**Self-Supervised Image Representation Learning with Geometric Set Consistency**

Nenglun Chen*1,2, Lei Chu2, Hao Pan2, Yan Lu2, and Wenping Wang3

1The University of Hong Kong 2Microsoft Research Asia 3Texas A&M University

**Abstract**

We propose a method for self-supervised image representation learning under the guidance of 3D geometric consistency. Our intuition is that 3D geometric consistency priors such as smooth regions and surface discontinuities may imply consistent semantics or object boundaries, and can act as strong cues to guide the learning of 2D image representations without semantic labels. Specifically, we introduce 3D geometric consistency into a contrastive learning framework to enforce the feature consistency within image views. We propose to use geometric consistency sets as constraints and adapt the InfoNCE loss accordingly. We show that our learned image representations are general. By fine-tuning our pre-trained representations for various 2D image-based downstream tasks, including semantic segmentation, object detection, and instance segmentation on real-world indoor scene datasets, we achieve superior performance compared with state-of-the-art methods.

**1. Introduction**

Self-supervised image representation learning is an important problem in the field of computer vision and has been rapidly developed in recent years. Existing works in this area mainly focus on designing various pretext tasks to learn general and intrinsic image features in self-supervised manners [14, 18, 37]. Those pretext tasks are usually low-level and can capture general image properties that favor many downstream tasks, like image classification, semantic segmentation, object detection, instance segmentation, etc. Due to its ability to learn from a large amount of unlabeled data, self-supervised representation learning has already become a standard training regime in many real-world applications [2, 3, 15, 55].

Recently, researchers have started to use 3D data, usually represented by point clouds, meshes, or voxel grids, as guidance for learning image representations [18, 30]. Compared with 2D images, 3D data has complementary advantages for learning discriminative image features. Since 3D data is usually acquired by real-world scanning and reconstruction, and has the same dimension as the real-world scenes, learning geometric structures from 3D data is much easier than 2D images. Moreover, complex occlusions, as well as the texture of objects will also affect the performance of image perception methods. Meanwhile, 3D data is occlusion-free, and geometric cues like smooth regions and sharp edges can be strong priors for semantic understanding. Thus it is natural to use 3D geometric cues to favor the learning of image...
representations.

Pioneering works, like Pri3D [18], mainly rely on multi-view pixel-level consistency or 2D-3D pixel to point consistency for learning image representations in a self-supervised manner. The image representations learned in this way are known to have significantly better performance than learning purely from 2D images for downstream tasks. Despite these great successes, the geometric consistency priors in 3D data (i.e. smooth regions or depth gaps) are not directly employed, which we demonstrate are strong cues and can significantly enhance the learning of image semantics.

In this paper, we propose to use geometric consistency to promote the learning of image representations. Our intuition is that 3D points within the same smooth or even planar regions may share similar semantics, while the discontinuities or depth gaps in 3D space may imply semantic changes. Such geometric cues are directly observable in unlabeled 3D data. However, due to complex textures, these cues can be hardly learned from purely unlabeled 2D images. Based on the above intuition, we design a simple yet effective method to learn the geometric consistency priors described above. Specifically, we leverage the continuous and discontinuities of 3D data, and use the clustering method to cluster the 3D data into many local small segments, termed geometric consistency sets, to guide the learning of image representations in a self-supervised contrastive way. Our method is simple and can be easily implemented. We show that by exploring geometric consistency in the self-supervised pre-training stage, the performance of downstream tasks can be significantly improved.

In the following, we summarize our main contributions:

1. We introduce geometric consistency into a contrastive learning framework for self-supervised image representation learning.
2. We propose a simple yet effective multi-view contrastive loss with geometric consistency sets to leverage the consistency within images.
3. We demonstrate superior performance on several downstream tasks compared with SOTA methods.

2. Related Works

2.1. Scene Understanding

Learning based scene understanding tasks, including semantic scene segmentation [19, 28, 32, 40], object detection [20, 38, 39] and instance segmentation [17, 24, 54], are fundamental tasks in computer vision and have been rapidly developed in recent years. Due to the available of large scale, real scanned 2D and 3D datasets [1, 8, 12, 47], these tasks have been widely explored.

View-based methods [9, 23] mainly rely on 2D images for scene understanding tasks, and the results produced with 2D networks can be further fused to 3D space based on multi-view consistency. In recent years, with the development of 3D deep learning networks [7, 13, 40, 49, 51], the performance of scene understanding tasks has been further promoted. The architectures of 3D networks can be roughly classified into two categories including point-based [40, 51], and sparse-voxel based [7, 13, 49]. These architectures are mainly designed for extracting features from sparse and unordered 3D data and have achieved great success in 3D scene understanding. Besides, since 2D images have clearer textures while 3D data is occlusion free from which extracting structure information is much easier, it is also an important research topic to study the joint training of 2D and 3D data [18, 30, 31, 48] for scene understanding tasks.

In this paper, we propose to use 3D geometric consistency as guidance to promote the learning of 2D image representations and improve the performance of 2D scene understanding tasks.

2.2. Self-Supervised Image Pre-training

Self-supervised image pre-training as a fundamental way for learning representations from a large amount of data has shown significant impact in the field of computer vision and has been proved to be useful for various applications. Researchers in this field mainly aim at designing pretext tasks [25, 35–37] to learn intrinsic and general image representations that may favor various downstream tasks. Contrastive learning [4, 6, 14, 37], due to its effectiveness and simplicity, has attracted great attention in recent years. The key idea of contrastive learning is to enforce the consistency between positive pairs while pushing away negative samples. Mainstream works mainly use various data augmentation strategies to get the positive pairs, including cropping, rotating, Gaussian blurring, etc. Besides, region-wise contrastive learning methods [29, 53] performance contrasting at region level to learn the region level similarities. Recently, researchers start to involve 3D data into contrastive pre-training [18, 30, 31], and the positive samples can be directly induced via multi-view consistency.

In our work, we also explore using 3D data to promote the learning of image representations. In addition to multi-view consistency, we further leverage geometric consistency to enhance the consistency within image views.

2.3. Multimodal Representation Learning

Multimodal representation learning aims at learning joint representations by interacting between the data from different modalities. By incorporating the advantages from different modalities, the representations learned in this way are usually much better than those learned with a single modality only. Vision language pre-training [27, 34, 41] is a successful example. The availability of a large amount of image and language pairs makes the learned representations generalizable to various downstream tasks. Recently,
with the availability of large scale RGB-D datasets, learning representations jointly from 2D and 3D data has attracted great attention in the research community [18, 30, 31, 48]. In particular, Pri3D [18] proposed to use 3D data as guidance for learning 2D to 2D and 2D to 3D pixel-level consistency. P4Contrast [31] proposes to use point-pixel pairs for contrastive learning, and constructs negative samples artificially based on disturbed RGB-D points. TupleInfoNCE [30] compose new negative tuples for contrastive learning using modalities from different scenes to ensure that the weak modalities are not being ignored.

Existing works mainly use pixel-level multi-view consistency of single or multiply modalities. In our work, we use geometric consistency sets as guidance to explore intra-view consistency for 2D representation learning.

2.4. Pseudo Labeling

Pseudo labeling [21, 26, 33, 42, 46], as an effective way for learning with unlabeled data, has been widely studied in various applications. The key to the success of pseudo labeling methods is to generate high-quality pseudo labels for the unlabeled data. Researchers in this field have explored various ways for acquiring pseudo labels, including predicting directly with trained networks [26, 33], neighborhood graph propagation [21] or confidence based selection [42, 46] etc.

The mainstream works mainly use pseudo labels in semi-supervised or weakly supervised learning scenarios. In our work, the geometric consistency sets we used serve as pseudo consistency labels to guide the learning of 2D image representations in a self-supervised manner.

3. Method

Given a large collection of unlabeled images \( \{I_i\} \) together with the corresponding 3D scenes \( \{S_i\} \) represented as 3D points, the goal of our method is to pre-train an image encoder \( f_\theta \) that can extract general and representative features from the input images with the guidance of geometric consistency in a self-supervised manner. And the pre-trained encoder \( f_\theta \) can improve the performance of various downstream tasks after fine-tuning on relatively small labeled datasets. Note that 3D data is only used for pre-training and is not available in the fine-tuning stage.

Figure 2 illustrates our overall framework. In order to learn the image representations given a set of multi-view unlabeled images together with the corresponding 3D data, our observation is that the consistency or inconsistency of the geometric features in 3D space can serve as strong prior that may imply consistent semantics or object boundaries. For example, points that lie in the same smooth or even planar region may share the same semantics, while those separated by geometry dis-continuities or depth gaps may infer semantic changes. Based on the above observation, we propose to incorporate geometric consistency sets, for example clustering the 3D data with similar geometry features into local smooth segments, into a contrastive learning framework to leverage the geometric consistency in learning image representations, and adapt the InfoNCE loss accordingly. In the following, we will first describe the formulation of the set-InfoNCE loss that generalizes the InfoNCE loss to enforce the consistency within image views. And then, we will provide the detailed implementation of our proposed contrastive training framework under the guidance of geometric consistency set.
3.1. Set-InfoNCE Loss

We visualize the geometric consistency sets (with different colors) as well as the corresponding 2D projections on two corresponding image views.

2D contrastive image representation learning mainly aims at learning consistent representations by contrasting multi-view images on either image level or pixel level. The multi-view images can be acquired via different data augmentation strategies or from real-world scanned sequences. Contrastive learning then enforces the feature consistency between different views that correspond to the same content while pushing away the others. For the pixel level contrastive representation learning, current works use pixel-level InfoNCE loss [18], which is defined as:

$$L_{\text{pixel}} = - \sum_{(i,j) \in M_p} \log \frac{\exp(f_i \cdot f_j / \tau)}{\exp(f_i \cdot f_j / \tau)}$$

where $M_p$ denotes the set of pixel-to-pixel matching pairs from one view to another, $f_i$ is the normalized feature vector of the pixel at the $i$th position, and $\tau$ is the temperature parameter for controlling the concentration of the features in the representation space.

Though the representations learned with pixel-level InfoNCE loss [18] has been proven to be beneficial for many downstream tasks, additional aggregated correspondences among the pixels like the geometric consistency prior could further improve the representation learning. In the following, we give the formulation of the set level InfoNCE loss that generalizes the pixel level InfoNCE to leverage such aggregated correspondences:

$$L_{\text{set}} = - \sum_{(i,j) \in M_s} \log \frac{\exp(F(P_i) \cdot F(P_j) / \tau)}{\exp(F(P_i) \cdot F(P_j) / \tau)}$$

where $P_i$ is a set of feature points that are likely to have the same semantics, $F$ is a mapping from a set to a feature vector in $\mathbb{R}^c$, and $M_s$ denotes the correspondence pairs among sets. Obviously Eqn. 1 is a special case to Eqn. 2 when the sets $P_i = \{f_i\}$ degenerate to one-element only. In our implementation, we set $F(P_i) = \frac{1}{|P_i|} \sum_{s \in P_i} f_s$ for aggregating the features from all the points within sets, and we will also discuss different variations of $F$ in Section 3.2.

Note that the set level InfoNCE loss defined in Eqn. 2 is general, and the strategies for acquiring the set $P_i$ might be different for different scenarios [29, 30, 53]. In this paper, we define our geometric consistency set under the guidance of 3D geometry, and will give the detailed description in Section 3.2.

3.2. Learning with Geometric Set Consistency

Intuitively, the geometric consistency can be a strong prior that can guide the learning of within image consistency. The geometric cues like smooth regions and depth gaps may imply the same semantics or object boundaries respectively. In the following, we give the formal definition of the proposed set-InfoNCE loss with geometric consistency sets under projection.

Formally, for a given equivalence relation $\sim$ among the points in all the scenes $\cup S_i$, we call the quotient set $\{P_i\} = \cup S_i / \sim$ as the collection of geometric consistency sets. For example, when we have some pre-defined geometric labels among the 3D points and define the equivalence relation as points that share the same label, then the geometric consistency set $P_i$ would contain all the points that have the $i$th label. Besides, for each image $I_m$, we call $P_m^i = \{proj(s) \in I_m | s \in P_i\}$ as the projection of $P_i$ from 3D onto 2D image view $I_m$. By adapting Eqn. 2, we could define our set-InfoNCE loss with geometric consistency set as:

$$L_{\text{geo-set}} = - \sum_{(i,m,n) \in M_s} \log \frac{\exp(F(P_i^m) \cdot F(P_i^n) / \tau)}{\exp(F(P_i^m) \cdot F(P_i^n) / \tau)}$$

where $i$ is the index of the geometric consistency set and $m,n$ are the view indices. $M_s = \{(i,m,n)\}$ maintains the matching pairs of projections $P_i^m,P_i^n$ from one view to another for each geometric consistency set $P_i$.

In general, any proper notion of spatial equivalence induces its corresponding geometric consistency sets; in this work, we use a simple 3D clustering method and find it already works well in our case. We leave the exploration of more sophisticated spatial partitioning methods as future work. Specifically, for a sequence of 2D images acquired by scanning around a specific scene, we first have the corresponding 3D surface $S$ via 3D reconstruction. To obtain geometric consistency sets, we use the surface over-segmentation results produced by a normal-based graph cut method [11, 22]. As shown in Figure 3, the 3D points that are likely to have the same semantics can be clustered into the same geometric consistency set.
3.3. Training Strategy

We use a two-stage training strategy to pre-train with our method. Besides our geometric guided set-InfoNCE loss that mainly considers set-level consistency and involves some high-level semantics such as smooth regions and surface discontinuities, we also include the pixel-level multi-view consistency to enforce the low-level distinction beneficial for the downstream tasks. Thus we propose to pre-train the network from low-level to high-level progressively with two stages. Specifically, in the first stage, we train the network with pixel to pixel and pixel to point multi-view contrastive loss termed as View and Geo loss in Pri3D [18]. And then, in the second stage, we continue to train the network with our geometric guided set-InfoNCE loss alone. In this way, we can achieve better performance on downstream tasks than training solely with our set-InfoNCE loss or with Pri3D losses.

4. Experimental Setup

In this section, we will provide a detailed description of our experimental setup, including the backbone architectures, datasets for pre-training and evaluation, and the implementation details. More experiment details are given in the supplementary materials.

4.1. Backbone Architecture

We use a UNet-style [44] architecture with residual connections as our network backbone since it has been widely used in various deep learning tasks. Specifically, it uses ResNet [16] as encoder, and the decoder part contains standard convolution blocks together with bi-linear interpolation layers for upsampling. In most of our experiments, we use ResNet50 as our encoder for its efficiency and effectiveness. In addition, we also use ResNet18 as the encoder to see the effectiveness of our method with a more lightweight architecture.

4.2. Dataset

We conduct our experiments on two widely used datasets ScanNet [8] and NYUv2 [47], where ScanNet is used for both pre-training and downstream task fine-tuning, and NYUv2 is used for demonstrating the transferability of the representations learned with our method on ScanNet. In the following, we give more details about the two datasets.

ScanNet ScanNet [8] contains large amount of RGB-D sequences together with the corresponding 3D scene reconstructions. It has 1513 scenes for training, which contains around 2.5M images in total. Following Pri3D [18], we regularly sample every 25th frame from the original ScanNet sequence for pre-training. And the frame pairs with more than 30% pixel overlap will be used for contrastive training. Totally, we have around 804k frame pairs in the pre-training stage. Note that we do not use any semantic labels for pre-training. For fine-tuning on downstream tasks, we use the standard labeled training and validation set of ScanNet benchmark [8], and the 2D images are sampled every 100th frame from the original sequence. Thus, in the fine-tuning stage, we have around 20k images for training and 5k images for validation.

NYUv2 NYUv2 [47] is a dataset consisting of 2D images scanned from various indoor scenes, and in this paper, it is used for fine-tuning downstream tasks. Totally it has 1449 labeled images, in which 795 images are used for training and 654 images for testing.

4.3. Comparison Methods

Supervised ImageNet Pre-training [10] The network is pre-trained with the supervised classification task on the ImageNet dataset.

MoCoV2 [5] For MoCoV2, there are various strategies for feeding the data, we only list the best performing result (i.e. MoCoV2-supIN → SN) from Pri3D [18]. To be specific, MoCoV2 is initialized with supervised ImageNet pre-training and then trained on ScanNet with randomly shuffled images.

Pri3D [18] Pri3D is the method most related to ours, in which 3D data and contrastive pre-training are used. Specifically, it comprises 2D to 2D and 2D to 3D point InfoNCE loss, termed View and Geo loss, respectively, and is also initialized with supervised ImageNet pre-training.

Depth Prediction The network is pre-trained with a single view depth prediction task, and is also initialized with supervised ImageNet pre-training. We obtain the results directly from Pri3D [18] where the method is implemented as a baseline.

4.4. Implementation Details

We use the following setting in our experiments unless explicitly specified. For pre-training with our method, Stochastic gradient descent (SGD) [43] with Polynomial Learning Rate Policy (PolyLR) [45] is used for optimization. The initial learning rate is 0.1 and the batch size is 64. Following Pri3D [18], we initialize the network with ImageNet pre-trained weights. In the first stage of pre-training, we train the network with View and Geo loss from Pri3D for 5 epochs. And then in the second stage, the network is trained with our proposed methods for 2 epochs. We use 8 NVIDIA V100 GPUs for training our network, and all the experiments are implemented with PyTorch.
Figure 4. Qualitative results of semantic segmentation task on ScanNet and NYUv2 datasets. All the methods are pre-trained on ScanNet with ResNet50 as the backbone. For Pri3d we reproduce the results with their released code.

| Method         | ResNet50 | ResNet18 |
|----------------|----------|----------|
| From Scratch   | 39.1     | 37.5     |
| ImageNet Pre-training | 55.7     | 51.0     |
| MoCoV2         | 56.6     | 52.9     |
| Depth Prediction | 58.4     | -        |
| Pri3D(View)    | 61.3     | 54.4     |
| Pri3D(Geo)     | 61.1     | 55.3     |
| Pri3D(View + Geo) | 61.7     | 55.7     |
| Ours           | 63.1     | 57.2     |

Table 1. 2D semantic segmentation results on ScanNet. We use ResNet50 and ResNet18 as the backbone, mIoU for evaluation.

5. Experimental Results

5.1. Fine-tuning for 2D Downstream Tasks

In the following, we demonstrate the effectiveness of our method by fine-tuning the pre-trained network on several 2D downstream tasks, including semantic segmentation, instance segmentation, and object detection.

Semantic Segmentation on ScanNet We demonstrate the effectiveness of our pre-trained representation by fine-tuning it with the ScanNet standard labeled training set, and mean IoU is used as the evaluation metric. ResNet50 or ResNet18 is used as the backbone encoder for all pre-training methods. In Table 1, we compare our results with other state-of-the-art pre-training methods. Our method outperforms all the state-of-the-art methods for both ResNet50 and ResNet18 backbones, which confirms the effectiveness of our proposed self-supervised pre-training strategy with geometry-guided set consistency. Figure 4 demonstrates some qualitative results, as we can see that the segmentation results of our method are less noisy and have clearer details.

We further test our method with different amounts of data for fine-tuning to study the performance under limited data scenarios. ResNet50 is used as the backbone. As shown in Table 2, by comparing our method with the best performing SOTA method, we can achieve superior performance consistently under different ratios of training data.

Moreover, we also test our pre-trained representations...
Table 3. **2D semantic segmentation results on ScanNet with different network architectures.** ResNet50 is used as the backbone for all architectures, and mIoU is used for evaluation.

| Method                  | ResNet50 |
|-------------------------|----------|
| DeepLabV3(ImageNet)     | 57.0     |
| DeepLabV3(Pri3D)        | 61.3     |
| DeepLabV3(Ours)         | **62.2** |
| DeepLabV3+(ImageNet)    | 57.8     |
| DeepLabV3+(Pri3D)       | 61.6     |
| DeepLabV3+(Ours)        | **62.7** |
| PSPNet(ImageNet)        | 59.7     |
| PSPNet(Pri3D)           | 62.8     |
| PSPNet(Ours)            | **63.7** |

Table 4. **2D object detection results on ScanNet.** We use ResNet50 as the backbone and average precision for evaluation.

| Method                  | AP@0.5 | AP@0.75 | AP  |
|-------------------------|--------|---------|-----|
| From Scratch            | 32.7   | 17.7    | 16.9|
| ImageNet Pre-training   | 41.7   | 25.9    | 25.1|
| MoCoV2                  | 43.5   | 26.8    | 25.8|
| Pri3D(View)             | 43.7   | 27.0    | 26.3|
| Pri3D(Geo)              | 44.2   | **27.6**| 26.6|
| Pri3D(View + Geo)       | 44.5   | 27.4    | 26.6|
| Ours                    | **45.1**| **27.6**| **26.9**|

Table 5. **2D instance segmentation results on ScanNet.** We use ResNet50 as the backbone and average precision for evaluation.

| Method                  | AP@0.5 | AP@0.75 | AP  |
|-------------------------|--------|---------|-----|
| From Scratch            | 25.8   | 13.1    | 12.2|
| ImageNet Pre-training   | 32.6   | 17.8    | 17.6|
| MoCoV2                  | 33.9   | 18.1    | 18.3|
| Pri3D(View)             | 34.3   | 18.7    | 18.3|
| Pri3D(Geo)              | 34.4   | 18.7    | 18.3|
| Pri3D(View + Geo)       | 35.8   | **19.3**| 18.7|
| Ours                    | **36.0**| **19.3**| **19.5**|

Table 6. **2D semantic segmentation results on NYUv2.** mIoU is used as the evaluation metric.

| Method                  | ResNet50 |
|-------------------------|----------|
| From Scratch            | 24.8     |
| ImageNet Pre-training   | 50.0     |
| MoCoV2                  | 47.6     |
| Pri3D(View)             | 54.2     |
| Pri3D(Geo)              | 54.8     |
| Pri3D(View + Geo)       | 54.7     |
| Ours                    | **55.4** |

Table 7. **2D object detection results on NYUv2.** We use ResNet50 as the backbone, and average precision for evaluation.

| Method                  | AP@0.5 | AP@0.75 | AP  |
|-------------------------|--------|---------|-----|
| From Scratch            | 21.3   | 10.3    | 9.0 |
| ImageNet Pre-training   | 29.9   | 17.3    | 16.8|
| MoCoV2                  | 30.1   | 18.1    | 17.3|
| Pri3D(View)             | 33.0   | 19.8    | 18.9|
| Pri3D(Geo)              | 33.8   | 20.2    | 19.1|
| Pri3D(View + Geo)       | 34.0   | 20.4    | 19.4|
| Ours                    | **34.6**| **20.5**| **19.7**|

**Instance Segmentation and Detection on ScanNet** We fine-tune our pre-trained representations on 2D ScanNet object detection and instance segmentation tasks to see the generalizability of the learned representations. Specifically, ResNet50 is used as the backbone encoder for all the pre-training methods. Mask-RCNN [15] implemented by Detectron2 [52] is used for object detection and instance segmentation tasks. As shown in Table 4 and 5, we can achieve comparable or superior performance compared with the other state-of-the-arts on different evaluation metrics.

**Transfer to NYUv2** We demonstrate the transferability of our pre-trained representations to other datasets. Specifically, we use ResNet50 as the backbone encoder for all pre-training methods. The network is pre-trained on ScanNet [8] dataset, and fine-tuned for downstream tasks on NYUv2 [47] dataset. Following Pri3D [18], we use the learning rate 0.01 instead of 0.1 for semantic segmentation task. As demonstrated in Table 6, 7 and 8, our pre-trained representations achieve superior performance compared with SOTA methods on most evaluation metrics, which further confirms that our pre-trained image representations are general and transferable across different datasets.

**5.2. Ablation Study**

**Different Implementation of Set Feature Aggregation** We study a different choice of the set feature aggregation function $F$ in Eqn. 3 to see whether it will influence the performance. Instead of using the average function to aggre-
Table 8. 2D instance segmentation results on NYUv2. We use ResNet50 as the backbone, and average precision for evaluation.

| Method              | AP@0.5 | AP@0.75 | AP  |
|---------------------|--------|---------|-----|
| From Scratch        | 17.2   | 9.2     | 8.8 |
| ImageNet Pre-training | 25.1   | 13.9    | 13.4|
| MoCoV2              | 27.2   | 14.7    | 14.8|
| Pri3D(View)         | 28.1   | 15.7    | 15.7|
| Pri3D(Geo)          | 29.0   | 15.9    | 15.2|
| Pri3D(View + Geo)   | 29.5   | 16.3    | 15.8|
| Ours                | 29.7   | 16.3    | 16.5|

Table 10. Ablation of Pri3D. Continuous training of different configurations of Pri3D can not lead to better results. Pri3D Continue means continuous training of Pri3D.

Ablation of Pri3D

Since in the first stage of our method we use View and Geo loss in Pri3D [18] to train the initial representations, one may wonder whether training Pri3D for more epochs will lead to better performance compared with the results listed in their original paper. To address this concern, we test thoroughly by continuing training from the Pri3D checkpoint with different combinations of their losses for two more epochs. And then fine-tune the network on ScanNet semantic segmentation task [8]. Totally, in this experiment, Pri3D is pre-trained for 7 epochs. As shown in Table 10, continuous training of different configurations of Pri3D can not lead to better results.

6. Limitations & Future Work

While our proposed pre-training method demonstrates superior performance compared with the state-of-the-art methods, there still exist several limitations. Since our approach relies on 3D geometric consistency to guide the learning of image representations, it can hardly be applied directly on the datasets such as ImageNet [10], where 3D reconstructions are not applicable. Besides, in this work, we use a simple 3D clustering method for computing geometric consistency sets; one may consider using more advanced clustering techniques to further improve the performance.

7. Conclusion

We propose a method that uses 3D geometric consistency as guidance for self-supervised image representation learning. We leverage the continuities and discontinuities of 3D data and use the 3D clustering results to produce geometric consistency sets for 2D image views. Then we incorporate the geometric consistency sets into a contrastive learning framework to make the learned 2D image representations aware of 3D geometric consistency. We show that the image representations learned in this way will lead to superior performance than state-of-the-arts after fine-tuning for downstream tasks. Moreover, our approach can also improve the performance of downstream tasks when fine-tuning with limited training data. Various ablation studies further verify the effectiveness of our method.
References

[1] Iro Armeni, Ozan Sener, Amir R Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1534–1543, 2016. 2

[2] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587, 2017. 1, 7

[3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European conference on computer vision (ECCV), pages 801–818, 2018. 1, 7

[4] Xinlei Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In International conference on machine learning, pages 1597–1607. PMLR, 2020. 2

[5] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297, 2020. 5

[6] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 15750–15758, 2021. 2

[7] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3075–3084, 2019. 2

[8] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In Proc. Computer Vision and Pattern Recognition (CVPR), IEEE, 2017. 2, 5, 7, 8

[9] Angela Dai and Matthias Nießner. 3dmv: Joint 3d-multiview prediction for 3d semantic scene segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), pages 468–468, 2018. 2

[10] Jià Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pages 248–255. Ieee, 2009. 5, 8

[11] Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image segmentation. International journal of computer vision, 59(2):167–181, 2004. 2, 8

[12] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. The International Journal of Robotics Research, 32(11):1231–1237, 2013. 2

[13] Benjamin Graham, Martin Engelcke, and Laurens Van Der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 9224–9232, 2018. 2

[14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9729–9738, 2020. 1, 2

[15] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In Proceedings of the IEEE international conference on computer vision, pages 2961–2969, 2017. 1, 7

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016. 5

[17] Ji Hou, Angela Dai, and Matthias Nießner. 3d-sis: 3d semantic instance segmentation of rgb-d scans. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4421–4430, 2019. 2

[18] Ji Hou, Saining Xie, Benjamin Graham, Angela Dai, and Matthias Nießner. Pr3d: Can 3d priors help 2d representation learning? In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 5693–5702, 2021. 1, 2, 3, 4, 5, 7, 8

[19] Wenbo Hu, Hengshuang Zhao, Li Jiang, Jiaya Jia, and Tien-Tsin Wong. Bidirectional projection network for cross dimension scene understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14373–14382, 2021. 2

[20] Zeyu Hu, Mingmin Zhen, Xuyang Bai, Hongbo Fu, and Chiew-lan Tai. Jsenet: Joint semantic segmentation and edge detection network for 3d point clouds. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16, pages 222–239. Springer, 2020. 2

[21] Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondrej Chum. Label propagation for deep semi-supervised learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5070–5079, 2019. 3

[22] Andrej Karpathy, Stephen Miller, and Li Fei-Fei. Object discovery in 3d scenes via shape analysis. In 2013 IEEE International Conference on Robotics and Automation, pages 2088–2095. IEEE, 2013. 4, 8

[23] Abhijit Kundu, Xiaoxi Yin, Alireza Fathi, David Ross, Brian Brewington, Thomas Funkhouser, and Caroline Pantofaru. Virtual multi-view fusion for 3d semantic segmentation. In European Conference on Computer Vision, pages 518–535. Springer, 2020. 2

[24] Jean Lahoud, Bernard Ghanem, Marc Pollefeys, and Martin R Oswald. 3d instance segmentation via multi-task metric learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 9256–9266, 2019. 2

[25] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Colorization as a proxy task for visual understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 6874–6883, 2017. 2

[26] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural net-
[53] Tete Xiao, Colorado J Reed, Xiaolong Wang, Kurt Keutzer, and Trevor Darrell. Region similarity representation learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 10539–10548, 2021.

[54] Li Yi, Wang Zhao, He Wang, Minhyuk Sung, and Leonidas J Guibas. Gspn: Generative shape proposal network for 3d instance segmentation in point cloud. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3947–3956, 2019.

[55] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In CVPR, 2017.