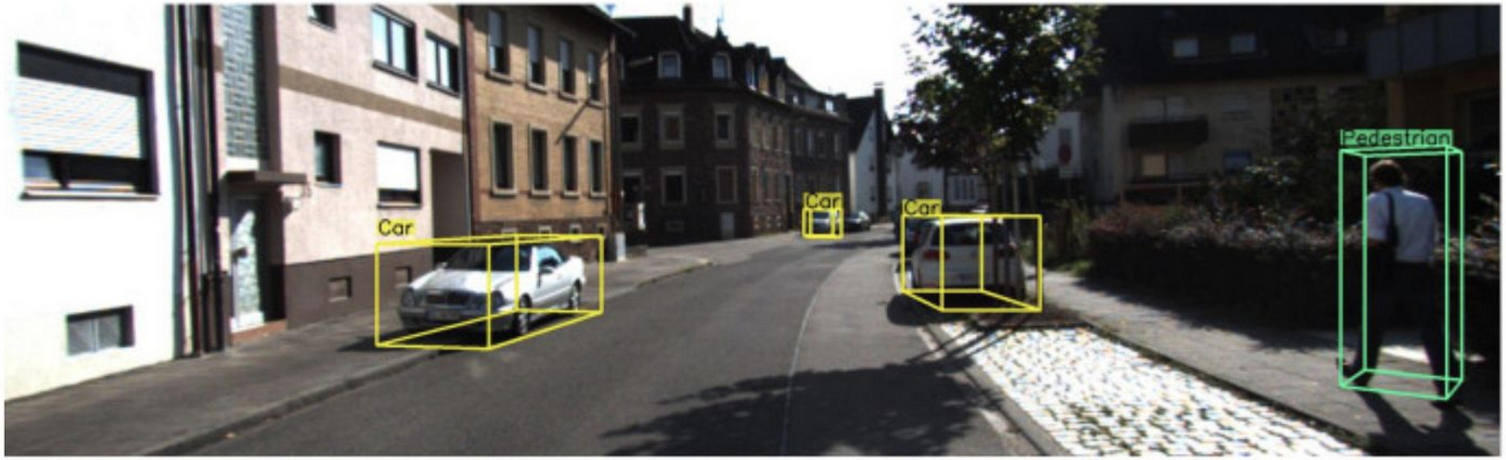
STONE: A Submodular Optimization Framework for Active 3D Object Detection

Ruiyu Mao, Sarthak Kumar Maharana ,Rishabh K Iyer, Yunhui Guo

UT Dallas
Erik Jonsson School of Engineering & Computer Science

Introduction

- **3D object detection is important** [1, 2]
  - autonomous driving
  - robotics
  - VR/ AR application
- **Challenges** [3, 4]
  - a dataset of significant scale
    - high cost of manual annotation

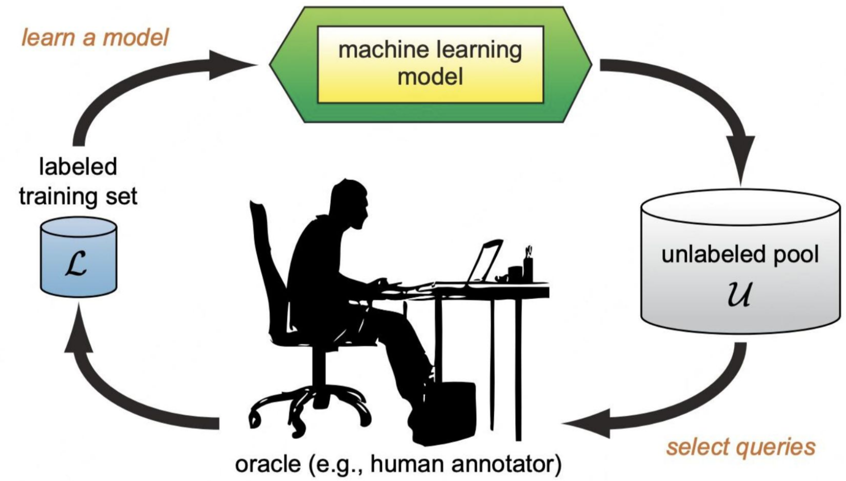


Introduction to Active Learning

- **Active learning** [5]
  - A technique in machine learning where the learning algorithm selectively queries the most informative instances from the unlabeled data pool to have them labeled by an expert.
- **Advantage** [6]
  - This technique helps to reduce the **labeling cost** and enhance the accuracy of the model with a smaller amount of labeled data.

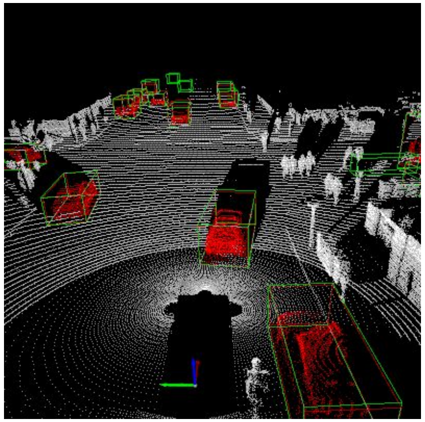
Active learning methods:

- Random
- Uncertainty [7]
- CORESET [8]
- BADGE [9]



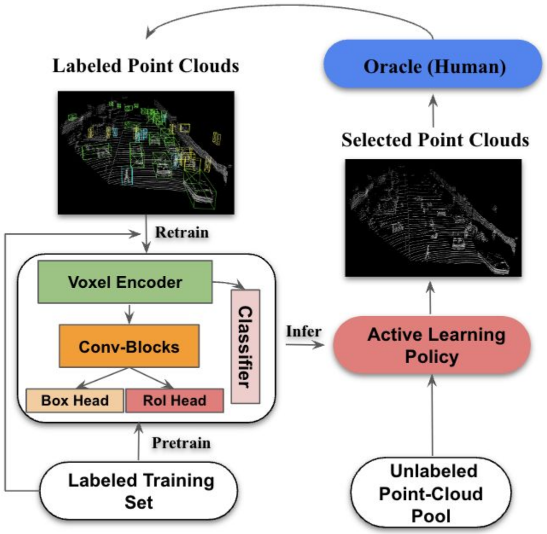
Background

- **3D Object Detection**
  - **Input**: Point Clouds (*LiDAR sensors*)
    - $P_i = \{(x, y, z, r)\}$
  - 3D object detector
  - **Output**:
    - Predicted Bounding box:
      - $\{b'_i\}_{i=1}^{N_i}$ spatial coordinates; box size; heading angle
    - Predicted Semantic Labels:
      - $\{c'_i\}_{i=1}^{N_i}$ $c'_i \in \{1, 2, \ldots, C\}$
        - N_i represents the number of bounding box in the i-th point cloud



Background

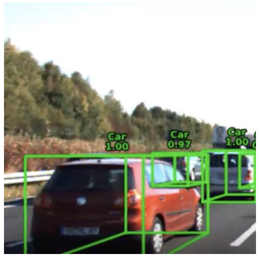
- **Active Learning for 3D Object Detection**
  - **Step 1**: Labeled small number of point clouds D_L randomly selected from D_U
    - Initialize backbone 3D object detection model.
  - **Step 2 (Retrain)**: Query Iteration $q \in \{1, 2, \ldots, Q\}$
    - Active learning method
      - select Γ number of unlabeled point clouds from D_U
    - Human annotator for labeling
    - New selected point clouds training set:
      - $D_L = D_L \cup D_S$
  - **Stop Active Learning**:
    - query round Q is reached.
    - Budget: N_Q number of bounding box is used.



Background

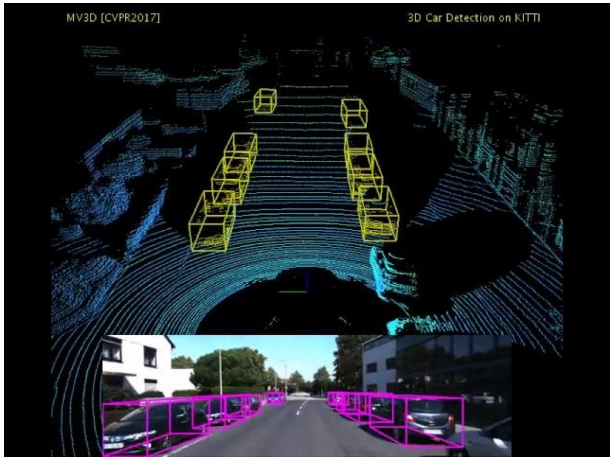
- **Submodular Function and optimization**
  - A set function *f*, in discrete $f\colon 2^D \to \mathbb{R}$
    - $f(A) + f(B) \geq f(A \cup B) + f(A \cap B)$
      - $\forall A, B \subseteq D$
    - $f(A \cup \{x\}) - f(A) \geq f(B \cup \{x\}) - f(B)$
      - $\forall A, B \subseteq D$, $A \subseteq B$ and $x \notin B$
  - Property of diminishing return
  - *f* is strictly monotone if $f(A) < f(B)$ for $A \subseteq B$
  - Why choose Submodular Function:
    - Example: Shannon Entropy [10]
    - *f* is monotone submodular
      - NP-complete greedy approximation algorithm [11]
      - Faster than k-medoids approach

Challenges:

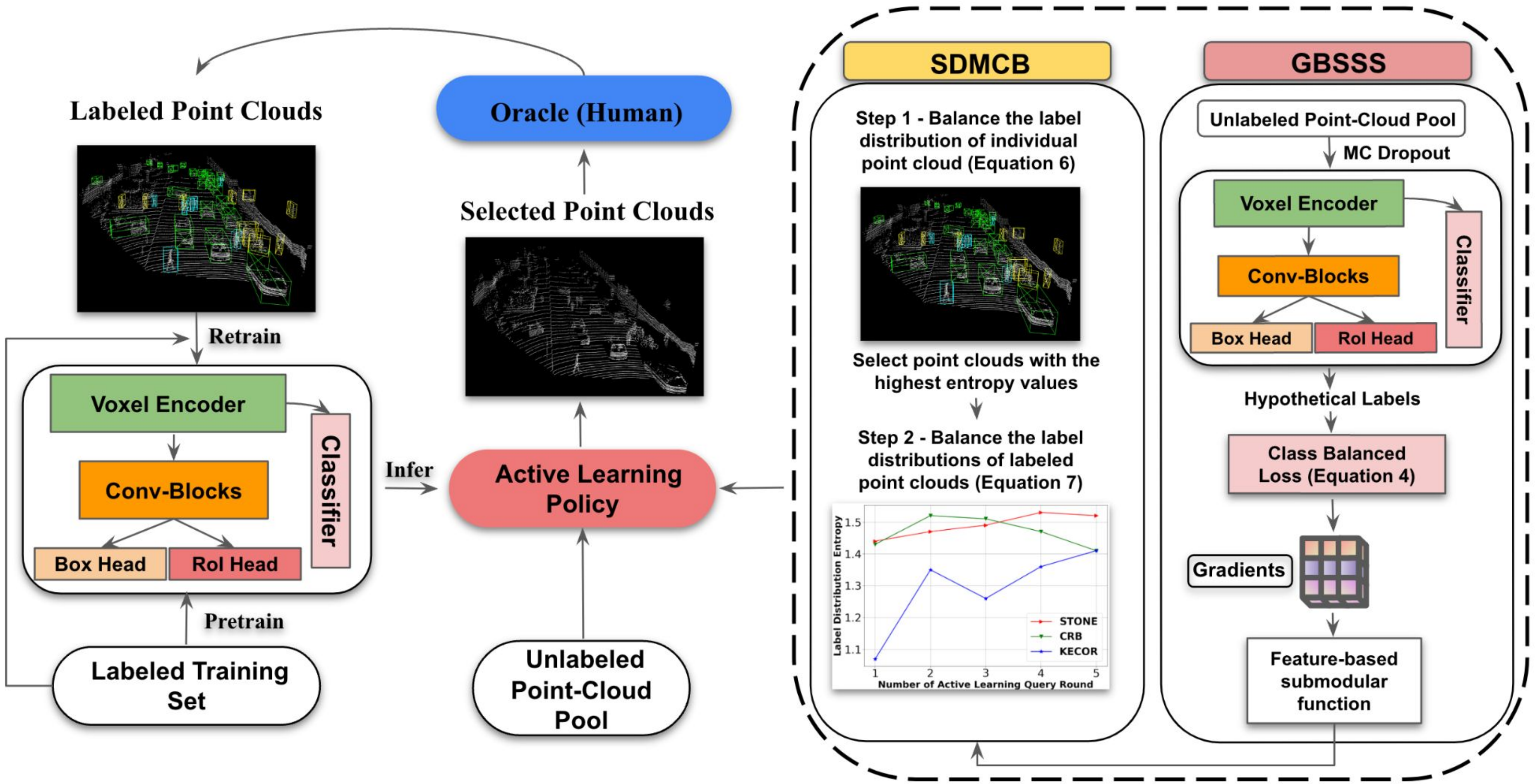


- **Various Difficulty levels**
  - Size, Occlusion Level, and Truncation of 3D Objects:
    - Eazy, Moderate, and Hard
    - Selected labeled point cloud data should include various difficulty levels.

- **Imbalanced Data**:
  - Each 3D scene can contain multiple objects, leading to highly imbalanced label distributions in the certain point cloud
    - Include cars, but not cyclists or pedestrians



STONE: An illustrative pipeline



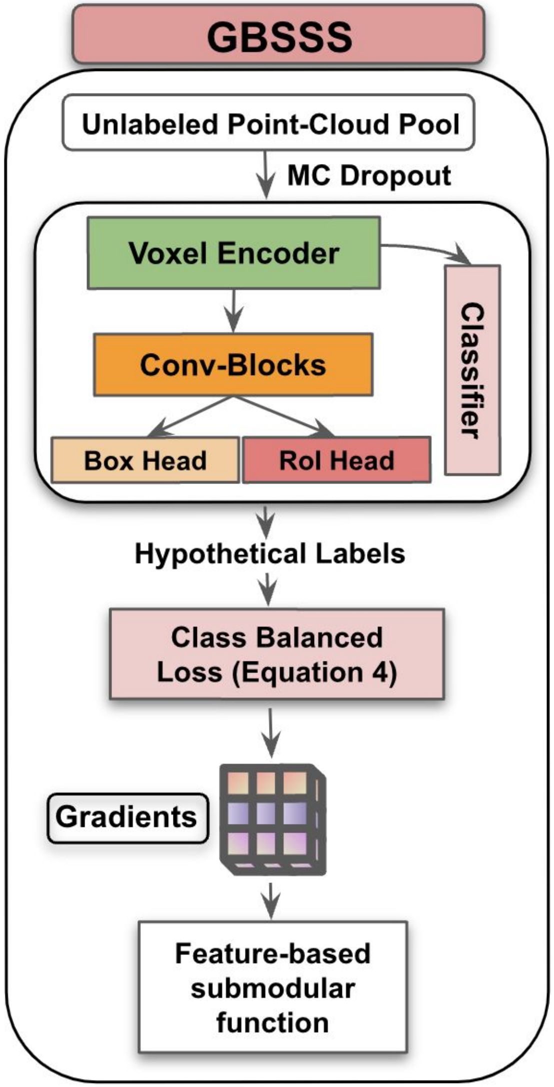
STONE Algorithm Overview

- **Revisiting Challenges**
  - Various Difficulty levels
    - representative and inclusive of various difficulty level (sample uncertainty [10])
      - Min absolute difference between $f_1(D_U)$ and $f_1(D_S)$
        - Since $D_S \subset D_U$
          - $\max_{D_S \subset D_U} [f_1(D_S) - f_1(D_U)]$
  - Imbalanced Data (clustering-based [12, 13, 14])
    - Label distribution balancing
      - selected unlabeled point clouds D_S are added to the labeled set D_L
        - label distribution quality not decrease

$$\max_{D_S \subset D_U} [f_1(D_S) - f_1(D_U)] + [f_2(D_L) - f_2(D_L \cup D_S)]$$

STONE:
Stage 1: Gradient-Based Submodular Subset Selection

- **GBSSS (Part 1)**
  - **Step 1**: Input D_U into backbone model
    - MC-dropout [15] at detector head for each point cloud P_i
      - Multiple regression and classification prediction [16]
        - Hypothetical label by average
        - Loss using hypothetical label as true label
          - backpropagation
            - gradients $\nabla_{\boldsymbol{\theta}} \mathcal{L}_i$ from FC layer
  - **Step 2**: After compute the gradient for each point cloud
    - Feature-based submodular function
      - $\max_{D_S \subset D_U, |D_S| = \Gamma_1} \sum_{P_i \in D_S} g\left(\mu(\nabla_\theta \hat{L}_i)\right)$
      - $g(x) = \log(1+x)$ concavity
      - $\mu(\cdot)$ is $H(\nabla_{\boldsymbol{\theta}} \mathcal{L})$ informativeness [17]



STONE:
Stage 1: Gradient-Based Submodular Subset Selection

- **GBSSS (Part 2)**
  - **Dataset highly imbalanced**
    - Gradients become biased and inaccurate (fewer classes) [23]
  - **Loss Re-weighing Module**
    - Regression Loss
      - Re-weighing Factor $w_c = \frac{1}{n_c}$ $\tilde{w}_c = \frac{w_c}{\max(w_c)}$
        - n_c number of bounding box of class C
      - $\hat{L}_{reg} = \frac{1}{C} \sum_{c=1}^{C} \tilde{w}_c \cdot L_{reg}^{c}$
    - Classification Loss
      - Rare classes penalize its distance to decision boundary (margin)
        - Margin vector $m_{i,c} = \frac{1}{\sqrt{n_c}}$
      - $\hat{L}_{cls} = L_{cls}(\hat{y}_i, f_i - m_i)$

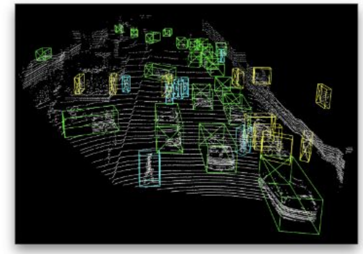
$$\hat{L} = \hat{L}_{reg} + \hat{L}_{cls}$$

STONE:
Stage 2: Submodular Diversity Maximization for Class Balancing

- **SDMCB**
  - **Multiple Semantic Classes Single 3D scene**
    - CRB [16], KECOR [18]
      - Partially Solve it
  - **Step 1**: Balance individual point cloud
    - $$H(P_i) = -\sum_{c=1}^{C} \boldsymbol{p}_{i,c} \log \boldsymbol{p}_{i,c}, \quad \boldsymbol{p}_{i,c} = \frac{e^{n_c\,/\,N_i}}{\sum_{c=1}^{C} e^{n_c/\,N_i}}$$
  - **Step 2**: Balance labeled point clouds
    - $$H(P_i) = -\sum_{c=1}^{C} \boldsymbol{p}_{i,c} \log \boldsymbol{p}_{i,c}, \quad \boldsymbol{p}_{i,c} = \frac{e^{n_c\,/\,N_i}}{\sum_{c=1}^{C} e^{n_c/\,N_i}}$$

SDMCB

Step 1 - Balance the label distribution of individual point cloud (Equation 6)



Select point clouds with the highest entropy values

Step 2 - Balance the label distributions of labeled point clouds (Equation 7)

Experimental Setup

- **Datasets**
  - KITTI Dataset [19]
    - 80, 256 labeled objects
    - cars, pedestrians, and cyclists
  - Waymo Open Dataset
    - 158, 361 training samples
    - 40, 077 testing samples
    - 2 difficulty levels
- **Baseline**
  - Generic
    - Random, Entropy, COREST
  - SOTA
    - CRB
    - KECOR
- **Evaluation Metrics**
  - Average Precision (AP)
  - Bird Eye View (BEV)

Experiment Results

- **KITTI Dataset** [19]
  - 3D AP(%) scores with 1% queried bounding boxes
  - PV-RCNN [21] as the backbone
- **CRB**
  - Eazy: **1.47%**, MODERATE: **0.84%**, Hard: **1.24%**
- **KECOR**
  - Eazy: **0.54%**, MODERATE: **0.08%**, Hard: **0.63%**

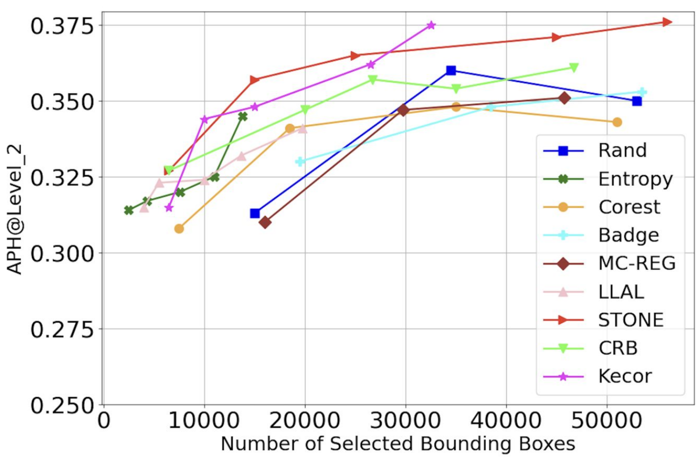
	CAR			Pedestrian			Cyclist			Average		
Method	EASY	MOD.	HARD	EASY	MOD.	HARD	EASY	MOD.	HARD	EASY	MOD.	HARD
CORESET	87.77	77.73	72.95	47.27	41.97	38.19	81.73	59.72	55.64	72.26	59.81	55.59
BADGE	89.96	75.78	70.54	51.94	46.24	40.98	84.11	62.29	58.12	75.34	61.44	65.55
LLAL	89.95	78.65	75.32	56.34	49.87	45.97	75.55	60.35	55.36	73.94	62.95	58.88
MC-REG	88.85	76.21	73.47	35.82	31.81	29.79	73.98	55.23	51.85	66.21	54.41	51.70
MC-MI	86.28	75.58	71.56	41.05	37.50	33.83	86.26	60.22	56.04	71.19	57.77	53.81
CONSENSUS	90.14	78.01	74.28	56.43	49.50	44.80	78.46	55.77	53.73	75.01	61.09	57.60
LT/C	88.73	78.12	73.87	55.17	48.37	43.63	83.72	63.21	59.16	75.88	63.23	58.89
CRB	90.98	79.02	74.04	64.17	54.80	50.82	86.96	67.45	63.56	80.70	67.81	62.81
KECOR	91.71	79.56	74.05	65.37	57.33	51.56	87.80	**69.13**	**64.65**	81.63	68.67	63.42
STONE	**92.09**	**80.27**	**75.44**	**66.1**	**58.84**	**52.70**	**88.31**	67.14	64.01	**82.17**	**68.75**	**64.05**

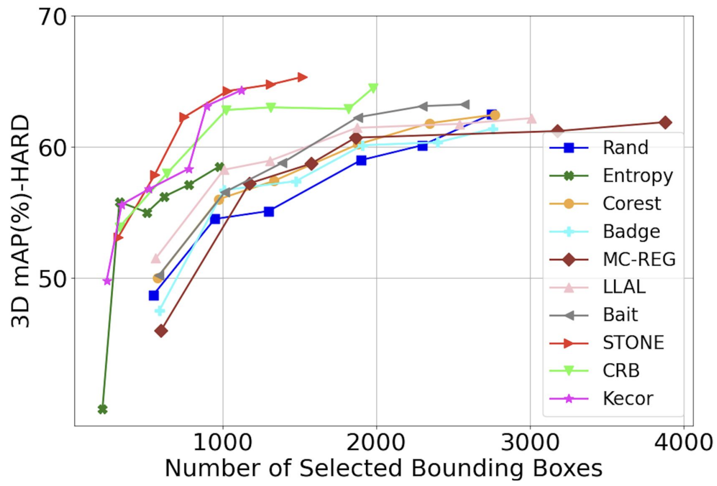
Experiment Results

- **KITTI Dataset**
  - 3D AP(%) scores with 1% queried bounding boxes
  - SECOND [21] as the backbone (one stage) good generalization ability
- **KECOR**
  - 3D Detection Hard: **3.4%** mAP
  - BEV Detection Hard: **2.43%** mAP

	3D Detection average mAP			BEV Detection average mAP		
Method	EASY	MOD.	HARD	EASY	MOD.	HARD
Random	66.33	55.48	51.53	75.66	63.77	59.71
CORESET	66.86	53.22	48.97	73.08	61.03	56.95
LLAL	69.19	55.38	50.85	76.52	63.25	59.07
BADGE	69.92	55.60	51.23	76.07	63.39	59.47
BAIT	69.45	55.61	51.25	76.04	63.49	53.40
CRB	72.33	58.06	53.09	78.84	65.82	61.25
KECOR	74.05	60.38	55.34	80.00	68.20	63.20
STONE	**76.86**	**64.04**	**58.75**	**82.14**	**70.82**	**65.68**

Experiment Results





- **Waymo Open Dataset** [20] **(left)** **KITTI Dataset** [19] **(right)**
  - Regardless of the detection difficulty level
  - STONE consistently surpasses other baseline methods

Citation

- [1] Deng, B., Qi, C.R., Najibi, M., Funkhouser, T., Zhou, Y., Anguelov, D.: Revisiting 3d object detection from an egocentric perspective. Advances in Neural Information Processing Systems 34, 26066–26079 (2021)
- [2] Wang, J., Lan, S., Gao, M., Davis, L.S.: Infofocus: 3d object detection for autonomous driving with dynamic information modeling. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16. pp. 405–420. Springer (2020)
- [3] Song, S., Lichtenberg, S.P., Xiao, J.: Sun rgb-d: A rgb-d scene understanding benchmark suite. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 567–576 (2015)
- [4] Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., et al.: Scalability in perception for autonomous driving: Waymo open dataset. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2446–2454 (2020)
- [5] Dasgupta, S.: Two faces of active learning. Theoretical computer science 412(19), 1767–1781 (2011)
- [6] Settles, B.: Active learning literature survey (2009)
- [7] Sharma, Manali, and Mustafa Bilgic. "Evidence-based uncertainty sampling for active learning." Data Mining and Knowledge Discovery 31 (2017): 164-202.
- [8] Sener, O., Savarese, S.: Active learning for convolutional neural networks: A core-set approach. arXiv preprint arXiv:1708.00489 (2017)
- [9] Ash, J.T., Zhang, C., Krishnamurthy, A., Langford, J., Agarwal, A.: Deep batch active learning by diverse, uncertain gradient lower bounds. arXiv preprint arXiv:1906.03671 (2019)
- [10] Shannon, C.E.: A mathematical theory of communication. ACM SIGMOBILE mobile computing and communications review 5(1), 3–55 (2001)
- [11] Feige, U.: A threshold of ln n for approximating set cover. Journal of the ACM (JACM) 45(4), 634–652 (1998)

Citation

- [12] Nguyen, H.T., Smeulders, A.: Active learning using pre-clustering. In: Proceedings of the twenty-first international conference on Machine learning. p. 79 (2004)
- [13] Wang, M., Min, F., Zhang, Z.H., Wu, Y.X.: Active learning through density clustering. Expert systems with applications 85, 305–317 (2017)
- [14] Bodó, Z., Minier, Z., Csató, L.: Active learning with clustering. In: Active Learning and Experimental Design workshop In conjunction with AISTATS 2010. pp. 127–139. JMLR Workshop and Conference Proceedings (2011)
- [15] Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: international conference on machine learning. pp. 1050–1059. PMLR (2016)
- [16] Luo, Y., Chen, Z., Wang, Z., Yu, X., Huang, Z., Baktashmotlagh, M.: Exploring active 3d object detection from a generalization perspective. In: The Eleventh International Conference on Learning Representations (2023)
- [17] Wei, K., Liu, Y., Kirchhoff, K., Bartels, C., Bilmes, J.: Submodular subset selection for largescale speech training data. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 3311–3315. IEEE (2014)
- [18] Luo, Y., Chen, Z., Fang, Z., Zhang, Z., Baktashmotlagh, M., Huang, Z.: Kecor: Kernel coding rate maximization for active 3d object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18279–18290 (2023)
- [19] Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 3354–3361. IEEE (2012)
- [20] Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., et al.: Scalability in perception for autonomous driving: Waymo open dataset. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2446–2454 (2020)
- [21] Shi, S., Guo, C., Jiang, L., Wang, Z., Shi, J., Wang, X., Li, H.: Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10529–10538 (2020)
- [22] Yan, Y., Mao, Y., Li, B.: Second: Sparsely embedded convolutional detection. Sensors 18(10), 3337 (2018)

Thank You

- Thank you for your time and attention during this presentation. We hope you found it informative and engaging!