International Journal of
Geo-Information

Article

Exploration of an Open Vocabulary Model on Semantic Segmentation for Street Scene Imagery

Zichao Zeng * and Jan Boehm

Department of Civil, Environmental and Geomatic Engineering, University College London, London WC1E 6BT, UK; j.boehm@ucl.ac.uk
* Correspondence: zichao.zeng.21@ucl.ac.uk

Abstract: This study investigates the efficacy of an open vocabulary, multi-modal, foundation model for the semantic segmentation of images from complex urban street scenes. Unlike traditional models reliant on predefined category sets, Grounded SAM uses arbitrary textual inputs for category definition, offering enhanced flexibility and adaptability. The model's performance was evaluated across single and multiple category tasks using the benchmark datasets Cityscapes, BDD100K, GTA5, and KITTI. The study focused on the impact of textual input refinement and the challenges of classifying visually similar categories. Results indicate strong performance in single-category segmentation but highlighted difficulties in multi-category scenarios, particularly with categories bearing close textual or visual resemblances. Adjustments in textual prompts significantly improved detection accuracy, though challenges persisted in distinguishing between visually similar objects such as buses and trains. Comparative analysis with state-of-the-art models revealed Grounded SAM's competitive performance, particularly notable given its direct inference capability without extensive dataset-specific training. This feature is advantageous for resource-limited applications. The study concludes that while open vocabulary models such as Grounded SAM mark a significant advancement in semantic segmentation, further improvements in integrating image and text processing are essential for better performance in complex scenarios.

Keywords: street view; semantic segmentation; foundation models; open vocabulary; multi-modal AI; GeoAI

check for updates

Citation: Zeng, Z.; Boehm, J. Exploration of an Open Vocabulary Model on Semantic Segmentation for Street Scene Imagery. *ISPRS Int. J. Geo-Inf.* **2024**, *13*, 153. https://doi.org/10.3390/ijgi13050153

Academic Editors: Wolfgang Kainz, Hartwig H. Hochmair, Hao Li and Levente Juhász

Received: 29 February 2024
Revised: 26 April 2024
Accepted: 2 May 2024
Published: 5 May 2024

1. Introduction

Street view imagery, rich in geospatial information, has emerged as a key resource for urban analysis and applications [1–3]. Scraping and understanding visual elements at the pixel level from street view imagery considerably affects many geospatial applications, due to their abundant real-world semantics [4,5]. The ability to understand street scenes and extract environmental components rapidly and accurately can significantly advance fields such as autonomous vehicles, health and well-being, greenery, urban morphology, transportation, and human mobility [2,6,7].

Recently, the advent of Large Language Models (LLMs) has ushered in a new era of possibilities for geospatial science. LLMs, exemplified by models like GPT-4, have the capacity to understand and generate text-based descriptions of geospatial data, thus enhancing geospatial analyses [8,9]. These models are proving instrumental in AI-based Spatial Data Analysis, enabling the automated interpretation of geospatial information from text data [8–11].

Multi-modal AI models combined with LLMs, including Contrastive Language-Image Pre-training (CLIP) [12] and Bootstrapped Language-Image Pre-training (BLIP) [13,14], have expanded the horizons of Geospatial Artificial Intelligence (GeoAI). These large multi-modal models align textual descriptions with visual data, facilitating a deeper understanding of geospatial context [7]. GeoAI leverages these models to extract geospatial insights from a wide range of sources, paving the way for more comprehensive urban greenery assessments and mobility solutions [11].

Semantic segmentation, which involves the classification of images at the pixel level, plays a vital role in capturing and processing the information of a user's surroundings [15–17]. With this pixel level classification, the system can better understand the various objects and scenes in the street view image such as roads, traffic signs, pedestrians, vehicles, buildings, etc. [6,18].

Before the widespread adoption of deep learning, segmentation primarily relied on hand-crafted techniques such as pixel colour, Histogram of Oriented Gradients (HOG), and Scale-Invariant Feature Transform (SIFT) [19]. However, the field of semantic segmentation underwent a transformation with the advent of deep learning. Fully Convolutional Networks (FCNs) marked a significant milestone by enhancing pixel-level classification in segmentation tasks, integrating transposed convolutional layers into an end-to-end trainable architecture [20].

The subsequent emergence of innovative models, including SegNet [21], U-Net [22], DeepLab [23], and PSPNet [24], further propelled the capabilities of semantic segmentation. U-Net, known for its unique skip connections and symmetric encoder–decoder structure, achieved notable success in medical image segmentation [22,25]. DeepLab, on the other hand, enhanced segmentation accuracy by introducing dilated convolutions to expand the learnable context [23,26–28]. These techniques proved invaluable in urban imagery understanding.

Despite the remarkable accuracy of deep learning methods, they encounter challenges related to training datasets, domain adaptation, and object-specific segmentation. For instance, traffic signs may exhibit variations between regions, necessitating domain adaptation techniques [29]. Moreover, semantic segmentation in street view scenarios often requires the identification of specific objects, such as "pedestrians" and "riders" rather than generic "people" [30].

This is where multi-modal foundation models in geographic context come into play. Because their vast training datasets include diverse geo-information and -location, these large models have a capacity for geographical understanding. This addresses the geographical limitations often encountered by traditional deep learning models when applied to street scene understanding [30–32]. They can adapt seamlessly to diverse urban environments and have the capability to detect specific targets, such as "pedestrians crossing the road", "cars with flat tires", and "no passing signs" [30,32].

In open vocabulary tasks, the inputs are not only images but also text describing the targeted set of classes, referred to as a "prompt" in Natural Language Processing (NLP) [30,32,33]. Prompts are texts, queries, or descriptive instructions input to language-based models to complete user-expected tasks. Slight differences in input prompts would cause entirely different outputs, so careful text design is needed in open vocabulary tasks [34]. Therefore, exploring appropriate prompts (i.e., words for expected classes) is also a key to this study.

Recent advancements in open vocabulary multi-modal models have shown impressive results in many zero-shot tasks. Grounding DINO, a model designed for open-vocabulary object detection, seeks to extend the understanding of open-set concepts by integrating both language and visual modalities [30]. The Segment Anything Model (SAM), introduced in a promptable model for segmentation, excels in generating high-quality masks from just a single foreground point and demonstrates robust performance across various downstream tasks through prompt engineering [35]. Grounded SAM, which merges the Grounding DINO and SAM models, can detect and segment relevant regions in images based on arbitrary textual inputs from users [36].

The purpose of this study is to investigate the performance of open vocabulary models in the task of the semantic segmentation of street scene imagery. Four street scene datasets from open benchmarks were used for testing using Grounded SAM [36]. The specific objectives were to (1) validate the ability of the open vocabulary model to reason directly about the visual semantics of street scenes without training; (2) explore the performance of the open vocabulary model for different text prompts and the segmentation results of the open vocabulary model for individual confusable categories; and (3) compare the performance to SOTA models on the four benchmark datasets.

2. Data and Task

2.1. Dataset

To evaluate the zero-shot capabilities of open vocabulary models, we selected benchmark datasets that were not used in the training of the pre-trained models. Grounding DINO was trained on datasets including Object365, GoldG, Cap4M, OI, and RefC, while Segment Anything was trained on a specially created dataset [30,31,37]. Consequently, we choose Cityscapes [18], BDD100K [38], KITTI [39], and GTA5 [40] for testing to avoid data leakage.

2.1.1. Cityscapes

The Cityscapes dataset focuses on the semantic understanding of urban street scenes, providing a large number of diverse stereoscopic video sequences from 50 different cities [18]. It incorporates 19 pixel-level semantic categories, namely Road, Sidewalk, Building, Wall, Fence, Pole, Traffic Light, Traffic Sign, Vegetation, Terrain, Sky, Pedestrian, Rider, Car, Truck, Bus, Train, Motorbike, and Bike [18]. We utilised the validation set of Cityscapes to explore the detailed performance of the open vocabulary model.

Due to the high-quality annotation within Cityscapes, numerous studies have utilised it for research, making a vast array of reference results available. Consequently, our analysis will primarily engage with, compare, and discuss the Cityscapes dataset. Given the general applicability and usefulness of the categories defined in Cityscapes, many datasets have adopted similar categories. Hence, we have additionally selected three other Cityscapes-style datasets for comprehensive testing, to validate the performance in different conditions.

2.1.2. Additional Test Datasets

In addition to Cityscapes, BDD100K [38], KITTI [39], and GTA5 [40] were selected for further testing. While akin to Cityscapes in terms of the categorisation into 19 classes, these datasets offer unique perspectives owing to varied data collection conditions and environments. This diversity enriches the testing landscape, providing a more rounded evaluation of the model's capabilities.

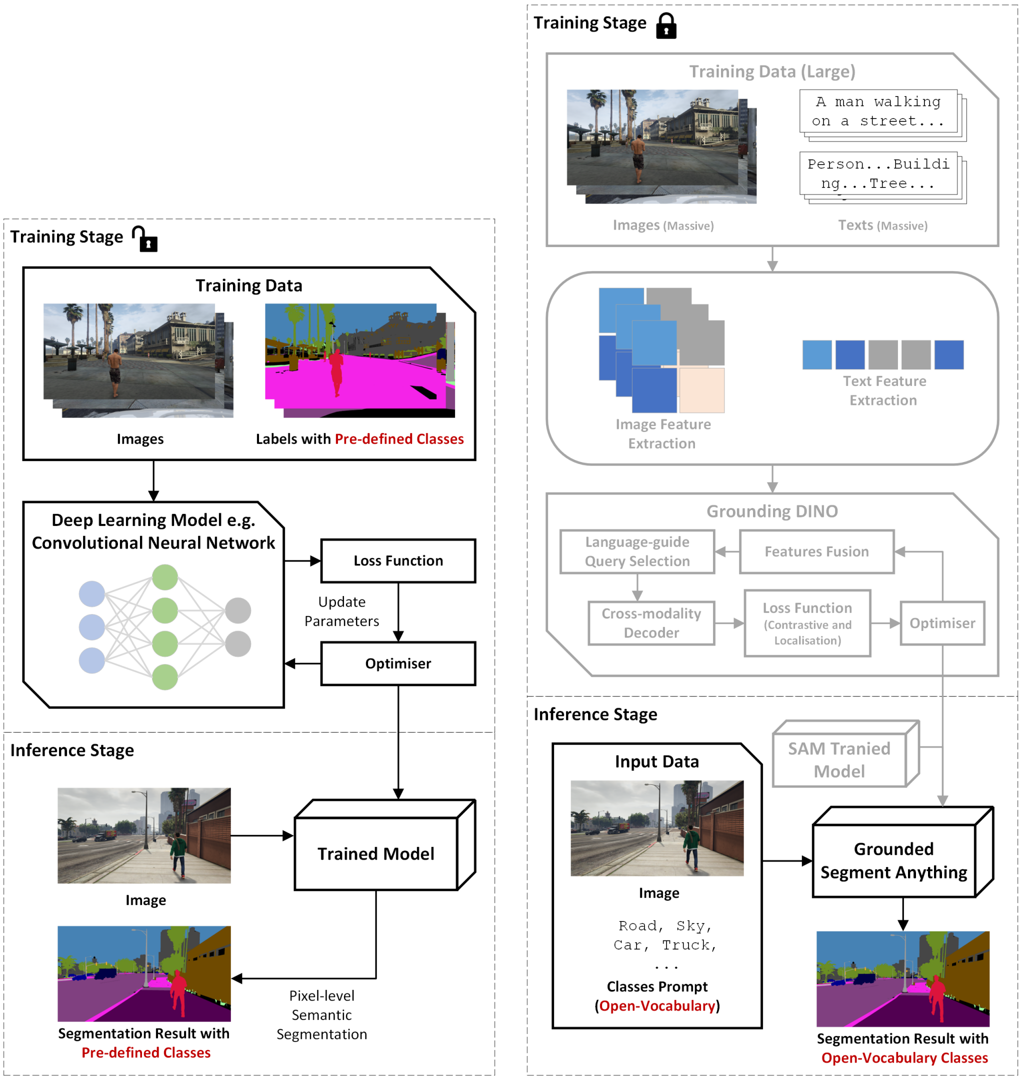
BDD100K consists of driving views and labels of various types for multitask learning. It was captured in the United States of America and its semantic segmentation set is Cityscapes-like [38]. KITTI, with its focus on both urban and rural environments in Europe, contains a Cityscapes-like semantic segmentation set which is tested in this study [39]. This offers an enriched context for model evaluation including urban and rural environments and different geographical locations. GTA5, a synthetic dataset derived from a video game engine, presents high-resolution images across a spectrum of simulated urban environments [40]. Its inclusion in the testing regimen is pivotal for assessing the model's adaptability from controlled, synthetic scenarios to real-world situations, a critical aspect of model generalisation.

The rationale for using these additional datasets alongside Cityscapes is their complementary nature. Each introduces different variables, such as varied lighting conditions and weather scenarios, and contrasts between synthetic and real imagery. This variety not only challenges but also reinforces the model's ability to generalise and perform accurately in diverse urban settings. Therefore, testing across these datasets offers a comprehensive evaluation of the semantic segmentation model's robustness and effectiveness.

2.2. Task Definition

2.2.1. Traditional Semantic Segmentation

Traditional semantic segmentation for images typically involves training a model such as a CNN on datasets similar to the one seen in Figure 1a. At the training stage, the training images are learned and their semantic labels with pre-defined classes are used to update the model's parameters. After incremental parameter optimisation, the model peaks with the best performance on the trained dataset. In the inference stage, the trained model receives an image which is of a similar scene to the training data. A segmentation result within the pre-defined classes is output.



(**a**) (**b**)

Figure 1. The framework of models in semantic segmentation tasks including training stage and inference stage. (**a**) Traditional deep learning model. (**b**) Grounded SAM. The unlocked symbol means that time must be spent on training; the locked symbol and grey boxes mean this part is unchangeable.

Traditional semantic segmentation tasks are usually based on a fixed, predefined vocabulary (or set of categories). Given an image I with size $W \times H$ (where W is the width of the image and H is the height of the image), the goal is to assign a category label $L_{i,j}$ to each pixel (i, j) (where $i \in [1, W]$ and $j \in [1, H]$).

Thus, we can define a mapping function $f : I \to L$, where I is an input image and L is a labelling matrix of size $W \times H$, where each element $L_{i,j}$ is a selection of labels from a predefined set of categories C; thus, $L_{i,j} \in C$.

The function f is usually a deep neural network that accepts an input of image I and outputs a labelling matrix L. The parameters of f are then updated iteratively by a loss function $\mathcal{L}$ that compares the ground truth L^*, i.e., the true label of I, with L from f, until L is closer to L^*. Normally, $\mathcal{L}$ is used to evaluate the prediction performance from the true labels.

2.2.2. Open Vocabulary Semantic Segmentation

Taking Grounded SAM as the example of open vocabulary semantic segmentation, the training data are normally not images only, as seen in Figure 1b. Besides the labelled images, textual data and image–text pairs are used during training. The model architecture is more complex, often not only a single network for visual learning but also additional structures for language processing and image–text fusion. In the inference stage, because the model has learned natural language and is able to understand text, the expected semantics are not limited to a fixed set of categories but have an open vocabulary. The input data are a proposed image and a set of expected classes. The output is the segmentation result within the open vocabulary list the user has provided.

Open vocabulary image understanding combines image and text information to deal with object categories that are not in a pre-defined vocabulary. Given an image I of size $W \times H$ and a set of several categories $T = \{t_1, t_2, \ldots, t_N\}$ (where N is the number of category texts) whose representation is Text, the aim is to segment each pixel (i, j) (where $i \in [1, W]$ and $j \in [1, H]$) to a category label $L_{i,j}$, and this category label is included in the prompted classes T. The mapping function is $f : (I, T) \to L$, where I is an input image and L is a labelling matrix of size $W \times H$ where each element $L_{i,j}$ is a selection of labels from the prompted categories T; thus, $L_{i,j} \in T$.

In open vocabulary semantic segmentation tasks, the function f is not a single network but a group of models. f consists of at least the following: (1) an image encoder for extracting image features; (2) a text encoder for extracting text features; and (3) a fusion mechanism that combines the image and text features to generate a category label for each pixel. The inputs are an image I and a labelling set T, and the output is a labelling matrix L ($L \in T$). The loss function $\mathcal{L}$ is applied to update the weights, but in multi-modal learning, a contrastive loss between predicted objects and language tokens for classification is implemented in addition to a loss function for visual modality. However, in this study, we focused on the evaluation of the inference performance of the open vocabulary model for the benchmark datasets without any training. We used common evaluation metrics to quantify the model performance by comparing L prompted from T to ground truths L^*.

3. Methodology

3.1. Framework of Implementation and Evaluation of Grounded SAM

To quantify the performance of open vocabulary models in the semantic segmentation of street views, we use four established image benchmark datasets and segment them with Grounded SAM. We then refine the text prompts for Grounded SAM based on our error analysis. Finally, we evaluate the performance on the benchmarks. The workflow shown in Figure 2 is structured into three stages: Initial Inference, Prompt Improvement, and Benchmarking.

Initially, we use the category names T_1, defined in the Cityscapes dataset, as text prompts. We apply single-word text prompts for each category to each input image I from the benchmark dataset. Using the pre-trained open vocabulary model Grounded SAM, we generate segmented images. We then compare these segmented outputs to the ground truth labels from the benchmarks. This phase involves analysing the direct inference results without any prompt tuning. We conduct experiments on each category separately, thus solving a series of bicategorical tasks. This explores the segmentation capabilities of Grounded SAM for each category separately. We follow this with multi-category experiments where all predefined categories T_1 are combined together in a single text prompt, with category words separated by commas. Effectively, the category sets T_i

are equivalent to the text prompts. For these multi-category experiments, we can generate confusion matrices which can help in identifying different types of errors.

Upon analysing the confusion matrices together with the input images, we discern two main error categories: textual confusion and image-based confusion. Considering that changing prompts can lead to different results (see Section 1), multiple prompts have been designed. We mitigate the detected error types by refining the set of category names: T_2 for fixing a textual error, T_3 for fixing a visual error, and T_4 for both. Again, these refined names are concatenated to form text prompts (separated by commas). The enhanced text prompts are then deployed with the model to produce new segmentation results. The effectiveness of the enhanced prompts is validated through comparison with ground truth data. We both visually and quantitatively checked the improvement of the new results over the initial outputs.

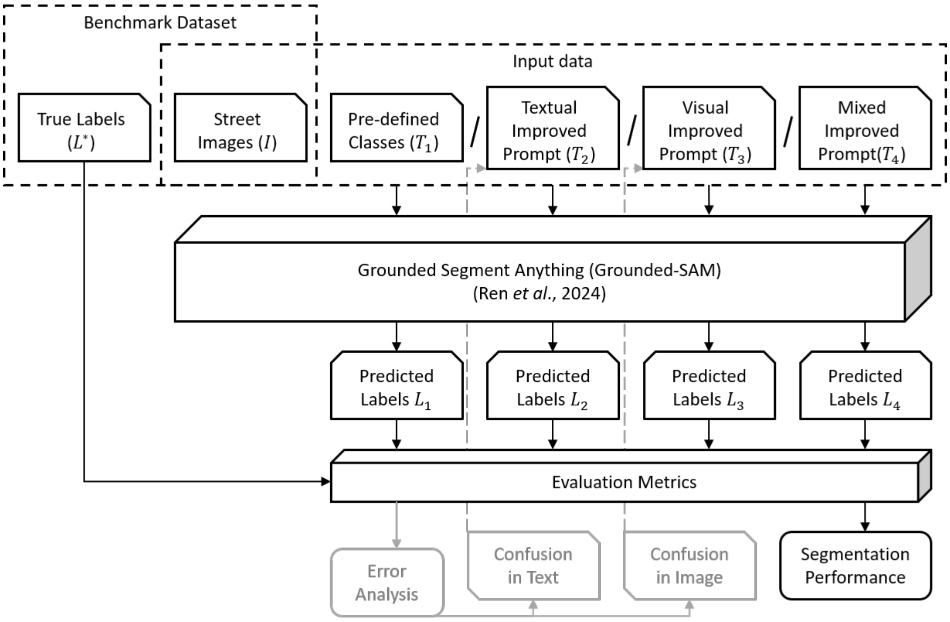


Figure 2. The workflow of implementing Grounded SAM [36] on benchmark datasets and evaluating segmented performance.

Finally, we compare outcomes derived from the open vocabulary model with state-of-the-art segmentation methods on the same benchmark datasets. This comparison clarifies the advantages and limitations of open vocabulary models in the context of street view image segmentation.

3.2. Prompt Design

To assess the impact of prompts on segmentation results, we modify the names of the corresponding categories, focusing on both the nature of the errors and the visual characteristics of the categories. The analysis of the preliminary results in Section 4 reveal distinct visual differences between Traffic Light (t_7) and Traffic Sign (t_8) and between Person (t_{12}) and Rider (t_{13}). However, Bus (t_{16}) and Train (t_{17}) present challenges in classification, even manually. This leads to the design of three different types of prompts besides the pre-determined names (T_1) by Cityscapes seen in Table 1:

- Prompt T_2: Designed to address confusions not related to the visual level. This prompt aims to rectify errors arising from textual misunderstandings or misclassifications that do not stem from visual similarities.
- Prompt T_3: Tailored to confusions at the visual level. This prompt is particularly focused on addressing the challenges in distinguishing visually similar categories, such as trains and buses.

- Prompt T_4: A combination prompt, developed to tackle both textual and visual confusions. This prompt incorporates elements from both T_2 and T_3 to provide a more comprehensive solution to the segmentation errors.

Additionally, within T_3, multiple prompt designs were considered, shown in Table 2. The classes Train (t_{16}) and Bus (t_{17}) are challenging to segment due to their visual similarity. This approach acknowledges the complexity of differentiating between certain categories and seeks to refine the model's accuracy through tailored prompt adjustments.

Table 1. Prompt design (i.e., input item of classes) for street view semantic segmentation using Grounded Segment Anything. **Bold** represents improved prompts to avoid textual misunderstandings, and <u>underline</u> represents improved prompts in addressing visual feature similarity.

Class	Pre-Defined Classes from Cityscapes (T_1)	T_2	T_3	T_4
t_1	Road	Road	Road	Road
t_2	Sidewalk	Sidewalk	Sidewalk	Sidewalk
t_3	Building	Building	Building	Building
t_4	Wall	Wall	Wall	Wall
t_5	Fence	Fence	Fence	Fence
t_6	Pole	Pole	Pole	Pole
t_7	Traffic Light	**Signal Light**	Traffic Light	**Signal Light**
t_8	Traffic Sign	**Signpost**	Traffic Sign	**Signpost**
t_9	Vegetation	Vegetation	Vegetation	Vegetation
t_{10}	Terrain	Terrain	Terrain	Terrain
t_{11}	Sky	Sky	Sky	Sky
t_{12}	Person	**Pedestrian**	Person	**Pedestrian**
t_{13}	Rider	Rider	Rider	Rider
t_{14}	Car	Car	Car	Car
t_{15}	Truck	Truck	Truck	Truck
t_{16}	Bus	Bus	Bus	Bus
t_{17}	Train	Train	<u>Tram/...</u>	<u>Tram</u>
t_{18}	Motorbike	Motorbike	Motorbike	Motorbike
t_{19}	Bike	Bike	Bike	Bike

Table 2. Additional category enhancements for confused categories caused by image feature similarity.

Prompt T	Class t_{16}	Class t_{17}
T_1	Bus	Train
T_3 (Train → On Rails)	Bus	On Rails
T_3 (Train → Locomotive)	Bus	Locomotive
T_3 (Train → Streetcar)	Bus	Streetcar
T_3 (Train → Tram)	Bus	Tram
T_3 (Bus → Coach)	Coach	Train
T_3 (Bus → Bus on Roads and Train → Train on Rails)	Bus on Roads	Train on Rails

The strategies used for prompt enhancement in this study are as follows: (1) Avoiding similar vocabulary for different categories, e.g., Traffic Light (t_7) and Traffic Sign (t_8) → Signal Light (t_7) and Signpost (t_8). (2) Specialising vocabulary, e.g., Person (t_{12}) → Pedestrian (t_{12}).

In summary, these prompt designs are integral to improving the segmentation accuracy of open vocabulary models. By addressing both textual and visual confusions, the model's performance in segmenting complex street scenes can be significantly enhanced. Section 4 will delve into the impact of these prompts in greater detail.

3.3. Experimental Setting

One of the benefits of open vocabulary models is that they allow inference on a particular dataset without any training on that dataset. The required resources at the inference stage are considerably lower than at the training stage.

The experiments were conducted on an Ubuntu server with Intel® Core™ i7-6850K CPU @ 3.60 GHz, 64 GB random-access memory, and a single GeForce RTX 2080 Ti with 11 GB memory. The Grounded SAM by Ren et al. [36] is implemented based on the PyTorch framework with PyTorch version 1.13 and CUDA version 11.7. Based on the above configuration, the inference time for 19 classes is around 20 s per image.

We keep the hyper-parameters of Grounded SAM to their default settings. Specifically, the box threshold is 0.25, the text threshold 0.25, and the Non-Maximum Suppression (NMS) threshold is 0.8. Initial experiments are conducted independently for each category defined in Cityscapes ($T \in T_1$). We apply single-word text prompts for each category to each input image (I). This explores the segmentation capabilities of Grounded SAM for each category separately. In the multi-category experiments, all predefined categories (T_1) are combined together in a single text prompt, with category words separated by commas. The textual improved prompt (T_2), visual improved prompt (T_3), and hybrid improved prompt (T_4) also contain all categories in a single text prompt.

3.4. Evaluation Metrics

To effectively assess the performance of semantic segmentation in image processing, the following evaluation metrics are employed:

- **Precision and Recall** are common metrics used for classification tasks [41]. Precision is the proportion of positive predictions that are true positives, i.e., how many of the positive predictions are correct. Recall is the proportion of all true samples that are correctly predicted, which assesses the model's ability to identify positive samples and how many positive samples are missed.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

  where TP is the number of true samples that are predicted as positive. FP is the number of false samples that are predicted as positive. FN is false negative samples.
- **Intersection over Union (IoU)** is typically used in segmentation tasks. It expresses the ratio of the intersection to the combination of predicted results and ground truth for a single class [42]. **Mean IoU** is the average value of IoU for all classes.

$$IoU_c = \frac{Area\ of\ Overlap}{Area\ of\ Union} = \frac{TP_c}{TP_c + FP_c + FN_c} \tag{3}$$

$$mIoU = \frac{1}{C}\sum_{c=1}^{C} IoU_c \tag{4}$$

  where C means the number of classes.

4. Experimental Results

4.1. Results for Individual Category Segmentation

This subsection presents the performance evaluation of Grounded SAM across individual categories (from t_1 to t_{19}). Our quantitative results are presented in Table 3. Some visual results are shown in Figure 3. An aggregate analysis of the results from four datasets indicates that the segmentation performance is exceptionally strong in most categories, particularly notable in Road (t_1), Building (t_3), Vegetation (t_9), Sky (t_{11}), Person (t_{12}), Car (t_{14}), Truck (t_{15}), Bus (t_{16}), and Mobile Bike (t_{18}), where the recall rates exceed 70%. This suggests that these categories are well-recognised by the model. However, apart from Road (t_1), the precision and IoU for other categories are somewhat disappointing, especially for the category Train (t_{17}), where the precision and IoU reach only 10% in the Cityscapes, BDD100K, and GTA5 datasets. The model's performance in other categories is reasonably good, with recall rates of around 50%, though the precision and IoU for Sidewalk (t_2) and Terrain (t_{10}) are relatively low. This performance pattern is somewhat expected. Considering that each category is treated as a binary classification task (presence or absence) in an open vocabulary model, some ambiguities in target definitions are likely. Overall, Grounded SAM shows robust segmentation capabilities across these categories without the necessity for category-specific training. This demonstrates the model's proficiency in handling a variety of urban elements, though it also highlights areas where further refinement, such as prompt design, could enhance performance.

Figure 3. Examples of the segmented prediction for each category in independent binary classification experiment. Red mask represents the ground truth; blue mask represents the prediction result; and purple mask represents the correct predicted result.

Table 3. Semantic segmentation performance on four benchmark datasets for each category in independent binary classification experiment.

Dataset	Road	S.walk	Build.	Wall	Fence	Pole	T.Light	T.Sign	Veget.	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	M.bike	Bike
Precision in %																			
Cityscapes val.	81.34	13.58	78.16	5.77	41.88	79.45	80.33	55.06	86.19	1.99	57.09	74.72	22.85	95.69	49.87	71.01	12.64	53.80	79.04
BDD100K val.	78.74	12.02	71.77	26.62	25.97	42.31	72.90	25.60	61.33	4.77	87.52	61.72	15.67	61.73	44.39	54.07	1.20	31.60	83.02
KITTI	91.16	26.87	78.48	16.30	30.70	78.68	76.80	23.64	83.73	14.05	94.33	73.10	43.61	96.23	24.20	70.20	60.04	56.54	59.96
GTA5	88.61	18.33	75.20	17.57	5.01	62.08	80.45	18.97	76.27	3.08	99.63	69.64	38.57	19.33	20.41	59.12	18.30	31.66	71.24
Recall in %																			
Cityscapes val.	97.04	69.03	88.21	65.71	66.85	33.95	69.37	57.28	70.15	55.49	95.52	83.05	61.52	87.05	92.73	92.23	85.17	80.78	73.36
BDD100K val.	94.46	73.08	92.70	93.44	73.23	75.35	84.19	71.11	79.75	52.41	95.70	86.07	85.77	89.99	92.39	96.77	13.36	71.94	85.88
KITTI	92.89	69.93	86.38	51.57	67.14	47.61	69.78	66.74	77.20	59.02	95.55	72.00	70.49	88.26	79.33	88.73	93.15	85.42	63.60
GTA5	84.02	74.41	88.43	80.29	57.83	66.78	68.16	62.96	75.56	45.23	90.14	89.88	96.99	90.38	88.84	97.81	85.45	91.96	92.13
IoU in %																			
Cityscapes val.	79.37	12.80	70.77	5.60	34.68	31.21	59.30	39.03	63.07	1.96	55.60	64.83	19.99	83.77	48.0	67.00	12.36	47.70	61.42
BDD100K val.	75.27	11.51	67.93	26.14	23.72	37.16	64.13	23.19	53.07	4.57	84.21	56.11	15.27	57.77	42.83	53.11	1.11	28.13	73.05
KITTI	85.22	24.08	69.84	14.13	26.69	42.17	57.63	21.15	67.13	12.80	90.36	56.92	36.87	85.31	22.77	64.45	57.50	51.57	44.64
GTA5	75.83	17.24	68.46	16.84	4.83	47.43	58.48	17.07	61.18	2.97	89.84	64.58	38.11	18.94	19.90	58.35	17.74	30.81	67.15

4.2. Result for Multi-Category Segmentation

This subsection discusses the performance of Grounded SAM in a multi-category classification context, focusing exclusively on the Cityscapes dataset. To evaluate the model's performance, a confusion matrix of the segmented results was constructed, as illustrated in Figure 4. The model performed relatively well on most of the categories.

Specifically, the categories of Road (t_1), Building (t_3), Sky (t_{11}), Person (t_{12}), Car (t_{14}), Truck (t_{15}), Bus (t_{16}), and Motorcycle (t_{18}) have very high recognition accuracies above 0.9, which implies that the model has strong discriminatory ability on these common and obvious categories. However, there are also some categories such as Fence (t_5), Traffic Sign (t_8), Terrain (t_{10}), Rider (t_{13}), and Train (t_{17}), which have significantly lower accuracies below 0.7, implying that the model may have recognition problems on these categories or confusion with other categories. In particular, the accuracies for Fence (t_5) and Terrain (t_{10}) ranged between 0.55 and 0.7, which may indicate that the model had some difficulty with these two categories, but the overall performance was acceptable. However, for the categories of Traffic Sign (t_8), Rider (t_{13}), and Train (t_{17}), the accuracy is as low as 0.04 to 0.01, which means that the model fails almost completely on these categories.

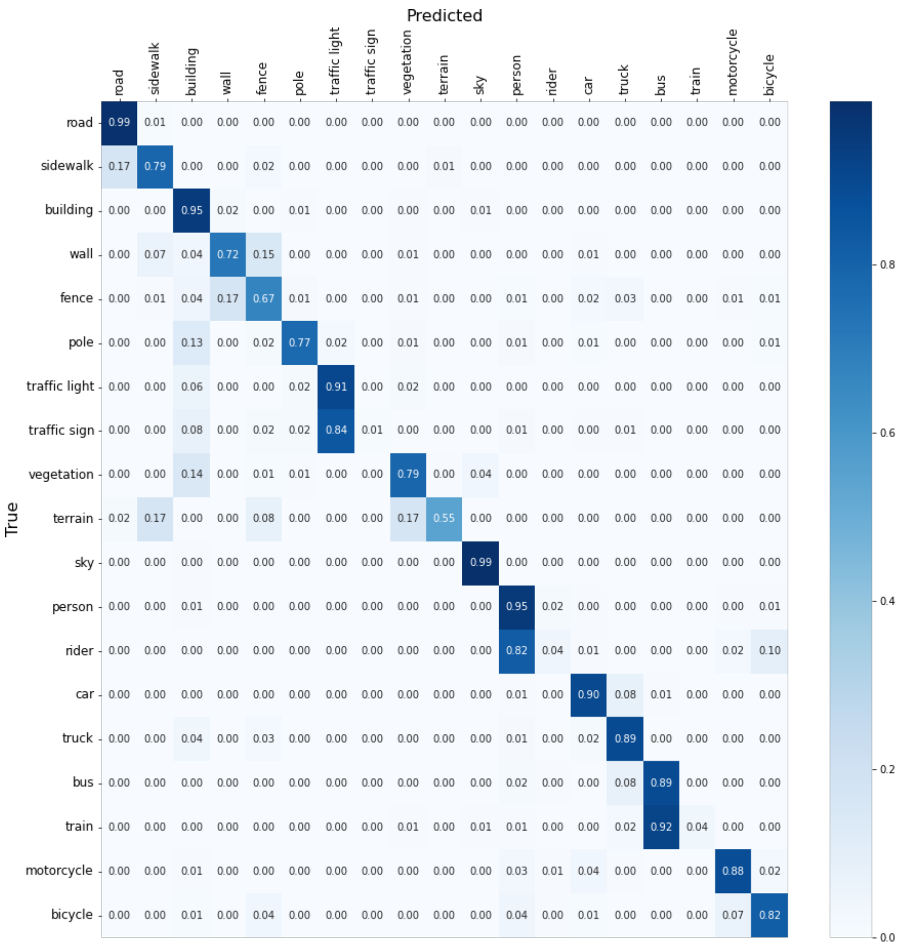


Figure 4. Confusion matrix of the segmented results for the multi-category experiment.

The confusion matrix analysis reveals that Grounded SAM excels in identifying and segmenting several key urban elements. However, it also highlights significant challenges in certain categories. The low accuracies in categories suggest a need for further model refinement, possibly through more sophisticated prompt engineering or additional training data to improve discrimination between these and other categories.

Overall, these results for multi-category segmentation provide valuable insights into the strengths and limitations of Grounded SAM in handling complex urban environments, laying the groundwork for future improvements in model accuracy.

4.3. Improvements in Addressing Textual Similarity

After improving input prompts, we have made significant progress in the categories that are prone to confusion due to text similarity shown in Figure 5. A notable enhancement was observed in the identification of Rider (t_{13}) seen in Figures 5a and 6. The accuracy for

Rider (t_{13}) significantly increased from 0.04 to 0.59. This substantial improvement indicates that the majority of previous misclassifications have been effectively addressed. However, it is worth noting that there was a slight decrease in the accuracy for the Person category (t_{12}). This trade-off suggests a shift in the model's ability to differentiate between closely related categories.

The adjusted prompts led to a remarkable improvement in distinguishing Traffic Sign (t_8), with their accuracy soaring from 0 to 0.80 seen in Figures 5b and 7. This improvement is particularly significant considering the previous challenges in differentiating Traffic Sign (t_8) from Traffic Light (t_7). Interestingly, the accuracy of Traffic Light (t_7) remained stable, indicating that the model's existing proficiency in recognising Traffic Light (t_7) was maintained while enhancing its ability to identify Traffic Sign (t_8).

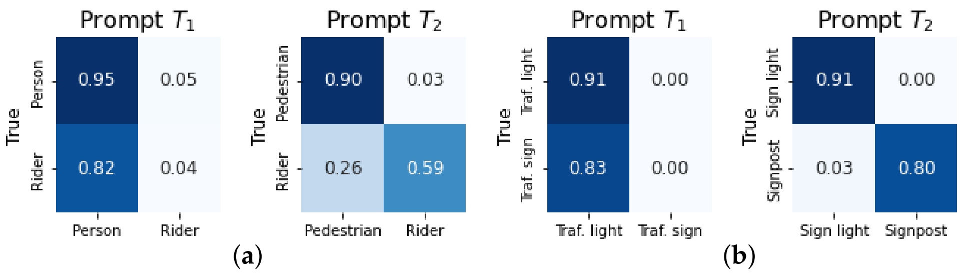


Figure 5. Confusion matrix based on original pre-defined names (left) and improved prompt (right) for categories caused by textual misunderstanding. (**a**) Person (t_{12}) and Rider (t_{13}). (**b**) Traffic Light (t_7) and Traffic Sign (t_8).

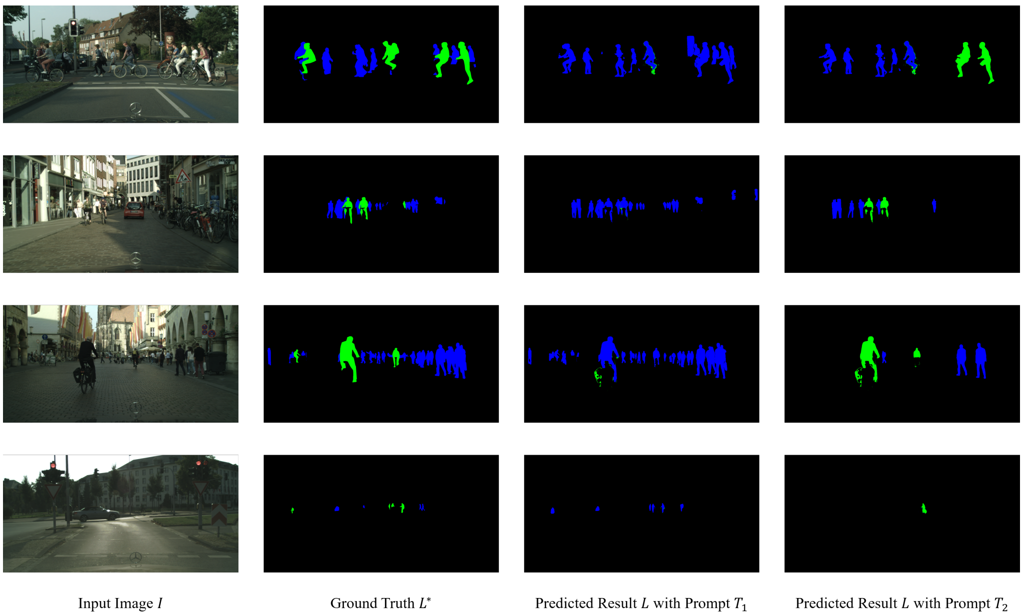


Figure 6. Examples of the segmented prediction for Person (t_{12}) and Rider (t_{13}). Green label represents Person (t_{12}) and blue label represents Rider (t_{13}).

These results demonstrate the effectiveness of prompt refinement in enhancing the model's performance, particularly in categories prone to textual confusion. The significant increase in accuracy for Rider (t_{13}) and Traffic Sign (t_8) underlines the potential of prompt engineering as a powerful tool to fine-tune semantic segmentation models. This improvement is crucial for applications where the precise identification of distinct but textually similar categories is essential.

In summary, the refinements in input prompts have led to substantial improvements in Grounded SAM's ability to accurately distinguish between categories that previously posed challenges due to textual similarities.

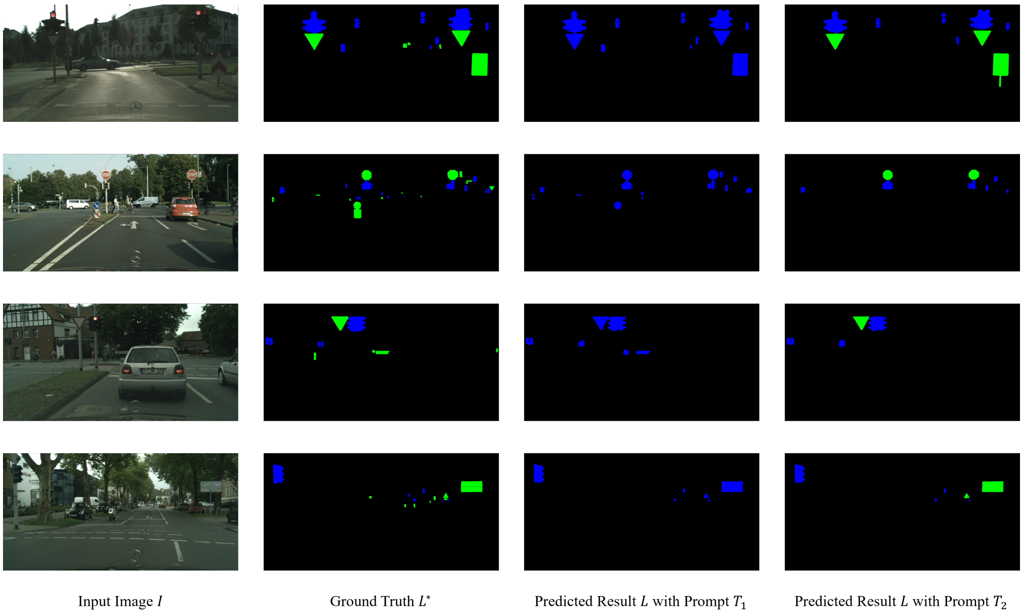


Figure 7. Examples of the segmented prediction for Traffic Light (t_7) and Traffic Sign (t_8). Green label represents Traffic Light (t_7); and blue label represents Traffic Sign (t_8).

4.4. Improvements in Addressing Visual Similarity

For categories confused by image similarity, we explored varying input prompts. The results are shown in Figure 8. When using the open vocabulary model for semantic segmentation, subtle changes to the prompt do affect the model's categorisation performance. Despite experimenting with a range of rail-related terms, the model continued to struggle with the Train (t_{17}) category. Different prompt pairings were tested, such as Bus (t_{16}) and On Rails (t_{17}), Bus (t_{16}) and Locomotive (t_{17}), Bus (t_{16}) and Streetcar (t_{17}), and Bus (t_{16}) and Tram (t_{17}). Unfortunately, these attempts resulted in the recognition accuracy for Train (t_{17}) remaining below 0.05, indicating persistent difficulties in differentiating trains from visually similar objects.

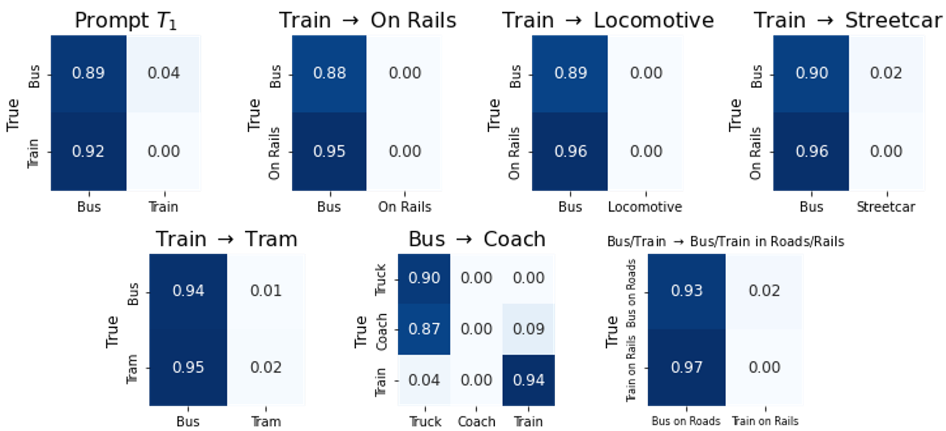


Figure 8. Confusion matrix based on original pre-defined names (T_1) and improved prompt (T_3) for categories caused by visual similarity.

An interesting observation was made when more descriptive prompts were employed, such as "Bus on Roads" (t_{16}) and "Train on Rails" (t_{17}). While the accuracy for Train (t_{17})

improved slightly, it remained relatively low. A significant breakthrough was achieved when the term "Coach" was introduced. Although this led to some new confusions, with the model misclassifying Coach (t_{16}) mainly as Truck (t_{15}), the recognition of Train (t_{17}) improved dramatically, reaching an accuracy of 0.94 shown in Figure 9.

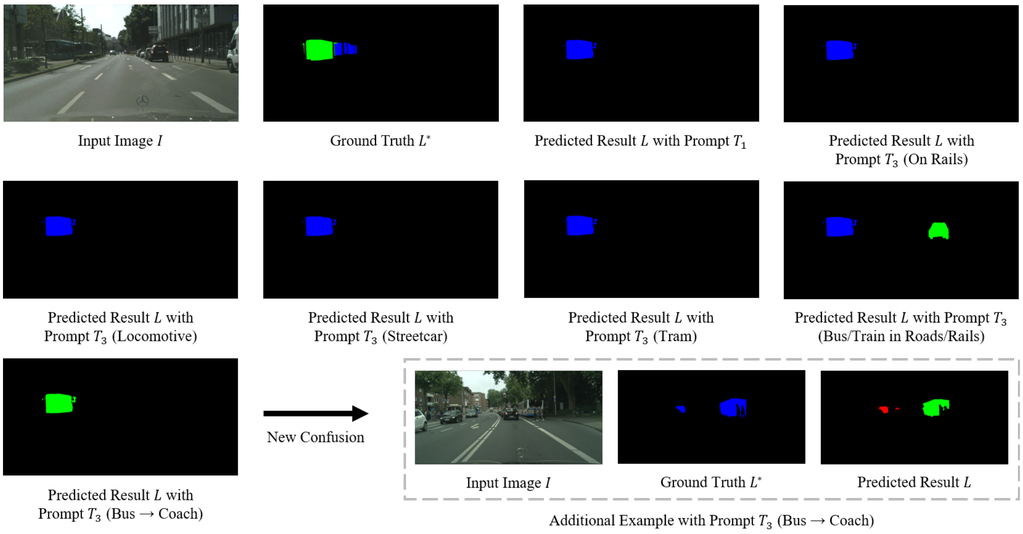


Figure 9. Examples of the segmented prediction for Bus (t_{16}) and Train (t_{17}). Red label represents Truck (t_{15}); green label represents Bus (t_{16}); and blue label represents Train (t_{17}).

These experiments underscore the nuanced impact that prompt design can have on the model's ability to categorise visually similar objects. While subtle changes in prompts can influence categorisation performance, certain categories that are too visually similar continue to pose challenges. The remarkable improvement in recognising trains using the term "Coach" highlights the potential of creative prompt engineering in overcoming these challenges.

To sum up, this study demonstrates that while prompt modifications can significantly enhance the recognition of certain visually similar categories in semantic segmentation, the process requires careful experimentation and creativity to identify the most effective terms.

4.5. Comprehensive Results

Much like in the initial test for single categories, the open vocabulary model Grounded SAM performed well on most categories, as shown in Table 4. On Road (t_1), Building (t_3), Sky (t_{11}), Car (t_{14}), and Bike (t_{19}), both precision and recall exceeded 80%, showing the model's ability to recognise and classify these categories well.

However, there is a significant difference in precision and recall between Traffic Light (t_7) and Traffic Sign (t_8) when using the category names defined by Cityscapes as prompts T_1. Traffic Sign (t_8) has a high precision of 90.35% but a surprisingly low recall of 0.96%. This implies that Grounded SAM may miss many instances of Traffic Sign (t_8) due to the confusing textual descriptions. Traffic Light (t_7) has only 23.54% precision, but its recall is as high as 90.73%, which implies that the model may have misclassified instances from other categories into this category. Combined with Figures 5b and 7 in Section 4.3, it can be found that the reason leading to these two extreme cases is that when applying T_1, most of Traffic Sign (t_8) is segmented into Traffic Light (t_7). After further adjustment using T_2 as input prompts, due to the textual similarity of the categories, Traffic Light (t_7) and Traffic Sign (t_8) improve. The precision of Traffic Light (t_7) increases to 69.91%, and the recall rate also increases slightly. After losing a little precision, the recall rate of Traffic Sign (t_8) increases to 80.22%.

For categories such as Rider (t_{13}) and Train (t_{17}), the model performs poorly with the application of prompt (T_1), and their precision and recall are very low. Figures 5a and 6 show that Rider (t_{13}) and Train (t_{17}) are extremely easy to confuse with Person (t_{12}) and

Bus (t_{16}), respectively. Based on previous analysis in Section 4.3, the reason for the Rider (t_{13}) error is due to similarity in the input text, as is the case for Traffic Sign (t_8). Similarly, with the updated prompt, the precision of Rider (t_{13}) increases from 17.70% to 56.44%, and its recall increases from 4.31% to 58.81%. The precision of Person (t_{12}) has also increased slightly, while its recall has decreased slightly. For Train (t_{17}), which is visually very similar to Bus (t_{16}), the recall is still almost zero with Prompt T_3's improvements, and the precision, which was 45.77%, plummets to 0%. The model still has significant challenges in this visually similar category.

These results illustrate Grounded SAM's effective performance in several categories and its challenges in dealing with textual and visual similarities. The adjustments made to the prompts led to notable improvements in some categories, but challenges persist, particularly in differentiating categories with close visual resemblances. This analysis underscores the importance of prompt design in optimising the performance of open vocabulary semantic segmentation models.

Table 4. Semantic segmentation performance with four prompt designs on Cityscapes validation dataset. **Bold values** indicate the improved categories and total results, and <u>underline values</u> highlight the best performance in different prompts.

Prompt	Road	S.walk	Build.	Wall	Fence	Pole	T.Light	T.Sign	Veget.	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	M.bike	Bike	Total
Precision in %																				
T_1	98.09	86.35	91.87	46.49	51.47	80.96	23.54	90.35	96.90	66.94	88.38	74.97	17.70	97.15	29.71	65.46	45.77	47.14	86.70	**67.68**
T_2	98.08	88.71	91.66	46.58	48.33	81.70	**69.91**	**85.22**	97.34	64.01	89.76	**88.26**	**56.44**	97.10	31.95	61.03	42.22	48.04	84.18	**<u>72.13</u>**
T_3	98.07	86.67	91.69	51.05	49.63	80.41	23.69	54.50	96.96	62.68	89.1	74.93	21.76	97.12	38.11	**56.95**	**0.00**	49.82	86.46	**63.66**
T_4	98.18	87.92	91.72	49.17	48.91	82.15	**69.03**	**83.59**	97.44	65.42	90.02	**87.92**	**56.13**	97.14	28.70	**58.33**	**0.06**	47.97	84.39	**69.69**
Recall in %																				
T_1	98.76	78.71	94.80	72.08	67.16	77.44	90.73	0.96	78.52	55.08	99.39	95.25	4.31	89.76	89.37	88.95	4.03	88.34	81.76	**71.34**
T_2	99.02	77.92	95.06	70.54	68.24	74.33	**91.16**	**80.22**	78.41	56.84	99.39	**89.51**	**58.81**	90.70	89.87	90.05	4.06	87.88	83.88	**<u>78.20</u>**
T_3	98.81	77.9	91.5	70.37	64.02	76.89	91.26	0.47	78.52	55.79	99.38	95.15	5.17	89.37	89.48	**90.34**	**0.00**	86.53	82.31	**70.70**
T_4	98.92	79.05	95.04	71.46	68.18	73.32	**91.30**	**79.54**	79.02	60.03	99.40	**89.44**	**57.78**	90.54	75.31	**89.82**	**0.03**	87.92	83.85	**77.37**
IoU in %																				
T_1	96.90	70.00	87.47	39.40	41.12	65.50	22.99	0.96	76.60	43.30	87.90	72.28	3.59	87.46	28.70	60.54	3.85	44.38	72.65	**52.93**
T_2	97.14	70.89	87.49	38.99	39.46	63.72	**65.48**	**70.42**	76.77	43.07	89.27	**79.99**	**40.45**	88.31	30.84	57.17	3.85	45.05	72.46	**<u>61.10</u>**
T_3	96.93	69.56	84.49	42.02	38.81	64.76	23.17	0.47	76.63	41.88	88.61	72.17	4.36	87.07	36.47	**53.68**	**0.00**	46.24	72.91	**52.64**
T_4	97.14	71.31	87.53	41.10	39.82	63.25	**64.77**	**68.80**	77.42	45.57	89.53	**79.65**	**39.80**	88.19	26.23	**54.71**	**0.02**	45.01	72.59	**60.65**

5. Discussion

In street view semantic segmentation tasks, traditional deep learning models usually rely on predefined sets of categories. In contrast, open vocabulary models lift this limitation by processing arbitrary textual descriptions to define the desired categories. This approach brings flexibility and adaptability to semantic segmentation tasks, enabling models to map and interpret complex visual scenes more accurately. Street scene understanding involves complex and diverse semantic elements with significant geographical variations. While traditional deep learning models are limited by their training conditions, the flexibility of open vocabulary models is particularly important in dealing with highly complex and variable environments. In scenarios such as urban environments, it is crucial to accurately recognise and classify numerous elements. Open vocabulary models provide an effective solution for this purpose. We explored the performance of the Grounded SAM model for the direct inference of street scene images without training, using Cityscapes and their defined categories as benchmarks. We devised multiple textual inputs to compare performance differences. In addition, we evaluated the model on four Cityscapes-style datasets and compared it to current state-of-the-art (SOTA) models.

5.1. Overall Performance of Open Vocabulary Models

Although Grounded SAM performs well when dealing with single-category street segmentation tasks, it still faces challenges in multi-category segmentation tasks, especially when there is a high degree of textual semantic or visual feature similarity between categories. For example, the model may produce misclassifications when distinguishing between visually highly similar objects such as pedestrians and cyclists. This phenomenon suggests that although the open vocabulary model is theoretically broadly adaptable, it still requires further fine-tuning or comparative learning when dealing with complex multi-classification problems.

5.2. Impact of Text Input Refinement on Results

Our study found that refined text inputs are crucial for reducing classification confusions. For example, different text inputs led to significantly different results when distinguishing between categories such as traffic signals and traffic signs or pedestrians and cyclists. This suggests that classification accuracy can be significantly improved by more specific and granular text input. However, for categories that are also difficult to distinguish visually, such as buses and trains, even fine-grained textual descriptions struggled to eliminate model confusion. This emphasises the need for open vocabulary models to be further enhanced with the combined capabilities of image processing and text understanding in future developments.

5.3. Comparison with Other SOTA Models

In our study, we compared Grounded SAM's performance on Cityscapes, BDD100K, GTA5, and KITTI to current state-of-the-art (SOTA) methods shown in Tables 5 and 6. Some visual results are shown in Figures 10 and 11. On Cityscapes, the highest performance of Grounded SAM with the prompt T_2 was 61.1%. In comparison, the highest performance of an SOTA model was 80.8% (Table 6 last column). Similarly, on BDD100K, GTA5, and KITTI, the highest performances of the models were 40.4%, 56.2%, and 50.6%, while those of the best SOTA model were 53.5%, 75.9%, and 72.8% (Table 5). On a single category, using Cityscapes as an example shown in Table 6, Grounded SAM performs close to SOTA's performance on most classes. For example, on Road, Building, Pole, Traffic Light, Person, and Bike, the difference is very small. However, on Fence, Terrain, Rider, Train, and Motorbike, the gap is still large, at over 20%.

Table 5. Semantic segmentation performance of Grounded SAM compared to state-of-the-art segmentation methods. We generated Grounded SAM results for BDD100K, GTA5, and KITTI. SOTA is taken from the respective literature.

Method	mIoU
Dataset: BDD100K (val)	
NiseNet [43]	53.5
Grounded SAM with T_1	39.1
Grounded SAM with T_2	40.4
Grounded SAM with T_3	38.2
Grounded SAM with T_4	39.2
Dataset: GTA5	
MIC [44]	75.9
HRDA [45]	73.8
DAFormer [46]	68.3
ProDA [47]	57.5
CCM [48]	49.9
Grounded SAM with T_1	53.1
Grounded SAM with T_2	56.2
Grounded SAM with T_3	51.4
Grounded SAM with T_4	53.6
Dataset: KITTI	
Deeplabv3+ + SDCNet [49]	72.8
MapillaryAI [50]	69.6
SIW [51]	68.9
AHiSS [52]	61.2
SegStereo [53]	59.1
APMoE-seg [54]	48.0
Grounded SAM with T_1	45.4
Grounded SAM with T_2	50.6
Grounded SAM with T_3	45.2
Grounded SAM with T_4	50.3

Table 6. Semantic segmentation performance per category compared to state-of-the-art methods on Cityscapes validation dataset. Results for Grounded SAM were generated by us using T_2, and SOTA values are collected from Takikawa et al. [55].

Method	Road	S.walk	Build.	Wall	Fence	Pole	T.Light	T.Sign	Veget.	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	M.bike	Bike	Total
LRR [56]	97.7	79.9	90.7	44.4	48.6	58.6	68.2	72.0	92.5	69.3	94.7	81.6	60.0	94.0	43.6	56.8	47.2	54.8	69.7	69.7
Deeplabv2 [26]	97.9	81.3	90.3	48.8	47.4	49.6	57.9	67.3	91.9	69.4	94.2	79.8	59.8	93.7	56.5	67.5	57.5	57.7	68.8	70.4
Piecewise [57]	98.0	82.6	90.6	44.0	50.7	51.1	65.0	71.7	92.0	72.0	94.1	81.5	61.1	94.3	61.1	65.1	53.8	61.6	70.6	71.6
PSP-Net [24]	98.2	85.8	92.8	57.5	65.9	62.6	71.8	80.7	92.4	64.5	94.8	82.1	61.5	95.1	78.6	88.3	77.9	68.1	78.0	78.8
Deeplabv3+ [28]	98.2	84.9	92.7	57.3	62.1	65.2	68.6	78.9	92.7	63.5	95.3	92.3	62.8	95.4	85.3	89.1	80.9	64.6	77.3	78.8
GSCNN [55]	98.3	86.3	93.3	55.8	64.0	70.8	75.9	83.1	93.0	65.1	95.2	85.3	67.9	96.0	80.8	91.2	83.3	69.6	80.4	80.8
Grounded SAM	97.1	70.9	87.5	39.0	39.5	63.7	65.5	70.4	76.8	43.1	89.3	80.0	40.5	88.3	30.8	57.2	3.9	45.1	72.5	61.1

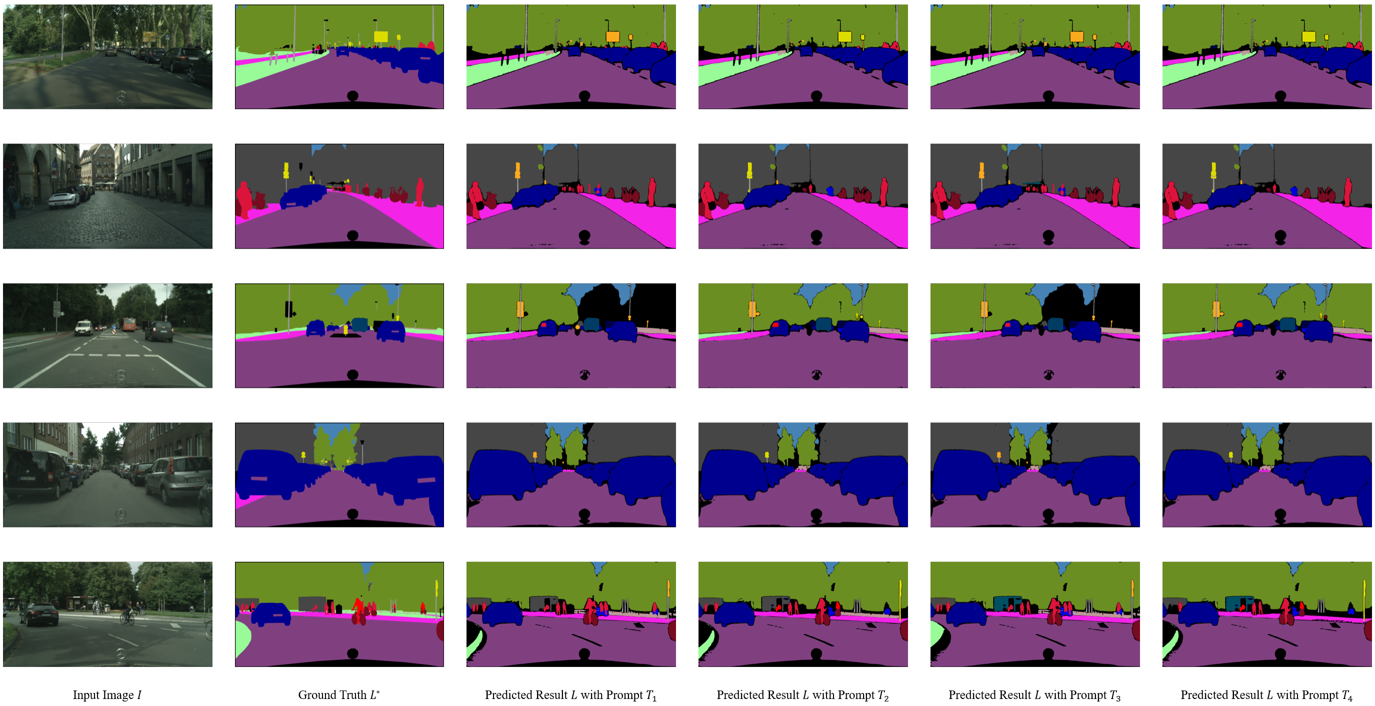


Figure 10. Examples of segmented prediction in multi-category classification experiment with Prompt T_1, T_2, T_3, and T_4 on Cityscapes. * The visualised segmented results apply the same colour map as Cityscapes.

When compared to current state-of-the-art models, the open vocabulary model may be slightly worse in terms of overall performance. However, it is worth noting that its performance in some specific categories is very close or even equal to these state-of-the-art models. Most importantly, instead of requiring extensive training on a specific dataset, the open lexical model allows for direct inference on lower-cost hardware. This feature provides significant advantages in terms of resource and time costs and is particularly important for resource-constrained research and application scenarios.

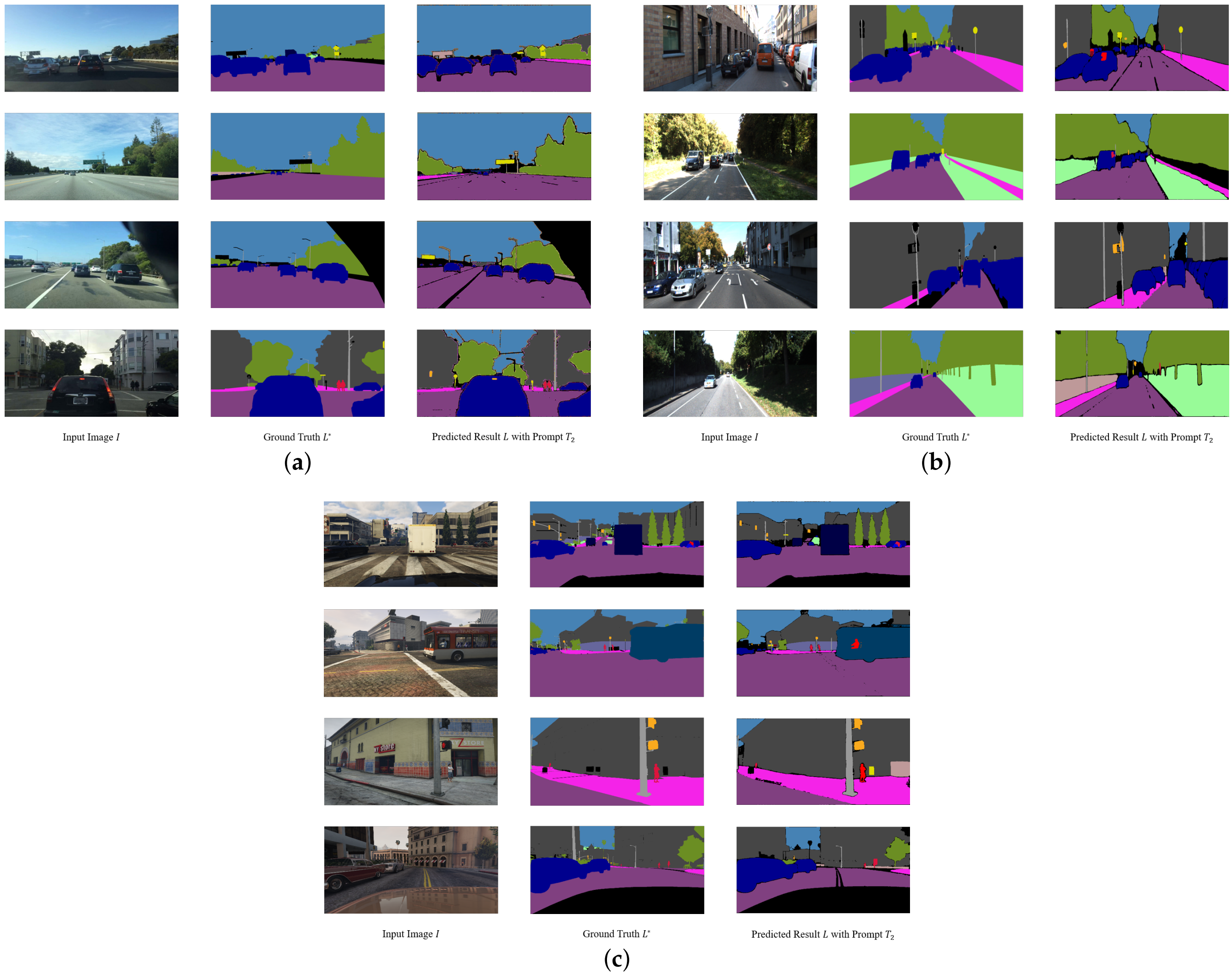


Figure 11. Examples of segmented prediction in multi-category classification experiment with Prompt T_2. * The visualised segmented results apply the same colour map as Cityscapes. (**a**) BDD100K. (**b**) KITTI. (**c**) GTA5.

6. Conclusions

This study explored the capabilities of Grounded SAM, an open vocabulary model, in the context of street view semantic segmentation. Grounded SAM demonstrates significant adaptability and flexibility over traditional deep learning models, particularly in its ability to process arbitrary textual descriptions for category definition. While it excels in single-category tasks, challenges arise in multi-category segmentation scenarios, especially with categories having high textual or visual similarities. Textual input refinement proved crucial in reducing classification errors, yet the model struggled with visually similar categories like buses and trains. Compared to state-of-the-art models, Grounded SAM shows competitive performance in specific categories without the need for extensive training, highlighting its efficiency and potential in resource-constrained settings. The findings underscore the need for further enhancements in integrating image processing and text understanding to improve the model's overall efficacy in complex urban environments. In addition, Grounded SAM performed well on datasets in different regions, which indicates that it has a strong adaptability in geography.

Finally, this study finds that open vocabulary models for segmentation such as Grounded SAM are capable of segmenting visual elements in street view imagery without any further training. Despite their significant robustness and generalisation, the performance in categories with similar visual appearance, which are also difficult to separate in traditional deep learning models or even human judgement, are still a considerable challenge. An additional finding is that minor changes in prompts can drastically influence predictions. Specific vocabulary for categories enables models to reduce confusion. The adaptability of foundation models in different geographic regions is impressive, which may supplant transfer learning for geographic regions. While we focus on street view imagery, the capabilities of open vocabulary models in the understanding of other crucial vision data such as remote sensing imagery await further exploration.

Author Contributions: Conceptualization, Zichao Zeng and Jan Boehm; methodology, Zichao Zeng and Jan Boehm; software, Zichao Zeng; validation, Zichao Zeng; formal analysis, Zichao Zeng; investigation, Zichao Zeng; resources, Jan Boehm; data curation, Zichao Zeng; writing—original draft preparation, Zichao Zeng; writing—review and editing, Jan Boehm and Zichao Zeng; visualization, Zichao Zeng; supervision, Jan Boehm; project administration, Jan Boehm; funding acquisition, Jan Boehm. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Engineering and Physical Sciences Research Council (EPSRC) through an industrial Cooperative Award in Science & Technology (iCASE) with Ordnance Survey (Grant number EP/W522077/1).

Data Availability Statement: Data are contained within the article.

Acknowledgments: The authors want to thank the Ordnance Survey for their support of this research.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zhang, F.; Zhou, B.; Liu, L.; Liu, Y.; Fung, H.H.; Lin, H.; Ratti, C. Measuring human perceptions of a large-scale urban region using machine learning. *Landsc. Urban Plan.* **2018**, *180*, 148–160. [CrossRef]
2. Biljecki, F.; Ito, K. Street view imagery in urban analytics and GIS: A review. *Landsc. Urban Plan.* **2021**, *215*, 104217. [CrossRef]
3. Liu, Y.; Chen, M.; Wang, M.; Huang, J.; Thomas, F.; Rahimi, K.; Mamouei, M. An interpretable machine learning framework for measuring urban perceptions from panoramic street view images. *Iscience* **2023**, *26*. [CrossRef] [PubMed]
4. Kang, Y.; Zhang, F.; Gao, S.; Lin, H.; Liu, Y. A review of urban physical environment sensing using street view imagery in public health studies. *Ann. GIS* **2020**, *26*, 261–275. [CrossRef]
5. Guan, F.; Fang, Z.; Zhang, X.; Zhong, H.; Zhang, J.; Huang, H. Using street-view panoramas to model the decision-making complexity of road intersections based on the passing branches during navigation. *Comput. Environ. Urban Syst.* **2023**, *103*, 101975. [CrossRef]
6. Feng, D.; Haase-Schütz, C.; Rosenbaum, L.; Hertlein, H.; Glaeser, C.; Timm, F.; Wiesbeck, W.; Dietmayer, K. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. *IEEE Trans. Intell. Transp. Syst.* **2020**, *22*, 1341–1360. [CrossRef]
7. Jongwiriyanurak, N.; Zeng, Z.; Wang, M.; Haworth, J.; Tanaksaranond, G.; Boehm, J. Framework for Motorcycle Risk Assessment Using Onboard Panoramic Camera (Short Paper). In Proceedings of the 12th International Conference on Geographic Information Science (GIScience 2023). Schloss Dagstuhl-Leibniz-Zentrum für Informatik, Leeds, UK, 12–15 September 2023.
8. Li, Z.; Ning, H. Autonomous GIS: The next-generation AI-powered GIS. *Int. J. Digit. Earth* **2023**, *16*, 4668–4686. [CrossRef]
9. Roberts, J.; Lüddecke, T.; Das, S.; Han, K.; Albanie, S. GPT4GEO: How a Language Model Sees the World's Geography. *arXiv* **2023**, arXiv:2306.00020.
10. Wang, X.; Fang, M.; Zeng, Z.; Cheng, T. Where would i go next? large language models as human mobility predictors. *arXiv* **2023**, arXiv:2308.15197.
11. Mai, G.; Huang, W.; Sun, J.; Song, S.; Mishra, D.; Liu, N.; Gao, S.; Liu, T.; Cong, G.; Hu, Y.; et al. On the opportunities and challenges of foundation models for geospatial artificial intelligence. *arXiv* **2023**, arXiv:2304.06798.
12. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the 38th International Conference on Machine Learning, Virtual Event, 18–24 July 2021; pp. 8748–8763.
13. Li, J.; Li, D.; Xiong, C.; Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Proceedings of the 39th International Conference on Machine Learning, Baltimore, MD, USA, 17–23 July 2022; pp. 12888–12900.
14. Li, J.; Li, D.; Savarese, S.; Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Proceedings of the 40th International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023; pp. 19730–19742.

15. Garcia-Garcia, A.; Orts-Escolano, S.; Oprea, S.; Villena-Martinez, V.; Garcia-Rodriguez, J. A review on deep learning techniques applied to semantic segmentation. *arXiv* **2017**, arXiv:1704.06857.
16. Siam, M.; Gamal, M.; Abdel-Razek, M.; Yogamani, S.; Jagersand, M.; Zhang, H. A comparative study of real-time semantic segmentation for autonomous driving. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–23 June 2018; pp. 587–597.
17. Badue, C.; Guidolini, R.; Carneiro, R.V.; Azevedo, P.; Cardoso, V.B.; Forechi, A.; Jesus, L.; Berriel, R.; Paixao, T.M.; Mutz, F.; et al. Self-driving cars: A survey. *Expert Syst. Appl.* **2021**, *165*, 113816. [CrossRef]
18. 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, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 3213–3223.
19. Liu, X.; Deng, Z.; Yang, Y. Recent progress in semantic image segmentation. *Artif. Intell. Rev.* **2019**, *52*, 1089–1106. [CrossRef]
20. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
21. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef]
22. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; proceedings, part III 18; Springer: Cham, Switzerland, 2015; pp. 234–241.
23. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv* **2014**, arXiv:1412.7062.
24. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
25. Zhou, Z.; Rahman Siddiquee, M.M.; Tajbakhsh, N.; Liang, J. Unet++: A nested u-net architecture for medical image segmentation. In Proceedings of the Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, 20 September 2018; Proceedings 4; Springer: Cham, Switzerland, 2018; pp. 3–11.
26. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef]
27. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
28. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
29. Kang, Y.; Cho, N.; Yoon, J.; Park, S.; Kim, J. Transfer learning of a deep learning model for exploring tourists' urban image using geotagged photos. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 137. [CrossRef]
30. Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Li, C.; Yang, J.; Su, H.; Zhu, J.; et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv* **2023**, arXiv:2303.05499.
31. Li, L.H.; Zhang, P.; Zhang, H.; Yang, J.; Li, C.; Zhong, Y.; Wang, L.; Yuan, L.; Zhang, L.; Hwang, J.N.; et al. Grounded language-image pre-training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 10965–10975.
32. Minderer, M.; Gritsenko, A.; Stone, A.; Neumann, M.; Weissenborn, D.; Dosovitskiy, A.; Mahendran, A.; Arnab, A.; Dehghani, M.; Shen, Z.; et al. Simple open-vocabulary object detection with vision transformers. *arXiv* **2022**, arXiv:2205.06230.
33. Zareian, A.; Rosa, K.D.; Hu, D.H.; Chang, S.F. Open-vocabulary object detection using captions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 14393–14402.
34. Du, Y.; Wei, F.; Zhang, Z.; Shi, M.; Gao, Y.; Li, G. Learning to prompt for open-vocabulary object detection with vision-language model. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 14084–14093.
35. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.Y.; et al. Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2023; pp. 4015–4026.
36. Ren, T.; Liu, S.; Zeng, A.; Lin, J.; Li, K.; Cao, H.; Chen, J.; Huang, X.; Chen, Y.; Yan, F.; et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv* **2024**, arXiv:2401.14159.
37. Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; Joulin, A. Emerging properties in self-supervised vision transformers. In Proceedings of the Pr IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 9650–9660.
38. 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, Seattle, WA, USA, 14–19 June 2020; pp. 2636–2645.

39. Abu Alhaija, H.; Mustikovela, S.K.; Mescheder, L.; Geiger, A.; Rother, C. Augmented reality meets computer vision: Efficient data generation for urban driving scenes. *Int. J. Comput. Vis.* **2018**, *126*, 961–972. [CrossRef]
40. Richter, S.R.; Vineet, V.; Roth, S.; Koltun, V. Playing for data: Ground truth from computer games. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part II 14. Springer: Cham, Switzerland, 2016; pp. 102–118.
41. Szeliski, R. *Computer Vision: Algorithms and Applications*; Springer Nature: Cham, Switzerland, 2022.
42. Everingham, M.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [CrossRef]
43. Nag, S.; Adak, S.; Das, S. What's there in the dark. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 2996–3000.
44. Hoyer, L.; Dai, D.; Wang, H.; Van Gool, L. MIC: Masked image consistency for context-enhanced domain adaptation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 11721–11732.
45. Hoyer, L.; Dai, D.; Van Gool, L. Hrda: Context-aware high-resolution domain-adaptive semantic segmentation. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Cham, Switzerland, 2022; pp. 372–391.
46. Hoyer, L.; Dai, D.; Van Gool, L. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 9924–9935.
47. Zhang, P.; Zhang, B.; Zhang, T.; Chen, D.; Wang, Y.; Wen, F. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 12414–12424.
48. Li, G.; Kang, G.; Liu, W.; Wei, Y.; Yang, Y. Content-consistent matching for domain adaptive semantic segmentation. In Proceedings of the European Conference on Computer Vision, Virtual Event, 23–28 August 2020 Springer: Cham, Switzerland, 2020, pp. 440–456.
49. Zhu, Y.; Sapra, K.; Reda, F.A.; Shih, K.J.; Newsam, S.; Tao, A.; Catanzaro, B. Improving semantic segmentation via video propagation and label relaxation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 8856–8865.
50. Bulo, S.R.; Porzi, L.; Kontschieder, P. In-place activated batchnorm for memory-optimized training of dnns. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5639–5647.
51. Yin, W.; Liu, Y.; Shen, C.; Hengel, A.v.d.; Sun, B. The devil is in the labels: Semantic segmentation from sentences. *arXiv* **2022**, arXiv:2202.02002.
52. Meletis, P.; Dubbelman, G. Training of convolutional networks on multiple heterogeneous datasets for street scene semantic segmentation. In Proceedings of the 2018 IEEE Intelligent Vehicles Symposium (IV), Changshu, China, 26–30 June 2018; pp. 1045–1050.
53. Yang, G.; Zhao, H.; Shi, J.; Deng, Z.; Jia, J. Segstereo: Exploiting semantic information for disparity estimation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 636–651.
54. Kong, S.; Fowlkes, C. Pixel-wise attentional gating for parsimonious pixel labeling. *arXiv* **2018**, arXiv:1805.01556.
55. Takikawa, T.; Acuna, D.; Jampani, V.; Fidler, S. Gated-scnn: Gated shape cnns for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 5229–5238.
56. Ghiasi, G.; Fowlkes, C.C. Laplacian pyramid reconstruction and refinement for semantic segmentation. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part III 14; Springer: Cham, Switzerland, 2016; pp. 519–534.
57. Lin, G.; Shen, C.; Van Den Hengel, A.; Reid, I. Efficient piecewise training of deep structured models for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 3194–3203.