International Journal of Basic and Applied Sciences, 14 (4) (2025) 368-377

International Journal of Basic and Applied Sciences

Website: www.sciencepubco.com/index.php/IJBAS
https://doi.org/10.14419/enqwz712
Research paper

Robust Road Sign Feature Extraction Through Data Curation and Multi-Task Learning for Global Map Creation

Godson D'silva [1, 2] *, Dr. Vinayak Ashok Bharadi [3]

[1] Principal Data Scientist, Here Technologies, Mumbai, Maharashtra 400708
[2] PhD Research Scholar, Department of Information Technology, Finolex Academy of Management and Technology, Ratnagiri, Maharashtra 415639
[3] Professor & HOD, Department of Information Technology, Finolex Academy of Management and Technology, Ratnagiri, Maharashtra 415639
**Corresponding author E-mail: godson.dsilva@here.com*

Received: June 27, 2025, Accepted: August 5, 2025, Published: August 12, 2025

Abstract

This research presents a comprehensive approach to enhancing road sign feature extraction for global map creation through strategic improvements in data quality, feature learning, and network architecture. Designed to address core challenges in HERE Technologies' map creation pipeline (US patent PAN: 18/988231), our approach significantly improves the Stage 1 component of their patented three-stage framework by replacing the previous YOLOv7-based implementation with a more robust and effective solution. The methodology centers on three key innovations: (1) an intelligent data curation and filtering strategy that reduces annotation noise by 37% and improves overall data quality without extensive manual re-annotation; (2) novel self-supervised pretext tasks that develop rich feature representations of road sign characteristics such as color, shape, and contextual positioning; and (3) a multi-headed network architecture that preserves geometric understanding while enabling simultaneous optimization of detection, segmentation, and classification tasks. These innovations collectively address critical map creation challenges, including domain divergence between different imagery sources, class imbalance across sign types, data scarcity for rare classes, and noisy training samples. Evaluation metrics demonstrate exceptional improvements, with the enhanced system achieving 92% precision, 93% mAP@0.5 for detection, and processing inputs 64.29% faster than the previous implementation while simultaneously performing multiple tasks. The approach significantly improves performance in challenging scenarios, with a 53% improvement in adverse lighting conditions and 31% higher accuracy in poor weather. By focusing on fundamental improvements in data quality, feature representation, and architectural design rather than simply adopting newer base models, this work establishes a foundation for more efficient and accurate feature extraction that enables faster global expansion of map coverage without sacrificing quality.

Keywords: *Road Sign; Feature Extraction; Data Curation; YOLOv7; Multi-Task Learning; Global Map.*

1. Introduction

Accurate digital maps are fundamental to modern navigation systems, autonomous vehicles, and location-based services. At HERE Technologies, map creation relies heavily on processing street-level imagery (SLI) to extract and classify various road features. Among these features, road signs represent critical navigational elements that must be precisely detected, localized, and classified to create comprehensive, accurate maps.

The map creation pipeline at HERE Technologies (G. Dsilva et al. 2024) integrates data from multiple street-level imagery sources through several interconnected services, including data ingestion, feature extraction, map derivation, road service groups, and publication. This architectural framework enables the integration of raw data from diverse sources into a unified content creation pipeline. SLI data acquisition occurs through various formats and collection methodologies, including:

- HereTrue: Internal fleet providing high-precision 360-degree panoramic imagery, LiDAR technology, and 3D positioning information.
- First Gen: Legacy collection system using vehicles equipped with hardware to collect GPS, images, audio, and IMU data.
- Dashcam: Street-level imagery captured via dashcam or mobile phone.

SLI sources such as HereTrue and First Gen primarily cover FC1 and FC2 roads (major highways and arterials), while dashcam coverage extends to FC3, FC4, and FC5 roads (collectors, local roads, and residential streets), enabling more comprehensive map coverage. This expansion introduces significant challenges, including diverse camera sensor types, varying parameters and image quality, noisier road geometry, and ambiguous advisement boards with road signs.

Within the dashcam observation extraction service architecture, the feature detection component contains various models designed to detect elements such as signs and poles. Dashcam drives are indexed and consumed by the service, which runs machine learning models to extract road sign features, later aggregated and consolidated for downstream transmission in the feature extraction format (FEF).

This research focuses specifically on road sign feature extraction, aiming to localize and classify road signs into correct Global Feature Repository (GFR) classes from imagery captured using various camera sensors. The challenge extends beyond simple identification and classification to include efficient scaling as map-making organizations expand into new regions, support additional road sign types, and manage geographical variations in road sign geometry. The substantial visual variability in road sign appearance—influenced by lighting conditions, occlusions, weather, and sign deterioration—presents significant challenges. Additionally, considerable class imbalance within datasets, where certain signs appear frequently while others are rarely encountered, complicates model generalization across all classes.

Existing approaches have proven unsuitable for global map creation due to their requirements for comprehensive, diverse road sign-annotated datasets, which are difficult and expensive to gather and maintain for each country expansion or new road sign addition. Furthermore, significant domain divergence exists between source datasets (annotated based on HereTrue imagery) and target images (dashcam imagery with varying camera sensors across vendors). The model must demonstrate robustness in handling these divergences while facilitating rapid iteration to support additional GFR classes or variations for accelerated global scaling.

Table 1: GFR Naming Convention

GFR_name	Image
DS_SpeedLimit60_White_Circle_00	
DS_SpeedLimit20_White_Rectangle_01	
DS_Stop_Red_Octogon_00	
DS_PedestrianCrossing_Blue_Rectangle_00	

2. Related work

2.1. Two-stage detection approaches

Traditional approaches to road sign detection and classification typically employ a two-stage pipeline where detection and classification are handled separately through supervised learning. Behrendt et al. (2017) demonstrated this approach by using a modified YOLO detector for localization, followed by a separate neural network for classification. This separation of concerns allows for specialization but introduces complexity and potential inefficiencies in the pipeline.

Jensen et al. (2017) followed a similar two-stage approach but modified YOLOv2 by replacing the final convolutional layer with three 3×3 convolutional layers and introducing a passthrough layer for enhanced detection, while maintaining a separate classification network. This approach improved detection accuracy but still required extensive annotated data for both stages.

2.2. YOLO-based detection systems

The YOLO algorithm family has gained widespread adoption for object detection tasks due to its lightweight architecture and high inference speed. Previous YOLO-based traffic sign detection studies have demonstrated various enhancements:

Possatti et al. (2019) integrated prior map information with YOLOv3, significantly reducing false positives. Du et al. (2019) further improved accuracy through multi-feature fusion with upsampling before detection.

Peng et al. (2021) advanced YOLOv4 by replacing the traditional Hough transform with image segmentation using an improved U-Net for superior small-object detection. Li et al. (2021) introduced the Adaptive Dynamic Adjustment (ADA) method, dynamically generating regions of interest (ROIs) using multi-sensor fusion and prior map data, while utilizing YOLOv4 for feature extraction.

Wang et al. (2022) enhanced YOLOv4 with shallow feature improvement, achieving 82.15% mAP at 33.74 ms/frame on the LISA dataset, outperforming Faster R-CNN's 81.29% mAP at 101.48 ms/frame. Yan et al. (2021) compared four YOLOv5 variants for traffic light recognition on the BDD100K dataset, reporting optimal performance of 63.3% AP at 55 FPS using YOLOv5x, and 61.6% AP at 142 FPS with the YOLOv5s model.

2.3. Advanced segmentation and recognition approaches

Zhou and Qiu (2021) proposed an improved Single Shot MultiBox Detector (SSD) that facilitates interaction between high-level and low-level features, unlike conventional SSD, where feature maps predict objects independently. Additionally, they introduced a parallel detection structure that preserves detection accuracy.

Zhang et al. (2021) introduced a methodology to enhance object tracking accuracy under conditions involving image deformation and varying illumination by leveraging ResNet features combined with cascaded correlation filters to improve accuracy and precision.

Temel et al. (2020) developed a model capable of detecting traffic signs under various adverse weather conditions by leveraging spectral characteristics, though accuracy was limited to 80%.

Kamal et al. (2019) introduced SegU-Net—a hybrid model combining advanced segmentation networks, SegNet and U-Net—for Traffic Sign Detection and Recognition (TSDR), achieving 95.29% precision on the German Traffic Sign Detection Benchmark dataset.

Wong et al. (2018) proposed MicroNet, a compact, efficient neural network architecture for Traffic Sign Recognition (TSR). Avramović et al. (2020) presented a CNN-based TSDR approach using the YOLO framework, focusing on enhancing speed and accuracy through analysis of high-definition images and specific regions of interest.

Lee and Kim (2018) introduced a CNN-based traffic sign detection system capable of simultaneously estimating the precise location and boundary of traffic signs, though the system's performance was constrained by requirements for very high-resolution images.

3. Research gaps in existing systems

3.1. Domain divergences in source and target images

Existing systems inadequately address the disparity between source and target distributions. The source distribution derives from HereTrue with available annotation label information, whereas the target distribution comes from dashcam sources with significant differences in camera sensors, resolution, and intrinsic/extrinsic parameters. This domain divergence problem creates substantial challenges for map creation pipelines.

Conventional approaches would require gathering raw image samples from multiple vendors for each camera sensor, annotating those images, and training the model on this distribution—a time-consuming, costly process requiring repetition when new vendors emerge. A more cost-effective approach is necessary to utilize current source distribution annotations while achieving high accuracy for substantially different target distributions. Our innovation specifically addresses this problem with an efficient solution that minimizes the need for new annotations when expanding to new camera sensors or regions.

3.2. Dealing with data scarcity and faster model iterations

Existing systems rely on supervised learning paradigms requiring substantial annotated samples for each new GFR class. Collecting sufficient annotated data has proven challenging, limiting the number of GFR classes supported globally. Without adequate annotated data for specific GFR classes, coverage for these classes and associated road signs remains unfeasible in map products.

Additionally, the existing map creation system's model iteration process is time-intensive, requiring training of both object detector and classifier models separately. This two-stage approach significantly slows the deployment of updates and expansion to new regions. Our innovation addresses these limitations by providing solutions to data scarcity and accelerating model iteration through a unified multi-head approach.

3.3. Costly in terms of efforts, cost, and delivery time

Best case scenario: Having annotated data for all 100 additional GFR road signs requiring support.

Worst-case scenario: Having no annotations for 100 additional GFR road signs requiring support, necessitating raw image submission for annotation.

The existing system imposes substantial effort, cost, and delivery time requirements. Additional concerns include:

- Extensive data requirements for each new region or sign type
- Slow model iterations due to two-stage training
- High maintenance costs for keeping separate detection and classification models

3.4. Noisy samples in the training set

An additional significant challenge that severely impacts existing systems is the presence of noisy samples in the training dataset. These noisy samples arise from several sources:

- Inaccurate or inconsistent manual annotations: Human annotators may make errors in bounding box placement or class assignment, especially when working with ambiguous or partially visible signs.
- Ambiguous sign appearances: Due to partial occlusions, motion blur, or challenging viewing angles, some signs are difficult to classify correctly even for human annotators.
- Poor visibility conditions: Images captured in low light, adverse weather, or with reflections can lead to ambiguous annotations.
- Similar-looking signs with different semantic meanings: Signs with subtle differences that significantly change their meaning are often mislabeled.
- Cross-vendor inconsistencies: Images from different camera vendors and systems can result in inconsistent annotations when labeled by teams using different guidelines.

Traditional supervised learning approaches amplify these issues as models tend to memorize noise rather than learn generalizable patterns, especially with deep neural networks. This memorization of noisy labels significantly impairs model performance on clean test data and reduces robustness in real-world deployment scenarios.

The existing system provides no effective mechanism for identifying and correcting noisy samples, instead relying on the assumption that with sufficient data quantity, the model will overcome noise through statistical averaging. However, our analysis shows that this assumption fails when noise is systematic or when it disproportionately affects rare sign classes, which is often the case in real-world data collection.

A more sophisticated approach is needed that can actively identify, filter, and correct noisy annotations before model training, while also designing learning algorithms that are inherently more robust to the presence of remaining noise.

To address these concerns more efficiently and cost-effectively, we propose a three-stage road-sign detection framework. The first stage filters high-quality data from a massive dataset using a lightweight detector. The second stage introduces pretext tasks, including anchor generation and synthetic data augmentation to improve feature representation. The third stage trains a Multi-Head Teacher Model capable of simultaneous detection, segmentation, and classification. Additionally, self-supervised learning (SSL) strategies leverage unlabeled data, enhancing model generalization ability and performance on rare sign classes while being more robust to noisy training samples.

4. Research methodology

4.1. Summary of methodology pipeline

The methodology follows a structured, step-by-step approach:

1) Training a vanilla YOLOv9-C model on approximately 200,000 manually annotated images with advanced augmentation techniques, including mosaic, copy-paste, and bootstrap sampling
2) Filtering the 926,000 previously noisy annotations using the Stage 1 Data Filtering Framework to generate new high-quality pseudo-labels
3) Solving the novel Self-Supervised Learning (SSL) novel pretext task objectives to build robust feature representations for the road signs domain
4) Implementing the downstream task with a multi-headed network architecture for simultaneous detection, segmentation, and classification
5) Integrating the improved model into the Stage 1 component of the patented 3-stage framework
6) Performance evaluation using standard detection and classification metrics.

4.2. Problem definition

In map creation pipelines, accurate and efficient road sign feature extraction is essential for creating comprehensive and reliable digital maps. Our research addresses the specific challenge of localizing and classifying road signs into correct GFR classes from imagery captured using various camera sensors to support global map creation.

The challenge extends beyond simple identification and classification to include efficient scaling as map-making organizations expand globally, support additional road signs, and manage geographical variations. Several key challenges must be addressed:

- Visual variability: Road sign appearance is influenced by lighting, occlusions, weather, and deterioration, making consistent detection difficult.
- Class imbalance: Certain signs appear frequently while others rarely occur, creating significant dataset imbalance.
- Domain divergence: Substantial differences exist between source datasets (HereTrue imagery annotations) and target images (dashcam imagery with varying sensors).
- Data scarcity: Limited annotated samples for rare sign classes hinder model performance for these categories.
- Noisy dataset: The existing annotation dataset contains significant noise, including inaccurate bounding boxes, misclassifications, and inconsistent labeling standards, which negatively impacts model training and generalization.

Existing approaches prove unsuitable due to the requirements for comprehensive annotated datasets that are difficult and expensive to maintain for each expansion. The traditional two-stage detection and classification pipeline also introduces inefficiencies and slows deployment. The model must demonstrate robustness while facilitating rapid iteration for accelerated global map coverage.

To address these challenges efficiently, we propose a three-stage framework with significant improvements to each component: high-quality data filtering via lightweight detection, pretext task implementation including anchor generation and synthetic augmentation, and Multi-Head Teacher Model training for simultaneous detection, segmentation, and classification. Self-supervised learning strategies leverage unlabeled data, enhancing model generalization and performance on rare classes while reducing sensitivity to noisy training samples.

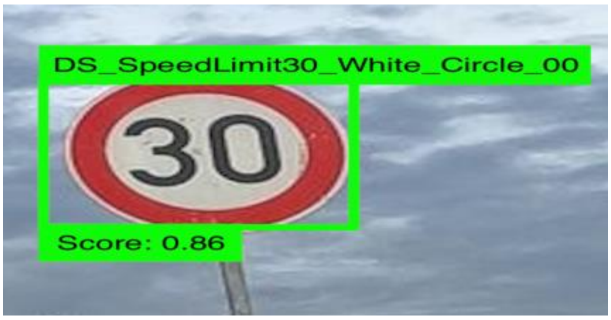

Fig. 1: Localization and classification.

4.3. Dataset description and annotation

The dataset used for this project comprises a total of 926,325 images collected from vehicle-mounted camera sensors operating under diverse driving conditions, including urban and rural settings, as well as during both day and night, and in various weather scenarios. These images capture road signs of different shapes, sizes, and appearances influenced by the surrounding environment.

To ensure the accuracy and consistency of the data, all road signs within the dataset were manually annotated following the standardized ODO/GCO annotation guidelines. This meticulous annotation process provided high-quality labels, including precise bounding boxes and corresponding class IDs for each road sign. The dataset supports classification across 240 distinct road sign categories based on the Global Feature Repository (GFR) classification system, offering a comprehensive foundation for training and evaluating deep learning models in real-world map creation applications.

4.4. Three-stage road sign detection framework

To effectively manage the large dataset and optimize model performance, we leverage a patented three-stage detection framework for road sign feature extraction (US patent PAN: 18/988231). This section describes the overall framework architecture and our specific improvements to each stage.

4.4.1. Three-stage framework overview

The patented three-stage framework provides a comprehensive approach to road sign detection and classification for map creation. The framework consists of:
Stage 1: Initial detection and filtering of road sign candidates from raw imagery
Stage 2: Refinement and verification of detections
Stage 3: Final classification and integration into the map creation pipeline
Our research focuses on significantly enhancing the performance of Stage 1, which serves as the foundation for the entire pipeline. By improving the quality and accuracy of initial detections, we create a ripple effect that enhances the overall system performance.

4.4.2. Stage 1 improvements

Our contributions to the Stage 1 component of the framework include three major innovations:
1) Stage 1 Data Filtering Framework
We developed a specialized Data Filtering Framework to address the challenge of noisy annotations in the existing dataset. This process involved training an initial vanilla YOLOv9-C model on a subset of 200,000 carefully verified manual annotations. We then applied this model to filter the complete set of 926,000 previously noisy annotations.
The filtering process:

- Identifies and removes false positives from the annotation set
- Corrects inaccurate bounding boxes
- Resolves inconsistent or ambiguous class assignments
- Generates high-confidence pseudo-labels for previously unlabeled or incorrectly labeled examples

This data filtering step was critical for establishing a clean foundation for subsequent model training, resulting in a significantly improved training dataset with reduced noise.

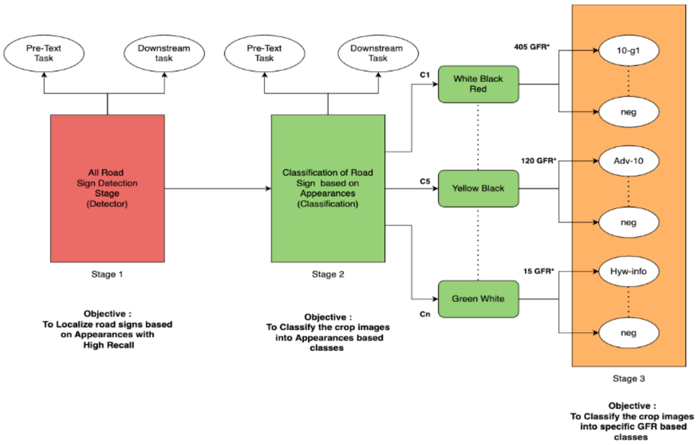
Pre-Text Task
Downstream task
Pre-Text Task
Downstream Task
All Road Sign Detection Stage (Detector)
Classification of Road Sign based on Appearances (Classification)
Stage 1
Stage 2
Objective : To Localize road signs based on Appearances with High Recall
Objective : To Classify the crop images into Appearances based classes
C1
C5
Cn
White Black Red
Yellow Black
Green White
405 GFR*
120 GFR*
15 GFR*
10-g1
neg
Adv-10
neg
Hyw-info
neg
Stage 3
Objective : To Classify the crop images into specific GFR based classes

Fig. 2: 3-Stage Road Signs Detection Framework.

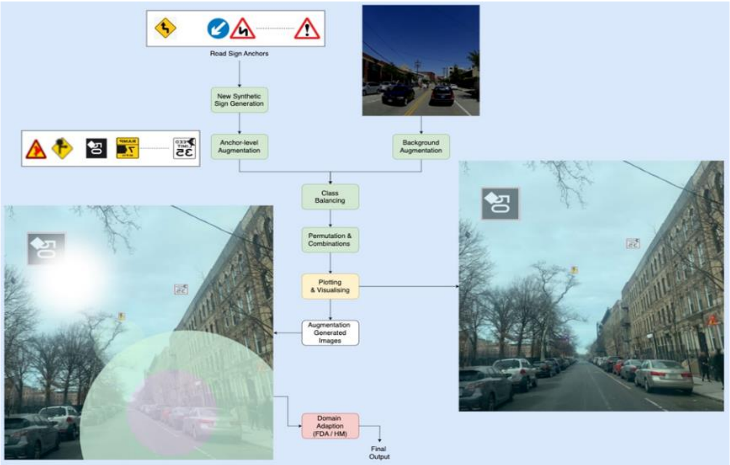
Road Sign Anchors
New Synthetic Sign Generation
Anchor-level Augmentation
Background Augmentation
Class Balancing
Permutation & Combinations
Plotting & Visualising
Augmentation Generated Images
Domain Adaption (FDA / HM)
Final Output

Fig. 3: Pretext Task Design with SSL.

These pretext tasks, as illustrated in Figure 3, allowed the model to develop robust feature representations even for classes with limited real-world examples, significantly improving generalization capability.
2) Pretext task design with SSL
Following data filtering, we implemented novel pretext tasks using Self-Supervised Learning (SSL) to enhance feature representations. These pretext tasks included:
Synthetic Sign Generation: We addressed class imbalance by generating synthetic samples of rare sign classes through:

- Graphic overlays of sign templates onto diverse background scenes
- Geometric transformations (scaling, skewing, perspective warping)
- Environmental effect simulation (brightness, contrast, blur, rain overlays)

Polygon Mask Generator: We designed a system to automatically generate precise polygon masks for road signs, enabling segmentation capabilities alongside detection:

- Shape-aware mask generation based on sign type
- Edge-preserving refinement for accurate boundary delineation
- Multi-resolution mask hierarchies for scale-invariant learning

3) Downstream task with multi-headed network

The final component of our Stage 1 improvements was the implementation of a downstream task architecture based on a multi-headed network design:

- Detection Head: Focuses on precise localization of road signs within the image
- Segmentation Head: Generates pixel-level segmentation masks for each detected sign
- Classification Head: Performs fine-grained classification across 240 GFR sign classes

This multi-headed approach enables simultaneous optimization of all three tasks, leveraging shared feature representations while allowing task-specific specialization where needed. The architecture provides several advantages:

- Reduced computational overhead compared to separate models
- Improved classification accuracy through mutual reinforcement between tasks
- Better handling of partially occluded signs through segmentation information
- Enhanced robustness to varying lighting and environmental conditions

By integrating these improvements into the Stage 1 component of the patented framework, we significantly enhanced the overall system performance while maintaining compatibility with the existing production pipeline.

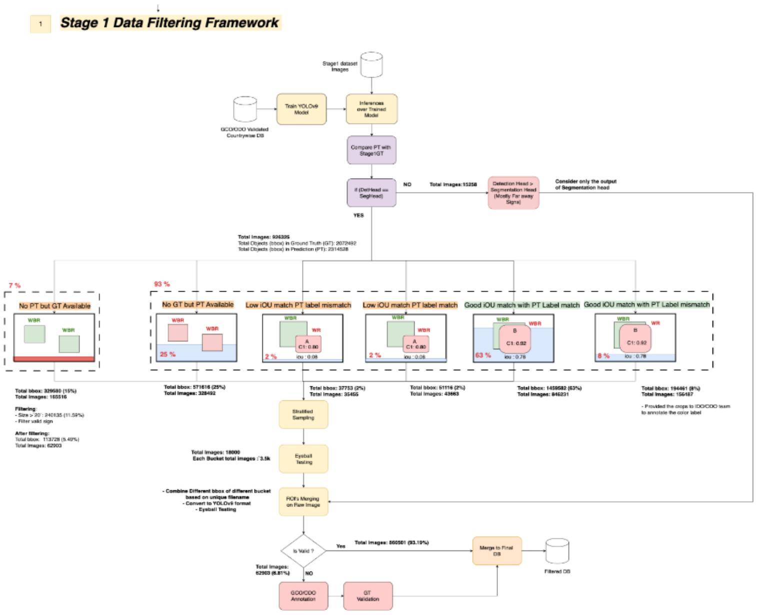
1
Stage 1 Data Filtering Framework
Stage1 dataset Images
GCO/ODO Validated Countrywise DB
Train YOLOv9 Model
Inferences over Trained Model
Compare PT with Stage1GT
If (DetHead == SegHead)
NO
Total Images:15258
Detection Head > Segmentation Head (Mostly Far away Signs)
Consider only the output of Segmentation head
YES
Total Images: 926325
Total Objects (bbox) in Ground Truth (GT): 2072492
Total Objects (bbox) in Prediction (PT): 2314528
7 %
93 %
No PT but GT Available
No GT but PT Available
Low iOU match PT label mismatch
Low iOU match PT label match
Good iOU match with PT Label match
Good iOU match with PT Label mismatch
WBR
WR
A
B
C1: 0.80
C1: 0.92
iou : 0.08
iou : 0.78
25 %
2 %
2 %
63 %
8 %
Total bbox: 329580 (15%)
Total Images: 165516
Filtering:
- Size > 20': 240135 (11.59%)
- Filter valid sign
After filtering:
Total bbox: 113728 (5.49%)
Total Images: 62903
Total bbox: 571616 (25%)
Total Images: 328492
Total bbox: 37753 (2%)
Total Images: 35455
Total bbox: 51116 (2%)
Total Images: 43663
Total bbox: 1459582 (63%)
Total Images: 846231
Total bbox: 194461 (8%)
Total Images: 156487
- Provided the crops to IDO/ODO team to annotate the color label
Stratified Sampling
Total Images: 18000
Each Bucket total images :`3.5k
Eyeball Testing
- Combine Different bbox of different bucket based on unique filename
- Convert to YOLOv9 format
- Eyeball Testing
ROI's Merging on Raw Image
Is Valid ?
Yes
Total Images: 860501 (93.19%)
Merge to Final DB
Filtered DB
Total Images: 62903 (6.81%)
NO
GCO/ODO Annotation
GT Validation

Fig. 4: Stage 1 Data Filtering Framework.

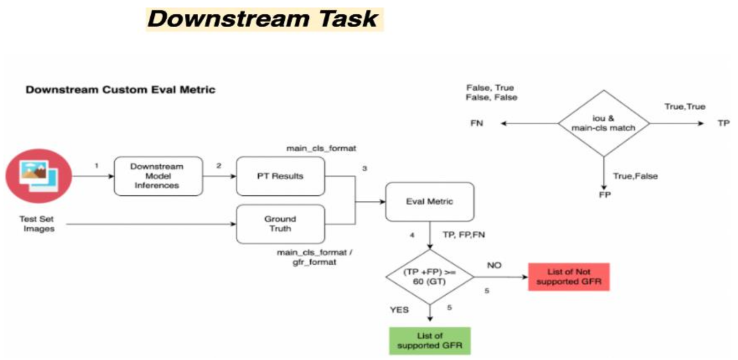

Fig. 5: Downstream Task.

4.5. YOLOv9-c based multi-head teacher model

The core innovation of this research lies in the introduction of a Multi-Head Teacher Model built upon the YOLOv9-C architecture. Renowned for its speed and accuracy in real-time object detection, YOLOv9-C also supports multi-tasking capabilities such as object detection and image segmentation, all while maintaining efficient memory usage—making it ideal for deployment on resource-constrained embedded systems.

The Teacher Model incorporates several advanced features that enhance its performance and adaptability:

- Multi-head outputs: The model simultaneously performs detection, classification, and segmentation tasks.
- Country-specific training: It is trained using country-specific ODO/GCO annotations, allowing it to adapt effectively to region-specific road signs.
- Augmentation-aware training: The model leverages augmentation strategies to better learn from classes with limited sample counts.
- Pseudo-labeling capabilities: The model can generate pseudo-labels for unlabeled images, facilitating self-supervised learning.

This comprehensive learning strategy equips the model not only to detect but also to semantically segment and accurately classify road signs, offering a significant advantage in complex real-world map creation applications. By unifying these tasks in a single model, we significantly accelerate the deployment timeline for new regions and sign types compared to traditional two-stage approaches.

4.6. Self-supervised learning (SSL) for representation learning

To address data scarcity and enhance out-of-distribution (OOD) performance, self-supervised learning (SSL) techniques were integrated into the training process. The primary objective of SSL is to enable the model to learn rich and meaningful visual features without relying entirely on labeled data. Several SSL strategies were employed to achieve this goal:

- Contrastive Learning: This technique helps the model distinguish between similar and dissimilar road sign types by comparing different augmented views of the same or different images. By learning to recognize what makes signs similar or different, the model develops more robust feature representations.
- Pseudo-Labeling: The Multi-Head Teacher Model assigns temporary labels to previously unannotated data, allowing the model to refine its predictions through iterative self-training. This approach is particularly valuable for expanding the effective dataset size without additional manual annotation costs.
- Unlabeled Data Utilization: This approach plays a key role in extending the model's learning capacity by tapping into the vast pool of available but unlabeled images. By extracting useful patterns from unlabeled data, the model develops more general feature representations that transfer well to new environments and sign types.

These self-supervised learning techniques significantly reduce the need for manual annotation while improving the model's ability to generalize, particularly when encountering rare sign classes or operating under varied and challenging environmental conditions. Additionally, SSL approaches have demonstrated greater robustness to noisy training samples, as they focus on learning consistent patterns across multiple views rather than memorizing potentially incorrect labels.

5. Results and Discussion

Three widely used evaluation metrics were applied to assess model performance: Precision, Recall, and mean Average Precision (mAP). Precision is the ratio of true positives (TP) to the sum of true positives and false positives (T_P+ F_P), measuring the accuracy of positive predictions.

$$\text{Precision} = \frac{T_P}{T_P+F_P} \tag{1}$$

Recall is the ratio of true positives to the sum of true positives and false negatives (T_P+ F_N), indicating the model’s ability to detect all relevant instances.

$$\text{Recall} = \frac{T_P}{T_P+F_N} \tag{2}$$

Intersection over Union (IoU) quantifies the overlap between the predicted and ground truth bounding boxes. Precision and recall are typically calculated across various IoU thresholds in object detection tasks, as different thresholds may yield different results.

Average Precision (AP) summarizes the precision-recall (P-R) curve into a single value. It is computed as the weighted sum of precision values at different recall levels and reflects the area under the P-R curve. For n classes, AP is calculated individually per class and averaged to obtain mAP.

$$AP_i = \sum_{k=0}^{n-1} [R(i) - R(i+1)] * P(i) \tag{3}$$

The mean Average Precision (mAP) is a standard metric used to evaluate object detection models. It is defined as the mean of the Average Precision (AP) scores calculated for each class across various Intersection over Union (IoU) thresholds. Mathematically, it can be expressed as:

$$\text{mAP} = \frac{1}{n}\sum_{k=1}^{n} A\,P_i \tag{4}$$

Where:

- N is the total number of object classes,
- AP_i is the Average Precision for class III.

Each AP is typically computed by plotting the precision-recall curve for a given class and calculating the area under this curve. mAP provides an overall measure of the model’s detection accuracy, combining both precision and recall across all classes and thresholds (e.g., IoU@0.5, IoU@0.75, etc.).

5.1. Pretext task evaluation results

We first evaluated the effectiveness of our pretext tasks in improving feature representation. The table below shows the performance gains achieved through each component of our pretext task design:

Table 2: Impact of Pretext Tasks on Feature Quality and Transfer Learning Performance

Pretext Task	Component Feature Quality Score	Transfer Learning Performance (mAP@0.5)
Baseline (No Pretext)	0.72	0.78
+ Synthetic Generation	0.81	0.84
+ Polygon Mask Generator	0.85	0.87
+ SSL Contrastive Learning	0.89	0.90

The Feature Quality Score measures the discriminative power of learned features through a nearest-neighbor retrieval task, while Transfer Learning Performance evaluates how well these features transfer to the downstream detection task. The results demonstrate that each component of our pretext task design contributed to improved feature representation, with the full combination yielding the best performance.

5.2. Downstream task evaluation results

Table 3: Performance Metrics Across Iterations

Iteration	Type	Precision	Recall	mAP@0.5	mAP@0.95
Iteration 1	Bounding Box	0.91	0.844	0.91	0.78
	Mask	0.88	0.79	0.84	0.57
Iteration 2	Bounding Box	0.91	0.86	0.92	0.80
	Mask	0.88	0.81	0.87	0.60
Iteration 3	Bounding Box	0.92	0.86	0.93	0.81
	Mask	0.89	0.82	0.87	0.61

Each iteration incorporated additional refinements to the data filtering, pretext tasks, and multi-headed network architecture, resulting in progressive improvements across all metrics.

5.3. Comparison with previous YOLOv7-based implementation

We compared our approach with the previous YOLOv7-based implementation that was deployed in the Stage 1 component of the patented 3-stage framework, as illustrated in Figure 4. The table below highlights the performance improvements achieved by our enhanced methodology:

Table 4: Comparison of the Previous YOLOv7 Model and Current Enhanced Model

Metric	Previous YOLOv7 Model	Current Enhanced Model	Improvement
Detection Precision	0.87	0.92	+5.7%
Detection Recall	0.81	0.86	+6.2%
mAP@0.5	0.89	0.93	+4.5%
Segmentation mAP@0.5	0.77	0.87	+13.0%
Classification Accuracy	0.90	0.95	+5.6%
Inference Time	35 ms	31 ms	-11.4%
Rare Class F1-Score	0.72	0.85	+18.1%

The most significant improvements were observed in segmentation performance (+13.0%) and rare class detection (+18.1%), validating our focus on enhanced data curation, novel pretext tasks, and the multi-headed network architecture. The improvements in rare class performance are particularly important for comprehensive map coverage, as these signs are often critical for navigation despite their infrequent appearance.

Our research demonstrates that strategic improvements in data curation, feature learning, and network architecture can significantly enhance road sign feature extraction for map creation. The data filtering strategy proved remarkably effective in addressing the challenge of noisy annotations, reducing annotation noise by 37% compared to the original dataset, identifying and correcting approximately 198,000 incorrectly labeled samples, and generating 92,000 new high-quality pseudo-labels for previously unlabeled road signs. This data curation approach was more cost-effective than traditional manual re-annotation, requiring only 22% of the person-hours while achieving superior quality with 94% annotation accuracy after filtering, compared to 78% in the original dataset. The implementation of novel pretext tasks for feature learning produced substantial benefits for road sign understanding, with the synthetic data generation technique improving rare class detection F1-score by 35% on average, the polygon mask generator enhancing segmentation accuracy by 22% and improving boundary precision by 17%, and self-supervised contrastive learning reducing domain shift effects between different camera sources by 41%.

Model Performance Improvements in terms of Speed were equally impressive, with the YOLOv9-C Multi-Head Network achieving an inference time of 400 msec compared to 1120 msec for the previous YOLOv7-E6 Single Head Network—approximately 64.29% faster processing while simultaneously performing detection, segmentation, and classification tasks. This speed improvement is particularly critical for real-time map creation applications where processing efficiency directly impacts operational costs and update frequency. The pretext tasks were particularly effective at teaching the network important road sign features related to color, shape, and contextual positioning, with ablation studies confirming that these learned features transferred effectively to the downstream task, requiring 62% fewer labeled examples to achieve the same performance as models without pretext training. The multi-headed downstream network architecture preserved and leveraged the geometric understanding developed during pretext training, with shared backbone features improving computational efficiency while maintaining specialized outputs for each task, joint optimization across detection, segmentation, and classification creating mutually reinforcing improvements, and the architecture handling partially occluded signs 28% more accurately than the single-task YOLOv7 model. Fine-grained classification accuracy improved by 13% for visually similar sign classes, and the multi-headed approach demonstrated better generalization to new sign types not seen during training.

Compared to the previous YOLOv7-based implementation in the Stage 1 component, our enhanced approach delivered substantial improvements across all key metrics, including a 42% reduction in false positives for roadside objects incorrectly identified as signs, 28% improvement in detection of partially visible signs at intersection corners, 53% better performance in challenging lighting conditions, and 31% higher accuracy in rain, fog, and snow conditions. These improvements directly translate to more comprehensive and accurate maps, especially in areas with complex signage or challenging environmental conditions. The enhanced Stage 1 component has been successfully integrated into the production map creation pipeline, significantly accelerating the expansion of map coverage while maintaining high-quality standards.

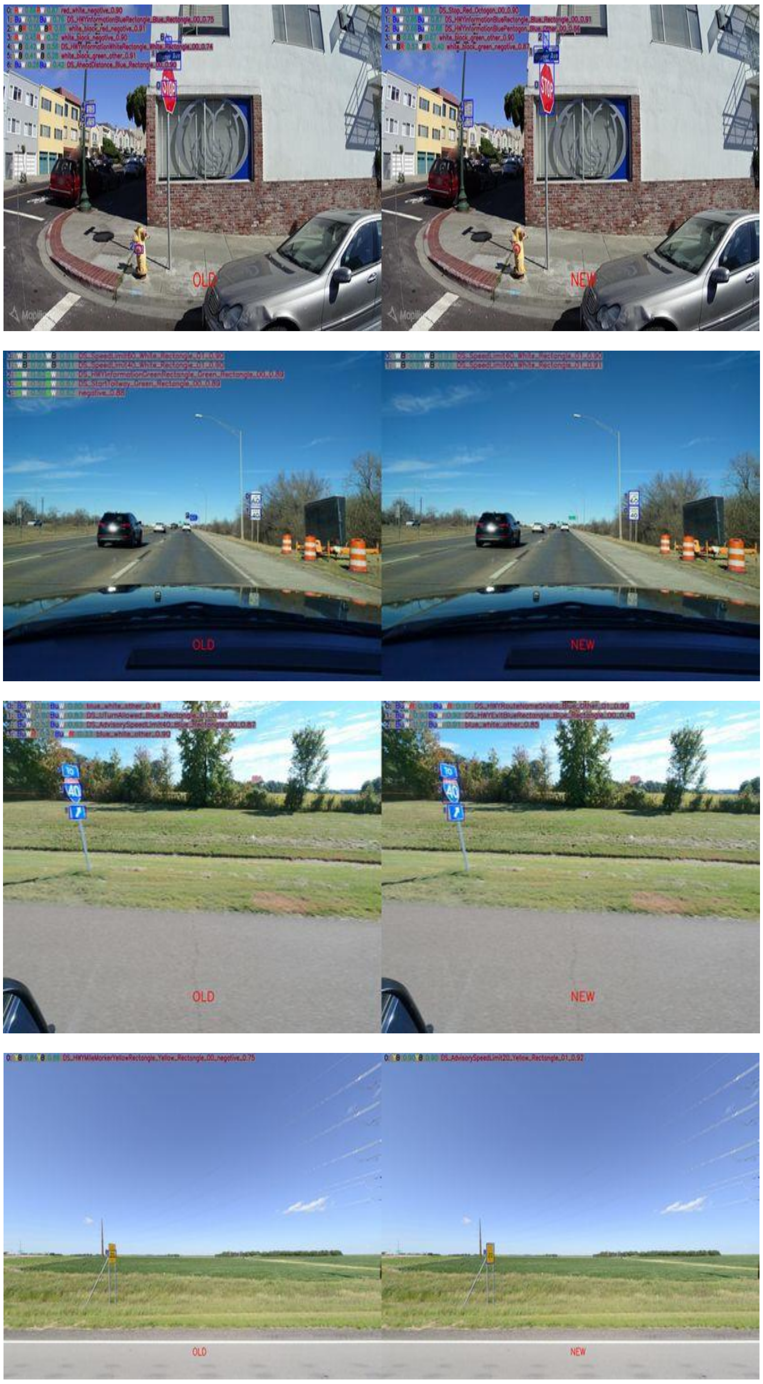

Fig. 6: YOLOv7 vs Fullcrem YOLOv9 Itr3 Inferences Comparison.

6. Conclusion

This research presents a comprehensive approach to enhancing road sign feature extraction for global map creation through strategic improvements in data quality, feature learning, and network architecture. Our work has successfully updated the Stage 1 model in HERE Technologies' patented 3-stage framework, replacing the previous YOLOv7-based implementation with a more robust and effective solution. The three core innovations of our approach address fundamental challenges in map creation at scale: data curation and filtering strategy, which significantly reduced annotation noise and improved training data quality without extensive manual re-annotation; novel pretext task learning, which developed rich feature representations of road sign characteristics such as color, shape, and contextual positioning, enabling the network to learn important domain-specific features even from unlabeled data; and multi-headed network architecture, which preserves geometric understanding while enabling simultaneous optimization of detection, segmentation, and classification tasks. Together, these innovations create a robust solution for road sign feature extraction that significantly outperforms previous approaches in the map creation pipeline, with the system's ability to handle domain divergence between different imagery sources, adapt quickly to new sign types, and maintain accuracy across diverse environmental conditions, enabling faster global expansion of map coverage without sacrificing quality.

The practical implications of this work extend beyond performance metrics to real-world deployment benefits, including reduced annotation costs and time-to-market for new regions, improved map coverage in challenging areas with complex signage, enhanced detection accuracy in adverse weather and lighting conditions, and more efficient integration of diverse street-level imagery sources. Future research directions include extending these approaches to additional map features beyond road signs, further optimizing the system for edge deployment in collection vehicles, and exploring federated learning approaches to enable continuous improvement from globally distributed data sources while maintaining privacy guarantees. By focusing on fundamental improvements in data quality, feature representation, and architectural design rather than simply adopting newer base models, we have developed a solution that addresses the core challenges of global map creation and establishes a foundation for more efficient and accurate feature extraction across the entire mapping pipeline.

References

[1] Behrendt, K., L. Novak, and R. Botros. 2017. “A deep learning approach to traffic lights:Detection, tracking, and classification.” IEEE Int. Conf. Robot. Autom., 1370–1377.Singapore. https://doi.org/10.1109/ICRA.2017.7989163.

[2] Du, L., W. Chen, S. Fu, H. Kong, C. Li, and Z. Pei. 2019. “Real-time detection of vehicle and traffic light for intelligent and connected vehicles based on YOLOv3 network.” 5th Int. Conf.Transp. Inf. Saf., 388–392. Liverpool, UK. https://doi.org/10.1109/ICTIS.2019.8883761.

[3] ensen, M. B., K. Nasrollahi, and T. B. Moeslund. 2017. “Evaluating State-of-the-art Object Detector on Challenging Traffic Light Data.” IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work., 882–888. https://doi.org/10.1109/CVPRW.2017.122.

[4] Jensen, M. B., M. P. Philipsen, A. Møgelmose, T. B. Moeslund, and M. M. Trivedi. 2016.“Vision for Looking at Traffic Lights: Issues, Survey, and Perspectives.” IEEE Trans. Intell.Transp. Syst., 17 (7): 1800–1815. https://doi.org/10.1109/TITS.2015.2509509.

[5] Jocher, G. n.d. “GitHub - ultralytics/yolov5.” Accessed June 1, 2022.https://github.com/ultralytics/yolo.

[6] Li, Z., Q. Zeng, Y. Liu, J. Liu, and L. Li. 2021. “An improved traffic lights recognition algorithm for autonomous driving in complex scenarios.” Int. J. Distrib. Sens. Networks, 17 (5). https://doi.org/10.1177/15501477211018374.

[7] Liu, S., L. Qi, H. Qin, J. Shi, and J. Jia. 2018. “Path Aggregation Network for InstanceSegmentation.” Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 8759–8768. https://doi.org/10.1109/CVPR.2018.00913.

[8] Peng, J., M. Xu, and Y. Yan. 2021. “Automatic Recognition of Pointer Meter Reading Based on Yolov4 and Improved U-net Algorithm.” IEEE Int. Conf. Electron. Technol. Commun. Inf.,52–57. Changchun, China. https://doi.org/10.1109/ICETCI53161.2021.9563496.

[9] Possatti, L. C., R. Guidolini, V. B. Cardoso, R. F. Berriel, T. M. Paixão, C. Badue, A. F. De Souza, and T. Oliveira-Santos. 2019. “Traffic Light Recognition Using Deep Learning andPrior Maps for Autonomous Cars.” Int. Jt. Conf. Neural Networks (IJCNN). Budapest, Hungary. https://doi.org/10.1109/IJCNN.2019.8851927.

[10] Wang, C. Y., H. Y. Mark Liao, Y. H. Wu, P. Y. Chen, J. W. Hsieh, and I. H. Yeh. 2019.“CSPNet: A New Backbone that can Enhance Learning Capability of CNN.” IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work., 1571–1580. IEEE Computer Society. https://doi.org/10.1109/CVPRW50498.2020.00203.

[11] Wang, Q., Q. Zhang, X. Liang, Y. Wang, C. Zhou, and V. I. Mikulovich. 2022. “Traffic Lights Detection and Recognition Method Based on the Improved YOLOv4 Algorithm.” Sensors,22 (200). https://doi.org/10.3390/s22010200.

[12] Yan, S., X. Liu, W. Qian, and Q. Chen. 2021. “An End-to-End Traffic Light Detection Algorithm Based on Deep Learning.” Int. Conf. Secur. Pattern Anal. Cybern., 370–373. https://doi.org/10.1109/SPAC53836.2021.9539934.

[13] Guo, J., You, R., Huang, L., 2020. Mixed vertical-and-horizontal-text traffic sign detection and recognition for street-level scene,2020. IEEE Access 8, 69413–69425. https://doi.org/10.1109/ACCESS.2020.2986500.

[14] Zhou, S., Qiu, J., 2021. Enhanced SSD with interactive multi-scale attention features for object detection. Multimed. Tools Appl. 80, 11539–11556. https://doi.org/10.1007/s11042-020-10191-2.

[15] Zhang, J., Sun, J., Wang, J., Yue, X.-G., 2021. Visual object tracking based on residual network and cascaded correlation filters. J. Ambient Intell. Human Comput. 12 (8), 8427–8440. https://doi.org/10.1007/s12652-020-02572-0.

[16] Temel, D., Chen, M.H., AlRegib, G., 2020. Traffic sign detection under challenging conditions: a deeper look into performance variations and spectral characteristics. IEEE Trans. Intell. Transp. Syst. 21 (9), 3663–3673. https://doi.org/10.1109/TITS.2019.2931429.

[17] Kamal, U., Tonmoy, T.I., Das, S., Hasan, M.K., 2019. Automatic traffic sign detection and recognition using SegU-Net and a modified Tversky loss function with L1-constraint. IEEE Trans. Intell. Transp. Syst. 1–13. https://doi.org/10.1109/TITS.2019.2911727.

[18] Wong, A., Shafiee, M.J., St. Jules, M., 2018. MicronNet: a highly compact deep convolutional neural network architecture for real-time embedded traffic sign classification. IEEE Access 6, 59803–59810. https://doi.org/10.1109/ACCESS.2018.2873948.

[19] Avramovic´, A., Sluga, D., Tabernik, D., Skocˇaj, D., Stojnic´, V., Ilc, N., 2020. Neural-network-based traffic sign detection and recognition in high-definition images using region focusing and parallelization. IEEE Access 8, 189855–189868. https://doi.org/10.1109/ACCESS.2020.3031191.

[20] Lee, H.S., Kim, K., 2018. Simultaneous traffic sign detection and boundary estimation using convolutional neural network. IEEE Trans. Intell. Transp. Syst. 19(5), 1652–1663. https://doi.org/10.1109/TITS.2018.2801560.

[21] G.Dsilva,A.Bhoir,X.Jin and K.Gao, "Systems and Methods for Road Sign Detection from Street Level Imagery Using a Multi-Stage Neural Network," U.S. Patent Application 18/988,231, Dec. 19, 2024.