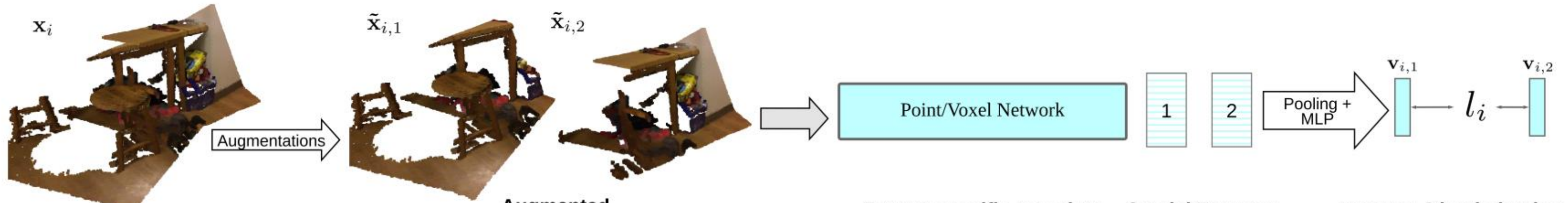
# DepthContrast: Learning Self-supervised 3D Features from Single-view Depth Scans

Zaiwei Zhang[1], Rohit Girdhar[2], Armand Joulin[2], Ishan Misra[2]
[1]University of Texas at Austin [2]Facebook AI Research

ICML International Conference On Machine Learning



Input Depth map — Augmented versions — Format-specific encoders — Spatial Features — Instance Discrimination

## Motivation

- Expensive 3D data labeling
- Availability of existing large collection of single-view depth scans
- More commercial 3D sensors will lead to more unlabeled single-view 3D data

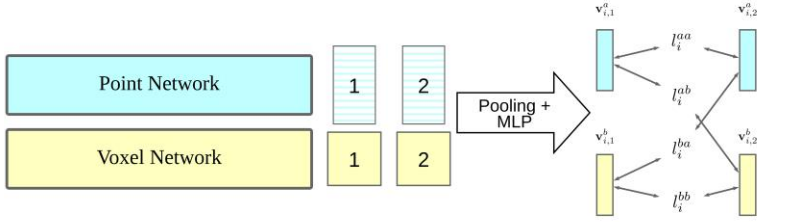
## Key Takeaways

- Works with single/multi-view depth scans acquired by varied sensors (Lidar or Kinect)
- Works on point cloud and voxel-based model architectures
- Improves label efficiency on downstream tasks

## Method & Formulation

$$l_i = -\log \frac{\exp(\mathbf{v}_{i,1}^\top \mathbf{v}_{i,2}/\tau)}{\exp(\mathbf{v}_{i,1}^\top \mathbf{v}_{i,2}/\tau) + \sum_{j\neq i}^{K} \exp(\mathbf{v}_{i,1}^\top \mathbf{v}_j/\tau)}$$

## Extension to Multiple 3D Input Formats



Format-specific encoders — Spatial Features — Instance Discrimination

***Final Loss:***

$$L_i = \underbrace{l_i^{ab} + l_i^{ba}}_{\text{across format}} + \underbrace{l_i^{aa} + l_i^{bb}}_{\text{within format}}$$

## Main Results

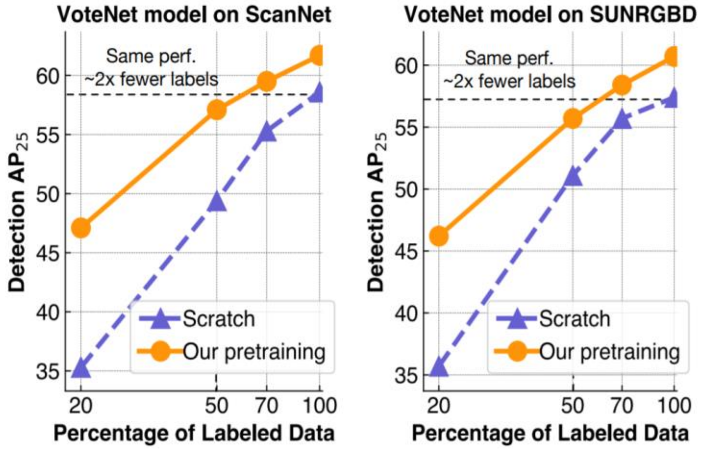
| Dataset | Stats | Task | Gain of DepthContrast |
|---|---|---|---|
| Self-supervised Pretraining | | | |
| ScanNet-vid (Dai et al., 2017) | 190K single-view depth maps (Indoor) | | |
| Redwood-vid (Choi et al., 2016) | 370K single-view depth maps (Indoor/Outdoor) | | |
| Transfer tasks | | | |
| ScanNet (Dai et al., 2017) | 1.2K train, 312 val (Indoor) | Det. | **+3.6% mAP** |
| | | Seg. | **+0.9% mIOU†** |
| SUNRGBD (Song et al., 2015) | 5.2K train, 5K val (Indoor) | Det | **+3.3% mAP** |
| S3DIS (Armeni et al., 2017) | 199 train, 67 val (Indoor) | Det | **+12.1% mAP** |
| | | Seg. | **+2.4% mIOU** |
| Synthia (Ros et al., 2016) | 19.8K train, 1.8K val (Synth.) | Seg. | **+2.4% mIOU** |
| Matterport3D (Chang et al., 2017) | 1.4K train, 232 val (Indoor) | Det. | **+3.9% mAP** |
| ModelNet (Wu et al., 2015) | 9.8K train, 2.4K val (Synth.) | Cls. | **+3.1% Acc†** |

## Baseline Comparison

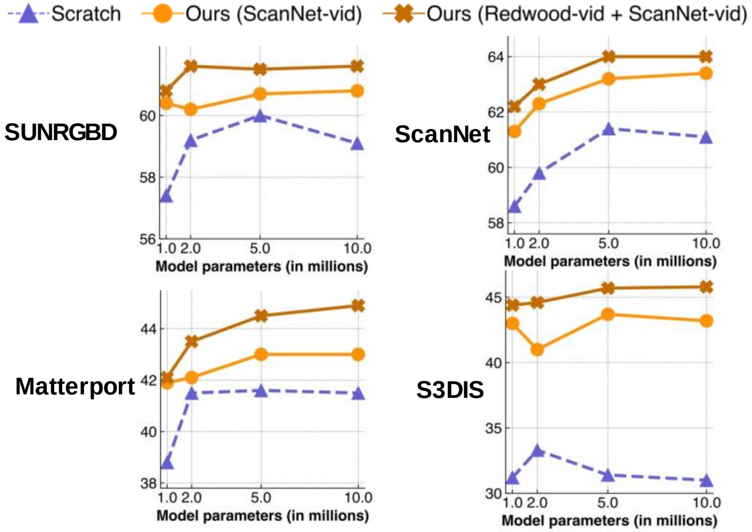
| Initialization | ScanNet | SUNRGBD | Matterport3D | S3DIS |
|---|---|---|---|---|
| Scratch | 58.6 | 57.4 | 38.8 | 31.2 |
| Supervised | - | 59.1 (+1.7) | 41.7 (+2.9) | **48.5** (+17.3) |
| DepthContrast (Ours) | **61.3** (+2.7) | **60.4** (+3.0) | **41.9** (+3.1) | 43.3 (+12.1) |
| PointContrast (Xie et al., 2020) | 59.2(+2.5) | 57.5(+1.9) | - | - |

| Loss | Point Transfer | | Voxel Transfer | |
|---|---|---|---|---|
| | SUNRGBD | ScanNet | S3DIS | Synthia |
| Scratch | 57.4 | 58.6 | 68.2 | 78.9 |
| Within Format only | 60.4 (+3.0) | 61.3 (+1.7) | 66.5 (-2.7) | 80.1 (+1.2) |
| Across format only | 60.0 (+2.6) | 61.1 (+2.5) | 69.9 (+1.7) | 81.2 (+2.3) |
| Both (Ours) | **60.7** (+3.3) | **62.2** (+3.6) | **70.6** (+2.4) | **81.3** (+2.4) |
| PointContrast (Xie et al., 2020) | 59.2(+2.5) | 57.5(+1.9) | **70.9**(+2.7) | **83.1**(+3.3) |

## Benefits on Label Efficiency


VoteNet model on ScanNet — VoteNet model on SUNRGBD

## Scaling on Model & Pretraining Data



Code Link: https://github.com/facebookresearch/DepthContrast