Datasets, Tasks, and Training Methods for Large-scale Hypergraph Learning

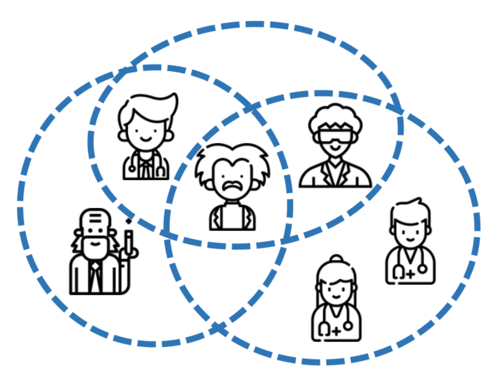
Sunwoo Kim[1]*, Dongjin Lee[2]*, Yul Kim[3], Jungho Park[3], Taeho Hwang[3], Kijung Shin[1,2]
[1]Kim Jaechul Graduate School of AI, [2] School of Electrical Engineering, KAIST, South Korea
[3]Samsung Research, South Korea
Code and Data: https://github.com/kswoo97/pcl

Summary

- **Goal: to find high-quality node representations on large-scale hypergraphs**
- **Previous Work:**
  - **Limited downstream tasks**: A few single-entity level downstream tasks have been used to evaluate the hypergraph neural network models.
  - **Small-scale benchmark datasets**: The evaluation of current hypergraph neural network models has been limited to small datasets (10k-scale).
  - **Underdeveloped training strategy for large-scale hypergraphs**: Most of the current hypergraph learning approaches are not scalable.
- **Contributions***:*
  - We present **two new pair-level prediction tasks**.
  - We construct and publicly release **two large-scale hypergraph datasets**.
  - We propose **PCL** (**P**artitioning-based **C**ontrastive **L**earning), **a scalable contrastive learning method** for hypergraph neural networks (HNNs).

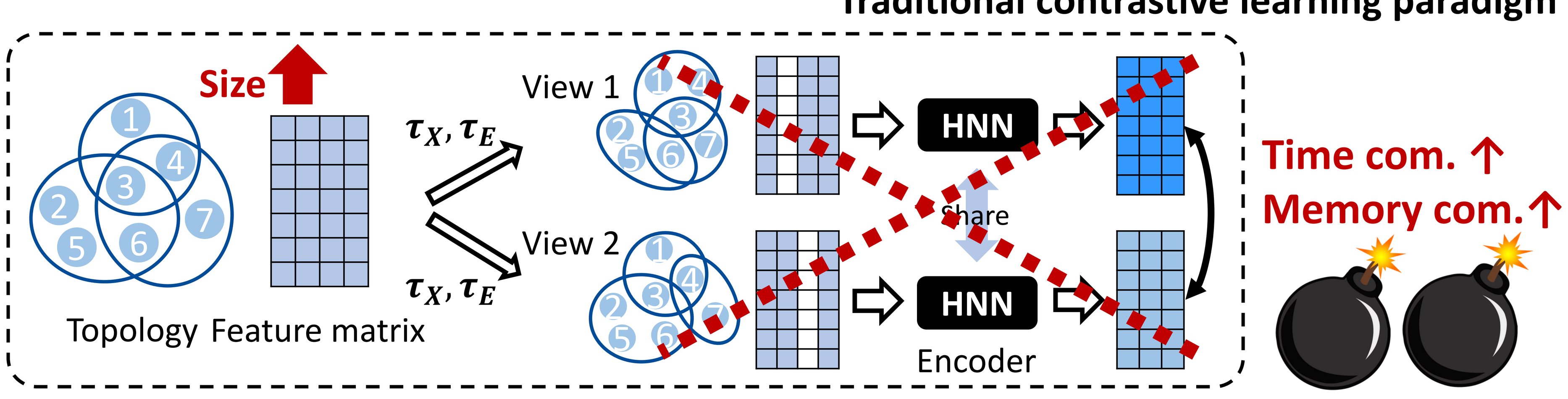
Background: Hypergraph & Contrastive Learning

- **Hypergraph:** A set of hyperedges that allow containing any number of nodes
- **Examples:** Collaborations of researchers, joint interactions of proteins, co-purchases of items
- **Properties:** A hypergraph naturally models group interactions.


Collaborations of researchers

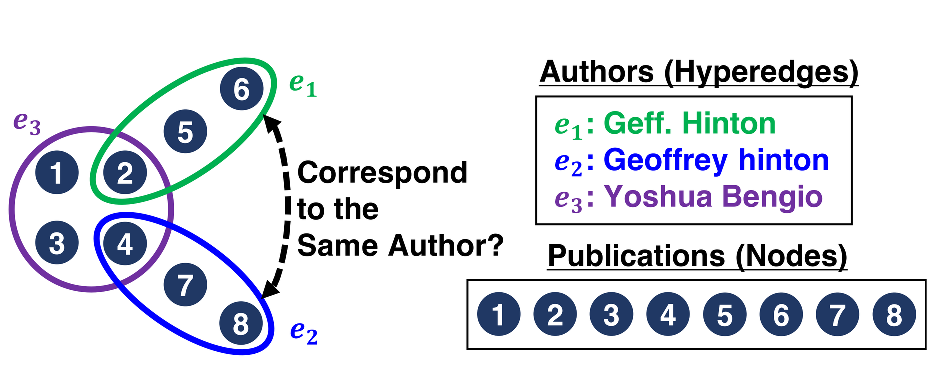
- **Contrastive learning** aims to maximize the agreement between differently augmented views of the same input.
- Previous research on hypergraph contrastive learning has a scalability issue when dealing with large-scale hypergraphs.

Traditional contrastive learning paradigm

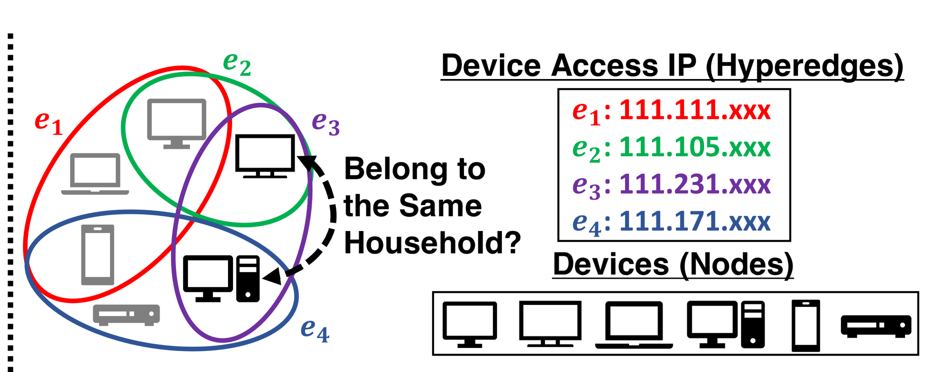


Proposed Tasks

- **Task1: Hyperedge Disambiguation**
  - **Setting**: every hyperedge is split into two hyperedges
  - **Goal**: To predict whether given two hyperedges are split from one or not
  - **Application**: researcher disambiguation, user identification
- **Task2: Local Clustering**
  - **Setting**: Each node is involved in one (or more) clusters
  - **Goal**: To predict whether given two nodes belong to the same cluster or not
  - **Application**: sub-field detection, household matching
- **Advantages over formulating these tasks as an entity classification task**:
  1. Does not require the number of labels (split hyperedges or number of clusters).
  2. Does not need to retrain a model whenever the number of labels changes.



Example of Task1



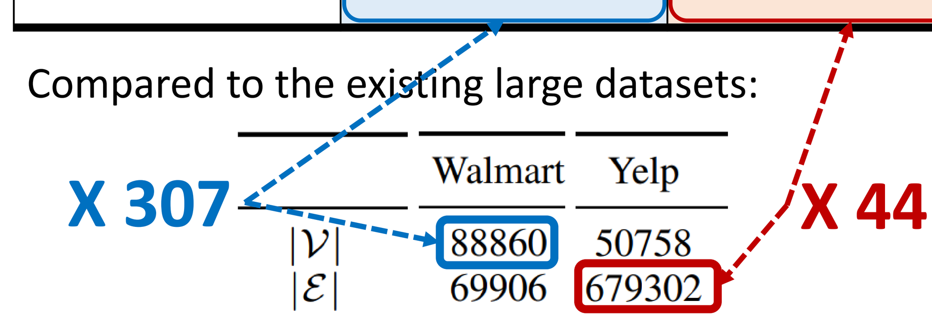
Example of Task2

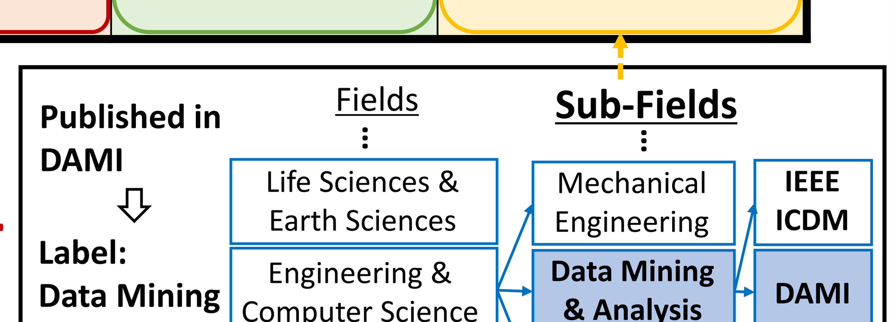
Proposed Datasets

- **Two 10M-scale co-authorship hypergraph datasets: MAG & AMiner**

Embeddings of corpus (title, keywords)

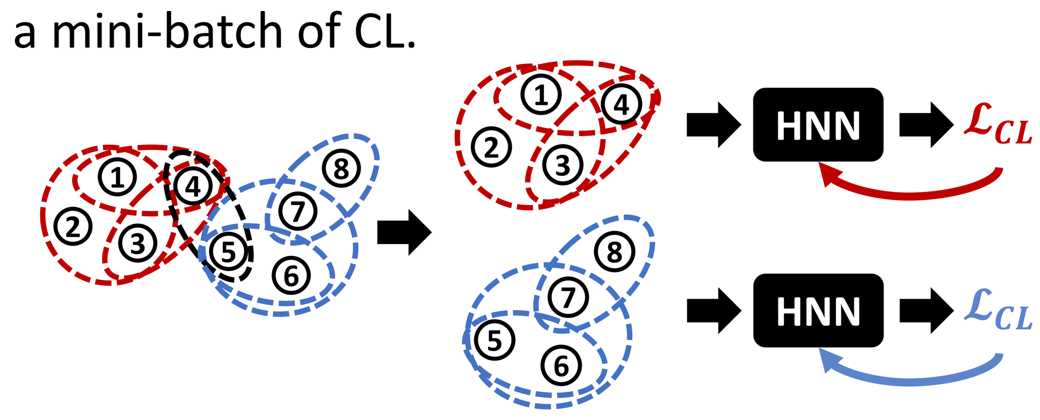
Dataset	# of Nodes (Publications)	# of Hyperedges (Authors)	Node Feature Dimension	# of Classes (Research Field)
AMiner	13,262,573	22,552,647	300	257
MAG	27,320,375	30,175,013	300	247



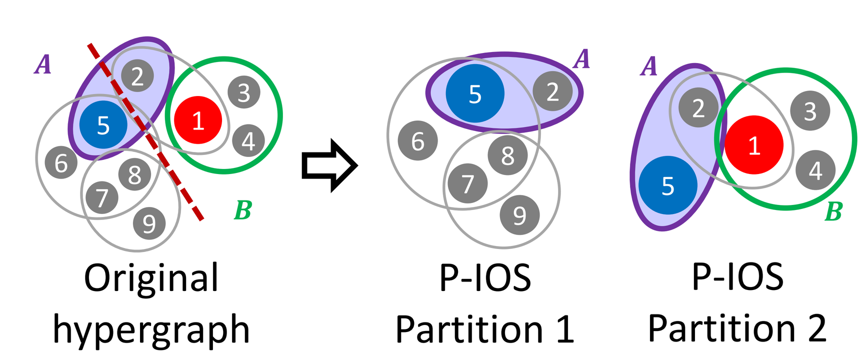


Proposed Learning Method: PCL

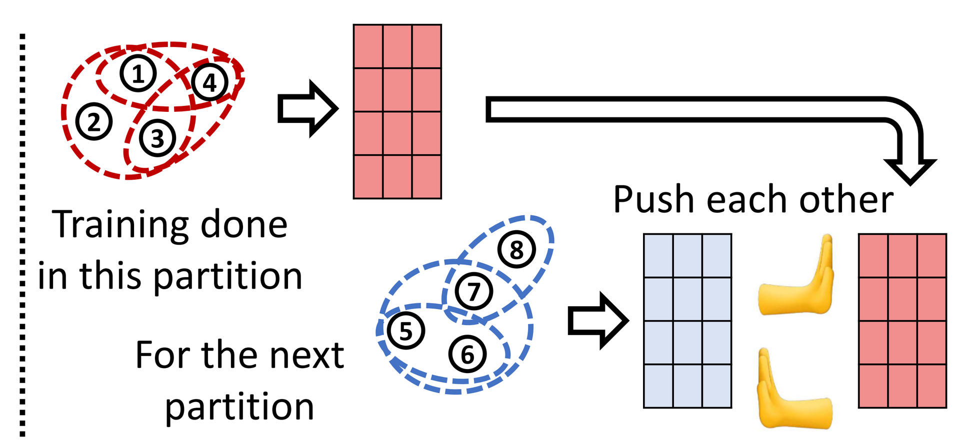
- **Challenge 1:** Large hypergraphs cannot be entirely loaded into GPU memory.
  - **Solution**: we use hypergraph partitioning to divide the large hypergraph into smaller partitions.
- **Challenge 2**: Information loss (e.g., split hyperedges) can be caused by partitioning.
  - **Solution**: we use contrastive learning and propose two additional techniques.
- **We propose PCL (Partitioning-based Contrastive Learning)**
  - We first divide input hypergraphs into several partitions.
  - Then, we train the HNN encoder via contrastive learning, by regarding each partition as a mini-batch of CL.



- **We propose two additional tools for PCL to mitigate information loss.**
  - **P-IOS**: Partitioning-technique that recovers lost topological information.
  - **PINS**: CL-technique that encourages the encoder to learn inter-partition dissimilarity.



Concept of P-IOS



Concept of PINS

Experiments

- **Accuracy of PCL on Task1 and Task2**

bold: best, underline: second-best

Data Type	Methods	DBLP AP	DBLP AUROC	Trivago AP	Trivago AUROC	OGBN-MAG AP	OGBN-MAG AUROC	AMiner AP	AMiner AUROC	MAG AP	MAG AUROC	Avg Rank
Only X	MLP	65.2 ± 3.4	62.2 ± 3.4	60.2 ± 1.6	60.7 ± 2.5	73.1 ± 2.6	71.9 ± 2.8	91.7 ± 0.3	92.9 ± 0.3	91.2 ± 0.6	92.6 ± 0.6	9.8
Full Hypergraph	HGNN	77.3 ± 2.5	80.0 ± 3.1	72.1 ± 5.1	76.6 ± 1.4	91.2 ± 0.7	92.9 ± 0.6	OOM	OOM	OOM	OOM	7.7
	UniGCNII	78.1 ± 3.4	77.1 ± 3.9	71.8 ± 1.5	75.6 ± 2.3	86.9 ± 5.6	88.3 ± 6.7	OOM	OOM	OOM	OOM	8.7
	ALLSET	53.1 ± 1.2	55.8 ± 2.2	53.7 ± 1.2	56.8 ± 2.2	65.3 ± 1.4	73.4 ± 1.6	OOM	OOM	OOM	OOM	13.1
	HCL	**91.0 ± 1.7**	**92.6 ± 2.1**	73.7 ± 0.7	78.8 ± 1.1	**94.2 ± 0.4**	**95.5 ± 0.3**	OOM	OOM	OOM	OOM	4.5
Partitioned Graph	GCN	84.5 ± 1.6	85.9 ± 1.6	65.9 ± 1.9	66.7 ± 2.0	82.3 ± 2.6	83.4 ± 2.8	81.2 ± 0.6	81.5 ± 0.6	OOM	OOM	7.9
	GAT	84.7 ± 1.3	86.4 ± 1.2	62.2 ± 3.1	62.5 ± 4.6	78.0 ± 1.8	79.6 ± 1.8	81.3 ± 0.5	81.0 ± 0.6	OOM	OOM	8.2
	BGRL	85.8 ± 1.7	85.6 ± 1.7	82.4 ± 2.2	83.5 ± 2.1	91.4 ± 0.5	92.8 ± 0.5	90.1 ± 0.8	90.9 ± 0.8	89.7 ± 0.7	90.8 ± 0.9	3.9
	GGD	80.1 ± 2.4	80.8 ± 2.3	76.9 ± 1.6	78.3 ± 1.5	88.7 ± 1.0	90.2 ± 0.9	82.5 ± 1.2	83.6 ± 1.5	80.1 ± 1.3	81.2 ± 1.4	6.2
Partitioned Hypergraph	HGNN	84.1 ± 2.0	86.2 ± 1.8	71.8 ± 0.8	75.8 ± 1.0	85.9 ± 1.1	88.1 ± 1.0	91.3 ± 0.3	92.5 ± 0.3	89.2 ± 0.7	91.6 ± 0.5	5.4
	UniGCNII	79.9 ± 1.8	80.0 ± 1.8	71.5 ± 2.6	74.2 ± 3.5	77.4 ± 0.9	75.7 ± 1.2	91.8 ± 0.5	92.1 ± 0.5	90.4 ± 0.4	91.9 ± 0.1	7.9
	ALLSET	54.2 ± 1.6	57.6 ± 2.7	53.6 ± 1.2	56.7 ± 2.0	59.3 ± 1.6	65.6 ± 2.4	61.5 ± 1.8	68.6 ± 2.5	OOM	OOM	13.2
Partitioned Hypergraph	**Basic PCL (proposed)**	87.1 ± 1.9	88.3 ± 2.0	**88.2 ± 0.9**	**88.9 ± 0.7**	94.1 ± 0.4	95.1 ± 0.3	**95.5 ± 0.7**	**96.0 ± 0.8**	**96.2 ± 0.3**	**96.9 ± 0.3**	**1.4**

Data Type	Methods	DBLP AP	DBLP AUROC	Trivago AP	Trivago AUROC	OGBN-MAG AP	OGBN-MAG AUROC	AMiner AP	AMiner AUROC	MAG AP	MAG AUROC	Avg. Rank
Only X	MLP	52.2 ± 2.3	51.3 ± 3.5	50.6 ± 0.4	50.9 ± 0.7	72.1 ± 1.0	73.9 ± 0.8	70.5 ± 0.9	72.9 ± 0.8	76.2 ± 1.3	79.4 ± 0.9	7.1
Full Hypergraph	HGNN	56.9 ± 5.3	54.9 ± 6.3	54.8 ± 3.7	54.8 ± 3.6	78.2 ± 1.0	80.1 ± 0.7	OOM	OOM	OOM	OOM	6.3
	UniGCNII	55.2 ± 4.6	52.9 ± 4.5	51.6 ± 1.0	51.1 ± 0.9	76.7 ± 1.2	79.0 ± 1.1	OOM	OOM	OOM	OOM	7.8
	ALLSET	54.2 ± 4.1	56.1 ± 5.4	51.0 ± 0.7	51.7 ± 0.7	60.0 ± 2.2	61.8 ± 2.5	OOM	OOM	OOM	OOM	8.6
	HCL	63.6 ± 6.9	61.4 ± 7.9	58.6 ± 2.6	58.8 ± 4.2	**78.5 ± 1.0**	**80.9 ± 0.7**	OOM	OOM	OOM	OOM	4.8
Partitioned Graph	GCN	51.9 ± 0.4	52.2 ± 0.4	50.1 ± 0.0	50.1 ± 0.0	51.6 ± 0.4	51.7 ± 0.4	54.9 ± 0.7	54.3 ± 0.7	OOM	OOM	12.7
	GAT	52.4 ± 0.6	53.1 ± 0.6	50.4 ± 0.7	50.5 ± 1.0	54.1 ± 0.7	54.3 ± 0.8	55.2 ± 0.4	54.6 ± 0.4	OOM	OOM	10.3
	BGRL	58.8 ± 7.7	56.8 ± 5.6	**61.2 ± 3.9**	**61.4 ± 3.9**	65.6 ± 1.0	66.4 ± 0.8	65.8 ± 0.5	67.1 ± 0.6	71.0 ± 1.0	72.7 ± 0.9	3.6
	GGD	**67.7 ± 12.7**	**67.2 ± 13.3**	59.4 ± 3.8	59.4 ± 3.5	64.8 ± 1.0	65.5 ± 0.6	62.8 ± 0.8	64.3 ± 0.8	68.5 ± 0.9	70.5 ± 0.9	3.8
Partitioned Hypergraph	HGNN	55.2 ± 6.5	55.2 ± 7.3	52.0 ± 1.0	52.4 ± 2.0	68.8 ± 1.3	69.9 ± 1.6	57.6 ± 1.6	57.5 ± 2.1	62.8 ± 3.0	62.5 ± 3.6	6.6
	UniGCNII	52.6 ± 1.7	52.3 ± 0.9	50.4 ± 0.3	50.1 ± 0.2	54.9 ± 0.6	54.6 ± 0.5	59.9 ± 1.2	59.6 ± 1.3	65.9 ± 2.2	66.0 ± 2.2	9.8
	ALLSET	53.2 ± 1.6	54.6 ± 2.1	51.5 ± 0.3	52.5 ± 0.5	57.5 ± 1.9	59.2 ± 2.8	58.1 ± 1.1	60.9 ± 1.4	OOM	OOM	8.2
Partitioned Hypergraph	**PCL+PINS (proposed)**	63.2 ± 10.6	64.0 ± 11.5	58.6 ± 1.0	59.2 ± 1.3	77.6 ± 0.8	79.8 ± 0.6	**80.2 ± 1.7**	**81.6 ± 0.8**	**83.1 ± 1.4**	**86.1 ± 1.5**	2.1

- **Effectiveness of P-IOS**

Task	Data	Metric	PCL w/o P-IOS	PCL+P-IOS
Task-I	DBLP	AP	87.1 ± 1.9	**90.8 ± 1.8**
		AUROC	88.3 ± 2.0	**92.9 ± 1.2**
	Trivago	AP	88.2 ± 0.9	**89.4 ± 1.2**
		AUROC	88.9 ± 0.7	**89.5 ± 0.9**
	OGBN-MAG	AP	94.1 ± 0.4	**94.8 ± 0.4**
		AUROC	95.1 ± 0.2	**96.2 ± 0.2**
Task-II	DBLP	AP	62.4 ± 11.5	**64.7 ± 11.5**
		AUROC	61.1 ± 9.9	**62.3 ± 12.3**
	Trivago	AP	58.6 ± 1.0	**59.2 ± 1.0**
		AUROC	58.7 ± 1.0	**59.3 ± 1.0**
	OGBN-MAG	AP	77.3 ± 0.7	**78.8 ± 0.9**
		AUROC	79.6 ± 0.3	**80.9 ± 0.7**

- **Effectiveness of PINS on Task2**

Data	Metric	PCL w/o PINS	PCL+PINS
DBLP	AP	62.4 ± 11.5	**63.2 ± 10.6**
	AUROC	61.1 ± 9.9	**64.0 ± 11.5**
Trivago	AP	**58.6 ± 1.0**	**58.6 ± 1.0**
	AUROC	58.7 ± 1.0	**59.2 ± 1.3**
OGBN-MAG	AP	77.3 ± 0.7	**77.6 ± 0.8**
	AUROC	79.6 ± 0.3	**79.8 ± 0.6**
AMiner	AP	78.6 ± 1.3	**80.2 ± 1.7**
	AUROC	81.3 ± 1.2	**81.6 ± 0.8**
MAG	AP	**83.3 ± 1.1**	83.1 ± 1.4
	AUROC	**86.3 ± 0.7**	86.1 ± 1.5

- **Time & space req. of PINS**

	Method	OGBN-MAG	AMiner	MAG
Average Running Time Per Epoch (Sec)	PCL w/o PINS	20.102	89.505	103.450
	PCL+PINS	24.025	118.338	133.739
Average GPU Memory Usage (GB)	PCL w/o PINS	2.344	10.573	23.942
	PCL+PINS	2.347	10.575	23.945

32% more

0.1% more