technologies

Communication

Towards Safe Visual Navigation of a Wheelchair Using Landmark Detection

Christos Sevastopoulos, Mohammad Zaki Zadeh, Michail Theofanidis, Sneh Acharya, Nishi Patel and Fillia Makedon *

Department of Computer Science and Computer Engineering, University of Texas at Arlington, Arlington, TX 76019, USA
* Correspondence: makedon@uta.edu

Abstract: This article presents a method for extracting high-level semantic information through successful landmark detection using 2D RGB images. In particular, the focus is placed on the presence of particular labels (open path, humans, staircase, doorways, obstacles) in the encountered scene, which can be a fundamental source of information enhancing scene understanding and paving the path towards the safe navigation of the mobile unit. Experiments are conducted using a manual wheelchair to gather image instances from four indoor academic environments consisting of multiple labels. Afterwards, the fine-tuning of a pretrained vision transformer (ViT) is conducted, and the performance is evaluated through an ablation study versus well-established state-of-the-art deep architectures for image classification such as ResNet. Results show that the fine-tuned ViT outperforms all other deep convolutional architectures while achieving satisfactory levels of generalization.

Keywords: landmark detection; multilabel classification; wheelchair navigation

check for updates

Citation: Sevastopoulos, C.; Zadeh, M.Z.; Theofanidis, M.; Acharya, S.; Patel, N.; Makedon, F. Towards Safe Visual Navigation of a Wheelchair Using Landmark Detection. *Technologies* **2023**, *11*, 64. https://doi.org/10.3390/technologies11030064

Academic Editor: Luc de Witte

Received: 14 December 2022
Revised: 9 March 2023
Accepted: 16 March 2023
Published: 25 April 2023

Copyright: © 2023 by the authors. 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

Estimating traversable paths is of crucial importance for the safe and precise indoor navigation of mobile units. An abundant number of applications in robotics consider the concept of traversability estimation as the cornerstone of extracting semantic information for motion planning. Deciding about the navigability of an area depends not only on the terrain's physical properties, such as slope, roughness, surface condition but also on the mechanical characteristics of the mobile unit traversing it [1]. Since different environments illustrate diverse amounts of uncertainty, it becomes apparent that the effort to collect and interpret data from various sensor modalities can lead to further predicaments as a result of the type and the volume of data acquired.

Determining traversable paths have an immediate application in building navigation systems for smart and powered wheelchairs. This is due to the fact that wheelchair users often face maneuvering difficulties [2] when accomplishing daily tasks due to the presence of uneven and rough terrains [3], small corridors and doorways [4], and environments that are described by various levels of stochasticity, e.g., due to the presence of humans. Additionally, staircases have been traditionally problematic due to the geometric threats they exhibit and also for the difficulties they pose to 3D laser scanners [5].

This work's primary aim was to perform some preliminary experiments to extract high-level semantic information regarding the scene's traversability, based on the landmarks' relative position with respect to the vicinity of a manual wheelchair. The proposed multilabel classification system, using RGB images as input, aimed to efficiently detect the presence of particular labels (open path, humans, staircase, doorways, obstacles). This can be a fundamental source of information enhancing scene understanding. Hence, the data collection process takes into account all the characteristics associated with the object's appearance (geometrical features, volume, environment's illumination, etc.) but also the objects' relative position with respect to the proximity of the wheelchair.

Moreover, the suggested method can be an indispensable component along any sensing or control modules that compose the navigation system of the mobile unit. Specifically, we exploit the strengths of a wide-lens camera that can provide valuable insight about whether an object is an obstacle or not, since it considers more angles of the surroundings than the standard lens does. Leveraging the concept of transfer learning, a vision transformer (ViT) [6] is fine-tuned towards performing a multilabel classification on a small dataset with a mere number of labels. An initial framework is proposed, which, through the prism of multilabel image classification using wide-lens images, detects important landmarks for safe wheelchair navigation. The focus of the approach is on the relative position of a landmark encountered with regards to the proximity of the mobile unit. The rest of the paper is structured as follows: In Section 2, the related work revolving around the paper's axes of interest is discussed, Section 3 gives an overview of the implemented method, Sections 4 and 5 outline the experimental setup and the performed ablation study, respectively, and ultimately, Sections 6 and 7 discuss the results and the conclusions derived.

2. Related Work

Mounting sensors on the right locations on a wheelchair's body is of paramount importance towards detecting obstacles and performing simultaneous localization and mapping (SLAM) [2]. Vision sensors on wheelchairs have been utilized in a interconnected fashion along with various modalities such as laser [7], ultrasound [8], and tactile sensors [9]. Using wide-lens cameras on wheelchairs has been associated with endeavors in navigation and assistance [10–12], as well as object detection and localization [13]. Ultra wide-lens images, such as obtained from fisheye cameras, have been used in people detection methods [14], robot traversability estimation [15], SLAM [16], pedestrian/vehicle detection and tracking [17], and autonomous driving [18].

Unsupervised learning has shown great potential with transfer learning due to its capacity to learn specific features that can be proven advantageous for the final tuning on the downstream task [19,20].

Contrastive learning approaches portray the ability to create representations among similar and dissimilar images in an unsupervised fashion. Thus, they present the ability to facilitate the task of distinguishing between images and they have been employed in research works incorporating determining traversable regions [21] and designing local traversability models [22]. It has been shown that transfer learning approaches initially require a dataset of considerable size for the initial training (Kitti [23], ImageNet [24], etc.) before transferring features from a new domain to initialize an existing trained network and thus enhance the levels of generalization performance on new unseen data [25]. Research efforts in exploiting transfer learning involving wheelchairs have been exploring tasks such as surface detection while using different wheelchair units [26] and sidewalk classification [19].

Using pretrained transformers [27,28] acts as a vital tool in creating rich feature representations that can be utilized for fine-tuning with respect to the pertinent downstream tasks. In the field of mobile robotics, ViTs have shown remarkable performance in extracting semantic information for applications that include terrain classification [29], navigation [30], recognition [31], bird's eye view segmentation [32], and object detection [33]. Furthermore, vision transformers have shown remarkable results on image classification [34–36] tasks over methods such as convolutional neural networks (CNNs), as described by Raghu et al. [37]. An important property that a ViT displays is the fact that it can preserve input spatial information at its higher layers. This is what makes a ViT a more promising direction than ResNet, which is less spatially discriminative. Due to the ability to retain spatial information, the ViT is considered as the backbone of the method in conjunction with the fact that the relative position of landmarks in the dataset is the main source of semantic information of the encountered scene.

3. Method Description

Mobile units that operate according to the traditional sense–plan–control loop rely on vision systems that can accurately understand the environment to detect traversable paths and objects [38]. In this study, a methodology that detects meaningful landmarks and subsequently affects the mobile unit's decision-making is presented. Thus, this approach enhances safe navigation by providing scene information that accounts for the scene's traversability by indicating the presence (or not) of hazardous objects.

The gist of the proposed method relies on the use of a ViT encoder that consists of a sequence of self-attention and feed-forward layers. Specifically, we employed a ViT pretrained on ImageNet-21k using the generative, self-supervised learning method of masked autoencoders (MAE) [39], which exhibited major amounts of effectiveness in generalization. The MAE process included the following steps:

- An input image was masked at random locations at a high masking ratio, roughly 75%;
- An encoder (ViT) was applied on the visible parts of the image;
- The decoder operated on both the encoded paths and the masked tokens;
- Missing pixels were constructed.

After the pretraining process was complete, the decoder was discarded and the encoder was used for image classification tasks. Masked autoencoders exhibit the potential to learn visual scene semantics in a holistic manner, and thus they can act as a powerful pretraining method for this article's multilabel classification task. They have also shown substantial efficiency in transfer learning tasks such as object detection, instance segmentation, etc. We also experimented with the ViT-base-patch16-224 base model, that was pretrained on ImageNet-21k. This standard ViT was chosen since it could be supported by the available computational resources and could provide a comparison against ViT_{MAE}.

For the supervised fine-tuning, a projection head was used, consisting of two fully connected layers. It was trained on both positive and negative data. The size of the output feature vector of the ViT was 768x1, and it was subsequently passed to the projection head that eventually classified the encountered scene with respect to the candidate classes (open path, doorways, staircase, humans, obstacle) (Figure 1). This simple network structure was used to prevent any overfitting occurrences given the fact that only a small quantity of annotated data was used. The *BCEWithLogitsLoss* loss function was employed, which combined a sigmoid layer and the BCELoss in one single class:

$$l_c(x, y) = L = \{l_1, ...l_N\}^T, l_n = -w_n[y_n log\sigma(x_n) + (1 - y_n)(1 - log\sigma(x_n))] \quad (1)$$

The reason for selecting this particular version of BCELoss was that the sequence of the log-sum-exp trick offered room for improved numerical stability.

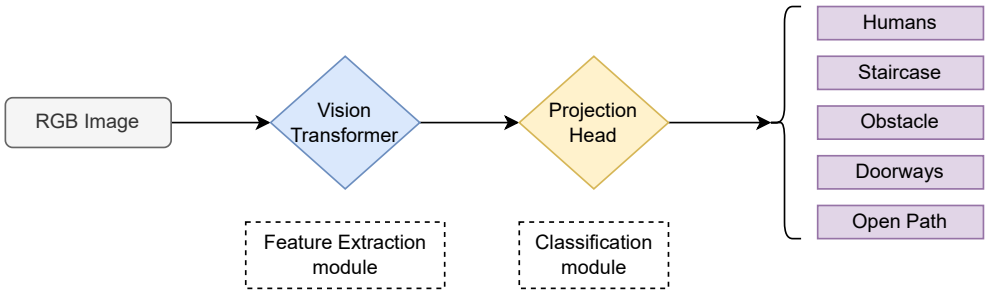

Figure 1. Pipeline of the proposed method.

Since a multilabel classification task was considered, the decision threshold value for each label needed to be carefully selected; by comparing against the probability value for each class label, it helped decide whether the encountered scene included that label or not. For the rest of the paper this threshold hyperparameter is denoted as τ. This threshold directly determined how conservative the method was towards the prediction of a certain label.

4. Experimental Setup

4.1. Hardware

Throughout the experimental process, a human operator navigated a standard wheelchair in four different buildings around the University of Texas, Arlington (UTA) campus. Data were recorded using a GoPro HERO10 camera, which recorded at 60 frames per second and was mounted on the wheelchair seat (Figure 2). For each building, the wheelchair was navigated in safe areas such as hallways and doorways, while encountering static (chairs, bins, tables, lockers) or dynamic (humans) obstacles. Moreover, ascending and descending staircases were targeted as additional areas of interest. Despite the fact that the environment was consistently academic, there were some distinct differences among the different buildings appearing in the dataset (Figure 3). Namely, by observing the four different buildings comprising our dataset, the following points were witnessed:

- Set 1: Hallway, desks, bright ambiance lighting, moving humans, wider staircases;
- Set 2: Hallway, desks/chairs, brick walls, static/moving humans, brighter ambiance lighting;
- Set 3: Normal ambiance lighting, moving humans, chairs/tables, narrower staircases;
- Set 4: Darker ambiance colors, bookshelves, conference room, desks.

Figure 2. The configuration used for the experiments consists of a GoPro HERO10 camera mounted on the seat of the manual wheelchair.

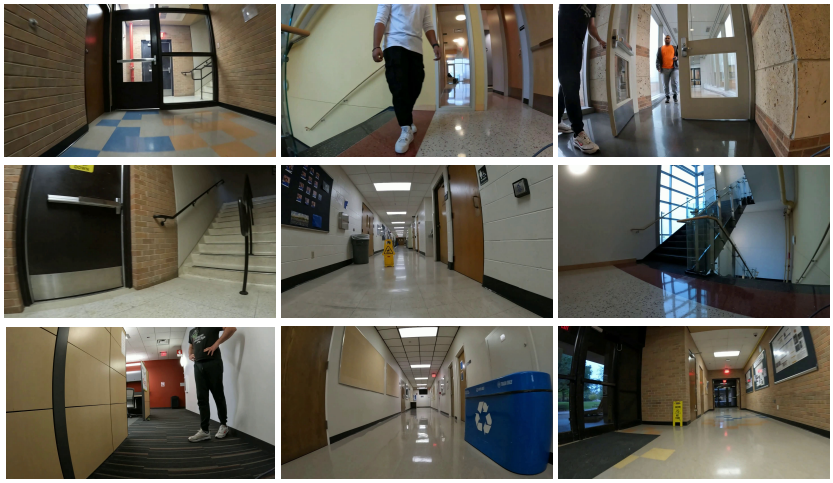

Figure 3. Various characteristic multilabeled scenes from the environment.

4.2. Data Collection and Processing

Data were recorded for approximately 150 min and created a dataset of 2704 images. The initial image size was 1920 × 1080 pixels before being resized to 224 × 224 pixels to match the resolution of the pretrained dataset. All images were manually labeled. The dataset included 2119 single-labeled images and 585 instances that comprised various combinations of the labels (open-path, humans, staircase, doorway, obstacles). Among the multilabeled images, 367 instances were described by two labels and 218 instances by three labels. Sets 1, 2, 3, 4 included 678, 697, 659, 670 image instances, respectively.

4.3. Fine-Tuning

For the conducted experiments, Pytorch (https://pytorch.org/, accessed on 13 December 2022) was used as the backbone framework. Training was done on a machine with two Titan RTX (24GB GDDR6 RAM, 4608 CUDA Cores) GPUs. Horizontal flips were performed as a means to augment the dataset. Training took place for 50 epochs using the BCE loss function, unless an early stopping callback terminated the trial upon observed convergence. Furthermore, the training parameters used were: batch size = 16, learning rate = 0.01, and weight decay = 5×10^{-4}. For the fine-tuning part, all transformer's deeper layers were frozen, and the classifier was replaced with two fully connected layers; the last one performed the classification. Layers were fine-tuned using stochastic gradient descent (SGD).

5. Ablation Study

To evaluate the performance of the proposed fine-tuned method on the custom dataset, an ablation study was conducted. A four-fold cross-validation was performed with three buildings selected for training and the remaining one for testing. The rationale behind folding on the buildings was to exploit the visual dissimilarity between semantically equivalent classes between buildings. This comparison helped us evaluate the ability of the proposed method to generalize beyond learning the visual representations of specific landmarks. Utilizing the same architecture for the projection head, a deep residual network (ResNet) [40] (ResNet50) that had been pretrained on ImageNet-21k was fine-tuned. The classifier was replaced with the projection head for the classification.

Additionally, a GAN ensemble network was trained following the methodology described by Hirose et al. in [15]. We used the GO Stanford (https://cvgl.stanford.edu/gonet/dataset, accessed on 13 December 2022) dataset and pretrained it on approximately 75k unlabeled fisheye images. Finally, a small convolutional network was trained, comprising four convolutional and two fully connected layers each followed by a ReLU activation function, except for the final layer. A Hamming loss was chosen as the performance metrics (as suggested in [41] since it only penalized the individual labels, and we experimented with

different values for τ. For both fine-tuned ViT and ResNet, the datasets that presented the highest (Set 4) and minimum (Set 3) hamming loss after performing four-fold cross validation were chosen. The results are shown in Figure 4.

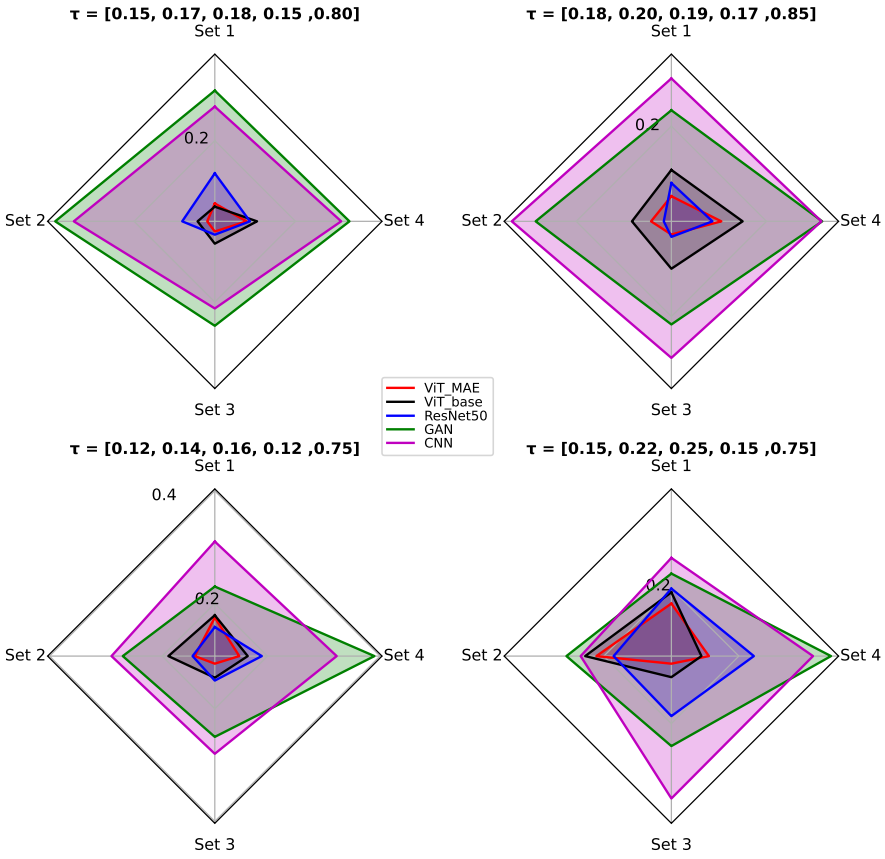

Figure 4. Methods' performance for various values of the threshold τ using the Hamming loss metric. A larger Hamming loss implies a lower network performance.

6. Results

The focus of this paper's approach was heavily dependent on landmarks' detection as this is crucial to ensure safe wheelchair navigation. The detection of staircases, humans, and miscellaneous static obstacles was prioritized by assigning a lower value for τ. Since humans' motion is governed by uncertainty and it is crucial to act in a conservative manner, given that predictions must align with the axis of safety, the best results, in terms of humans detection, were achieved when $\tau_{humans} = 0.15$. Similarly, the best detection results for staircases, static obstacles, doorways, and open paths were achieved when $\tau_{stairs} = 0.17$, $\tau_{obstacles} = 0.18$, $\tau_{doorways} = 0.15$, and $\tau_{open} = 0.80$, respectively.

Figure 4 presents the results of the ablation study. The fine-tuned ViT$_{MAE}$ outperformed all other networks while displaying critical levels of consistency. This was in agreement with the results from the literature [6,37] in which a ViT's performance can significantly outperform CNNs' in image classification tasks. This argument was also supported by the fact that the MAE training includes the notion of learning visual semantics holistically. With regards to the ViT-base-patch16-224 network, it did not demonstrate a significant improvement compared to ResNet50. GAN's performance was lower, due to the difficulty in training the ensemble's networks with an adequate number of data, whereas the custom fully supervised CNN did not exhibit a major amount of efficiency for practical tasks.

The lowest values of the Hamming loss, implying high levels of performance, were observed for Set 3. This was due to the fact that Set 3 displayed a considerable amount of balance with respect to varying illumination and object features. Contrariwise, Set 4 presented the largest amount of hamming loss because it was the one with the most uniquely distinct features in terms of visual information. Compared to the others sets, Set 4 was significantly more differentiated including the darkest illumination as well as areas with a dense concentration of bulky objects. The best performance of ViT$_{MAE}$ was achieved when using τ values = (0.15, 0.17, 0.18, 0.15, 0.80) for humans, staircases, static obstacles, doorways, and open paths, respectively.

Figure 5 displays a comparison between the Hamming loss as computed by fine-tuning the MAE and ResNet50 on Set 3 that exhibited the best performance. Specifically, the fine-tuned ViT$_{MAE}$ convincingly outperformed a fine-tuned ResNet50, with the performance margin, described by the Hamming loss, widening as the fraction of training data increased. Additionally, it was noticed that even for a small quantity of training data available, ViT$_{MAE}$'s Hamming loss was smaller than that of ResNet50. This showed that ViT$_{MAE}$ could be largely beneficial in scenarios where only a small number of training instances is available. In Figure 6, the recall was examined as observed in Set 3 for the images that included the "humans" label. ViT$_{MAE}$ consistently achieved a recall of around 86% for training sets larger than 40%, while ResNet50 achieved lower performance. Hence, it can be inferred that ViT$_{MAE}$ could sufficiently address the presence of humans in the scene. Overall, the attribute of our dataset that construed an object as an obstacle given its relative position seemed to be exploited at full extent with the use of a vision transformer pretrained with MAE.

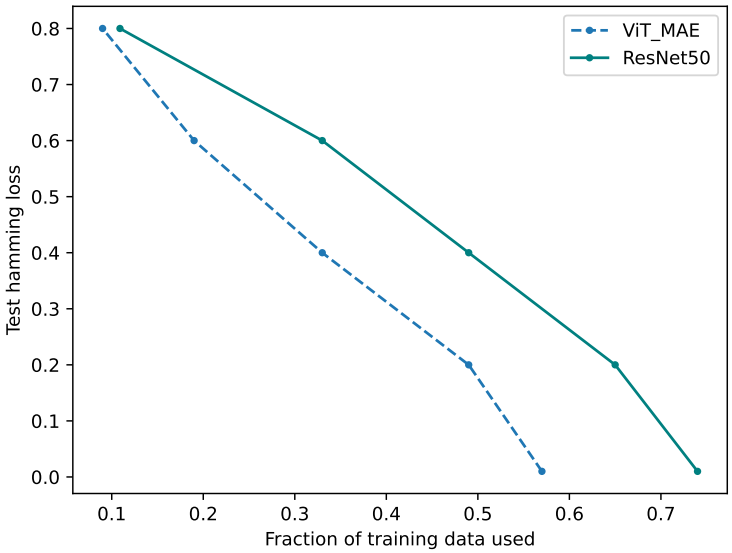

Figure 5. Graph of test Hamming loss against fraction of training data used for Set 3.

The confusion matrices depicted in Figure 7 provide an illustrative representation of the ViT$_{MAE}$'s best performance as noted on Set 3. Overall, the detection performance achieved high levels of efficiency. In addition, the results were consistent along the various labels irrespective of the notable differences among the sets, which were collected in different buildings. This can be attributed to the presence of pretrained self-attention layers along with the property that masked autoencoders portray, which is to learn visual scene semantics in a comprehensive manner. The aforementioned arguments reinforced the claim that ViTs provide generalizable solutions to the multilabel classification problem for small datasets.

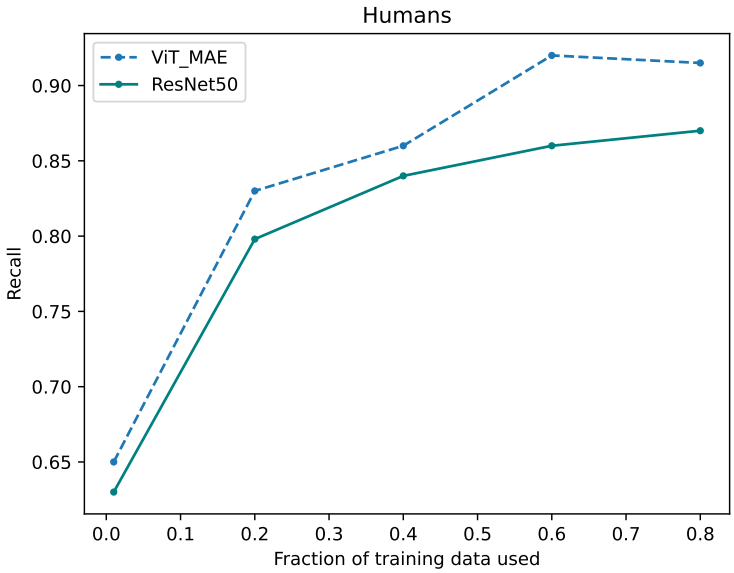

Figure 6. Comparison between the two prevalent fine-tuning methods for the "humans" label when testing on Set 3 for different quantities of training data.

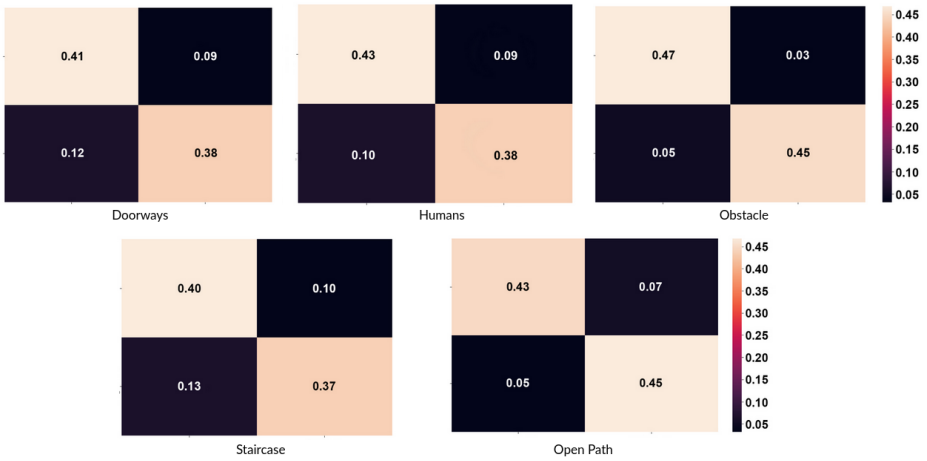

Figure 7. Confusion matrices for each label as observed in ViT$_{MAE}$'s best performance on Set 3.

7. Conclusions and Future Work

A method that extracted high-level semantic information regarding the scene's navigability through landmark detection was proposed. Experiments were conducted in different indoor environments using a manually driven wheelchair and a wide-lens camera. The results indicated that our multilabel classification method achieved a high performance without the loss of generalization and enriched scene understanding. Therefore, the proposed approach can act as a preceding step before designing the motion planning (autonomous or not) of a manual wheelchair.

Furthermore, the results showed that fine-tuning a vision transformer could act as a powerful tool for multilabel classification tasks in small datasets. We showed that fine-tuning a vision transformer pretrained with MAE led to a stronger performance compared to state-of-the-art deep architectures for image classification such as ResNet. Avenues for further research and improvement involve the utilization and fusion of additional modalities (depth, laser), which, along with RGB images, can lead to a deeper evaluation and understanding of the semanticity of the predicted scene labels.

Author Contributions: Conceptualization, C.S.; Methodology, C.S.; Software, S.A., N.P. and M.Z.Z.; Supervision, F.M.; Writing—review editing, C.S. and M.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Code and dataset available at https://github.com/ChristosSev/Towardssafe-visual-navigation-of-a-wheelchair-using-landmark-detection, accessed on 15 February 2023.

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

References

1. Sevastopoulos, C.; Konstantopoulos, S. A survey of traversability estimation for mobile robots. *IEEE Access* **2022**, *10*, 96331–96347. [CrossRef]
2. Leaman, J.; La, H.M. A comprehensive review of smart wheelchairs: Past, present, and future. *IEEE Trans. Hum.-Mach. Syst.* **2017**, *47*, 486–499. [CrossRef]
3. Podobnik, J.; Rejc, J.; Slajpah, S.; Munih, M.; Mihelj, M. All-terrain wheelchair: Increasing personal mobility with a powered wheel-track hybrid wheelchair. *IEEE Robot. Autom. Mag.* **2017**, *24*, 26–36. [CrossRef]
4. Pasteau, F.; Narayanan, V.K.; Babel, M.; Chaumette, F. A visual servoing approach for autonomous corridor following and doorway passing in a wheelchair. *Robot. Auton. Syst.* **2016**, *75*, 28–40. [CrossRef]
5. Delmerico, J.A.; Baran, D.; David, P.; Ryde, J.; Corso, J.J. Ascending stairway modeling from dense depth imagery for traversability analysis. In Proceedings of the 2013 IEEE International Conference on Robotics and Automation, Karlsruhe, Germany, 6–10 May 2013; pp. 2283–2290.
6. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
7. Trahanias, P.E.; Lourakis, M.I.; Argyros, S.; Orphanoudakis, S.C. Navigational support for robotic wheelchair platforms: An approach that combines vision and range sensors. In Proceedings of the International Conference on Robotics and Automation, Albuquerque, NM, USA, 25 April 1997; Volume 2, pp. 1265–1270.
8. Horn, O.; Kreutner, M. Smart wheelchair perception using odometry, ultrasound sensors, and camera. *Robotica* **2009**, *27*, 303–310. [CrossRef]
9. Trujillo-León, A.; Vidal-Verdú, F. Driving interface based on tactile sensors for electric wheelchairs or trolleys. *Sensors* **2014**, *14*, 2644–2662. [CrossRef]
10. Kurata, J.; Grattan, K.T.; Uchiyama, H. Navigation system for a mobile robot with a visual sensor using a fish-eye lens. *Rev. Sci. Instrum.* **1998**, *69*, 585–590. [CrossRef]
11. Ha, V.K.L.; Chai, R.; Nguyen, H.T. A telepresence wheelchair with 360-degree vision using WebRTC. *Appl. Sci.* **2020**, *10*, 369. [CrossRef]
12. Delmas, S.; Morbidi, F.; Caron, G.; Albrand, J.; Jeanne-Rose, M.; Devigne, L.; Babel, M. SpheriCol: A Driving Assistance System for Power Wheelchairs Based on Spherical Vision and Range Measurements. In Proceedings of the 2021 IEEE/SICE International Symposium on System Integration (SII), Iwaki, Japan, 11–14 January 2021; pp. 505–510.
13. Lecrosnier, L.; Khemmar, R.; Ragot, N.; Decoux, B.; Rossi, R.; Kefi, N.; Ertaud, J.Y. Deep learning-based object detection, localisation and tracking for smart wheelchair healthcare mobility. *Int. J. Environ. Res. Public Health* **2021**, *18*, 91. [CrossRef]
14. Duan, Z.; Tezcan, O.; Nakamura, H.; Ishwar, P.; Konrad, J. RAPiD: Rotation-aware people detection in overhead fisheye images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 636–637.
15. Hirose, N.; Sadeghian, A.; Vázquez, M.; Goebel, P.; Savarese, S. Gonet: A semi-supervised deep learning approach for traversability estimation. In Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, Spain, 1–5 October 2018; pp. 3044–3051.
16. Caruso, D.; Engel, J.; Cremers, D. Large-scale direct slam for omnidirectional cameras. In Proceedings of the 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Hamburg, Germany, 28 September–3 October 2015; pp. 141–148.
17. Bertozzi, M.; Castangia, L.; Cattani, S.; Prioletti, A.; Versari, P. 360 detection and tracking algorithm of both pedestrian and vehicle using fisheye images. In Proceedings of the 2015 IEEE Intelligent Vehicles Symposium (iv), Seoul, Republic of Korea, 28 June–1 July 2015; pp. 132–137.
18. Yogamani, S.; Hughes, C.; Horgan, J.; Sistu, G.; Varley, P.; O'Dea, D.; Uricár, M.; Milz, S.; Simon, M.; Amende, K.; et al. Woodscape: A multi-task, multi-camera fisheye dataset for autonomous driving. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9308–9318.
19. Yoon, H.Y.; Kim, J.H.; Jeong, J.W. Classification of the Sidewalk Condition Using Self-Supervised Transfer Learning for Wheelchair Safety Driving. *Sensors* **2022**, *22*, 380. [CrossRef] [PubMed]

20. Goh, E.; Chen, J.; Wilson, B. Mars Terrain Segmentation with Less Labels. *arXiv* **2022**, arXiv:2202.00791.
21. Gao, B.; Hu, S.; Zhao, X.; Zhao, H. Fine-grained off-road semantic segmentation and mapping via contrastive learning. In Proceedings of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Prague, Czech Republic, 27 September–1 October 2021; pp. 5950–5957.
22. Shah, D.; Levine, S. Viking: Vision-based kilometer-scale navigation with geographic hints. *arXiv* **2022**, arXiv:2202.11271.
23. Wang, W.; Wang, N.; Wu, X.; You, S.; Neumann, U. Self-paced cross-modality transfer learning for efficient road segmentation. In Proceedings of the 2017 IEEE International Conference on Robotics and Automation (ICRA), Singapore, 29 May–3 June 2017; pp. 1394–1401.
24. Huh, M.; Agrawal, P.; Efros, A.A. What makes ImageNet good for transfer learning? *arXiv* **2016**, arXiv:1608.08614.
25. Yosinski, J.; Clune, J.; Bengio, Y.; Lipson, H. How transferable are features in deep neural networks? *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 3320–3328. [CrossRef]
26. Mokrenko, V.; Yu, H.; Raychoudhury, V.; Edinger, J.; Smith, R.O.; Gani, M.O. A Transfer Learning Approach to Surface Detection for Accessible Routing for Wheelchair Users. In Proceedings of the 2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC), Madrid, Spain, 12–16 July 2021; pp. 794–803.
27. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
28. Chen, H.; Wang, Y.; Guo, T.; Xu, C.; Deng, Y.; Liu, Z.; Ma, S.; Xu, C.; Xu, C.; Gao, W. Pre-trained image processing transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 12299–12310.
29. Bednarek, M.; Łysakowski, M.; Bednarek, J.; Nowicki, M.R.; Walas, K. Fast haptic terrain classification for legged robots using transformer. In Proceedings of the 2021 European Conference on Mobile Robots (ECMR), Bonn, Germany, 31 August–3 September 2021; pp. 1–7.
30. Chen, K.; Chen, J.K.; Chuang, J.; Vázquez, M.; Savarese, S. Topological planning with transformers for vision-and-language navigation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 11276–11286.
31. Wang, R.; Shen, Y.; Zuo, W.; Zhou, S.; Zheng, N. TransVPR: Transformer-based place recognition with multi-level attention aggregation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 13648–13657.
32. Dutta, P.; Sistu, G.; Yogamani, S.; Galván, E.; McDonald, J. ViT-BEVSeg: A Hierarchical Transformer Network for Monocular Birds-Eye-View Segmentation. *arXiv* **2022**, arXiv:2205.15667.
33. Antonazzi, M.; Luperto, M.; Basilico, N.; Borghese, N.A. Enhancing Door Detection for Autonomous Mobile Robots with Environment-Specific Data Collection. *arXiv* **2022**, arXiv:2203.03959.
34. Bhojanapalli, S.; Chakrabarti, A.; Glasner, D.; Li, D.; Unterthiner, T.; Veit, A. Understanding robustness of transformers for image classification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10231–10241.
35. Chen, C.F.R.; Fan, Q.; Panda, R. Crossvit: Cross-attention multi-scale vision transformer for image classification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 357–366.
36. Chen, X.; Hsieh, C.J.; Gong, B. When vision transformers outperform ResNets without pre-training or strong data augmentations. *arXiv* **2021**, arXiv:2106.01548.
37. Raghu, M.; Unterthiner, T.; Kornblith, S.; Zhang, C.; Dosovitskiy, A. Do vision transformers see like convolutional neural networks? *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12116–12128.
38. Beer, J.M.; Fisk, A.D.; Rogers, W.A. Toward a framework for levels of robot autonomy in human-robot interaction. *J. Hum.-Robot Interact.* **2014**, *3*, 74. [PubMed]
39. He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; Girshick, R. Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 16000–16009.
40. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
41. Tsoumakas, G.; Katakis, I. Multi-label classification: An overview. *Int. J. Data Warehous. Min. (IJDWM)* **2007**, *3*, 1–13. [CrossRef]

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.