insitro

Channel Vision Transformers: An Image Is Worth 1 x 16 x 16 Words

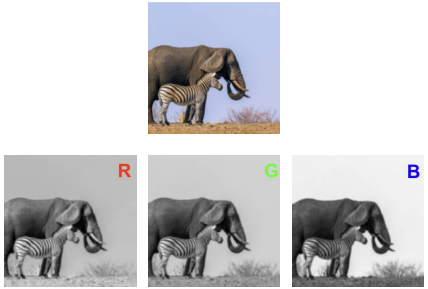
Yujia Bao*, Srinivasan Sivanandan*, Theofanis Karaletsos

ICLR 2024

*equal contribution

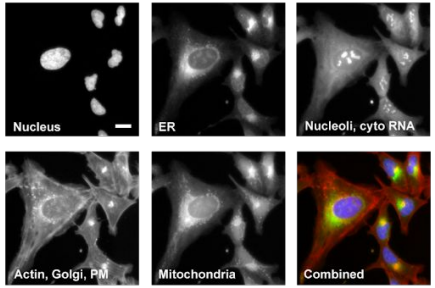
Multi-Channel Imaging: Challenge 1

RGB imaging

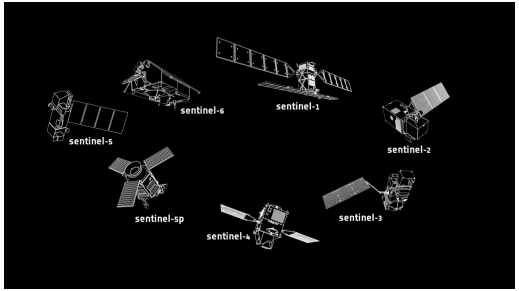


The example zebra image exhibits high correlations across its RGB channels.
Source: ImageNet

Multi-channel imaging



In **cell imaging**, different channels correspond to different stains. They reveal different biological properties.
Source: JUMP-CP



In **satellite imaging**, different signals are acquired from different satellites.
Source: GIS Geography

Challenge 1: In multi-channel imaging, different channels often exhibit independent knowledge → modeling cross-channel interactions

insitro

Multi-Channel Imaging: Challenge 2

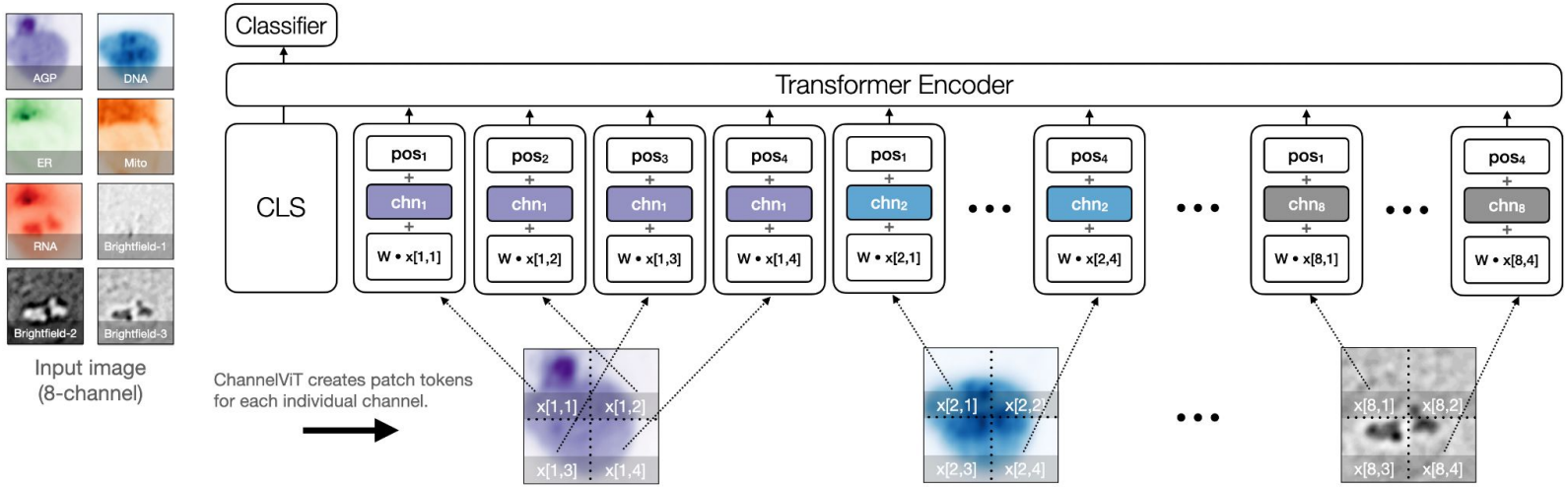
Cell imaging datasets	Available channels								
	Hoechst	ConA	Phalloidin	Syto 14	MitoTracker	WGA	Brightfield 1	Brightfield 2	Brightfield 3
JUMP-CP	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
RxRx1	Yes	Yes	Yes	Yes	Yes	Yes			
RxRx19a	Yes	Yes	Yes	Yes		Yes			

References: https://jump-cellpainting.broadinstitute.org/ https://www.rxrx.ai/datasets

Challenge 2: Channel availabilities can be very different across datasets
→ modeling data with different input channels
& enhancing robustness when there are missing channels

insitro

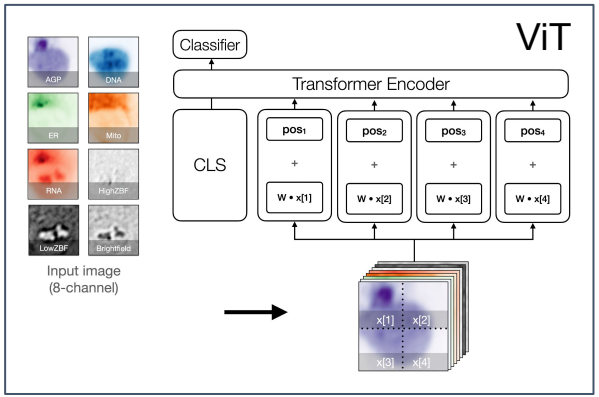
ChannelViT for cross-positional and cross-channel reasoning

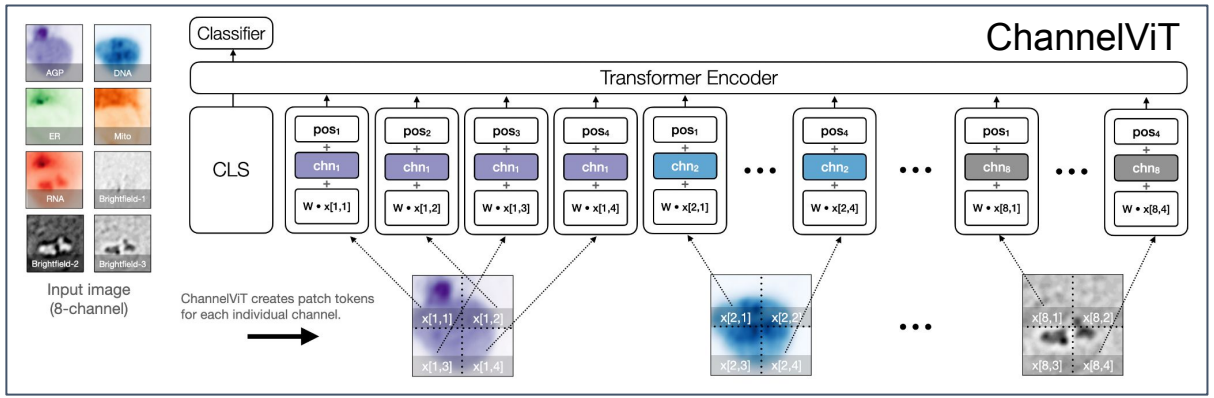


- ChannelViT creates `image tokens` by looking at all **1-channel** image patches.
- Hierarchical channel sampling (HCS) for **efficient** training and **robust** generalization.

insitro

ViT vs. ChannelViT





ViT (left):

$$\text{PatchEmbedding}(x[i]) = \text{PosEmbedding}_i + W \cdot x[i]$$

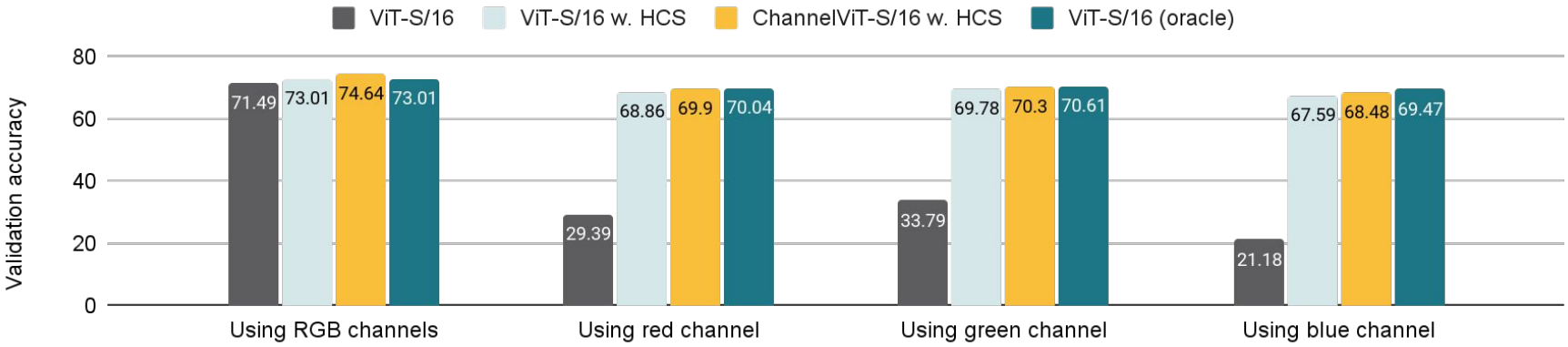
- Map information across all channels linearly into a single patch token;
- Parameters are not shared across channels.

ChannelViT (right):

$$\text{PatchEmbedding}(x[c, i]) = \underbrace{\text{PosEmbedding}_i + W \cdot x[c, i]}_{\text{(parameters shared across channels)}} + \underbrace{\text{ChannelEmbedding}_c}_{\text{(channel-specific parameters)}}$$

insitro

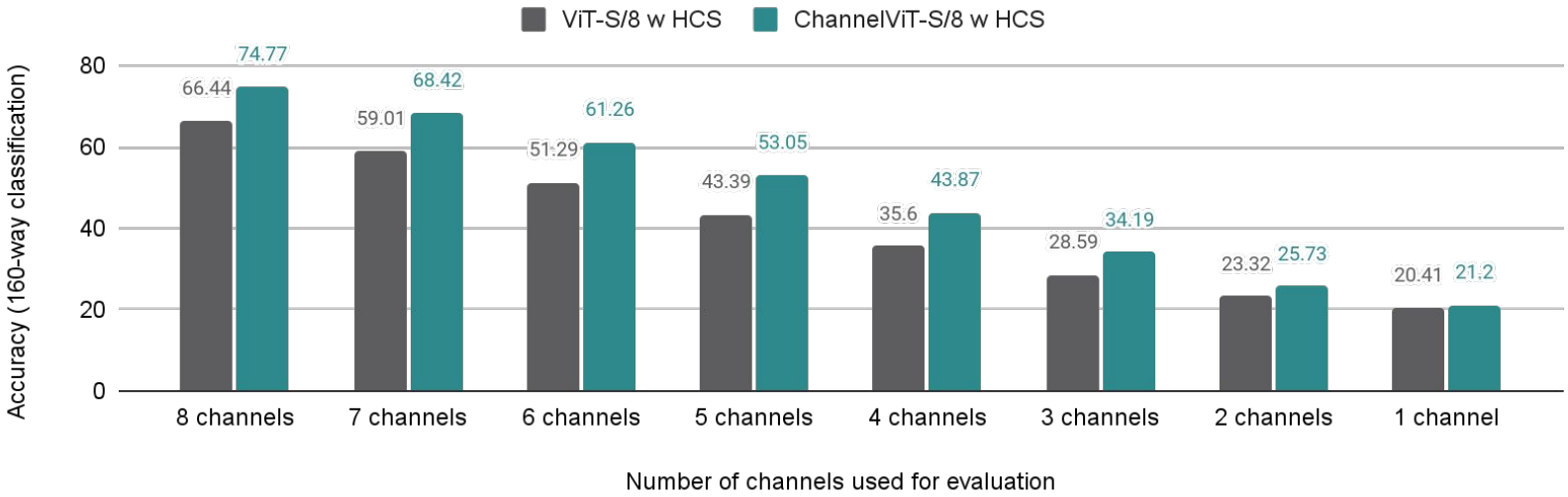
ImageNet



1. HCS is crucial for improving test-time channel robustness;
2. ChannelViT consistently outperforms ViT and closes the gap to the oracle experts which are trained on the specific channel configuration that the model is tested on.

insitro

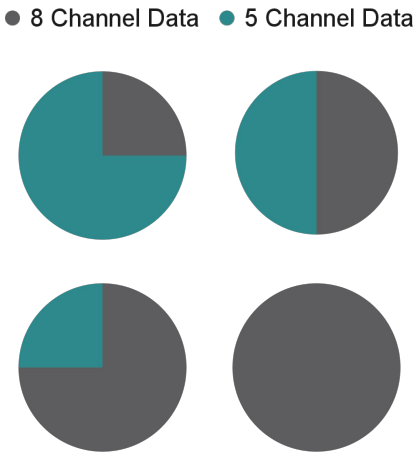
JUMP-CP: cell imaging benchmark

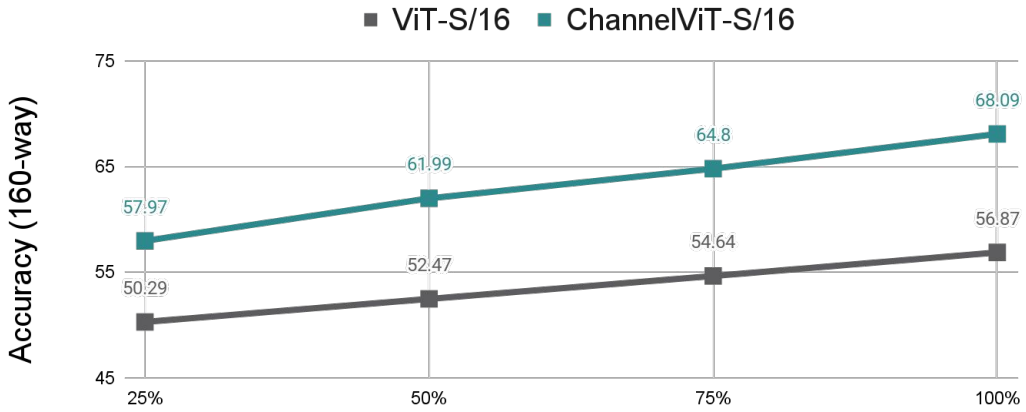


ChannelViT consistently and significantly outperforms ViT across all evaluation settings.

insitro

Training on datasets with different channels





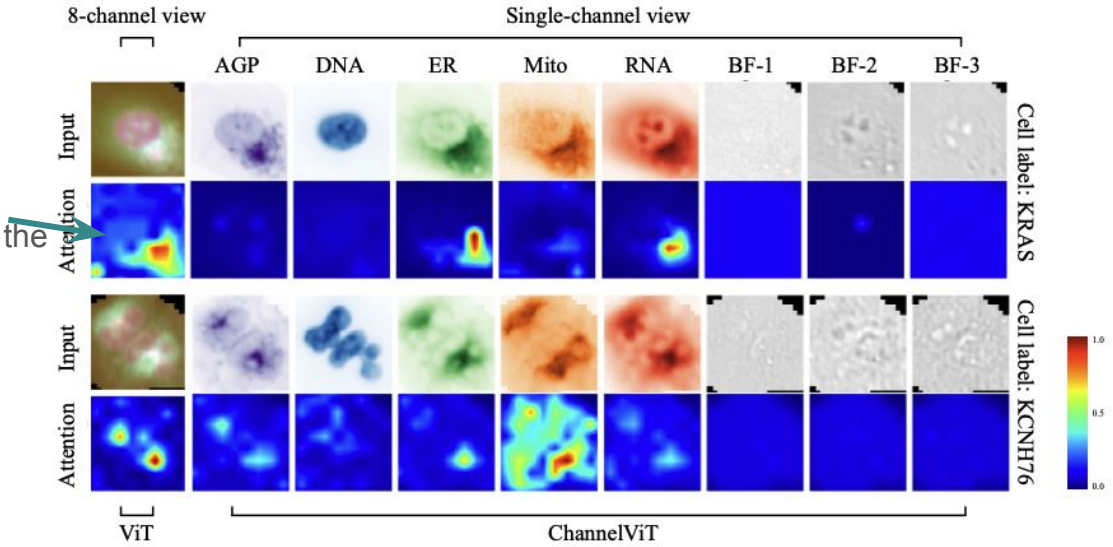
Percentages of 8-channel JUMP-CP data. The remaining data

ChannelViT effectively integrates datasets with different channel configurations.

insitro

ChannelViT offers extra interpretability

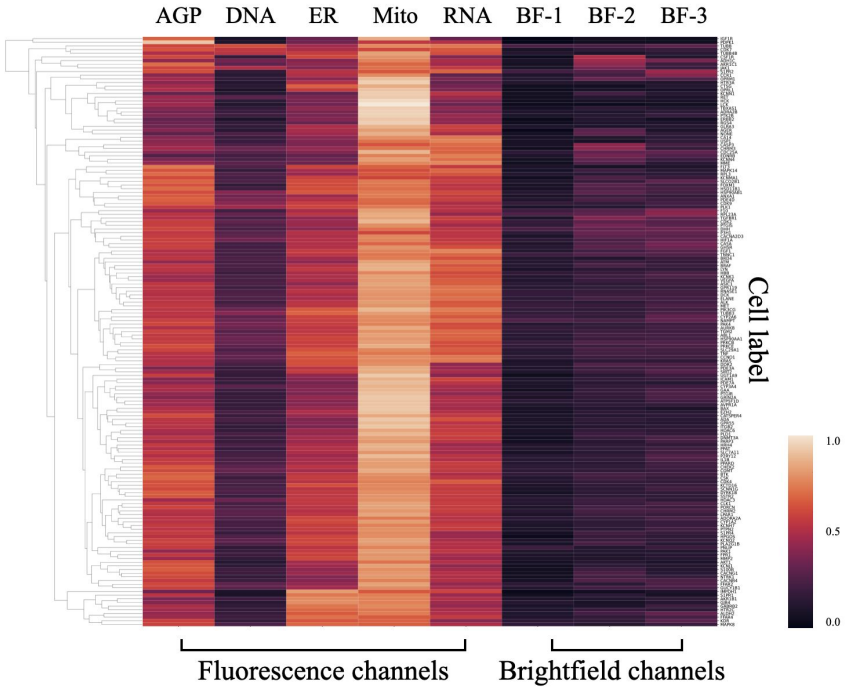
For ViT, it is difficult to understand what part of the input contributes to the prediction.



Channel highlights the contributions made by each individual channel, providing better interpretation to the user.

insitro

Visualizing perturbed-gene specific channel attribution



ChannelViT focuses on different input channels depending on the gene that has been perturbed in the cell.

This enables us to understand the underlying biological relationship between different genes.

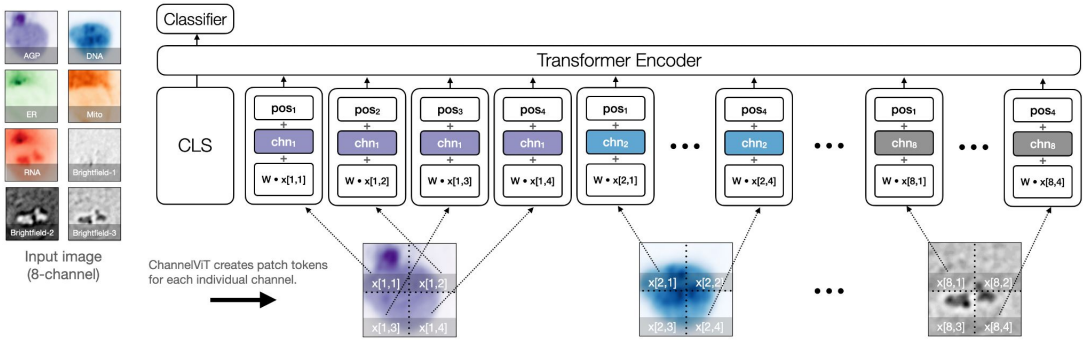
insitro

Conclusion

Channel Vision Transformers: An Image is Worth 1 x 16 x 16 Words

Yujia Bao*, Srinivasan Sivanandan*, Theofanis Karaletsos

Paper: arxiv.org/abs/2309.16108
Code + Pretrained Model Weights: https://github.com/insitro/ChannelViT



- cross-channel and cross-positional reasoning;
- significant performance gains on multi-channel imaging applications and standard imaging benchmarks;
- extra interpretability to the models' decisions.