Proceedings of the Canadian Society for Mechanical Engineering International Congress
32nd Annual Conference of the Computational Fluid Dynamics Society of Canada
Canadian Society of Rheology Symposium
CSME/CFDSC/CSR 2025
May 25–28, 2025, Montréal, Québec, Canada

Development of a dual attention model for soil type semantic segmentation in construction environments

Mitchell Vella[1], Jaho Seo[1*], Hanmin Lee[2], Ha-Young Jang[3]

[1]Department of Automotive and Mechatronics Engineering, Ontario Tech University, Oshawa, Canada
[2]Department of Industrial Machinery DX, Korea Institute of Machinery and Materials (KIMM), Daejeon, Republic of Korea
[3]SURROMIND, Seoul, Republic of Korea
[*]Jaho.Seo@ontariotechu.ca

***Abstract*—**This paper studies the development of a machine vision model to classify soil types commonly encountered in construction environments. The aim of this research is to create an implementation that can aid autonomous heavy machinery operating in unregulated environments by informing task management systems of surface properties and material locations. The model proposed – which we call DaViT-Dilate — is a modification of Dual Attention Visual Transformer (DaViT) that replaces the channel attention with Dilate attention and introduces a heterogeneous layer structure. The results show that these modifications result in drastic improvements in all metrics on our dataset, with minimal cases where SegFormer or the original DaViT performs better. This model would allow a lowered amount of necessary supervision and setup in new construction locations for simple tasks, freeing labour for complex operations.

***Keywords-component*—soil classification; construction; semantic segmentation; transformers; computer vision**

I. INTRODUCTION

Autonomous vehicles used in constrained environments can easily rely on simple obstacle detection for most navigation tasks, as the surface they travel on is generally reliable. Autonomous heavy machinery (AHM) must operate in environments that are significantly less reliable, especially when performing operations such as digging or lifting where stable footing is necessary. For this purpose, a machine vision model can be implemented to determine the category of soils in the operating envelope. This has a few benefits over simply having an operator dictate acceptable footing locations: the AHM can operate in diverse locations with little setup; material used for backfill does not have to be explicitly located for task completion; instead a material area and type can be specified; newly unstable surfaces can be avoided in changing ground environments; expected digging characteristics can be determined and below-surface impacts with unexpected obstructions can be more easily detected.

The focus of this initial research is to determine only the category of soil within one of the following: Dry Soil, Sand, Stone, Water, Wet Soil. Approximately 1000 images were collected from a variety of sources and manually labeled, then augmentation was applied for approximately 3000 images total. Models were tested against this dataset after being pre-trained on

The rest of this paper is organized as follows: Existing research briefly, methodology, results, and conclusions.

II. PRIOR RESEARCH

Most of the research that pertains to soil classification or segmentation is focused on lab-based methods; however, there has been recent interest in AHM from the construction, space, and land surveying industries, and research has begun to come out which focuses on segmenting soil in the field. One of the earlier works from Du et al. explores the use of a convolutional neural network (CNN) based model to segment mountainous regions for the risk of landslides [1]. This is considered especially difficult as landslide identification is usually done by surveying. The paper concluded that DeepLabV3 [2] had the best generalization capabilities, but was outperformed by GCN [3] on some landslide types.

A more recent paper from Zamini et al. has also explored the use of DeepLabV3 on soil segmentation tasks. The paper applied various decoder modifications to determine which architecture has the best performance. It found that DeepLabV3 is capable of very robust soil segmentation against their dataset, achieving 90% Jaccard index in many complex scenes. Ultimately, it was determined that the base model performed the most reliably, while those with the modified decoder tended to favor specific classes based on their architecture.

Liu et al. researched soil segmentation for navigation of a mars rover, making it the first soil segmentation task to utilize transformers [4]. The model that they implemented is a hybrid attention model that used a channel attention and a channel group attention mechanism in parallel. They then utilized a complex loss function in order to combine the two attentions together for the final segmentation. It was shown that this method was exceedingly effective at soil segmentation tasks, achieving 59.5 mIoU on their dataset, outperforming other tested models.

Research in this area and typical computer vision tasks is still dominated by CNNs due to their superior local contextual awareness. While transformers are typically targeted at tasks that favor global context, previous research has shown that they can perform well in a local context with some drawbacks [5], [6]. Increasingly, researchers have utilized more complex attention mechanisms in order to combat these deficiencies [7], [8]. These come at the cost of increased computational complexity and diminished global awareness. Dual attention models introduce the possibility of obtaining both global and local awareness with minimal additional complexity when run in parallel.

Several individual attention mechanisms have been introduced since this research to attempt to retain the long-range connectivity while limiting the computational time required – such as 'Shift' [6], 'Overlapping Patches' [7], and 'Sparse' [9], [5]. These single attention models present trade-offs between density, global context capture, and complexity, often leading to a reduction in performance against some tasks or with loss to real-time performance. An architecture that promises to limit these trade-offs is the Dual Attention Visual Transformer (DaViT), which utilizes two attention mechanisms per layer applied alternately: a 'channel' based self-attention, and a 'channel group' based self-attention. The channel based self-attention is similar to sliding window attention, and provides local context, while the channel group-based self-attention attends across the channels to provide global context. This means that the global attention mechanism is DaViT performs very well on a wide range of tasks with minimal additional computational increase due to the simplified global attention mechanism [10]. This model is similar to that introduced in [4], but uses a significantly simplified loss function, which reduces the computational complexity but does not grossly reduce performance.

III. Methodology

Within soil detection, the very small surface texture differences are important locally, but the scene also informs individual segmentation regions. The proposed DaViT-based architecture attempts to improve performance in this difficult task by addressing some pitfalls in the original model.

A. Model Architecture

While the Segformer architecture performs well for this task under most circumstances, it notably performs poorly when differentiating similar soil types such as sand and gravel, or wet soil and water. The DaViT model performs better on these tasks, within a few percentage points of Segformer under some circumstances but often performs significantly better, and infrequently performs worse. To improve the performance of the DaViT model further, we propose a modification that replaces the channel attention with the Dilate attention from [11].

Fig. 1 shows the proposed modifications to the stage and block structures. The beginning of each stage contains a common patch embedding layer and feeds the embeddings to the next block. While the original DaViT utilized only dual attention blocks, the proposed model allows for defining heterogeneous architectures containing arbitrary stacks of dual or dilate-only blocks. At the block level, a block can consist of either a dilated self-attention and channel group self-attention layer, or only a dilated self-attention layer with layer norm and FFN implemented between layers and the end of the block. While this architecture suggests that FFNs may not be required due to the improved global performance, this was not explored in this paper.

The DilateFormer from Hassani et al. in [11] performs very well on a variety of tasks while maintaining low computational complexity through the use of a sliding window based dilated attention. The general working principle is similar to sliding window attention, aiming to attend to across the entire image using patches of fixed sizes that do not overlap; however, dilated attention differs in that the window begins as a solid dense patch, but implements a dilation factor to control additional sparse attention over larger regions without a quadratic increase in complexity. While it may be assumed that utilizing dilate attention limits the applicability of the DaViT architecture, through our testing this was not shown to be the case. Models were tested with a variety of layer configurations, both with the Dilate attention dominating the architecture, and with channel group attention dominating the architecture. While increasing the dominance of Dilate attention improved model performance to a point, past a threshold, the performance regressed. The optimal layout (the proposed model) was found to have most layers operating normally – i.e. containing both a Dilate attention and a channel group attention component – with 1 in 4 layers containing only Dilate attention components. In the best-performing model, there were 12 layers, and 3 of those contain only dilate attention components; the layers were split between 4 stages with a distribution of $\{1,2,8,1\}$.

B. Image Processing

As transformers are notoriously data intensive, the model used was trained on two available datasets before being fine-tuned and the fine-tuning data was collected from multiple sources. Models were first trained on ADE20k [12], then subsequently trained on Cityscapes [13]. Models trained in this way performed better on post fine-tuning tasks. Additional training was attempted with COCO Stuff [14], however, this resulted in either negligible improvements or a decline over using only ADE20k and Cityscapes for pre-training.

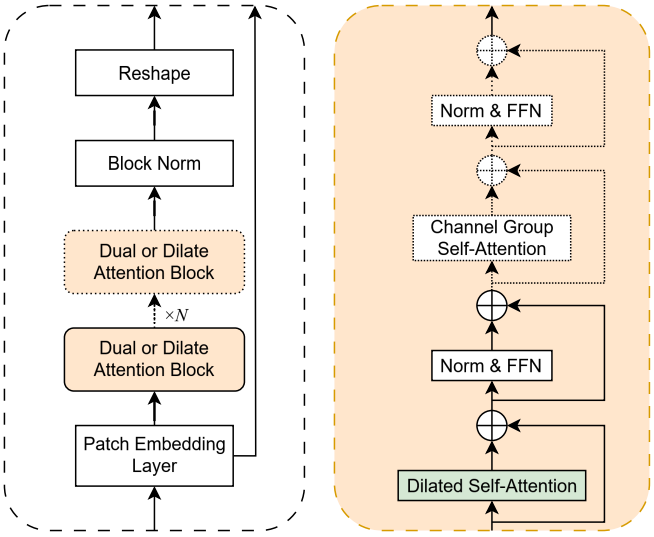

Figure. 1. Proposed changes to model architecture at the stage (left) and block (right) levels. Optional portions have a dotted outline. **Left:** Proposed modifications within each stage apply to the dual attention block. In the proposed model, block composition within a stage is a heterogeneous composition of dual attention blocks containing both dilated and channel group self attention, and dilate attention blocks containing only dilated self-attention. **Right:** Proposed modifications to block structure includes replacing channel self-attention with dilated self attention and allowing channel group attention to be optionally skipped.

Task images were collected from online sources: general image databases with the license filtering for free to use, and the dataset of Stagnant Water and Wet Surfaces [15]. The images were first automatically filtered for duplicates, low resolutions, and extreme aspect ratios using findimagedupes [16], GraphicsMagick [17], and custom tools. The images were then resized and center cropped for a resolution of 1024x1024, manually filtered for appropriate examples, and labeled by hand; images that were not visually distinguishable but passed the duplicate detection were eliminated as well. After initial labeling, the dataset consisted of 1000 images allocated as 80% for training, 10% for testing, and 10% for validation.

The base augmentation done to images consists of hue shift ±11°, and up to 3% added noise. Additionally, a modified brightness augmentation was done that ensures the images were not under or over exposed. This is done by first evaluating the "average brightness" of the image using the formulas [18]:

$$\sqrt{0.299 * r^2 + 0.587 * g^2 + 0.114 * b^2} = I_{pixel} \quad (1)$$

$$RMS(I_{pixels}) = I_{avg} \quad (2)$$

Where r, g, and b are the values of a single pixel's red, green, and blue channel respectively. I_{pixel} is the intensity of a single pixel, I_{pixels} is an vector holding the intensity values for every pixel, and I_{avg} is the calculated average intensity across the entire image.

The brightness augmentation was then clamped such that the image would not become darker or lighter than set thresholds. The value maps to the range 0-255, values of 50 and 205 were chosen as the lower and upper bounds, respectively. This eliminated cases where already overexposed or underexposed images in the dataset would become unrecognizable, tainting the training data. The final dataset consisted of 3000 images after augmentation maintaining the training, testing, and validation split.

IV. Results

When analyzed on individual images, the proposed model performs within several percentage points (%pts) of DaViT and SegFormer on most soil types, but outperforms the others in cases where the most similar classes are present together. The metrics in Table I are cumulative metrics across the test datasets, it is clear from the table that DaViT-Dilate performs better than both the SegFormer and original DaViT models with between 5 and 10 percentage points over SegFormer. In specific cases, such as the wet-soil and water case as shown in Figs. 2-4, the DaViT-Dilate shows up to a 4%pts improvement with SegFormer achieving 90.6% Accuracy (ACC), 80.1% mIoU, and 88.8% DICE; DaViT achieving 92.8% ACC, 83.9% mIoU, and 91.1% DICE; and DaViT-Dilate achieving 95.7% ACC, 90.3% mIoU, and 94.9% DICE.

The proposed model in basically all cases outperforms the original DaViT model on this task, but in some situations does fall behind SegFormer. These cases could likely be minimized through more architecture tuning to ensure the layout of the model is optimal. SegFormer was shown to be more versatile that DaViT due to its strong global attention, but suffers when classifying fine aggregates, water, and wet soil due to not retaining enough of the local features. DaViT performs very well on those classes where SegFormer falters on due to its strong local performance, but underperforms otherwise. It is speculated that while the original DaViT model was able to diminish some of the inherent locality from the channel attention by attending over each channel that the output from the channel attention did not contain sufficient feature information to fully provide the global context. DaViT-Dilate is able to retain enough global context through the dilated attention that the channel group attention can better express those features.

TABLE. I
Overall Performance Of Tested Models On Soil Dataset (%)

Model	ACC	mIOU	DICE
SegFormer	76.6	72.8	58.1
DaViT	70.6	65.4	49.8
DaViT-Dilate (ours)	81.9	81.1	68.3

V. Conclusion

With the increasing interest in AHM methods, ensuring safe traversal and operation on a variety of surfaces is necessary to allow workers to confidently operate alongside them. One tasks that is important, especially to navigation, is the classification and segmentation of soil types in the working envelope.

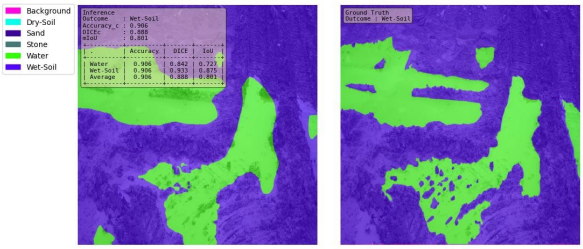

Figure. 2. SegFormer **Left:** Inference **Right:** Ground truth

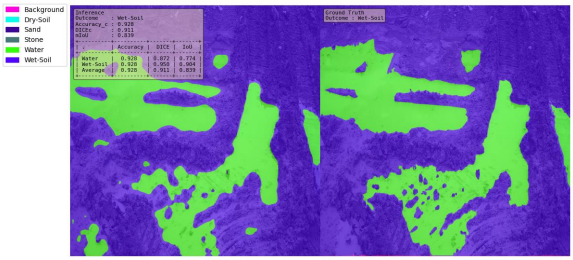

Figure. 3. DaViT **Left:** Inference **Right:** Ground truth

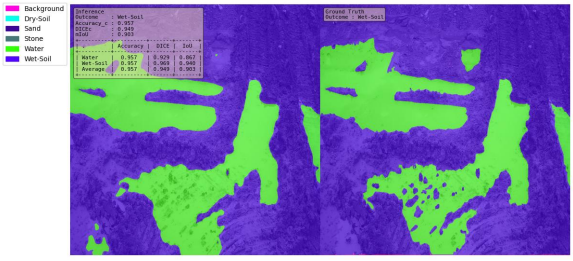

Figure. 4. Ours **Left:** Inference **Right:** Ground truth

We introduced a new model structure that is capable of strong performance across a variety of scenes, improving significantly over both SegFormer and the base DaViT models at this task. Future work should be done to increase the size and regional variation of the training dataset to make models more robust against new environments. Improvements to DaViT-Dilate may be possible through parameter tuning and layer structure modifications, especially to work towards real-time applications.

ACKNOWLEDGMENT

The Korea Evaluation Institute of Industrial Technology (KEIT) Industrial Technology Innovation Program (20023824) and National Research Foundation of Korea (Brain Pool Fellowship Program 2023) funded this research.

REFERENCES

[1] Bowen Du, Zirong Zhao, Xiao Hu, Guanghui Wu, Liangzhe Han, Leilei Sun, and Qiang Gao. Landslide susceptibility prediction based on image semantic segmentation. *Computers & Geosciences*, 155:104860, October 2021.

[2] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking Atrous Convolution for Semantic Image Segmentation, December 2017. arXiv:1706.05587 [cs].

[3] Thomas N. Kipf and Max Welling. Semi-Supervised Classification with Graph Convolutional Networks, February 2017. arXiv:1609.02907 [cs].

[4] Haiqiang Liu, Meibao Yao, Xueming Xiao, and Hutao Cui. A hybrid attention semantic segmentation network for unstructured terrain on Mars. *Acta Astronautica*, 204:492–499, March 2023.

[5] Xuanyao Chen, Zhijian Liu, Haotian Tang, Li Yi, Hang Zhao, and Song Han. SparseViT: Revisiting Activation Sparsity for Efficient High-Resolution Vision Transformer, March 2023. arXiv:2303.17605 [cs].

[6] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows, August 2021. arXiv:2103.14030 [cs].

[7] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. PVT v2: Improved baselines with Pyramid Vision Transformer. *Comp. Visual Media*, 8(3):415–424, September 2022.

[8] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. CvT: Introducing Convolutions to Vision Transformers, March 2021. arXiv:2103.15808 [cs].

[9] Tianlong Chen, Yu Cheng, Zhe Gan, Lu Yuan, Lei Zhang, and Zhangyang Wang. Chasing Sparsity in Vision Transformers: An End-to-End Exploration, October 2021. arXiv:2106.04533 [cs].

[10] Mingyu Ding, Bin Xiao, Noel Codella, Ping Luo, Jingdong Wang, and Lu Yuan. DaViT: Dual Attention Vision Transformers. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, volume 13684, pages 74–92. Springer Nature Switzerland, Cham, 2022. Series Title: Lecture Notes in Computer Science.

[11] Ali Hassani and Humphrey Shi. Dilated Neighborhood Attention Transformer, January 2023. arXiv:2209.15001 [cs].

[12] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene Parsing through ADE20K Dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5122–5130, Honolulu, HI, July 2017. IEEE.

[13] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes Dataset for Semantic Urban Scene Understanding, April 2016. arXiv:1604.01685 [cs].

[14] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. COCO-Stuff: Thing and Stuff Classes in Context, March 2018. arXiv:1612.03716 [cs].

[15] Sonali Bhutad and Kailas Patil. Dataset of Stagnant Water and Wet Surface Label Images for Detection. *Data in Brief*, 40:107752, February 2022.

[16] Jonathan H. N. Chin. jhnc/findimagedupes, February 2025. original-date: 2019-01-26T19:04:40Z.

[17] GraphicsMagick Image Processing System.

[18] HSP Color Model - Alternative to HSV (HSB) and HSL.