Spatio-temporal Self-Supervised Representation Learning for 3D Point Clouds

Siyuan Huang[1,*], Yichen Xie[2,*], Song-Chun Zhu[3,4,5], Yixin Zhu[3,4]

[1] University of California, Los Angeles [2] Shanghai Jiao Tong University
[3] Beijing Institute for General Artificial Intelligence [4] Peking University [5] Tsinghua University

ICCV 2021 OCTOBER 11-17 VIRTUAL

Project Page: https://siyuanhuang.com/STRL

Objective

Practical and generalizable pre-trained 3D models that capable of learning from unlabeled 3D point clouds in a self-supervised fashion.

Challenge

- **Simplicity.** How to design a simple self-supervised learning model without dense reconstruction of the 3D point cloud?
- **Invariance.** How could we introduce and leverage the invariance in 3D point clouds for self-supervised learning?
- **Generalizability.** How to demonstrate sufficient generalizability to higher-level tasks (e.g., 3D object detection)?

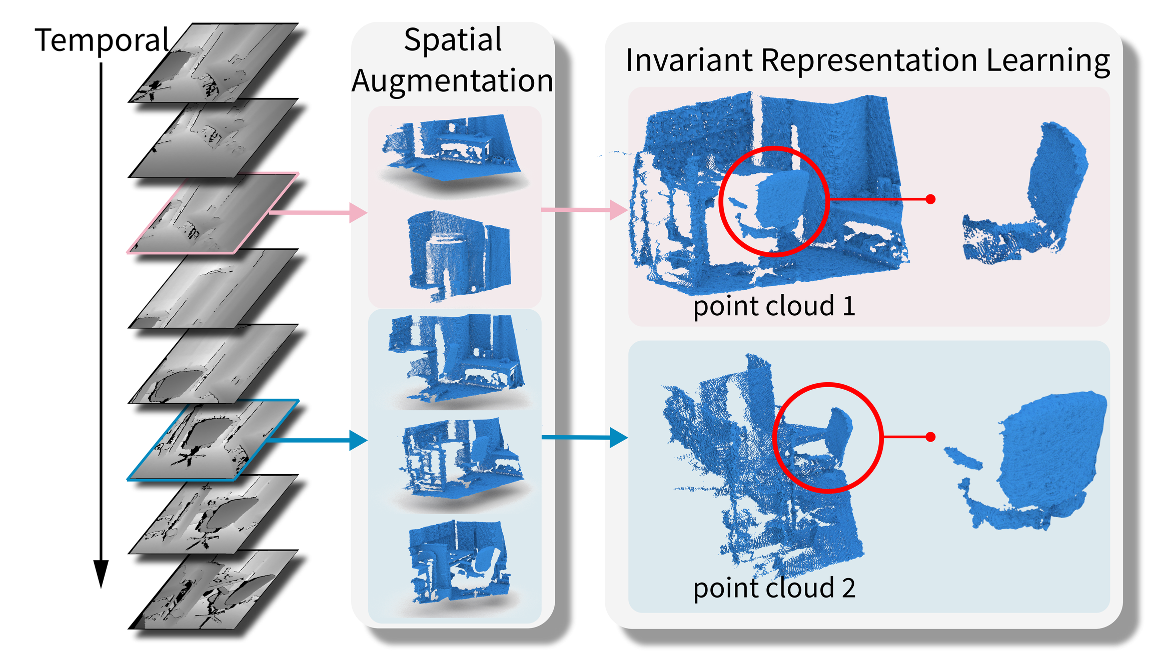
Approach

- A **spatio-temporal representation learning (STRL)** framework to learn from unlabeled 3D point clouds.
- **Remarkably simple** by learning only from the positive pairs. STRL uses two neural networks, referred to as online and target networks, that interact and learn from each other.
- STRL takes two **temporally-correlated** frames from a 3D point cloud sequence as the input, transforms it with the **spatial data augmentation**, and learns the **invariant representation** self-supervisedly.
- **Effective generalization** to downstream 3D scene understanding tasks directly or with additional fine-tuning.

Contribution & Discovery

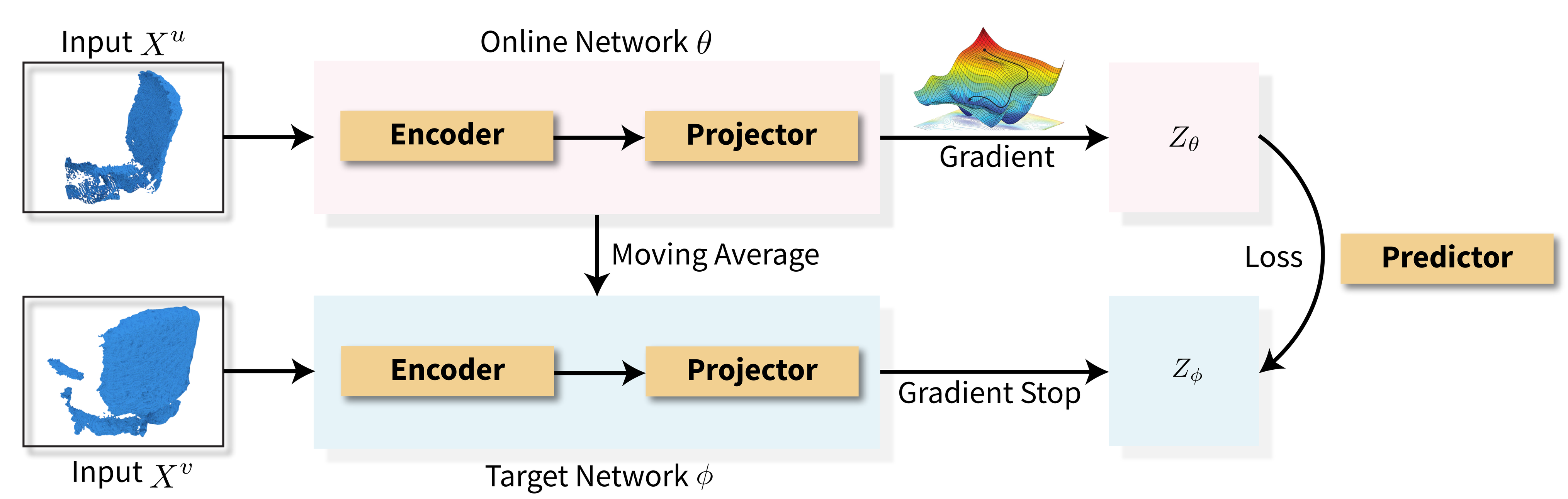
1. **Our method outperforms prior arts** in (i) unsupervised 3D shape learning, (ii) semi-supervised 3D shape learning with limited data, and (iii) transferring to downstream tasks such as 3D object detection and semantic segmentation.
2. **Simple learning and augmentation strategy** leads to the satisfying performance of learned 3D representation.
3. **The spatio-temporal cues boost the performance of learned representation.** Relying on spatial or temporal augmentation alone only yield relatively low performance.
4. **Pre-training on synthetic 3D shapes (ShapeNet) is indeed helpful** for real-world applications.

Overview



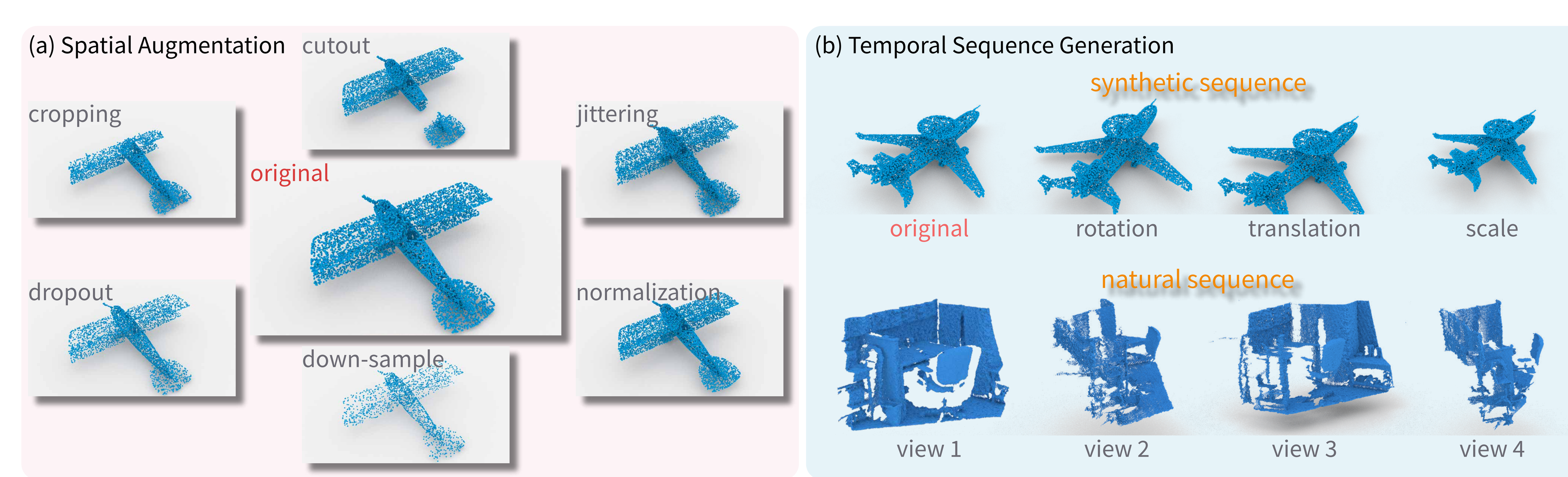
By learning the **spatio-temporal data invariance** from a point cloud sequence, our method self-supevisedly learns an effective representation.

Framework



Given two **spatio-temporal correlated 3D point clouds**, the online network predicts the target network's representation via a predictor. Parameters of the target network are updated by the online network's moving average.

Data Augmentation



Spatial data augmentation and temporal sequence generation. Except for the natural sequence generation, each type of augmentation transforms the input point cloud data stochastically with certain internal parameters.

Shape Classification on ModelNet with Pre-trained Model

Method	ModelNet40
3D-GAN [69]	83.3%
Latent-GAN [1]	85.7%
SO-Net [38]	87.3%
FoldingNet [75]	88.4%
MRTNet [21]	86.4%
3D-PointCapsNet [75]	88.9%
MAP-VAE [75]	88.4%
Sauder *et al.* + PointNet [53]	87.3%
Sauder *et al.* + DGCNN [53]	90.6%
Poursaeed *et al.* + PointNet [46]	88.6%
Poursaeed *et al.* + DGCNN [46]	90.7%
STRL + PointNet (ours)	88.3%
STRL + DGCNN (ours)	**90.9%**

(a) Linear evaluation results.

(a) Fine-tuned on Full Training Set

Category	Method	Accuracy
Supervised	PointNet [48]	89.2%
	PointNet++ [49]	90.7%
	PointCNN [39]	92.2%
	DGCNN [67]	92.2%
	ShellNet [78]	93.1%
Self-supervised	Sauder *et al.* + DGCNN [53]	92.4%
	STRL + DGCNN (ours)	**93.1%**

(b) Fine-tuned on Few Training Samples

Method	1%	5%	10%	20%
DGCNN	58.4%	80.7%	85.2%	88.1%
STRL + DGCNN	**60.5%**	**82.7%**	**86.5%**	**89.7%**

(b) Fine-tuned results using limited training samples.

Results on Scene Understanding Tasks

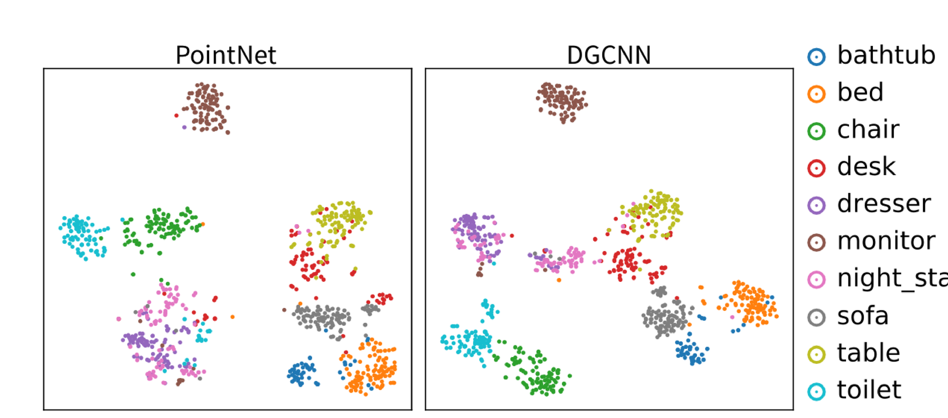
Model	Method	Input	mAP@0.25 IoU
VoteNet	*from scratch*	Geo+Height	57.7
		Geo	57.0
SR-UNet [9]	PointContrast [73]	Geo	57.5
VoteNet	STRL (ours)	Geo	**58.2**

(a) 3D object detection fine-tuned on SUN RGB-D.

Method	Car (IoU=0.7) 3D	Car (IoU=0.7) BEV	Pedestrian 3D	Pedestrian BEV	Cyclist 3D	Cyclist BEV
PV-RCNN (*from scratch*)	84.50	90.53	57.06	59.84	70.14	75.04
STRL + PV-RCNN (*frozen backbone*)	81.63	87.84	39.62	42.41	69.65	74.20
STRL + PV-RCNN	**84.70**	**90.75**	**57.80**	**60.83**	**71.88**	**76.65**

(b) 3D object detection fine-tuned on KITTI.

Visualization of Learned Feature



The extracted features for each sample in ModelNet10 test set using t-SNE. Both models are pre-trained on ShapeNet.