Application of Self Supervised Vision Transformers for Multiplexed Microscopy Images and Its Challenges

Gantugs Atarsaikhan[1,3] **Isabel Mogollon**[1] **Katja Välimäki**[1]
Teijo Pellinen[1] **Tuomas Mirtti**[2,3] **Lassi Paavolainen**[1,3]

[1]Finnish Institute for Molecular Medicine (FIMM), Unversity of Helsinki
[2]Department of Pathology, University of Helsinki
[3]iCAN Digital Precision Cancer Medicine Flagship, University of Helsinki
gantugs.atarsaikhan@helsinki.fi, lassi.paavolainen@helsinki.fi

Abstract

Multiplexed microscopy imaging enables the simultaneous use of numerous fluorescent markers on one biological sample. This technique is especially useful in cancer research, cellular and molecular biology, and drug discovery. Studying these microscopic images is challenging due to the large scale of datasets, number of channels that exceeds the natural imaging domain, and the lack of annotations. In this work, we applied a self-supervised learning method for representation learning, and then studied the quality of the learned representations visually and by classification tasks. Results show that although the model creates similar feature embeddings for the same metadata labels, the model also captures some technical variation between slides.

1 Introduction

Fluorescent microscopy imaging [15] is a widely used technique in biological and biomedical fields. Researchers use fluorescent markers to attach onto specific molecules or biological structures, that become visible under the microscope. Nowadays, high throughput imaging (HTI) technologies use robotic automation and are able to create large number of high-resolution images rapidly. These images are analysed for observing previously unknown biological phenomena, discovering novel drugs, and even for clinical diagnoses.

Cyclic multiplexed immunofluorescence microscopy imaging technique [3, 16] enables the simultaneous use of numerous (dozens at a time) fluorescent markers on one biological sample compared to only 3-5 markers of traditional fluorescent microscopy imaging. Moreover, it enables imaging of samples from multiple sources in one experiment, minimizing the risk of batch effects. Fig. 1 shows three examples from multiplexed image datasets that consists of around 2,500 cancer sample images from around 870 patients. Each multiplex image has six or more fluorescent markers representing the image channels. In these examples, there are three structural markers: DAPI channel for cell nuclei, epithelial channel for epithelial cells, and stroma channel for stromal cells. SMA, pSTAT3, FAP, and PDFGRB channels present different proteins of interest. Typically, the fluorescent markers to use for the tissue or cell samples are selected based on the research question from thousands of known markers.

Working with fluorescent images, especially multiplex images is challenging for a few reasons. Firstly, raw images are typically collected using a 16-bit range compared to 8-bit natural images. The intensity of fluorescent images can vary drastically from low background noise to high intensity from clustered markers or autofluorescent (artefacts). The 8-bit range with 256 different values does not capture this intensity variance and results in loss of data, hence, a 16-bit range is preferred. However,

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

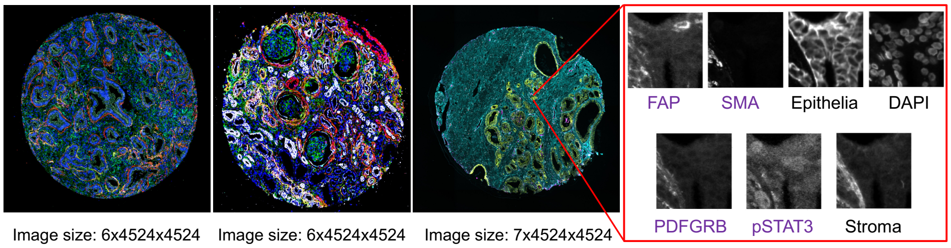


Figure 1: Example of fluorescent microscopic images. Images are converted from multiple channels into RGB images for visualization purposes. There are three structural markers: **DAPI** channel for cell nuclei, **Epithelial** channel for epithelial cells, and **Stroma** channel for stromal cells. **SMA**, **pSTAT3**, **FAP**, and **PDFGRB** represent the imaged proteins.

processing 16-bit multi-channel images requires more resources compared to natural RGB images. Secondly, only limited annotations or incomplete ground truths are available due to the lack of resources or privacy reasons. The sheer number of images makes it hard to annotate all images even at the image level. Therefore, an unsupervised or self-supervised method to learn the representation of the image dataset is our preferred solution.

In this paper, we present our approach to train a vision transformer (ViT) [10] using small patch images with self-supervised "Self-Distillation with No Labels (DINO)" [6] method. Experiments are done by training on multi-channel images and extracting the embeddings directly. Then, the feature embeddings of patch images are aggregated by another self-supervised method [7]. We provided qualitative results by reducing dimensions and visualizing with Uniform Manifold Approximation and Projection (UMAP) [18], and quantitative results by predicting the clinical information with a simple k-NN classifier.

2 Related Work

Image-based cell profiling field (bioimage profiling) quantifies biological phenotype differences among a variety of cell populations under different perturbations [4]. Hundreds of morphological features (shape, intensity, texture measurements) from thousands of cells can be measured and aggregated to create phenotype profiles at the cell population level or even patient level. Compared with other profiling methods (e.g., mRNA, protein, etc [11]), image-based profiling methods have a wider range of applications that include identifying the mechanism-of-action of drugs for drug discovery, specifying characters of a specific disease, and understanding functions of cellular organelles [13, 5]. Although many studies have been done on multiplex imaging data, such as measuring classical morphological features for analyzing cancer micro-environments from multiplexed images [19], the large number of channels makes measuring classical morphological features complex and expensive [12].

Due to common problem of lack-of-annotation, many unsupervised and weakly-supervised machine learning methods are proposed for biological applications. Chen et al. [7] used multi-stage self-supervised training to create features from low to high resolutions. Cross-Zamirski et al. [9] combined DINO [6] with weak labels with impressive results. Han et al. [12] created weak labels from nuclei channel for cell segmentation on multiplex imaging. Others have tried fully unsupervised methods with various levels of success [17, 20].

We chose DINO [6] method over other successful self-supervised learning methods, such as VICReg [1] and VICRegL [2] or SimCLRv2 [8] due to its simplicity and adaptability. Moreover, when trained with a ViT as the backbone, it is somewhat easy to interpret the outcome by extracting the attention maps [14] without using any class activation mapping techniques.

3 Methodology

Data preparation. Our multiplex image dataset is prepared in-house using human cancer tissue samples. Tissue Microarray (TMA) spots with a diameter of around 1 mm are cut from the cancer tissues. Around one hundred TMA spots can be prepared and imaged in one slide. The actual imaging

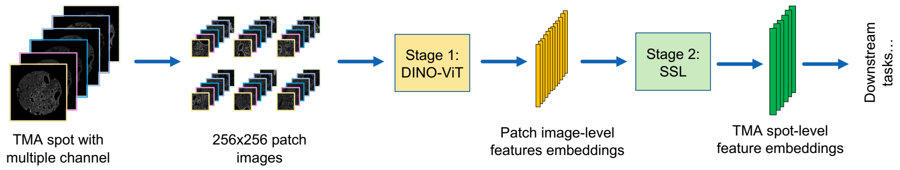

TMA spot with multiple channel
256x256 patch images
Stage 1: DINO-ViT
Patch image-level features embeddings
Stage 2: SSL
TMA spot-level feature embeddings
Downstream tasks...

Figure 2: Two-stage self-supervised neural network for training multiplex microscopic images. The input TMA spots are large images having 4k pixel dimensions that include six to seven channels. Stage 1 trains on the small patch images from the TMA spot images, and then Stage 2 aggregates patch image-level feature embeddings to a TMA spot-level embedding.

is done with multiple rounds of staining with fluorescent markers, microscopic imaging, and washing the TMA spots for the next staining round. Three types of cancer samples are imaged in this way: prostate cancer, renal (kidney) cancer, and lung cancer. These datasets are collected using two rounds of imaging resulting in six markers for renal and lung cancer, and seven markers for prostate cancer. In addition, one marker is used in all rounds of imaging to enable image registration.

The pixel dimension of resulting images is around 4500×4500px for each TMA spot. Contrary with [9], we cut off pixel values lower than the median value of the respective slide to remove the unnecessary background. Then, non-overlapped patches with the size of 256×256px are taken from the tissue area. The outcome is a multi-channel dataset with 179k renal, 124k prostate, and 98k lung cancer patch images taken from 2,500 TMA spots that belongs to 870 patients.

Stage 1. Training on the patch images. Fig. 2 shows a two-stage method that we are experimenting with. In Stage 1, a ViT is trained with self-supervised strategy using multi-channel patch images to capture the raw image patterns. We used DINO [6] training scheme with a modified ViT-B backbone for multi-channel images. Feature embeddings for each patch image are extracted after self-supervised training of the model.

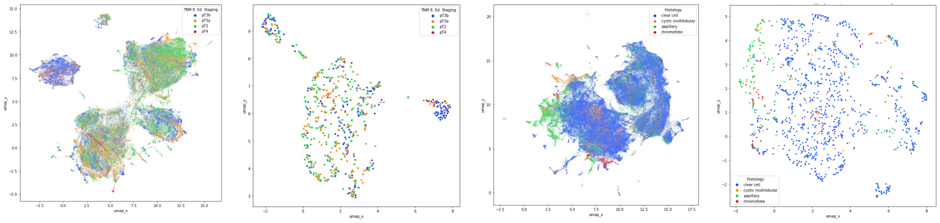
Stage 2: Aggregating features. Feature embeddings from Stage 1 correspond to individual patch images, and need to be aggregated into TMA spot-level features, then potentially to patient-level features. One TMA spot is a complex mixture of various types of cells. Merely mean or median aggregation from patch image features does not capture the complexity. To solve this challenge, a self-supervised aggregation approach is used [7]. The method takes 256 patches with 768 features for each TMA spot as input, and reorganizes these features as channel dimension by considering each ordered feature from all patch images as one image $(256 \times 768- > 768 \times 16 \times 16)$. Then, the reorganized feature images are trained using a modified DINO-ViT method to extract TMA spot-level features.

Downstream tasks. Downstream tasks with TMA spot-level features include classification for clinical information (cancer histology subtype, aggressiveness, survival status, etc.), clustering for discovery purposes, and outlier detection. Even though it is possible to use the patch-level feature embeddings for downstream tasks, using TMA spot-level or patient-level information is more meaningful for biomedical applications.

Training. We used the LUMI supercomputer with AMD MI250X GPUs for training. For all Stage 1 experiments, the main parameters used were ViT-B/16 backbone, 2 global crops of 224×224, 8 local crops of 96×96, total batch size of 1000. The student ViT uses the AdamW optimizer with a learning rate starting from $1e - 5$ after 10 warm-up epochs, and the teacher temperature as 0.04. Data augmentation is limited to random horizontal flips, and random but small changes in brightness, contrast, hue, saturation, and blurring. Images from each type of cancer are trained separately until convergence. Dimensions of extracted feature vector is 768 for each patch image. For Stage 2, we used modified ViT-S, 2 global crops of 14×14, and 2 local crops of 6×6. The output dimension is 384.

4 Results

Visualization of feature embeddings. We used UMAP [18] for dimensionality reduction of the feature embeddings to visualize these in two dimensions (Fig. 3). On prostate cancer samples, we



(a) Prostate cancer patch image (b) Prostate cancer TMA spot (c) Renal cancer patch image (d) Renal cancer TMA spot

Figure 3: Feature embedding visualization with UMAP dimensionality reduction technique. On prostate cancer figures, colors show different staging of the cancer. On renal cancer figures, color code shows histology or subtype of the cancer.

Table 1: Average top-1 accuracy with the k-NN classifier for cancer stage classification task for prostate cancer samples, and histology classification task for lung and renal cancer samples. "mean agg." refers to classifications on aggregated TMA spot features that are created by averaging the patch features.

Embedding name	**NSS** (%)	**NST** (%)	**No restriction** (%)	**NSS mean agg.** (%)	**No restriction mean agg.** (%)
Prostate patch image	30.45	50.28	85.52	-	-
Prostate TMA spot	33.08	-	57.79	27.08	48.00
Renal patch image	69.35	79.63	90.35	-	-
Renal TMA spot	71.96	-	83.91	72.89	84.13
Lung patch image	53.45	60.95	79.13	-	-
Lung TMA spot	57.07	-	65.02	55.62	58.04

studied the staging of the cancer. On renal and lung cancer, we focused on cancer histology sub-types. Patch-level feature embeddings present more clearer clusters compared to TMA spot-level feature embeddings. The reason is that patch images cover only a small area of the TMA spots, thus there is much more variance in the patches. Moreover, there are similar patterns in patch images in different TMA spots representing similar cellular neighborhoods.

Classification of feature embeddings. We used k-NN classifier (k=10) to study similarities of feature embeddings of samples from the same class. Table 1 displays the average top-1 classification accuracy achieved with the k-NN classifier trained on cancer stage for prostate, and histology for lung and renal datasets. In the "Not-same-slide" (NSS) scenario, only feature embeddings from different slides are used for training the k-NN classifier. Moreover, "Not-same-TMA spot" (NST) employs only patch image feature embeddings from different spots. However, the classifier performs most accurately when there are no restrictions. This could indicate the possibility of a batch effect between slides, and even between TMA spots. The NSS classification is the most restrictive and, consequently, less accurate than the other metrics.

5 Conclusion and Discussion

In this work, we used self-supervised training approach to learn the representation in multiplexed microscopy images. UMAP visualization of extracted feature embeddings show promising clustering by the cancer stage or histological subtype showing that the approach is able to learn biologically meaningful representations. However, results of the classification tasks indicate that there seems to be batch or slide effects between TMA slides and TMA spots.

The overall goal of this work is to study a pan-cancer dataset with unbiased self-supervised learning approaches to find meaningful patterns that are relevant for cancer diagnostics and treatment. The current results indicate that self-supervised learning is capable of recognizing patterns that represent for instance cancer grading and aggressiveness, however, more thorough model interpretation is needed to link these findings into biomarkers for diagnostics in the future.

6 Acknowledgement

We acknowledge CSC for awarding this project access to the LUMI supercomputer, owned by the EuroHPC Joint Undertaking, hosted by CSC (Finland). We acknowledge FIMM Digital microscopy and molecular pathology unit for sample preparation and imaging.

References

[1] A. Bardes, J. Ponce, and Y. LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In *ICLR*, 2022.

[2] A. Bardes, J. Ponce, and Y. LeCun. Vicregl: Self-supervised learning of local visual features. In *NeurIPS*, 2022.

[3] S. Blom, L. Paavolainen, D. Bychkov, R. Turkki, P. Mäki-Teeri, A. Hemmes, K. Välimäki, J. Lundin, O. Kallioniemi, and T. Pellinen. Systems pathology by multiplexed immunohistochemistry and whole-slide digital image analysis. *Scientific reports*, 7(1):15580, 2017.

[4] J. C. Caicedo, S. Cooper, F. Heigwer, S. Warchal, P. Qiu, C. Molnar, A. S. Vasilevich, J. D. Barry, H. S. Bansal, O. Kraus, et al. Data-analysis strategies for image-based cell profiling. *Nature methods*, 14(9):849–863, 2017.

[5] J. C. Caicedo, S. Singh, and A. E. Carpenter. Applications in image-based profiling of perturbations. *Current opinion in biotechnology*, 39:134–142, 2016.

[6] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021.

[7] R. J. Chen, C. Chen, Y. Li, T. Y. Chen, A. D. Trister, R. G. Krishnan, and F. Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16144–16155, 2022.

[8] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020.

[9] J. O. Cross-Zamirski, G. Williams, E. Mouchet, C.-B. Schönlieb, R. Turkki, and Y. Wang. Self-supervised learning of phenotypic representations from cell images with weak labels. *arXiv preprint arXiv:2209.07819*, 2022.

[10] 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.

[11] Y. Feng, T. J. Mitchison, A. Bender, D. W. Young, and J. A. Tallarico. Multi-parameter phenotypic profiling: using cellular effects to characterize small-molecule compounds. *Nature Reviews Drug Discovery*, 8(7):567–578, 2009.

[12] W. Han, A. M. Cheung, M. J. Yaffe, and A. L. Martel. Cell segmentation for immunofluorescence multiplexed images using two-stage domain adaptation and weakly labeled data for pre-training. *Scientific Reports*, 12(1):4399, 2022.

[13] L. Li, Q. Zhou, T. C. Voss, K. L. Quick, and D. V. LaBarbera. High-throughput imaging: Focusing in on drug discovery in 3d. *Methods*, 96:97–102, 2016.

[14] Y. Li, H. Mao, R. Girshick, and K. He. Exploring plain vision transformer backbones for object detection. In *European Conference on Computer Vision*, pages 280–296. Springer, 2022.

[15] J. W. Lichtman and J.-A. Conchello. Fluorescence microscopy. *Nature methods*, 2(12):910–919, 2005.

[16] J.-R. Lin, B. Izar, S. Wang, C. Yapp, S. Mei, P. M. Shah, S. Santagata, and P. K. Sorger. Highly multiplexed immunofluorescence imaging of human tissues and tumors using t-cycif and conventional optical microscopes. *eLife*, 7:e31657, jul 2018.

[17] A. X. Lu, O. Z. Kraus, S. Cooper, and A. M. Moses. Learning unsupervised feature representations for single cell microscopy images with paired cell inpainting. *PLoS computational biology*, 15(9):e1007348, 2019.

[18] L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[19] T. Pellinen, L. Paavolainen, A. Martín-Bernabé, R. Papatella Araujo, C. Strell, A. Mezheyeuski, M. Backman, L. La Fleur, O. Brück, J. Sjölund, et al. Fibroblast subsets in non-small cell lung cancer: Associations with survival, mutations, and immune features. *JNCI: Journal of the National Cancer Institute*, 115(1):71–82, 2023.

[20] A. Perakis, A. Gorji, S. Jain, K. Chaitanya, S. Rizza, and E. Konukoglu. Contrastive learning of single-cell phenotypic representations for treatment classification. In C. Lian, X. Cao, I. Rekik, X. Xu, and P. Yan, editors, *Machine Learning in Medical Imaging*, pages 565–575, Cham, 2021. Springer International Publishing.