International Journal of Geo-Information

MDPI

Article

Investigating the Performance of Open-Vocabulary Classification Algorithms for Pathway and Surface Material Detection in Urban Environments

Kauê de Moraes Vestena [1,*], Silvana Phillipi Camboim [1], Maria Antonia Brovelli [2] and Daniel Rodrigues dos Santos [3]

[1] Department of Geomatics, Federal University of Paraná, Curitiba 80060-000, Brazil; silvanacamboim@ufpr.br
[2] Politecnico di Milano, Department of Civil and Environmental Engineering, 20133 Milan, Italy; maria.brovelli@polimi.it
[3] Military Institute of Engineering, Cartography Section, Rio de Janeiro 22290-270, Brazil; daniel.rodrigues@ime.eb.br
* Correspondence: kauemv2@gmail.com

Abstract: Mapping pavement types, especially in sidewalks, is essential for urban planning and mobility studies. Identifying pavement materials is a key factor in assessing mobility, such as walkability and wheelchair usability. However, satellite imagery in this scenario is limited, and in situ mapping can be costly. A promising solution is to extract such geospatial features from street-level imagery. This study explores using open-vocabulary classification algorithms to segment and identify pavement types and surface materials in this scenario. Our approach uses large language models (LLMs) to improve the accuracy of classifying different pavement types. The methodology involves two experiments: the first uses free prompting with random street-view images, employing Grounding Dino and SAM algorithms to assess performance across categories. The second experiment evaluates standardized pavement classification using the Deep Pavements dataset and a fine-tuned CLIP algorithm optimized for detecting OSM-compliant pavement categories. The study presents open resources, such as the Deep Pavements dataset and a fine-tuned CLIP-based model, demonstrating a significant improvement in the true positive rate (TPR) from 56.04% to 93.5%. Our findings highlight both the potential and limitations of current open-vocabulary algorithms and emphasize the importance of diverse training datasets. This study advances urban feature mapping by offering a more intuitive and accurate approach to geospatial data extraction, enhancing urban accessibility and mobility mapping.

Keywords: open-vocabulary algorithms; pavement segmentation; surface material detection; street-view imagery; urban mobility

check for updates

Citation: de Moraes Vestena, K.; Phillipi Camboim, S.; Brovelli, M.A.; Rodrigues dos Santos, D. Investigating the Performance of Open-Vocabulary Classification Algorithms for Pathway and Surface Material Detection in Urban Environments. *ISPRS Int. J. Geo-Inf.* **2024**, *13*, 422. https://doi.org/10.3390/ijgi13120422

Academic Editor: Wolfgang Kainz

Received: 17 September 2024
Revised: 1 November 2024
Accepted: 21 November 2024
Published: 24 November 2024

Copyright: © 2024 by the authors. Published by MDPI on behalf of the International Society for Photogrammetry and Remote Sensing. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction and Related Work

The study of sidewalks is crucial for accessibility and quality of life in cities [1–4]. As the required level of detail is not available in satellite imagery and in situ surveys are costly, mapping such features is challenging. Using street-level imagery for this purpose is, therefore, a potential improvement. However, defining the features to be classified in street-level imagery is challenging because landscape features are conceptualized in natural language. Thus, large language models (LLMs) can be used to improve the process of image pattern retrieval from street-level imagery by allowing the use of flexible and natural language prompts. This characteristic makes the extraction of map features more intuitive and accurate by exploiting the language-understanding capabilities of LLMs, which can be applied in many contexts, such as in urban planning and transport studies, such as in [5], where it was used to solve a community-level land-use task through using a feedback iteration. Recently, there has been considerable interest in the accessibility of these models,

as demonstrated by ChatGPT, which serves as an interface for user interaction with the latest iterations of the Generative Pre-trained Transformer (GPT). This development is the most recent in a continuing series of advances in natural language processing. The study of natural language processing (NLP) can be traced back to the 1950s [6], with early developments culminating in the establishment of the essential subtasks, such as sentence boundary detection, tokenization, part-of-speech assignment to individual words, morphological decomposition, chunking, and problem-specific segmentation [6]. Later developments resulted in the tailoring of conventional algorithms, such as support vector machines and hidden Markov models.

Among these developments, the extension of open-vocabulary prompting for image pattern retrieval represents a substantial leap forward in NLP, driven by advances in LLMs. This approach has recently gained more attention due to its capabilities of richer expressiveness and more extensive flexibility due to its generalization capabilities [7], its ability to capture nuances of processes [8], and its enabling of transfer learning [9]. The work of Zareian et al. [10] claims to be the first to pose the problem of "Open Vocabulary Object Detection" in opposition to other similar approaches. One example of that alternative approach was "Zero Shot-Detection", which aims to generalize from seen to unseen classes using semantic descriptions or attributes [11–13]. Another approach would be weakly supervised detection, which focuses on detecting classes with limited information, such as image-level or noisy labels [14,15]. It is worth noticing that there are some main drawbacks to open-vocabulary approaches, such as a higher computational cost due to their larger complexity [7] and the risk of misinterpretation due to the lack of constraints [8].

This paper aims to evaluate open-vocabulary algorithms within a domain-specific pipeline for pathway segmentation and surface material classification. To achieve these objectives, we designed a two-stage pipeline that establishes a workflow for large-scale, georeferenced pathway surface classification. In Experiment 1, we segment pathways in street-level imagery, while in Experiment 2, we classify segmented pathways by surface material type. Performance is evaluated regarding the pipeline's accuracy, adaptability to various pathway types, and efficiency in processing large-scale georeferenced images, aiming to develop a robust workflow for urban infrastructure mapping. Identification of pavement material types is crucial for maintaining road safety and ensuring the well-being of people [16]. This process is vital for mobility studies as it influences safety, skid resistance, and road noise [17].

Particularly for sidewalks, identifying suitable pavement types is essential for improving accessibility, urban mobility, and safety for all users, including those with reduced mobility [18]. Additionally, using specific pavement types, like exposed aggregate concrete (EAC) and porous concrete, can significantly reduce noise and enhance safety on both roadways and sidewalks [19]. Zeng & Boehm [20] exploited a broader investigation of open-vocabulary algorithms, which obtained averagely good results compared to SOTA closed-vocabulary ones, observing a prompt-dependent accuracy.

Although "street" and "sidewalk" are common classes amongst many classification/segmentation datasets [21–24], being one of the reference classes for testing many proposed algorithms [24–26], pavement type identification is a far less commonly undertaken task: convolutional neural networks were used to identify a few classes, limited to asphalt, gravel, and cement [27], and also for the identification of asphalt damage traits, such as "pothole" or "patch" [28], and pavement damage assessment [29], as well as for pixel-level segmentation, limited to "paved" and "unpaved" streets [30], and have only recently included eight specific classes, such as "granite blocks" and "hexagonal asphalt paver" [31]. Each of these studies developed specific models with a closed set of classes, thus not allowing a wide use of the algorithms, in contrast to open-vocabulary counterparts, such as CLIP [32] and Grounding Dino [33], which are designed to be generalist algorithms.

Furthermore, none of these previous studies on pavement material identification follow a global standard. This work innovates by integrating its application into OpenStreetMap (OSM), the world's most popular open and collaborative geographic database [34].

The feature attribute standards created by the OSM community have typically been agreed upon over many years, including surface materials, although they are often not defined by domain experts. Adopting this strategy helps maintain interoperability between the developments of this research and other tools and applications linked to OSM. In this context, this work aims to test the behaviour of some state-of-the-art algorithms for open-vocabulary image classification tasks in identifying pavement types according to the OSM conceptual model, thus aiming to help design valuable building blocks for producing compliant geographic data.

This study presents a valuable proposition to the community of urban planners and geospatial analysts. By using open-vocabulary algorithms and large language models (LLMs) to segment and identify pavement materials in street-level imagery, this work advances the methodology for mapping urban features. It bridges the gap between natural language processing and geospatial data extraction, providing a more intuitive and granular approach to feature classification. By integrating with OpenStreetMap (OSM), the study improves data interoperability and sets a precedent for using open and collaborative platforms for urban data management. We also contribute to creating building blocks for producing compliant geodata by providing open resources, such as the Deep Pavements dataset and a fine-tuned CLIP-based model. These innovations contribute to more accessible, scalable, and flexible solutions for urban infrastructure analysis, with potential applications in transport planning, accessibility improvement, and public safety.

2. Methodology

The task of using NLP to extract features from images has as its main challenge the process called "grounding", whose level means the capability of the model to associate elements of language with the proper regions in the images [35–37]. It was shown that the relevance of the hierarchical relationships between concepts improves the grounding level by establishing ontologies [38–40]. Ontologies bridge the gap between natural language and logical reasoning by providing a machine-readable representation (class) of real-world concepts [41]. Our study's methodological approach is presented in Figure 1, which includes the main steps and processes to fulfil our study's goals.

The motivational task behind this study can be formalized as follows: "Given a set of street view images, the objective is to segment all visible paths in each image using free-text input algorithms, identify their surface materials, and ensure that the process is not affected by the use of synonyms in the input". Considering the hierarchical relationships between the entities of interest and their properties, an ontology was created using the Web Ontology Language (OWL) standard [42]. This ontology is presented as a graph in Figure 2. The hierarchy naturally starts with "pathway", which is then specialized into "Road" and "Footway" categories. The branching follows up to two main interest categories: "Road" and "Sidewalk". All of them, at the same time, have their main differences. Still, they fundamentally share the characteristic of being walkable, which is essentially made possible through having a surface material, hereby treated as a fundamental shared characteristic of all "pathways". After that, "surface material" has its branching sub-ontology that is composed of two main categories, namely "unpaved material" and "paved material", the latter being subdivided into "continuous" and "block-based". This property sharing is expected to be visible in the results of any algorithm classifying bitmaps among these categories. With the ontology set, it is possible to establish different degrees of semantic separation among classes, highlighting pavement surfaces. Therefore, "asphalt" is more similar to "concrete" (0 degrees, same hierarchical level) than it is to "sett" (1 degree) and compacted pavement (2 degrees). This ontology is basilar to the proposed experiment, with its hierarchical relationships being called off along the study. As far as our knowledge goes, no similar ontology has been proposed before.



Figure 1. The study methodology workflow.

To undertake the present study, we designed two sequential experiments as part of a comprehensive methodology to obtain surface-labeled, segmented pathways from georeferenced terrestrial imagery. The first experiment, "Free Prompting and Random Imagery Evaluation", focuses on pathway segmentation. Here, we aimed to test the feasibility of open-vocabulary prompting across a diverse set of 296 randomly selected images, covering multiple categories and prompts to observe general trends and the model's responses in a realistic, large-scale processing scenario. In cases where an image was excessively blurry, it was discarded from the sample to ensure reliable segmentation results. The employed algorithm was a combination of Grounding Dino [33] and the foundation model SAM—Segment Anything [43]. Grounding Dino provided output bounding boxes based on the prompts, which were then used as inputs for SAM to obtain the segmented features. We used Mappillary, an ever-growing platform with billions of community-contributed CC BY-SA licensed street-view images [44], as the data source.

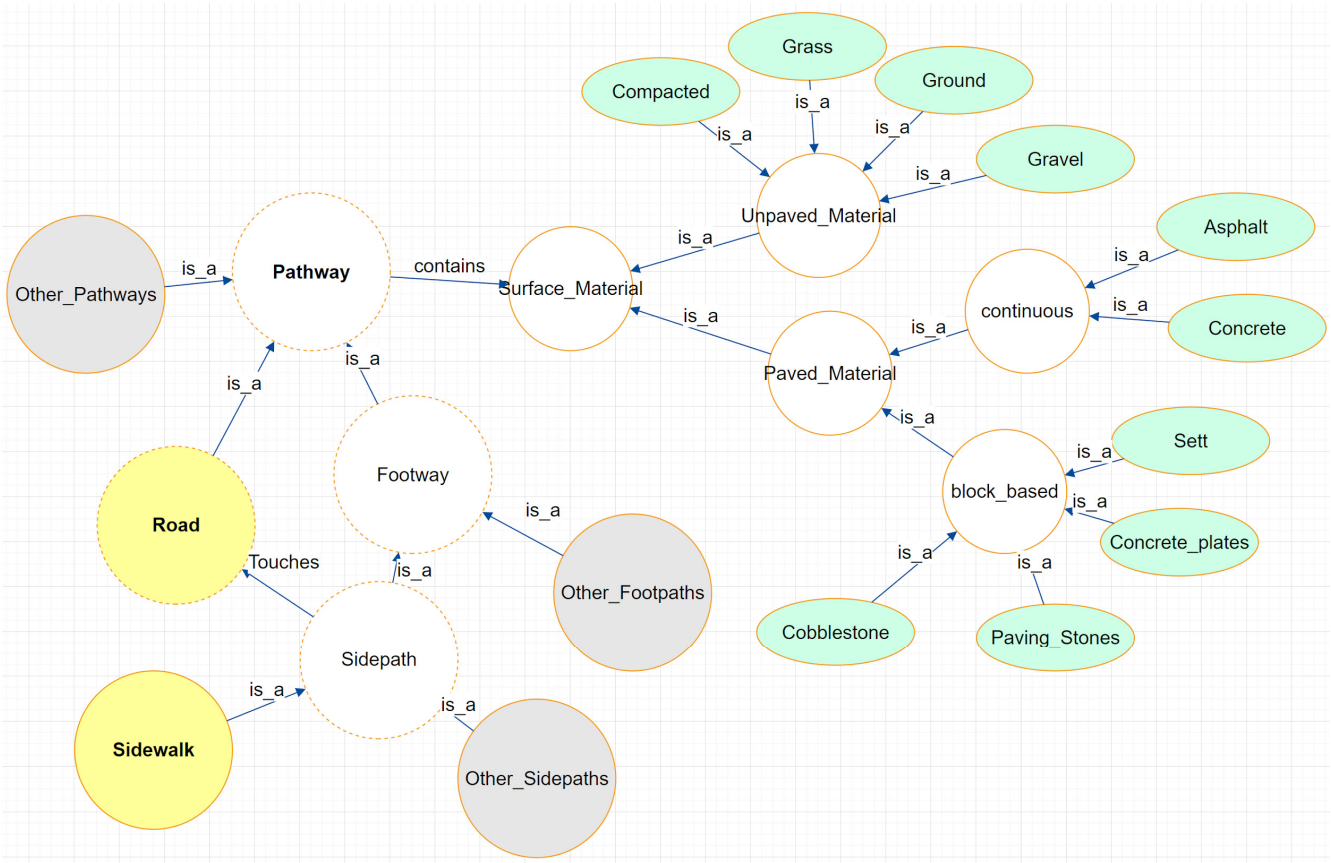


Figure 2. The ontology of the phenomena of interest.

The second experiment, designed for surface-labeling, namely "Evaluation of Open-Vocabulary Algorithms for Standardized Pavement Classification", builds upon the segmentation results of the first experiment. Here, we tested standardized pavement classes prompting against images of surface patches to assess the classification ability of an open-vocabulary algorithm in a constrained setup. We employed the CLIP algorithm, known for assigning probabilities to each entry of a set of sentences that can be changed at each query, constituting a fundamental advantage, as it allows adaptation to different realities. We tested eight variations of the algorithm, each employing different datasets and training strategies, as compared in Table 1 regarding their origin and computational costs. The data source for this experiment was the "deep-pavements-dataset", a tailored dataset of surface patches collaboratively maintained on GitHub [45], containing 500 samples for OSM-compliant pavement categories at the time of the experiments. The testing included all categories in the ontology, with examples shown in Figure 3. These experiments establish a pipeline where Experiment 1 segments pathways and Experiment 2 assigns surface labels, demonstrating a workflow for large-scale pathway surface classification using georeferenced images.

Table 1. Tested CLIP algorithm variants.

Package	Model Name	Base Algorithm	Main Image Dataset	Model Size (GB)	Image Parameters (Millions)	Text Parameters (Millions)	Image GFLOPS	Text GFLOPS	Package	Model Name
CLIP	RN101	CNN	OpenAI	0.28	56.26	63.43	19.54	5.96	CLIP	RN101
CLIP	RN50x64	CNN	OpenAI	1.26	420.38	202.88	529.11	23.55	CLIP	RN50x64
CLIP	ViT-B/32	VT	OpenAI	0.34	87.85	63.43	8.82	5.96	CLIP	ViT-B/32
CLIP	ViT-L/14	VT	OpenAI	0.89	303.97	123.65	162.03	13.3	CLIP	ViT-L/14
Open CLIP	ViT-H-14-378-quickgelu	VT	dfn5b	3.95	632.68	354.03	1006.96	47.09	Open CLIP	ViT-H-14-378-quickgelu
Open CLIP	EVA02-E-14-plus	VT	laion2b_s9b_b144k	10.1	4350.56	694.33	2264.33	97.86	Open CLIP	EVA02-E-14-plus



Figure 3. Some snapshots of the Deep Pavements dataset.

It is valuable to elaborate on the reasons for selecting this set of algorithms for our study. SAM is particularly noted for its unpaired segmentation capabilities, supported by a billion-level sample size, which enhances its applicability to real-world scenarios and provides superior generalization performance [46,47]. Grounding Dino offers complementary strengths and significantly outperforms comparable algorithms. This advantage is attributed to its use of a transformer architecture that integrates multi-level text information [33,48]. CLIP marked a significant breakthrough in the vision-language domain by employing a shared embedding space for text and images created through a contrastive learning approach [49]. It has been recognised for its robustness across various scenarios [50–52]. These algorithms are considered zero-shot learners capable of performing tasks without being specifically optimized [11]. Furthermore, despite their relatively recent introduction, these models have already seen widespread use in the industry for various applications, such as in automated image data annotation, image search engines, accessibility tools providing image descriptions, and enhancing content recommendation systems [53].

Regarding the first experiment and the proposed ontology, we tested the following categories with their corresponding prompts:

1. Auxiliary: These entities are detected to determine if a pavement detection failure occurred due to occlusion rather than incorrect detection. Prompts are “car”, “vehicle”, “pole”, and “tree”.
2. Sidepaths: This query is primarily aimed at detecting sidewalks, but it may also retrieve other sidepaths, such as paved shoulders. Prompts are “sidewalk”, “sidepath”, “sideway”, “sidetrack”, and “lateral”.
3. Roads: Focused on detecting motorized pathways. Prompts are “road” and “street”.
4. Pathways: These are intended to detect any kind of traversable way. Prompts are “way”, “path”, “pathway”, “pavement”, and “track”.
5. Surface pavement types: Prompts directly target the property, with additional ones included for broader testing. Prompts are “sett”, “grass”, “cobblestone”, “earth”, “soil”, “dirt”, “sand”, “concrete”, “paving stones”, “chip seal”, “gravel”, “compacted”, “asphalt”, “concrete plates”, and “ground”.
6. Abstract concepts: Words that do not have a unique or physical representation are used to test the model’s responses. Prompts include anything, nothing, something, void, and thing.

For the second experiment, we conducted two tests with different strategies to compare the classification accuracy with the pre-trained models. In the first “Model Aggregation” test, we summed the class probabilities across all models to create an overall score. The class

with the highest aggregated score was selected as the final prediction. This aggregation was performed internally, using all models initially before being repeated with the three best overall performers.

In the second approach, termed the "Fine-Tuning" test, we selected 60% of the samples for fine-tuning and used the remaining 200 samples per category for testing. Following the specifications in Table 1, we chose the lightest model to evaluate the extent of improvement relative to the model size and computational burden. Previous studies [49] showed that CLIP's fine-tuning is highly sensitive to the optimizer's algorithm hyperparameters. Therefore, we empirically tested multiple scenarios, presenting the worst and the best outcomes side by side. All the produced models and analyses are published in a repository at the HuggingFace community [54] to ensure reproducibility.

3. Results and Discussion

3.1. Free Prompting and Random Imagery Evaluation

Following the methodology presented in the previous section, eight images were tested for each prompt. As such, for categories that received 5 prompts, the experiment involved 40 images. Table 2 presents the results obtained for the "auxiliary" category. In the forthcoming tables, a "**" sign indicates that the targeted prompt does not exist in the image sample; thus, the correct outcome should be none.

Table 2. Prompt results for the "auxiliary" category.

Row	Prompt	1	2	3	4	5	6	7	8	Hit %
I	car	car	truck	car	car	car panel	car	car	car	100
II	poles	street lamp	power pole	building pole	street lamp	street lamp	guard rail **	street lamp	street lamp	87.5
III	vehicle	car	truck	car	truck	fence **	car	car	truck	87.5
IV	tree	tree	Tree	tree	tree	tree	vegetation	tree	tree	100

**: it was't present on the sample, thus the proper answer would be "none".

The desired results were achieved in all cases, except when the targeted feature was absent in the picture, causing the algorithm to hallucinate. Table 3 presents the results for the "Sidepaths" category.

Table 3. Prompt results for the "sidepaths" category.

Row	Prompt	1	2	3	4	5	6	7	8	Hit %
I	sidewalk	sidewalk	sidewalk	road	sidewalk	sidewalk	a sidepath	kerbs	sidewalk	75
II	sidepath	road **	road **	kerb	road **	road	road **	road	road	0
III	sideway	car panel	car	truck	car	car	car panel	road	car	0
IV	sidetrack	car	car panel	building	car	car panel **	truck container	car	car	0
V	lateral	road **	road	road and sidewalk	car **	car	vegetation	road	car **	0
VI	walkway	road and sidepath	road **	road **	road **	partial sidewalk	all pavements	road	road **	12.5

**: it was't present on the sample, thus the proper answer would be "none".

There is considerable evidence supporting the hypothesis of a lack of generalization capability. While "sidewalk" produced consistently good results, none of the other prompts were recognised as similar concepts, which should have led to proper detections. Recall that "sidewalk" is common in many well-spread training datasets. Table 4 presents the results obtained for the "roads" category, reinforcing the hypothesis of a stronger capability for road identification, likely due to the prevalence of roads in terrestrial imagery.

Table 4. Prompt results for the "sidepaths" category.

Row	Prompt	1	2	3	4	5	6	7	8	Hit %
I	street	road	road	road	road	road	road	road	road	100
II	road	road	road	road	road	road	road	road	road	100

Table 5 presents the results for the "pathway" category. Although the results were generally good, there was a noticeable bias towards detecting roads.

Table 5. Prompt results for the "pathway" category.

Row	Prompt	1	2	3	4	5	6	7	8	hit %
I	track	car	road	road	road	road	road	road	road	87.5
II	pavement	road	road	a path	road	road --	road	walkway	road	100
III	way	bus	road markings	road	road	bus	road	sidewalk	road	62.5
IV	path	road	road	road	road	road	road	road	road	100
V	pathway	road	road	sidewalk	road	road	road	road	road	100

Although, on average, we can see the expected results, there is a clear bias towards detecting roads. Table 6 presents the results for the "surface" category, where only "asphalt" and "grass" yielded satisfactory results. Other prompts, such as "sand" and "concrete", were less successful due to the nature of the test, which involved random images that often did not correspond to the prompts. The correct behaviour of returning "nothing" was only observed in the "concrete plates" category.

Table 6. Prompt results for the "surface" category.

Row	Prompt	1	2	3	4	5	6	7	8	Hit %
I	sett	car **	car **	car **	car **	asphalt road **	tree **	car **	car **	0
II	grass	grass sidepath	grass sidewalk	grass sidewalk	grass in a garden	grass sidepath	plants **	grass stripe	grass sidepath	87.5
III	cobblestone	grass sidepath	a pathway with stones	bush row	dirt sidepath	sidepath with vegetation	asphalt road	asphalt road	asphalt road	0
IV	earth	car **	gravel sidewalk **	asphalt road **	car panel **	asphalt road **	a tire **	asphalt road **	car **	0
V	soil	an unpaved sidepath	asphalt road **	asphalt road	soy crop	car	soil sidepath	grass sidewalk	grass sidepath	12.5
VI	dirt	asphalt road **	asphalt road **	car **	unpaved sidewalk **	asphalt road **	asphalt road and car panel **	a grass slope **	asphalt road **	0
VII	sand	car **	asphalt road **	asphalt road **	sand sidewalk	asphalt road **	asphalt road **	asphalt road (partial) **	car	12.5
VIII	concrete	asphalt road **	compacted pathway **	asphalt road **	asphalt road **	concrete sidewalk	guard rail **	asphalt road **	asphalt road **	12.5
IX	paving-stones	asphalt road	soil sidepath **	concrete sidewalk **	asphalt road	asphalt road **	asphalt road	cobblestone sidewalk stripe	asphalt road	0
X	chipseal	car	truck	car	car	car	car	car	car	0

Table 6. *Cont.*

Row	Prompt	1	2	3	4	5	6	7	8	Hit %
XI	gravel	asphalt road **	asphalt pathway **	asphalt pathway **	sidepath with vegetation	car	asphalt road **	car	wall	0
XII	compacted	truck	car	car	car	car	car	car	car	0
XIII	asphalt	asphalt road	asphalt road	asphalt road	asphalt road	asphalt pathway	asphalt road	asphalt road	asphalt road	100
XIV	concrete plates	truck	nothing	nothing	truck	truck	truck	truck	a stairway	25
XV	ground	asphalt road **	asphalt road **	asphalt road and ground sidewalk	asphalt road and asphalt sidepath	asphalt road **	asphalt road **	asphalt road	asphalt road **	0

**: it was't present on the sample, thus the proper answer would be "none".

We obtained suitable matches only for 'asphalt' and 'grass'. Some categories, such as "sand" and "concrete", were affected by the nature of the test, where random images were selected and often had no correspondence. Although the correct behaviour should be "nothing", this was only seen in the "concrete plates" category. Table 7 presents the "abstract" category results, where a bias towards vehicles was observed. The prompt "anything" yielded more varied answers than "something", while "void" and "nothing" still produced responses, indicating that the language model struggles with abstract concepts.

Table 7. Prompt results for the "abstract" category.

Row	Prompt	1	2	3	4	5	6	7	8	Hit %
I	anything	car	landscape	road	person	road and sky	car panel	vegetation and sky	road and car panel	N/A
II	nothing	asphalt road and noise	asphalt road and noise	car	asphalt road and noise	noisy asphalt road and sky	sky and noise	truck	road, wall, and noise	N/A
III	something	car hood	car panel	traffic signal	car	car	nothing	truck	truck	N/A
IV	anything	road, sidewalk, and vegetation	noisy asphalt road and noisy sky	car	tree	landscape and noisy road	nothing	car	building	N/A
V	void	asphalt road	car	car	nothing	car	car	car	car	N/A

Once again, the bias towards vehicles can be observed. Somehow, "anything" gives more varied answers than "something". "Void" and "nothing" still produce matches, so the language model is not handling this abstract concept correctly. Figure 4 summarises the initial experiment's results, showing the distribution of detected classes with a hierarchical categorization approach similar to that outlined in Figure 3. The most frequently identified classes are "road" and "vehicle", with "sidepath" as the third most common, followed by categories, like "vegetation" and "amalgamation". While "sidepath" appears prominently, there is a possible bias toward road- and vehicle-related features in the dataset, which may reflect the characteristics of the urban environment sampled.

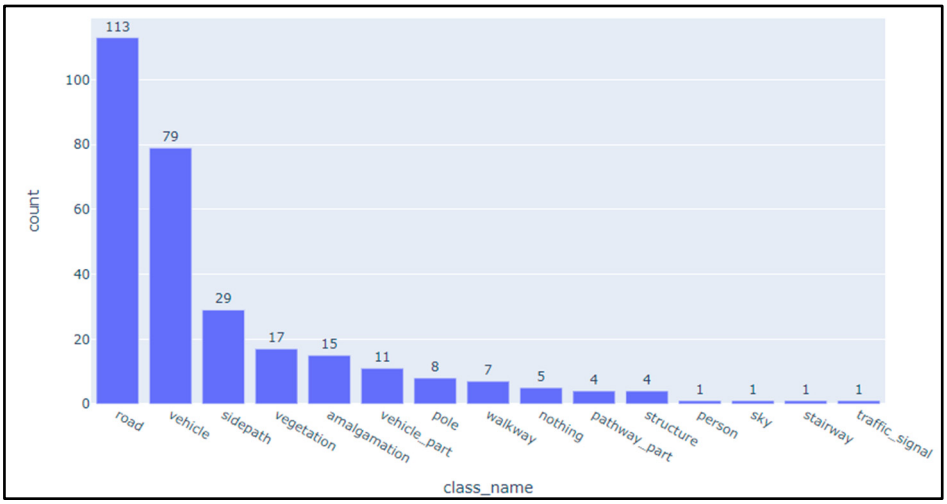


Figure 4. Free prompting and random imagery evaluation summarized results.

3.2. Evaluation of Open-Vocabulary Algorithms for Standardized Pavement Classification

We mainly relied on confusion matrices for the second experiment to summarize the algorithms' performances. In Figure 5, we computed those confusion matrices for all the presented versions.

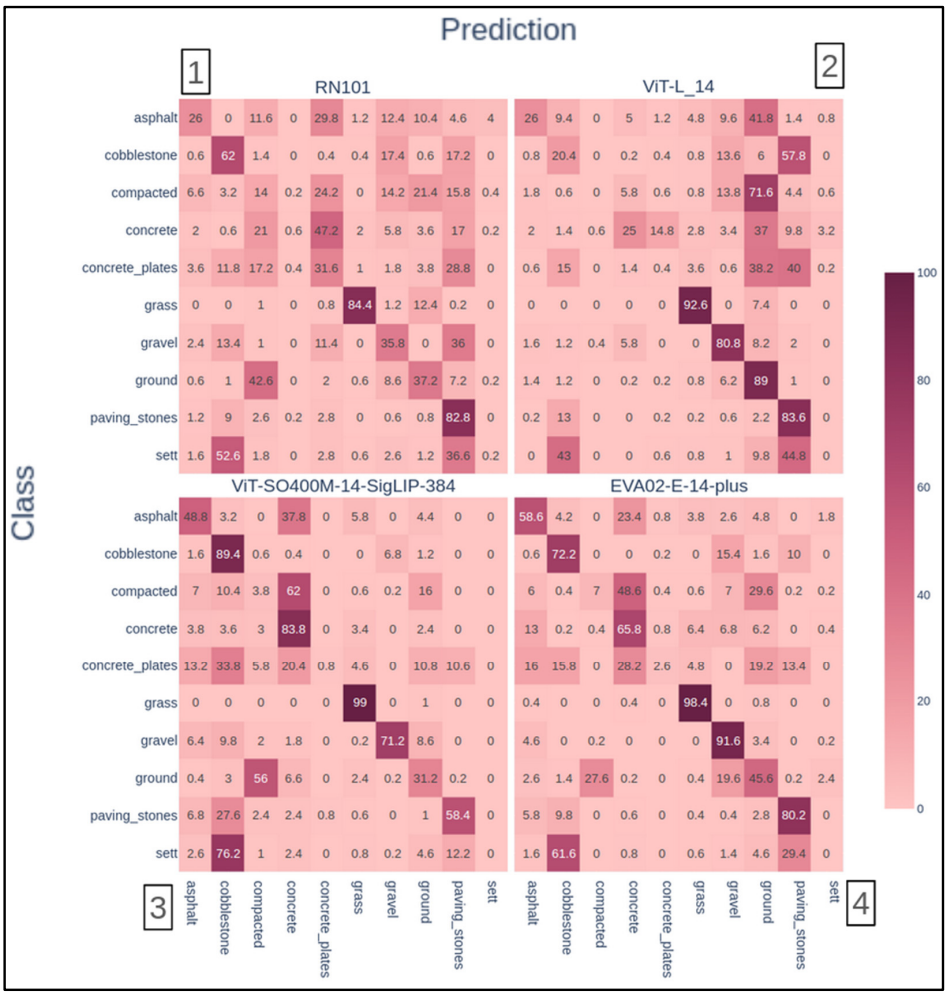


Figure 5. *Cont.*

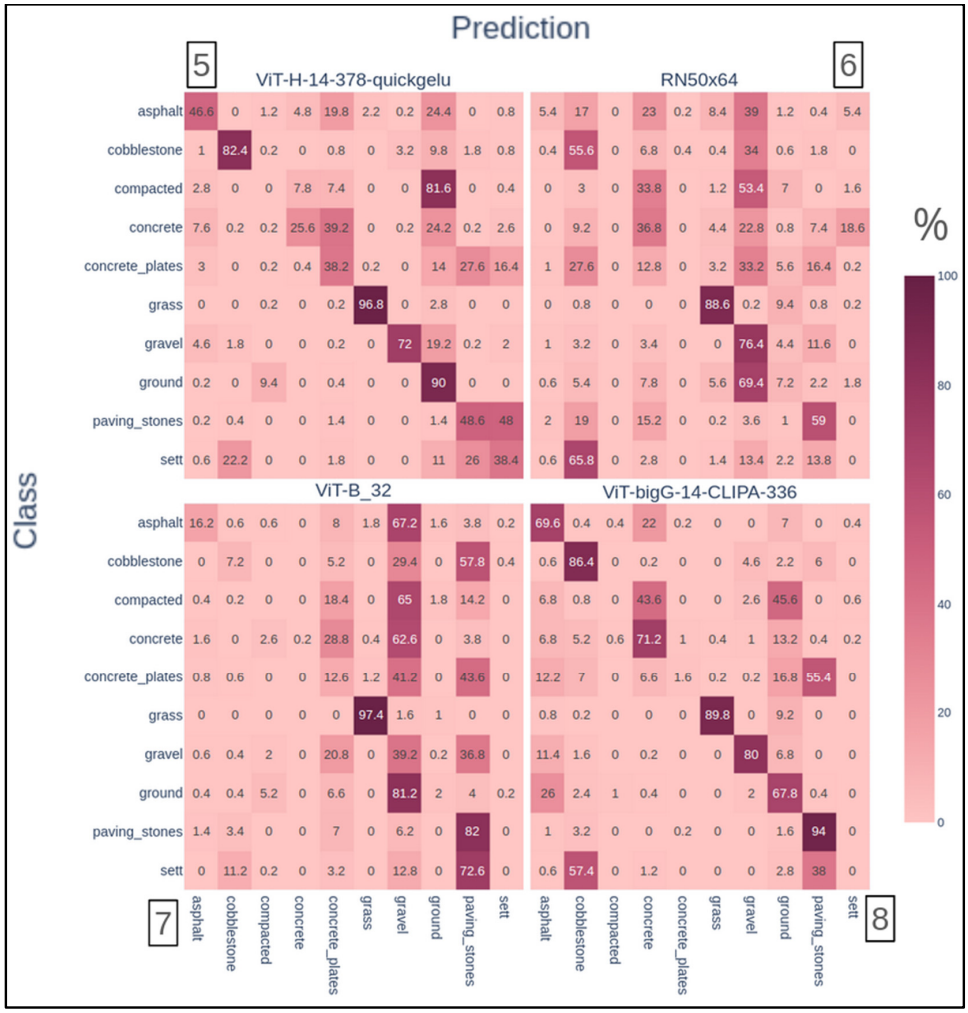


Figure 5. The resulting confusion matrices for the variants of the algorithm.

All true positives are located on the main diagonal in the confusion matrices, with the rows summing to 100%. The columns reveal biases towards certain classes, such as the "Sett" category, which had notably low prediction rates (e.g., only 38% true positives in model 5 and 0% in model 3). Extreme cases were observed in model 7, where the "gravel" class summed to 406% and the "ground" class summed to 318%, indicating significant misclassification issues. It is worth recalling that different datasets were used in each model's training. Therefore, both extremes can be due to imbalanced datasets. In this context, despite their differences, all models are variations of an algorithm that claims to be widely capable of generalizing correctly both in the vision and language domains. It is not only far from happening but also occurs in very different ways throughout all the models.

3.3. Model Aggregation and Fine-Tuning

The results of the aggregation strategy, shown in Figure 6, were obtained by summing the class probabilities from all models to create a composite arbitrary score for each class. The class with the highest score was then selected as the final prediction. This score is used only as an internal ranking mechanism for the purpose of classification and does not represent probabilistic values, as the summed values may exceed one. This test was conducted using all models and the three best overall performers.



Figure 6. Confusion matrices for the aggregation strategy.

While a general improvement was observed, some classes still showed overlap: for instance, "sett" overlapped with "cobblestone" and "paving stones", "concrete plates" overlapped with "paving stones" and "ground", and "compacted" overlapped with "concrete" and "ground". Some classes are still completely hindered, mainly "sett", "concrete plates", and "compacted".

In the fine-tuning experiments (Figure 7), we adjusted various optimizer hyperparameters, as detailed in Table 8, to refine the model performance. The fine-tuning and model selection were carried out empirically, reflecting the complexity of AI models and the absence of a universal closed-form solution [55]. As stated in the methodology, we selected the model ViT-B/32 (number 7), as it obtained the worst overall performance, as shown in Figure 8.

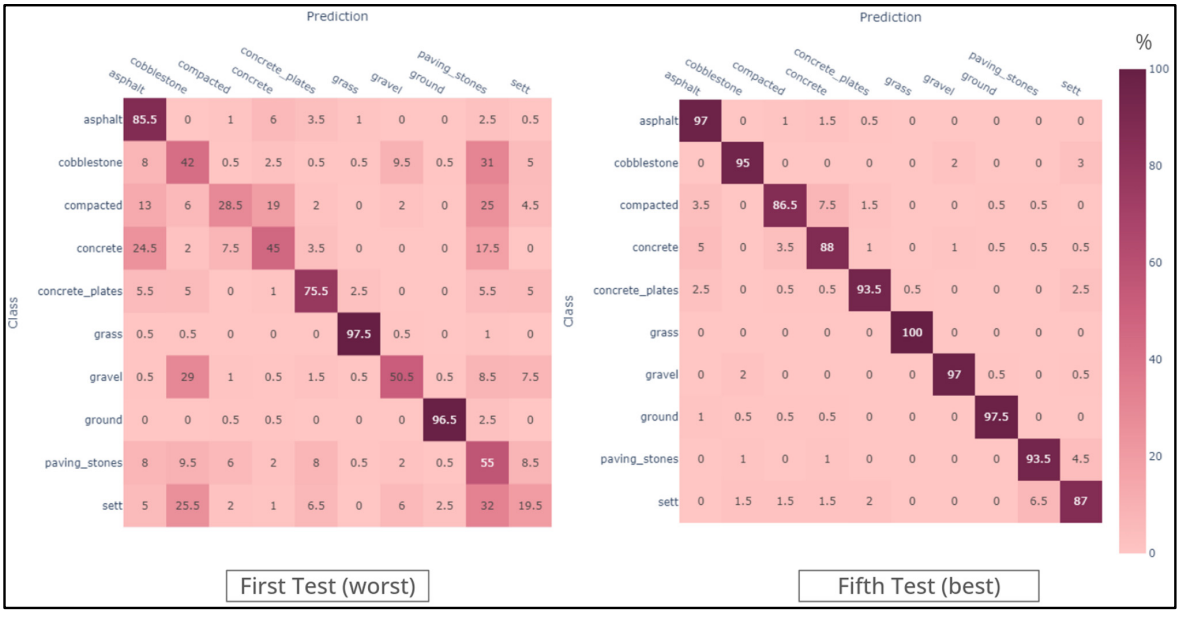


Figure 7. Confusion matrices for the fine-tuning strategy.

Table 8. Hyperparameter variation in the fine-tuning results.

Hyperparameter	Original	Best Result
Batch size	256	320
Learning rate	5.00×10^{-5}	5.00×10^{-6}
Weight decay	0.2	0.5
AMSGrad method	Deactivated	Activated
Betas	(0.9,0.98)	(0.9,0.98)
Epsilon	1.00×10^{-6}	1.00×10^{-6}
Training epochs	100	100

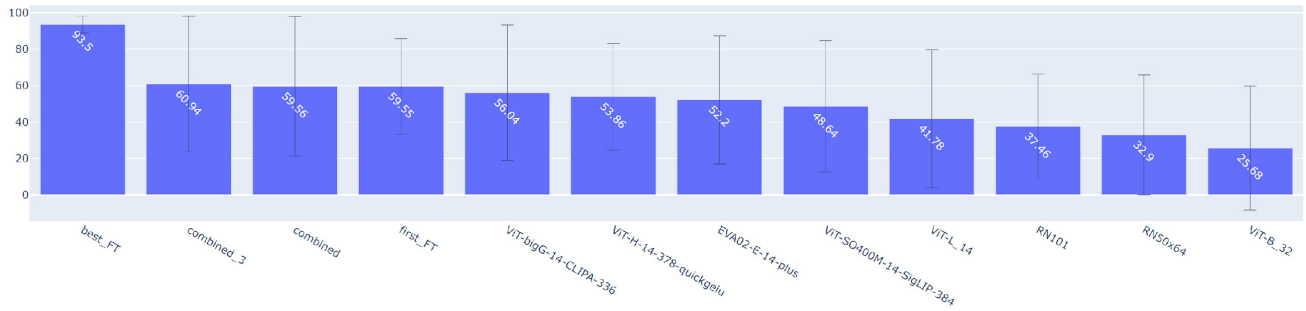


Figure 8. Average TPR of the different variants of the algorithm.

Figure 7 shows that the improvement was quite prominent after reducing the learning rate by a whole order of magnitude, which allowed for better avoidance of local minima, as foretold in the literature [56]. The use of the AMS grad [57] variant also resulted in adaptive capabilities, enabling the variation in the learning rate when mathematically viable. It performs slightly better than the state-of-the-art performance, which, as of April 2024, is an average TPR (true positive rate) of 92.4% on ImageNet [58], according to [59]. It is worth recalling that the best improvement comparison of the final result is with image 5, confusion matrix number 7, which was the starting spot. To compare all the tested scenarios directly, we created the chart in Figure 8, which contains the average TPR and the standard deviation for each tested variation.

The chart in Figure 8 was created to directly compare all tested scenarios, containing the average TPR and the standard deviation for each tested variation. Enforcing a specific meaning for a word with diverse meanings, such as "compacted", played a central role in achieving these groundbreaking results, which are also very consistent among classes. Additionally, defining "sett" as quarried and roughly parallelepipedal bricks was essential, as image searches for "cobblestone" often yielded results that matched this description, reflecting an algorithmic bias corroborated by Figures 5 and 6. These results mostly corroborate the expected that was established on the ontology, with correlations generally appearing on categories with 0 degrees of semantic separation, such as "sett" and "paving stones". One important exception was between the "compacted" and "concrete" categories, in which the proposed ontology fails, and this can be explained by the smoothness frequently present on the bitmaps belonging to these two classes.

In summary, the first experiment's "Free Prompting and Random Imagery Evaluation" demonstrated that while the desired results were generally achieved, the algorithm struggled with hallucinations when targeted features were absent. This result was particularly evident in the "sidepaths" and "surface" categories, where a bias towards detecting roads was observed. The second experiment, "Evaluation of Open-Vocabulary Algorithms for Standardized Pavement Classification", highlighted the strengths and weaknesses of different CLIP algorithm variants. The confusion matrices revealed biases towards certain classes and the challenges in accurately classifying less common pavement types. Both the aggregation and fine-tuning strategies led to improvements, though some classes remained difficult to distinguish. These findings underscore the importance of diverse training datasets and the potential for further enhancements in open-vocabulary models.

4. Conclusions

This study makes key contributions to geospatial analysis and urban infrastructure management. Firstly, we created the Deep Pavements dataset, a robust and expandable dataset specifically designed for training models in standardized (OSM-compatible) pavement material classification. This open dataset is a valuable resource for future research, enabling more comprehensive training and benchmarking of algorithms in this specialized area.

Secondly, we developed a fine-tuned, lightweight CLIP-based model optimized for pavement type detection. This model demonstrates how adapting large, state-of-the-art language models to specific, real-world applications can improve accuracy and efficiency. It also highlights the potential of using LLMs for tasks that require nuanced understanding and classification, offering a more intuitive approach to complex geospatial data challenges.

While this study provides significant insights, there are limitations that highlight opportunities for future research. The model's performance could benefit from incorporating preprocessing techniques to handle low-quality or blurry images, which are common in real-world data. Automating such preprocessing steps would make the pipeline more robust and better suited to large-scale applications. Additionally, our study focused solely on terrestrial imagery. Future research could explore the integration of satellite or sensor data to improve model accuracy and add complementary perspectives, particularly for broader urban infrastructure analysis. Further investigation into optimal train–test split ratios could also provide additional performance insights.

Our findings emphasize the importance of diversifying training datasets to improve the performance of open-vocabulary models in specialized tasks, like pavement type detection. This result aligns with broader challenges in AI, where the reliance on widely used datasets may lead to skewed or limited results in less common applications. Addressing these biases by creating more diverse and representative datasets can significantly enhance model performance and reliability.

Future research should focus on enhancing the CLIP-based model by incorporating different or additional approaches, such as data augmentation and fine-tuning with heavier versions, to compare their performance on generic tasks. Exploring robust solutions that combine specialized CLIP models with constrained vocabularies to verify hallucinations independently could further enhance the reliability of these models. Achieving a truly generalist open-vocabulary algorithm for specialized tasks, like pavement type detection, is an ambitious goal. Still, it is essential for advancing the capabilities of AI in diverse and practical applications.

In conclusion, this study contributes to developing specialized AI models for urban planning and geospatial analysis and provides valuable insights into the broader implications of dataset biases and model generalization. These contributions lay the groundwork for future advancements in creating more flexible, accurate, and representative AI models for real-world applications.

Supplementary Materials: The deep pavements dataset is widely available at https://github.com/kauevestena/deep_pavements_dataset [45], accessed on 1 May 2024; the fine-tuned CLIP-based model is widely available at https://huggingface.co/kauevestena/clip-vit-base-patch32-finetuned-surface-materials [54], accessed on 1 May 2024.

Author Contributions: Conceptualization, Kauê de Moraes Vestena, Silvana Phillipi Camboim, Maria Antonia Brovelli, and Daniel Rodrigues dos Santos; Data curation, Kauê de Moraes Vestena; Formal analysis, Kauê de Moraes Vestena; Funding acquisition, Silvana Phillipi Camboim, and Maria Antonia Brovelli; Investigation, Silvana Phillipi Camboim; Methodology, Kauê de Moraes Vestena, Silvana Phillipi Camboim, Maria Antonia Brovelli, and Daniel Rodrigues dos Santos; Project administration, Silvana Phillipi Camboim; Resources, Silvana Phillipi Camboim, and Maria Antonia Brovelli; Software, Kauê de Moraes Vestena; Supervision, Silvana Phillipi Camboim, Maria Antonia Brovelli, and Daniel Rodrigues dos Santos; Validation, Kauê de Moraes Vestena; Visualization, Kauê de Moraes Vestena; Writing—original draft, Kauê de Moraes Vestena; Writing—review and editing,

Silvana Phillipi Camboim, Maria Antonia Brovelli, and Daniel Rodrigues dos Santos. All authors have read and agreed to the published version of the manuscript.

Funding: This study was financed in part by the Coordination for the Improvement of Higher Education Personnel—Brazil (CAPES)—Finance Code 001.

Data Availability Statement: The Deep Pavements dataset is available under an MIT license: https://github.com/kauevestena/deep_pavements_dataset/blob/main/LICENSE, accessed on 1 May 2024. The fine-tuned CLIP-based model is subject to original OPENAI's licensing terms, with an MIT-based license: https://github.com/openai/CLIP/blob/main/LICENSE, accessed on 1 May 2024. Data are contained within the Supplementary Materials.

Acknowledgments: We acknowledge the Politecnico di Milano's GeoLAB, where most of the present research was physically developed.

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

References

1. Hamim, O.F.; Kancharla, S.R.; Ukkusuri, S. Mapping Sidewalks on a Neighborhood Scale from Street View Images. *Environ. Plan. B Urban Anal. City Sci.* **2023**, *51*, 823–838. [CrossRef]
2. Serna, A.; Marcotegui, B. Urban Accessibility Diagnosis from Mobile Laser Scanning Data. *ISPRS J. Photogramm. Remote Sens.* **2013**, *84*, 23–32. [CrossRef]
3. de Moraes Vestena, K.; Camboim, S.P.; dos Santos, D.R. OSM Sidewalkreator: A QGIS Plugin for an Automated Drawing of Sidewalk Networks for OpenStreetMap. *Eur. J. Geogr.* **2023**, *14*, 66–84. [CrossRef]
4. Wood, J. Sidewalk City: Remapping Public Spaces in Ho Chi Minh City. *Geogr. Rev.* **2016**, *108*, 486–488. [CrossRef]
5. Zhou, Z.; Lin, Y.; Li, Y. Large Language Model Empowered Participatory Urban Planning. *arXiv* **2024**, arXiv:2402.17161.
6. Nadkarni, P.M.; Ohno-Machado, L.; Chapman, W.W. Natural Language Processing: An Introduc-Tion. *J. Am. Med. Inform. Assoc.* **2011**, *18*, 544–551. [CrossRef]
7. Wang, X.; Ji, L.; Yan, K.; Sun, Y.; Song, R. Expanding the Horizons: Exploring Further Steps in Open-Vocabulary Segmentation. In *Pattern Recognition and Computer Vision*; Liu, Q., Wang, H., Ma, Z., Zheng, W., Zha, H., Chen, X., Wang, L., Ji, R., Eds.; Springer Nature: Singapore, 2024; pp. 407–419.
8. Eichstaedt, J.C.; Kern, M.L.; Yaden, D.B.; Schwartz, H.A.; Giorgi, S.; Park, G.; Hagan, C.A.; Tobolsky, V.A.; Smith, L.K.; Buffone, A.; et al. Closed- and Open-Vocabulary Approaches to Text Analysis: A Review, Quantitative Comparison, and Recommendations. *Psychol. Methods* **2021**, *26*, 398–427. [CrossRef]
9. Zhu, C.; Chen, L. A survey on open-vocabulary detection and segmentation: Past, present, and future. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 8954–8975. [CrossRef]
10. Zareian, A.; Dela Rosa, K.; Hu, D.H.; Chang, S. Open-Vocabulary Object Detection Using Captions. *arXiv* **2020**, arXiv:2011.10678. Available online: https://arxiv.org/abs/2011.10678 (accessed on 1 May 2024).
11. Yang, G.; Ye, Z.; Zhang, R.; Huang, K. A Comprehensive Survey of Zero-Shot Image Classification: Methods, Implementation, and Fair Evaluation. *Appl. Comput. Intell.* **2022**, *2*, 1–31. [CrossRef]
12. Lampert, C.H.; Nickisch, H.; Harmeling, S. Attribute-Based Classification for Zero-Shot Visual Object Cat-Egorization. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 453–465. [CrossRef] [PubMed]
13. Rohrbach, M.; Stark, M.; Schiele, B. Evaluating Knowledge Transfer and Zero-Shot Learning in a Large-Scale Setting. In Proceedings of the CVPR 2011, Colorado Springs, CO, USA, 20–25 June 2011; IEEE: New York, NY, USA, 2011; pp. 1641–1648.
14. Zhang, D.; Han, J.; Cheng, G.; Yang, M.-H. Weakly Supervised Object Localisation and Detection: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 5866–5885. [CrossRef]
15. Vo, H.V.; Siméoni, O.; Gidaris, S.; Bursuc, A.; Pérez, P.; Ponce, J. Active Learning Strategies for Weakly-Supervised Object Detection. In *European Conference on Computer Vision*; Springer Nature: Cham, Switzerland, 2022.
16. Blasiis, M.D.; Benedetto, A.; Fiani, M. Mobile Laser Scanning Data for the Evaluation of Pavement Surface Distress. *Remote Sens.* **2020**, *12*, 942. [CrossRef]
17. Praticò, F.; Vaiana, R. A Study on the Relationship between Mean Texture Depth and Mean Profile Depth of Asphalt Pavements. *Constr. Build. Mater.* **2015**, *101*, 72–79. [CrossRef]
18. Fidalgo, C.D.; Santos, I.M.; Nogueira, C.d.A.; Portugal, M.C.S.; Martins, L.M.T. Urban Sidewalks, Dysfunction and Chaos on the Projected Floor. The Search for Accessible Pavements and Sustainable Mobility. In Proceedings of the 7th International Congress on Scientific Knowledge, Virtual, 20–24 September 2021.
19. Vaitkus, A.; Andriejauskas, T.; Šernas, O.; Čygas, D.; Laurinavičius, A. Definition of concrete and composite precast concrete pavements texture. *Transport* **2019**, *34*, 404–414. [CrossRef]
20. 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. [CrossRef]

21. 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, 27–30 June 2016; pp. 3213–3223.
22. Geiger, A.; Lenz, P.; Stiller, C.; Urtasun, R. Vision Meets Robotics: The KITTI Dataset. *Int. J. Robot. Res.* **2013**, *32*, 1231–1237. [CrossRef]
23. Ros, G.; Sellart, L.; Materzynska, J.; Vazquez, D.; Lopez, A.M. The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3234–3243.
24. Yu, H.; Yang, Z.; Tan, L.; Wang, Y.; Sun, W.; Sun, M.; Tang, Y. Methods and Datasets on Semantic Seg-Mentation: A Review. *Neurocomputing* **2018**, *304*, 82–103. [CrossRef]
25. Hao, S.; Zhou, Y.; Guo, Y. A brief survey on semantic segmentation with deep learning. *Neurocomputing* **2020**, *406*, 302–321. [CrossRef]
26. Mo, Y.; Wu, Y.; Yang, X.; Liu, F.; Liao, Y. Review the State-of-the-Art Technologies of Semantic Segmenta-Tion Based on Deep Learning. *Neurocomputing* **2022**, *493*, 626–646. [CrossRef]
27. Zou, J.; Guo, W.; Wang, F. A Study on Pavement Classification and Recognition Based on VGGNet-16 Transfer Learning. *Electronics* **2023**, *12*, 3370. [CrossRef]
28. Zhang, C.; Nateghinia, E.; Miranda-Moreno, L.F.; Sun, L. Pavement Distress Detection Using Convolu-Tional Neural Network (CNN): A Case Study in Montreal, Canada. *Int. J. Transp. Sci. Technol.* **2022**, *11*, 298–309. [CrossRef]
29. Riid, A.; Lõuk, R.; Pihlak, R.; Tepljakov, A.; Vassiljeva, K. Pavement Distress Detection with Deep Learning Using the Orthoframes Acquired by a Mobile Mapping System. *Appl. Sci.* **2019**, *9*, 4829. [CrossRef]
30. Mesquita, R.; Ren, T.I.; Mello, C.; Silva, M. Street Pavement Classification Based on Navigation through Street View Imagery. *Ai Soc.* **2022**, *39*, 1009–1025. [CrossRef]
31. Hosseini, M.; Miranda, F.; Lin, J.; Silva, C.T. CitySurfaces: City-scale semantic segmentation of sidewalk materials. *Sustain. Cities Soc.* **2022**, *79*, 103630. [CrossRef]
32. 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 International Conference on Machine Learning, Virtual, 18–24 July 2021.
33. 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.
34. Grinberger, A.Y.; Minghini, M.; Juhász, L.; Yeboah, G.; Mooney, P. OSM Science—The Academic Study of the OpenStreetMap Project, Data, Contributors, Community, and Applications. *ISPRS Int. J. Geo-Inf.* **2022**, *11*, 230. [CrossRef]
35. Zeng, Y.; Huang, Y.; Zhang, J.; Jie, Z.; Chai, Z.; Wang, L. Investigating Compositional Challenges in Vision-Language Models for Visual Grounding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle WA, USA, 17–21 June 2024; pp. 14141–14151.
36. Rajabi, N.; Kosecka, J. Q-GroundCAM: Quantifying Grounding in Vision Language Models via GradCAM. *arXiv* **2024**, arXiv:2404.19128.
37. Wang, S.; Kim, D.; Taalimi, A.; Sun, C.; Kuo, W. Learning Visual Grounding from Generative Vision and Language Model. *arXiv* **2024**, arXiv:2407.14563.
38. Quarteroni, S.; Dinarelli, M.; Riccardi, G. Ontology-Based Grounding of Spoken Language Understanding. In Proceedings of the 2009 IEEE Workshop on Automatic Speech Recognition & Understanding, Moreno, Italy, 13–17 December 2009; IEEE: New York, NY, USA, 2009; pp. 438–443.
39. Baldazzi, T.; Bellomarini, L.; Ceri, S.; Colombo, A.; Gentili, A.; Sallinger, E. Fine-Tuning Large Enterprise Language Models via Ontological Reasoning. In *International Joint Conference on Rules and Reasoning*; Springer Nature: Cham, Switzerland, 2023.
40. Jullien, M.; Valentino, M.; Freitas, A. Do Transformers Encode a Foundational Ontology? Probing Abstract Classes in Natural Language. *arXiv* **2022**, arXiv:2201.10262.
41. Larionov, D.; Shelmanov, A.; Chistova, E.; Smirnov, I. Semantic Role Labeling with Pretrained Language Models for Known and Unknown Predicates. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019), Varna, Bulgaria, 2–4 September 2019; Incoma Ltd.: Shoumen, Bulgaria, 2019; pp. 619–628.
42. Smith, M.K.; Welty, C.; McGuinness, D.L. *OWL Web Ontology Language Guide*; W3C: Cambridge, MA, USA, 2004.
43. 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, Paris, France, 2–6 October 2023.
44. Meta Platforms, Inc. Mapillary. 2024. Available online: https://www.mapillary.com/ (accessed on 20 November 2024).
45. Vestena, K. GitHub—Kauevestena/deep_pavements_dataset. GitHub. Available online: https://github.com/kauevestena/deep_pavements_dataset (accessed on 20 November 2024).
46. Fan, Q.; Tao, X.; Ke, L.; Ye, M.; Zhang, Y.; Wan, P.; Wang, Z.; Tai, Y.-W.; Tang, C.-K. Stable Segment Anything Model. *arXiv* **2023**, arXiv:2311.15776.
47. Hetang, C.; Xue, H.; Le, C.; Yue, T.; Wang, W.; He, Y. Segment Anything Model for Road Network Graph Extraction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle WA, USA, 17–21 June 2024.

48. Son, J.; Jung, H. Teacher–Student Model Using Grounding DINO and You Only Look Once for Multi-Sensor-Based Object Detection. *Appl. Sci.* **2024**, *14*, 2232. [CrossRef]
49. Dong, X.; Bao, J.; Zhang, T.; Chen, D.; Gu, S.; Zhang, W.; Yuan, L.; Chen, D.; Wen, F.; Yu, N. CLIP Itself Is a Strong Fine-Tuner: Achieving 85.7% and 88.0% Top-1 Accuracy with ViT-B and ViT-L on ImageNet. *arXiv* **2022**, arXiv:2212.06138.
50. Nguyen, T.; Ilharco, G.; Wortsman, M.; Oh, S.; Schmidt, L. Quality Not Quantity: On the Interaction between Dataset Design and Robustness of Clip. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 21455–21469.
51. Fang, A.; Ilharco, G.; Wortsman, M.; Wan, Y.; Shankar, V.; Dave, A.; Schmidt, L. Data Determines Distributional Robustness in Contrastive Language Image Pre-Training (Clip). In Proceedings of the International Conference on Machine Learning, PMLR, Baltimore, MD, USA, 17–23 July 2022; pp. 6216–6234.
52. Tu, W.; Deng, W.; Gedeon, T. A Closer Look at the Robustness of Contrastive Language-Image Pre-Training (Clip). *Adv. Neural Inf. Process. Syst.* **2024**. [CrossRef]
53. Mumuni, F.; Mumuni, A. Segment Anything Model for Automated Image Data Annotation: Empirical Studies Using Text Prompts from Grounding DINO. *arXiv* **2024**, arXiv:2406.19057.
54. Kaue-Vestena/Clip-Vit-Base-Patch32-Finetuned-Surface-Materials. Hugging Face. 2024. Available online: https://huggingface.co/kauevestena/clip-vit-base-patch32-finetuned-surface-materials (accessed on 20 November 2024).
55. Eimer, T.; Lindauer, M.; Raileanu, R. Hyperparameters in Reinforcement Learning and How To Tune Them. In Proceedings of the 40th International Conference on Machine Learning, PMLR, Honolulu, HI, USA, 23–29 July 2023.
56. Tong, Q.; Liang, G.; Bi, J. Calibrating the Adaptive Learning Rate to Improve Convergence of ADAM. *Neurocomputing* **2022**, *481*, 333–356. [CrossRef]
57. Reddi, S.J.; Kale, S.; Kumar, S. On the Convergence of Adam and Beyond. *arXiv* **2019**, arXiv:1904.09237.
58. Yu, J.; Wang, Z.; Vasudevan, V.; Yeung, L.; Seyedhosseini, M.; Wu, Y. CoCa: Contrastive Cap-Tioners Are Image-Text Foundation Models. *arXiv* **2022**, arXiv:2205.01917.
59. Code, P.W. Papers with Code—ImageNet Benchmark (Image Classification). *GitHub*. 2024. Available online: https://paperswithcode.com/sota/image-classification-on-imagenet?metric=GFLOPs (accessed on 20 November 2024).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.