The 30th International Conference on
Technologies and Applications of Artificial Intelligence

December 13-14, 2025

Synergistic Pseudo-Labeling: Harmonizing Heterogeneous Datasets with a Foundation Model

Jing-Qiao Chen[1], Cheng-Hsueh Hu[1], Kai-Jun Liang[1]
Chien-Yao Wang[2], Yi-Ting Huang[1]
1.Department of Electrical Engineering, National Taiwan University of Science and Technology, Taipei, Taiwan
2.Institute of Information Science, Academia Sinica, Taipei, Taiwan

Paper ID: 170

Outline

- Introduction
- Related Work
- Background
- Method
- Experimental Methodology
- Conclusion
- Reference

Introduction

Challenge - Introduction

- Challenge 1: High Annotation Cost
  - Semantic segmentation relies heavily on large-scale, pixel-level labeled data.
  - Mitigation: Using existing heterogeneous datasets.

- Challenge 2: Heterogeneous Label Spaces
  - Root Cause: Inconsistent class taxonomies and varying levels of granularity.
  - Our Solution: Proposing the **Integrated Pseudo-Labeling (*IPL*) Pipeline**.

Main Contribution - Introduction

- **New Framework:** Proposed a new, effective framework for automating pixel-level labeling and addressing data scarcity with minimal manual effort.

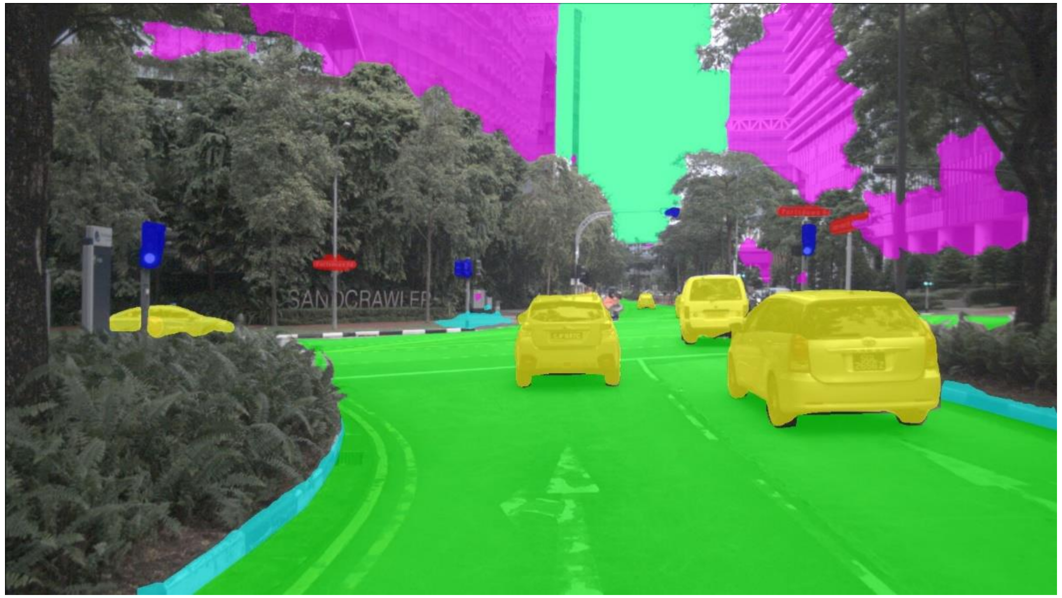
- **Intelligent Aggregation:** Designed a new weighted voting mechanism that leverages both class-specific expertise and general model reliability.

- **Preservation Strategy:** Introduced a rule-based integration strategy that preserves the quality of original ground-truth labels.

nuImages ground truth

EOV-Seg Prediction

OpenSeed Prediction



Our Full Method

Related Work

Related Work

- Pseudo-Labeling
    - Focuses on adaptation within a **unified label taxonomy**.[7]

- Traditional Ensemble Methods
    - Relies on **simple majority voting**; fails to capture models' **specific class expertise**. [4]

- Multi-Dataset Learning
    - A central challenge in this area is **resolving conflicts in datasets** with heterogeneous label spaces. (e.g., Cityscapes: person; nuImages: adult, child, …) [3][2]

Related Work

- Label Harmonization Framework
  - **Explicitly resolves heterogeneous label conflicts** across multiple datasets.

- Performance-Aware Weighted Voting
  - Intelligently balances **"Specific Confidence"** with **"General Reliability"**.

- Model-Agnostic Post-Processing
  - **High efficiency and low cost**, integrating predictions directly via rule-based policies.

Background

Datasets - Background

- Datasets Utilized
  - We involve mainstream semantic segmentation datasets, notably nuImages and BDD100K. [10]

- Key Relationship
  - Datasets like Cityscapes, KITTI, and BDD100K share a common set of 19 evaluation classes. [3][5][10]

- R_{manual} (Class Mapping Rules)
  - We define these rules to formalize the parent-child and conceptual overlap relationships.

Datasets - Background

Cityscapes	nuImages	Relationship
person	human.pedestrian.adult human.pedestrian.child human.pedestrian.construction_worker human.pedestrian.personal_mobility human.pedestrian.police_officer human.pedestrian.stroller human.pedestrian.wheelchair	Parent-Child (nuImages is more granular)
rider	(No direct equivalent)	Unique to Cityscapes
car	vehicle.car vehicle.construction vehicle.emergency.ambulance vehicle.emergency.police	Parent-Child (nuImages is more granular)
bus	vehicle.bus.rigid, vehicle.bus.bendy	Parent-Child
truck	vehicle.truck	Direct Mapping
trailer	vehicle.trailer	Direct Mapping
caravan	(No direct equivalent)	Unique to Cityscapes
motorcycle	vehicle.motorcycle	Direct Mapping
bicycle	vehicle.bicycle	Direct Mapping
train	(No direct equivalent)	Unique to Cityscapes

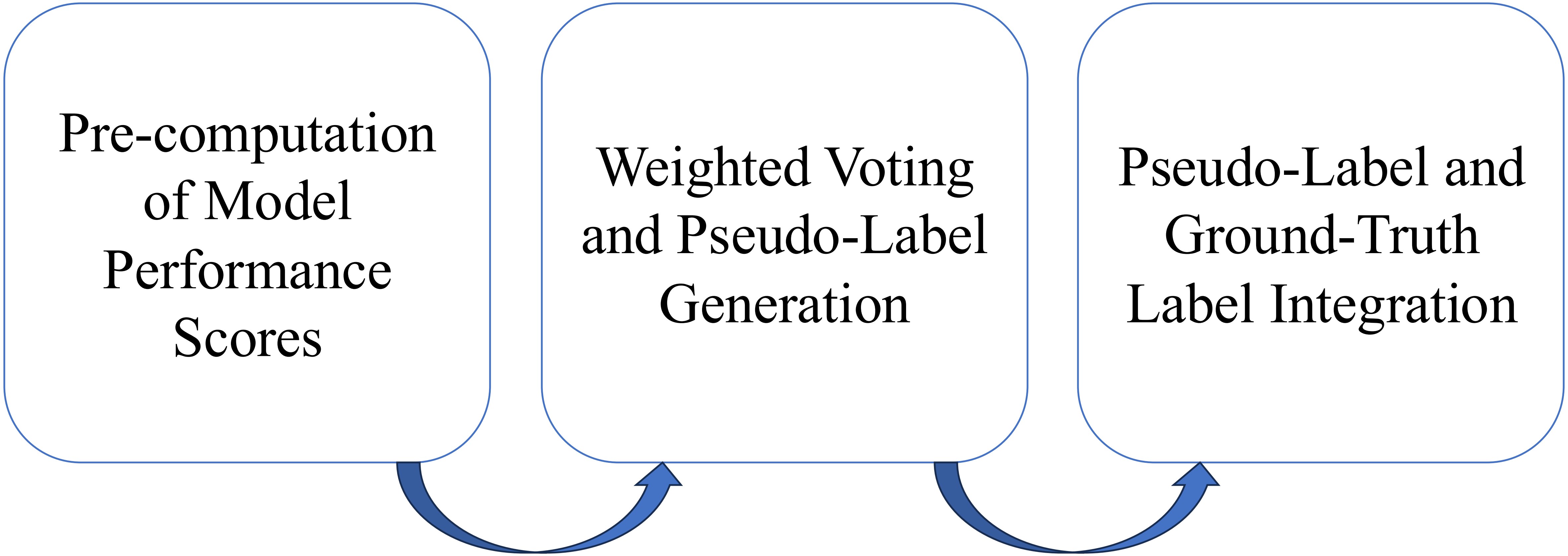
Foundation Models - Background

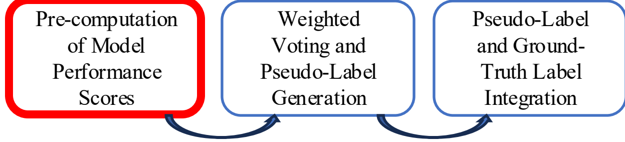
- **OpenSeed**
  - It offers strong generalization for open-vocabulary tasks.
  - It is suitable for multi-dataset integration.

- **EOV-Seg**
  - It provides high efficiency with reduced computational cost.
  - It is ideal for fast, high-quality pseudo-label generation in constrained-category settings

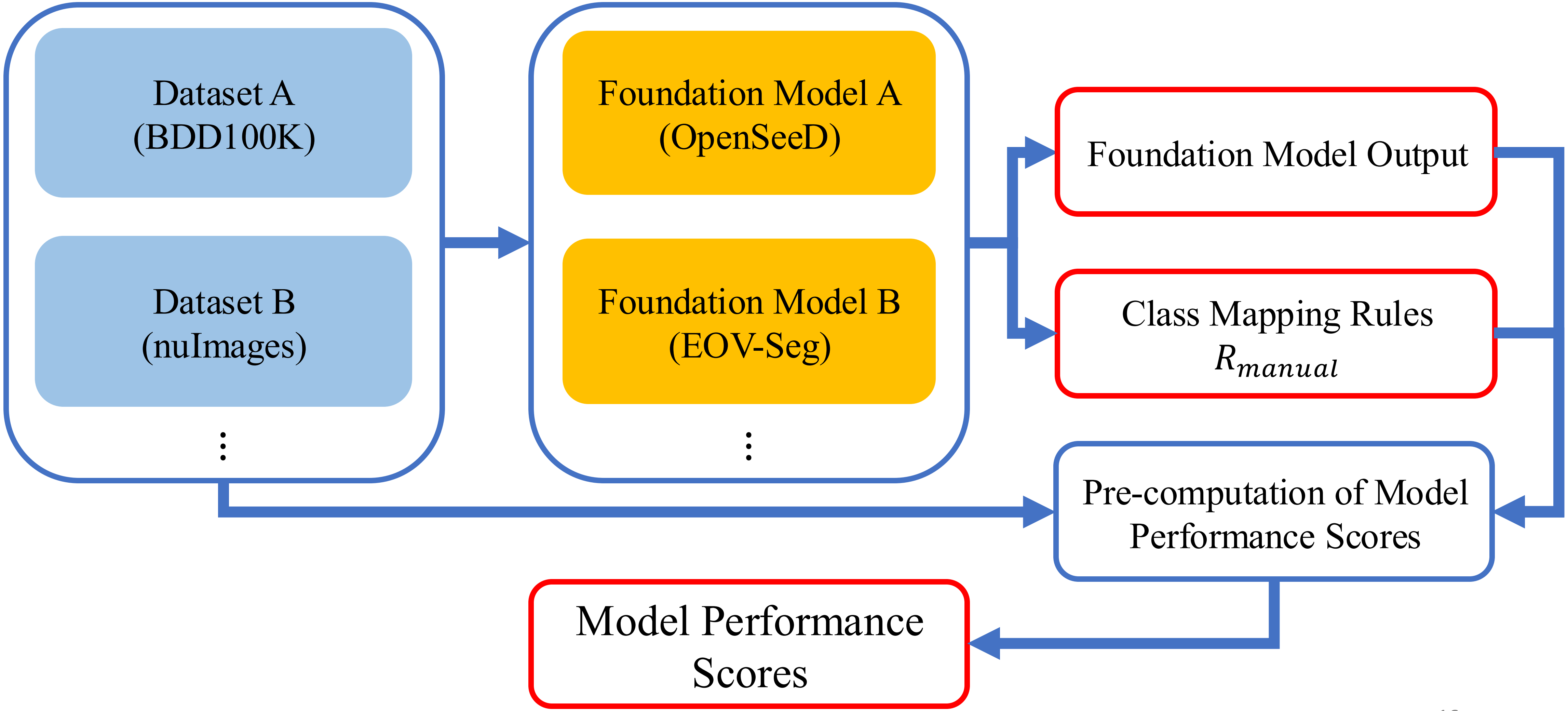
Method

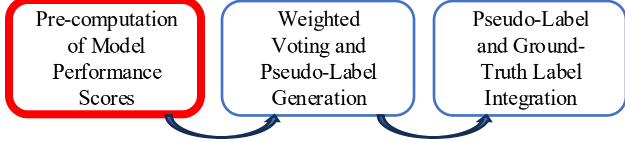
Integrated Pseudo-Labeling (IPL) Pipeline





Pre-computation - Implementation



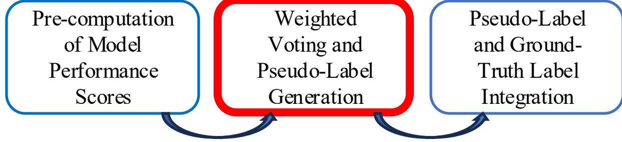


Pre-computation of Model Performance Scores

- Goal:
  - Establish a Performance-Aware weighting basis for the subsequent voting scheme.

- Class Variance:
  - A single model exhibits varying performance across different semantic classes (e.g., better at detecting car than tree).

- Dataset Variance:
  - The same class may different reliability scores when predicted by models trained on heterogeneous datasets.



Weighted Voting and Pseudo-Label Generation

$$W_{Model,specific} = F1_{specific} + F1_{avg}$$

$F1_{specific}$

Define: The F1-score of the specific class currently being voted on in another dataset.

Function: Rewards the model for its domain expertise.

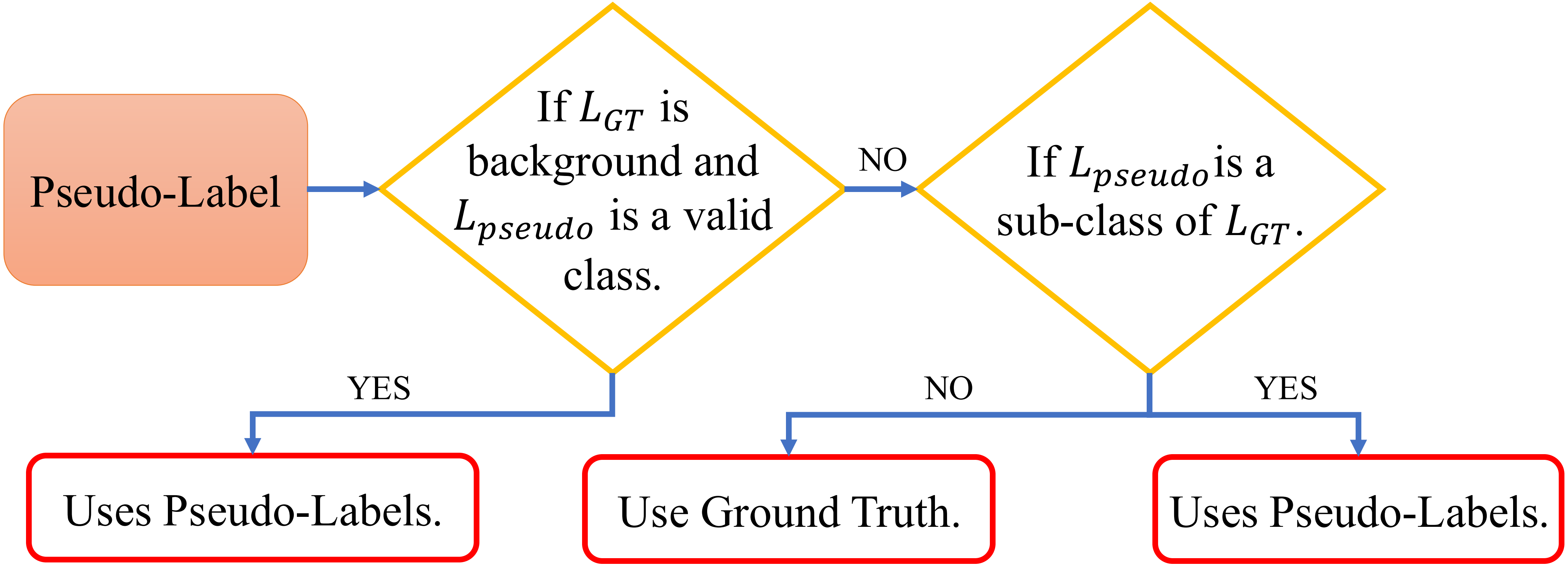
$F1_{avg}$

Define: The average F1-score of the shared class in this dataset.

Function: Ensures stable baseline quality.

Synergistic Balance: This way synergistically balances the model's overall stability with its specific competence, resulting in higher quality pseudo-labels .

Pseudo-Label and Ground-Truth Label Integration



To expand and **fill in missing annotations** in the original data.

Ensures the **integrity** of the **original high-quality** labels is maintained.

Uses pre-defined R_{manual} to **retain finer granularity** in the final label map.

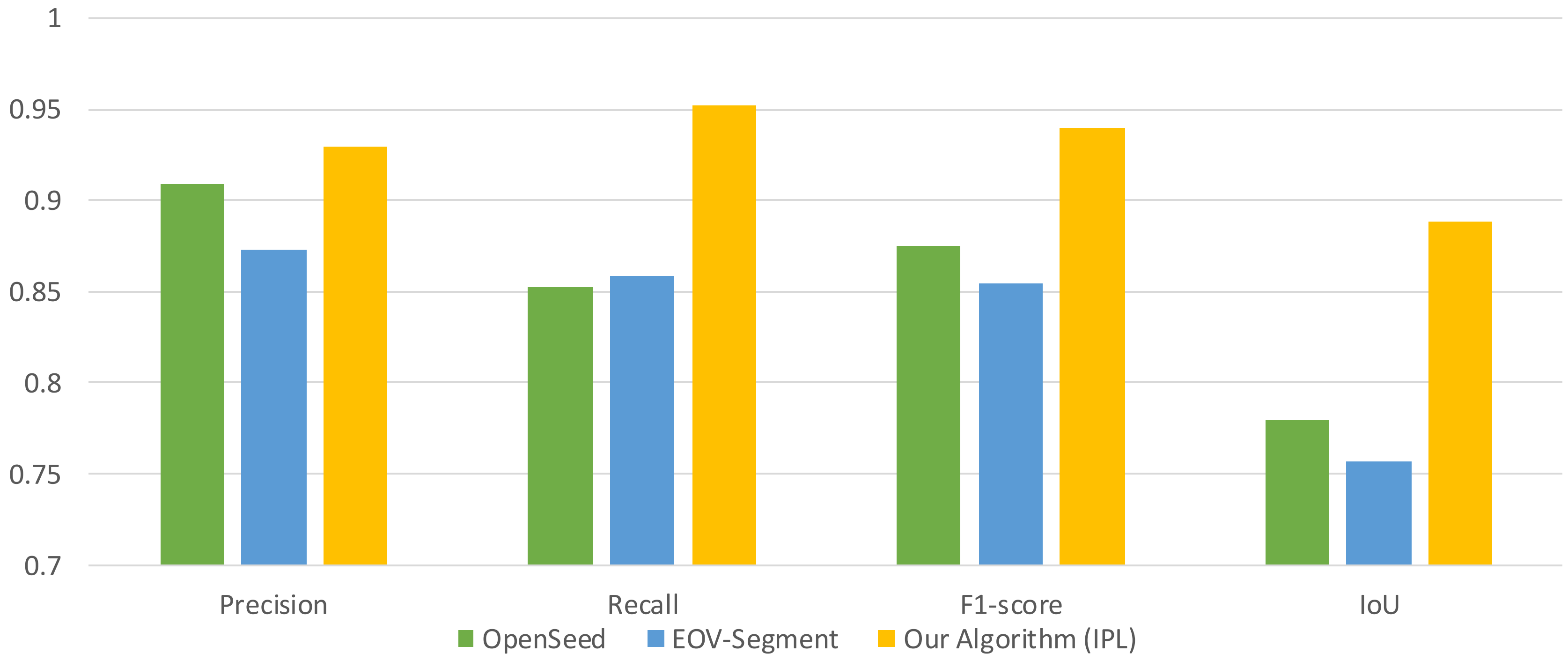
Experimental Methodology

Experimental Methodology

- Goal: Complement the categories of nuImages to align them with other mainstream datasets.
- Assume that the nuImages dataset lacks annotations for the three categories, "bicycle," "car," and "bus".
- Use our algorithm to generate pseudo-labels for these missing categories.

Experimental Results

	OpenSeeD				EOV-Segment				Our Algorithm (IPL)			
Category	Precision	Recall	F1-score	IoU	Precision	Recall	F1-score	IoU	Precision	Recall	F1-score	IoU
bicycle	**0.9392**	0.7599	0.8399	0.7243	0.9284	0.7393	0.8231	0.7001	0.9364	**0.9077**	**0.9217**	**0.8550**
bus	0.8291	0.9327	0.8779	0.7823	**0.9328**	0.9317	**0.9323**	**0.8731**	0.8725	**0.9664**	0.9170	0.8470
car	0.9580	0.8643	0.9086	0.8324	0.7573	0.9053	0.8068	0.6965	**0.9807**	**0.9829**	**0.9818**	**0.9642**
Average	0.9088	0.8523	0.8755	0.7797	0.8728	0.8588	0.8541	0.7566	**0.9297**	**0.9523**	**0.9402**	**0.8887**



Ablation Study - Two Step

To further dissect the effectiveness of our proposed IPL pipeline, we conduct an ablation study to analyze the individual contributions of our two core components:

1. The weighted voting scheme.
2. The ground-truth (GT) integration policy.

Method	bicycle	bus	car	Avg. IoU
OpenSeeD (Baseline)	0.7243	0.7823	0.8324	0.7797
EOV-Segment (Baseline)	0.7001	**0.8731**	0.6965	0.7566
IPL (Voting Only)	0.8288	0.8368	0.7484	0.8047
IPL (Full Method)	**0.8549**	0.8469	**0.9642**	**0.8887**

Ablation Study - Metric

To validate the selection of the core metric (F1-score, Precision, or Recall) for our performance-aware weighted voting scheme.
The F1-score-based Weight consistently yields superior performance across all categories, achieving the highest Average IoU (0.8047).

Voting Weight Metric	bicycle	bus	car	Avg. IoU
Precision-based Weight	0.8237	0.8324	**0.7489**	0.8017
Recall-based Weight	0.8194	0.8015	0.7425	0.7878
F1-score-based Weight	**0.8288**	**0.8368**	0.7484	**0.8047**

Ablation Study - Metric

To validate the selection of the core metric ($F1_{specific} + F1_{avg}$, $F1_{specific}$ *only*, *or* $F1_{avg}$ *only*) for our performance-aware weighted voting scheme.
The $F1_{specific} + F1_{avg}$ Weight consistently yields superior performance across all categories, achieving the highest Average IoU (0.8047).

Voting Weight Component	**bicycle**	**bus**	**car**	**Avg. IoU**
Specialist Competence ($F1_{specific}$ only)	0.8162	0.8368	**0.7489**	0.8006
Generalist Reliability ($F1_{avg}$ only)	0.8194	0.8012	0.7443	0.7883
Balanced Combination ($F1_{specific} + F1_{avg}$)	**0.8288**	**0.8368**	0.7484	**0.8047**

Conclusion

Core Conclusion

- **Successfully Addressed Label Inconsistency:** Introduced the IPL pipeline, resolving the critical challenge of harmonizing heterogeneous semantic segmentation datasets.

- **Automated High-Quality Pseudo-Labeling:** Achieved significantly higher quality pseudo-labels than individual models.

- **Validated Components:** Ablation studies confirmed that both **Weighted Voting** and **Rule-Based Integration** are crucial for superior performance.

Main Contribution

- **New Framework:** Proposed a new, effective framework for automating pixel-level labeling and addressing data scarcity with minimal manual effort.

- **Intelligent Aggregation:** Designed a new weighted voting mechanism that leverages both class-specific expertise and general model reliability.

- **Preservation Strategy:** Introduced a rule-based integration strategy that preserves the quality of original ground-truth labels.

Future Work

- **Adaptive Rule Derivation:** Explore using Large Language Models (LLMs) to automatically construct semantic hierarchies between class labels .

- **Task Extension:** Extend the ***IPL*** framework to other dense prediction tasks, such as Instance Segmentation or Depth Estimation, to validate its versatility.

Reference

[1] Bousselham, W., Thibault, G., Pagano, L., Machireddy, A., Gray, J., Chang, Y.H., Song, X.: Efficient self-ensemble for semantic segmentation. In: Proceedings of the British Machine Vision Conference (BMVC) (2022), https://bmvc2022.mpi-inf .mpg.de/0892.pdf

[2] Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuScenes: A Multimodal Dataset for Autonomous Driving . In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11618–11628 (Jun 2020). https://doi.org/10.1109/CVPR42600.2020.01164

[3] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The Cityscapes Dataset for Semantic Urban Scene Understanding . In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3213–3223 (Jun 2016). https://doi.org/10.1109/CVPR.2016.350

[4] Dietterich, T.G.: Ensemble methods in machine learning. Multiple classifier systems (MCS) pp. 1–15 (2000)

[5] Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3354–3361 (2012). https://doi.org/10.1109/CVPR.2012.6248074

[6] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Doll´ar, P., Girshick, R.: Segment anything (2023), https://arxiv.org/abs/2304.02643

[7] Lee, D.H.: Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks. ICML 2013 Workshop : Challenges in Representation Learning (WREPL) (07 2013)

[8] Niu, H., Hu, J., Lin, J., Jiang, G., Zhang, S.: Eov-seg: Efficient open-vocabulary panoptic segmentation (2024), https://arxiv.org/abs/2412.08628

[9] Xie, Q., Dai, Z., Hovy, E., Luong, T., Le, Q.: Unsupervised data augmentation for consistency training. In: Advances in Neural Information Processing Systems. vol. 33, pp. 6256–6268 (2020), https://proceedings.neurips.cc/paper_files /paper/2020/file/44feb0096faa8326192570788b38c1d1-Paper.pdf

[10] Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., Darrell, T.: Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2636–2645 (June 2020)

[11] Zhang, H., Li, F., Zou, X., Liu, S., Li, C., Yang, J., Zhang, L.: A Simple Framework for Open-Vocabulary Segmentation and Detection . In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1020–1031 (Oct 2023). https://doi.org/10.1109/ICCV51070.2023.00100

[12] Zhang, Y., Jiao, R., Liao, Q., Li, D., Zhang, J.: Uncertainty-guided mutual consistency learning for semi-supervised medical image segmentation. Artificial Intelligence in Medicine 138, 102476 (2023). https://doi.org/https://doi.org/10.1016/j.artmed.2022.102476

Thank you for listening!

Integrated Pseudo-Labeling (IPL) Pipeline

Pre-computation of Model Performance Scores

Weighted Voting and Pseudo-Label Generation

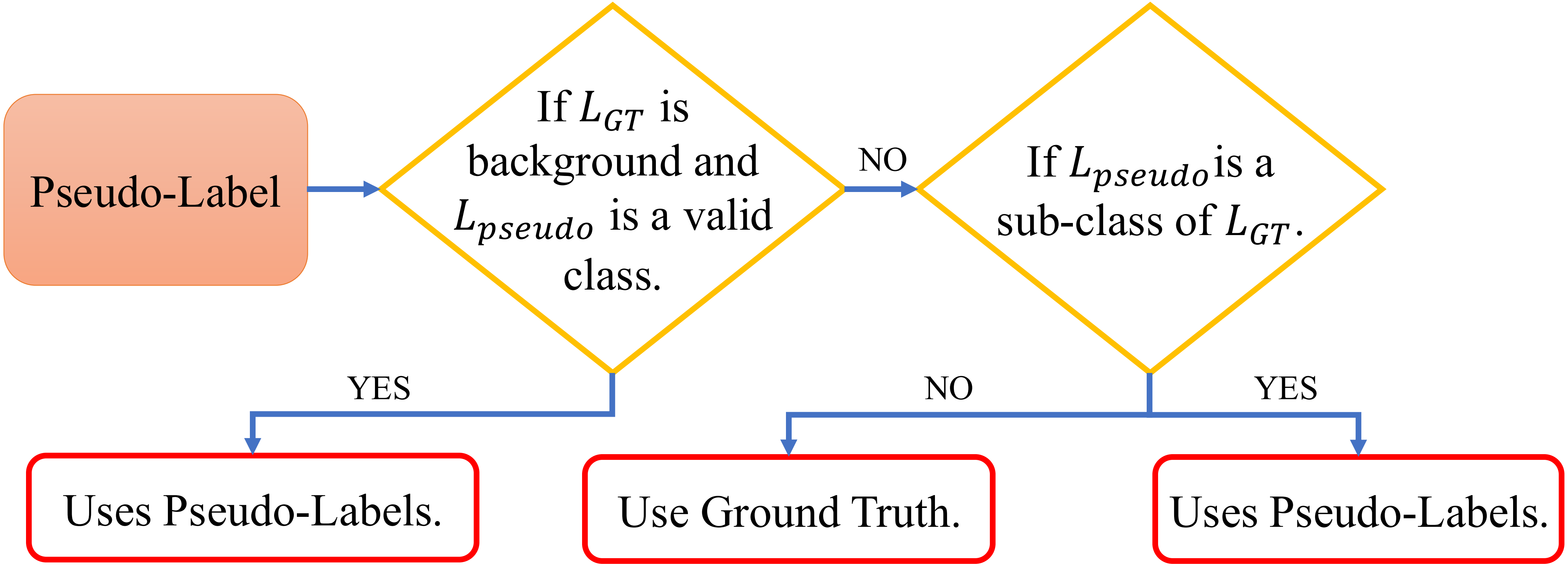
Pseudo-Label and Ground-Truth Label Integration

Pre-computation of Model Performance Scores

Weighted Voting and Pseudo-Label Generation

Pseudo-Label and Ground-Truth Label Integration

Pseudo-Label and Ground-Truth Label Integration



To expand and **fill in missing annotations** in the original data.

Ensures the **integrity** of the **original high-quality** labels is maintained.

Uses pre-defined R_{manual} to **retain finer granularity** in the final label map.

Pre-computation - Implementation