Scalable Deep Learning Framework for Public Transit Sign Detection

Nicolò Savioli[1], Erkin Turkoz[1], Tobias Neumann[1], Lu Liu[1], Yongliang Wang[1], Yanfeng Zhang[1]

[1]Riemann Lab, Huawei - (nicolo.savioli, erkin.turkoz, tobias.neumann, luliu1, wangyongliang775, zhangyanfeng8)@huawei.com

Keywords: Object Detection, Transit Sign Detection, Deep Learning, Urban Navigation, Template Matching, Scalability

Abstract

Detecting public transit signs in urban environments is a complex and challenging task due to the significant variability of these signs across different regions and cities. Unlike standardized traffic signs, transit signs often vary considerably, requiring adaptable and robust solutions for effective detection. In this work, we propose a novel Domain-Specific Agnostic Segmentation Method with Contrastive Learning, integrated into a scalable two-phase detection pipeline. This innovative approach enables our model to adapt to city-specific visual patterns without manual prompts, achieving accurate detection even for small and distant transit signs. As a significant contribution, we introduce the first dataset dedicated to public transit signs, consisting of 2,300 manually annotated images from multiple European cities. Our method significantly outperforms existing Swin Transformer-based detection models in real-world tests. These results establish a new benchmark for public transit sign detection, highlighting the effectiveness of our approach in providing a robust, scalable solution for intelligent transportation systems and reducing the need for retraining in diverse urban environments.

1. Introduction

Public transit signs, such as those that indicate bus stops, metro entrances, and tram stations, are integral to urban navigation and optimization of route systems (Liu et al., 2023). These signs provide essential information to pedestrians and autonomous vehicles (Ashraf and Idrisi, 2024), enhancing spatial awareness and facilitating efficient movement within densely populated areas. Therefore, accurate detection of public transit signs is a critical component in the development of intelligent transportation systems.

Unlike standardized traffic signs, which maintain consistent designs across regions (for example, stop signs or speed limit indicators), public transit signs exhibit significant variability. They often differ not only between countries, but even among cities within the same country. This variability poses substantial challenges for detection algorithms, as transit signs can closely resemble traffic signs, increasing the risk of misclassification. For example, a tram stop sign in one city might be visually similar to a pedestrian crossing sign in another. Consequently, effective detection of transit signs requires models that are flexible and adaptive, capable of handling diverse visual patterns while minimizing confusion with traffic signs.

Detecting public transit signs involves several key challenges. First, transit signs are subject to frequent changes due to updates in routes or transit lines, which requires real-time detection methods (Liu et al., 2022). Second, significant design variability across different regions complicates the detection task, as models must generalize across a wide range of visual appearances. Third, the scarcity of publicly available large-scale data sets specifically dedicated to public transit signs hinders the development of robust detection models (Ertler et al., 2020).

Although recent methods like Co-DETR (Zong et al., 2023) have been introduced, they predominantly rely on the Swin Transformer backbone and focus on enhancing detection heads rather than fundamentally altering backbone architectures. Consequently, performance improvements over traditional deep learning methods remain limited, especially when addressing the core challenges of transit sign detection, such as design variability and visual similarity to other signs.

Traditional deep learning object detection approaches often struggle to generalize in diverse urban environments without extensive manual labeling or retraining (Zhu and Yan, 2022). Fine-tuning pre-trained models on small, localized datasets can lead to overfitting, resulting in poor generalization to different regions (Malpani et al., 2023). Furthermore, visual similarities between traffic and transit signs can lead to frequent misclassifications (Zhang and Sabuncu, 2018), which undermines the reliability of detection systems.

Recent advancements in unsupervised learning, such as DETReg (Bar et al., 2022) and JoinDet (Wang et al., 2022), have shown promise in object localization and embedding, achieving encouraging results in general object detection tasks. Similarly, models such as LOST (Siméoni et al., 2021) and TokenCut (Wang et al., 2023) leverage vision transformers and graph-based algorithms to localize and segment salient objects without relying on extensive labeled data. However, these unsupervised approaches face significant challenges when applied to the detection of public transit signs. The small size and high visual similarity of the transit signs, often embedded in complex urban scenes, make it difficult to differentiate them from the traffic signs and other background elements. This is particularly problematic in fine-grained recognition tasks, where unsupervised methods struggle to accurately distinguish subtle visual details (Yang et al., 2012). Even recent improvements, such as EAGLE (Kim et al., 2024), which enhance object-level semantics through spectral clustering and contrastive learning, continue to face limitations when distinguishing small, visually similar objects due to the inherent challenges of vision transformer architectures.

To address these limitations, we propose a scalable deep learning framework for public transit sign detection that minimizes the need for extensive fine-tuning and enables robust detection across regions with minimal retraining. Using advanced tech-

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-G-2025-779-2025 | © Author(s) 2025. CC BY 4.0 License.

niques in computer vision, our framework achieves high accuracy in diverse urban environments.

Our key contributions are as follows:

1. **Two-Phase Detection Method:** We propose a scalable two-phase detection method that combines agnostic segmentation with a custom matcher network to accurately identify and match detected signs, achieving high accuracy with minimal training data.
2. **Scalability:** Our system addresses the challenges of scaling by eliminating the need to retrain models for each new city, significantly reducing deployment time and effort.
3. **Improved Detection Performance:** Our method surpasses state-of-the-art detection models, demonstrating substantial enhancements in the detection and classification of public transit signs across diverse urban settings.
4. **Novel Dataset:** We introduce a new dataset comprising approximately 2,300 manually annotated public transit signs collected from ten European cities, providing a critical resource for advancing research in transit sign detection.

2. Method

Our framework integrates advanced techniques for the detection of scalable transit signs in diverse urban environments, as illustrated in Figure 1(b) and Figure 1(c). The methodology is divided into two main stages.

1. **Detection Stage:** This stage includes:
   (a) **Domain-Specific Agnostic Segmentation:** A Swin Transformer-based Mask R-CNN, pre-trained on diverse urban datasets and fine-tuned without class labels, detects small and distant sign-like objects, capturing potential transit and traffic signs.
   (b) **Class-Specific Interference Suppression (CSIS):** The CSIS network filters out irrelevant traffic signs, isolating transit signs for more accurate detection.
2. **Classification Stage:** Detected signs are matched with predefined transit sign templates using a matching network for accurate classification in cities.

2.1 Domain-Specific Agnostic Segmentation

Agnostic segmentation focuses on segmenting objects in an image without relying on specific class labels. Models like Segment Anything Model 2 (SAM 2) (Ravi et al., 2024) generate segmentation masks using prompts, but rely heavily on grid-based embeddings and manual input, which are not ideal for detecting small and distant objects like public transit signs. These signs often lack prominent features and vary significantly in urban environments, making manual prompting impractical.

To overcome these limitations, we propose a novel Domain-Specific Agnostic Segmentation method that eliminates the need for manual prompts and grid-based embeddings. As illustrated in Figure 1(a), our approach employs an automated domain-adaptive process optimized to detect small objects such as transit signs in diverse urban landscapes. The key innovation lies in using contrastive learning tailored to each specific city.

Our method operates in two main steps. First, the model undergoes contrastive learning on city-specific data to learn unique urban visual patterns. This allows it to capture subtle differences and features characteristic of a particular city's transit signs without requiring explicit annotations or prompts. Second, we fine-tune the model using a Swin Transformer-based Mask R-CNN (He et al., 2018), deliberately excluding the classification loss during training. By focusing solely on segmentation and regression losses, the model achieves true class-agnostic segmentation, identifying sign-like objects based on visual cues alone.

This approach offers several advantages over existing methods like SAM 2:

- **Automated Segmentation**: By eliminating manual prompts, our method streamlines the segmentation process, making it fully automated and efficient for real-world applications.
- **Optimized for Small Objects**: The combination of city-specific contrastive learning and exclusion of classification loss enhances the detection of small and distant transit signs, addressing a key limitation of SAM 2.
- **Domain Adaptability**: Iterative training in different cities enables the model to generalize effectively across diverse urban environments, capturing a wide range of visual patterns associated with transit signs.

We pre-train our model extensively on street view images from multiple cities, utilizing datasets like VISTAS (Neuhold et al., 2017) and the Mapillary Traffic Sign Dataset (Ertler et al., 2020). During training, we integrate both segmentation masks and bounding boxes by converting bounding boxes into pseudo-segmentation masks, treating the bounding box area as a positive mask. This hybrid approach allows the model to learn more precise object boundaries, improving its ability to detect small objects with high precision.

After segmentation, identified sign-like objects are converted back into bounding boxes for integration with the two-phase method, ensuring compatibility with downstream processes that require bounding-box input.

2.2 Class-Specific Interference Suppression (CSIS)

The Class-Specific Interference Suppression (CSIS) network is used to remove traffic signs from input images, preventing misclassification with transit signs. Using Faster R-CNN (Ren et al., 2016) trained on the Mapillary Traffic Sign Dataset to generate a traffic sign mask M_{traffic}. The mask is applied to the input image I as follows:

$$I_{\text{clean}} = I \odot (1 - M_{\text{traffic}}), \tag{1}$$

where:

- I_{clean} = cleaned image without traffic signs,
- I = original input image,

- M_{traffic} = predicted traffic sign mask,
- $\odot$ = element-wise multiplication.

This operation removes traffic signs, leaving only transit signs and relevant background, thus improving classification accuracy in subsequent steps.

2.3 Matcher Network: Feature Correlation and Classification

The Matcher Network classifies transit signs by using Vision Transformers (ViT) (Dosovitskiy et al., 2021) for feature extraction and cosine similarity for matching. The embedding, $\mathbf{E}^{(L)}$, of the input image is compared with template embeddings, $\mathbf{E}^{(k)}_{\text{template}}$, corresponding to each transit sign class using cosine similarity:

$$S_k = \frac{\left\langle \mathbf{E}^{(L)}, \mathbf{E}^{(k)}_{\text{template}} \right\rangle}{\|\mathbf{E}^{(L)}\| \left\|\mathbf{E}^{(k)}_{\text{template}}\right\|}, \tag{2}$$

where S_k is the similarity score for class k, $\langle \cdot, \cdot \rangle$ denotes the dot product, and $\| \cdot \|$ is the Euclidean norm. The class with the highest similarity score is selected as:

$$k^* = \arg\max_k S_k. \tag{3}$$

To ensure reliable classification, a similarity threshold τ is applied, meaning a match is accepted only if:

$$S_{k^*} \geq \tau. \tag{4}$$

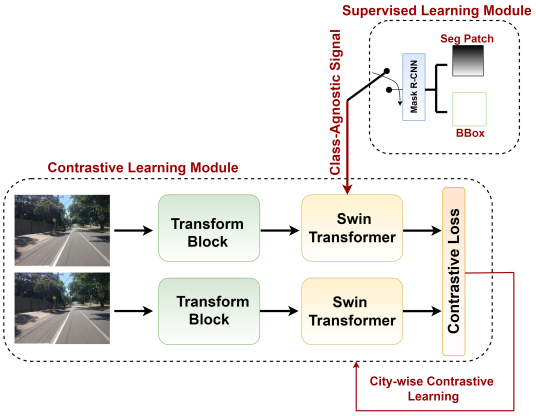
Choosing an appropriate threshold is crucial for balancing detection performance. As observed in our experiments, setting τ too low (below 0.5) results in excessive matches, increasing false positives. Conversely, increasing τ above 0.5 makes the classifier too restrictive, leading to a high rate of false negatives. The optimal value $\tau = 0.5$ achieves a balance between precision and recall, ensuring robust classification.

For feature extraction, we compared two pre-trained backbones: DINO (Caron et al., 2021) and CLIP (Radford et al., 2021). Both models were fine-tuned using contrastive learning on our custom Mapillary Traffic Sign dataset. DINO, a self-supervised Vision Transformer, provides solid visual representations. However, CLIP, trained with image-text pairs, outperformed DINO in capturing finer details, leading to higher precision in classification, as shown in our ablation studies.

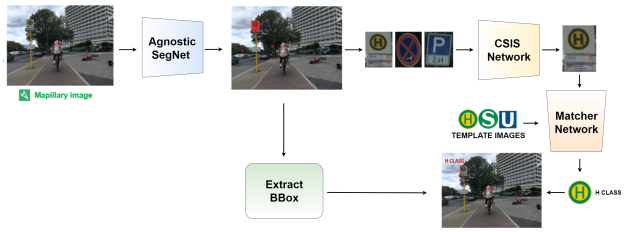
The final step computes the cosine similarity between the extracted features and template embeddings, assigning the class if the similarity exceeds the threshold. The complete computational process is described in Algorithm 1, and the architecture is shown in Figure 1(c), where dual ViT backbones are used for feature extraction, and cosine similarity is employed for classification.

Algorithm 1 Matcher Network Computational Process

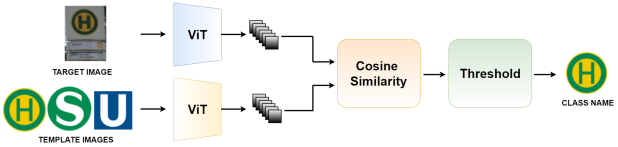
Require: Input image I, Template embeddings $\{\mathbf{E}^{(k)}_{\text{template}}\}_{k=1}^{K}$, Similarity threshold τ
Ensure: Predicted class k^* or negative detection
Step 1: Divide I into patches and compute input embedding $\mathbf{E}^{(L)}$ using ViT.
Step 2: **For** each class k **do**
  Compute similarity score $S_k = \dfrac{\left\langle \mathbf{E}^{(L)}, \mathbf{E}^{(k)}_{\text{template}} \right\rangle}{\|\mathbf{E}^{(L)}\| \left\|\mathbf{E}^{(k)}_{\text{template}}\right\|}$
End For
Step 3: Determine $k^* = \arg\max_k S_k$
Step 4: **If** $S_{k^*} \geq \tau$ **then**
  Return Predicted class k^*
Else
  Return Negative detection
End If



(a) Domain-Specific Agnostic Segmentation with Contrastive Learning.



(b) Overview of our framework for urban transit sign detection.



(c) Workflow of the Matcher Network using dual ViT backbones for feature extraction.

Figure 1: Illustration of our urban transit sign detection framework. (a) Shows our innovative Domain-Specific Agnostic Segmentation method, (b) presents the overall architecture, and (c) details the Matcher Network.

3. Experiments

We conducted extensive experiments to validate the effectiveness and scalability of our proposed method, using diverse datasets and state-of-the-art models.

3.1 Datasets

The performance of object detection models is highly dependent on the diversity and quality of the datasets used. We used three data sets in this study: the Mapillary Traffic Sign Dataset, the VISTAS Dataset, and our custom In-House Transit Sign Dataset. These datasets offer a complementary blend of global and localized perspectives for traffic and transit sign detection under various environmental conditions.

3.1.1 Mapillary Traffic Sign Dataset The Mapillary Traffic Sign Dataset (Ertler et al., 2020) includes more than 100,000 high-resolution images, with 52,000 fully annotated and 48,000 partially annotated images. The data set features over 300 traffic sign classes and covers various environmental conditions such as weather, seasons, and camera types, making it ideal for training models that need to generalize across multiple real-world scenarios.

3.1.2 VISTAS Dataset The VISTAS Dataset (Neuhold et al., 2017) contains approximately 25,000 images annotated with 124 semantic object categories. Unlike Mapillary, VISTAS captures a wide range of environmental conditions across various geographic regions, making it suitable for urban and suburban object detection tasks.

3.1.3 In-House Dataset Our custom In-House Transit Sign Dataset focuses on transit signs from ten major European cities, including London, Paris, Berlin, and Zurich. Comprising 2,300 manually annotated images, the dataset captures a diverse range of environmental conditions, lighting variations, and transit sign designs. Unlike existing datasets, which primarily focus on traffic signs, our dataset provides a dedicated benchmark for transit sign detection, addressing a crucial gap in urban object recognition. The images were collected using the Mapillary API and meticulously annotated with bounding boxes using the CVAT tool (CVAT.ai, 2024). To ensure broader applicability, the dataset accounts for regional disparities in transit sign designs, enabling robust model generalization across different urban settings. To support future research and development, the dataset will be publicly available at the following link: `https://www.kaggle.com/datasets/nsavioli/urbantransitsigns`

3.2 Evaluation Metrics and Implementation Details

We evaluated our urban sign detection model using three metrics: Intersection of Union (IoU), mean average precision (mAP) and precision, providing a thorough performance assessment. The framework was implemented using MMDetection (Chen et al., 2019), with Mask R-CNN and a Swin Transformer backbone for Domain-Specific Agnostic Segmentation (Figure 1(a)), and Faster R-CNN for Class-Specific Interference Suppression (CSIS).

The Two-Phase Method, as well as the entire framework (Figures 1(b) and 1(c)), were developed in PyTorch and trained on 8 NVIDIA RTX 3090 GPUs using the Adam optimizer (learning rate: 0.001, batch size: 100). The segmentation model was fine-tuned with AdamW, using random cropping and data augmentation to improve adaptability across different urban environments.

3.3 Limitations of Current State-of-the-Art Methods

Despite advancements in detection methods with limited data, current state-of-the-art models still encounter significant challenges in real-world scenarios. We conducted extensive experiments using both supervised learning (SL) and self-supervised learning (SSL) pre-training strategies to assess model performance across various object sizes and accuracy thresholds.

As shown in Table 1, pre-trained models with SSL techniques, such as the MoBY framework (Xie et al., 2021), were tested on datasets like the Munich street view images, VISTAS, and the Mapillary Traffic Sign Dataset (TSD). Although SSL in the Mapillary TSD dataset achieved the highest mAP scores in the validation set, these improvements did not consistently translate to practical applications. In real-world deployments, particularly in cities with varying transit sign designs, the models exhibited poor generalization and high false positive rates.

Figure 2 illustrates these limitations. Even with advanced SSL pretraining methods such as SSL (Mapillary TSD), the models struggled to accurately distinguish transit signs from other visually similar objects, resulting in numerous false positives. These results highlight that, while pretraining with large datasets improves validation performance, it is insufficient to address the complexity of detecting transit signs in diverse and complex urban environments.

Our findings demonstrate that relying solely on SSL and SL pretraining methods is not enough for robust transit sign detection across multiple cities. This emphasizes the need for more advanced methods, such as our proposed two-phase method, which overcomes these limitations and delivers more accurate and reliable detection results in real-world settings.

3.4 Effectiveness and Scalability of Our Proposed Two-Phase Method

To overcome the limitations of current detection methods, we have developed a scalable two-phase detection pipeline that significantly enhances the detection of transit signs in multiple cities. The scalability of our approach is inherently built into the design: By using a single, generalizable model, we eliminate the need to retrain or fine-tune the model for each new city, regardless of variations in transit sign designs. Instead, we directly apply our two-phase method, which adapts to different urban environments without additional training.

We evaluated the performance of our method using an internal dataset comprising transit signs from ten major European cities. The images were sourced from Mapillary and meticulously annotated to ensure diversity and representativeness. As shown in Figure 4, our two-phase method consistently outperformed state-of-the-art models such as Faster R-CNN with a Swin Transformer backbone (Liu et al., 2021), achieving higher mean Average Precision (mAP) and Intersection over Union (IoU) scores across all cities evaluated.

Tables 2 and 3 provide detailed comparisons of these improvements. For example, in Paris, our method achieved an IoU of 71% and an mAP of 81%, while the baseline model only achieved 6. 5% IoU and 1% mAP. Similar enhancements were observed in London, with significant increases in both IoU and mAP. These results highlight not only the effectiveness but also the inherent scalability of our approach. By eliminating the need for city-specific retraining or fine-tuning, our method significantly reduces deployment time and resource requirements, making it highly practical for real-world applications where transit signs vary between regions.

Figure 2: This figure shows two images from different cities: London (top) and Paris (bottom). It visualizes the results using three different methods: no pre-training (NP), SSL (self-supervised learning) using state-of-the-art (SOTA) detection methods, specifically SSL (Mapillary TSD, which is the Mapillary Traffic Sign Detection dataset), and our proposed Two-Phase Method. Even with the best pre-training methods, such as SSL (Mapillary TSD), we observe detection errors and false positives that are not present when using our approach. The pre-training methodology applied across different cities replicates the process used for Munich, as shown in Table 1. This demonstrates that even with the best SSL (Mapillary TSD) pre-training and SOTA detection methods, real-world performance can still suffer from false positives in complex urban settings, which our Two-Phase Method overcomes, ensuring more accurate and robust detection results.

Figure 3: Comparison of segmentation results between SAM 2 (Ravi et al., 2024) and our Domain-Specific Agnostic Segmentation with Contrastive Learning method (see Figure 1(a)). SAM 2 tends to over-segment, including irrelevant objects like traffic signs. Our method, by focusing on city-specific contrastive learning and fine-tuning, avoids this issue and achieves more accurate transit sign detection.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-G-2025-779-2025 | © Author(s) 2025. CC BY 4.0 License.

Metrics	NP	SL (VISTAS)	SSL (Munich SV)	SSL (VISTAS)	SSL (Mapillary TSD)	SSL (Munich+VISTAS)
bbox mAP	37.3%	38.1%	38.8%	39.5%	**39.9%**	38.9%
bbox mAP_50	51.9%	52.5%	53.6%	54.1%	**54.8%**	53.6%
bbox mAP_75	43.3%	44.3%	45.0%	**46.5%**	46.1%	45.5%
bbox mAP_s	17.3%	17.5%	17.5%	18.1%	**18.9%**	18.2%
bbox mAP_m	39.7%	40.6%	41.8%	42.5%	**42.8%**	41.7%
bbox mAP_l	56.6%	57.6%	58.4%	58.7%	**58.8%**	57.9%

Table 1: Performance comparison of object detection models using various pre-training strategies on validation data. NP = No Pretrain, SL = Supervised Learning, SSL = Self-Supervised Learning, VISTAS = VISTAS Dataset, SV = Street View, TSD = Traffic Sign Dataset. “NP” refers to models trained from scratch without pre-training, “SL (VISTAS)” denotes supervised learning pre-training on the VISTAS dataset, while the “SSL” variants use self-supervised learning techniques like MoBY (Xie et al., 2021) applied to datasets such as Munich street view images (randomly extracted from the city of Munich), VISTAS, and Mapillary. Although SSL on the Mapillary Traffic Sign Dataset achieves the highest mAP scores in validation, the SSL (Munich SV) results are very close, indicating that Munich street view images provide valuable information for pre-training. However, the presence of noisy or irrelevant images in the Munich dataset slightly impacted the performance, but only marginally. As seen in Figure 2, the left image shows results with NP, while the center image displays results from the best pre-training (SSL on Mapillary TSD). Even though SSL on Mapillary TSD uses the best pre-training methods and state-of-the-art detection networks, these gains do not consistently translate into better real-world results, as highlighted by the remaining detection challenges in complex urban settings and the presence of false positives.

City	Baseline [IoU]	Baseline + Pretraining [IoU]	Our Model [IoU]	Improvement (%)
London	3.5%	3.8%	**36%**	**+32.2%**
Paris	6.5%	7.0%	**71%**	**+64.0%**
Rome+Milan	1%	1.2%	**49%**	**+47.8%**
Lyon	0.9%	1.0%	**12%**	**+11.0%**
Edinburgh	2%	2.3%	**59%**	**+56.7%**
Berlin+Munich	1%	1.1%	**29%**	**+27.9%**
Vienna	2%	2.2%	**36%**	**+33.8%**
Zurich	0.2%	0.3%	**16%**	**+15.7%**

Table 2: Intersection over Union (IoU) comparison across various cities between the baseline detection model, the baseline model with pretraining, and our proposed two-phase method. IoU values are normalized and presented as percentages. The “Improvement” column represents the absolute increase in IoU achieved by our model compared to the best-performing baseline model for each city. These results were obtained by evaluating the models on our in-house dataset of transit signs from ten European cities, ensuring diverse urban conditions.

City	Baseline [mAP]	Baseline + Pretraining [mAP]	Our Model [mAP]	Improvement (%)
London	0.5%	0.6%	**50%**	**+49.4%**
Paris	1%	1.2%	**81%**	**+79.8%**
Rome+Milan	0%	0.1%	**27%**	**+26.9%**
Lyon	0%	0.1%	**52%**	**+51.9%**
Edinburgh	0.4%	0.5%	**34%**	**+33.5%**
Berlin+Munich	0.1%	0.2%	**17%**	**+16.8%**
Vienna	0.4%	0.5%	**47%**	**+46.5%**
Zurich	0%	0.1%	**7%**	**+6.9%**

Table 3: Mean Average Precision (mAP) comparison across various cities between the baseline detection model, the baseline model with pretraining, and our proposed two-phase method. mAP values are normalized and presented as percentages. The “Improvement” column represents the absolute increase in mAP achieved by our model compared to the best-performing baseline model for each city. These results highlight the significant performance gains achieved by our method in diverse urban settings.

Model	Precision
DINO (Caron et al., 2021)	0.83
CLIP (Radford et al., 2021)	0.96

Table 4: Precision comparison of the Matcher Network fine-tuned with DINO and CLIP as backbone models. CLIP achieves higher precision, leveraging its multimodal capabilities, demonstrating the advantage of using a multimodal backbone in diverse visual understanding tasks.

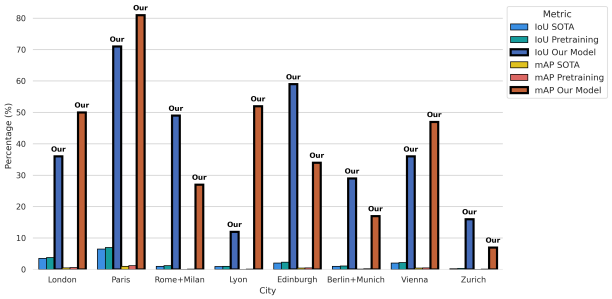


Figure 4: Comparison of performance between the baseline model (Faster R-CNN with Swin Transformer) and our proposed two-phase method. The chart highlights significant improvements achieved by our method across various cities, demonstrating its superior ability to handle complex and diverse urban environments.

Substantial improvements demonstrate that our two-phase method can effectively handle the variability in transit sign designs without additional training for each city. This inherent scalability addresses one of the main challenges in the deployment of detection systems in diverse urban environments, showcasing the practicality and robustness of our approach.

3.5 Ablation Study

To further validate our approach, we conducted an ablation study focusing on two key components of our method: the agnostic segmentation and the Matcher Network’s backbone.

3.5.1 Comparison with Existing Agnostic Segmentation Methods We compared our Domain-Specific Agnostic Segmentation method with the Segment Anything Model 2 (SAM 2) (Ravi et al., 2024), a leading model in the literature for agnostic segmentation.

As shown in Figure 3, our method provides superior detection of transit signs compared to SAM 2. While SAM 2 offers a generalized object segmentation approach, it often over-segments and includes irrelevant objects like traffic signs. This over-segmentation occurs because SAM 2 relies on grid-based embeddings and requires prompts to generate segmentation masks, which may not be precise enough for small, distant objects such as transit signs. Additionally, SAM 2 struggles with small objects due to its architectural design and dependence on prompts, making it unsuitable for scenarios where precise sign detection is required.

In contrast, our Domain-Specific Agnostic Segmentation with Contrastive Learning approach, as illustrated in Figure 1(a), eliminates the need for prompts and focuses on city-specific contrastive learning. By first applying contrastive learning to each city’s unique urban environment, followed by a supervised signal that only considers segmentation and regression losses, our method adapts more effectively to real-world urban settings. This allows the model to better detect small and distant objects without over-segmenting irrelevant features.

Unlike SAM 2, our method iteratively trains on multiple cities, generalizing across diverse urban environments. This leads to a more accurate and scalable solution for transit sign detection. Our ablation study shows that even though SAM 2 is considered a leading method in the literature, it is not sufficient for the specific challenges of detecting small transit signs in complex urban environments. Our Domain-Specific Agnostic Segmentation with Contrastive Learning proves to be a more robust and effective approach for this task.

3.5.2 Impact of Backbone Selection in the Matcher Network A key component of our Matcher Network is the feature extractor, which generates embeddings for both the detected sign and the template images. These embeddings are compared using cosine similarity to determine a match. To evaluate which backbone performs best for this task, we experimented with different pre-trained models.

We compared two types of backbones: DINO v1 (Caron et al., 2021), a vision transformer model pre-trained using self-supervised learning solely on visual data, and CLIP (Radford et al., 2021), a multimodal model pre-trained on a large dataset of image-text pairs, integrating both visual and textual modalities.

To ensure a fair comparison and isolate the influence of the backbone architecture, we fine-tuned both models on our custom Mapillary Traffic Sign dataset using contrastive learning, adapting them to our specific domain.

As shown in Table 4, CLIP achieved a higher matching precision of 94%, outperforming DINO v1, which reached 83%. This result suggests that the multimodal pre-training of CLIP, which incorporates language information, enhances its ability to extract more discriminative visual features for matching. Even though the matching process involves only visual features, the pre-training with language data enriches CLIP’s feature space, enabling better differentiation between subtle variations in transit signs.

Our ablation study confirms that using a multimodal backbone like CLIP enhances the Matcher Network’s capability to accurately match transit signs, highlighting the unexpected benefit of multimodal pre-training in visual matching tasks.

4. Conclusion and Future Work

We have developed a scalable two-phase detection framework that effectively integrates domain-specific agnostic segmentation with a matcher network, enabling precise detection of public transit signs with minimal training data requirements. Unlike previous studies that rely on highly curated datasets, our evaluation was conducted in real-world scenarios, using images collected from multiple European cities with varying environmental conditions, occlusions, and sign variability. This ensures that our approach is robust to real-world deployment challenges rather than being optimized solely for idealized datasets. Our framework demonstrated significant improvements in mean Average Precision (mAP) and robustness, setting a new benchmark for public transit sign detection.

Additionally, we addressed a major gap in the field by providing a dataset of 2,300 manually annotated transit sign images,

offering a valuable and realistic resource for further research in urban object detection.

Looking ahead, we plan to enhance our model's adaptability by incorporating adaptive decision networks to further reduce false positives and improve detection accuracy in complex urban scenes. We also envision extending our framework to broader applications such as smart city infrastructure, real-time urban mapping, and augmented reality navigation.

Furthermore, to make our dataset a global benchmark, we plan to expand its coverage beyond Europe, incorporating transit signs from multiple continents and diverse urban environments. This worldwide dataset extension will facilitate the development of more generalizable and robust models, reducing geographic biases and ensuring scalability across different transit systems.

References

Ashraf, A., Idrisi, M. J., 2024. Smart and Sustainable Public Transportation-A Need of Developing Countries. *International Journal of Networked and Distributed Computing*, 12(1), 144–152.

Bar, A., Wang, X., Kantorov, V., Reed, C. J., Herzig, R., Chechik, G., Rohrbach, A., Darrell, T., Globerson, A., 2022. Detreg: Unsupervised pretraining with region priors for object detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14605–14615.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A., 2021. Emerging properties in self-supervised vision transformers. *Proceedings of the IEEE/CVF international conference on computer vision*, 9650–9660.

Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., Zhang, Z., Cheng, D., Zhu, C., Cheng, T., Zhao, Q., Li, B., Lu, X., Zhu, R., Wu, Y., Dai, J., Wang, J., Shi, J., Ouyang, W., Loy, C. C., Lin, D., 2019. Mmdetection: Open mmlab detection toolbox and benchmark.

CVAT.ai, 2024. Cvat: Computer vision annotation tool. `https://github.com/cvat-ai/cvat`. Accessed: 2024-09-26.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An image is worth 16x16 words: Transformers for image recognition at scale.

Ertler, C., Mislej, J., Ollmann, T., Porzi, L., Neuhold, G., Kuang, Y., 2020. The mapillary traffic sign dataset for detection and classification on a global scale. *European Conference on Computer Vision*, Springer, 68–84.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2018. Mask r-cnn.

Kim, C., Han, W., Ju, D., Hwang, S. J., 2024. Eagle: Eigen aggregation learning for object-centric unsupervised semantic segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3523–3533.

Liu, C., Wang, K., Lu, H., Cao, Z., Zhang, Z., 2022. Robust object detection with inaccurate bounding boxes. *European Conference on Computer Vision*, Springer, 53–69.

Liu, L., Porr, A., Miller, H. J., 2023. Realizable accessibility: evaluating the reliability of public transit accessibility using high-resolution real-time data. *Journal of Geographical Systems*, 25(3), 429–451.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows.

Malpani, V., Shukla, S., Gyanchandani, M., Shrivastava, S., 2023. Customized cnn for traffic sign recognition using keras pre-trained models. *International Conference On Innovative Computing And Communication*, Springer, 91–98.

Neuhold, G., Ollmann, T., Rota Bulo, S., Kontschieder, P., 2017. The mapillary vistas dataset for semantic understanding of street scenes. *Proceedings of the IEEE international conference on computer vision*, 4990–4999.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I., 2021. Learning transferable visual models from natural language supervision.

Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K. V., Carion, N., Wu, C.-Y., Girshick, R., Dollár, P., Feichtenhofer, C., 2024. Sam 2: Segment anything in images and videos.

Ren, S., He, K., Girshick, R., Sun, J., 2016. Faster r-cnn: Towards real-time object detection with region proposal networks.

Siméoni, O., Puy, G., Vo, H. V., Roburin, S., Gidaris, S., Bursuc, A., Pérez, P., Marlet, R., Ponce, J., 2021. Localizing objects with self-supervised transformers and no labels. *arXiv preprint arXiv:2109.14279*.

Wang, Y., Chen, M., Tang, S., Zhu, F., Yang, H., Bai, L., Zhao, R., Yan, Y., Qi, D., Ouyang, W., 2022. Unsupervised Object Detection Pretraining with Joint Object Priors Generation and Detector Learning. *Advances in Neural Information Processing Systems*, 35, 12435–12448.

Wang, Y., Shen, X., Yuan, Y., Du, Y., Li, M., Hu, S. X., Crowley, J. L., Vaufreydaz, D., 2023. Tokencut: Segmenting objects in images and videos with self-supervised transformer and normalized cut. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Xie, Z., Lin, Y., Yao, Z., Zhang, Z., Dai, Q., Cao, Y., Hu, H., 2021. Self-supervised learning with swin transformers. *arXiv preprint arXiv:2105.04553*.

Yang, S., Bo, L., Wang, J., Shapiro, L., 2012. Unsupervised template learning for fine-grained object recognition. *Advances in neural information processing systems*, 25.

Zhang, Z., Sabuncu, M., 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31.

Zhu, Y., Yan, W. Q., 2022. Traffic sign recognition based on deep learning. *Multimedia Tools and Applications*, 81(13), 17779–17791.

Zong, Z., Song, G., Liu, Y., 2023. Detrs with collaborative hybrid assignments training.