Upsampling DINOv2 features for unsupervised vision tasks and weakly supervised materials segmentation

Ronan Docherty[1,2], Antonis Vamvakeros[2,3], Samuel J. Cooper[2]

1 Department of Materials, Imperial College
London, London SW7 2A7, UK
2 Dyson School of Design Engineering, Imperial College
London, London SW7 2DB, UK
3 Finden Limited, The Oxford Science Park, Magdalen Centre, Robert Robinson
Avenue, Oxford, OX4 4GA, UK

ronan.docherty18@imperial.ac.uk

Segmentation – the assigning of user-specified class to each pixel in an image – is a prerequisite for most downstream analysis in microscopy, like phase quantification, physical simulations, particle size analysis *etc*. These analyses are key in materials science for understanding the structure-processing-property relations and therefore for improving material performance.

An existing micrograph segmentation technique is ‘interactive pixel classification’, where features like average greyscale intensity, edge intensity and texture are extracted on multiple length scales around each pixel. A user then draws class labels onto pixels with a brush tool, and a classifier (commonly a random forest) is trained to map from the feature vector for that pixel to its class. Example tools that use this workflow include Trainable Weka Segmentation [1] and ilastik [2]; see Figure 1 for more explanation.

This technique offers a few advantages: it is fast, works with only a few, sparse labels (also called ‘weakly supervised segmentation’) and generalizes well to new materials, as a new classifier can be trained each time. However, it suffers when applied to more complex materials with tertiary or quaternary phases that may have a similar appearance.

In this work [3] we improve this weakly supervised segmentation by supplementing these classical features with deep features extracted from a recent feature foundation model, DINOv2 [4]. DINOv2, a vision transformer (ViT) model, was trained using self-supervised learning on a large dataset of natural images, and has learned features that are consistent between “changes of pose, style or even objects” [4]. Despite their semantic richness, these features are learnt at the “patch” level (a 14x14px square) in order to be computationally tractable. We introduce a technique for upsampling these features to better achieve the desired resolution, and examine existing upsampling techniques [5] in this context. We further study the impact of combining the deep features with the classical features and find they produce high-qulaity segmentations.

We demonstrate the effectiveness of our new technique as compared to the classical approach over two case studies. The first is a dataset of Transmission Electron Microscopy (TEM) images of human T-cells, where the goal is to segment the nucleus, cell and background. We train a classifier over the features of 6 sparsely labelled cells, and apply unseen to the other 130 cells, finding the deep features to produce segmentations that align far better with the ground truth labels.

The second case study (shown partly in Figure 2) is the application of the technique to a series of micrographs which cover different material systems and instruments: ranging from NMC battery cathodes to alloys to polymorphs of organic crystals. We achieve good segmentations of complex phases and effects, like the “pore-back” [6] in the cathode, graphite inclusions in the alloy or separating the polymorphs in the crystal.

We also explore using these features without any supervision or labels, and perform automated segmentation using clustering and attention maps from the deep features – potentially relevant in autonomous setups or for dataset creation. We continue to explore the limitations, including high GPU memory and time cost, a blurring effect introduce by the upsampling and the implict positional bias in the deep features, as well as recommending mitigations.

References

[1] I. Arganda-Carreras, V. Kaynig, C. Rueden, K. W. Eliceiri, J. Schindelin, A. Cardona, and H. Sebastian Seung, “Trainable Weka Segmentation: a machine learning tool for microscopy pixel classification,” *Bioinformatics,* (2017)

[2] S. Berg, D. Kutra, T. Kroeger, C. N. Straehle, B. X. Kausler, C. Haubold, M. Schiegg, J. Ales, T. Beier, M. Rudy, K. Eren, J. I. Cervantes, B. Xu, F. Beuttenmueller, A. Wolny, C. Zhang, U. Koethe, F. A. Hamprecht, and A. Kreshuk, “ilastik: interactive machine learning for (bio)image analysis,” *Nature Methods,* (2019)

[3] R. Docherty, A. Vamvakeros, S. J. Cooper, 'Upsampling DINOv2 features for unsupervised vision tasks and weakly supervised materials segmentation', *NeurIPS AI4Mat Workshop,* (2024)

[4] M. Oquab *et al.* ‘DINOv2: Learning Robust Visual Features without Supervision’, *arXiv preprint* (2023)

[5] S. Fu, M. Hamilton, L. Brandt, A. Feldman, Z. Zhang, and W. T. Freeman, “FeatUp: A Model-Agnostic Framework for Features at Any Resolution,” *arXiv preprint*, (2024)

[6] S. J. Cooper, S. A. Roberts, Z. Liu, and B. Winiarski, “Methods—Kintsugi Imaging of Battery Electrodes: Distinguishing Pores from the Carbon Binder Domain using Pt Deposition,” *Journal of The Electrochemical Society*, (2022)

Figures

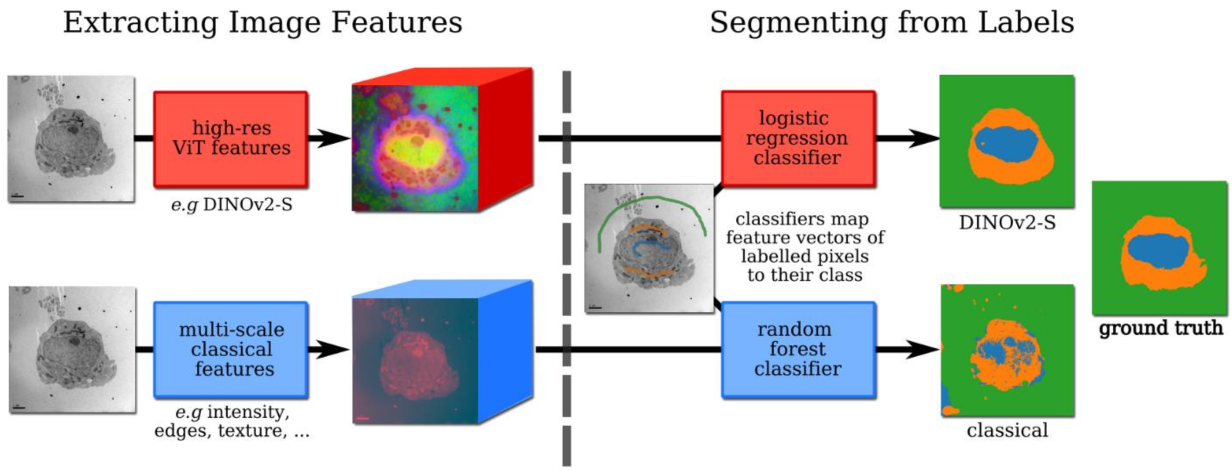
Extracting Image Features
Segmenting from Labels
high-res ViT features
e.g DINOv2-S
multi-scale classical features
e.g intensity, edges, texture, ...
logistic regression classifier
classifiers map feature vectors of labelled pixels to their class
random forest classifier
DINOv2-S
classical
ground truth

Figure 1: weakly-supervised micrograph segmentation using deep or classical features. A classfier (random forest or linear regressor) is trained to map from a labelled pixel's feature vector to its class (drawn with a label). The deep ViT features are more semantically rich, admitting a better segmentation of the nucleus from the rest of the cell. Taken from [3].

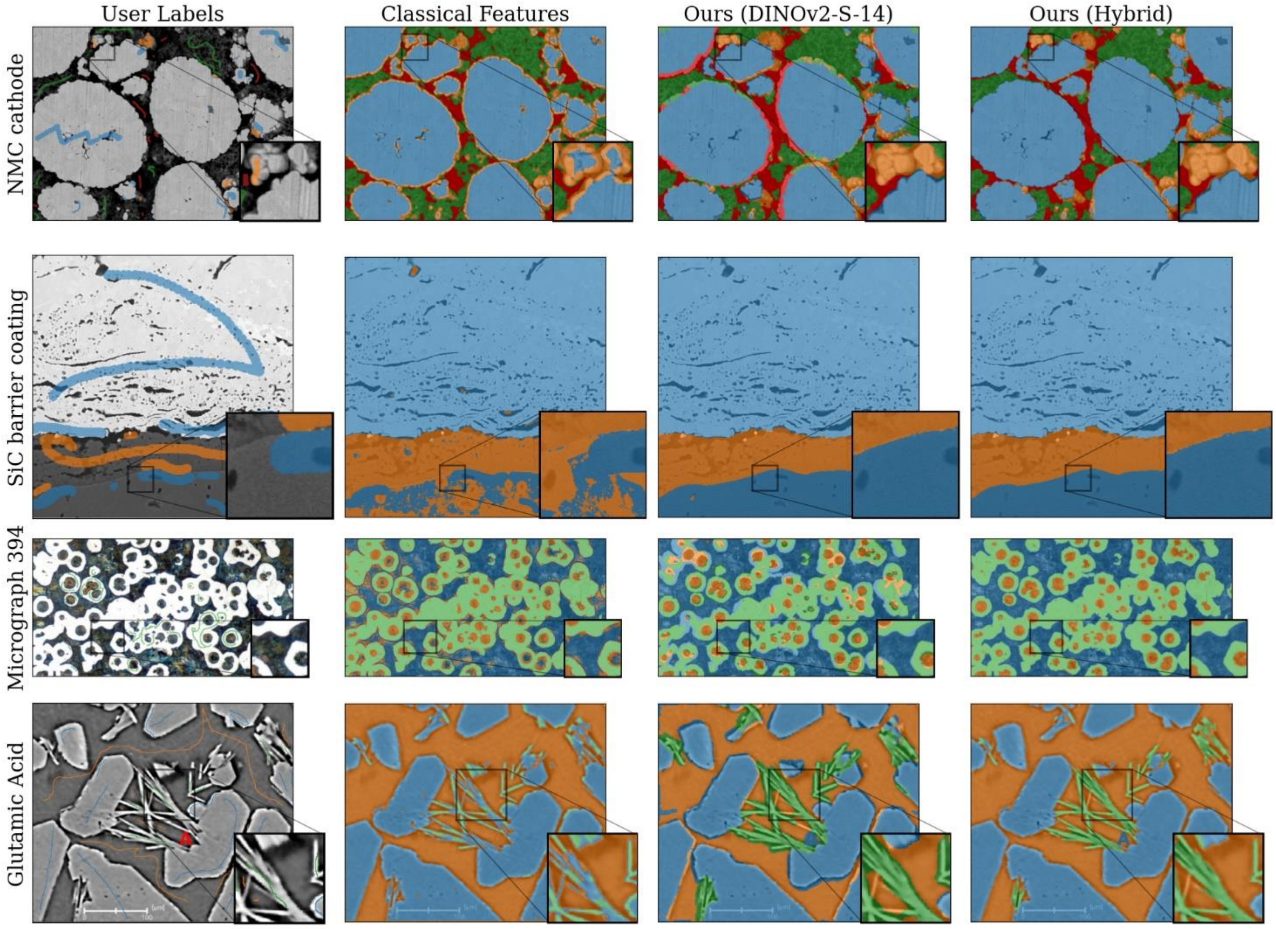

Figure 2: example segmentations on a series of micrographs captured with various instruments using classical features, deep ViT features and the combination of both ('hybrid'). We see that the hybrid scheme produces high quality segmentations of complex phases with only a few user labels. Adapted from [3].