based on 3D point cloud data are proposed to fully utilize 3D information. One early thought for detection on 3D point cloud data is to migrate 2D detection methods to 3D detection. (Chen et al. 2017; Ku et al. 2018; Yang, Luo, and Urtasun 2018; Yang, Liang, and Urtasun 2018; Liang et al. 2018) reconstruct 2D bird's eye view with 3D point clouds and use 2D CNNs to extract features for 3D bounding box generation. But these methods are still facing information loss when constructing bird's eye view images or voxels.

Another way to solve this problem is directly working on point clouds (Shi, Wang, and Li 2019; Shi et al. 2020; Qi et al. 2019; Xie et al. 2020a; Zhang et al. 2020; Cheng et al. 2021). VoteNet (Qi et al. 2019) uses a well-designed deep hough voting network to estimate not only oriented 3D bounding boxes but also semantic classes of objects directly from point clouds. Some further research (Xie et al. 2020a; Zhang et al. 2020; Cheng et al. 2021) enhance VoteNet with extra modules and achieve better performance.

Instance-level data augmentation for 3D detection. Traditional 3D data augmentation on point clouds such as global scaling/rotation and point-level random jittering are widely used to prevent overfitting, and some research (Cheng et al. 2020) introduces the idea of AutoAugment (Cubuk et al. 2019) to automatically search for improved data augmentation policies from traditional strategies.

To fully exploit the potential of data augmentation, researchers (Yan, Mao, and Li 2018; Lang et al. 2019; Hu et al. 2020; Fang et al. 2021) begin to utilize instance-level labels such as bounding boxes and instance ids for better performance. Yan et al. (Yan, Mao, and Li 2018) create a database of target objects from training dataset, randomly select several ground truths from this database and introduce them into the current training point cloud via concatenation. However, they ignore the functional consistency between the pasted object and the original scene.

A similar line of related work is indoor scene synthesis (Zhou, While, and Kalogerakis 2019; Fisher and Hanrahan 2010; Li et al. 2019; Wang et al. 2018), where they retrieve models from a 3D database and place the furniture in a reasonable manner. However, a **manually designated position** is required as precondition and then they analyze the most likely object category to appear in the given location, which contradicts to automatically finding a suitable pasting location for cropped object in crop-and-paste data augmentation. Moreover, our work directly augments the point cloud in data level and does not require 3D models of the objects, observing the deep learning architectures used in these works are time consuming for online data augmentation.

Method

Overview

Generally, given a 3D scene S and a designated object instance I, a crop-and-paste augmentation process can be defined as

$$I^{'} = \mathbf{T}(I|S) \tag{1}$$

where function $\mathbf{T}$ refers to a transformation operated on I under the given S condition. To ensure the shape invariance of designated objects, the transformation $\mathbf{T}$ is usually defined by a 3D affine transformation

$$\begin{bmatrix} X' \\ Y' \\ Z' \end{bmatrix} = s \prod_{i=z,y,x} \begin{bmatrix} \cos\theta_i & -\sin\theta_i & 0 \\ \sin\theta_i & \cos\theta_i & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \\ t_z \end{bmatrix} \tag{2}$$

where t_x, t_y, t_z denote the coordinate shift in x, y, z-axis respectively, s denotes the scale variance and θ_x, θ_y, θ_z denote the rotation around x, y, z-axis in degrees. Assuming (x_0, y_0, z_0) is the object's original coordinate and (x, y, z) is the pasted coordinate, t_x, t_y, t_z can be calculated by $t_x = x - x_0$, $t_y = y - y_0$, $t_z = z - z_0$. Thus, transformation $\mathbf{T}$ can be uniquely determined by a 7D tuple

$$\mathcal{B} = (x, y, z, \theta_x, \theta_y, \theta_z, s) \tag{3}$$

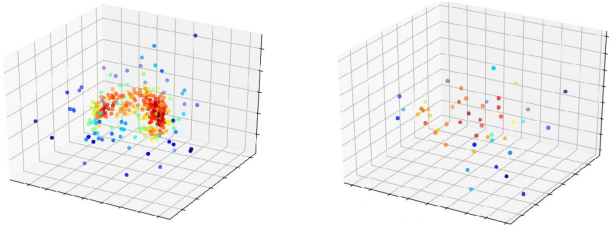
We denote each $\mathcal{B}$ as a transformation configuration. Among positioning, rotation and scaling which are three key components of $\mathcal{B}$, positioning contributes more for data augmentation as it can bring richer diversity to a scene.

To explore feasible and reasonable positions for transformation through the whole scene, an important prerequisite is that the pasted object should be functionally consistent with other objects. For example, a lamp should be on a desk rather than under it. We introduce a novel insight of Correlation Field to ensure this prerequisite. We first assume that there is an interactive force acting on each pair of objects to keep them at a proper relative position, which expresses as repulsive force when their distance is too close and attractive force vice versa. We then define the Correlation Field, an energy field induced by this force. With this formulation, proper relative positions for a pair of functional objects should be at the equilibrium points where the energy of Correlation Field is at a low level. Specifically, since every object is most likely to appear at its original location in the original dataset, the relative positions in the original dataset should be assigned as equilibrium points.

For real scenes that usually have multiple objects, the full Correlation Field can be calculated by superimposing sub fields between two objects. In this aspect, positions with low energies are able to represent places which are proper for pasting, and the position exploration problem is interpreted into a much easier optimization problem.

Following this insight, we propose two augmentation strategies. One is called *Category Consistent Exchanging*. We exchange pairs of objects in the same category with appropriate shapes and orientations. Since objects in the same category exhibit similar natural occurrence frequency and share similar functionality, this approach will still keep the energy of this system at a low level. Another is *Energy Optimized Transformation*. We sample positions with low energy according to a probability function. Experimental results in Sec. show the surprising effectiveness of these augmentation strategies.

In Sec. we introduce the formulation of the Correlation Field, and the adoption of it with two augmentation strategies will be detailed in Sec. and Sec. . Finally, we describe the whole pipeline and implementation details in Sec. .



A. nightstand--bed B. bookshelf--sofa

Figure 2: Illustration of correlation origin sets and Correlation Fields. Each plot illustrates a point set $\mathcal{P}_{(c_1,c_2)}$. Different colors indicate relative strength of Correlation Field in corresponding positions, where red denotes low and blue denotes high.

Correlation Field

Functional objects affect each other in a particular scene which can be regarded as a kind of interactive force. We generalize Correlation Field based on this force to determine how well the target object fits with others when placed in any position in the 3D scene.

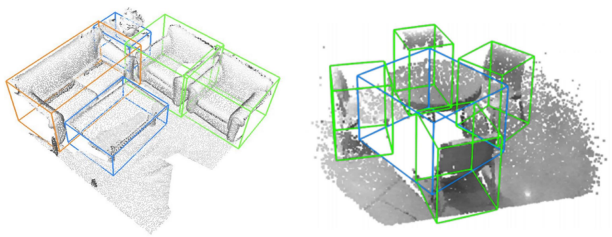
Correlation Origin Set Before we define the Correlation Field, the abstracted interactive force which ensures functional objects at proper relative positions are necessary to be first defined. Observing the property that the force should attract objects to those proper positions, we formulate the force as gravitational effects with gravitational centers at these positions. We name these gravitational centers as correlation origins p, as these positions show the inherent correlation of two objects.

In this paper, we assume that objects in the same category share similar functionality, and mainly focus on the category to identify a correlation origin set $\mathcal{P}$ for each category pair (c_1, c_2). Considering objects appear at its original location naturally in the original training dataset, positions of c_2 objects in the c_1 object coordinate system can represent correlation origins.

To this end, a correlation origin set $\mathcal{P}$ for (c_1, c_2) can be represented by

$$\mathcal{P}_{(c_1,c_2)} = \{p = \overrightarrow{\mathbf{C}_{I_j}\mathbf{C}_{I_i}} | I_i \in c_1, I_j \in c_2\} \tag{4}$$

where $\mathbf{C}_{I_k}$ refers to the geometric center (a.k.a. center of the bounding box) of object I_k $(k = i, j)$. By traversing all objects in category c_1 and c_2 through all scenes in the training set, we can obtain $\mathcal{P}_{(c_1,c_2)}$ including all correlation origins, which represents a statistical distribution of potential gravitational centers. Following such a process above, we can get correlation origin set $\mathcal{P}_{(c_i,c_j)}$ for each category pair (c_i, c_j). Fig. 2 visualizes several correlation origin sets of some typical pairs, where the beginnings of all vectors are set at the coordinate origin. Figure A reveals a strong relation that the nightstand is often near the front of the bed. For pair *bookshelf–sofa* in figure B, they do not have such a strong correlation and thus the points nearly randomly distribute in the space.



A. shape discrepancy B. orientation inconsistent

Figure 3: Special cases for Category Consistent Exchanging. Figure A demonstrates a shape discrepancy case between the sofas in green boxes and the sofa in brown box. Figure B demonstrates an orientation inconsistent case among chairs in green boxes. The illustration is from (Qi et al. 2019).

Correlation Field After obtaining correlation origins, we then define the Correlation Field. Given a particle set of weights $\{M_1, ..., M_n\}$, the gravitational field in physical world is defined as

$$E_{(x)} = -\sum_{i=1}^{n} \frac{GM_i}{R_i} \tag{5}$$

where G is a constant and R_i is the distance from position x to particle i. Draw on this expression, Correlation Field between category pair (c_1, c_2) can be formulated as

$$E_{(x,c_2|c_1)} = -\sum_{p \in \mathcal{P}_{(c_1,c_2)}} \frac{G_{(c_1,c_2)}}{R_p^{\gamma} + k} \tag{6}$$

where R_p denotes the distance from position x to correlation origin p, γ denotes a distance attenuation index for the whole dataset, $G_{(c_1,c_2)}$ is a category correlation index which is constant for each category pair, and k is a constant to balance the value when $R \to +0$. Now, given an object in a certain category c, various Correlation Field $E_{(x,\cdot|c)}$ centered on this object can be formed accordingly. Note that a Correlation Field between (c_1, c_2) is directional, where we name objects in c_1 as initiators ans c_2 as receptors.

Under this formulation, the Correlation Field has three key properties:

- **Superimposable.** Multiple individual Correlation Field on a single receptor can be superimposed together by a symmetric function.
- **Optimal initial state.** A scene in the original dataset without any augmentation is at a low energy level.
- **Category oriented.** The dependent variables of Correlation Field are object categories and distance.

Based on these three properties, we derive two different data augmentation strategies on Correlation Field.

Category Consistent Exchanging

Based on the category oriented and optimal initial state property, exchanging objects in the same category will keep the energy of full Correlation Field still at a low level. However, random pairwise exchanges may fail in some special cases. As shown in Fig. 3, sofas in figure A illustrate shape discrepancy between objects of the same kind while chairs in figure B depict an inconsistent problem in orientation.

To solve these defects, we propose a category consistent similarity $s(I_a, I_b)$ to measure the shape and orientation similarity between two objects in the same class, which can be written as

$$s(I_a, I_b) = \lambda_s \cos(\overrightarrow{\mathbf{S}}_a, \overrightarrow{\mathbf{S}}_b) + \lambda_o \cos(\overrightarrow{\mathbf{CG}}_a, \overrightarrow{\mathbf{CG}}_b) \tag{7}$$

where shape vector $\overrightarrow{\mathbf{S}} = (l, w, h)$ represents the shape of bounding box (length, width and height respectively); $\mathbf{C}$ denotes the geometric center (center of the bounding box), and $\mathbf{G}$ denotes the gravity center (center of point cloud) of object I; λ_s, λ_o are importance weights for shape and orientation. In this manner, we use $\overrightarrow{\mathbf{S}}$ and $\overrightarrow{\mathbf{CG}}$ to indicate the shape and orientation of an object respectively, and the category consistent similarity scores the consistency of two objects in a same class.

Then we assign a mapping function $f(\cdot)$ to map the category consistent similarity to the probability of exchanging the object, written as

$$f(s) = \log(s) \tag{8}$$

As a weighted choice instead of random selection is applied, our method will show a tendency to exchange objects with similar shape and orientation.

Energy Optimized Transformation

Former exchanging strategy provides extensive different permutations of objects, but it actually exploits only a small part of Correlation Field and fails to explore extra reasonable positions. In this section, we propose a complementary strategy called Energy Optimized Transformation to fill this gap, and show the most essential capability of our proposed Correlation Field.

Following Sec. , Correlation Fields $E_{(x,\cdot|\cdot)}$ between any pair of categories can be first calculated and saved when preprocessing. In the process of data augmentation, after giving a scene S and object I of category c, an overall Correlation Field $\mathbf{E}$ for I can be superimposed by

$$\mathbf{E}(x|S, I) = \sum_{I' \in S, I' \neq I} E_{(x,c|c')} \tag{9}$$

where c' is the category of object I'.

Then we generate a 3D probability map on it by a mapping function $g(\cdot)$ which should be negative correlated. In our implementation, it is formulated as

$$g(\mathbf{E}) = \log(-\mathbf{E}) \tag{10}$$

After values in the probability map are normalized, we sample candidate positions for pasting via Monte Carlo method. Finally, we introduce a collision test to adjust the position in a small neighborhood. Such operation on Correlation Field explores a large number of potential positions for pasting and brings huge diversity for data augmentation.

Framework	Method	ScanNet		SUN RGB-D	
		mAP@0.25	mAP@0.5	mAP@0.25	mAP@0.5
VoteNet	vanilla	58.6	33.5	57.7	32.7
	exchanging	59.4	36.9	58.9	35.1
	transformation	60.1	38.8	59.8	36.3
	full	**61.0**	**41.2**	**61.0**	**37.7**
	△↑	2.4	7.7	3.3	5.0
MLCVNet	vanilla	64.5	41.4	59.8	36.3
	exchanging	65.0	44.6	60.5	37.4
	transformation	65.4	45.9	61.1	37.8
	full	**65.8**	**47.0**	**61.7**	**38.4**
	△↑	1.3	5.6	1.9	2.1
H3DNet w/o refine	vanilla	60.2	37.3	58.5	34.2
	exchanging	61.4	39.7	59.1	35.6
	transformation	62.3	40.8	59.5	36.4
	full	**63.4**	**42.3**	**59.9**	**37.1**
	△↑	3.2	5.0	1.4	2.9
H3DNet w/ refine	vanilla	67.2	48.1	60.1	39.0
	exchanging	67.6	48.7	60.5	39.9
	transformation	67.8	48.8	60.8	40.6
	full	**68.1**	**49.3**	**61.0**	**41.1**
	△↑	0.9	1.2	0.9	2.1

Table 1: Boosted 3D object detection results on both ScanNetV2 and SUN RGB-D dataset, evaluated with mAP@0.25 IoU and mAP@0.5 IoU. *vanilla* denotes the original framework, *exchanging* denotes Category Consistent Exchanging, *transformation* denotes Energy Optimized Transformation and *full* denotes the full pipeline of CorrelaBoost.

Method	ScanNet		SUN RGB-D	
	mAP@0.25	mAP@0.5	mAP@0.25	mAP@0.5
vanilla	58.6	33.5	57.7	32.7
SECOND	59.2	35.9	58.3	34.2
SGN	59.4	36.2	58.4	34.5
*PointContrast	59.2	38.0	57.5	34.8
CorrelaBoost	**61.0**	**41.2**	**61.0**	**37.7**

Table 2: Comparison with baselines on ScanNetV2 and SUN RGB-D dataset on VoteNet. *vanilla* denotes the original framework.

Pipeline and Implementation Details

Our **online** data augmentation approach can be implemented in an efficient manner. We first pre-compute the Correlation Field for each category pair. In the training process, both strategies are incorporated into the dataloader, after common online augmentation methods such as global rotation and flipping (facebookresearch 2019). We assign P_E and P_T to denote probabilities whether an object should perform category consistent exchanging and energy optimized transformation respectively. Then, we traverse all the objects in the scene and apply both strategies according to the probabilities. Finally, we introduce a collision test (Yan, Mao, and Li 2018) and fine-tune the transformation configuration to achieve better spatial coherence if a collision happens. We fill the empty area caused by crop operation with a hole-filling algorithm (Lucas 2019).

In our implementation, augmentation probabilities P_E and P_T are set as 0.5 and 0.3 respectively. For exchanging strategy, $\lambda_s = \lambda_o = 0.5$. For Correlation Field in transformation strategy, we treat γ as 2, k as 0.2 and $G_{(c_1,c_2)}$ as 1 for all category pairs for simplification.

Experiments

Datasets, Frameworks and Baselines

Datasets Performance of 3D detection models has been tested on two popular datasets, including ScanNetV2 (Dai et al. 2017) and SUN RGB-D (Song, Lichtenberg, and Xiao 2015). Experiments on these datasets prove that CorrelaBoost is effective to exploit functional relationships in structured scenes.

Frameworks We implement our augmentation method on three different 3D detection frameworks. VoteNet (Qi et al. 2019) is a classic 3D object detection pipeline with deep Hough voting, and MLCVNet (Xie et al. 2020a) improves it with three hierarchical context modules. Recently, H3DNet (Zhang et al. 2020) further push the performance with better representations and refinement modules.

Baselines We compare with previous augmentation methods, as well as an unsupervised learning baseline PointContrast (Xie et al. 2020b).

SECOND (Yan, Mao, and Li 2018). The data augmentation in SECOND is a classic one and has been widely used in 3D detection frameworks. This approach includes three elements: *sample ground truths from the database*, *object noise*, and *global rotation and scaling*.

SGN (Zhou, While, and Kalogerakis 2019). Scene Graph Net is a state-of-the-art approach which introduces graph

neural networks to augment an input 3D indoor scene with new objects matching their surroundings. Note that SGN is an only scene augmentation approach rather than data augmentation approach, as it requires a manually designated position as input. Here, we sample several positions without collision as input for SGN to generalize it into a data augmentation method.

***PointContrast (Xie et al. 2020b).** PointContrast is a state-of-the-art unsupervised learning approach to boost the performance of 3D detection frameworks, which is a much stronger baseline than approaches above. PointContrast introduces plenty of extra data to pre-train the models with a contrastive loss. Since it is not a data augmentation approach, we mark it with *.

Substantial Improvement

CorrelaBoost is evaluated on both ScanNetV2 and SUN RGB-D dataset for all three frameworks with mAP@0.25 IoU and mAP@0.5 IoU as evaluation metrics. Experimental results are in Tab. 1.

VoteNet. The performance of VoteNet could be further elevated with CorrelaBoost. In detail, we achieve 2.4 and 7.7 improvement on ScanNetV2 dataset, as well as 3.3 and 5.0 improvement on SUN RGB-D dataset for two metrics. It is worth noting that our proposed augmentation methods bring a huge boost on the more stringent metric.

MLCVNet. Similar as VoteNet, CorrelaBoost also brings amazing free-lunch improvement on both datasets. Considering MLCVNet is a much stronger network to capture implicit features of the training set, the relative lifts brought by CorrelaBoost have a reasonably narrowing comparing with VoteNet.

H3DNet. Refinement modules are used in H3DNet to fine-tune object bounding boxes locally at object-level. Since our augmentation approach operates at scene-level globally and does not change the object itself, CorrelaBoost is ineffective for the refinement module. For H3DNet without refinement, great improvement is achieved. After applying the refinement module, our approach can still achieve substantial improvement, which proves that our data augmentation is complementary with the refinement process and can be used together for better performance.

Comparison with Augmentation Baselines

In order to show the superior performance of CorrelaBoost in terms of data augmentation, we compare our method with several representative baselines, including previous state-of-the-art. Results are given in Tab. 2. It shows that our data augmentation approach can achieve better performance on both datasets and both metrics.

Further, our method also greatly outperforms PointContrast (Xie et al. 2020b), an unsupervised learning approach, in the case of boosting 3D detection performance. It is worth noting that PointContrast requires extra training data for unsupervised learning, while our method can provide free-lunch improvement without introducing any other information to the dataset.

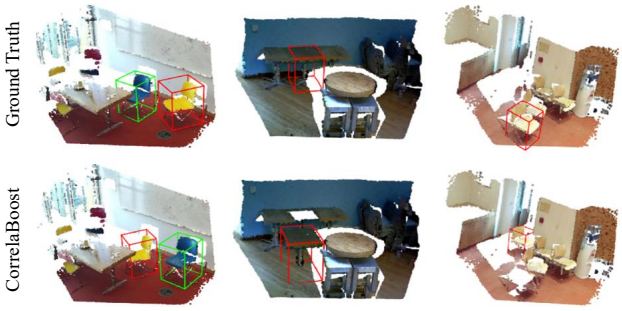


Figure 4: Visualization of 3D objects augmented by CorrelaBoost, where the top row illustrates the original scenes and the bottom row illustrates augmented scenes. The left sample is operated by Category Consistent Exchanging and others are operated by Energy Optimized Transformation. Target objects are marked by bounding boxes.

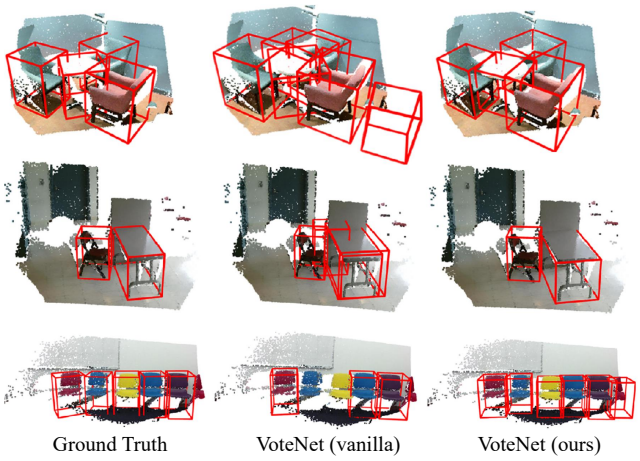


Figure 5: 3D Detection results of vanilla VoteNet (Qi et al. 2019) *vs.* VoteNet trained with CorrelaBoost. Ground truths and detection results are annotated by red bounding boxes. CorrelaBoost makes progress in both recall and precision.

Qualitative Results.

Augmented Scenes. Fig. 4 visualizes several augmented samples with CorrelaBoost. The left two samples are augmented by Category Consistent Exchanging strategy and the right two samples are operated by Energy Optimized Transformation strategy. These cases prove that our approach with Correlation Field can leverage functional relations between objects, and consequently generate diverse and realistic results.

Detection results after Augmentation. We visualize some of the detection results of vanilla VoteNet and CorrelaBoosted VoteNet (see in Fig. 5). The visualization results indicate that CorrelaBoost makes progress in higher recall and precision. For cases in the first and second row, augmented VoteNet gives more accurate regression results and is able to remove erroneously detected objects in the vanilla framework. Case in the bottom row indicates that augmented framework can identify previously unrecognized

Method	ScanNet		SUN RGB-D	
	AP_{25}	AP_{50}	AP_{25}	AP_{50}
vanilla	58.6	33.5	57.7	32.7
Random Paste	58.9	34.2	57.9	33.6
transformation	60.1	38.8	59.8	36.3
CorrelaBoost	**61.0**	**41.2**	**61.0**	**37.7**

Table 3: Comparison with random paste on ScanNetV2 and SUN RGB-D dataset on VoteNet.

λ_s	0	1	0.7	0.5	0.3	0
λ_o	0	0	0.3	0.5	0.7	1
AP_{25}	59.0	59.2	59.3	**59.4**	59.3	59.1
AP_{50}	35.5	36.3	36.7	**36.9**	36.6	36.1

Table 4: Analysis for category consistent exchanging on ScanNetV2 dataset. The reported results are from VoteNet boosted only by our exchanging strategy.

γ	2	2			1	3
k	0.20	0.10	0.15	0.25	0.20	
AP_{25}	**60.1**	59.8	60.0	60.0	60.0	59.8
AP_{50}	**38.8**	38.4	38.6	38.7	38.6	38.3

Table 5: Sensitive analysis for Correlation Field on ScanNetV2 dataset on VoteNet. The reported results are from VoteNet boosted only by our transformation strategy.

Method	workers	vanilla	CorrelaBoost	Dec.
VoteNet	4	0.79	0.90	14%
	8	0.78	0.83	6%
	16	0.78	0.83	6%
MLCVNet	4	0.74	0.86	16%
	8	0.74	0.81	9%
	16	0.74	0.80	8%

Table 6: Speed comparison on different detection frameworks. Column *vanilla* and *CorrelaBoost* record the time for training one epoch by vanilla framework and CorrelaBoost augmented framework in minute. *workers* denotes the number of workers for dataloader and *Dec.* denotes the deaccleration ratio after applying CorrelaBoost to the framework.

objects in vanilla framework, which shows a better recall. These cases also prove that our augmentation method improves the model's immunity to incomplete point clouds and background noise.

Analysis

Contribution of Two Strategies. Tab. 1 also exhibits the contribution of both strategies. When either strategy is applied alone, notable improvement is still achieved. The Energy Optimized Transformation strategy contributes more since it fully exploits reasonable pasted positions through the whole scene and brings more diversity.

Comparison with Random Paste. Random paste is a most basic crop-and-paste augmentation that first crops an object and then pastes it to a random position on ground planes without collision, and it ignores the functional relationships between objects. In order to figure out that the guidance of Correlation Field on pasted positions is effective, we compare our method with a random paste strategy. Tab. 3 shows that a random paste method brings a little improvement for the performance. After applying Correlation Field to ensure the functional consistency between objects, promising improvement can be achieved.

Analysis for Category Consistent Exchanging. To demonstrate the effectiveness of category consistent exchanging, we compare with a random pair exchanging baseline (without category consistent similarity as $\lambda_s = \lambda_o = 0$). Results in Tab. 4 show that the performance improves when the exchanging is guided by our proposed similarity. Further, a sensitive analysis for the similarity is also conducted in Tab. 4, which shows the weights of shape and orientation are not sensitive in the non-extreme range.

Sensitive Analysis for Correlation Field. We analyze both constant k and attenuation index γ in the Correlation Field to show how they influence the performance of our transformation strategy. Results are in Tab. 5. The performance is stable when k and γ are in a proper range. As k goes pretty small or γ goes large, highly probable pasted positions concentrate near Correlation Origins which harms the diversity of our transformation strategy. Note that our transformation strategy will degenerate into random paste as k goes super large or γ goes super small.

Online Speed Analysis. We discuss the time efficiency of CorrelaBoost in this paragraph. We train VoteNet and MLCVNet on ScanNetV2 dataset with batchsize 8 on RTX 2080Ti. Note that our data augmentation is online and integrated into the dataloader, we compare the time in one epoch for simplicity. Results are shown in Tab. 6, where both frameworks exhibit a similar trend in speed when the number of workers grows. With only 8 workers which is easily affordable for most cpus, the deceleration of time caused by our data augmentation can be reduced to only 6% for VoteNet.

Conclusion

This paper studies an instance-level crop-and-paste data augmentation method for better 3D detection performance. We propose a novel pipeline called CorrelaBoost to explore positions with high natural occurrence frequency of designated objects and scenes. To guarantee reasonable functional relationships among different objects, we design the Correlation Field and two augmentation strategies correspondingly. Exhaustive experiments illustrate that our method brings huge free-lunch improvement and surpasses previous data augmentation approaches. Our online augmentation can be easily implemented into existing frameworks with little cpu overhead. Currently, our approach has limitations on cluttered scenes where objects do not show significant functional relationships. We will improve this as our future work.

Acknowledgements

This work was supported by the National Key Research and Development Project of China (No. 2021ZD0110700), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), Shanghai Qi Zhi Institute, and SHEITC (2018-RGZN-02046).

References

Chabot, F.; Chaouch, M.; Rabarisoa, J.; Teuliere, C.; and Chateau, T. 2017. 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*, 2040–2049.

Chen, J.; Lei, B.; Song, Q.; Ying, H.; Chen, D. Z.; and Wu, J. 2020. A hierarchical graph network for 3D object detection on point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 392–401.

Chen, X.; Kundu, K.; Zhang, Z.; Ma, H.; Fidler, S.; and Urtasun, R. 2016. Monocular 3d object detection for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2147–2156.

Chen, X.; Kundu, K.; Zhu, Y.; Berneshawi, A. G.; Ma, H.; Fidler, S.; and Urtasun, R. 2015. 3d object proposals for accurate object class detection. In *Advances in Neural Information Processing Systems*, 424–432.

Chen, X.; Ma, H.; Wan, J.; Li, B.; and Xia, T. 2017. Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1907–1915.

Cheng, B.; Sheng, L.; Shi, S.; Yang, M.; and Xu, D. 2021. Back-tracing Representative Points for Voting-based 3D Object Detection in Point Clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8963–8972.

Cheng, S.; Leng, Z.; Cubuk, E. D.; Zoph, B.; Bai, C.; Ngiam, J.; Song, Y.; Caine, B.; Vasudevan, V.; Li, C.; et al. 2020. Improving 3d object detection through progressive population based augmentation. In *European Conference on Computer Vision*, 279–294. Springer.

Cubuk, E. D.; Zoph, B.; Mane, D.; Vasudevan, V.; and Le, Q. V. 2019. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 113–123.

Dai, A.; Chang, A. X.; Savva, M.; Halber, M.; Funkhouser, T.; and Nießner, M. 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5828–5839.

Dan, X.; Todorovic, S.; and Zhu, S. C. 2013. Inferring "Dark Matter" and "Dark Energy" from Videos. In *2013 IEEE International Conference on Computer Vision (ICCV)*.

Dwibedi, D.; Misra, I.; and Hebert, M. 2017. Cut, paste and learn: Surprisingly easy synthesis for instance detection. In *ICCV*.

facebookresearch. 2019. VoteNet. https://github.com/facebookresearch/votenet. Accessed: 2021-09-10.

Fang, H.-S.; Sun, J.; Wang, R.; Gou, M.; Li, Y.-L.; and Lu, C. 2019. Instaboost: Boosting instance segmentation via probability map guided copy-pasting. In *Proceedings of the IEEE International Conference on Computer Vision*, 682–691.

Fang, H.-S.; Wang, C.; Gou, M.; and Lu, C. 2020. Graspnet-1billion: A large-scale benchmark for general object grasping. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11444–11453.

Fang, J.; Zuo, X.; Zhou, D.; Jin, S.; Wang, S.; and Zhang, L. 2021. LiDAR-Aug: A General Rendering-Based Augmentation Framework for 3D Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4710–4720.

Fisher, M.; and Hanrahan, P. 2010. Context-based search for 3D models. In *ACM SIGGRAPH Asia 2010 papers*, 1–10.

Helbing, D.; and Molnar, P. 1998. Social Force Model for Pedestrian Dynamics. *Physical Review E Statistical Physics Plasmas Fluids & Related Interdisciplinary Topics*, 51(5): 4282.

Hu, P.; Ziglar, J.; Held, D.; and Ramanan, D. 2020. What you see is what you get: Exploiting visibility for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11001–11009.

Ku, J.; Mozifian, M.; Lee, J.; Harakeh, A.; and Waslander, S. L. 2018. Joint 3d proposal generation and object detection from view aggregation. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 1–8. IEEE.

Lang, A. H.; Vora, S.; Caesar, H.; Zhou, L.; Yang, J.; and Beijbom, O. 2019. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 12697–12705.

Li, J.; Bian, S.; Zeng, A.; Wang, C.; Pang, B.; Liu, W.; and Lu, C. 2021a. Human Pose Regression with Residual Log-likelihood Estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11025–11034.

Li, J.; Xu, C.; Chen, Z.; Bian, S.; Yang, L.; and Lu, C. 2021b. HybrIK: A Hybrid Analytical-Neural Inverse Kinematics Solution for 3D Human Pose and Shape Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3383–3393.

Li, M.; Patil, A. G.; Xu, K.; Chaudhuri, S.; Khan, O.; Shamir, A.; Tu, C.; Chen, B.; Cohen-Or, D.; and Zhang, H. 2019. Grains: Generative recursive autoencoders for indoor scenes. *ACM Transactions on Graphics (TOG)*, 38(2): 1–16.

Liang, M.; Yang, B.; Wang, S.; and Urtasun, R. 2018. Deep continuous fusion for multi-sensor 3d object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 641–656.

Lucas, C. 2019. Point Cloud Hole Filling. https://github.com/Geodan/fill-holes-pointcloud.

Mahler, J.; Liang, J.; Niyaz, S.; Laskey, M.; Doan, R.; Liu, X.; Ojea, J. A.; and Goldberg, K. 2017. Dex-net 2.0: Deep learning to plan robust grasps with synthetic

point clouds and analytic grasp metrics. *arXiv preprint arXiv:1703.09312*.

Mousavian, A.; Anguelov, D.; Flynn, J.; and Kosecka, J. 2017. 3d bounding box estimation using deep learning and geometry. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7074–7082.

Qi, C. R.; Litany, O.; He, K.; and Guibas, L. J. 2019. Deep hough voting for 3d object detection in point clouds. In *Proceedings of the IEEE International Conference on Computer Vision*, 9277–9286.

Shi, S.; Guo, C.; Jiang, L.; Wang, Z.; Shi, J.; Wang, X.; and Li, H. 2020. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10529–10538.

Shi, S.; Wang, X.; and Li, H. 2019. Pointrcnn: 3d object proposal generation and detection from point cloud. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–779.

Song, S.; Lichtenberg, S. P.; and Xiao, J. 2015. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 567–576.

Wang, K.; Savva, M.; Chang, A. X.; and Ritchie, D. 2018. Deep convolutional priors for indoor scene synthesis. *ACM Transactions on Graphics (TOG)*, 37(4): 1–14.

Xie, Q.; Lai, Y.-K.; Wu, J.; Wang, Z.; Zhang, Y.; Xu, K.; and Wang, J. 2020a. MLCVNet: Multi-Level Context VoteNet for 3D Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10447–10456.

Xie, S.; Gu, J.; Guo, D.; Qi, C. R.; Guibas, L.; and Litany, O. 2020b. PointContrast: Unsupervised pre-training for 3D point cloud understanding. In *European Conference on Computer Vision*, 574–591. Springer.

Yan, Y.; Mao, Y.; and Li, B. 2018. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10): 3337.

Yang, B.; Liang, M.; and Urtasun, R. 2018. Hdnet: Exploiting hd maps for 3d object detection. In *Conference on Robot Learning*, 146–155.

Yang, B.; Luo, W.; and Urtasun, R. 2018. Pixor: Real-time 3d object detection from point clouds. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 7652–7660.

You, Y.; Ye, Z.; Lou, Y.; Li, C.; Li, Y.-L.; Ma, L.; Wang, W.; and Lu, C. 2020. Canonical Voting: Towards Robust Oriented Bounding Box Detection in 3D Scenes. *arXiv preprint arXiv:2011.12001*.

Zhang, Z.; Sun, B.; Yang, H.; and Huang, Q. 2020. H3dnet: 3d object detection using hybrid geometric primitives. In *European Conference on Computer Vision*, 311–329. Springer.

Zhou, Y.; While, Z.; and Kalogerakis, E. 2019. SceneGraphNet: Neural Message Passing for 3D Indoor Scene Augmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, 7384–7392.