Infection Team at UCC AI Quest 2023: High Vegetation Semantic Segmentation in Aerial Images

Minh-Khoi Pham
School of Computing
Dublin City University
Dublin, Ireland

Thu-Uyen Nguyen
School of Computing & Statistics
Trinity College Dublin
Dublin, Ireland

Hoai-Nam Trinh
School of Computing
Dublin City University
Dublin, Ireland

Tu-Khiem Le
School of Computing
Dublin City University
Dublin, Ireland

Van-Tu Ninh
School of Computing
Dublin City University
Dublin, Ireland

***Abstract*—The integration of artificial intelligence (AI) into environmental monitoring for sustainable development has gained momentum. In response, the University College Cork (UCC) initiated the UCC AI Quest, leveraging advanced drone technology to collect a realistic dataset of high-resolution drone images, serving as a benchmark for high vegetation patches segmentation model evaluation. Team "Infection" from Dublin City University and Trinity College Dublin actively participated, achieving a first-place ranking on the private scoreboard with a High Vegetation IoU score of 86.67%. Our solution code is available at: https://github.com/kaylode/ucc-ai-quest-2023.**

***Index Terms*—semantic segmentation, high vegetation patch segmentation, deep learning models**

I. Introduction

In recent years, the integration of artificial intelligence (AI) into environmental monitoring has emerged as a promising avenue for sustainable development. Recognizing the critical role of AI in advancing environmental sustainability, the University College Cork (UCC) initiated the UCC AI Quest, which aims to harness the potential of high-resolution aerial imagery, captured by advanced drone technology, for the recognition of vegetation patches in diverse Irish natural landscapes. A new and realistic dataset of drone images captured from various above-ground levels is released and serves as a benchmark for evaluating the performance of semantic segmentation models in accurately identifying and classifying vegetation areas. Our team, named "Infection," is composed of four Ph.D. students from the School of Computing at Dublin City University and one undergraduate student from the School of Computer Science & Statistics at Trinity College Dublin. We enthusiastically participate in the challenge to construct robust deep-learning models for the detection of vegetation patches in Ireland. Through this endeavor, we aim to contribute our skills and knowledge to the realization of Sustainable Development Goals (SDGs).

II. Proposed Approach

A. Models

Three models were chosen in our approach that includes DeepLabv3+ [1], UNet++ [2], and DinoV2 [3]. The training of the DeepLabv3+ model involves utilizing the backbone architecture of EfficientNet-b4 and EfficientNet-b5, both pre-trained on the ImageNet dataset [4], as outlined by Tan et al. [5]. Conversely, during the training stage of the UNet++ model, only EfficientNet-b4 is employed as the backbone. In the training of the DinoV2 model [3], we specifically employed its ViT-B14 (Vision Transformer) backbone architecture as the encoders, rather than utilizing the entire model due to the small-scaled of the dataset. For the model to work on the semantic segmentation task, we attach a segmentation head to the backbone, by incorporating a simple two-layer Fully Convolutional Network (FCN) for the generation of segmentation masks.

B. Loss functions

The loss function of our model was the sum of Dice loss [6] and Online Hard Example mining Cross-entropy (OHEMCE) [7] with equal weights. The Dice loss encourages the model to improve the overlap between predicted and ground truth regions while the OHEMCE loss acts as traditional Cross-entropy loss but considers only top-k highest loss pixels in the predicted masks to help the deep network not to be overconfident in the void pixels.

C. Data augmentation

Owing to the small scale of the dataset, we employed different data augmentation techniques to create more variation of training samples to prevent over-fitting issues. Since aerial images are captured from above, most spatial augmentation methods (vertical/horizontal flip/random crop) are applicable without affecting the meaningfulness of the image. Additionally, we randomly augment the input color to simulate different levels of brightness, contrast, hue, and saturation. We also inherit Mosaic augmentation from Yolov4 [8] to blend multiple training images into one. This introduces a variety of situations where different plants are present in a same scene, which also helps the model generalize better.

D. Training methods

When fine-tuning large models on this such small-scaled dataset, we employ a two-stage training method to ensure the stability and fast convergence of the algorithms. Firstly, the models' backbone layers are frozen while only the segmentation head (or decoder layers) are trained with a normal learning rate. After an appropriate number of iterations, we unfreeze all the layers of the networks, then we fine-tune the whole model on the same dataset with a 10 times smaller learning

TABLE I
PRIVATE LEADERBOARD.

Model name	High Vegetation IoU	High Vegetation Acc
Ours (DinoV2 & UNet++ & DeepLabv3+)	**86.67**	93.97
Ours (DinoV2)	86.5	**94.00**
Ours (UNet++ & DeepLabv3+)	84.6	92.43
vantuan5644	85.61	92.85
philip1	83.35	89.60

rate. In the training phase, the size of input images is set to 672×672, and the batch size is set to 16. The reported models were all trained by using a machine with 1 RTXA6000 (48GB VRAM). The initial learning rate to train the head is set to 0.0001 and optimized using Adam optimizer.

E. Models ensemble

We exhaustively trained several models with slightly different setting modifications and evaluated them on the public validation set. Afterwards, we ranked them based on the IoU score of the high vegetation class, using the provided evaluation script. Top-5 models with the highest validation scores were then chosen to generate the probability masks on the private set. The chosen models are:

- DeeplabV3+ with EfficientNet-B4 as backbone
- DeeplabV3+ with EfficientNet-B5 as backbone
- Unet++ with EfficientNet-B4 as backbone
- DinoV2 with ViT-B14 as the backbone, 1 layer of FCN as segmentation head
- DinoV2 with ViT-B14 as the backbone, 2 layers of FCN as segmentation head

Finally, all the masks are averaged to generate the final submission.

III. RESULTS

Table I demonstrates the private leaderboard of the competition. It can be seen that our DinoV2 model beats the second-best team with 1% IoU and accuracy score higher. This confirms the effectiveness of our designed architecture and training policies.

IV. DISCUSSION & INSIGHTS

In this section, we discuss some challenges that we were facing during the contest and propose what we have yet tried that can potentially improve the results. Firstly, the provided data is not at its best quality of annotation, especially since details of the boundaries of the objects are not guaranteed to be perfect. This leads to the instability and degradation of the model's performance. Several ill-performing models that we have trained might be the victims of this assumption: YOLOv8, OneFormer, SegFormer, and MaskFormer. Thus, we tried to "smoothen" the boundary of the prediction masks, however, the results did not improve. While our best performing model is DinoV2 with two-layer FCN as an attached head, we still think there is much room to improve it. As suggested by the original paper [3], one can attach a much more complex segmentation head into DinoV2 to boost its performance, for example, the Mask2Former head. Furthermore, we have not yet tried tuning the hyperparameters thoroughly.

V. CONCLUSION

In summary, in this technical report, we reported 5 main contributions that we made to achieve first place in the competition:

1) Training a DinoV2-ViTB14 with a customized two-layer FCN for semantic segmentation task.
2) Combining both region-based and class-based loss functions as objective functions, namely Dice loss and OHEMCE loss
3) Applying Mosaic augmentation to generate a variety of complex data scenarios to enhance models' training
4) Introducing simple yet effective technique to finetune small-scaled dataset on such large state-of-the-art model
5) Employing ensemble methods that further boost the precision of predicted masks

REFERENCES

[1] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[2] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*. Springer, 2018, pp. 3–11.

[3] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[5] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[6] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*. Springer, 2017, pp. 240–248.

[7] Z. Wu, C. Shen, and A. v. d. Hengel, "High-performance semantic segmentation using very deep fully convolutional networks," *arXiv preprint arXiv:1604.04339*, 2016.

[8] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.