JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

ISBNet: a 3D Point Cloud Instance Segmentation Network with Instance-aware Sampling and Box-aware Dynamic Convolution

Tuan Duc Ngo Binh-Son Hua Khoi Nguyen

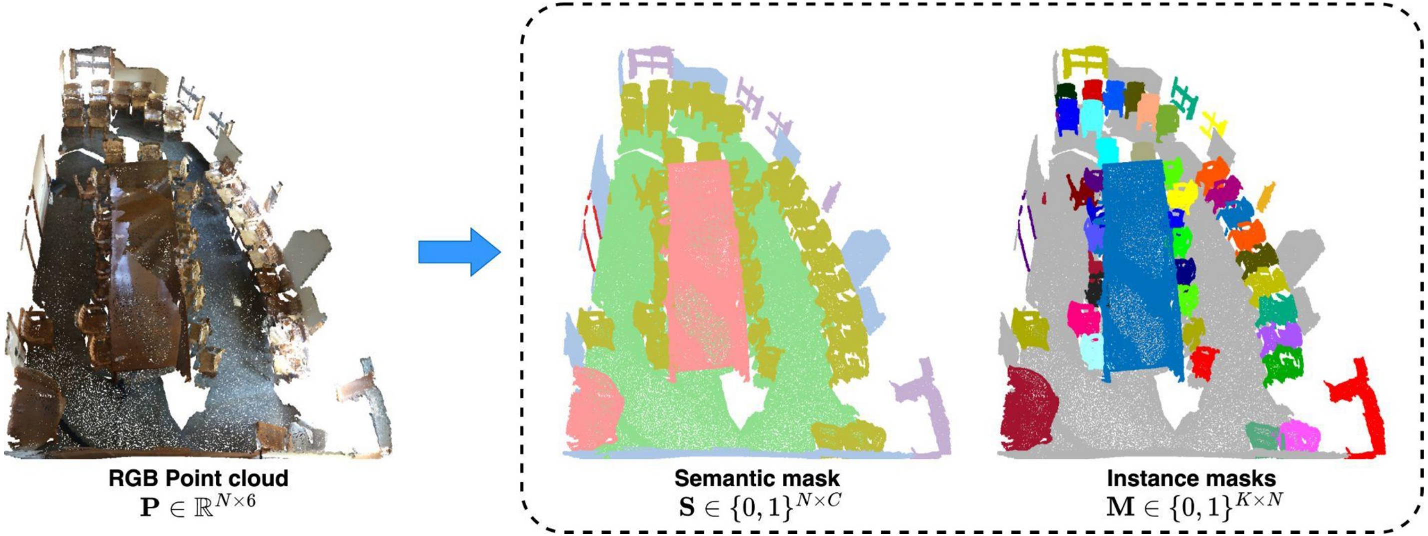
VinAI

Our code:

Paper Tag: **WED-PM-114**

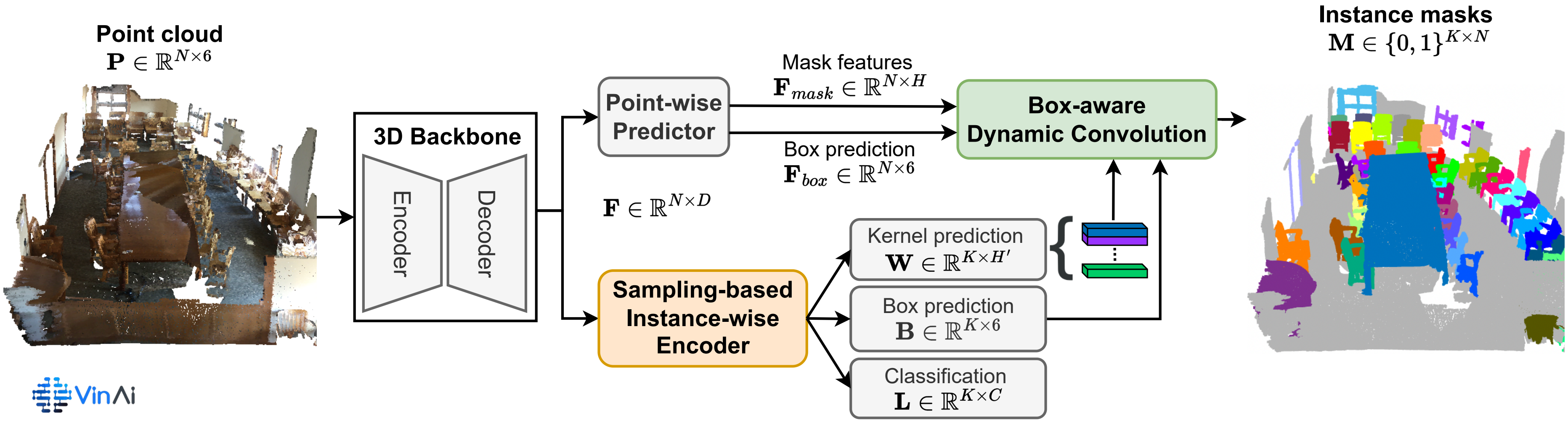
Overview

- **ISBNet** is a novel framework for **3D point cloud instance segmentation (3DIS)**



Overview

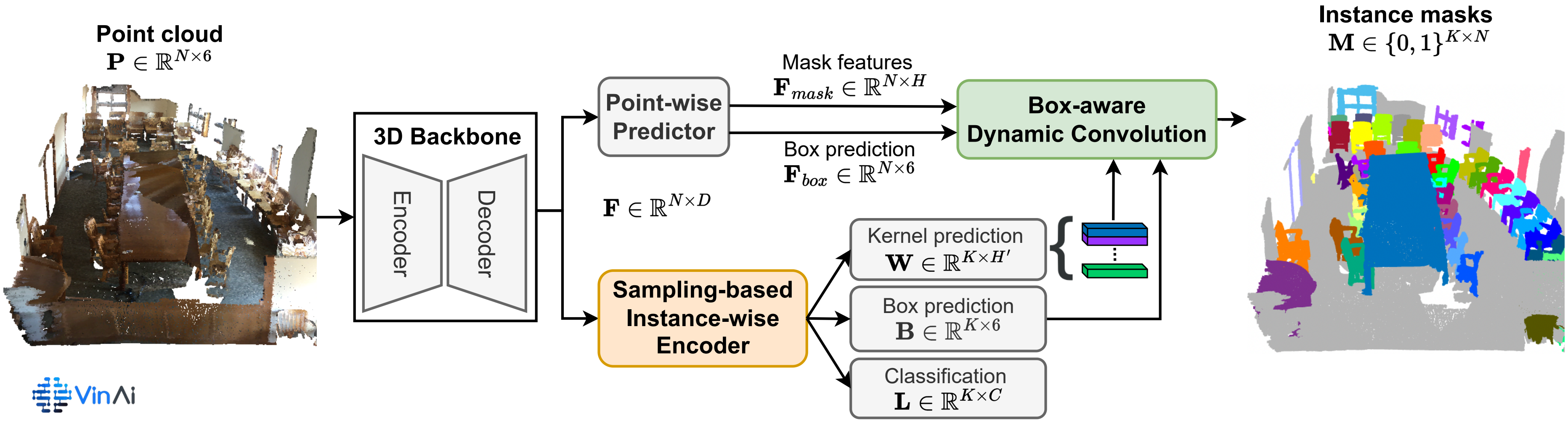
- **ISBNet** is a novel framework for **3D point cloud instance segmentation (3DIS)**



VinAi

Overview

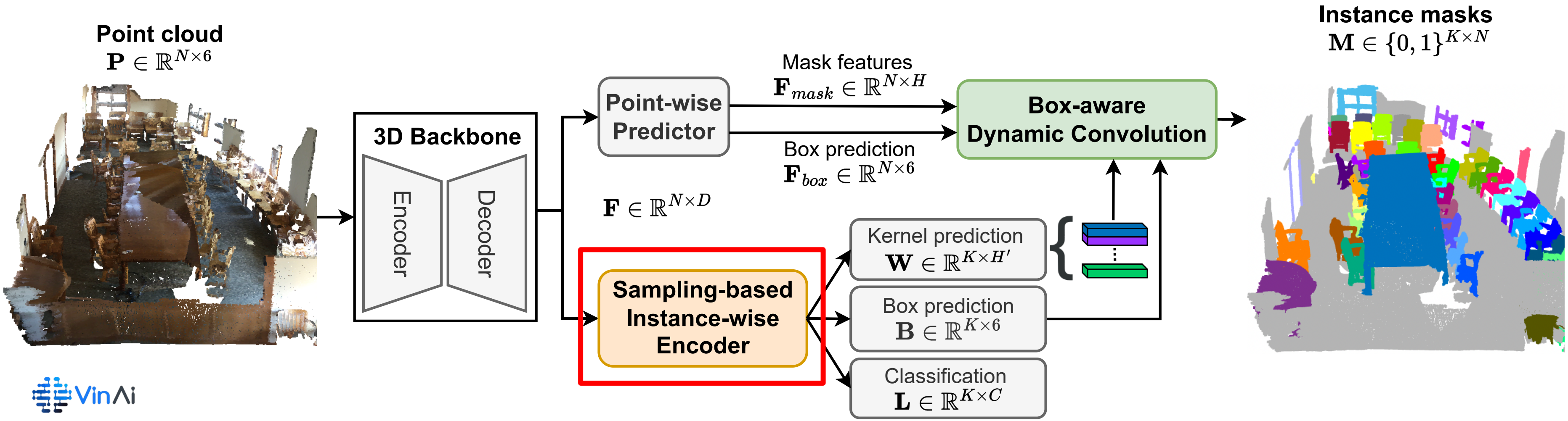
- **ISBNet** is a novel framework for **3D point cloud instance segmentation (3DIS)**
- We replace the clustering algorithm in existing 3DIS methods with a simple strategy to sample instance candidates



VinAi

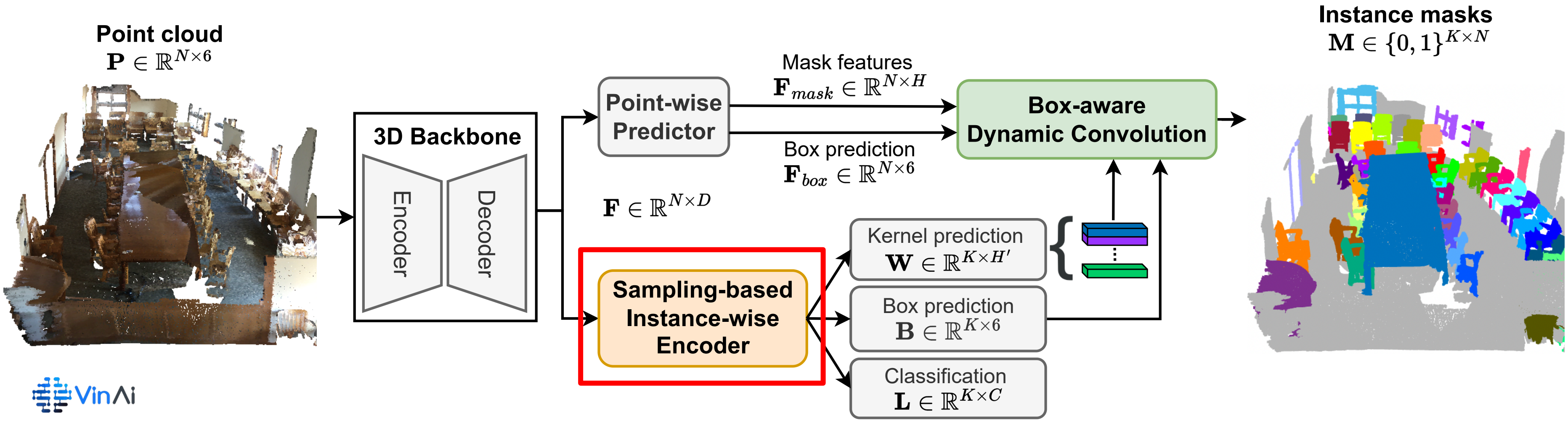
Overview

- **ISBNet** is a novel framework for **3D point cloud instance segmentation (3DIS)**
- We replace the clustering algorithm in existing 3DIS methods with a simple strategy to sample instance candidates



Overview

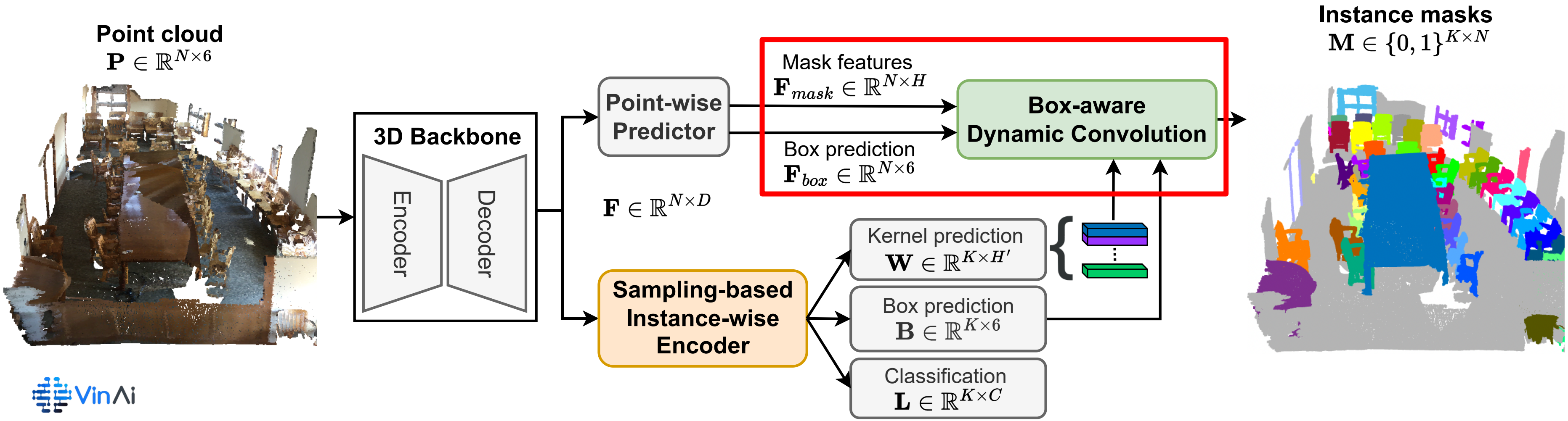
- **ISBNet** is a novel framework for **3D point cloud instance segmentation (3DIS)**
- We replace the clustering algorithm in existing 3DIS methods with a simple strategy to sample instance candidates
- We leverage the bounding box as a strong geometric cue to further boost performance



VinAI

Overview

- **ISBNet** is a novel framework for **3D point cloud instance segmentation (3DIS)**
- We replace the clustering algorithm in existing 3DIS methods with a simple strategy to sample instance candidates
- We leverage the bounding box as a strong geometric cue to further boost performance



VinAi

Overview

- **ISBNet** is a novel framework for **3D point cloud instance segmentation (3DIS)**
- We replace the clustering algorithm in existing 3DIS methods with a simple strategy to sample instance candidates
- We leverage the bounding box as a strong geometric cue to further boost performance
- Our method set new SOTA results on various datasets and retains fast inference time

ScanNetV2

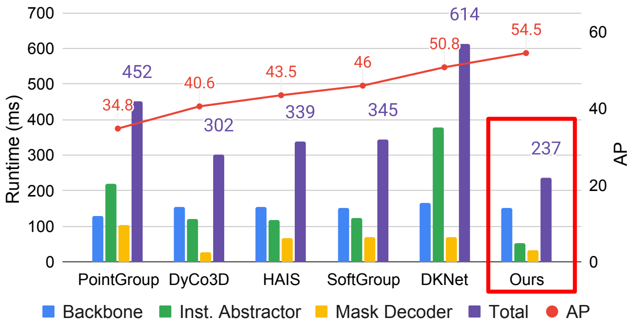
Method	Venue	AP	AP_{50}	AP_{25}
SGPN [35]	CVPR 18	4.9	14.3	26.1
MTML [39]	ICCV 19	28.2	54.9	73.1
3D-BoNet [39]	NeurIPS 19	25.3	48.8	68.7
PointGroup [21]	CVPR 20	40.7	63.6	77.8
OccuSeg [14]	CVPR 20	44.3	67.2	74.2
DyCo3D [16]	CVPR 21	39.5	64.1	76.1
PE [41]	CVPR 21	39.6	64.5	77.6
HAIS [5]	ICCV 21	45.7	69.9	80.3
SSTNet [26]	ICCV 21	50.6	69.8	78.9
SoftGroup [34]	CVPR 22	50.4	76.1	**86.5**
RPGN [9]	ECCV 22	42.8	64.3	80.6
PointInst3D [17]	ECCV 22	43.8	-	-
Di&Co3D [42]	ECCV 22	47.7	70.0	80.2
DKNet [38]	ECCV 22	53.2	71.8	81.5
ISBNet	-	**55.9**	**76.3**	84.5

S3DIS

Method	AP	AP_{50}	mCov	mWCov	$mPrec_{50}$	$mRec_{50}$
SGPN† [40]	-	-	32.7	35.5	36.0	28.7
PointGroup† [21]	-	57.8	-	-	61.9	62.1
HAIS† [5]	-	-	64.3	66.0	71.1	65.0
SSTNet† [26]	42.7	59.3	-	-	65.6	64.2
SoftGroup† [34]	51.6	**66.1**	66.1	68.0	73.6	66.6
RPGN† [9]	-	-	-	-	64.0	63.0
PointInst3D† [17]	-	-	64.3	65.3	73.1	65.2
Di&Co3D† [42]	-	-	65.5	66.1	63.9	67.2
DKNet† [38]	-	-	64.7	65.6	70.8	65.3
ISBNet†	**54.0**	65.8	**71.6**	**70.9**	**74.2**	**72.7**
SGPN‡ [40]	-	54.4	37.9	40.8	38.2	31.2
3D-BoNet‡ [39]	-	-	-	-	65.6	47.7
PointGroup‡ [21]	-	64.0	-	-	69.6	69.2
OccuSeg‡ [14]	-	-	-	-	72.8	60.3
HAIS‡ [5]	-	-	67.0	70.4	73.2	69.4
SSTNet‡ [26]	54.1	67.8	-	-	73.5	73.4
SoftGroup‡ [34]	54.4	68.9	69.3	71.7	75.3	69.8
RPGN‡ [9]	-	-	-	-	**84.5**	70.5
PointInst3D‡ [17]	-	-	71.5	74.1	76.4	74.0
DKNet‡ [38]	-	-	70.3	72.8	75.3	71.1
ISBNet‡	**60.8**	**70.5**	**74.9**	**76.8**	77.5	**77.1**

STPLS3D

	AP	AP_{50}
PointGroup [21]	23.3	38.5
HAIS [5]	35.1	46.7
SoftGroup [34]	46.2	61.8
ISBNet	**49.2**	**64.0**



ISBNet: a 3D Point Cloud Instance Segmentation Network with Instance-aware Sampling and Box-aware Dynamic Convolution

More details

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.

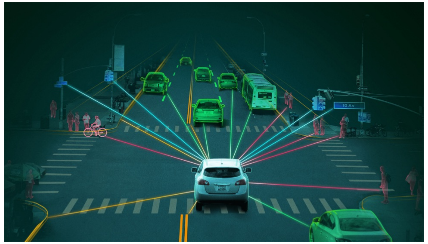
K: # object instances,
N: # points, C: # categories

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

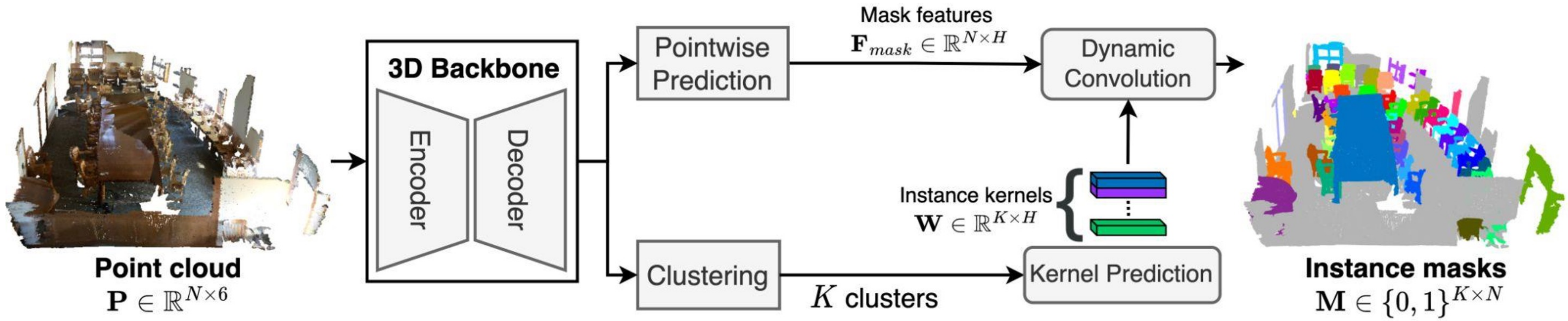


Challenges

- 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

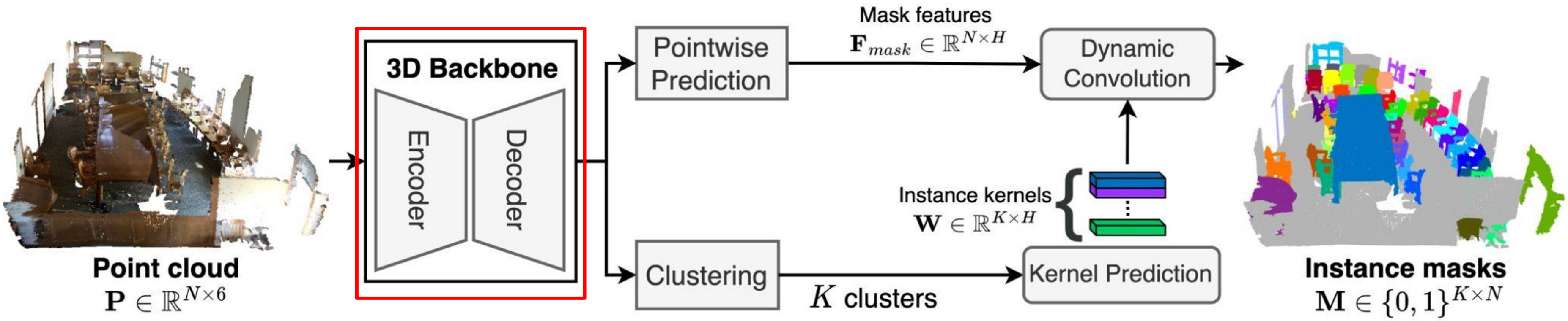
➔ It is not trivial to apply 2D instance Segmentation approaches to 3DIS

A Typical Approach for 3DIS: DyCo3D [1]



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

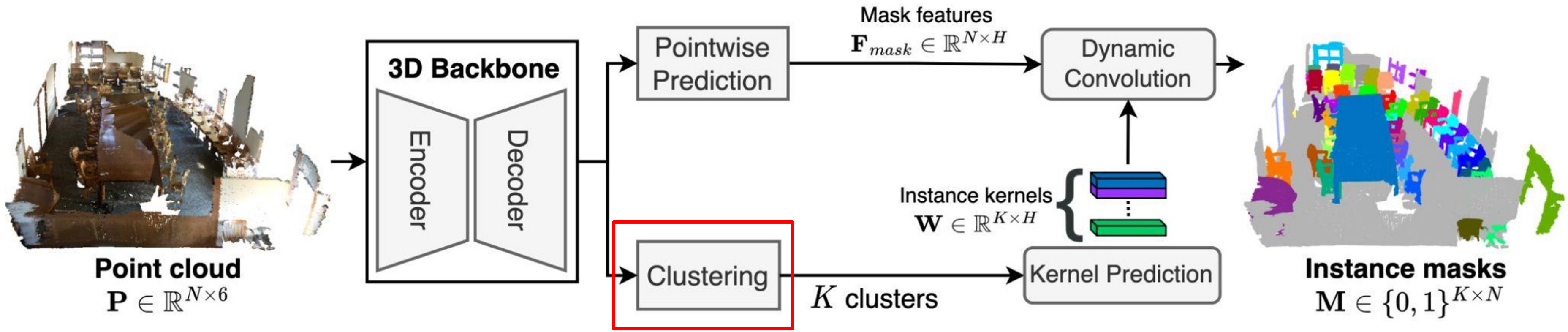
A Typical Approach for 3DIS: DyCo3D [1]



1. Use a 3D backbone to extract pointwise features

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

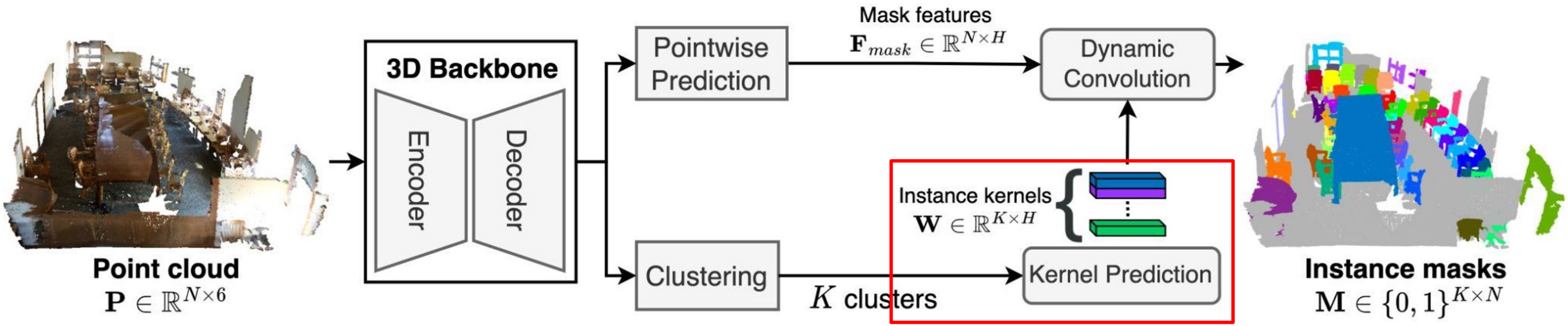
A Typical Approach for 3DIS: DyCo3D [1]



1. Use a 3D backbone to extract pointwise features
2. Predict instance masks:
    a. Group points into **clusters** for object candidates

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

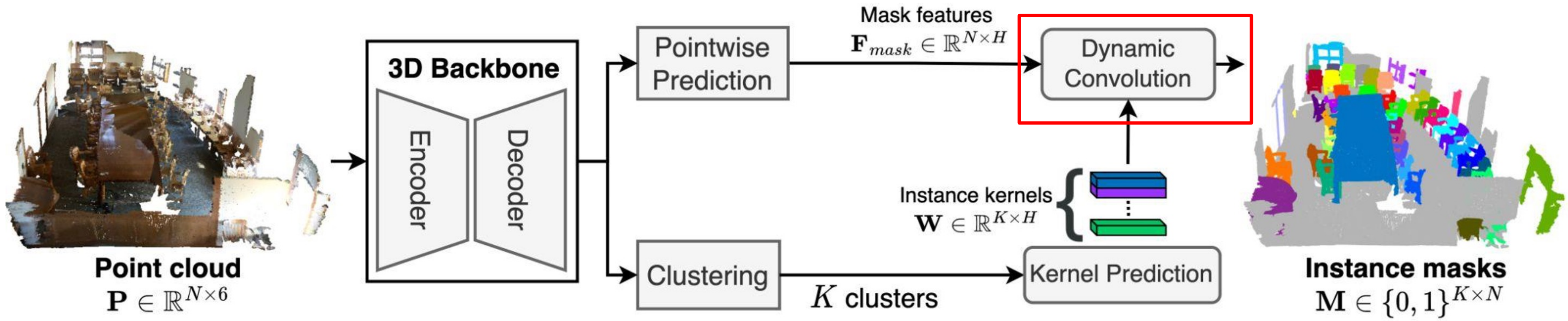
A Typical Approach for 3DIS: DyCo3D [1]



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

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

A Typical Approach for 3DIS: DyCo3D [1]



1. Use a 3D backbone to extract pointwise features

2. Predict instance masks:

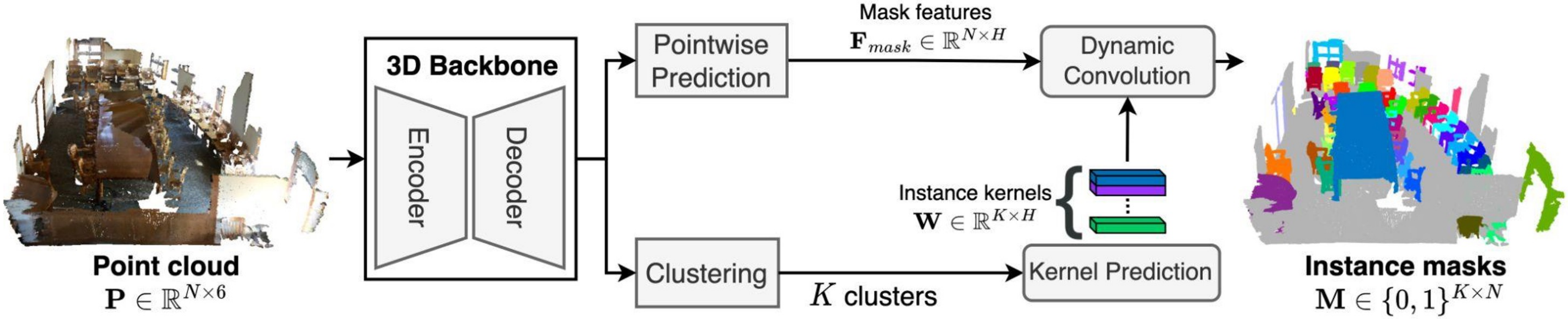
   a. Group points into **clusters** for object candidates

   b. Generate an **instance kernel** for each object candidate

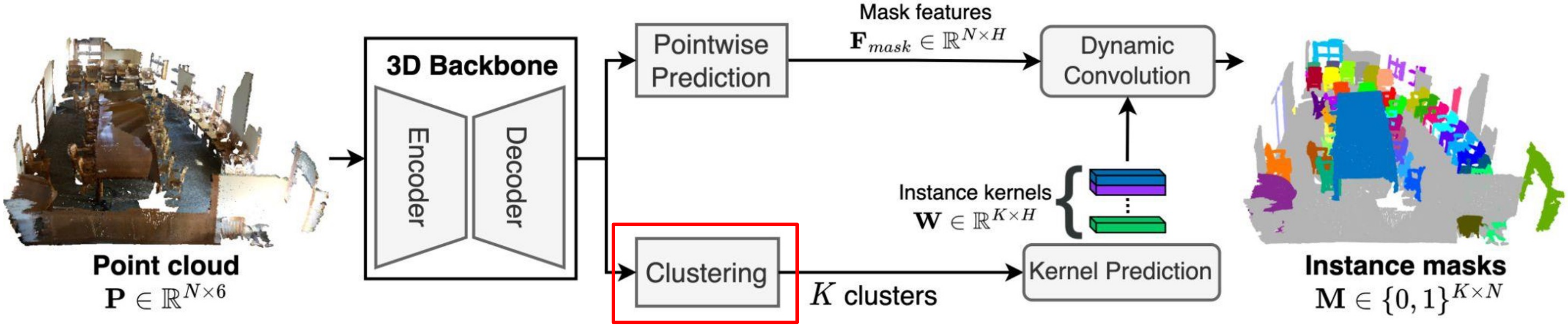
   c. **Dynamic convolution:** Convolve each generated kernel with mask features to obtain a binary instance mask for each object

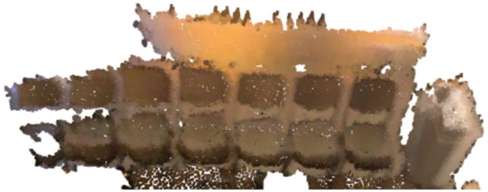
[1] DyCo3D: Robust Instance Segmentation of 3D Point Clouds through Dynamic Convolution (CVPR 2021)

Limitations of DyCo3D

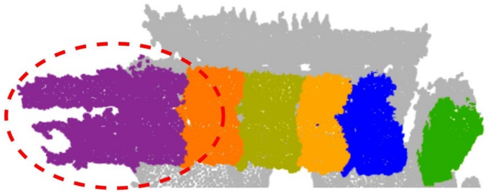


Limitations of DyCo3D





RGB point cloud

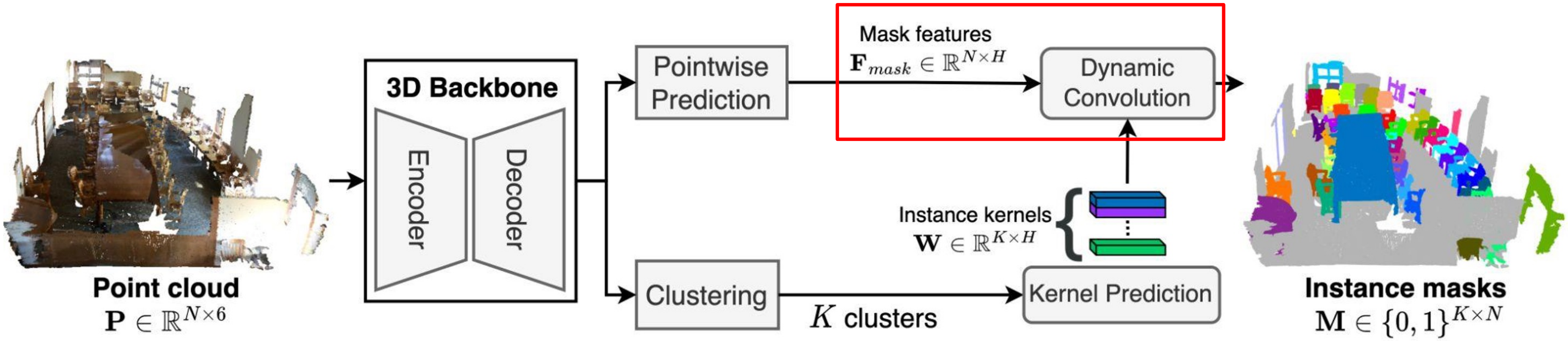


DyCo3D's predictions

Mis-grouping points when similar objects are adjacent

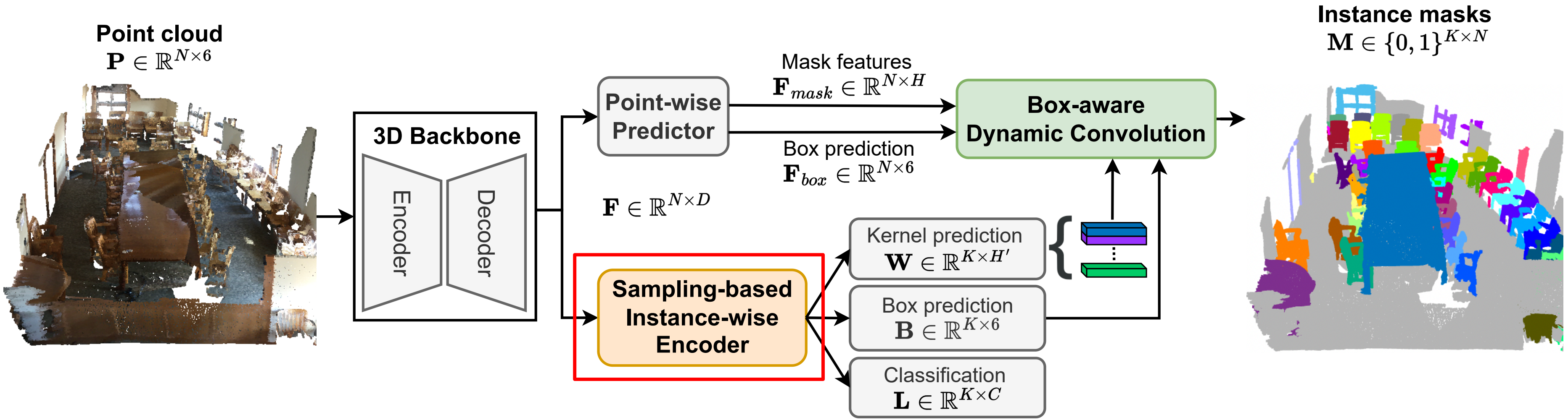
➔ **Low-recall object proposals**

Limitations of DyCo3D



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

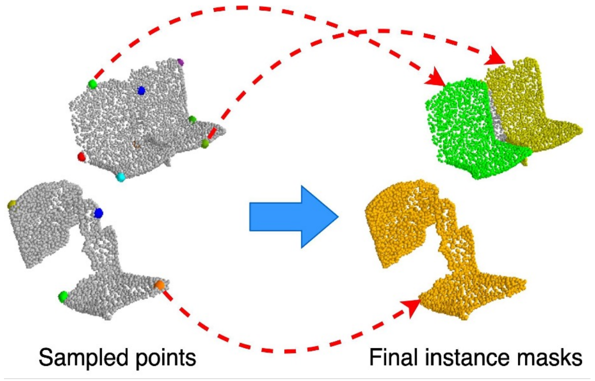
Our ISBNet



Replace clustering by **instance-aware sampling**:

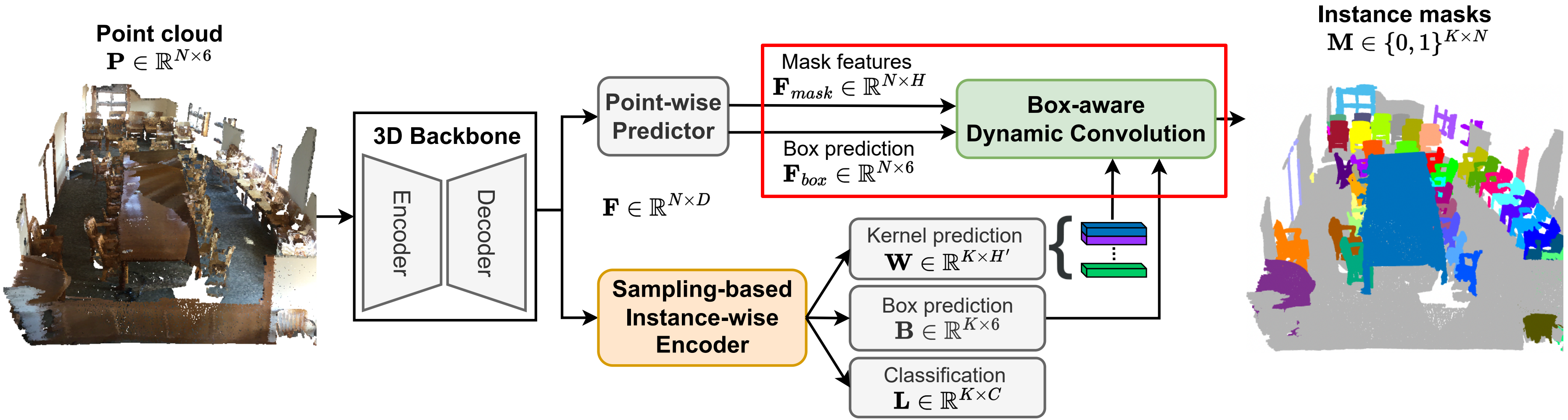
Each sampled point represents a **candidate object** to obtain instance mask

➔ **high-recall object proposals**



VinAi

Our ISBNet



Propose **Box-aware Dynamic Convolution**:

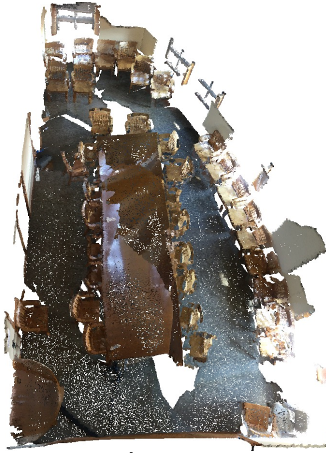
Enhance appearance feature with geometric cue, i.e., **object bounding box**

Instance-aware Sampling (IA-FPS)

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

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**.

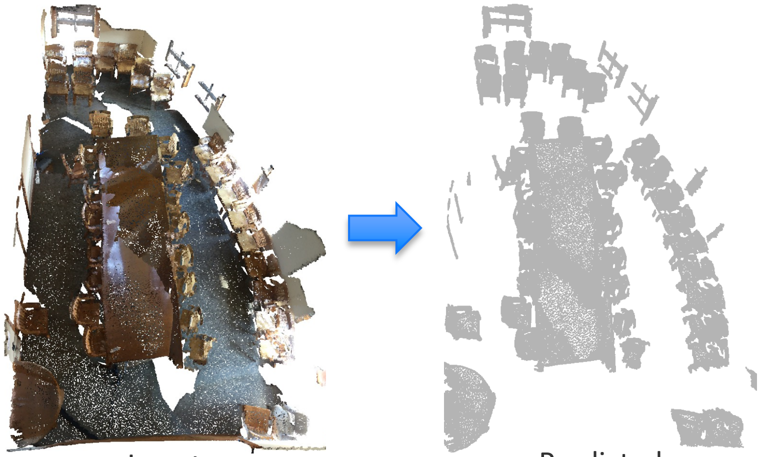


Input
point cloud

Instance-aware Sampling (IA-FPS)

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

- Only sample from **foreground points**



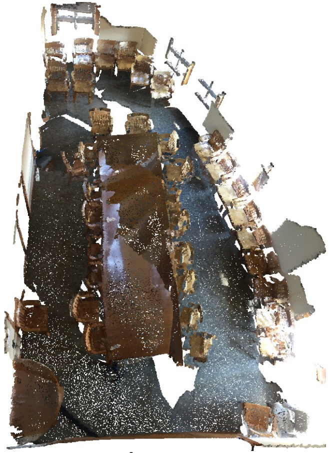
Input point cloud

Predicted foreground points

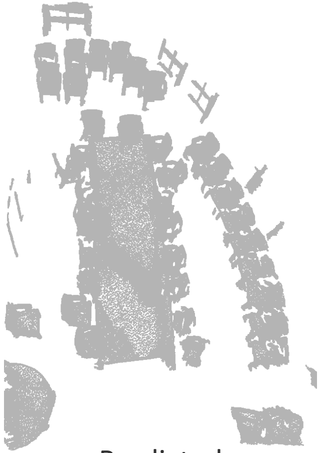
Instance-aware Sampling (IA-FPS)

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 sampling and object mask prediction: Avoid **points belonging to previous predicted instance masks**



Input
point cloud

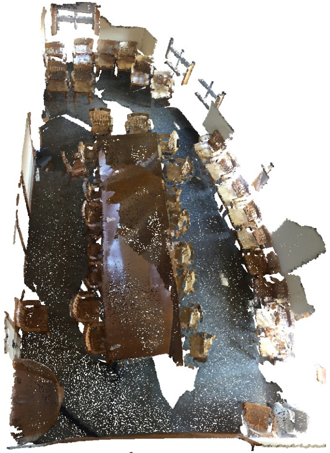


Predicted
foreground points

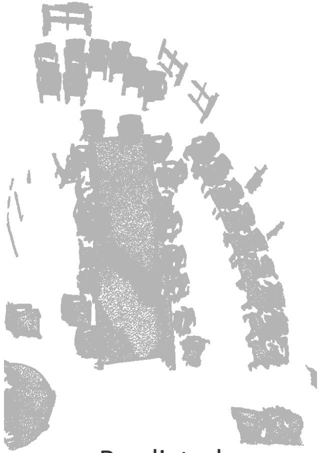
Instance-aware Sampling (IA-FPS)

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

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



Input
point cloud



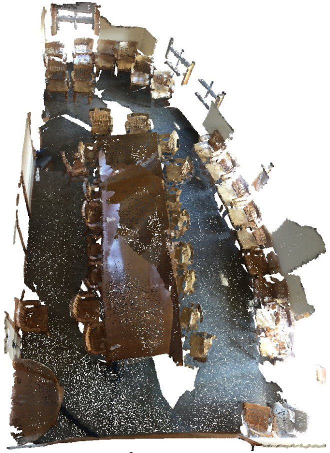
Predicted
foreground points

VinAi

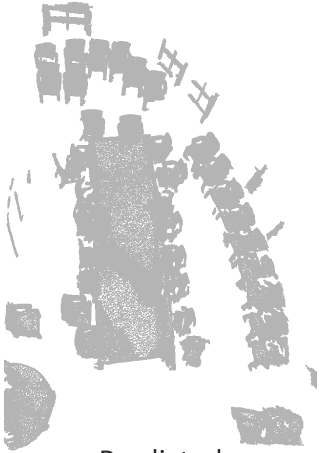
Instance-aware Sampling (IA-FPS)

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

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



Input point cloud



Predicted foreground points

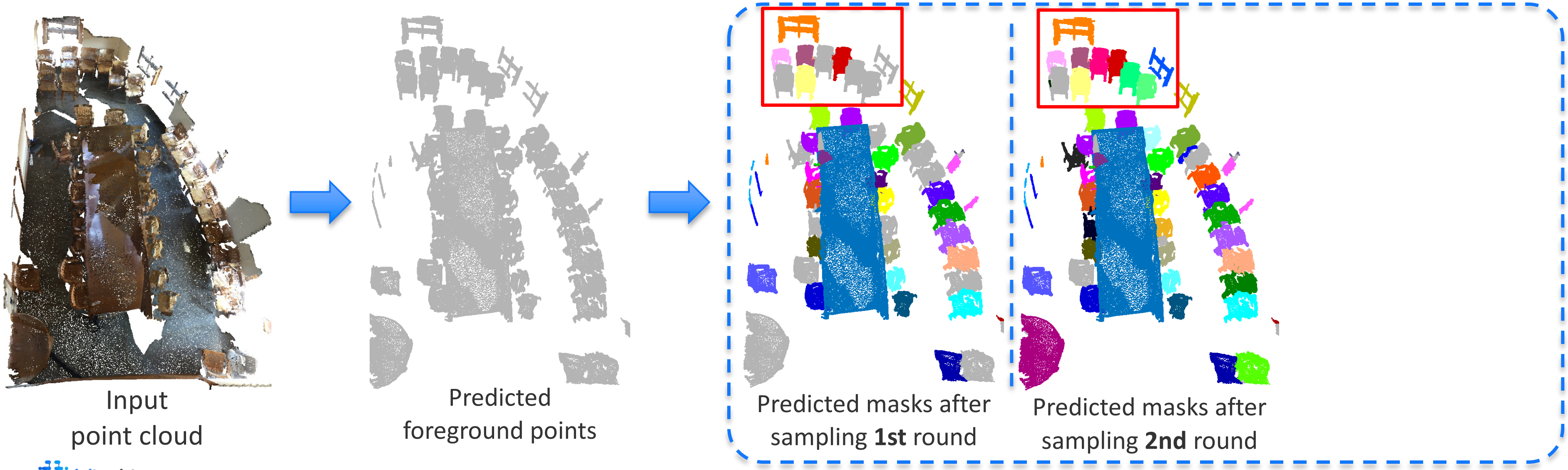
Predicted masks after sampling **1st** round

VinAi

Instance-aware Sampling (IA-FPS)

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

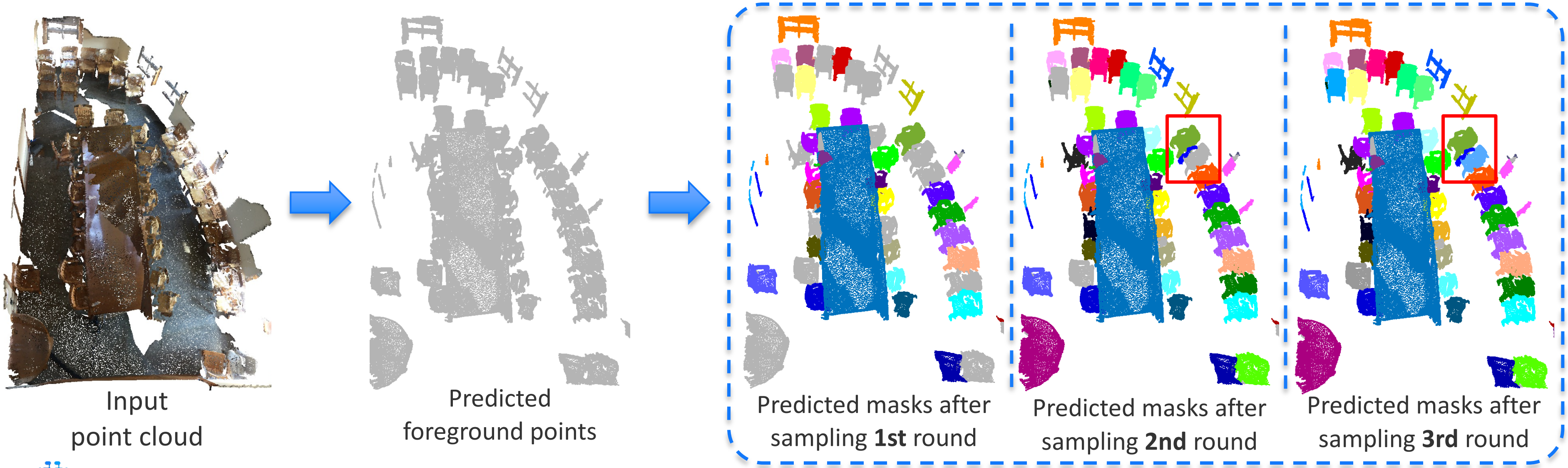
- Only sample from **foreground points**
- Multiple-rounds sampling and object mask prediction: Avoid **points belonging to previous predicted instance masks**



Instance-aware Sampling (IA-FPS)

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

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



Input point cloud

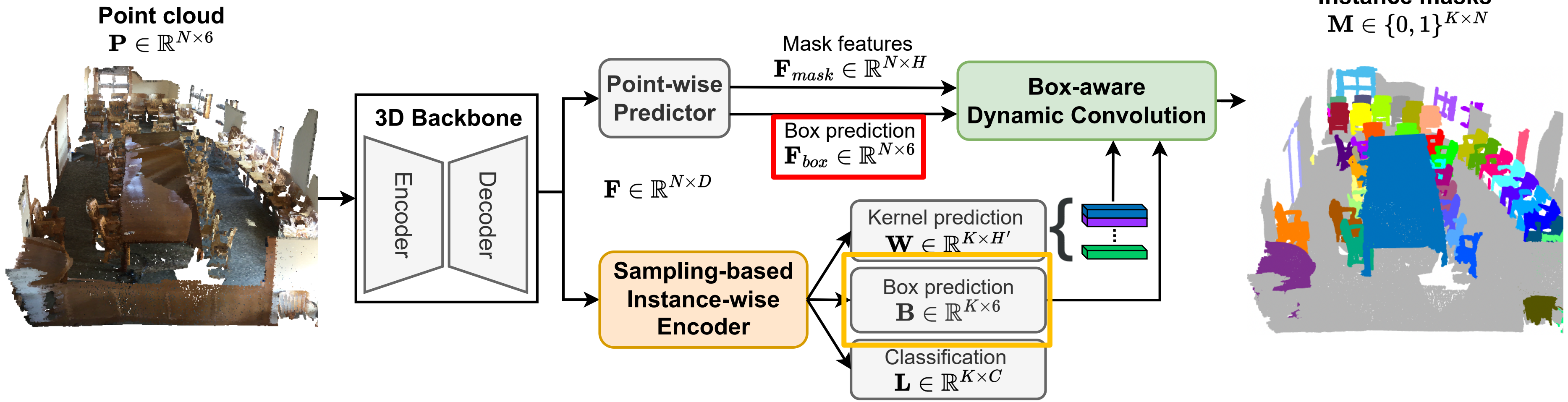
Predicted foreground points

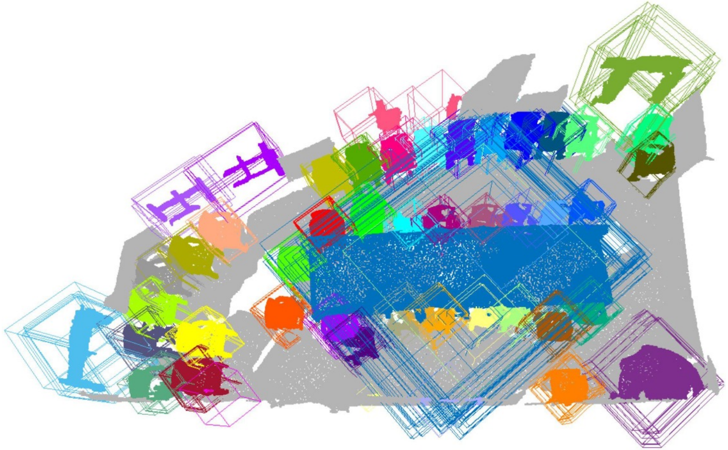
Predicted masks after sampling **1st** round

Predicted masks after sampling **2nd** round

Predicted masks after sampling **3rd** round

Box-aware Dynamic Convolution

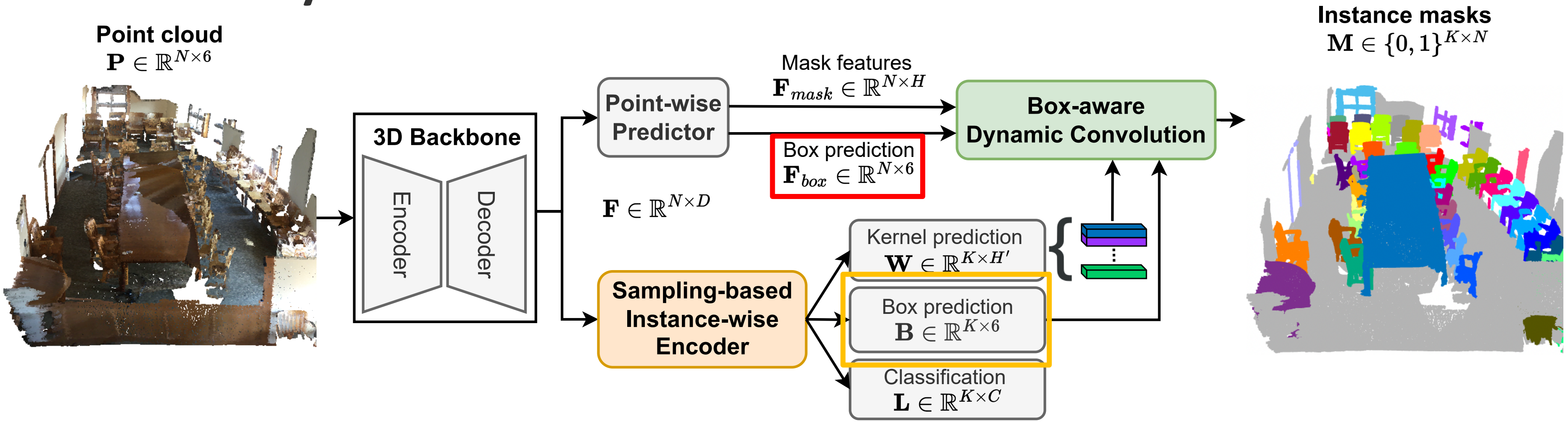




Intuition: an object candidate **(2)** "attracts" points **(1)** predicting similar boxes

VinAI

Box-aware Dynamic Convolution



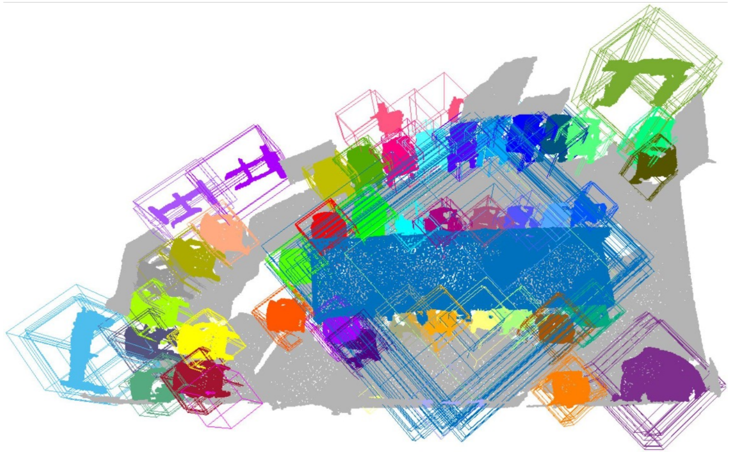
$$\widehat{m}^{(k)} = \text{Sigmoid}\Big(\text{Conv}\Big(\big[\mathbf{F}_{mask}; \mathbf{F}^{(k)}_{box}\big]; w^{(k)}\Big)\Big)$$

Instance kernel

Box difference

Appearance feature

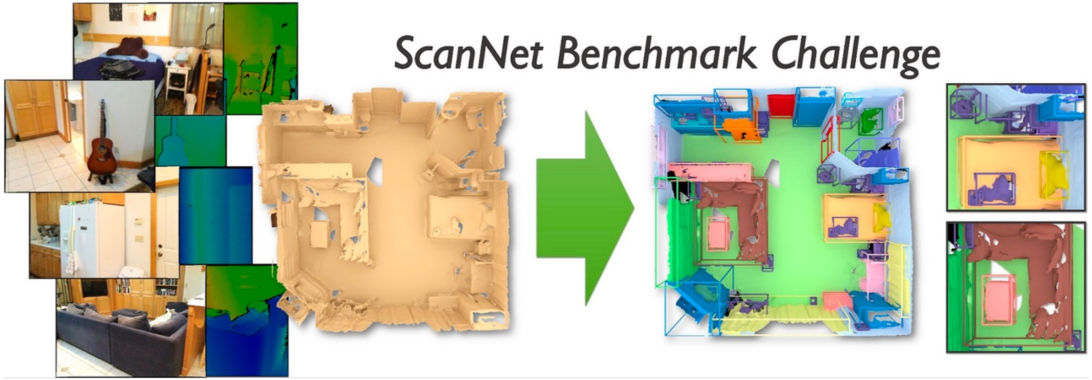
Box difference: difference in box size and box center between pointwise predicted box and object candidate's predicted box



VinAI

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**

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	-	-
+13.9	**Ours**	CVPR 23	**54.5**	**73.1**	74.2	72.7	49.2	64.0
			+4.3	+5.5				

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	+9.9
		+4.3	+5.5	+0.6	+5.5			

VinAI

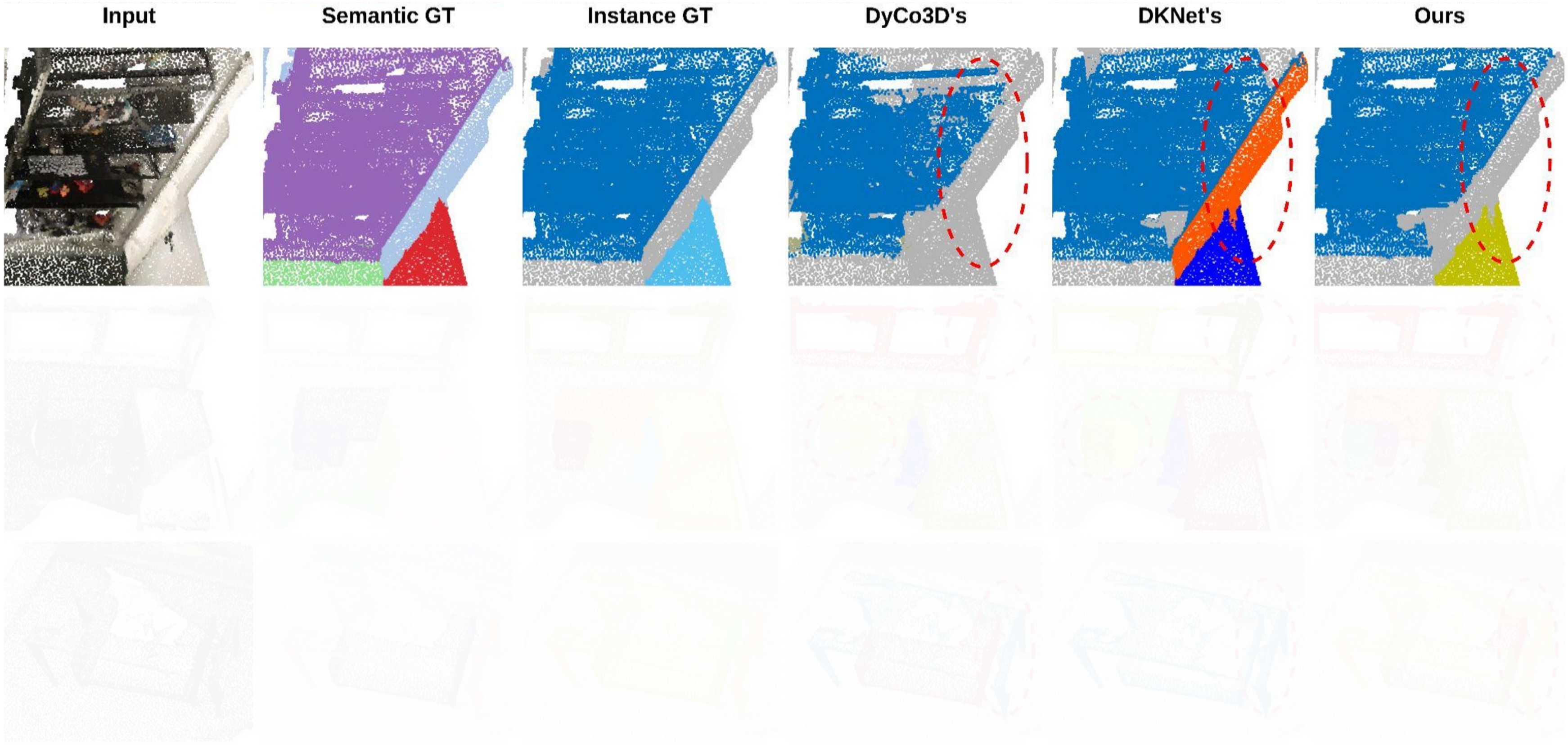
Comparison with SOTA

		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

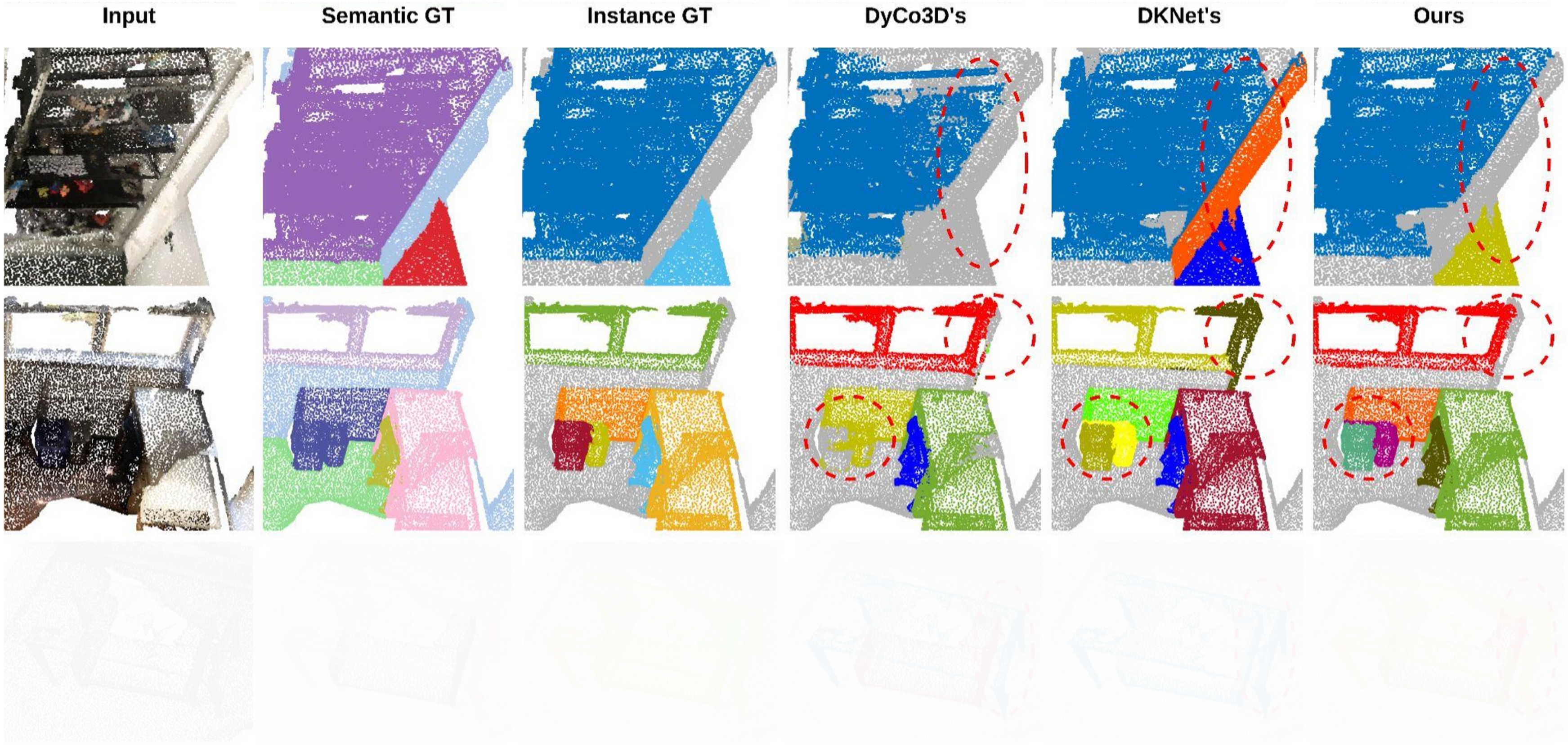
VinAI

Qualitative Results on ScanNetV2

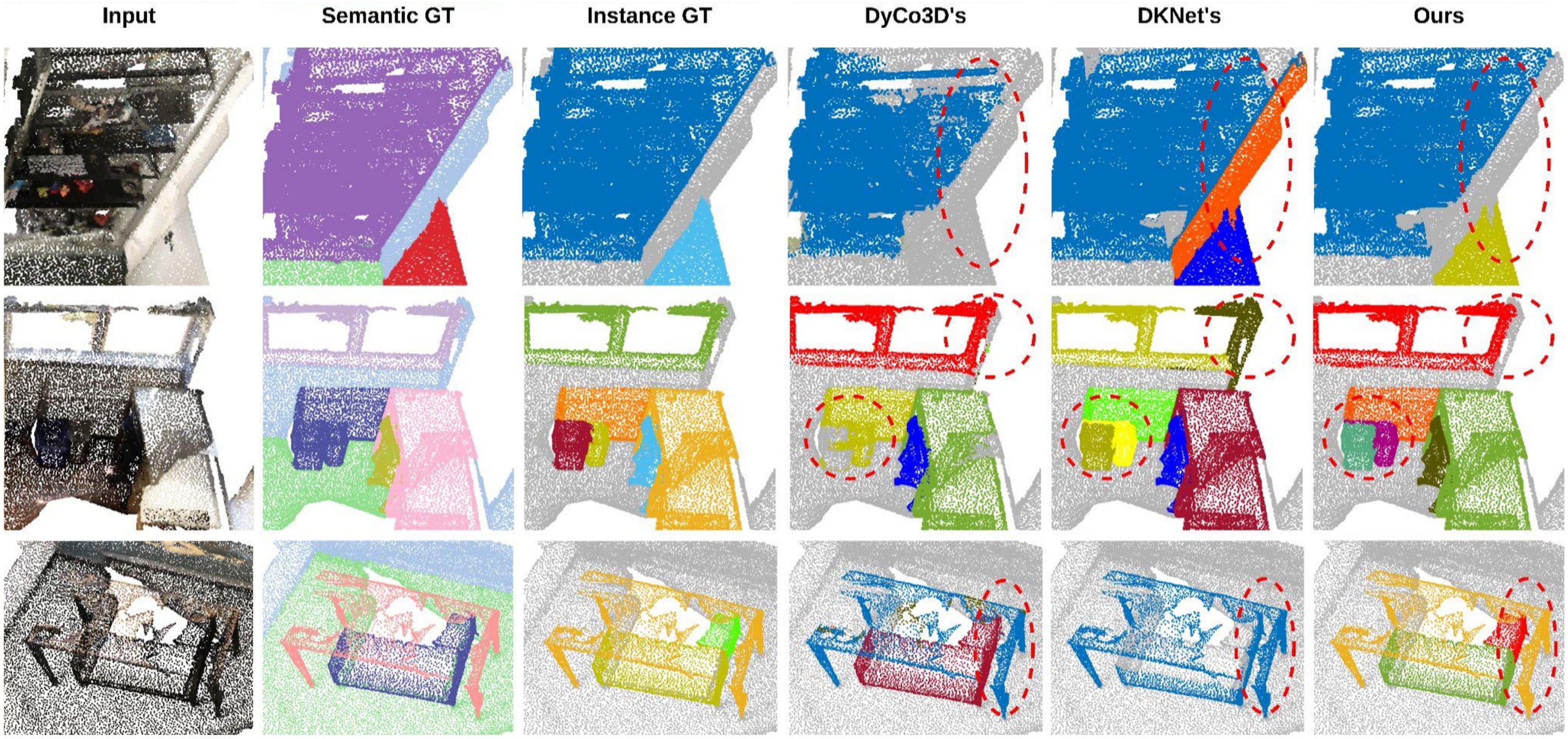
Qualitative Results on ScanNetV2



Qualitative Results on ScanNetV2



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**

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

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**

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

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**
			+6.6	+6.7	+5.4

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

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**
			+6.6	+6.7	+5.4

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

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**

Multiple rounds sampling

VinAi

Ablation Study on ScanNetV2

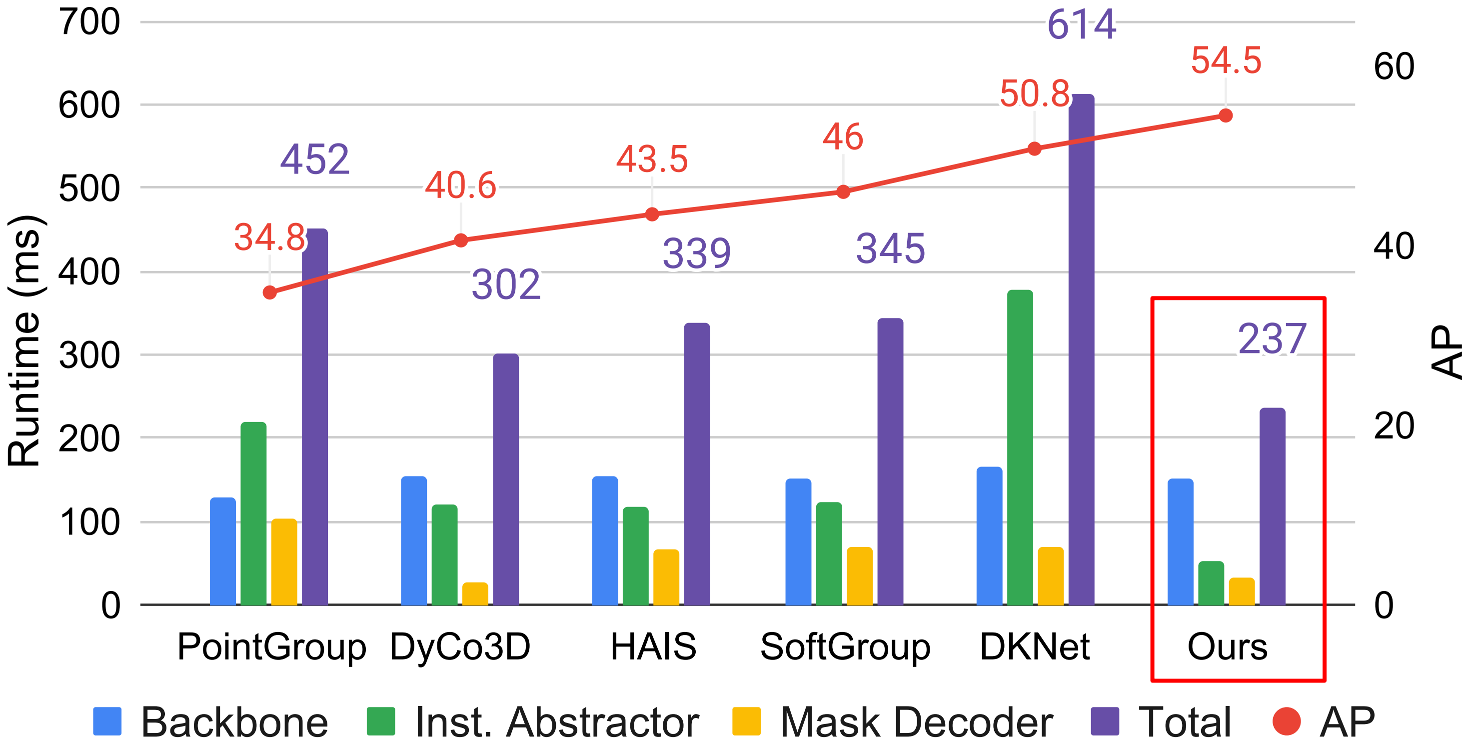
	IA-FPS	BA-DyCo	AP	AP_{50}	AP_{25}
Baseline			47.9	66.4	77.1
	✓		53.4	71.9	81.8
		✓	48.6	67.7	77.8
ISBNet	✓	✓	**54.5**	**73.1**	**82.5**

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**

Multiple rounds sampling

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

Runtime Analysis



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

That’s it!

Want more? Check out our code
https://github.com/VinAIResearch/ISBNet

VinAi