ChAda-ViT : Channel Adaptive Attention for Joint Representation Learning of Heterogeneous Microscopy Images

Supplementary Material

A. IDRCell100k Construction

We acquired the IDRCell100K dataset using a distributed High-Performance Computing (HPC) cluster managed by HTCondor. Employing multi-processing and multi-threading techniques, the dataset was efficiently downloaded over two weeks. IDRCell100K serves as an initial, diverse channel configuration dataset, a pioneering resource in this domain to the best of our knowledge. Although currently small in scale, it provides an excellent basis for self-supervised learning applications, including training on high-quality microscopy images and evaluating channel-adaptive architectures with a multichannel dataset.

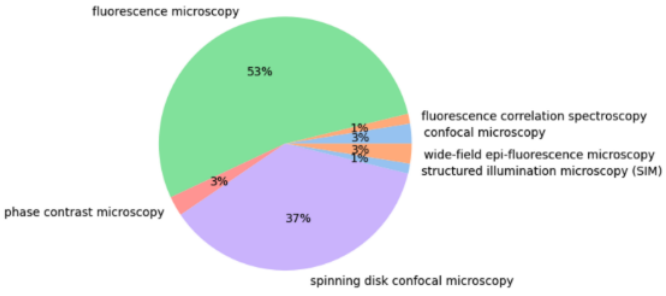

Figure 6. The imaging method distribution within IDRCell-100k, highlighting the dataset's variety in microscopy techniques. This diversity is crucial for developing models capable of interpreting cells imaged under various conditions, despite certain methods being less represented due to higher acquisition costs and technical complexities.

To ensure a broad data distribution, we randomly selected 1,300 images from each study with at least this amount of data available. These images were uniformly chosen from the 8,050,408 images in the "Cell" section of the Image Data Resources datalake. The curated dataset comprises 104,093 multiplexed images, corresponding to 308,898 channel, and features 1 to 10 channels per image. IDRCell100k encompasses microscopy images from 79 different assays and 7 distinct imaging methods (see Fig. 6). It will be made freely accessible under an open license, and would require 183GB of disk space for storage.

B. Evaluation Tasks and Datasets

CYCLoPs

Dataset Description. The CYCLoPs (Collection of Yeast Cells Localization Patterns) dataset* is a specialized collection of 27,058 single-cell yeast images, designed to support research in eukaryotic cell biology. This dataset amalgamates systematic genetics, high-throughput microscopy, and image analysis to reveal protein interactions within cells. A distinctive feature of CYCLoPs is its dual-channel imaging: one channel highlights the protein of interest, and the other visualizes the cytosol. This setup is advantageous for precise visualization and subsequent computational analysis. The dataset's standardization and detailed annotations facilitate its application in machine learning, particularly in deep learning-based classification tasks, without necessitating extensive prior domain knowledge.

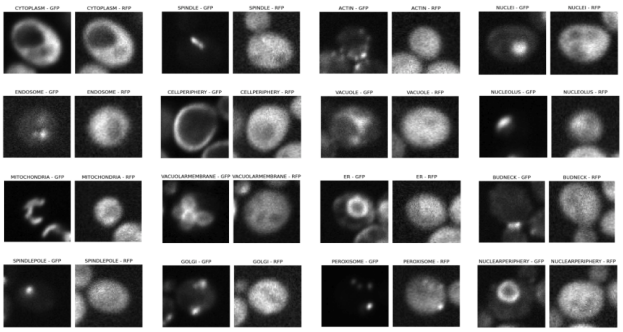

Figure 7. The CYCLoPs dataset: A depiction of 16 distinct classes representing diverse protein localizations within the yeast cell, each class corresponding to a unique subcellular region.

Task Description. The primary analytical task with the CYCLoPs dataset is the classification of proteins' subcellular localizations in yeast cells. This task is critical for understanding protein network dynamics and cellular functions. Accurate classification of protein localizations provides insights into their roles and interactions within the cell, essential for advancing knowledge in cellular biology and proteomics. The dataset's high-quality, dual-channel images enable precise localization, making it a valuable resource for developing and testing deep learning models for protein localization prediction in eukaryotic cells.

*The dataset can be accessed on Kaggle: CYCLoPs Dataset.

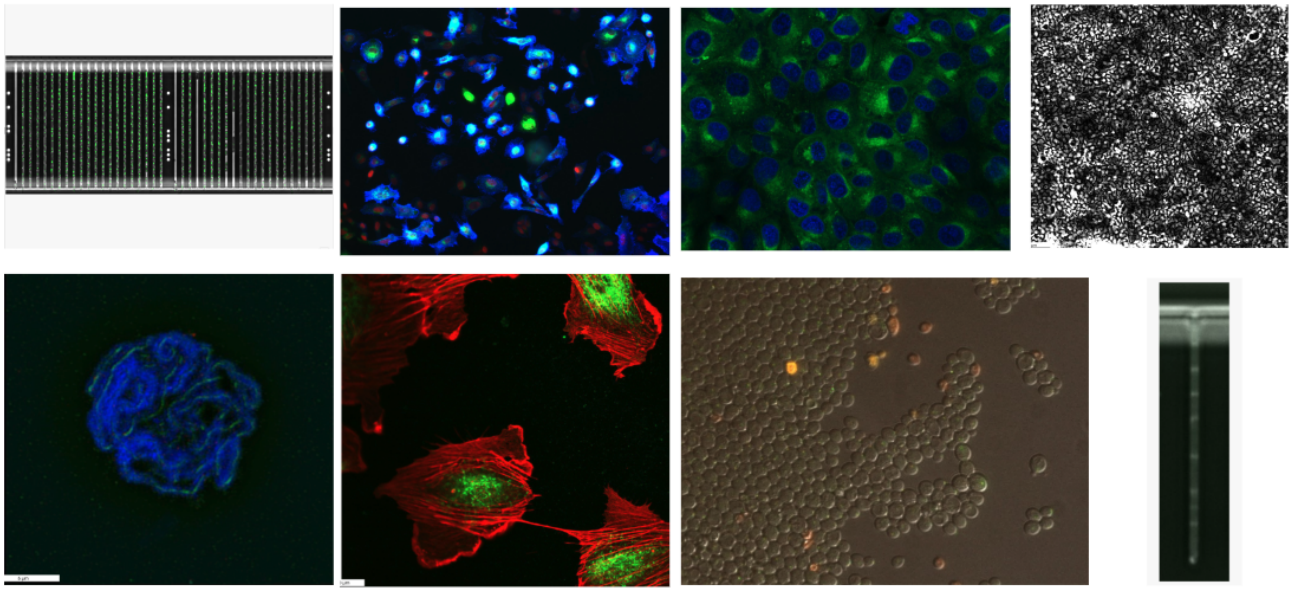

Figure 8. A selective overview of the diverse microscopy image types in IDRCell-100k, illustrating the dataset's variety. The aim is to achieve a broad and heterogeneous distribution of training data.

BBBC048

Dataset Description. The BBBC048 dataset[†], part of the Broad Bioimage Benchmark Collection, comprises 32,266 images of Jurkat cells, a type of human immune cell. These cells were grown asynchronously and imaged using the ImageStream platform. The dataset's uniqueness lies in its dual staining approach: Propidium Iodide for DNA content quantification and MPM2 antibody for identifying cells in the mitotic phase of the cell cycle. Created by the Flow Cytometry Core Facility at Newcastle University, this dataset is specifically designed to support cell cycle reconstruction and disease progression analysis, particularly through deep learning methods.

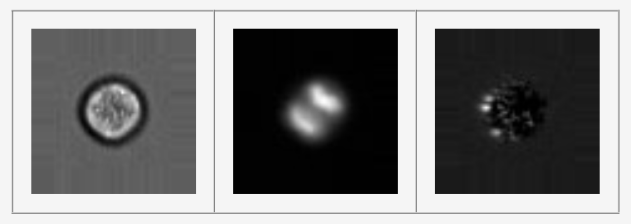

Figure 9. Representative Image from BBBC048: A Jurkat cell stained with Propidium Iodide and MPM2 antibody, exemplifying the dataset's dual-staining technique for cell cycle analysis.

Task Description. The primary task with the BBBC048 dataset is to classify discrete stages of the cell cycle. This task is crucial for understanding cellular dynamics and disease progression, particularly in the context of cancer research and other biological processes. The dataset has demonstrated its utility in this regard by achieving a six-fold reduction in error rate for cell cycle stage classification compared to previous boosting-based approaches. Accurate cell cycle stage classification using deep learning models not only advances our understanding of cellular mechanisms but also has potential applications in therapeutic interventions. The BBBC048 dataset thus serves as a pivotal resource for developing robust and generalizable models for cell cycle prediction and analysis from raw image data.

[†]More details on the dataset are found here : BBBC048.

BloodMNIST

Dataset Description. BloodMNIST, a part of the MedMNIST collection[48], is a dataset focused on images of normal blood cells. It was created from blood samples collected from individuals free of infections, hematologic or oncologic diseases, and not undergoing any pharmacologic treatments. The dataset contains a total of 17,092 images, which are divided into 7 classes. Each class represents a different type of normal cell found in human blood. The original high-resolution images ($3 \times 360 \times 363$ pixels) (see Fig.11) have been resized to fit the standardized MedMNIST format of 3 x 28 x 28 pixels. This resizing maintains the dataset's utility while ensuring compatibility with the broader MedMNIST framework.

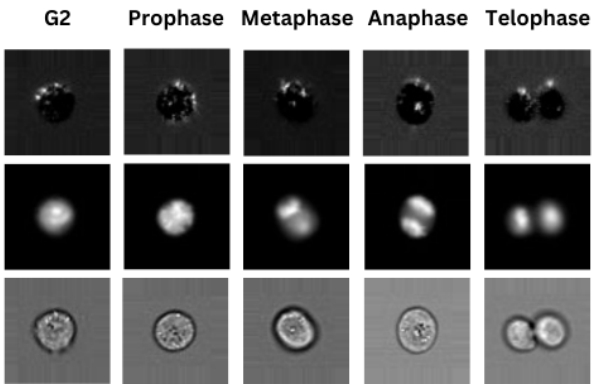

Figure 10. Randomly sampled BBBC048 images across various Cell Cycle Stages. Each column represents a multiplexed image from a distinct cell cycle, with each row corresponding to a different channel for the same cell. The visualization suggests specific channels (predominantly the middle row) and their Intra-Channel interactions is sufficient for effective cell cycle classification, lowering down the need to extract Inter-Channel information.

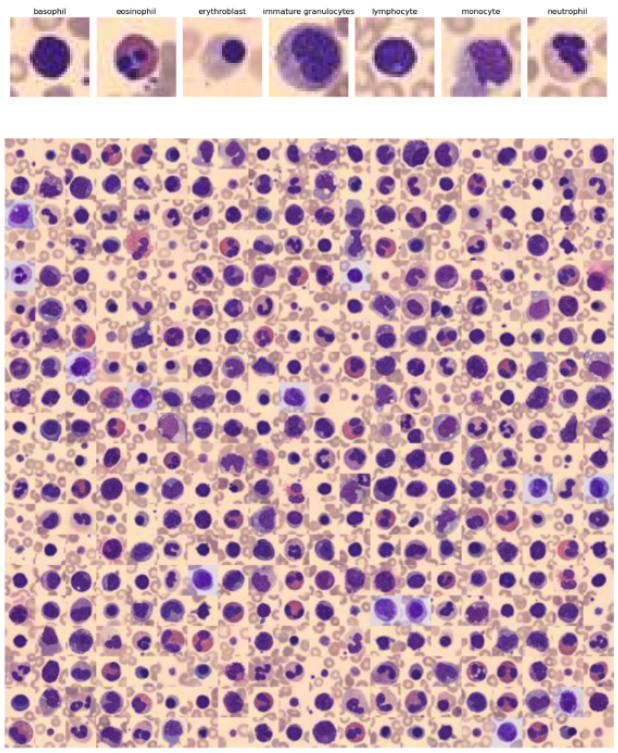

Figure 11. Sample Observations from BloodMNIST: The top row displays the 7 different classes of normal blood cells, showcasing the dataset's variety in cellular types.

Task Description. The primary application of the BloodMNIST dataset is in various biological and medical image analysis tasks, prominently including the classification of blood cell types. These tasks are crucial for understanding the characteristics of normal blood cells, which can have implications for diagnosing and studying blood-related diseases. The dataset's structure and diverse cell types make it an excellent resource for training and testing machine learning models, especially for classification tasks. Utilizing BloodMNIST provides a means to assess the performance and generalization capabilities of pretrained models in biological imaging, offering insights into how well these models can adapt and apply their learned knowledge to a range of biological tasks.

NF-kB Nuclear Translocation Assay

Dataset Description. The NF-kB Nuclear Translocation Assay in HCC1143 Cancer Cells dataset focuses on the translocation of the NF-kB protein from the cytoplasm to the nucleus within HCC1143 cancer cells. This process is central to activating the NF-kB pathway, a critical element in cellular responses to various stimuli, especially in cancer biology. The assay encompasses several stages: cell culture, cell seeding, treatment, fixation and permeabilization, and staining. During treatment, cells are exposed to TNF-alpha to induce NF-kB translocation. Staining is performed using specific antibodies for NF-kB and DAPI for nuclear labeling, ensuring clear differentiation between cytoplasmic and nuclear regions. The images acquired through high-content screening fluorescence microscopy provide detailed visualizations of NF-kB translocation post-treatment.

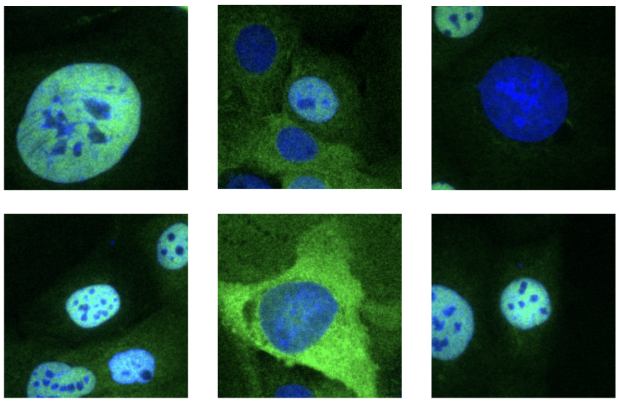

Figure 12. Representative Images from the NF-kB Nuclear Translocation Assay: Visualizing NF-kB protein movement in HCC1143 cancer cells, post TNF-alpha treatment.

Task Description. This assay focuses on quantifying NF-kB translocation from the cytoplasm to the nucleus—a crucial regression task in understanding cancer biology. It provides a standardized approach to assess NF-kB activation in response to TNF-alpha. By comparing treated and untreated cells, the assay enables the delineation of NF-kB activation's degree and kinetics. This is pivotal for under-

standing the dynamics of NF-kB movement and its implications in cancer biology. The data derived from this assay can be used across different cell lines and under varying conditions, making it a valuable tool for future studies aimed at quantifying translocation extent and deciphering its broader implications in the field of cancer biology.

BBBC021

Dataset Description. The BBBC021 dataset[‡] from the Broad Bioimage Benchmark Collection contains 39,600 images of MCF-7 breast cancer cells. These cells, a relevant model for p53-wildtype breast cancer research, were treated with 113 small molecules at eight different concentrations. The imaging process employed fluorescent microscopy to highlight DNA, F-actin, and Beta-tubulin. This dataset offers an extensive view into the cellular morphology of MCF-7 cells under various pharmacological conditions, making it an invaluable resource for drug discovery and cell biology research.

MoA Task Description. The primary task associated with the BBBC021 dataset is the prediction of drug mechanisms of action (MoA) through image-based phenotypic profiling. This involves analyzing cellular morphological changes induced by a diverse range of chemical compounds. The dataset facilitates the identification of 12 distinct primary mechanisms, providing critical insights into how different compounds affect cellular morphology and behavior. This information is essential in drug discovery, as it helps in understanding the therapeutic effects and MoA of various compounds. The detailed cellular response analysis to these compounds is key to identifying new drug candidates and comprehending the cellular impact of existing drugs, thereby contributing significantly to cancer research and treatment.

Channel Prediction Task Description. In the realm of cell morphology analysis, the ability to predict cellular components imaged in different channels is pivotal. Given the practical limitations in the number of channels that can be imaged before cell degradation, predicting one channel based on others could enhance cell understanding at reduced costs and augment existing datasets. For the BBBC021 dataset, a key task is predicting the Actin component (second channel) using the other two channels. We specifically approach this task in this paper by utilizing only the CLS token representation from the input channels for Actin channel prediction, allowing an assessment of the specific effects and fidelity of our model representations in channel prediction.

Bray et al

Dataset Description. The Bray et al. dataset[5], detailed in Gigascience, presents an extensive array of images and morphological profiles derived from the Cell Painting assay. It encompasses 919,265 five-channel fields of view, featuring images from 30,000 small-molecule treatments and covering 30,616 distinct compounds. This dataset is particularly valuable for its detailed morphological features extracted from individual cells and at population levels, including quality-control metrics and chemical annotations for the compounds used. It stands as a significant resource for comparing cellular states under various chemical perturbations, offering a comprehensive view of cell morphology and function under diverse treatment conditions.

MoA Task Description. The primary task associated with the Bray et al. dataset is, similarly to BBBC021, to explore Mechanisms of Action (MoA) and cellular responses to a wide range of chemical treatments. This involves analyzing how different small molecules impact cell morphology and function. The dataset's extensive collection of images and detailed morphological profiles allow for in-depth studies of the effects of these treatments at both the single-cell and population levels. Such analyses are crucial in cellular biology research, providing insights into the diverse impacts of small molecules on cells. This dataset serves as a rich foundation for developing and testing computational methods in drug discovery and cell biology, aiding in the advancement of research in these fields.

C. Model details

Architectural Differences

In the field of bioimaging, different approaches are adopted to encode images with multiple channels into a unified representation, each with its distinct methodology and outcomes, as shown in Figure 13.

The One Channel Encoding approach, prevalent in existing works, treats each channel of an image independently, creating a latent representation for each channel which are then concatenated. This method, while effective for images with uniform channel counts, faces challenges with heterogeneous datasets. It lacks the capacity to encode various datasets into a single representation, limiting its applicability for diverse bioimage datasets.

ChAda-ViT, or Channel Adaptive Vision Transformer, presents a novel approach. It encodes images with varying channel dimensions into a single fixed-size embedding space, incorporating both Inter-Channel and Intra-Channel attention. This method leverages token padding and masking techniques, adapted from NLP transformers, to handle images with different numbers of channels. Additionally, it introduces channel embeddings to preserve channel-specific information. ChAda-ViT's architecture enables it to process a greater number of input tokens compared to standard Vision Transformer models, thus accommodating the variability in channel count while maintaining the integrity of the

[‡]Further information is available at BBBC021.

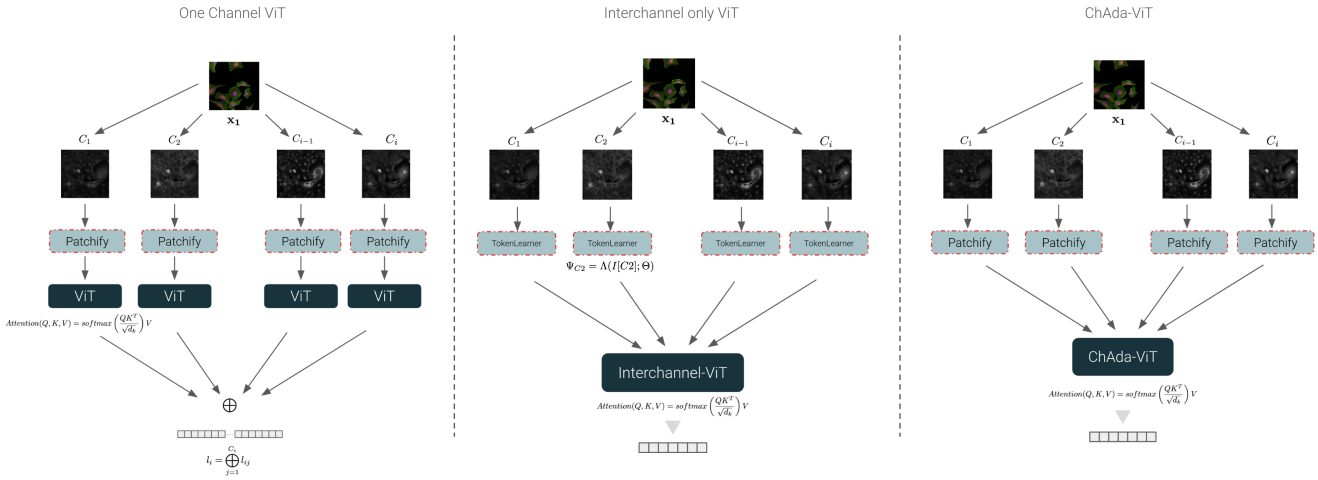

Figure 13. Comparative Architectural Overview: This figure illustrates the distinct methodologies of One Channel Encoding (left), Inter-Channel Only (center) and ChAda-Vit (right) approaches in bioimage processing. It highlights the unique processing strategies and attention mechanisms employed by each method to handle multi-channel bioimages.

self-attention mechanism.

Another variant in this realm is the Inter-Channel only approach. This method tokenizes each channel as a single large patch, compelling the model to focus solely on inter-channel relationships. By omitting Intra-Channel attention, this approach concentrates on the features derived from the relationships between individual channels, differing from the one-channel approach and ChAda-ViT, which consider both intra and inter-channel dynamics.

Patch-wise Channel Processing

Given an image I of dimensions $H \times W \times C$, where H, W, and C denote the height, width, and number of channels respectively, we dissect each channel into ***non-overlapping*** patches. For each channel c, the patch at spatial location (i, j) is denoted as $P_{c,i,j}$, and is of dimensions $p \times p$ where $p = 16$ in our optimal model configuration.
Formally, the process of patch extraction and linear projection can be expressed as follows:

$$P_{c,i,j} = I[c, i \cdot p : (i+1) \cdot p, j \cdot p : (j+1) \cdot p]$$

Subsequently, each patch $P_{c,i,j}$ is linearly projected using a shared Conv2D layer f with learned parameters θ, yielding the projected patch $\hat{P}_{c,i,j}$:

$$\hat{P}_{c,i,j} = f(P_{c,i,j}; \theta)$$

Thus, ***each channel of image I is transformed into a set of linearly projected patches***, and the operation is performed identically across all channels to maintain a consistent projection space. This results in a set of projected patches $\hat{P}$ for the entire image I, which are then further processed through subsequent stages of the Channel-Adaptive Vision Transformer architecture.

Positional Embeddings

Positional information is crucial for retaining the spatial layout of an image through the transformation process. To encode this information, we introduce a differentiable positional embedding to each patch. The positional embedding for a patch at spatial location (i, j) is denoted as $\text{pos}_{i,j}$.
Formally, the positional embedding is added to the linearly projected patch $\hat{P}_{c,i,j}$ from the previous stage as follows:

$$\tilde{P}_{c,i,j} = \hat{P}_{c,i,j} + \text{pos}_{i,j}$$

This operation is performed for each patch, across all channels, ensuring that patches located at the same spatial coordinates, albeit in different channels, receive the same positional embedding.

Through this mechanism, the spatial coherence of the original image is preserved across its transformation into the joint embedding space.
The positional embeddings $\text{pos}_{i,j}$ are learnable parameters that are optimized during the training process to better capture the spatial relationships among patches.

Channel Embeddings

Channel embeddings are also introduced to encapsulate the channel order information, ensuring the model can discern patches from different channels even when they are at the same spatial location. Let chan_c denote the channel embedding for channel c.
The channel embedding is added to the previously obtained representation $\tilde{P}_{c,i,j}$ from the positional embedding stage as follows:

$$\bar{P}_{c,i,j} = \tilde{P}_{c,i,j} + \text{chan}_c$$

This operation augments the patch representation with channel-specific information, ensuring that the model can differentiate patches from distinct channels. The channel embeddings chan_c are learnable parameters that are optimized during the training process, allowing the model to learn and represent channel-wise order information effectively.
Through the integration of channel embeddings, our model can robustly handle the differentiation between patches across channels, addressing a critical challenge in processing multichannel images.

Padding Mechanism for Uniform Sequence Generation

In order to standardize the sequence length across different multichannel images, a padding strategy is employed. Let C_{max} denote the maximum number of channels across the dataset, and C denote the number of channels in a given image I. The difference $D = C_{\text{max}} - C$ denotes the number of missing channels that need to be padded.
For each missing channel d where $d = 1, 2, \ldots, D$, a padding token pad_d is generated and ***appended*** to the sequence of patches. Formally, the padded sequence of patches for image I is denoted as $\text{Seq}_{\text{pad}}(I)$ and is given by:

$$\text{Seq}_{\text{pad}}(I) = \{\bar{P}_{c,i,j}\}_{c=1}^{C} \oplus \{\text{pad}_d\}_{d=1}^{D}$$

, where $\{\bar{P}_{c,i,j}\}_{c=1}^{C}$ denotes the sequence of channel-augmented patches from the previous stage, $\{\text{pad}_d\}_{d=1}^{D}$ denotes the set of padding tokens for the missing channels and where $\oplus$ denotes the concatenation operation.

This padding mechanism ensures that the sequence length is ***uniform*** across all images, facilitating a consistent input structure for the subsequent processing within the Transformer architecture.

Handling Padded Sequences in Self-Attention

The self-attention mechanism is central to the Transformer architecture. However, the presence of padded tokens can distort the attention computation. To tackle this, we utilize a `src_key_padding_mask` to indicate the locations of the padded tokens, ensuring they are excluded from the attention computation.

Let $\text{Seq}_{\text{pad}}(I)$ denote the padded sequence of patches for image I from the previous stage. The `src_key_padding_mask` is a binary mask of dimensions $N \times T$, where N is the batch size and T is the sequence length, with ones indicating the locations of padded tokens and zeros elsewhere.

This `src_key_padding_mask` for image I is constructed as follows:

$$\text{mask}_{n,t} = \begin{cases} 0 & \text{if } \text{Seq}_{\text{pad}}(I)[t] \text{ is a padded token} \\ 1 & \text{otherwise} \end{cases}$$

The self-attention computation is then modified to exclude the influence of padded tokens. Let Q, K, V denote the query, key, and value matrices respectively, and $\text{Attn}(Q, K, V)$ denote the original attention computation. The modified attention computation $\text{Attn}_{\text{mask}}(Q, K, V)$ is given by:

$$\text{Attn}_{\text{mask}}(Q, K, V) = \text{Attn}(Q \odot \text{mask}, K \odot \text{mask}, V \odot \text{mask})$$

, where $\odot$ denotes element-wise multiplication. Through this modification, the self-attention mechanism effectively ignores the padded tokens, ensuring accurate attention computation across the real patches.

Class Token Integration

A distinct class token, denoted as $[CLS]$, is ***prepended*** to the sequence of patches to obtain a global representation of the image. The class token is a *differentiable* parameter that is optimized during the training process. Let $\text{Seq}_{\text{pad}}(I)$ denote the padded sequence of patches for image I from the previous stage.

Thus the sequence with the class token, denoted as $\text{Seq}_{\text{CLS}}(I)$, is constructed as follows:

$$\text{Seq}_{\text{CLS}}(I) = [CLS] \oplus \text{Seq}_{\text{pad}}(I)$$

, where $\oplus$ denotes the concatenation operation.

This updated sequence $\text{Seq}_{\text{CLS}}(I)$ is then fed into the Transformer, which processes the sequence through its multiple layers of self-attention and feed-forward networks. The final representation of the [CLS] token captures a global representation of the image, which is utilized for

downstream tasks.

TokenLearner in Interchannel only ViT

The TokenLearner mechanism plays a pivotal role in the interchannel-only Vision Transformer architecture. It is essentially a series of convolutional layers with the primary objective of the TokenLearner to segment each channel of an image into one condensed token. This segmentation is akin to creating a mosaic, where each token represents a concentrated summary of information from a specific channel of the image.
Formally, the TokenLearner operates as follows:

$$\Psi_c = \Lambda(I[c]; \Theta)$$

Here, Ψ_c represents the transformation of channel c into a tokenized output. The function Λ denotes the convolutional layers within the TokenLearner, characterized by learnable parameters Θ. The input $I[c]$ represents the channel c of the image I.
The convolutional layers within the TokenLearner are meticulously designed to process each channel c of the image I, converting them into a series of large tokens. These tokens are the essence of the image's information condensed into more manageable and informative segments.
This approach allows the TokenLearner to maintain a consistent tokenization process across all channels of each image, ensuring uniformity and coherence in the representation of the image's information. The resultant tokenized channels are then seamlessly integrated into the ViT architecture, enabling the model to process and interpret the image data with enhanced efficiency and precision.

Through this method, the TokenLearner was meant to learn meaningful feature maps for any channel in order to "tokenize" them into a smaller embedding space.

D. Experimental Evaluation

Our experimental approach encompasses several distinct methods to assess the performance of our models across a range of tasks. All experiments were performed under 5 different seeds, and the mean and standard deviation of the results were reported :

Linear Probing: For datasets BloodMNIST, BBBC048, CYCLoPs, and NF-kB Nuclear Translocation Assay (Transloc), we employed linear probing. This involved freezing the encoder of our model and appending a trainable linear layer to the CLS token. The linear layer was then trained specifically for the task associated with each dataset. This method allows us to evaluate the representational quality of the encoded features in a variety of biological contexts.

Channel Reconstruction: In this task, we focus on reconstructing a target channel from given input channels. The encoder is kept frozen, and a simple decoder is added atop the CLS token. The decoder comprises of two fully connected layers, five convolutional layers, followed by a sigmoid function, to predict the target channel accurately. We ensured that the decoder was scaled to maintain an equivalent number of parameters (approximately 5.2 million) for both the One Channel Approach and ChAda-ViT, facilitating a fair comparison of their reconstruction capabilities.

Performance Metrics: For classification tasks, we utilize top-1 accuracy as our primary metric. For regression tasks, the R2 score is employed to measure the accuracy of our predictions. Specifically, for the Channel Prediction task, we use Mean Absolute Error (MAE) and Mean Squared Error (MSE), along with the R2 score. Here, the prediction involves comparing the flattened predicted channel against the flattened target channel. This comparison is conducted on the entire dataset in a single evaluation, providing a comprehensive view of our model's prediction accuracy.

E. Additional Evaluations

	Linear Eval		KNN Eval	
	BBBC048	Cyclops	BBBC048	Cyclops
ChAda-ViT Moyen16	77±0.38	71.89±3.49	78.37±0.22	51.12±3.87
One Channel ViT Tiny/16	77.48±0.14	70.9±3.02	78.43±0.04	38.82±0.08
3 Ch. ViT Tiny/16	72.08±0.14	40.2±2.36	77.79±0.02	37.24±2.38

Table 3. Using All Token Evaluation setting, ChAda outperforms both One Channel Approach and the standard 3 Channel ViT (trained on the 3 channel subset of the IDRCell100k), proving the added value of integrating channel level attention over classic approaches.

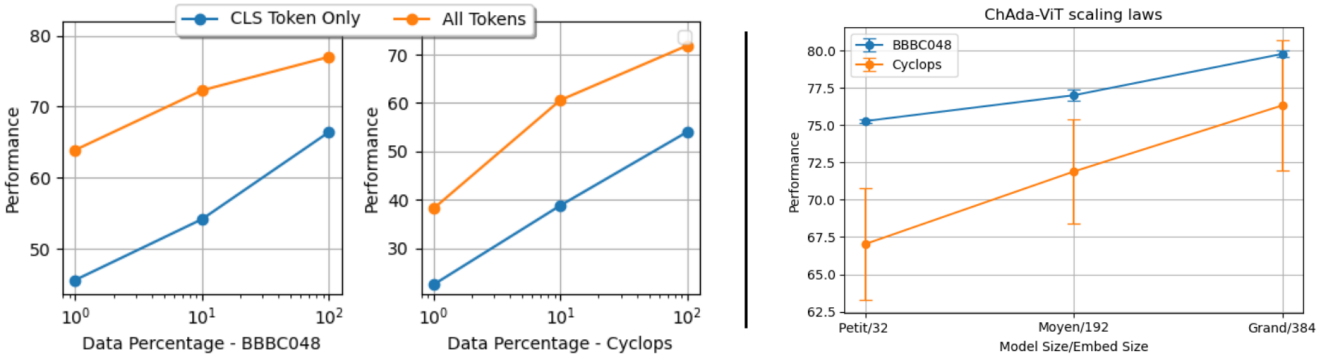

Figure 14. **Left :** Comparison of evaluation results using ChAda-ViT Moyen/16 with CLS token eval and All output token eval. All token evaluation shows a higher performance than using CLS token. **Right :** Scaling laws of the performance of the model with All Token Eval, when trained with more parameters and larger embedding sizes, confirming the scalability of the ChAda-ViT architecture.

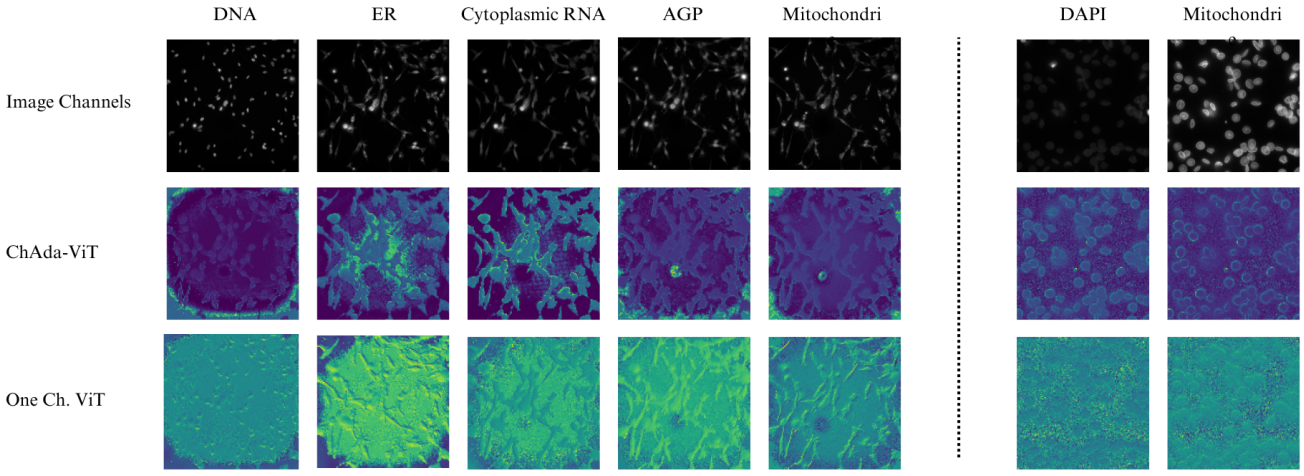

Figure 15. Additional Attention Maps for images with 2 channels (right) and 5 channels (left), with One Channel ViT and ChAda-ViT.

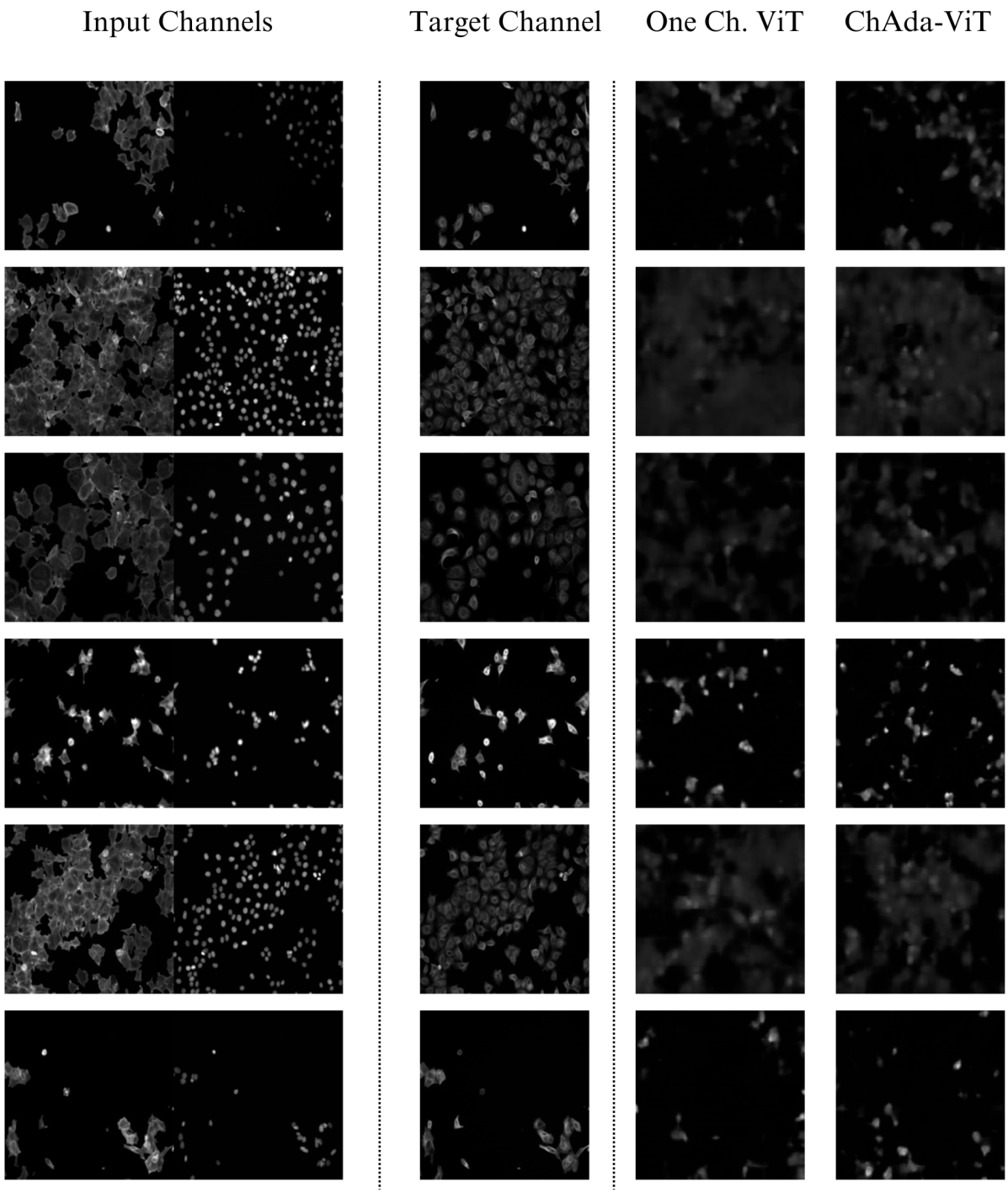

Figure 16. Channel Prediction in BBBC021: Demonstrating ChAda-ViT's enhanced spatial distribution accuracy in reconstructing cell images, compared to the One Channel approach. This superiority is evident even with a basic convolutional decoder utilizing only the *CLS* token, as highlighted in the first row of examples.