Leveraging DINOv2 Features for the GOOSE 2D Semantic Segmentation Challenge

Nils Schacknat[1]

***Abstract*— This technical report outlines my approach and results for the GOOSE 2D Semantic Segmentation Challenge. It showcases how to train a DPT decoder on top of frozen DINOv2 features, which scores an overall mIoU value of 0.82, ranking third in the competition.**
Code is available at github.com/nils-schacknat/dpt-decoder.

I. INTRODUCTION

The GOOSE (German Outdoor and Offroad Semantic) dataset [1], along with its GOOSE-Ex extension [2], provides a diverse collection of images captured from various platforms (including cars, excavators, and quadruped robots) across a wide range of environments such as roads, forests, landfills, and quarries. The dataset includes annotations for 64 fine-grained semantic classes, which are grouped into 8 superclasses for the purpose of this challenge. Evaluation is conducted on the withheld annotations from the test set.

In this work, I utilize DINOv2 [3], a self-supervised Vision Transformer (ViT) [4] as a frozen feature extractor. DINOv2 provides general-purpose visual features that achieve state-of-the-art performance on a wide range of tasks without the need for finetuning. On top, I train a Dense Prediction Transformer (DPT) [5] decoder to fuse multi-scale features for accurate pixel-wise predictions.

By combining the state-of-the-art DINOv2 backbone with a well-established decoding architecture, the proposed method effectively adapts to the challenges posed by the GOOSE and GOOSE-Ex datasets. The results underscore the strong performance of DINOv2 features in complex, unstructured real-world environments.

II. RELATED WORK

DINOv2 (knowledge **di**stillation with **no** labels) is a self-supervised learning framework that utilizes Vision Transformers to learn rich and general-purpose visual representations from unlabeled data. Notably, training lightweight decoders on top of frozen DINOv2 features has been shown to achieve state-of-the-art or near state-of-the-art performance in a wide range of tasks. This approach removes the need for resource-intensive training procedures and task-specific large-scale datasets, making it both effective and efficient for real-world applications.

The Dense Prediction Transformer (DPT) architecture produces high-resolution predictions by combining the global context modeling capabilities of ViTs with a convolutional decoder. DPT assembles features from multiple stages of the transformer into multi-scale representations, which are then progressively combined to produce high-resolution predictions. This design enables the recovery of fine-grained details and has demonstrated substantial improvements in tasks like semantic segmentation and depth estimation.

The DINOv2 paper itself successfully showcases how to apply a DPT decoder for depth estimation.

III. METHOD

A. Preprocessing and Feature Extraction

All input images were resized such that the shorter side was set to 518 pixels while maintaining the aspect ratio. The longer side was adjusted to the next higher multiple of 14, which is the patch size of the DINOv2-ViT. The resized images were then embedded using the DINOv2 giant model with register tokens [6]. The feature maps have shapes $(h/14, w/14, 1536)$ and were extracted from layers 9, 19, 29, and 39 of the model. To save on computational resources, the embeddings were stored on disk and subsequently loaded during training and evaluation.

B. Decoder Training

A DPT decoder was trained on top of the extracted embeddings. To align the spatial dimensions across varying image sizes, both the embeddings and the corresponding segmentation masks were randomly cropped to fixed-size squares with lengths 518 and 37 $(37 \cdot 14 = 518)$ respectively. Note that recomputing the embeddings from the cropped image regions, as opposed to cropping the precomputed embeddings, would likely yield greater feature diversity and improved generalization. However, this approach would require running inference with a ViT-G model at every training step, resulting in a substantial computational overhead.

The decoder was trained using a cross-entropy loss with a learning rate of 10^{-4} and optimized using the AdamW optimizer [7] with a batch size of 16.

Since it provides significantly more images than the other platforms, the decoder was first trained on MuCAR-3 images for 64 epochs. Afterwards, the pre-trained decoder was fine-tuned separately on the images of the Spot and ALICE platforms for 32 epochs each.

IV. RESULTS

My method scores an overall mIoU value of 0.82 on the test set, which is the average across all platforms weighted by the number of images, see Table I and Fig. 1.

[1]Master's student at Heidelberg University (Data and Computer Science)
schacknat.nils@gmail.com

TABLE I

mIoU on the GOOSE test set across all platforms.

Overall	MuCAR-3	ALICE	Spot V1	Spot V2
0.819	0.825	0.811	0.803	0.794

Fig. 1. Examples of the predicted segmentation masks from the unseen test set.

REFERENCES

[1] P. Mortimer, R. Hagmanns, M. Granero, T. Luettel, J. Petereit, and H.-J. Wuensche, "The goose dataset for perception in unstructured environments," 2024. [Online]. Available: https://arxiv.org/abs/2310.16788

[2] R. Hagmanns, P. Mortimer, M. Granero, T. Luettel, and J. Petereit, "Excavating in the wild: The goose-ex dataset for semantic segmentation," 2024. [Online]. Available: https://arxiv.org/abs/2409.18788

[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] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[5] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 12 179–12 188.

[6] T. Darcet, M. Oquab, J. Mairal, and P. Bojanowski, "Vision transformers need registers," *arXiv preprint arXiv:2309.16588*, 2023.

[7] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.