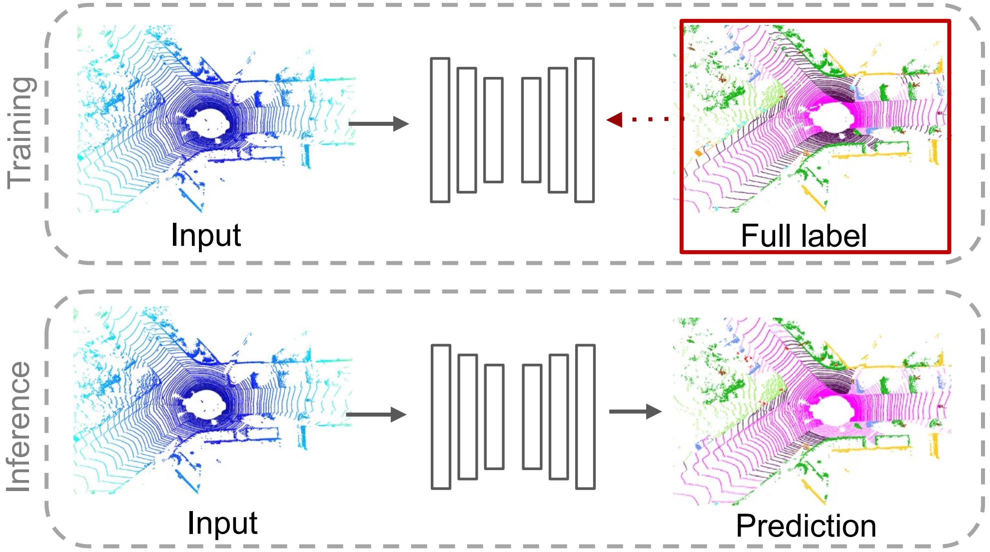
1 2

# COARSE3D: Class-Prototypes for Contrastive Learning in Weakly-Supervised 3D Point Cloud Segmentation

Rong Li[1] Anh-Quan Cao[2] Raoul de Charette[2]

**Arxiv:** https://arxiv.org/abs/2210.01784
**GitHub:** https://github.com/cv-rits/COARSE3D

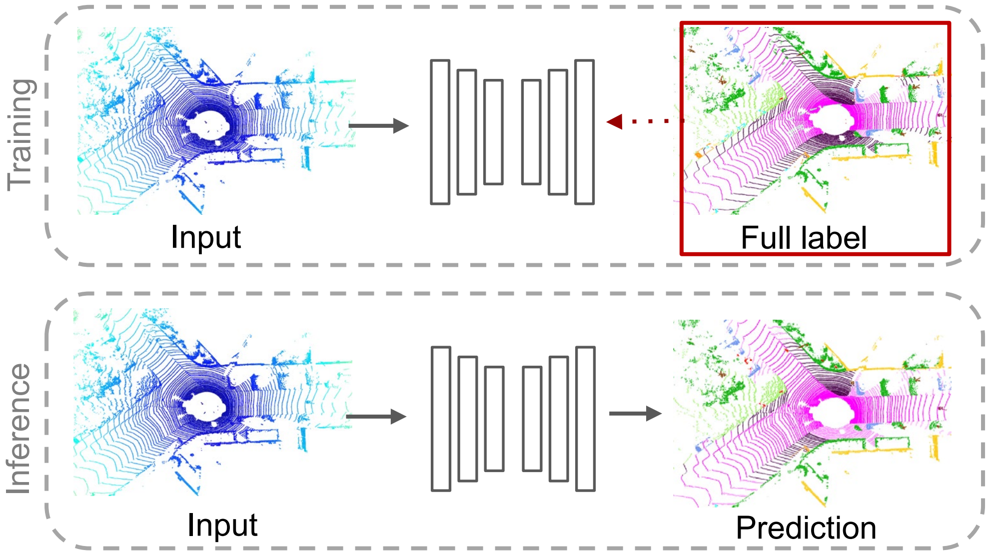
# Problem Statement

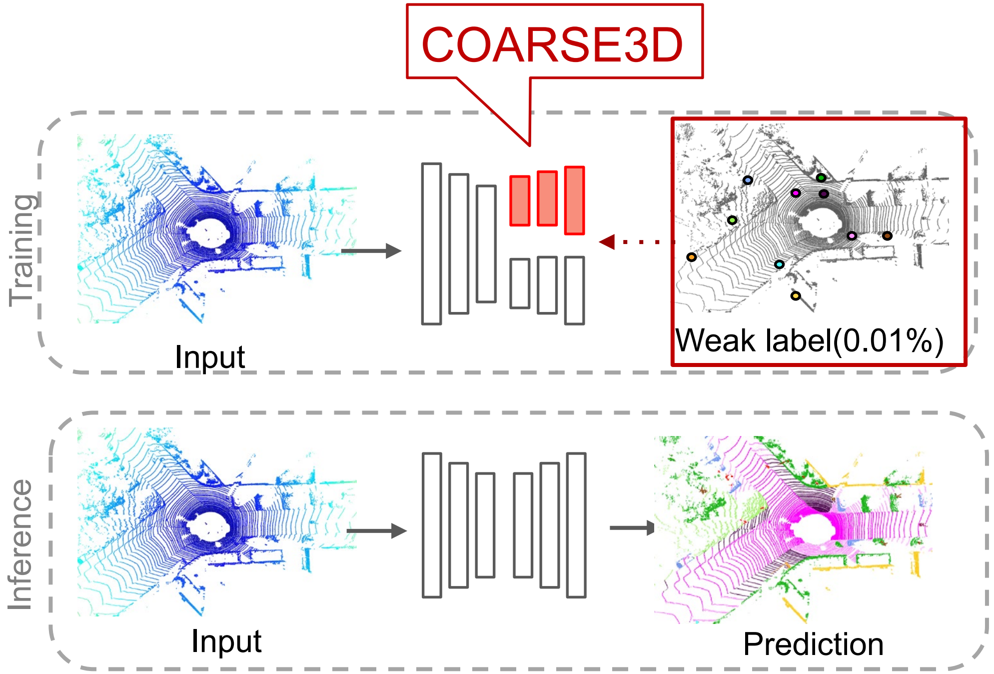



**Fully supervised LiDAR semantic segmentation**

- Annotation of large-scale 3D data is cumbersome and costly (e.g. 1700 hours for SemanticKITTI. )
- 3D annotation requires constant view rotation, more complex than 2D.

# Problem Statement



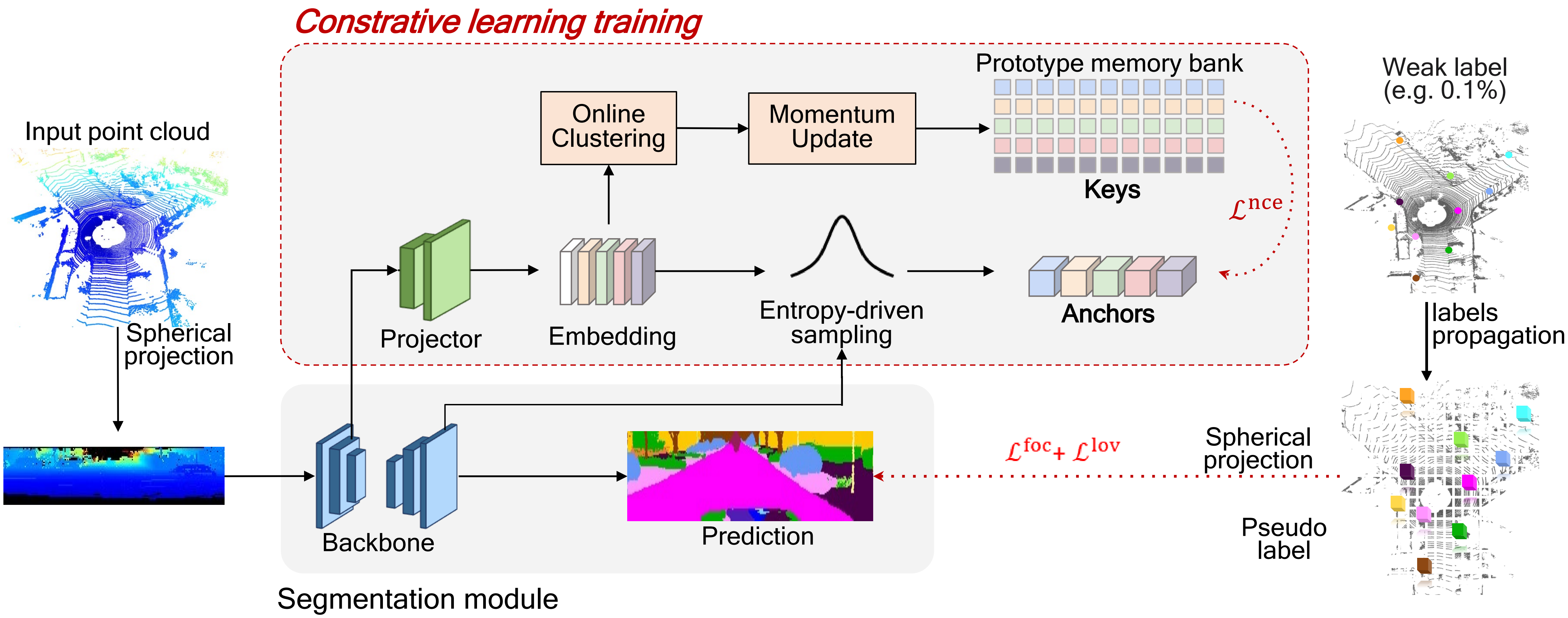

**Fully supervised LiDAR semantic segmentation**

- Annotation of large-scale 3D data is cumbersome and costly (e.g. 1700 hours for SemanticKITTI. )
- 3D annotation requires constant view rotation, more complex than 2D.

**Weakly supervised LiDAR semantic segmentation**

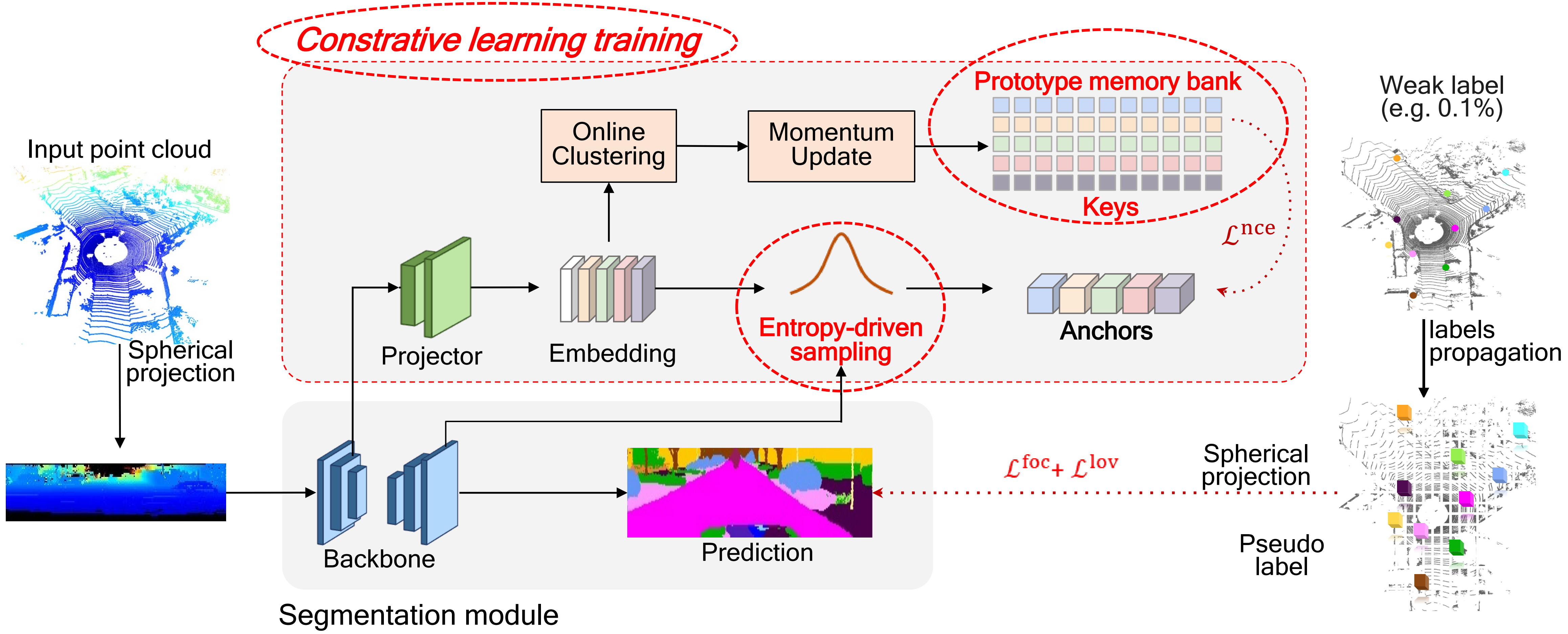
- Train the model using weak label, e.g. 0.01%.
- Prediction at full 100% ratio.

# Overview



Constrative learning training
Input point cloud
Spherical projection
Online Clustering
Momentum Update
Prototype memory bank
Keys
$\mathcal{L}^{\text{nce}}$
Weak label (e.g. 0.1%)
Projector
Embedding
Entropy-driven sampling
Anchors
labels propagation
$\mathcal{L}^{\text{foc}} + \mathcal{L}^{\text{lov}}$
Spherical projection
Pseudo label
Backbone
Prediction
Segmentation module

# Overview

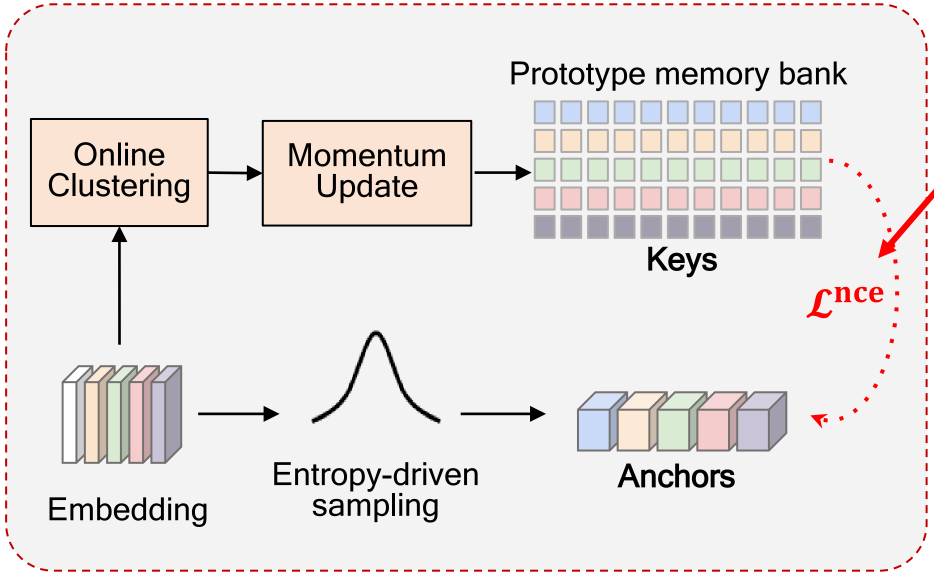

Constrative learning training
Input point cloud
Spherical projection
Online Clustering
Momentum Update
Prototype memory bank
Keys
Weak label (e.g. 0.1%)
$\mathcal{L}^{nce}$
Projector
Embedding
Entropy-driven sampling
Anchors
labels propagation
Backbone
Prediction
$\mathcal{L}^{foc} + \mathcal{L}^{lov}$
Spherical projection
Pseudo label
Segmentation module

# Method

## 1. Pixel-prototype-based contrastive loss ( $\mathcal{L}^{\text{nce}}$ )

From [27, 54, 81], contrastive learning helps 3d label-limited tasks.

*Constrative learning training*




$$\mathcal{L}^{\text{nce}} = \frac{1}{N_a} \sum_{a_i \in \mathcal{A}} -\log \frac{\sum_{p_j^+ \in \mathcal{P}^+} \exp(a_i \cdot p_j^+)}{\sum_{p_j^+ \in \mathcal{P}^+} \exp(a_i \cdot p_j^+ / \tau) + \sum_{p_j^+ \in \mathcal{P}^-} exp(a_i \cdot p_j^- / \tau)}$$

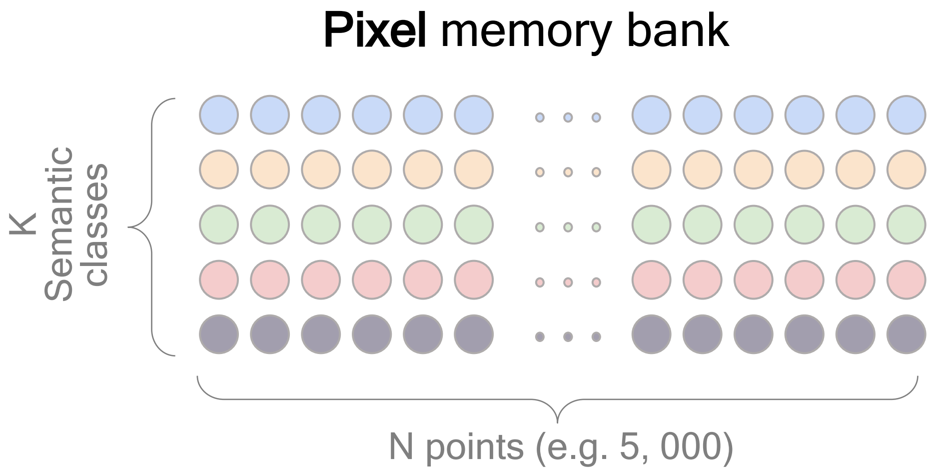
- Anchors ($\mathcal{A}$)
  - Sampled point-wise features from prediction
- Keys
  - Positive keys ($\mathcal{P}^+$): prototypes with same semantic
  - Negative keys ($\mathcal{P}^-$): prototypes with different semantic


[27] Hou et al. Exploring data-efficient 3d scene understanding with contrastive scene contexts. CVPR 2021
[54] David et al. Language-grounded indoor 3d semantic segmentation in the wild. ECCV 2022
[81] Xie et al. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. ECCV 2020

# Method

## 2. Prototype memory bank

From [26, 70], contrastive learning requires massive data to learn good representation




- Semantically redundant
- Costly in memory and computation
  e.g. (K, N, dim)


[26] He et al. Momentum contrast for unsupervised visual representation. CVPR 2020
[70] Wang et al. Exploring cross-image pixel contrast for semantic segmentation. ICCV 2021

# Method

## 2. Prototype memory bank

From [26, 70], contrastive learning requires massive data to learn good representation


- Semantically redundant
- Costly in memory and computation
  e.g. (K, N, dim)

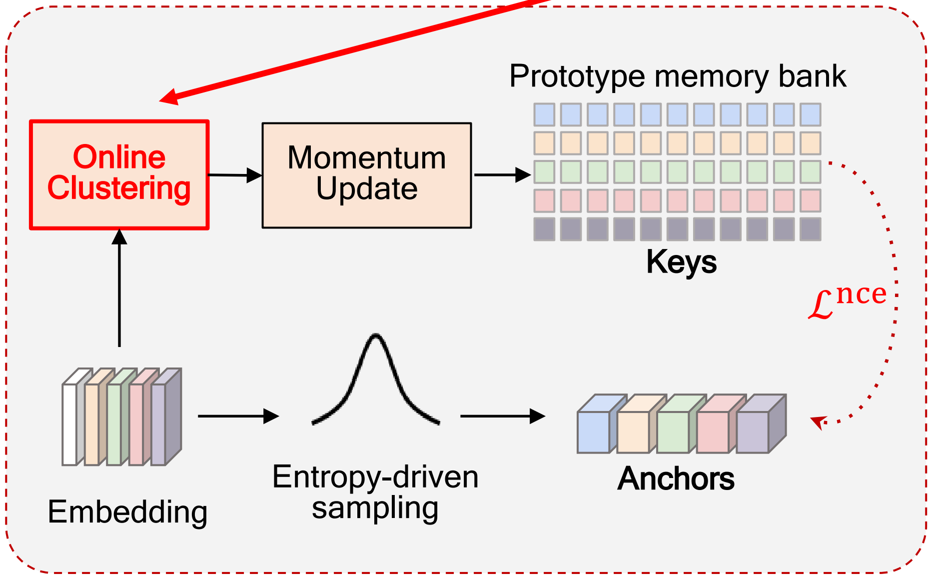
- Efficient in memory and computation
  e.g. (K, M, dim)


[26] He et al. Momentum contrast for unsupervised visual representation. CVPR 2020
[70] Wang et al. Exploring cross-image pixel contrast for semantic segmentation. ICCV 2021

# Method

## 2. Prototype memory bank

*Constrative learning training*




### I. Online prototype clustering

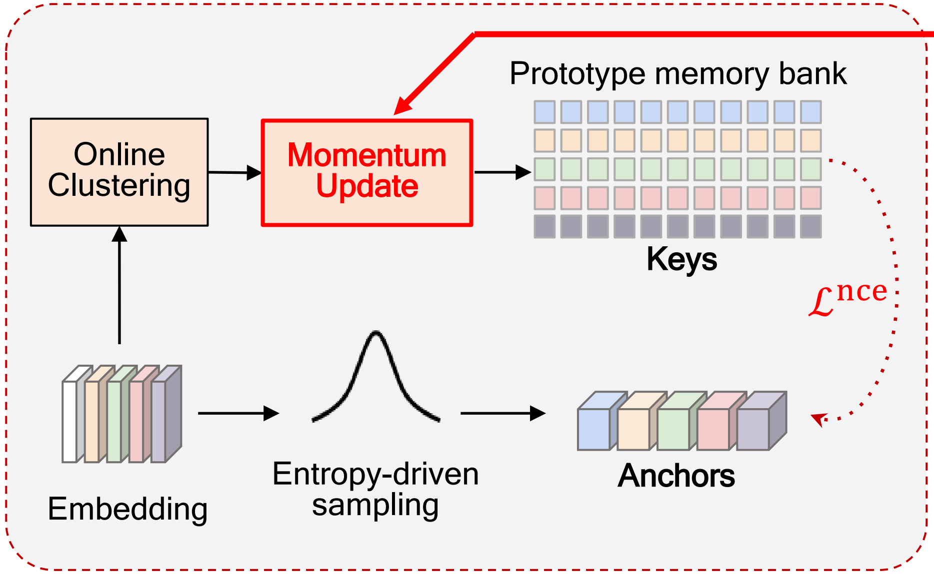
- Compute pixel-prototypes mapping framed as an optimal transport problem using Sinhorn algorithm[18].

[18] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. NeuRIPS 2013

# Method

## 2. Prototype memory bank

*Constrative learning training*




### I. Online prototype clustering

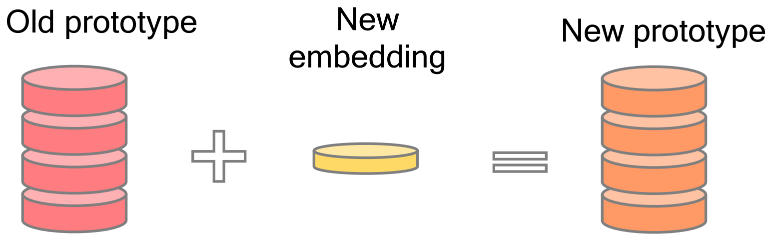
- Compute pixel-prototypes mapping framed as an optimal transport problem using Sinhorn algorithm[18].

### II. Online prototype update

- With momentum ($\sigma = 0.999$), $j^{th}$ prototype $\{P_k\}_j$ of class $k$ is updated as

$$\{P_k\}_j = \sigma\{P_k\}_j + (1-\sigma)\frac{1}{\sum_{i=1}^{N_k}[\![m(x_i) = j]\!]}\sum_{i=1}^{N_k} x_i[\![m(x_i) = j]\!]$$

- $m(x_i)$ is the prototype mapping of point $x_i$.

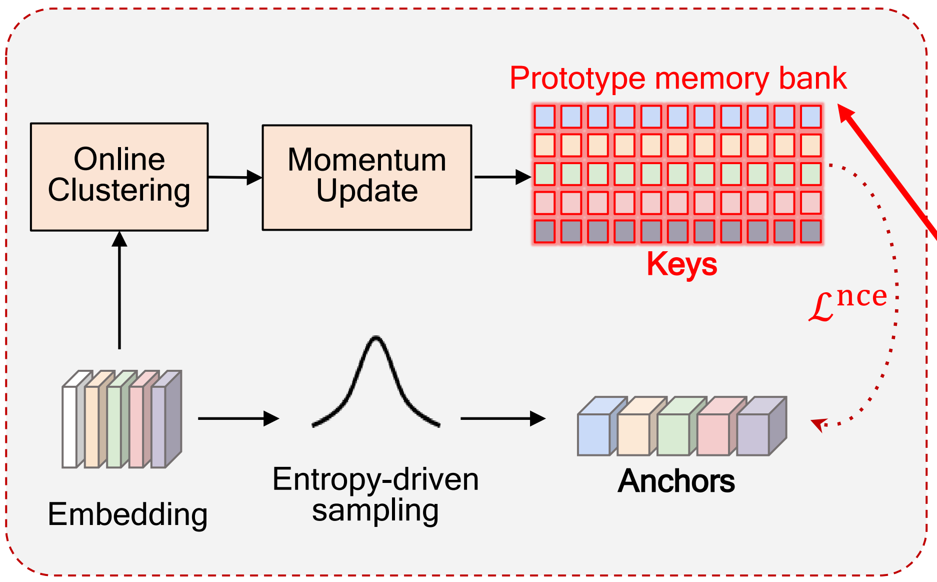



[18] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. NeuRIPS 2013

# Method

## 2. Prototype memory bank

*Constrative learning training*




### I. Online prototype clustering

- Compute pixel-prototypes mapping framed as an optimal transport problem using Sinhorn algorithm[18].

### II. Online prototype update

- With momentum ($\sigma = 0.999$), $j^{th}$ prototype $\{P_k\}_j$ of class $k$ is updated as

$$\{P_k\}_j = \sigma\{P_k\}_j + (1-\sigma)\frac{1}{\sum_{i=1}^{N_k}[\![m(x_i)=j]\!]}\sum_{i=1}^{N_k} x_i[\![m(x_i)=j]\!]$$

- $m(x_i)$ is the prototype mapping of point $x_i$.

### III. Compute contrastive loss

- Prototypes $\{P_k\}_j$ serves as keys in the training.

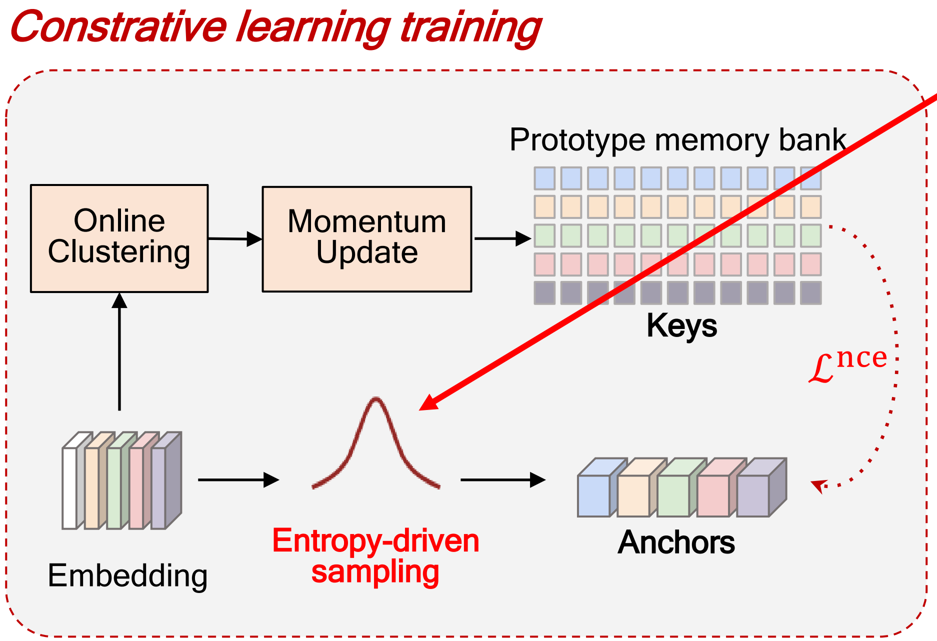
$$\mathcal{L}^{\text{nce}} = \frac{1}{N_a}\sum_{a_i\in\mathcal{A}} -\log\frac{\sum_{p_j^+\in\mathcal{P}^+}\exp\left(a_i\cdot p_j^+\right)}{\sum_{p_j^+\in\mathcal{P}^+}\exp\left(a_i\cdot p_j^+ / \tau\right)+\sum_{p_j^+\in\mathcal{P}^-} exp\left(a_i\cdot p_j^- / \tau\right)}$$

[18] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. NeuRIPS 2013

# Method

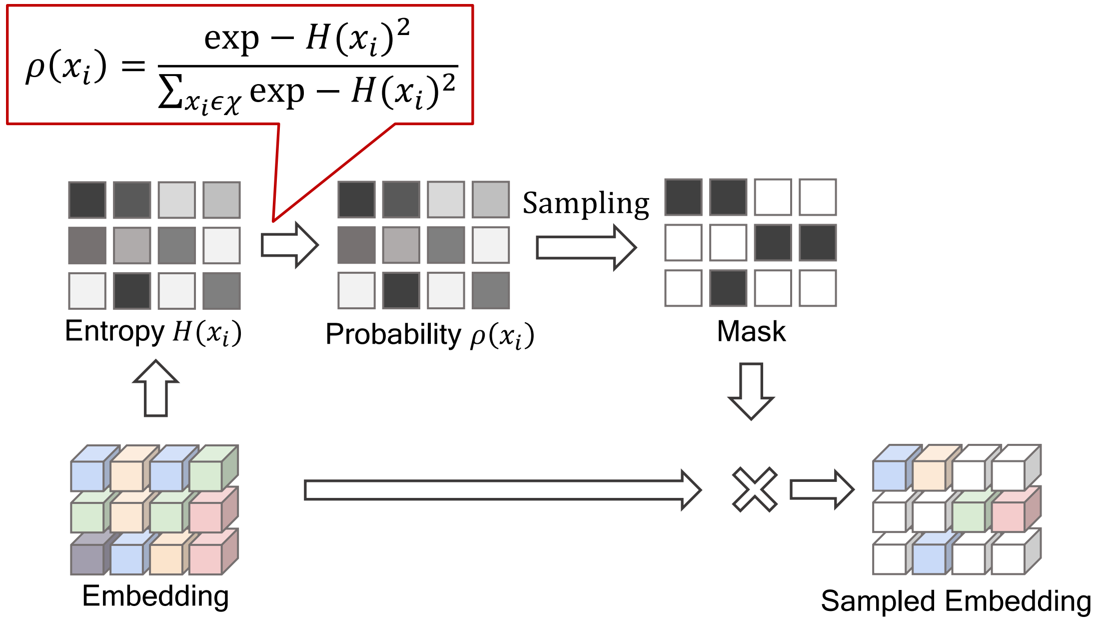
## 3. Entropy-driven sampling

From [58], Shannon entropy can evaluate the prediction quality




### I. Entropy-driven sampling

- Sample relevant pseudo-labels predictions based on Shannon entropy $H(x_i)$ of point $x_i$.



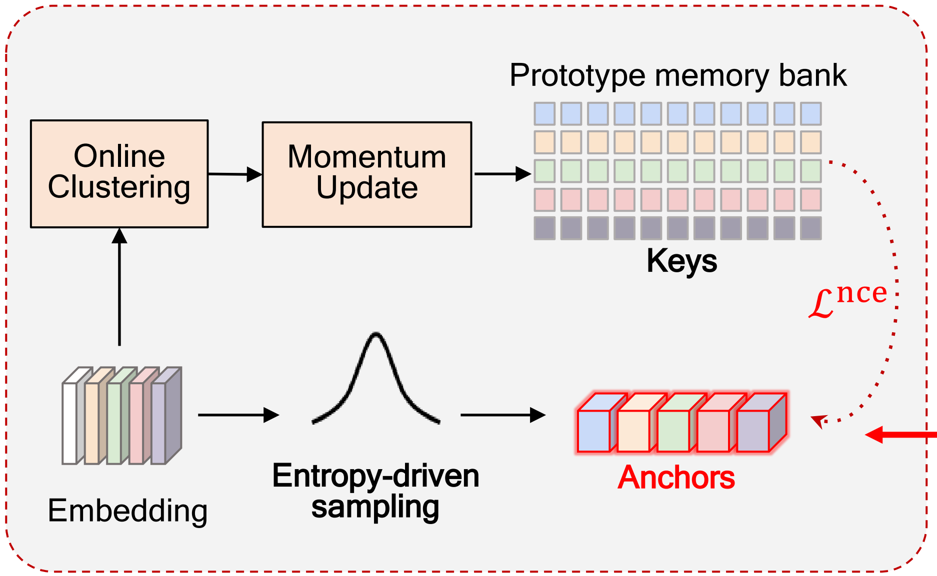

[58] Claude Elwood Shannon. A mathematical theory of communication. SIGMOBILE 2001.

# Method

## 3. Entropy-driven sampling

From [58], Shannon entropy can evaluate the prediction quality

*Constrative learning training*




### I. Entropy-driven sampling

- Sample relevant pseudo-labels predictions based on Shannon entropy $H(x_i)$.

$$\rho(x_i) = \frac{\exp - H(x_i)^2}{\sum_{x_i \epsilon \chi} \exp - H(x_i)^2}$$

- $\rho(x_i)$ is the sampling probability of point $x_i$.

### II. Compute contrastive loss

- Sampled embedding serves as anchors in the training.

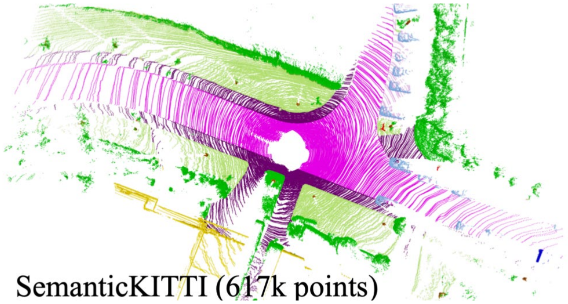
$$\mathcal{L}^{nce} = \frac{1}{N_a} \sum_{a_i \in \mathcal{A}} - \log \frac{\sum_{p_j^+ \in \mathcal{P}^+} \exp\left(a_i \cdot p_j^+\right)}{\sum_{p_j^+ \in \mathcal{P}^+} \exp\left(a_i \cdot p_j^+ / \tau\right) + \sum_{p_j^+ \in \mathcal{P}^-} exp\left(a_i \cdot p_j^- / \tau\right)}$$


[58] Claude Elwood Shannon. A mathematical theory of communication. SIGMOBILE 2001.

# Results

## I. SemanticKITTI

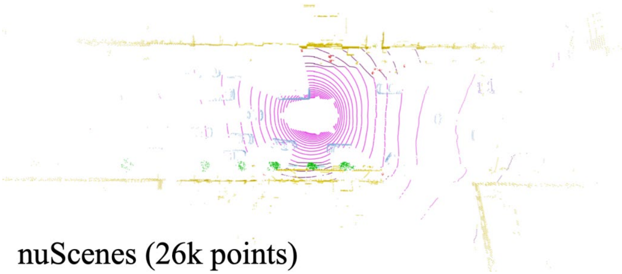
- 64 beams LiDAR
- Collected in Germany
- Most popular benchmark



SemanticKITTI (617k points)

## II. nuScenes

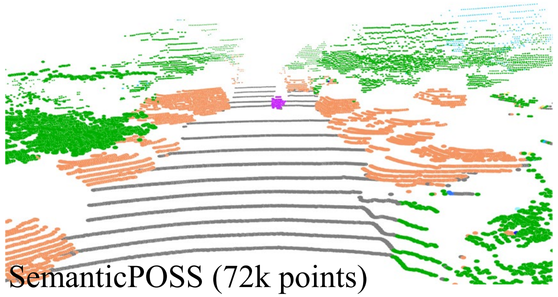
- 32 beams LiDAR
- Collected in America and Singapore
- Different weather and season



nuScenes (26k points)

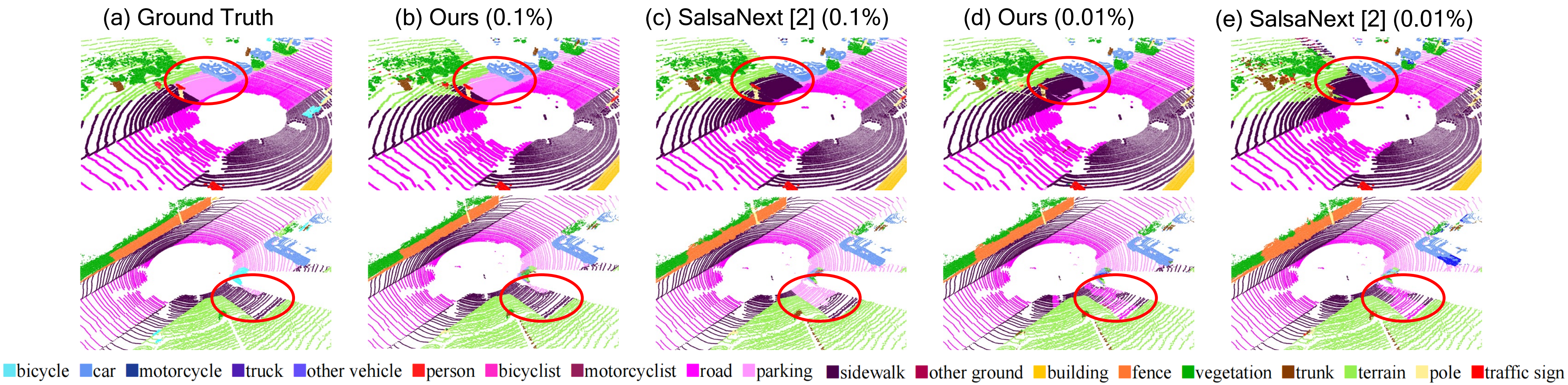
## III. SemanticPOSS

- 40 beams LiDAR
- Collected in China
- Denser and smaller



SemanticPOSS (72k points)

# Results on SemanticKITTI




**Analysis**

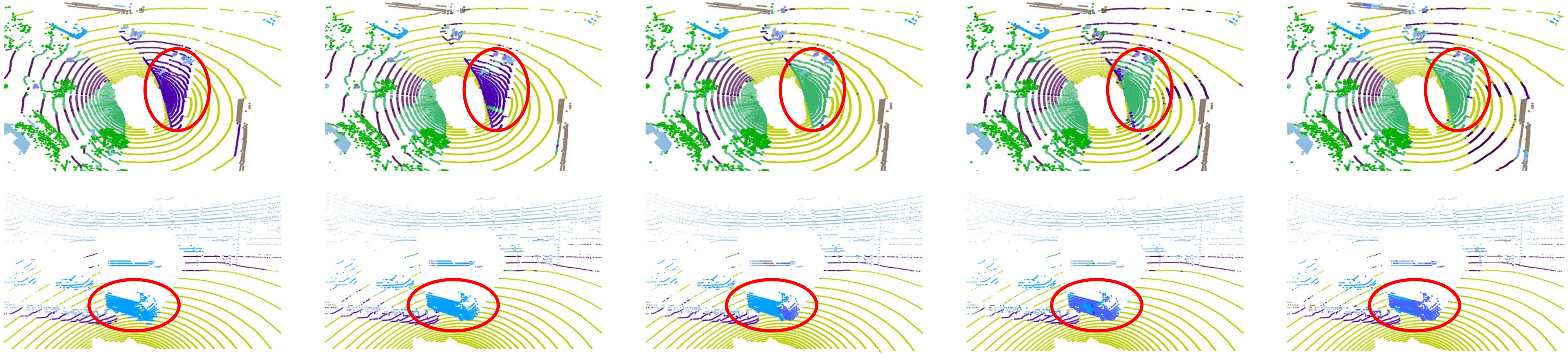
- Improve ~5% compared to SQN and reach SOTA
- Outperforms baseline method SalsaNext

| Anno. (%) | Method | Proj | mIoU (%) |
|---|---|---|---|
| 100 | (AF)$^2$S3Net [5] | × | 69.7 |
| | SquSegV3 [4] | √ | 55.9 |
| | SalsaNext [2] | | 59.5 |
| 0.1 | SQN [1] | × | <u>50.8</u> |
| | SalsaNext [2] | √ | 50.1 |
| | Ours | | **55.7** |
| 0.01 | SQN [1] | × | 39.1 |
| | SalsaNext [2] | √ | <u>42.6</u> |
| | Ours | | **46.2** |


[1] SQN. Hu et al. ECCV 2022.
[2] SalsaNext. Tiago et al. ISVC 2020.
[4] SqueezeSegV3. Xu et al. ECCV 2020.
[5] (AF)2S3Net. Cheng et al. CVPR 2021.

# Results on nuScenes

(a) Ground Truth (b) Ours (0.1%) (c) SalsaNext (0.1%) (d) Ours (0.01%) (e) SalsaNext (0.01%)



barrier bicycle bus car construction vehicle motorcycle pedestrian traffic cone trailer truck driveable surface other flat sidewalk terrain manmade vegetation

## Analysis

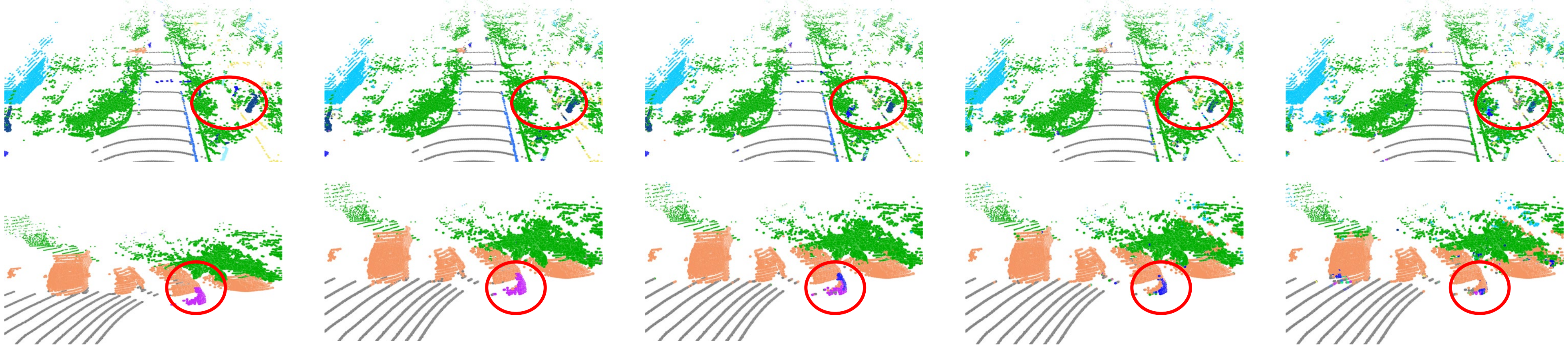
- Better than SalsaNext in 0.1% annotation
- Clustering fails to associate labels/prototypes in 0.01% annotation

| Anno. (%) | Method | Proj | mIoU (%) |
|---|---|---|---|
| 100 | PolarNet [6] | | 72.2 |
| | Cylinder3D [7] | × | 76.1 |
| | (AF)$^2$S3Net [5] | | 78.0 |
| | RangeNet[3] | √ | 65.5 |
| | SalsaNext [2] | | 72.2 |
| 0.1 | SalsaNext [2] | √ | 56.5 |
| | Ours | | **58.7** |
| 0.01 | SalsaNext [2] | √ | **44.5** |
| | Ours | | 42.9 |

[2] SalsaNext. Tiago et al. ISVC 2020.
[3] Rangenet. Milioto et al. IROS 2019.
[5] (AF)2S3Net. Cheng et al. CVPR 2021.
[6] PolarNet. Zhang et al. CVPR 2020.
[7] Cylinder3D. Zhu et al. CVPR 2021.

# Results on SemanticPOSS

(a) Ground Truth (b) Ours (0.1%) (c) SalsaNext [2] (0.1%) (d) Ours (0.01%) (e) SalsaNext [2] (0.01%)



people rider car trunk plants traffic-sign pole trashcan building cone/stone fence bike road

## Analysis

- Outperform SalsaNext (baseline) in both 0.1% and 0.01%

| Anno. (%) | Method | Proj | mIoU (%) |
|---|---|---|---|
| 100 | RandLANet[8] | × | 53.5 |
| | KPConv [9] | | 55.2 |
| | JS3C-Net [10] | | 60.2 |
| | SquSegV2[11] | √ | 29.8 |
| | SalsaNext [2] | | 45.0 |
| 0.1 | SalsaNext [2] | √ | 38.9 |
| | Ours | | **43.0** |
| 0.01 | SalsaNext [2] | √ | 27.4 |
| | Ours | | **31.1** |

[2] SalsaNext. Tiago et al. ISVC 2020.
[8] RandLANet. Hu et al. CVPR 2020.
[9] KPConv. Thomas et al. ICCV 2019.
[10] JS3C-Net. Yan et al. AAAI 2021.
[11] SqueezeSegV2. Wu et al. ICRA 2018.

# Ablation Study

Choice of backbone

| Methods | SemPOSS mIoU (%) | SemKITTI mIoU (%) |
|---|---|---|
| Rangenet-21 [3] | 25.1 | 40.7 |
| **Ours (Rangenet-21)** | **28.9 (+3.8)** | **44.5 (+3.8)** |
| SqueezeSegV3-21 [4] | 30.4 | 42.5 |
| **Ours (SqueezeSegV3-21)** | **36.7 (+6.3)** | **48.5 (+6.0)** |
| SalsaNext [2] | 38.9 | 52.4 |
| **Ours (SalsaNext)** | **43.0 (+4.1)** | **57.6 (+5.2)** |

COARSE3D performs consistently with different backbones.

[2] Tiago et al. Salsanext: Fast, uncertaintyaware semantic segmentation of lidar point clouds. ISVC 2020.
[3] Milioto et al. Rangenet ++: Fast and accurate lidar semantic segmentation. IROS 2019.
[4] Xu et al. Squeezesegv3: Spatially-adaptive convolution for efficient point-cloud segmentation. ECCV 2020.

# Ablation Study

Architecture ablation

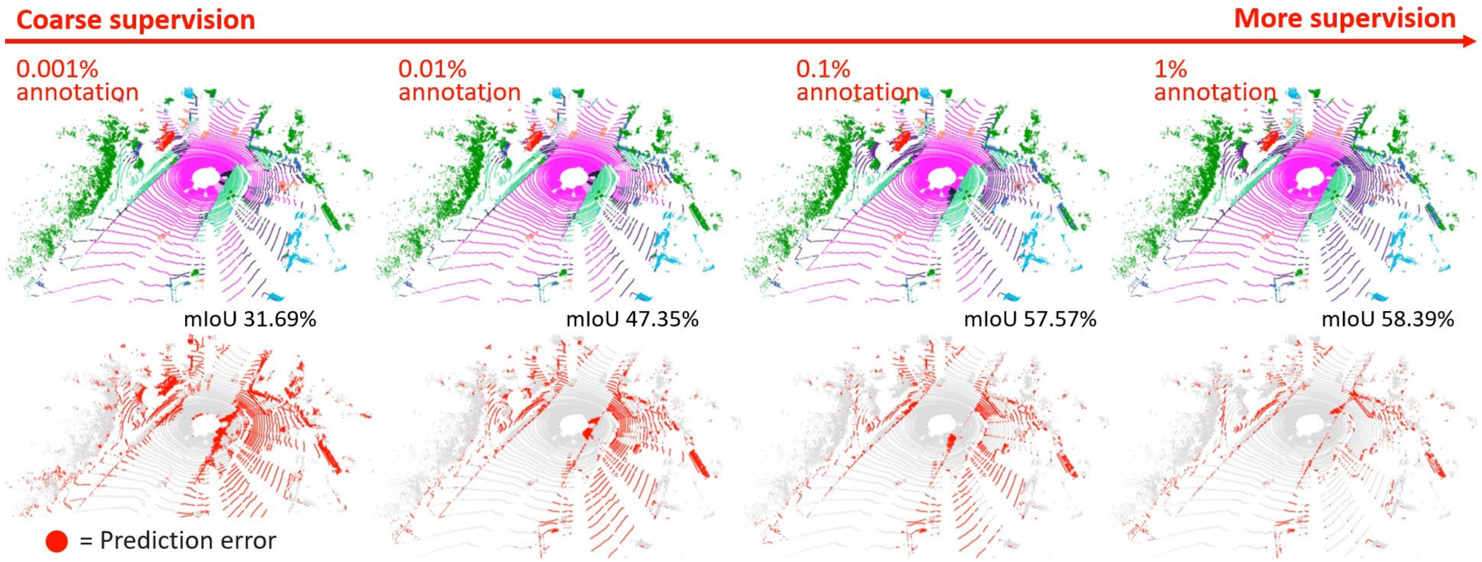
| Methods | mIoU (%) |
|---|---|
| **Ours** | **57.57** |
| w/o contrast module | 55.44 |
| w/o anchor sampling | <u>56.32</u> |
| w/o prototype (5k pxl) | 56.10 |
| w/o voxel propagation | 56.26 |
| w/o Focal loss | 42.41 |
| w/o Lovasz loss | 56.10 |

# Ablation Study

Annotation ablation

| Anno. | mIoU (%) | |
|---|---|---|
| | SalsaNext [2] | Ours |
| 0.001% | 30.39 | 31.69 |
| 0.01% | 44.00 | 47.13 |
| 0.1% | 52.43 | 56.61 |
| 1% | 56.16 | 58.30 |
| 100% | 56.44 | 58.39 |

- Outperform the baseline method in the different annotations.
- Reach the comparable performance with 100% label at 0.1%




[2] Tiago et al. Salsanext: Fast, uncertaintyaware semantic segmentation of lidar point clouds. ISVC 2020.

# Conclusion

- An **architecture-agnostic framework** for weakly-supervised LiDAR semantic segmentation.
- A **prototype memory bank** that captures per-class dataset information with an **entropy-driven sampling** technique to sample more confident pixels as anchors.
- Results on **3 baseline architectures** and **3 datasets** demonstrate the effectiveness.

Code is available !

https://github.com/cv-rits/COARSE3D