SimIPU: Simple 2D Image and 3D Point Cloud Unsupervised Pre-Training for Spatial-Aware Visual Representations

Zhenyu Li,[1] Zehui Chen,[2] Ang Li,[3] Liangji Fang,[3] Qinhong Jiang,[3]
Xianming Liu,[1] Junjun Jiang,[1]* Bolei Zhou,[4] Hang Zhao[5]
[1] Harbin Institute of Technology, [2] University of Science and Technology,
[3] SenseTime Research, [4] The Chinese University of Hong Kong, [5] IIIS, Tsinghua University
{zhenyuli17, csxm, jiangjunjun}@hit.edu.cn, lovesnow@mail.ustc.edu.cn,
{liang1, fangliangji, jiangqinhong}@senseauto.com, bzhou@ie.cuhk.edu.hk, hangzhao@mail.tsinghua.edu.cn

Abstract

Pre-training has become a standard paradigm in many computer vision tasks. However, most of the methods are generally designed on the RGB image domain. Due to the discrepancy between the two-dimensional image plane and the three-dimensional space, such pre-trained models fail to perceive spatial information and serve as sub-optimal solutions for 3D-related tasks. To bridge this gap, we aim to learn a spatial-aware visual representation that can describe the three-dimensional space and is more suitable and effective for these tasks. To leverage point clouds, which are much more superior in providing spatial information compared to images, we propose a simple yet effective 2D Image and 3D Point cloud Unsupervised pre-training strategy, called **SimIPU**. Specifically, we develop a multi-modal contrastive learning framework that consists of an intra-modal spatial perception module to learn a spatial-aware representation from point clouds and an inter-modal feature interaction module to transfer the capability of perceiving spatial information from the point cloud encoder to the image encoder, respectively. Positive pairs for contrastive losses are established by the matching algorithm and the projection matrix. The whole framework is trained in an unsupervised end-to-end fashion. To the best of our knowledge, this is the first study to explore contrastive learning pre-training strategies for outdoor multi-modal datasets, containing paired camera images and LIDAR point clouds.

1 Introduction

Large-scale models have achieved significant success in deep learning, where fine-tuning after pre-training has become a well-established and commonly used paradigm, such as ELMo (Peters et al. 2018), GPT (Brown et al. 2020), and BERT (Devlin et al. 2018) in NLP. As for computer vision, benefiting from the massive amount of labeled data, the supervised pre-trained models on ImageNet (Deng et al. 2009) have long dominated. In recent years, unsupervised pre-training strategies have drawn much more attention. Various successful methods (He et al. 2020; Chen et al. 2020; Grill et al. 2020) have achieved comparable or better results compared to supervised pre-trained ones on a few 2D tasks, including image classification, object detection, and image segmentation (Jaiswal et al. 2021). However, all these methods are designed on the two-dimensional image plane, which exists a large discrepancy between the three-dimensional space. As a result, such pre-trained models can not perceive spatial information and demonstrate limited performance improvement on 3D-related downstream tasks. Therefore, learning a spatial-aware representation that can describe three-dimensional space is much more essential.

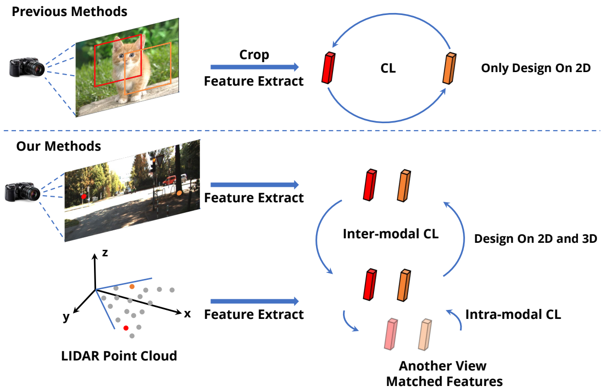


Figure 1: **Motivation**. SimIPU is designed on both the 2D image plane and the 3D space. Intra-modal module learns the spatial-aware representation with point clouds. Inter-modal module transfers the capability of extracting spatial-aware representations to the image-feature extractor. 'CL' is the abbreviation of the contrastive learning.

Compared to images, point clouds are much more superior in providing spatial information (Qi et al. 2017), which may lead them to be more suitable for learning such representations. PointContrast (Xie et al. 2020) is the first study to explore pre-training strategies for point clouds. They utilize different scene views to generate positive pairs and adopt a PointInfoNCE loss to learn useful dense/local representations. Motivated by the success of 2D image and 3D point cloud pre-training, Prior3D (Hou et al. 2021b) propose a geometric prior contrastive loss to imbue the prior 3D information to image representations. However, there is no intra-modal constraint on point clouds, which can lead to trivial solutions. Furthermore, all of these methods focus on in-

*Corresponding Author

door RGB-D data, where point clouds are reconstructed by the depth value. As for the outdoor scene, the point cloud provided by LIDAR contains more noise and massive background points. It also lacks point-to-point correspondences, which makes the design of pre-training methods tougher.

In this paper, we develop a simple yet effective 2D image and 3D point cloud unsupervised pre-training framework for outdoor multi-modal data (*i.e.,* paired images and LIDAR point clouds) to learn spatial-aware visual representations. To solve the aforementioned problems, our method explicitly imposes the contrastive loss on point-cloud features to guarantee models to learn spatial-aware representations. We harness more robust and informative global features and apply the Hungarian algorithm and the projection matrix to associate the matching correspondences. To the best of our knowledge, this is the first study to explore pre-training strategies for outdoor multi-modal data.

Specifically, the framework consists of an intra-modal spatial perception module and an inter-modal feature interaction module. In terms of the spatial perception module, we adopt an intra-modal contrastive learning method to learn spatial-aware representations with point clouds. We utilize global transformation to yield two views of a point cloud and adopt an encoder to extract the global features. Then, the Hungarian algorithm is applied to establish the matching correspondences between the downsampled points. A contrastive loss serves to push the distances of matched point features closer. The certain equivariance learned from random geometric transformation leads to a spatial-aware representation. As for the feature interaction module, we adopt a similar contrastive learning strategy to transfer prior spatial knowledge from the point-cloud encoder to the image encoder. Therefore, positive pairs between image features and point-cloud global features are established through the projection process from LIDAR to the camera. Since both features are global representations, alignment is achieved on the fly. Benefiting from the inter-modal interaction, the image encoder can gradually acquire the capability of extracting spatial-aware representations. In the pre-training stage, the whole framework is trained in an unsupervised end-to-end fashion. Our contributions are summarized as follows:

- We propose a simple yet effective pre-training method, termed **SimIPU**. It exerts the advantages of massive unlabeled multi-modal data to learn spatial-aware visual representations that can further improve the model performance on downstream tasks. To the best of our knowledge, this is the first study to explore contrastive learning pre-training strategies for outdoor multi-modal data.

- We develop a Multi-Modal Contrastive Learning framework, which consists of an intra-modal spatial perception module and an inter-modal feature interaction module.

- Our method significantly outperforms other pre-training counterparts when transferring the models to 3D-related downstream tasks, including 3D object detection (2.3% AP), monocular depth estimation (0.12m RMSE) and monocular 3D object detection (0.6% AP).

2 Related Work

2.1 2D Self-supervised Representation Learning

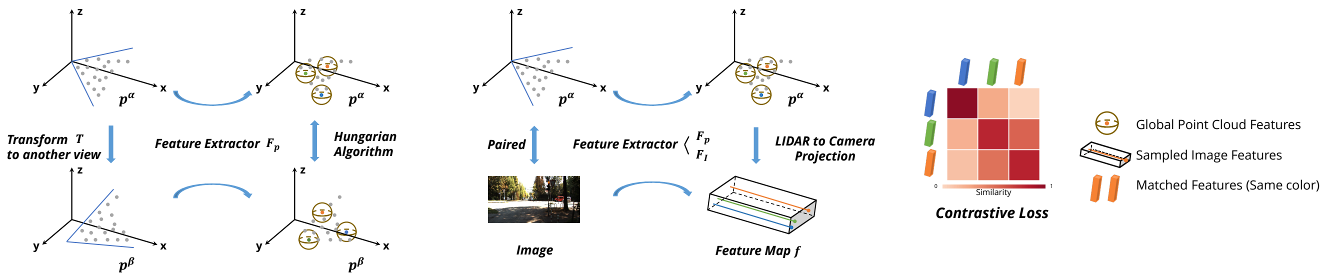
Pretext task and contrastive learning are two key points of 2D self-supervised representation learning. There is a wide range of tasks that have been designed to learn useful visual representations, including colorization (Zhang, Isola, and Efros 2016), inpainting (Pathak et al. 2016), spatial jigsaw puzzles (Noroozi and Favaro 2016), and discriminate orientation (Gidaris, Singh, and Komodakis 2018). Although the improvement is limited, these methods provide a possibility to achieve performance gains from pre-training strategies. SimCLR and SimCLR v2 (Chen et al. 2020) makes a breakthrough. They groundbreakingly proposed a discrimination pretext task. Contrastive loss is applied for pushing away the feature distances of different instances. MoCo and its improved version MoCo v2 (He et al. 2020) further utilize a memory bank to alleviate the constraints on large batch size. Beyond contrastive learning, BYOL (Grill et al. 2020) relies only on positive pairs, but it does not collapse in case a momentum encoder is used. All these methods are designed on the image plane and can be suboptimal solutions for 3D-related downstream tasks. To circumvent the need for spatial-aware representations, some methods propose to learn representations from videos by using ego-motion as supervisory signal (Jayaraman and Grauman 2015; Agrawal, Carreira, and Malik 2015; Lee et al. 2019) and self-supervised depth estimation (Jiang et al. 2018). In this paper, we aim to further explore contrastive pre-training strategies following the successful trend of contrastive learning.

2.2 3D Self-supervised Representation Learning

Inspired by the success of 2D self-supervised representation learning, PointContrast (Xie et al. 2020) introduces a contrastive pretext task in a 3D paradigm. Driven by indoor point cloud data properties, the same points in different frames compose positive pairs and are used for contrastive learning. To make full use of the point cloud data, Contrast Context (Hou et al. 2021a) proposes to adopt a ShapeContext descriptor to divide the scene, which provides more negative pairs for contrastive learning and improve the effectiveness of pre-training models. CoCoNets (Lal et al. 2021) further explores self-supervised learning of amodal 3D feature representations agnostic to object and scene semantic content. The above methods focus on indoor RGB-D data. As for outdoor LIDAR point clouds, Pillar-Motion (Luo, Yang, and Yuille 2021) propose a self-supervised pillar representation learning method that makes use of the optical flow extracted from camera images. Since the LIDAR point clouds lack point-to-point correspondences, there are fewer contrastive learning methods for pre-training.

2.3 Multi-modal Representation Learning

Much effort has been made into multi-modal representation learning. Based on paired image and text, (Yuan et al. 2021) propose a unified multi-modality contrastive learning framework to learn useful visual representations. Motivated by the success of 2D image and 3D point cloud pre-training,



(a) Spatial Perception Module

(b) Feature Interaction Module

(c) Contrastive Losses

Figure 2: **Framework of SimIPU**. Matched pairs are in the same color. The whole framework is trained in an end-to-end manner. (a) **Intra-Modal Spatial Perception Module**: We utilize set abstraction layers to extract global point cloud features and downsample points (results are in color) from different views. The Hungarian Algorithm is applied to match the downsampled points according to locations. (b) **Inter-Modal Feature Interaction Module**: We adopt a standard ResNet-50 to extract global image features. Projection matrix from point cloud to image plane establish the association between positive pairs. (c) **Contrastive Loss**: Contrastive losses are applied to push closer the distances of matched pair features.

Pri3D (Hou et al. 2021b) further explore multi-modal pre-training methods to enhance the visual representation learning with indoor RGB-D data. Dense/local representations are learned, which boosts the performance of downstream tasks. However, there is no intra-modal constraint on point clouds, which can lead to trivial solutions and limit performance improvement. On the contrary, (Liu et al. 2021) propose a method to imbue the image prior to the 3D representation. All these methods motivate us to further explore multi-modal pre-training strategies for more challenging outdoor multi-modal data.

2.4 3D Visual Tasks

We utilize three 3D visual tasks to evaluate the effectiveness of our method, which are fusion-based 3D object detection, monocular depth estimation, and monocular 3D object detection. Fusion-based 3D object detection methods (Zhang, Wang, and Change Loy 2020; Sindagi, Zhou, and Tuzel 2019; Liang et al. 2019; Wang et al. 2020) combine the image and point cloud modality and learn the interaction between them. By exert multi-modal data, they achieve satisfying results compared to methods that only utilize point cloud data. In this paper, we mainly focus on this task and design extensive experiments to show the effectiveness of our method. For monocular depth estimation and monocular 3D object detection, which are two challenging 3D-related visual tasks on a single image modality, many methods (Eigen and Fergus 2015; Bhat, Alhashim, and Wonka 2021; Lyu et al. 2020; Wang et al. 2021) are proposed to improve the model performance. We execute experiments on these two tasks, which only utilize the single image modality to further evaluate the generalization of SimIPU.

3 Method

In this section, we introduce our self-supervised pre-training pipeline in detail. First, to motivate the necessity of this multi-modal method, we conduct a pilot study to determine what kind of pre-training strategy we need in the downstream task that we are mainly focusing on: fusion-based 3D objection detection (Section 3.1). Then, we introduce our multi-modal self-supervised pre-training framework, including an Intra-Modal Spatial Perception module (Section 3.2), and an Inter-Modal Feature Interaction module (Section 3.3). The overview of the proposed framework is shown in Fig. 2.

3.1 Pilot Study: Is 2D pre-training Useful?

Previous fusion-based 3D objection detection methods utilize different kinds of pre-trained 2D feature extractors, including scratch models, ImageNet supervised classification pretrained models in (Chen et al. 2017; Liang et al. 2019; Wang et al. 2020), and 2D detection pre-trained models in (Sindagi, Zhou, and Tuzel 2019), to initialize the backbone. However, there has been no discussion about which pre-training strategy can further improve the model performance on 3D object detection. To fill this blank, we execute this pilot study to assess the effect of different pre-trained models on fusion-based 3D object detection.

We adopt the state-of-the-art fusion-based 3D object detection method MVXNet (Sindagi, Zhou, and Tuzel 2019) with Moca (Zhang, Wang, and Change Loy 2020) as our baseline, and train models on the KITTI (Geiger, Lenz, and Urtasun 2012) dataset. We only change different pre-trained 2D feature extractor weights when initialization and keep all the other training settings the same. The results are shown in Fig. 3. Critically, one can observe that pre-training models cannot improve the performance of the downstream task. These results suggest that the discrepancy between the two-dimensional image plane and three-dimensional space exists. All these pre-trained methods are designed on the 2D domain, which leads to sub-optimal solutions for the fusion-based 3D task.

Is there any way to learn a spatial-aware visual representation that is beneficial to 3D tasks? Without any label, it is ex-

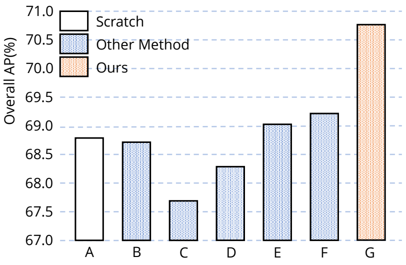


Figure 3: Pilot Study Results. In KITTI 3D object detection experiments, we adopt different pre-trained models, such as A. Scratch, B. 2D detection on CityScapes, C. 2D detection on KITTI, D. Supervised pre-trained on ImageNet, E. Moco-v2 on ImageNet, F. DenseCL on ImageNet, and G. Our method on KITTI, to initialize the backbones.

tremely tough to directly learn such representations with the single image modality. Driven by massive multi-modal data, we can achieve it via multi-modal contrastive learning in an indirect manner. Specifically, we propose an intra-modal spatial perception module to learn a spatial-aware representation from point clouds and an inter-modal feature interaction module to transfer the capability of space perception to the image encoder.

3.2 Intra-Modal Spatial Perception Module

We design the intra-modal contrastive learning module to pre-train a spatial-aware global representation with point clouds. The framework is shown in Fig. 2a. A key observation is that features at the same location in different views should be similar.

To yield two different views of a point cloud, we sample a random 3D geometric transformation T to transform a given point cloud $p^{\alpha} \in R^{n\times c}$ into another view $p^{\beta} \in R^{n\times c}$:

$$p^{\beta} = T(p^{\alpha}), \tag{1}$$

where n is the point number in a scene, and c is the channel of raw point features, which normally includes the 3D location and reflection rate. The superscript α and β indicate two different views. In this work, we mainly consider the rigid transformation T, including rotation, translation, and scaling.

On completion of constituting two different views, we extract the global point cloud features by a PointNet++ (Qi et al. 2017) encoder. Specifically, we apply several set abstraction layers for downsampling and extracting global context representations, which can be mathematically written as:

$$l^{\alpha}, f^{\alpha} = F_P(p^{\alpha}),\ l^{\beta}, f^{\beta} = F_P(p^{\beta}), \tag{2}$$

where l and f are the location and feature of downsampled points, respectively. F_P is the point cloud feature extractor. Note that our method is different from pre-training methods for indoor point clouds, which mainly focus on dense/local representations (Xie et al. 2020; Hou et al. 2021a,b). Since the outdoor data contains much more noise and massive background points, we utilize the global features to enhance the quality of the extracted representations. Therefore, during the feature extraction, meaningless information can be gradually filtered by the random sample strategy and the set abstraction layers. As a result, the spatial-aware representations will be well preserved, which results in certain spatial prior knowledge to transfer to the image encoder. Furthermore, random sampling can make the same point cloud generate different sampled points, which can improve the utilization of point cloud data and the effectiveness of the representation learning.

However, extracting global features can introduce random properties, which leads to the inevitable mismatching of downsampled points. To construct positive pairs for contrastive learning, we utilize the Hungarian algorithm to achieve the positive correspondence matching M_1. The cost matrix of the bipartite assigning algorithm is computed by the l_2-norm distance between the downsampled points in two different views:

$$M_1 = assign(cost = distance(T(l^{\alpha}),\ l^{\beta})). \tag{3}$$

Here, we apply the same transformation T in Eq.1 to align the coordinates for the distance computation. The Hungarian algorithm can guarantee a favorable global-optimal matching and solves the problem of the lack of correspondences in outdoor multi-modal data.

After establishing the correspondence, we adopt the constrastive loss to push in the distance between the features of matched points. As for each matched pair $(i,\ j) \in M_1$, point feature f_i^{α} will serve as the query and f_j^{β} will serve as the positive key k^{+}. We treat point feature f_k^{β} where $(\cdot,\ k) \in M_1$ and $k \neq j$ as negative keys. We calculate the intra-modal contrastive loss as:

$$L_{intra} = -\sum_{(i,\ j)\in M_1} \log \frac{\exp\left(f_i^{\alpha} \cdot f_j^{\beta}/\tau\right)}{\sum_{(\cdot,\ k)\in M_1} \exp\left(f_i^{\alpha} \cdot f_k^{\beta}/\tau\right)}, \tag{4}$$

where τ is the temperature factor.

The intra-modal contrastive loss can push in the feature distance on a similar location of different views. Such a certain equivariance learned from random geometric transformation engenders a spatial-aware representation.

3.3 Inter-Modal Feature Interaction Module

To enable the image encoder the capability of perceiving spatial space, we propose the Inter-Modal Feature Interaction module, where the image feature extractor can gradually learn spatial-aware representations by embracing the inter-modal interaction. The framework is shown in Fig. 2b.

Following most of the contrastive learning methods, we adopt a standard ResNet-50 as the default image backbone to extract global feature maps from given images:

$$f = F_I(I), \tag{5}$$

where f, F_I, I are the feature map, image feature extractor, and the input image, respectively.

Given the camera parameter C, we can establish the positive correspondences between the downsampled points $(l^{\alpha},\ f^{\alpha})$ in Eq.2 and the image feature maps f in Eq.5

Pre-train	Car AP$_{3D}$(%)			Pedestrian AP$_{3D}$(%)			Cyclist AP$_{3D}$(%)			Overall AP$_{3D}$(%)		
	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
Scratch	86.18	76.57	74.08	67.95	62.18	57.24	83.37	66.99	63.11	79.17	68.58	64.81
Ours-K	87.87	77.36	74.30	71.25	66.18	60.24	84.83	69.11	64.04	**81.32**	**70.88**	**66.19**
Gain	+1.69	+0.79	+0.22	+3.30	+4.00	+3.00	+1.46	+2.12	+0.93	+2.15	+2.30	+1.38
MoCo-v2-IN	87.98	77.40	74.08	69.33	62.03	57.14	82.66	67.38	62.42	79.99	68.94	64.55
MoCo-v2-K	87.66	77.10	74.33	69.24	62.42	57.86	80.55	65.33	60.97	79.15	68.28	64.39
DenseCL-IN	88.11	77.56	74.62	66.56	61.42	57.08	83.86	68.81	64.74	79.51	69.26	65.48

Table 1: Camera-lidar fusion based 3D object detection fine-tuned on KITTI validation set. We show the bounding box AP of each class in details. 'K' and 'IN' indicates pre-trained models are trained on KITTI and ImageNet datset. Best is in **bold**.

through the projection matrix. Specifically, we project the downsampled points onto the image plane and sample from the image feature maps to get the corresponding image features f^γ:

$$f^\gamma = f\Big\langle proj(l^\alpha, C)\Big\rangle, \tag{6}$$

where $proj()$ are the resulting 2D coordinates of the projected points. $\langle\rangle$ is the sampling operator. In our work, we utilize bi-linear interpolation to sample features.

Above Operations lead to positive matches M_2. Similar to Eq. 4, for each matched pair $(i,\ j) \in M_2$, we calculate the inter-modal contrastive loss as:

$$L_{inter} = -\sum_{(i,\ j)\in M_2} \log \frac{\exp\left(f_i^\alpha \cdot f_j^\gamma/\tau\right)}{\sum_{(\cdot,\ k)\in M_2} \exp\left(f_i^\alpha \cdot f_k^\gamma/\tau\right)}. \tag{7}$$

Here, we crop the gradient of f^α to avoid influences on Intra-Modal Spatial Perception Module. The contrastive loss imbues the prior spatial knowledge of the lidar feature extractor to the image encoder.

Finally, we train the whole framework in an end-to-end fashion with the total loss:

$$L_{total} = \lambda L_{intra} + \mu L_{inter}, \tag{8}$$

where λ, μ are hyperparameters that balances between the two parts of the loss.

Compared with methods that focus on the indoor RGB-D data (Hou et al. 2021b; Liu et al. 2021), which utilize a U-Net (Ronneberger, Fischer, and Brox 2015) shape backbone to align the dense/local feature extracted by point-cloud feature extractor, our method only adopts a standard ResNet-50 encoder to extract the global features. Since the modules apply to global representations, alignment is achieved on the fly. Such characteristic indicates that our method is more general to downstream tasks because there is no more limitation on the downstream network design.

4 Experiments

In this section, we introduce our experimental settings (Section 4.1) and downstream results (Section 4.2) in detail. The ablation study is applied to prove the effectiveness of key components in the framework and explore the influence of pre-training data scale in Section 4.3.

4.1 Experimental Settings

We study pre-training strategies on multi-modal datasets, including the KITTI dataset and the Waymo Open Dataset. Both the datasets contain paired image and point cloud data. All the experiments are based on MMDetection3d (Contributors 2020).

KITTI Dataset (K). The KITTI Dataset (Geiger, Lenz, and Urtasun 2012) contains 7481 training images and 7518 test images, both with their corresponding point cloud. To make full use of the dataset, we utilize both the training set and the testing set data to pre-train the model. Note that we filter out the validation set to avoid the information leak.

Waymo Open Dataset (W). The Waymo Dataset (Sun et al. 2020) contains ∼0.15 million training images with corresponding point clouds. The testing set is relatively smaller than the training set. We only use the training set data to pre-train models. Because we need to project lidar points to image plane during the pre-training stage, for simplicity, we filter out points beyond the image field of view (FOV).

Pre-training. In terms of the backbone settings, we use three set abstraction layers (Qi et al. 2017; Yang et al. 2020) to downsample the points and extract point cloud global features. We combine two downsampling strategies to make full use of the data, which are the 3D Euclidean distance sample (D-FPS) (Qi et al. 2017) and the feature distance sample (F-FPS) (Yang et al. 2020). Following most of the contrastive learning methods, a ResNet-50 (He et al. 2016) is adopted as the image feature extractor, which can also be replaced by any other image backbone. Compared with the Prior3D (Hou et al. 2021b), our method has fewer constraints on the image backbone and is more general to downstream tasks.

As for contrastive learning settings, the temperature factor τ in Eq. 4 and Eq.7 is 0.07. Following (Chen et al. 2020), we use the MLP projection head to map the dimension of features to 128-d for either the intra-modal contrastive learning and the inter-modal contrastive learning. We use 4096 matched pairs for faster training (Xie et al. 2020). In addition, we implement the Moco-v2 (He et al. 2020) on both the KITTI dataset and the Waymo dataset to make a fair comparison with other strong unsupervised methods. The data augmentation pipeline of MoCo-v2 consists of random color jittering, random gray-scale conversion, Gaussian blurring, and random horizontal flip.

Pre-train	Vehicle L1/L2		Pedestrian L1/L2		Cyclist L1/L2		Overall L1/L2	
	mAP(%)	mAPH(%)	mAP(%)	mAPH(%)	mAP(%)	mAPH(%)	mAP(%)	mAPH(%)
Scratch	65.0/61.0	64.6/60.5	67.6/62.9	58.8/54.7	64.0/61.2	61.1/58.5	65.57/61.53	61.75/57.93
Ours-W	66.5/62.4	66.1/62.0	69.4/64.7	60.5/56.3	64.7/62.3	62.3/60.0	66.92/63.01	63.18/59.47
Gain	+1.5/+1.4	+1.5/+1.5	+1.8/+1.8	+1.2/+1.6	+0.7/+1.1	+1.2/+1.5	+1.35/+1.48	+1.43/+1.54

Table 2: Camera-lidar fusion based 3D object detection performance comparison on Waymo validation set. ‘W’ indicates the pre-trained models are trained on Waymo datset.

Pre-train	REL↓	Sq R↓	RMS↓	log↓	δ_1↑	δ_2↑	δ_3↑
Scratch	0.096	0.493	3.575	0.146	0.896	0.977	0.994
Ours-W	0.073	0.285	2.840	0.113	0.935	0.990	**0.998**
Gain	-0.023	-0.208	-0.735	-0.033	+0.039	+0.013	+0.004
Super-IN	0.068	0.247	2.712	0.104	0.946	**0.993**	**0.998**
Ours-IN/W	**0.067**	**0.235**	**2.592**	**0.102**	**0.949**	**0.993**	**0.998**
Gain	-0.001	-0.012	-0.120	-0.002	+0.003	0	0

Table 3: Monocular depth estimation performance comparison on KITTI dataset. ‘IN/W’ indicates the double fine-tuning pre-trained models on ImageNet and Waymo.

Pre-train	AP(%)	ATE	ASE	AOE	AVE	AAE	NDS(%)
Scratch	17.90	0.92	0.30	0.84	1.33	0.19	26.27
Ours-W	26.18	0.84	0.27	0.67	1.31	0.17	33.50
Gain	+8.28	-	-	-	-	-	+7.23
Super-IN	27.71	0.83	0.26	0.59	1.34	0.16	35.23
Ours-IN/W	**28.36**	0.82	0.26	0.62	1.33	0.16	**35.36**
Gain	+0.65	-	-	-	-	-	+0.13

Table 4: Monocular 3D object detection performance comparison on Nuscenes dataset.

4.2 Experimental Results

3D Object Detection. We evaluate the pre-trained models by fine-tuning on the target 3D-related tasks. For early-fusion based 3D object detection, we utilize two challenging and popular datasets, *i.e.*, KITTI and Waymo. We fine-tune the pre-trained models with the state-of-the-art algorithms: MVX-Net (Sindagi, Zhou, and Tuzel 2019) with Moca (Zhang, Wang, and Change Loy 2020) on KITTI. For the Waymo dataset, we find that the Moca can not improve the MVX-Net performance. Therefore, we choose to only apply the MVX-Net as our default protocol, which is still a strong baseline of early-fusion based methods. When evaluating, we use the standard 2× schedule, which is more effective on the waymo dataset.

The KITTI 3D object detection performance comparison is shown in Tab. 1. We utilize the scratch one as our baseline method. Compared with it, our method achieves the significant 2.3% moderate overall AP_{3D} gains. Furthermore, 1.38% gains on the hard overall AP_{3D} shows that our pre-training method improves the localization accuracy. To further dig into the effectiveness of the pre-training methods, we report the per-class comparison results. Our fine-tuned model can localize small objects more accurately, which is reflected in significant 4% gains on pedestrian moderate AP_{3D}, compared with other classes. More results compared with other state-of-the-art counterpart pre-training methods can be found in Fig. 3 and the appendix:.

In Fig. 4, we report the 3D object detection results on Waymo Dataset. Waymo dataset is much larger than KITTI. Although the performance difference between pre-trained and scratch models is not obvious for larger dataset (Xie et al. 2020), our method still achieves a slight improvement even using limited data. Compared with training from scratch, our method achieves a relatively significant 1.35% mAP improvement. Similar to the results on KITTI, our fine-tuned model has the capability to localize small objects more accurately. We provide more results in the appendix.

KITTI Monocular Depth Estimation. As for the monocular depth estimation, we design a simple yet strong baseline to evaluate the model performance. We provide more information of the baseline in the appendix. We evaluate the effectiveness of our method by fine-tuning pre-trained the model on the KITTI Eigen split (Eigen and Fergus 2015). All settings are the same when doing fine-tuning to make a fair comparison.

In Tab. 3, we report the KITTI monocular depth estimation performance. Compared with training from scratch, our methods achieve great improvement on all of the metrics. However, we can only obtain ∼0.15M (Million) data to pre-train our model, which is extremely smaller than the ImageNet supervised pre-training methods that can utilize ∼1M labeled data to pre-train models, therefore, there is a small performance gap between the results of Ours-W over Super-IN. To fill this gap, we propose a simple double fine-tuning strategy. We load the pre-trained Super-IN model at the beginning of our multi-modal contrastive pre-training stage and double fine-tune the model on the downstream task. As a result, our method can learn useful spatial-aware visual representations and preserve part of the semantic representations, which leads to further performance gains, especially for the 0.12m (4.4%) improvement on RMS.

Nuscenes Monocular 3D Object Detection. In terms of the monocular 3D object detection, FCOS3D (Wang et al. 2021) is fine-tuned on Nuscenes training set and evaluated on Nuscenes validation set. For simplicity, we only replace the image encoder ResNet-101 in the default config of FCOS3D with ResNet-50. When evaluating, we use the standard 1× schedule. The batch size is set to 8. Synchronized batch normalization is used. All settings are default.

We report the performance of Nuscenes monocular 3D object detection in Tab. 4. Our methods achieve significant improvement on all of the metrics compared with training from scratch. Applying the same double fine-tuning strategy, our method further boosts the performance of the Super-IN model by 0.65% mAP.

Pre-train	Overall AP$_{3D}$(%)		
	Easy	Mod.	Hard
Scratch (Equivalent to $\lambda = 1$, $\mu = 0$)	79.17	68.58	64.81
SimIPU w/o intra-module ($\lambda = 0$, $\mu = 1$)	79.36	69.19	65.17
SimIPU w/o inter-module ($\lambda = 1$, $\mu = 0$)	79.17	68.58	64.81
SimIPU ($\lambda = 1$, $\mu = 1$)	**81.32**	**70.88**	**66.19**
Greedy Assignment	78.92	68.20	64.72
Hungarian Algorithm	**81.32**	**70.88**	**66.19**

Table 5: Ablation study on intra-modal contrastive learning module and matching strategies. We report 3D object detection performance results on KITTI validation set.

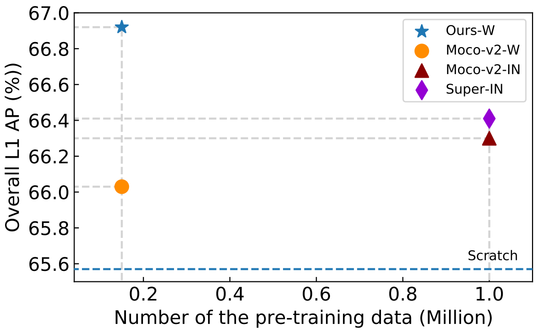


Figure 4: Fusion-based 3D object detection performance comparison on Waymo dataset. SimIPU achieves comparable results with limited multi-modal pre-training data.

4.3 Ablation Study

Intra-Modal Module. Viewing the whole framework, our method can be treated as an indirect method for image feature extractors to learn a spatial-aware visual representation. Knowledge transferred from the point cloud encoder is crucial and determines the effectiveness of our method. We design the first ablation study to prove if the intra-modal contrastive learning module does have learned a useful representation. The results shown in the first block of Tab. 5 indicate that the intra-modal branch is essential in the multi-modal contrastive learning framework. With the intra-modal branch, the point cloud feature extractor can learn a spatial-aware visual representation and transfer it to the image encoder via inter-modal contrastive learning, which can further boost the performance of 3D-related tasks.

Matching Algorithm. We use the Hungarian Algorithm as our default assignment method, which can achieve the optimal global solution for the positive pair association. In this ablation study, we compare it with another commonly used matching algorithm: Greedy Assignment. During the matching process, this algorithm takes the nearest optimal options and repeats them. The ablation experimental results are shown in the last block of Tab. 5. Greedy assignment hampers the downstream performance. The reason may be caused by poor quality matches. It hurts the effectiveness of intra-modal contrastive learning, which is essential to learn useful representations for downstream tasks.

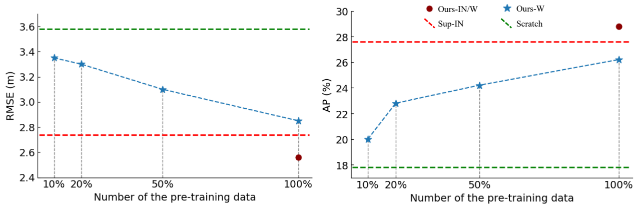


Figure 5: Ablation study on different pre-training data scale. We report the RMSE values on KITTI monocular depth estimation (left) and the AP results on Nuscenes monocular 3D object detection (right).

Number of Pre-training Data. In this ablation experiment, we use different amounts of training data to pre-train models on the Waymo dataset and fine-tune them on both the monocular depth estimation and the monocular 3D object detection task, to explore the influence of the number of pre-training data on downstream tasks. Results are shown in Fig. 5. One can easily observe that the performance of downstream tasks is significantly improved by the increase of pre-training data. It is in line with our intuition: a larger scale of pre-training data will further boost the performance of downstream tasks. Note that we only use ∼0.15M unlabeled data to pre-train our model. Compared with Super-IN, which utilizes ∼1M labeled data to pre-train models, our method undoubtedly shows significant competitiveness. The curves of the experimental results can thus be suggested that our method will easily surpass the Super-IN with more pre-training data. In addition, by a simple double fine-tuning strategy, our method can further boost the performance of baseline models, which indicates the friendly compatibility and generalization of our method.

5 Conclusion

In this paper, we propose SimIPU, a simple yet effective 2D Image and 3D Point cloud Unsupervised pre-training method, and develop a multi-modal contrastive learning framework to learn spatial-aware visual representation for 3D-related tasks in the outdoor environment. This method fills the blank of pre-training methods for outdoor multi-modal datasets and achieves significant performance gains on different 3D-related downstream tasks, including fusion-based 3D object detection, monocular depth estimation, and monocular 3D object detection, with a limited number of multi-modal pre-training data. However, an inadequate of this method is that it only focuses on spatial-aware visual representation while ignores the semantic information, but even notwithstanding this limitation, our approach still shows great generalization and effectiveness on downstream tasks. In the long term, associating both the spatial and the semantic information would be a fruitful area for further work, and we will dig into more effective methods to achieve this purpose. We hope our work will encourage more research on visual representation learning for a suitable design of the cross-modal pre-training paradigm.

6 Acknowledgments

The research was supported by the National Natural Science Foundation of China (61971165, 61922027), in part by the Fundamental Research Funds for the Central Universities (FRFCU5710050119). The authors would like to thank Jinghuai Zhang from the Computer Science Department of Duke University for helpful discussions on topics related to this work.

References

Agrawal, P.; Carreira, J.; and Malik, J. 2015. Learning to see by moving. In *International Conference on Computer Vision (ICCV)*, 37–45.

Bhat, S. F.; Alhashim, I.; and Wonka, P. 2021. Adabins: Depth estimation using adaptive bins. In *Computer Vision and Pattern Recognition (CVPR)*, 4009–4018.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NIPS)*, volume 33, 1877–1901.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, 1597–1607.

Chen, X.; Ma, H.; Wan, J.; Li, B.; and Xia, T. 2017. Multi-view 3D object detection network for autonomous driving. In *Computer Vision and Pattern Recognition (CVPR)*, 1907–1915.

Contributors, M. 2020. MMDetection3D: OpenMMLab next-generation platform for general 3D object detection. https://github.com/open-mmlab/mmdetection3d.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition (CVPR)*, 248–255.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 4171–4186.

Eigen, D.; and Fergus, R. 2015. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *International Conference on Computer Vision (ICCV)*, 2650–2658.

Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Computer Vision and Pattern Recognition (CVPR)*, 3354–3361.

Gidaris, S.; Singh, P.; and Komodakis, N. 2018. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations (ICLR)*, volume 1.

Grill, J.-B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P. H.; Buchatskaya, E.; Doersch, C.; Pires, B. A.; Guo, Z. D.; Azar, M. G.; et al. 2020. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Computer Vision and Pattern Recognition (CVPR)*, 9729–9738.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Hou, J.; Graham, B.; Nießner, M.; and Xie, S. 2021a. Exploring data-efficient 3D scene understanding with contrastive scene contexts. In *Computer Vision and Pattern Recognition (CVPR)*, 15587–15597.

Hou, J.; Xie, S.; Graham, B.; Dai, A.; and Nießner, M. 2021b. Pri3D: Can 3D Priors Help 2D Representation Learning? In *International Conference on Computer Vision (ICCV)*.

Jaiswal, A.; Babu, A. R.; Zadeh, M. Z.; Banerjee, D.; and Makedon, F. 2021. A survey on contrastive self-supervised learning. *Technologies*, 9(1): 2.

Jayaraman, D.; and Grauman, K. 2015. Learning image representations tied to ego-motion. In *International Conference on Computer Vision (ICCV)*, 1413–1421.

Jiang, H.; Larsson, G.; Shakhnarovich, M. M. G.; and Learned-Miller, E. 2018. Self-supervised relative depth learning for urban scene understanding. In *European Conference on Computer Vision (ECCV)*, 19–35.

Lal, S.; Prabhudesai, M.; Mediratta, I.; Harley, A. W.; and Fragkiadaki, K. 2021. CoCoNets: Continuous contrastive 3D scene representations. In *Computer Vision and Pattern Recognition (CVPR)*, 12487–12496.

Lee, S.; Kim, J.; Oh, T.-H.; Jeong, Y.; Yoo, D.; Lin, S.; and Kweon, I. S. 2019. Visuomotor Understanding for Representation Learning of Driving Scenes. In *The British Machine Vision Conference (BMVC)*.

Liang, M.; Yang, B.; Chen, Y.; Hu, R.; and Urtasun, R. 2019. Multi-task multi-sensor fusion for 3D object detection. In *Computer Vision and Pattern Recognition (CVPR)*, 7345–7353.

Liu, Y.-C.; Huang, Y.-K.; Chiang, H.-Y.; Su, H.-T.; Liu, Z.-Y.; Chen, C.-T.; Tseng, C.-Y.; and Hsu, W. H. 2021. Learning from 2D: Pixel-to-point knowledge transfer for 3D pretraining. *arXiv preprint arXiv:2104.04687*.

Luo, C.; Yang, X.; and Yuille, A. 2021. Self-supervised pillar motion learning for autonomous driving. In *Computer Vision and Pattern Recognition (CVPR)*, 3183–3192.

Lyu, X.; Liu, L.; Wang, M.; Kong, X.; Liu, L.; Liu, Y.; Chen, X.; and Yuan, Y. 2020. HR-Depth: high resolution self-supervised monocular depth estimation. In *AAAI Conference on Artificial Intelligence (AAAI)*.

Noroozi, M.; and Favaro, P. 2016. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision (ECCV)*, 69–84.

Pathak, D.; Krahenbuhl, P.; Donahue, J.; Darrell, T.; and Efros, A. A. 2016. Context encoders: Feature learning by inpainting. In *Computer Vision and Pattern Recognition(CVPR)*, 2536–2544.

Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep contextualized word representations. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, volume 1, 2227–2237.

Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems (NIPS)*, volume 30.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 234–241.

Sindagi, V. A.; Zhou, Y.; and Tuzel, O. 2019. Mvx-net: Multimodal voxelnet for 3D object detection. In *International Conference on Robotics and Automation (ICRA)*, 7276–7282.

Sun, P.; Kretzschmar, H.; Dotiwalla, X.; Chouard, A.; Patnaik, V.; Tsui, P.; Guo, J.; Zhou, Y.; Chai, Y.; Caine, B.; et al. 2020. Scalability in perception for autonomous driving: Waymo open dataset. In *Computer Vision and Pattern Recognition (CVPR)*, 2446–2454.

Wang, G.; Tian, B.; Zhang, Y.; Chen, L.; Cao, D.; and Wu, J. 2020. Multi-view adaptive fusion network for 3D object detection. *arXiv preprint arXiv:2011.00652*.

Wang, T.; Zhu, X.; Pang, J.; and Lin, D. 2021. FCOS3D: Fully convolutional one-stage monocular 3D object detection. In *International Conference on Computer Vision Workshops (ICCVW)*.

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

Yang, Z.; Sun, Y.; Liu, S.; and Jia, J. 2020. 3DSSD: Point-based 3D single stage object detector. In *Computer Vision and Pattern Recognition (CVPR)*, 11040–11048.

Yuan, X.; Lin, Z.; Kuen, J.; Zhang, J.; Wang, Y.; Maire, M.; Kale, A.; and Faieta, B. 2021. Multimodal contrastive training for visual representation learning. In *Computer Vision and Pattern Recognition (CVPR)*, 6995–7004.

Zhang, R.; Isola, P.; and Efros, A. A. 2016. Colorful image colorization. In *European Conference on Computer Vision (ECCV)*, 649–666.

Zhang, W.; Wang, Z.; and Change Loy, C. 2020. Multimodality cut and paste for 3D object detection. *arXiv preprint arXiv:2012.12741*.