An RGB-D Fusion System for Indoor Wheelchair Navigation

Christos Sevastopoulos
christos.sevastopoulos@mavs.uta.edu
University of Texas at Arlington
USA

Sneh Acharya
sxa6003@mavs.uta.edu
University of Texas at Arlington
USA

Fillia Makedon
makedon@uta.edu
University of Texas at Arlington
USA

ABSTRACT

We present a method for extracting high-level semantic information through successful landmark detection using feature fusion between RGB and depth information. We focus on the classification of specific labels (open path, humans, staircases, doorways, obstacles) in the encountered scene, which can be a fundamental source of information enhancing scene understanding, and acting towards the safe navigation of the mobile unit. Experiments are conducted using a manual wheelchair equipped with a stereo RGB-D camera that captures image instances consisting of multiple labels before fine-tuning on a pre-trained Vision Transformer (ViT).

CCS CONCEPTS

• **Computer systems organization** → **Navigation Systems**; • **Computing methodologies** → *Semi-Supervised learning*.

KEYWORDS

Wheelchair navigation, Multi-label classification

ACM Reference Format:
Christos Sevastopoulos, Sneh Acharya, and Fillia Makedon. 2023. An RGB-D Fusion System for Indoor Wheelchair Navigation. In *Proceedings of the 16th International Conference on PErvasive Technologies Related to Assistive Environments (PETRA '23), July 05–07, 2023, Corfu, Greece.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3594806.3594851

1 INTRODUCTION

Identifying accessible routing through vision sensors, has an immediate implementation on building navigation systems for smart and powered wheelchairs. Wheelchair users face an array of challenges [17] in accomplishing daily tasks. This can be pertaining to the presence of uneven and rough terrains [20], small corridors and doorways [19], and also stochastic environments depicted by uncertainty e.g due to the presence of humans. Furthermore, staircases have been traditionally problematic due to the geometric threats they present [10].

In this article we aim to perform some preliminary experiments to extract high-level semantic information regarding the scene's navigability, based on the landmarks' relative position with respect to the vicinity of a manual wheelchair. The proposed multi-label classification system, using both RGB and depth input, aims to efficiently detect the presence of particular labels (open path, humans, staircase, doorways, obstacles) by fusing the information from the two aforementioned modalities. Integrating depth with RGB information has been shown to enhance the performance of image classification tasks due to the additional semantic information that the depth channel provides [13]. Despite the fact that our dataset is relatively small, we collect data instances combining 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. We aim to leverage the strengths of a fused system consisting of RGB and depth information, in order to to enhance scene perception by critically identifying the presence of obstacles or not.

Exploiting the concept of transfer learning, we fine-tune a Vision Transformer (ViT) [12] towards performing multi-label classification on a small indoors dataset. We propose a framework that, through the viewpoint of multi-label image classification, can detect important landmarks for wheelchair navigation. Our approach focus is on the relative position of a landmark encountered with regards to the proximity of the mobile unit.

2 RELATED WORK

In order to increase the levels of perception towards safe navigation, the use of cameras on wheelchairs [17] has been examined in conjunction with various modalities such as laser [23], ultrasound [16] and tactile sensors [24]. Pre-trained transformers [11], [7] act 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 been the state-of-the-art method exhibiting vast amounts of efficiency for applications that include object detection [1], terrain classification [2], navigation [8] and recognition [27]. Furthermore, Vision Transformers have shown remarkable results on image classification [4, 6, 9] tasks over methods such as Convolutional Neural Networks (CNNs) as described by Raghu et al. [21]. An important property that a ViT displays, is the fact that it can preserve input spatial information at its higher layers. This is what renders ViT as a more promising direction than ResNet which is less spatially discriminative.

Recent transformer-based depth estimation methods [3], [28], [22] have been employed for pixel-wise prediction. Liu et al. [18] propose a a cross-modalfusion framework for RGB-X semantic segmentation, where X is any additional modality. Multi-Head Self-Attention [26] can be a powerful tool in controlling the mixture of information among parts of an input sequence and thus leading to richer representations. As described in the work of Tsai et al. [25], the multi-head cross-modal attention module is responsible for updating each modality's sequence(in their case video, audio, and language) via low-level external information. Eventually, they infer

PETRA '23, July 05–07, 2023, Corfu, Greece

ACM ISBN 979-8-4007-0069-9/23/07.
https://doi.org/10.1145/3594806.3594851

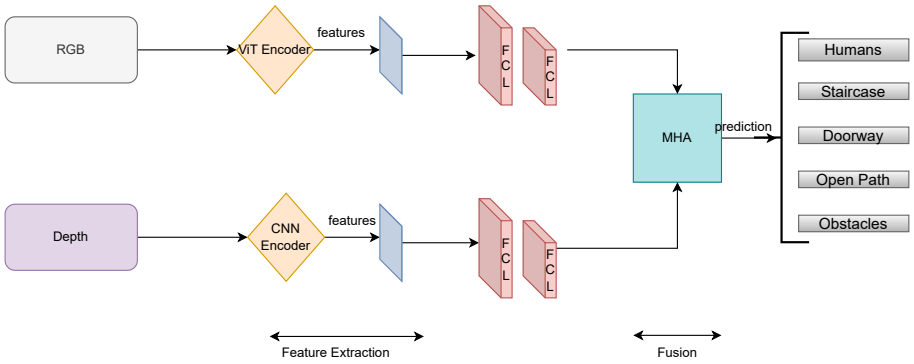

Figure 1: Overall Architecture

that the cross-modal transformer learns to correlate meaningful elements across different modalities. Other endeavors in robotics fusing modalities through the use of the MHSA module include, natural instructions and navigation graph [5], multi-robot collaboration for unknown exploration [30], UAV-driven segmentation [29] etc. Since ViT has shown remarkable performance in maintaining spatial information [21], we consider it as the backbone of our method. Additionally, the Multi-Head Self-Attention module has been shown great efficiency in modality fusion, and in this regard we aim to fuse RGB and depth information as a means to enrich the source of semantic information of the encountered scene.

3 METHOD DESCRIPTION

The methodology pursued in this article aims, by fusing modalities, to identify meaningful landmarks that the mobile unit is encountering. Thus, our approach aims to act towards safe wheelchair navigation by providing scene information regarding the presence of obstacles or not.

We are using a ViT pre-trained on ImageNet-21k using the generative, self-supervised learning method of Masked Autoencoders(MAE) [14] that has exhibited major amounts of effectiveness in generalization. The MAE process includes the following steps:

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

After the pre-training process is complete, the decoder is discarded and the encoder is used for image classification tasks. Masked Autoencoders exhibit the potential to learn visual scene semantics in a holistic manner and thus they act as a powerful pre-training method for our multi-label classification task. They have also shown substantial efficiency in transfer learning tasks such as object detection, instance segmentation etc.

The Multi-Head Self-Attention (MHSA) mechanism [26] obtains a number of different representations (as many as the heads h) of (Query, Key, Value), it then computes scaled dot-product attention for each representation, concatenates the results, and projects the concatenation through a feed forward layer. MHSA finds keys that matches the query, and gets the values of those keys. Intuitively, the rationale behind choosing multiple attention heads module is that it allows operating on parts of the given sequence differently (for instance longer-term dependencies versus shorter-term dependencies).

With regards to the supervised fine-tuning, the output feature vector of ViT is 768x1 for each modality and it is then passed to a projection head, consisting of two fully-connected (FC) layers. Depth images are fed to a CNN Encoder consisting of four convolutional layers of four convolutional and two fully connected layers that each, except for the final, is followed by a ReLu activation function. The Depth output feature vector is of 128x1 dimension. RGB and depth features are fed to each FC layer matching their dimensions before getting fused using the MHSA module with two heads. Afterwards, the fused features are fed to a linear classifier that classifies the encountered scene with respect to the candidate classes (open path, doorways, staircase, humans, obstacle) (Figure 1). We are using this simple network structure to prevent any overfitting since our dataset is relatively small. We use the *BCEWithLogitsLoss* loss function which combines a Sigmoid layer and the BCELoss in one single class.

The reason for selecting this particular version of BCELoss is that the sequence of the log-sum-exp trick offers room for improved numerical stability. Due to the fact that we are addressing a multi-label classification task, we need to determine a decision threshold value for each label, that by evaluating the probability value for each class label, decides whether the encountered scene includes this label or not. For the rest of the paper we denote this threshold hyper-parameter as d. This threshold directly determines how conservative our method is towards the prediction of a certain label.

4 EXPERIMENTAL SETUP

4.1 Data collection and processing

Throughout the experimental process, a human operator navigated a standard wheelchair in three different buildings around the University of Texas, Arlington (UTA) campus. For each building, we navigated the wheelchair in safe areas such as hallways, ascending and descending staircases, doorways while encountering static (chairs, bins, tables, lockers) or dynamic (humans) obstacles. With respect to the data gathering process, we mounted an OAK-D depth

camera on a manual wheelchair (Figure 2). The camera captures both RGB and depth data simultaneously and processes the data using a Python script along with the OpenCV[1] and depthai[2]. RGB stereo and depth data (Figure 3) are captured in real-time with a size of 640x480 pixels and a frame rate of 30 frames per second. Data were recorded for approximately 110 minutes and created a dataset of 12610 images. All images were manually labeled and then resized to 224x224 pixels, to match the resolution of the pretrained network. The dataset includes 8965 single-labeled images and 3645 instances that comprise of various combinations of the labels (open-path, humans, staircase, doorway, obstacles). Among the multi-labeled images, we notice 3219 two-labeled and 426 are described by three-labels in total. Sets 1, 2, 3 include 4243, 4078, 4289 image instances respectively. As far as the distinguishing features that each of the three sets presents, we observe the following: Set 1 includes dark ambience colours, voluminous objects, wide staircases, moving humans. Set 2 includes scenes of bright illumination, desks/chairs, brick walls while finally Set 3 presents more balanced ambience lighting, chairs/tables moving humans and narrow staircases.

4.2 Fine-tuning

For the experiments we used the Pytorch[3] framework. Training was done on a machine with 2 Titan RTX (24GB GDDR6 RAM, 4608 CUDA Cores) GPUs. We performed horizontal flip as a means to augment the dataset. We trained for 50 epochs, using the BCE loss function unless an early stopping callback terminated the trial upon observed convergence. Furthermore, as training parameters we used: batch size = 16, learning rate = 0.01 and weight decay = 5e-4. For the fine-tuning part, we freeze all transformer's deeper layers and replace the classifier with two fully-connected layers; the last one performs the classification. We fine-tuned the layers using stochastic gradient descent (SGD).

[1]https://opencv.org/
[2]https://docs.luxonis.com/en/latest/libraries
[3]https://pytorch.org/

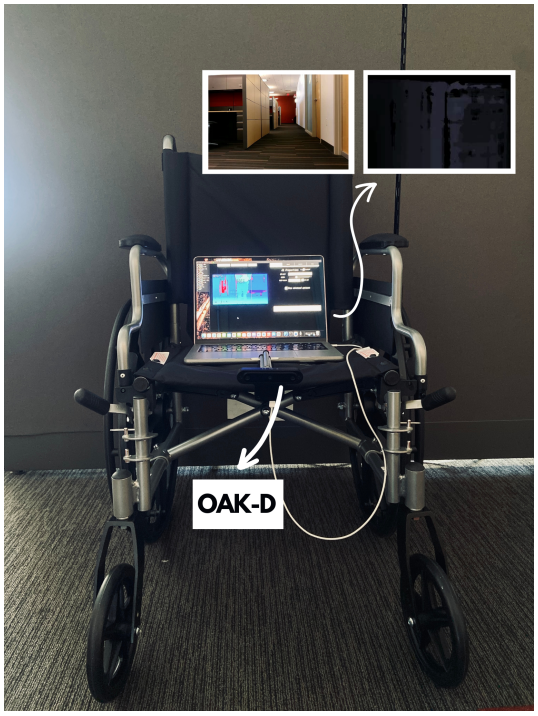

Figure 2: Wheelchair setup for data collection

4.3 Ablation Study

We perform an ablation study to evaluate the performance of the proposed fine-tuned method on our dataset. We perform 3-fold cross validation on three buildings selected for training and the remaining one for testing. The rationale behind folding on the buildings is to exploit the visual dissimilarity between semantically equivalent classes between buildings. This comparison is going to help us evaluate the ability of the proposed method to generalize beyond learning visual representations of specific landmarks. We also fine-tune, utilizing the same architecture for the projection head, a deep residual network (ResNet) [15], in particular the ResNet50 variant, that has been pre-trained on ImageNet-21k. We replace the classifier with the projection head for the multi-label classification.

5 RESULTS AND DISCUSSION

Our method's aim is to perform efficient landmarks' detection towards safe wheelchair navigation. For the detection of staircases, humans and static obstacles, we assign a lower value for *d*. Since

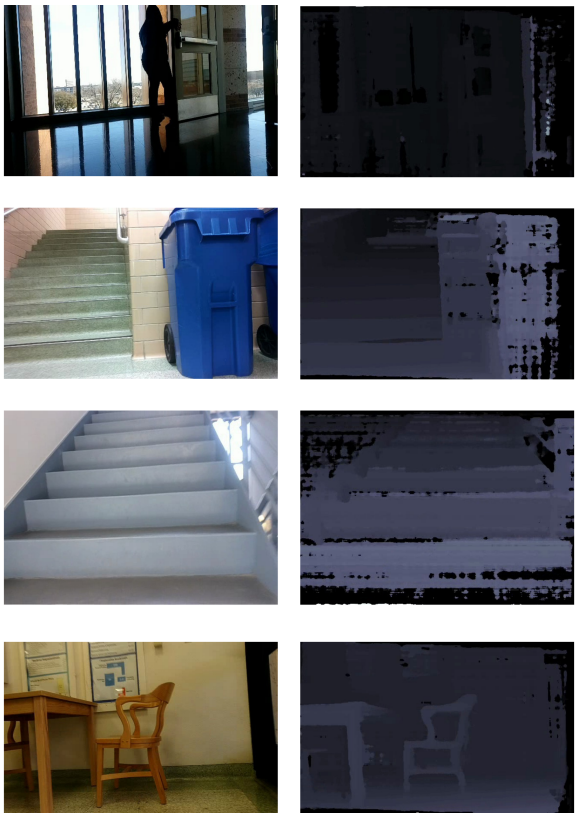

Figure 3: Examples of RGB and Depth pairs

Table 1: 3-fold cross-validation results

Hamming Loss[%]	Testing on Set 1	Testing on Set 2	Testing on Set 3
ViT_{RGB-D}	12.6	9.7	11.2
ViT_{RGB}	14.9	12.5	13.1
ResNet50_{RGB-D}	15.1	15.9	16.7
ResNet50_{RGB}	17.1	16.0	17.4

humans' moves can be unpredictable, we assign a lower threshold value for humans' detection. The best results were achieved when $d_{humans} = 0.12$. Likewise, the best detection results for staircases, doorways, obstacles and open paths were achieved when $d_{staircases} = 0.16$, $d_{doorways} = 0.16$, $d_{obstacles} = 0.17$, $d_{open} = 0.85$ respectively. Table 1 presents the results of the ablation study. We can notice that the fused ViT_{RGB-D} outperforms all other networks while displaying critical levels of consistency across all three sets. This observation can be supported by results in literature in which ViT's performance is significantly increased due to the: 1)the depth integration [13] and 2) the argument that ViT can outrun CNNs in image classification tasks [12], [21]. This argument is also supported by the fact that pre-training with Masked Autoencoders includes the notion of learning visual semantics holistically.

The lowest values of hamming loss which imply higher levels of performance, are observed for Set 2. This is due to the fact that Set 2 displays considerable amounts of balance with respect to varying illumination and object features. Contrariwise, Set 1 presents the largest amounts of hamming loss because it is the one with the most uniquely distinct features in terms of visual information. Compared to the others sets, Set 1 is significantly more differentiated than Sets 2 and 3 due to the presence of more voluminous objects as well as darker illumination. Figure 4 displays a comparison between the hamming loss as computed by fine-tuning the ViT_{RGB-D} and ResNet50_{RGB-D} on Set 2 that exhibits the best overall performance. In specific, the fine-tuned $_{RGB-D}$ convincingly outperforms fine-tuned ResNet50_{RGB-D}, with the performance margin, described by the hamming loss, widening as the fraction of training data increases. Moreover, we notice that even for a small amount of training data available, ViT_{RGB-D}'s hamming loss is smaller than the ResNet50_{RGB-D} one. This shows that integration of RGB with depth information for ViT, can be largely beneficial in scenarios where only a small amount of training instances is available. Figure **??** presents the recall performance as observed in Set 2 for images that include the "humans" label and the "staircases" label. It can be inferred that the utilization of depth information seems to have a substantial effect on the recall rates especially while the fraction of training data is getting increased.

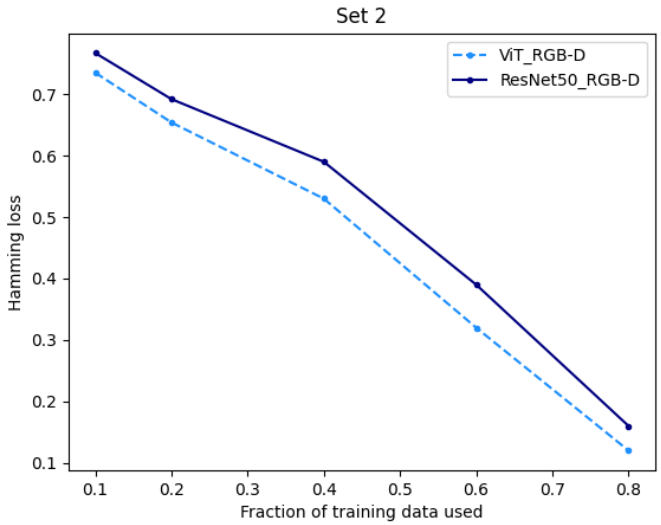

Figure 4: Graph of test hamming loss against between ViT_{RGB-D} and ResNet50_{RGB-D} with respect to the fraction of training data used for Set 2

6 CONCLUSIONS

We propose a RGB-D fusion method that extracts high-level semantic information regarding the scene's navigability for a wheelchair through landmark detection. Experiments were conducted in different indoors environments using a manually driven wheelchair and the OAK-D camera. The results present an improvement on multi-label classification when fusing with depth information rather than solely relying on RGB. Additionally, it is shown that fine-tuning a Vision Transformer can act as a powerful tool for multi-label classification tasks in small datasets. We show that fine-tuning a Vision Transformer on RGB-D information pre-trained with MAE, led to a stronger performance compared to state-of-the-art deep architecture for image classification such as ResNet. Avenues for further research include experimenting with more instances and different Vision Transformer architectures.

REFERENCES

[1] Michele Antonazzi, Matteo Luperto, Nicola Basilico, and N Alberto Borghese. 2022. Enhancing Door Detection for Autonomous Mobile Robots with Environment-Specific Data Collection. *arXiv preprint arXiv:2203.03959* (2022).
[2] Michał Bednarek, Mikołaj Łysakowski, Jakub Bednarek, Michał R Nowicki, and Krzysztof Walas. 2021. Fast haptic terrain classification for legged robots using transformer. In *2021 European Conference on Mobile Robots (ECMR)*. IEEE, 1–7.
[3] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. 2021. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4009–4018.
[4] Srinadh Bhojanapalli, Ayan Chakrabarti, Daniel Glasner, Daliang Li, Thomas Unterthiner, and Andreas Veit. 2021. Understanding robustness of transformers for image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10231–10241.
[5] Patricio Cerda-Mardini, Vladimir Araujo, and Alvaro Soto. 2020. Translating natural language instructions for behavioral robot navigation with a multi-head attention mechanism. *arXiv preprint arXiv:2006.00697* (2020).
[6] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. 2021. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*. 357–366.
[7] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. 2021. Pre-trained image

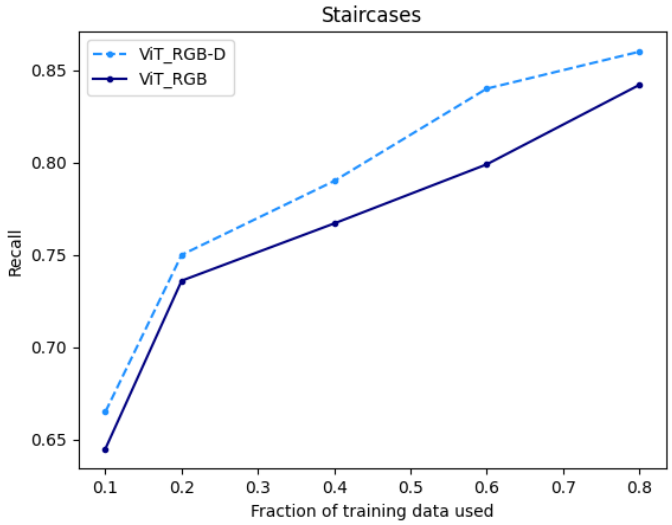

Figure 5: Recall performance with depth integration against only RGB as noted for the "Staircases" labels

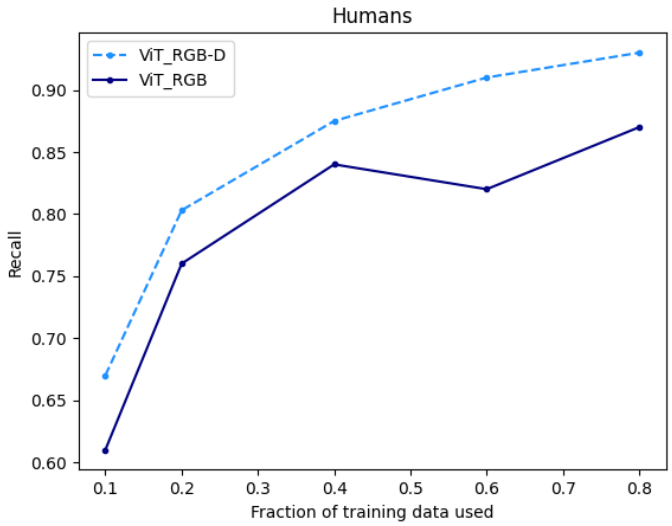

Figure 6: Recall performance with depth integration against only RGB as noted for the "Humans" labels

processing transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12299–12310.

[8] Kevin Chen, Junshen K Chen, Jo Chuang, Marynel Vázquez, and Silvio Savarese. 2021. Topological planning with transformers for vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11276–11286.

[9] Xiangning Chen, Cho-Jui Hsieh, and Boqing Gong. 2021. When vision transformers outperform ResNets without pre-training or strong data augmentations. *arXiv preprint arXiv:2106.01548* (2021).

[10] Jeffrey A Delmerico, David Baran, Philip David, Julian Ryde, and Jason J Corso. 2013. Ascending stairway modeling from dense depth imagery for traversability analysis. In *2013 IEEE International Conference on Robotics and Automation*. IEEE, 2283–2290.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[13] Caner Hazirbas, Lingni Ma, Csaba Domokos, and Daniel Cremers. 2017. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In *Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part I 13*. Springer, 213–228.

[14] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16000–16009.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[16] Odile Horn and M Kreutner. 2009. Smart wheelchair perception using odometry, ultrasound sensors, and camera. *Robotica* 27, 2 (2009), 303–310.

[17] Jesse Leaman and Hung Manh La. 2017. A comprehensive review of smart wheelchairs: past, present, and future. *IEEE Transactions on Human-Machine Systems* 47, 4 (2017), 486–499.

[18] Huayao Liu, Jiaming Zhang, Kailun Yang, Xinxin Hu, and Rainer Stiefelhagen. 2022. CMX: Cross-modal fusion for RGB-X semantic segmentation with transformers. *arXiv preprint arXiv:2203.04838* (2022).

[19] François Pasteau, Vishnu K Narayanan, Marie Babel, and François Chaumette. 2016. A visual servoing approach for autonomous corridor following and doorway passing in a wheelchair. *Robotics and Autonomous Systems* 75 (2016), 28–40.

[20] Janez Podobnik, Jure Rejc, Sebastjan Slajpah, Marko Munih, and Matjaz Mihelj. 2017. All-terrain wheelchair: Increasing personal mobility with a powered wheel-track hybrid wheelchair. *IEEE Robotics & Automation Magazine* 24, 4 (2017), 26–36.

[21] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. 2021. Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems* 34 (2021), 12116–12128.

[22] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. 2021. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 12179–12188.

[23] Panos E Trahanias, Manolis IA Lourakis, SA Argyros, and Stelios C Orphanoudakis. 1997. Navigational support for robotic wheelchair platforms: an approach that combines vision and range sensors. In *Proceedings of International Conference on Robotics and Automation*, Vol. 2. IEEE, 1265–1270.

[24] Andrés Trujillo-León and Fernando Vidal-Verdú. 2014. Driving interface based on tactile sensors for electric wheelchairs or trolleys. *Sensors* 14, 2 (2014), 2644–2662.

[25] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, Vol. 2019. NIH Public Access, 6558.

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[27] Ruotong Wang, Yanqing Shen, Weiliang Zuo, Sanping Zhou, and Nanning Zheng. 2022. TransVPR: Transformer-based place recognition with multi-level attention aggregation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13648–13657.

[28] Guanglei Yang, Hao Tang, Mingli Ding, Nicu Sebe, and Elisa Ricci. 2021. Transformer-based attention networks for continuous pixel-wise prediction. In *Proceedings of the IEEE/CVF International Conference on Computer vision*. 16269–16279.

[29] Shi Yi, Xi Liu, Junjie Li, and Ling Chen. 2023. UAVformer: A Composite Transformer Network for Urban Scene Segmentation of UAV Images. *Pattern Recognition* 133 (2023), 109019.

[30] Chao Yu, Xinyi Yang, Jiaxuan Gao, Jiayu Chen, Yunfei Li, Jijia Liu, Yunfei Xiang, Ruixin Huang, Huazhong Yang, Yi Wu, et al. 2023. Asynchronous Multi-Agent Reinforcement Learning for Efficient Real-Time Multi-Robot Cooperative Exploration. *arXiv preprint arXiv:2301.03398* (2023).