3D Point Cloud Instance Segmentation with Various Types of Supervision

Khoi Nguyen

Machine Learning Research School
DEPA, Bangkok, August 2023

About Me

https://khoinguyen.org

- **PhD in Computer Science** from Oregon State University, USA
- **Research Scientist** at VinAI Research working on CV
- **Research interests:**
  - 2D and 3D Scene Understanding: Detection, Segmentation
  - 2D and 3D Generation using Diffusion Models and NeRFs
  - Few-shot Learning, Vision-language Model
- **Best paper - Honorable Mention** at IEEE/CVF WACV 2023

Past Projects

During my PhD study

- Few-shot Object Segmentation in Images: ICCV'19, CVPR'21, ICCV'21, and CVPR'22

- Few-shot Counting, Classification & Detection: ICPR'20, NeurIPS'21, and ECCV'22a&b
- Image Enhancement: WACV'23 (**Best paper - Honorable Mention**)
- 3D Point Cloud Instance Segmentation: ECCV'22c, CVPR'23, ICCV'23

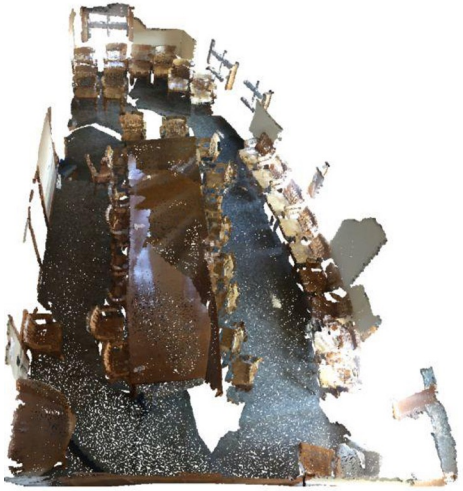
With VinAI Residents as 1st authors!

This talk!

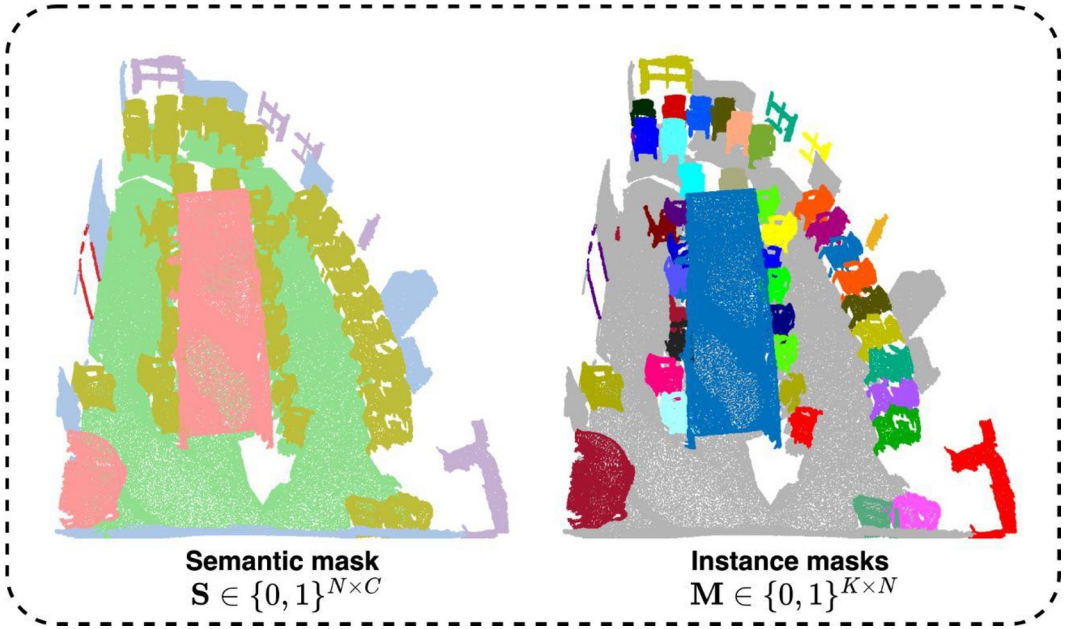
- **Geodesic-Former: a Geodesic-Guided Few-shot 3D Point Cloud Instance Segmenter,** *Tuan Ngo, Khoi Nguyen, ECCV 2022*
- **ISBNet: a 3D Point Cloud Instance Segmenter with Instance-aware Sampling and Box-aware Dynamic Convolution,** *Tuan Ngo, Binh-Son Hua, Khoi Nguyen, CVPR 2023*
- **GaPro: Box-Supervised 3D Point Cloud Instance Segmentation Using Gaussian Processes as Pseudo Labelers,** *Tuan Ngo, Binh-Son Hua, Khoi Nguyen, ICCV 2023*

3D Point Cloud Instance Segmentation (3DIS)

Given a **3D RGB point cloud** (3D coordinate + RGB), we seek to obtain **semantic** and **object instance masks** of specific categories of interest.



RGB Point cloud
$\mathbf{P} \in \mathbb{R}^{N \times 6}$



Semantic mask
$\mathbf{S} \in \{0,1\}^{N \times C}$

Instance masks
$\mathbf{M} \in \{0,1\}^{K \times N}$

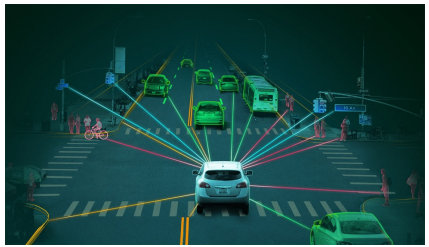
K: # object instances,
N: # points, C: # categories

table chair floor

Applications

Where 3D point cloud data can complement the information provided by 2D images

- Robot navigation in indoor environment
- Autonomous driving in outdoor environment
- Augmented reality applications

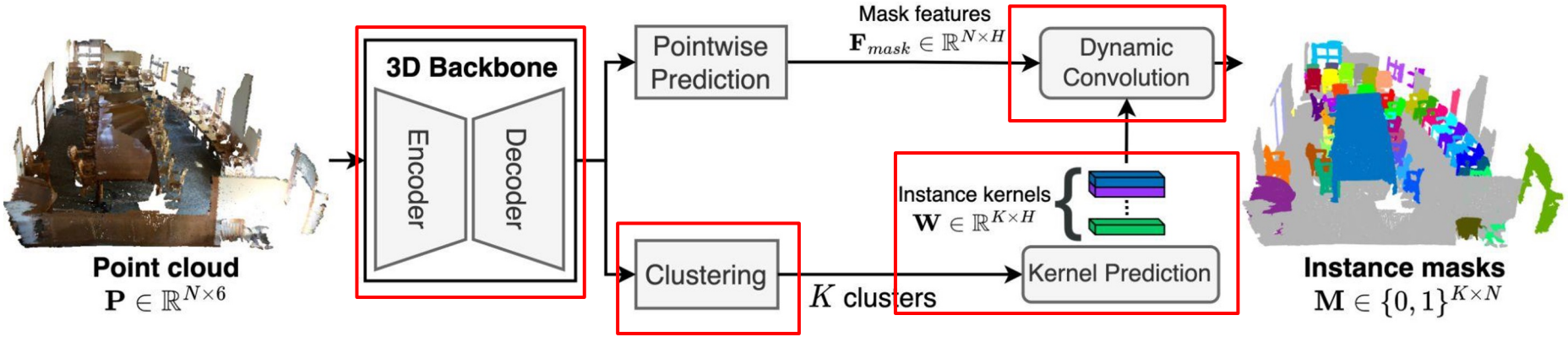


Challenges

- Objects in 3D have much higher variations in **appearance and shape** than 2D images.
- 3D point clouds are **unevenly distributed**, i.e., dense near object surface and sparse elsewhere

➔ It is not trivial to apply 2D instance Segmentation approaches to 3DIS

A Typical Approach for 3DIS: DyCo3D [1]



1. Use a 3D backbone to extract pointwise features
2. Predict instance masks:
    a. Group points into **clusters** for object candidates
    b. Generate an **instance kernel** for each object candidate
    c. Dynamic convolution: Convolve each generated kernel with pointwise mask features to obtain a binary instance mask for each object

[1] DyCo3D: Robust Instance Segmentation of 3D Point Clouds through Dynamic Convolution (CVPR 2021)

Limitations of DyCo3D

Appearance feature is **not distinct enough to distinguish** objects of the same class

Point cloud
$\mathbf{P} \in \mathbb{R}^{N \times 6}$

3D Backbone
Encoder
Decoder

Pointwise Prediction

Mask features
$\mathbf{F}_{mask} \in \mathbb{R}^{N \times H}$

Dynamic Convolution

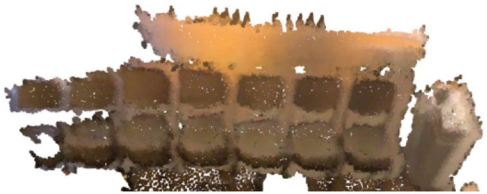
Instance kernels
$\mathbf{W} \in \mathbb{R}^{K \times H}$

Clustering

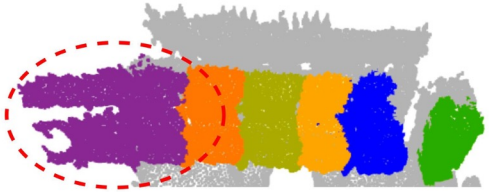
K clusters

Kernel Prediction

Instance masks
$\mathbf{M} \in \{0,1\}^{K \times N}$



RGB point cloud



DyCo3D's predictions

Mis-grouping points when similar objects are adjacent
➔ **Low-recall object proposals**

Our ISBNet

Enhance appearance feature with geometric cue, i.e., **object bounding box**

Point cloud
$\mathbf{P} \in \mathbb{R}^{N \times 6}$

3D Backbone
Encoder
Decoder

Pointwise Prediction

Box predictions

Mask features
$\mathbf{F}_{mask} \in \mathbb{R}^{N \times H}$

Box-aware Dynamic Convolution

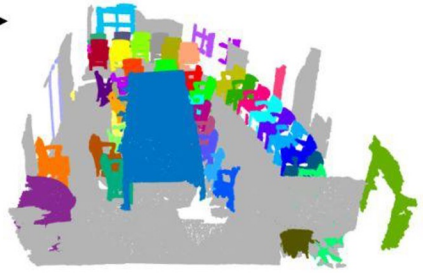
Instance kernels
$\mathbf{W} \in \mathbb{R}^{K \times H}$

Instance-aware Sampling

K candidate points

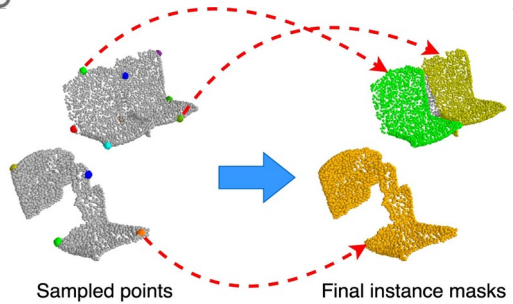
Kernel Prediction

Box Prediction



Instance masks
$\mathbf{M} \in \{0,1\}^{K \times N}$

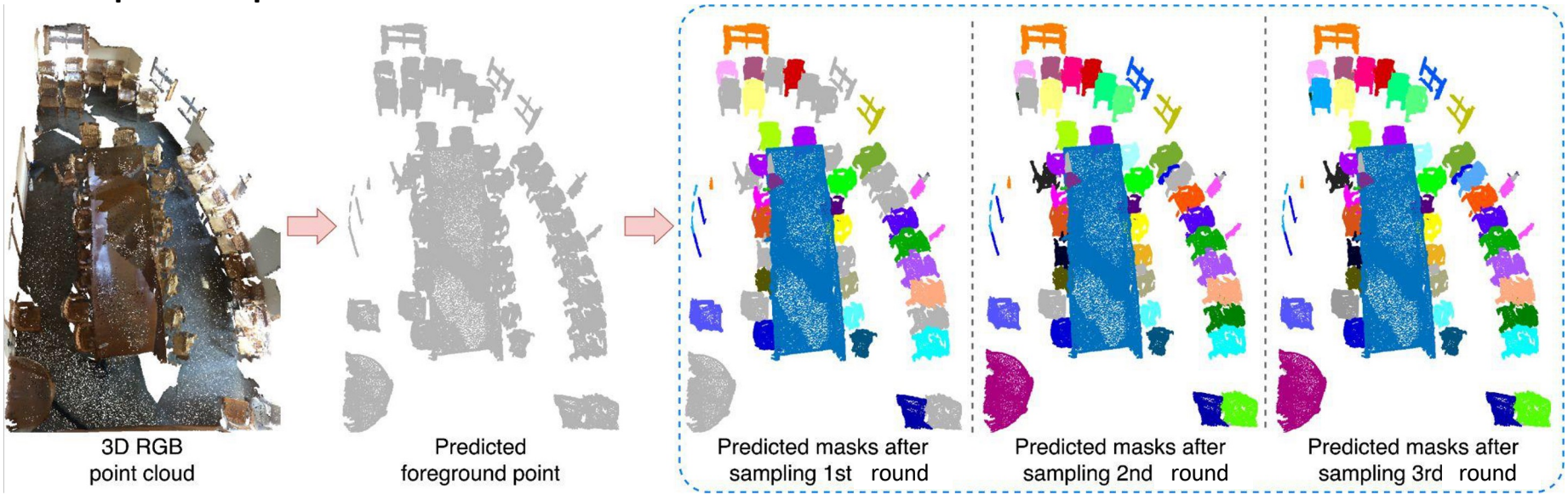
Replace clustering by **instance-aware sampling**:
Each sampled point represents a **candidate object** to obtain instance mask
=> **high-recall object proposals**



Instance-aware Sampling (IA-FPS)

Goal: sample a set of K candidate points from initial N points ($\boldsymbol{K \ll N}$) to maximize the **instance recall rate**.

- Only sample from **foreground points**
- Multiple-rounds of sampling and object mask prediction: Avoid **points belonging to previous predicted instance masks**



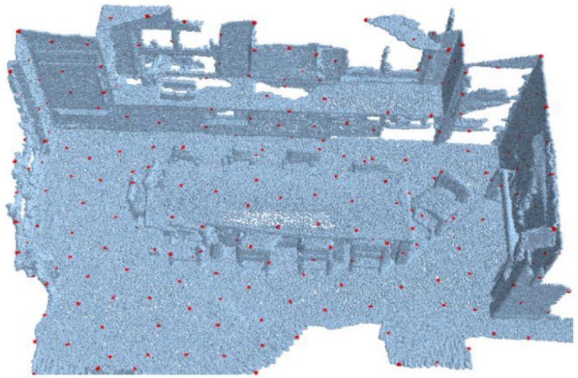
Instance-aware Sampling (IA-FPS)

Goal: sample a set of K candidate points from initial N points ($\boldsymbol{K \ll N}$) to maximize the **instance recall rate**.

- Only sample from **foreground points**
- Multiple-rounds of sampling and object mask prediction: Avoid **points belonging to previous predicted instance masks**
- For each round, sample points based on Farthest Point Sampling: a greedy algorithm in which next point is the farthest point away from all previous sampled points

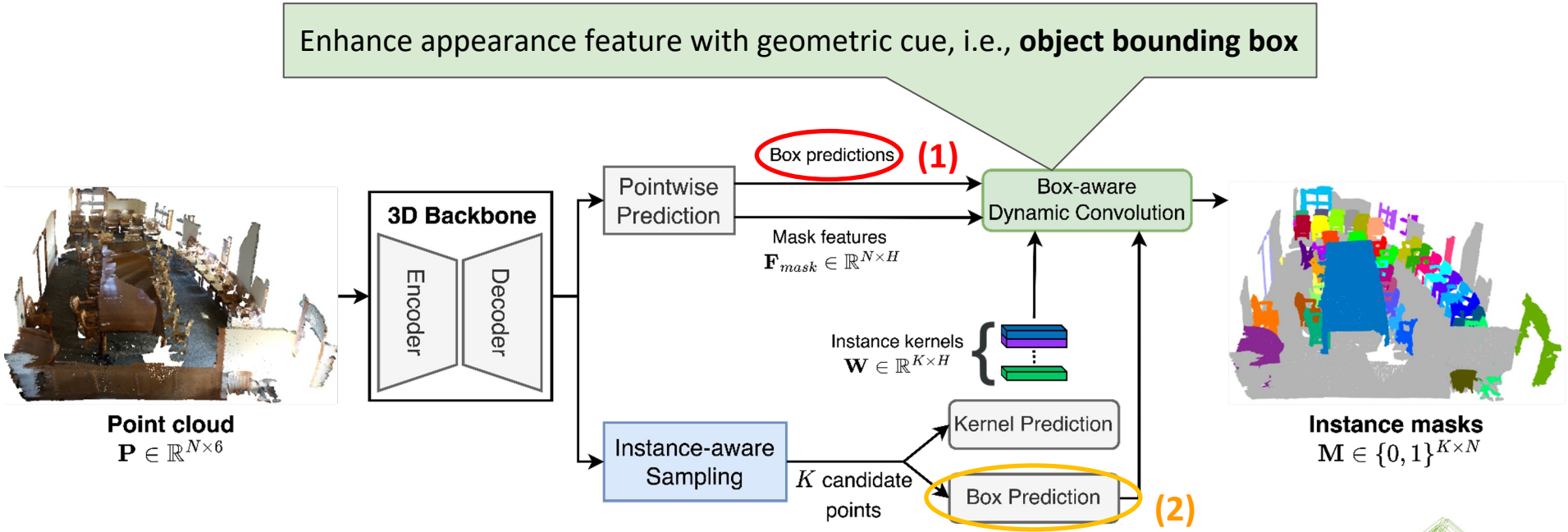
# sampling points	2048	512	256	128
FPS	99.3%	93.3%	85.4%	71.3%
IA-FPS	100%	98.4%	94.5%	89.2%
Clustering (fixed)	75.5%	75.5%	75.5%	75.5%

Instance recalls of different sampling strategies

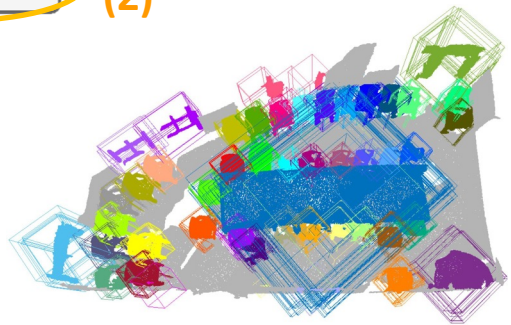


Farthest point sampling (256 points)

Box-aware Dynamic Convolution



Intuition: an object candidate (2) "attracts" points (1) predicting similar boxes



Box-aware Dynamic Convolution

Use **box difference** as an important geometric cue to enhance the appearance feature

$$\widehat{m}^{(k)} = \text{Sigmoid}\left(\text{Conv}\left(\left[\mathbf{F}_{mask}; \mathbf{F}_{box}^{(k)}\right]; w^{(k)}\right)\right)$$

Instance kernel

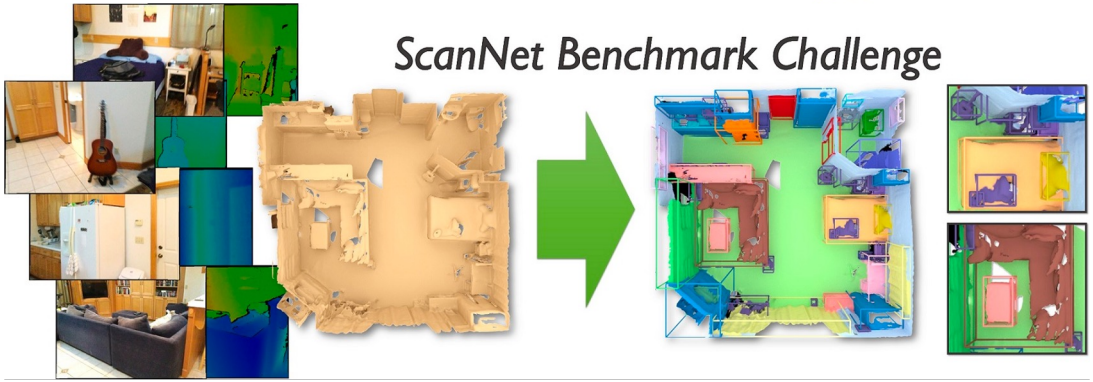
Box difference

Appearance feature

Box difference: difference in box size and box center between pointwise predicted box and object candidate's predicted box

Experiments

- **Datasets**:
  - Indoor: ScanNetV2 (18 classes), S3DIS (13 classes)
  - Outdoor: STPLS3D (15 classes)
- **Metrics**:
  - AP, AP50 (Average Precision) on ScanNetV2 and STPLS3D
  - mPrec (mean precision), mRec (mean recall) on S3DIS



ScanNetV2

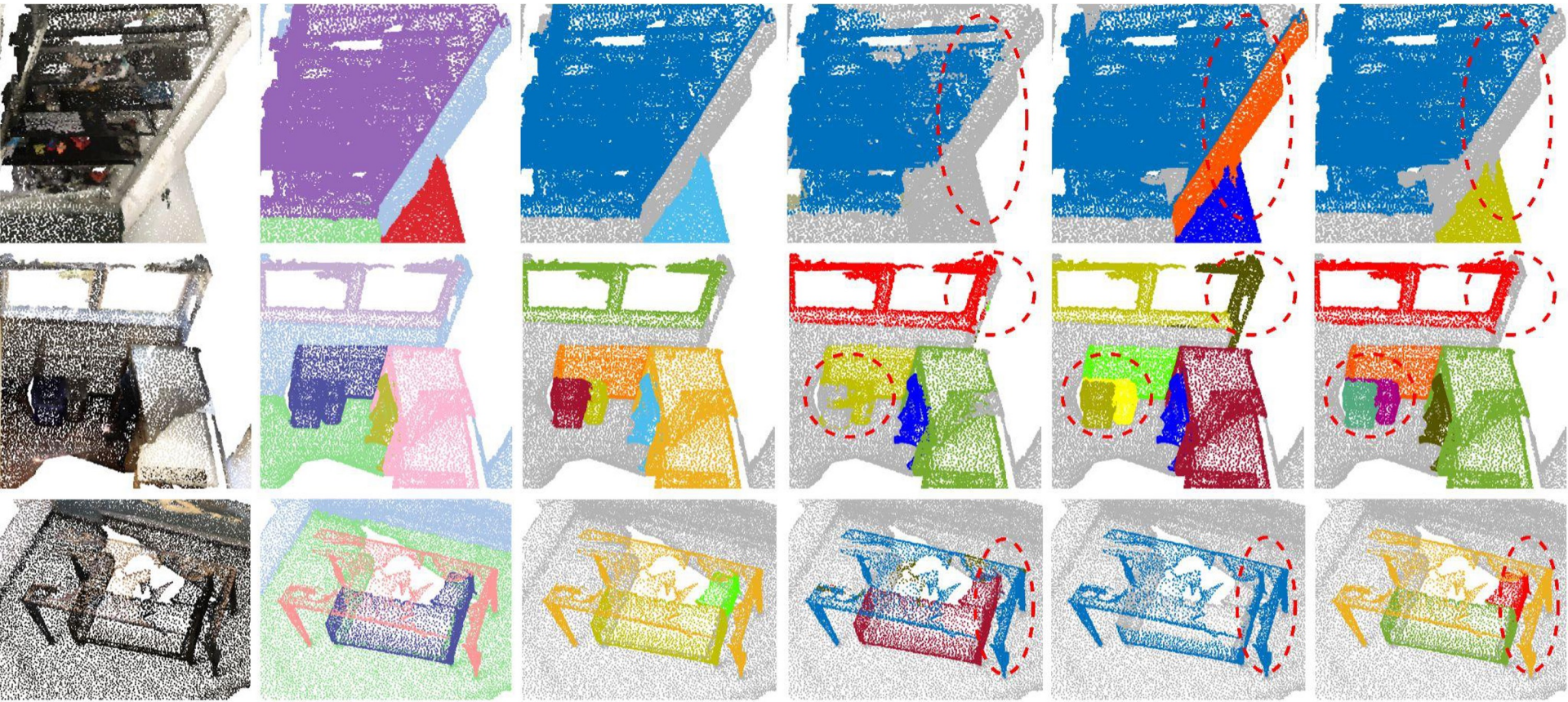
STPLS3D

Comparison with SOTA

		ScanNetV2 (indoor)		S3DIS (indoor)		STPLS3D (outdoor)	
		AP	AP50	mPrec	mRec	AP	AP50
GSPN	CVPR 19	19.3	37.8	36.0	28.7	-	-
PointGroup	CVPR 20	34.8	51.7	61.9	62.1	23.3	38.5
DyCo3D	CVPR 21	40.6	61.0	64.3	64.2	-	-
HAIS	ICCV 21	43.5	64.4	71.1	65.0	35.1	46.7
SoftGroup	CVPR 22	46.0	67.6	73.6	66.6	46.2	61.8
Di&Co3D	ECCV 22	47.7	67.2	63.9	67.2	-	-
PointInst3D	ECCV 22	45.6	63.7	73.1	65.2	-	-
DKNet	ECCV 22	50.8	66.7	70.8	65.3	-	-
Ours	CVPR 23	**54.5**	**73.1**	**74.2**	**72.7**	**49.2**	**64.0**
		+4.3	+5.5	+0.6	+5.5	+3.0	+2.2

+14

Qualitative Results on ScanNetV2



Ablation Study on ScanNetV2

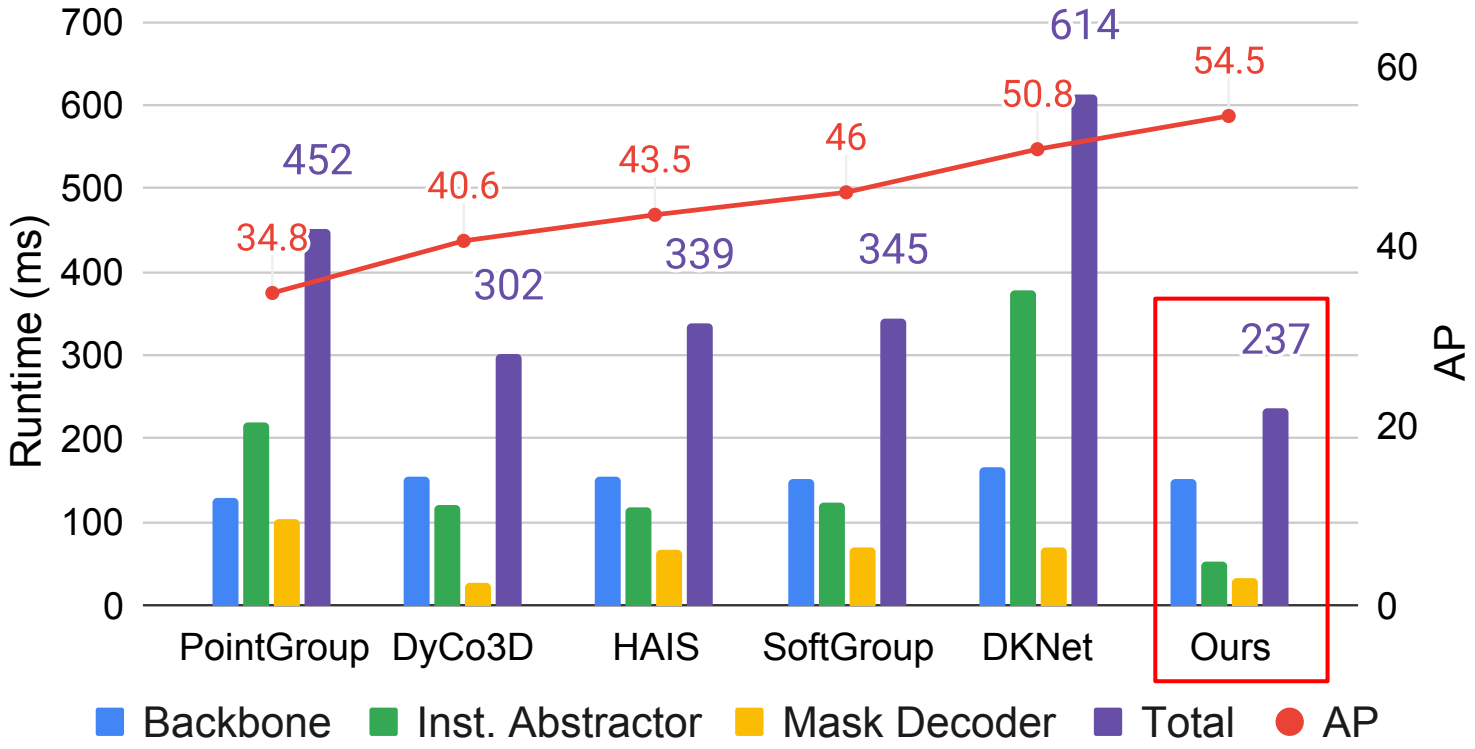
	IA-FPS	BA-DyCo	AP	AP_{50}	AP_{25}
Baseline			47.9	66.4	77.1
	✓		53.4	71.9	81.8
		✓	48.6	67.7	77.8
ISBNet	✓	✓	**54.5**	**73.1**	**82.5**

Chunk size	Total samples K	AP	AP_{50}	AP_{25}
(256)	256	53.9	72.2	80.8
(384)	384	54.2	72.4	81.4
(512)	512	53.6	71.9	81.1
(128,128,128)	384	54.0	72.8	81.0
(192,128,64)	384	**54.5**	**73.1**	**82.5**

- **IA-FPS**: Instance-aware Sampling
- **BA-DyCo**: Box-aware Dynamic Convolution

Multiple rounds sampling

Runtime Analysis



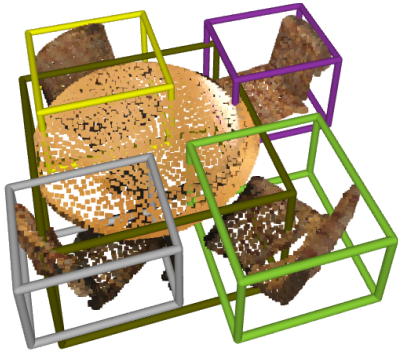
➔ Our method achieves **SOTA performance** while being the **fastest runtime**.

However, 3DIS requires costly pointwise annotations

Can we leverage cheaper annotations like 3D boxes instead?

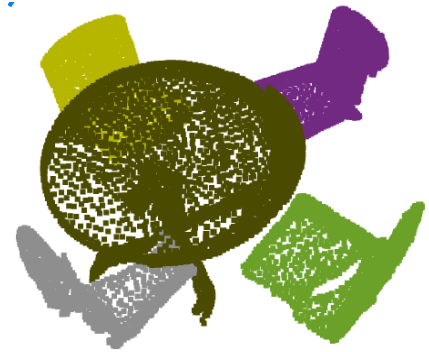
Box-supervised 3D Point Cloud Instance Segmentation

Box-Supervised 3DIS



GT axis-aligned bounding boxes

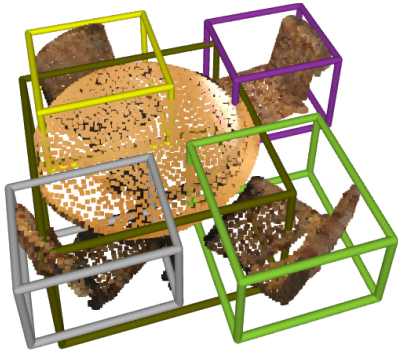
Convert



Pseudo mask annotation

Then use the pseudo mask annotation to train a regular 3DIS network

Prior Work on Box-Supervised 3DIS: Box2Mask [1]



GT axis-aligned bounding boxes

Convert

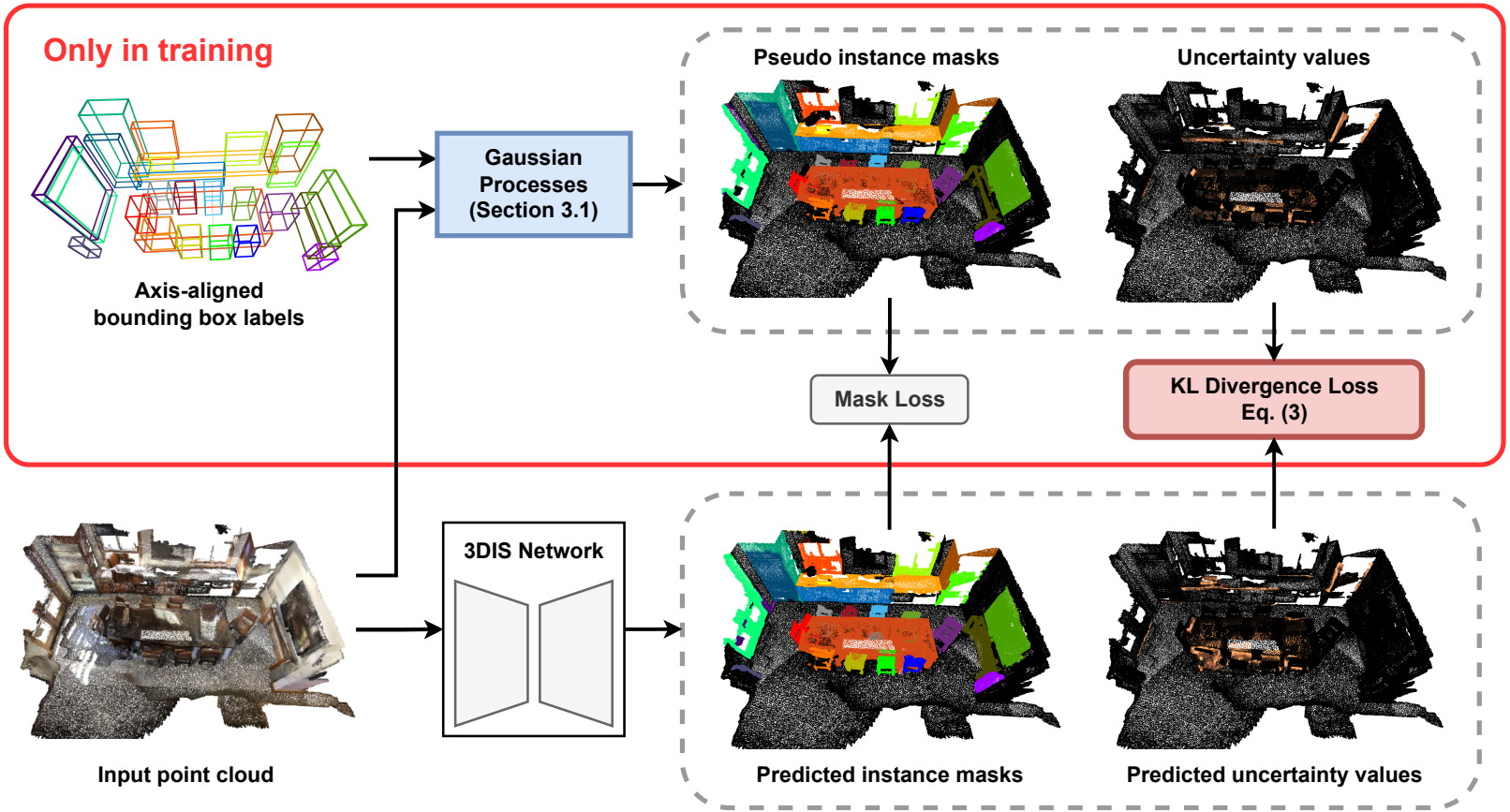
Box2Mask's pseudo mask annotation

Use heuristic rule to assign unlabeled points to box:

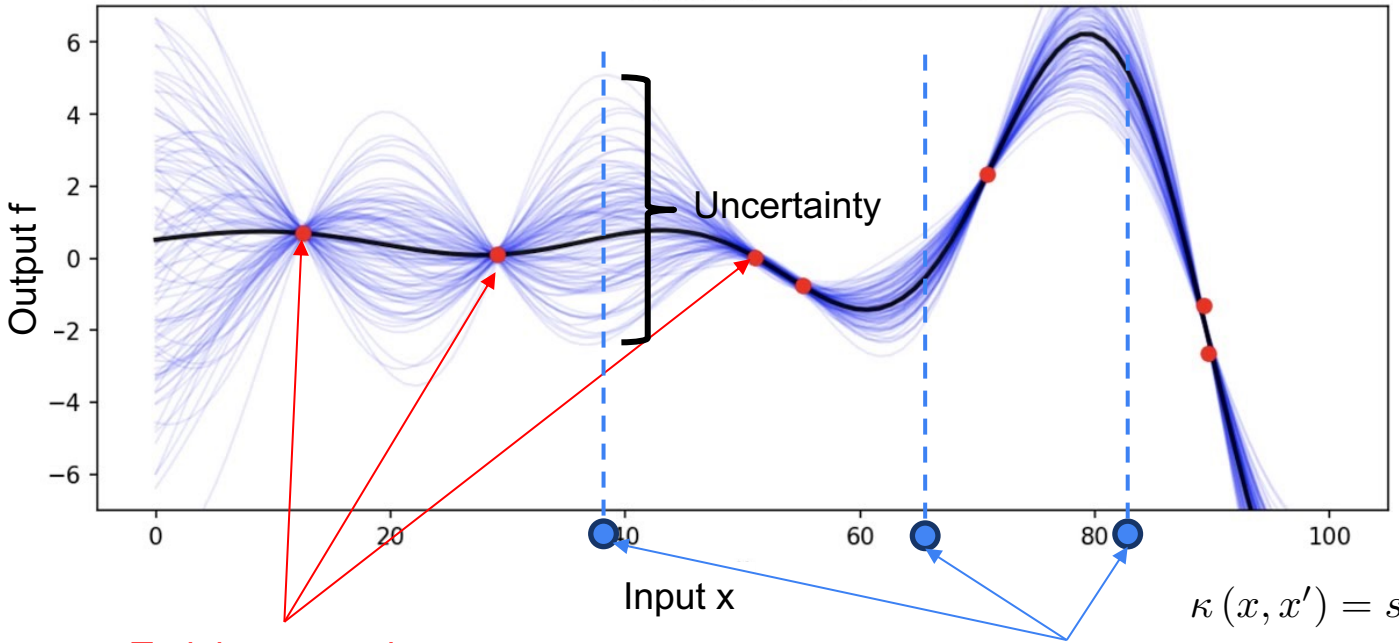
- If points belong to one box ➔ assign points to that box (object)
- If points belong to more than one box ➔ assign to the smaller box

[1] Chibane, Julian, et al. "Semantic Instance Segmentation of 3D Scenes Through Weak Bounding Box Supervision." *ECCV 2022*

Our Approach: GaPro (ICCV 2023)



Gaussian Process



$$p(\mathbf{f}) = \mathcal{N}(\mathbf{f}, \mathbf{0}).$$

$$p(\mathbf{f}_*) = \mathcal{N}(\mathbb{E}[\mathbf{f}_*], \mathrm{var}[\mathbf{f}_*])$$

$$\kappa(x, x') = s^2 \exp\left(-\frac{1}{2\ell^2}(x - x')^2\right)$$

$$\begin{pmatrix} \mathbf{f} \\ \mathbf{f}_* \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} \mathbf{K} & \mathbf{K}_* \\ \mathbf{K}_*^T & \mathbf{K}_{**} \end{pmatrix}\right)$$

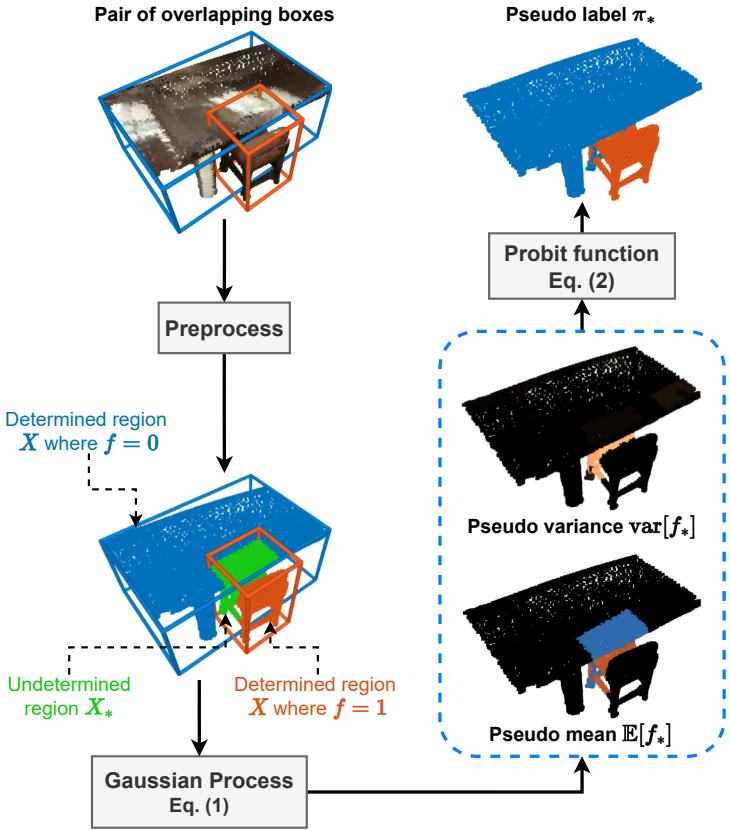
Joint Gaussian distribution

$\mathbf{K} = \kappa(\mathbf{X}, \mathbf{X})$ Similarity between between training examples

$\mathbf{K}_* = \kappa(\mathbf{X}, \mathbf{X}_*)$ Similarity between between train and test examples

$\mathbf{K}_{**} = \kappa(\mathbf{X}_*, \mathbf{X}_*)$ Similarity between between test examples

Applying to Box-Supervised 3DIS



Input: RGB and Coord or Deep feature of 3D points
Output: Label indicating which box the point belong to

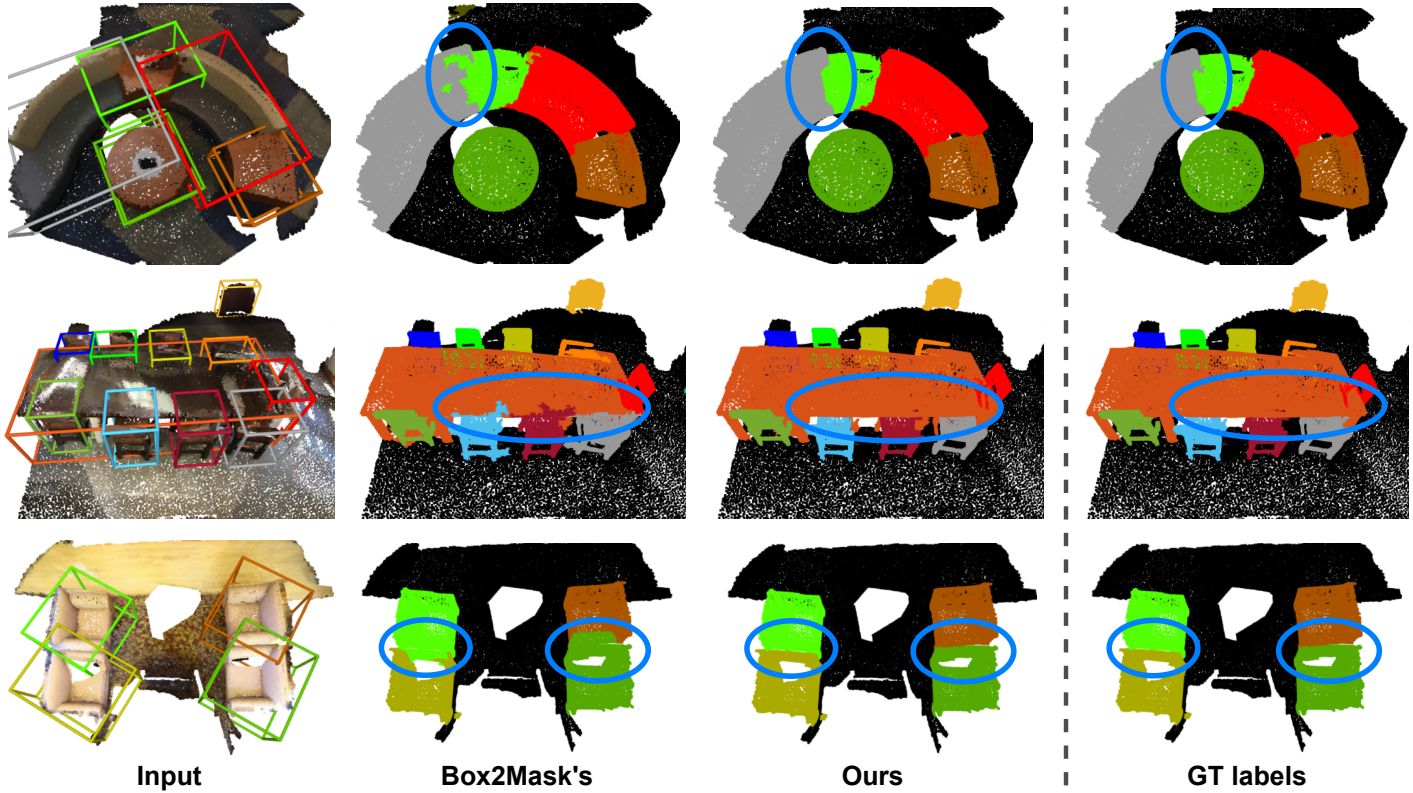
Training examples: points in **determined** region
Test examples: points in **undetermined** region

Convert regression results to classification results using **Probit approximation:**

$$\boldsymbol{\pi}_* = p\left(\mathbf{f}_* = 1 \mid \mathbf{X}_*, \mathbf{X}, \mathbf{f}\right) \approx \int \sigma\left(\mathbf{f}_*\right) p\left(\mathbf{f}_*\right) d\mathbf{f}_*,$$

$$\approx \sigma\left(\frac{\mathbb{E}\left[\mathbf{f}_*\right]}{\sqrt{1 + \frac{\pi}{8}\mathrm{var}[\mathbf{f}_*]}}\right),$$

Quality of Pseudo Labels on Training Set



Results on ScannetV2

Method	Sup.	Backbone	Test set				Val set			
			AP	% full	AP_{50}	AP_{25}	AP	% full	AP_{50}	AP_{25}
Mask3D [33]	Mask	Minkowski	56.6	-	78.0	87.0	55.2	-	73.7	83.5
PointGroup [19]		SPConv	40.7	-	63.6	77.8	34.8	-	51.7	71.3
SSTNet [26]		SPConv	50.6	-	69.8	78.9	49.4	-	64.3	74.0
SoftGroup [40]		SPConv	50.4	-	76.1	86.5	46.0	-	67.6	78.9
ISBNet [30]		SPConv	55.9	-	76.3	84.5	54.5	-	73.1	82.5
SPFormer [36]		SPConv	54.9	-	77.0	85.1	56.3	-	73.9	82.9
CSC [16]	Point	Minkowski	29.3	51.8%	59.2	70.2	15.9	28.8%	28.9	49.6
PointContrast [44]		Minkowski	27.8	49.1%	47.1	64.5	27.8	50.4%	47.1	64.5
Box2Mask [5] (stand-alone)	Box	Minkowski	43.3	-	67.7	80.3	39.1	-	59.7	71.8
WISGP [9] + PointGroup [19]		SPConv	-	-	-	-	31.3	89.9%	50.2	64.9
WISGP [9] + SSTNet [26]		SPConv	-	-	-	-	35.2	71.2%	56.9	70.2
GaPro + PointGroup [19]	Box	SPConv	39.4	96.8%	62.3	74.5	33.4	96.0%	53.7	69.8
GaPro + SSTNet [26]		SPConv	45.8	90.5%	65.2	75.0	43.9	88.9%	60.1	70.8
GaPro + SoftGroup [40]		SPConv	42.1	83.5%	62.9	79.4	41.3	89.8%	62.7	77.3
GaPro + ISBNet [30]		SPConv	49.3	88.2%	69.8	81.0	50.6	92.8%	69.1	79.3
GaPro + SPFormer [36]		SPConv	48.2	87.7%	69.2	82.4	51.1	90.8%	70.4	79.9

Ablation Study

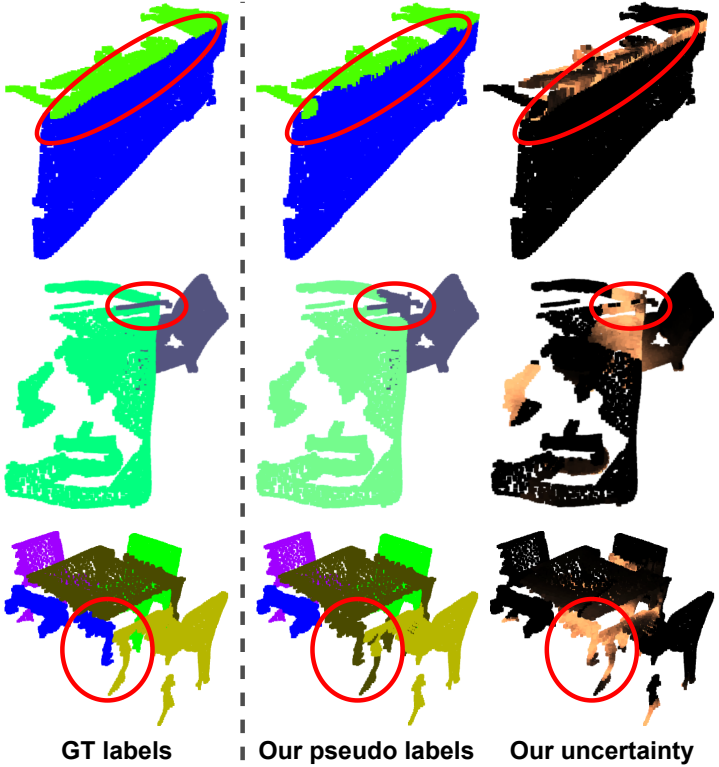
GP parameters	Superpoint	AP	AP_{50}
Fixed		46.3	66.3
Fixed	✓	48.0	67.2
Learnable		48.5	67.7
Learnable	✓	**50.6**	**69.1**

Handling of undetermined points	AP	AP_{90}
A: No pseudo labels in overlapped regions	53.6	22.5
B: Box2Mask: assign points to smaller box	64.4	27.6
C: Linear classifier with points	69.4	34.1
D3: **GaPro (ours)**	**85.9**	**63.1**
E: Ours w/ uncertainty-guided GT replacement	88.0	67.2

Pseudo Mask Prediction

	Handling of undetermined points	AP	AP_{50}
	A: No pseudo labels in overlapped regions	38.1	59.1
	B: Box2Mask: assign points to smaller boxes	41.8	64.8
	C: Linear Classifier with points	44.2	64.5
GaPro	D1: GP Classification with points	45.7	67.2
GaPro	D2: GP Regression with superpoints	47.8	67.7
GaPro	D3: **GP Classification with superpoints**	**48.9**	**68.4**

Results of Mask Prediction using Pseudo Mask

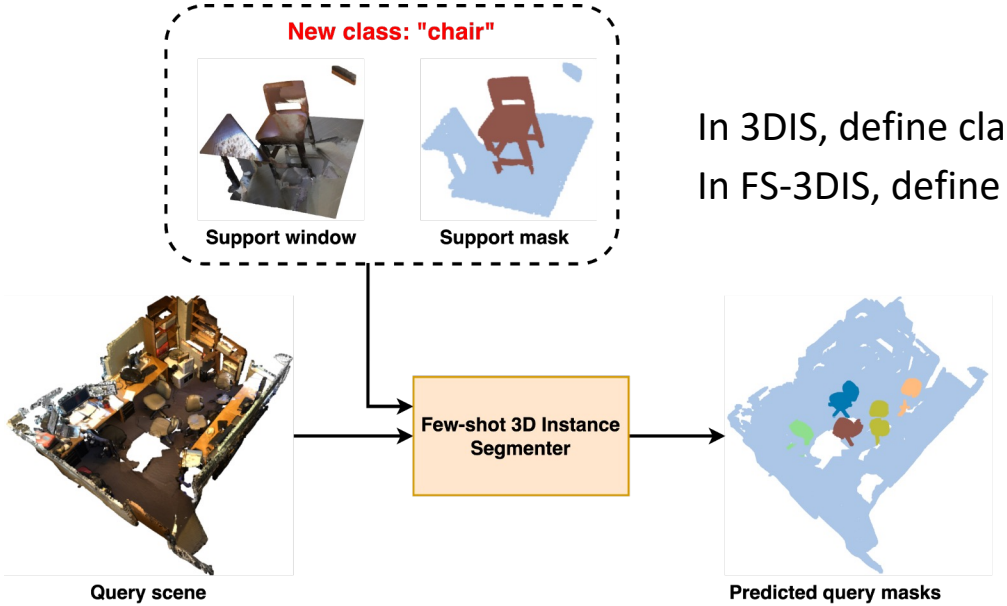


What if in testing we encounter new object categories?
Do we need to collect a lot of labeled examples for new classes?

$\Rightarrow$ We introduce a new task:

Few-shot 3D Instance Segmentation (FS-3DIS)

Few-shot 3D Instance Segmentation (FS-3DIS) – ECCV'22



In 3DIS, define classes of interest with full labels in **training**

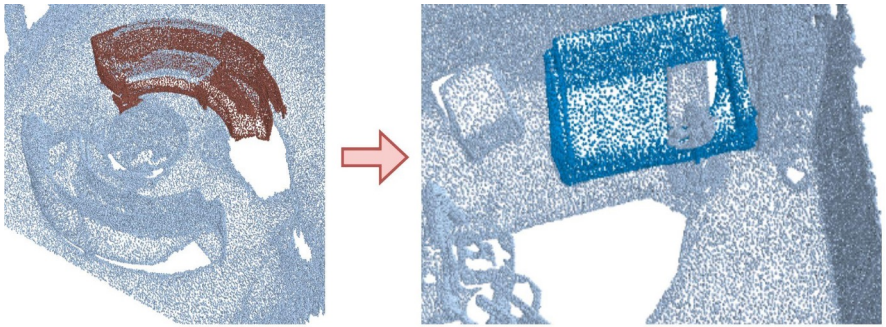
In FS-3DIS, define new classes by a few examples in **testing**

Problem setup:

- **In training**: abundant annotated examples of **base** classes are provided.
- **In testing**: given a **new** target class defined by K support examples, the goal is to segment all the instances of this target class in a query scene.

Challenges of Few-shot learning on 3D point clouds

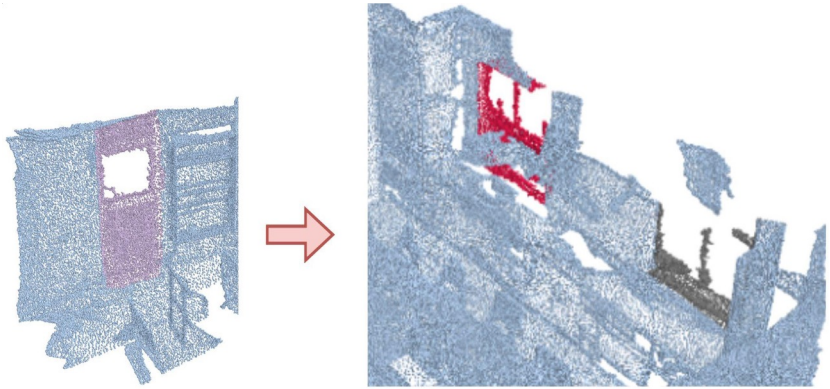
- **Insufficient support examples** for a **reliable training** of new classes.
- A few support examples cannot capture a **huge intra-class variation** in shape, size, and appearance of the objects.



Support sample

Test sample

“sofa”

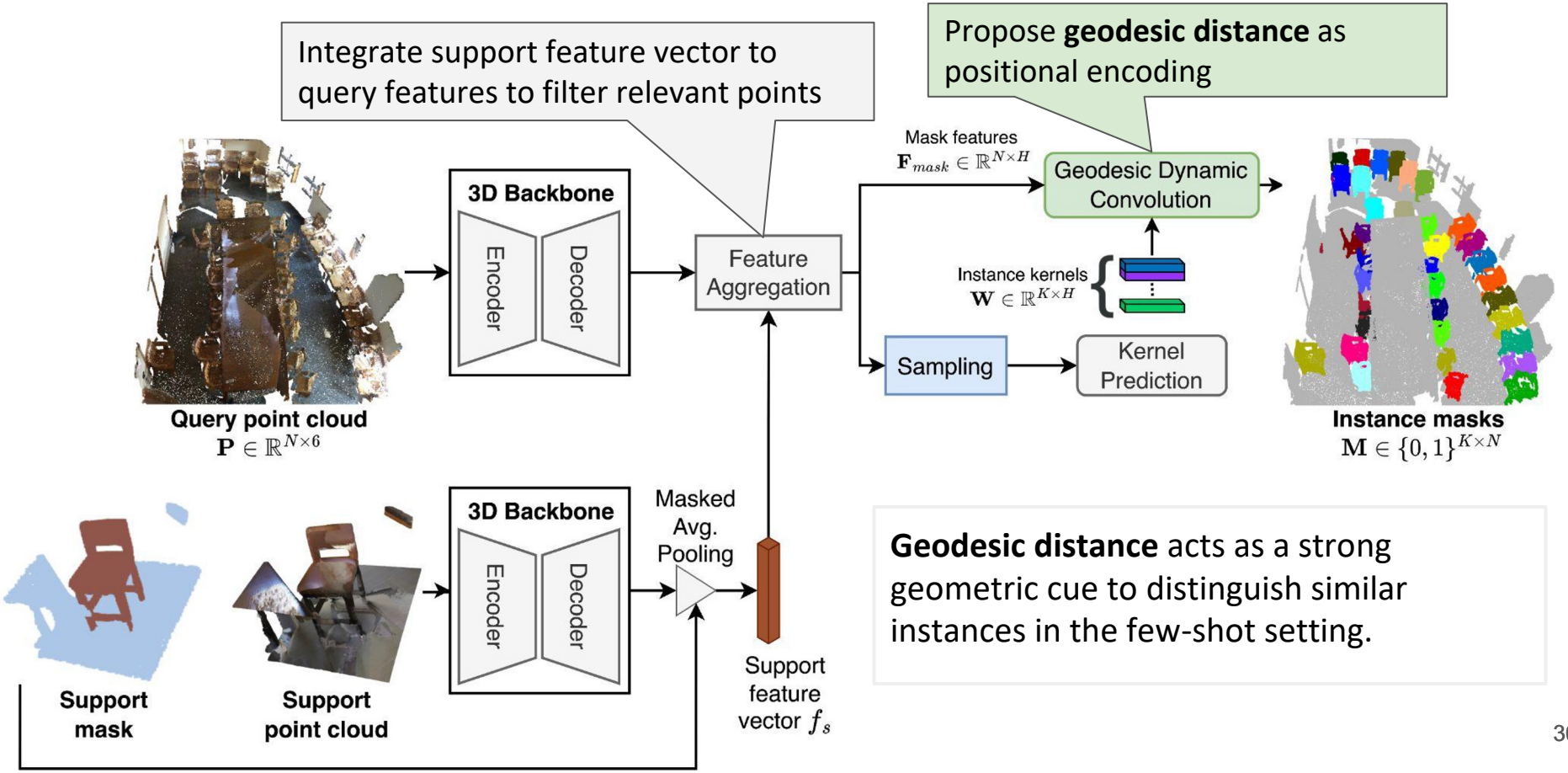


Support sample

Test sample

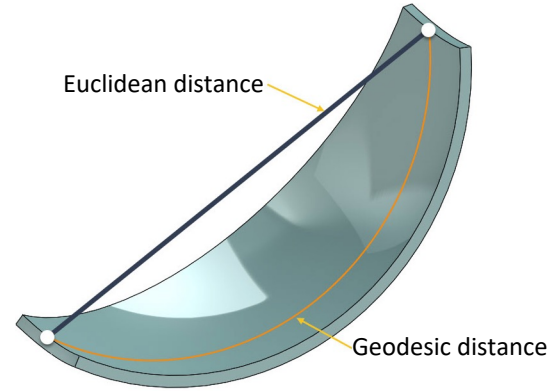
“window”

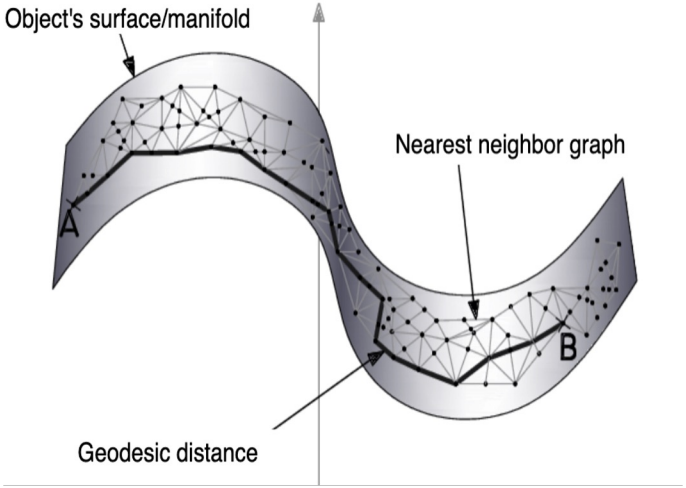
Our Approach

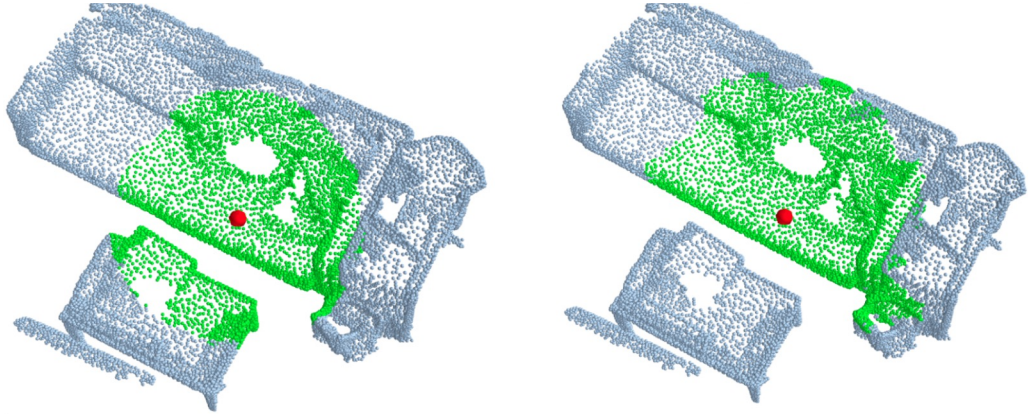


Geodesic distance acts as a strong geometric cue to distinguish similar instances in the few-shot setting.

Geodesic Distance







Top-2000 nearest neighbors of the red point in **Euclidean distance** (Left) and **Geodesic distance** (Right)

The geodesic distance between 2 points is approximately the **shortest distance** between their vertices in the **KNN graph** constructed from the point cloud.

Experiments

- Datasets:
  1. **ScanNetV2:** 18 categories (9 base + 9 novel classes).
  2. **S3DIS:** 12 categories (6 base + 6 novel classes).

- Metrics:
  1. **ScanNetV2**: AP, AP50 (average precision)
  2. **S3DIS:** mPrec, and mRec (mean precision, and mean recall).

- Setting:
  - Test on **1-shot** and **5-shot** settings.
  - Randomly sample **10 test sets** and report the average results.

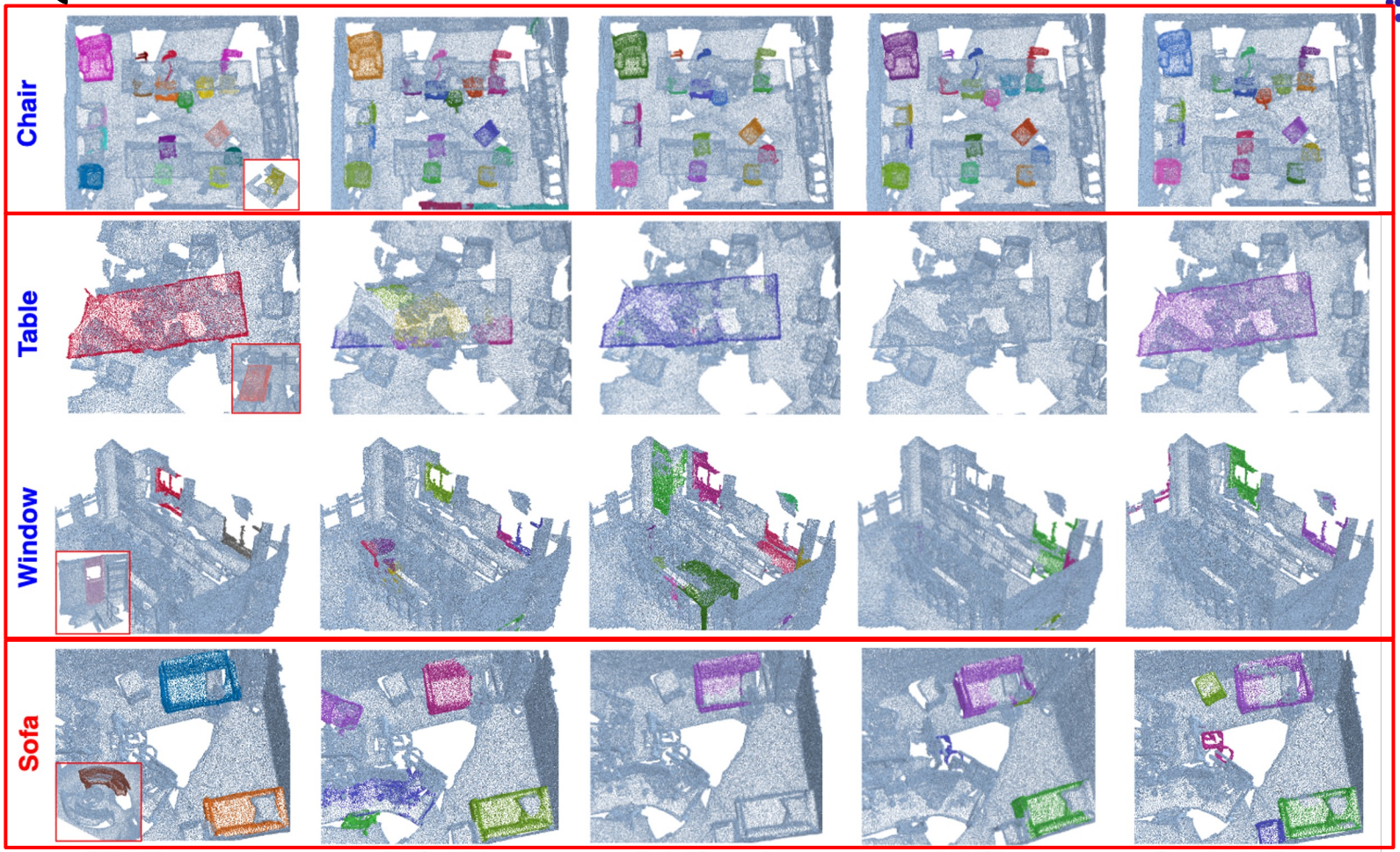
ScannetV2		S3DIS	
Fold 0	Fold 1	Fold 0	Fold 1
cabinet	sofa	beam	door
bed	table	board	floor
chair	window	bookcase	sofa
door	picture	ceiling	table
bookshelf	shower curtain	chair	wall
counter	refrigerator	column	window
desk	toilet		
curtain	sink		
bathtub	other furniture		
base	novel	base	novel

Comparison with adapted SOTA methods for FS-3DIS

Since we are the first to address FS-3DIS, we adapt the SOTA methods of 3DIS to the few-shot settings and compare these versions with our proposed approach

	ScanNetV2				S3DIS			
	1-shot		5-shot		1-shot		5-shot	
	AP	AP50	AP	AP50	mPrec	mRec	mPrec	mRec
DyCo3D's	6.2 ± 2.0	11.7 ± 3.1	6.4 ± 1.2	11.9 ± 2.2	2.9 ± 1.0	4.1 ± 1.4	3.1 ± 0.5	4.1 ± 1.4
PointGroup's	5.3 ± 1.2	10.3 ± 2.5	5.3 ± 0.5	11.7 ± 0.8	4.6 ± 1.4	3.8 ± 1.3	4.6 ± 0.6	3.8 ± 0.8
HAIS's	1.6 ± 0.6	3.5 ± 0.8	1.0 ± 0.2	2.3 ± 0.4	**8.1 ± 0.9**	3.9 ± 1.3	**11.8 ± 2.0**	4.1 ± 0.4
Ours	**10.6 ± 0.7**	**19.8 ± 1.4**	**13.2 ± 0.3**	**24.8 ± 1.3**	7.0 ± 0.4	**8.5 ± 1.7**	10.8 ± 1.3	**12.2 ± 1.8**
Ours (w/o geodesic)	7.6 ± 0.9	14.3 ± 1.0	11.7 ± 0.3	20.2 ± 0.7	6.7 ± 0.3	5.8 ± 1.4	8.1 ± 1.0	9.6 ± 1.3

Qualitative results on ScanNetV2



base class

novel classes

failure cases

Mis-segment
sofa stool to sofa

Conclusion

- Address **3DIS** with 2 contributions:
  - Replace the clustering by **Instance-aware Farthest Point Sampling**
  - Propose the **Box-aware Dynamic Convolution**

- We propose a new approach for **Box-supervised 3D Point Cloud Instance Segmentation (BS-3DSIS)** using **Gaussian Processes** to reduce annotation cost from mask to box!

- Propose a novel task: **Few-shot 3D Point Cloud Instance Segmentation (FS-3DIS)** and address this task by leveraging **Geodesic distance** as important geometric cue for segmenting new classes with a few examples

Thank you!