ChannelSFormer: A Channel Agnostic Vision Transformer for Multi-Channel Cell Painting Images

Jingwei Zhang[*]
Stony Brook University
jingwezhang@cs.stonybrook.edu

Srinivasan Sivanandan
Insitro
srinivasan@insitro.com

Abstract

High-content imaging using the Cell Painting assay is a cornerstone of modern drug discovery, generating multi-channel images where each channel reveals distinct cellular components. Existing Vision Transformers (ViTs) struggle with this data, as their global self-attention mechanisms are computationally expensive and become hard-coded to a specific number of channels, limiting flexibility. To address this, we introduce ChannelSFormer, a channel-agnostic Transformer architecture. ChannelSFormer decomposes the standard self-attention into two distinct steps: spatial-wise attention, which learns spatial relationship within each channel, and channel-wise attention, which learns relationships across channels. We also use per-channel class (CLS) token for each channel, which are duplicated from a single CLS token, to better capture per-channel information. ChannelSFormer eliminates the need for fixed channel embeddings, making the model adaptable to varying channels. Evaluation on the JUMP-CP dataset shows that ChannelSFormer surpasses SOTA methods by 4.12% - 7.58% in accuracy and is 27% - 281% faster.

1 Introduction

High-content imaging using the Cell Painting assay is a cornerstone of modern drug discovery [1–3]. This technique generates images with multiple channels, typically five or more, where each channel uses a specific fluorescent dye to reveal a distinct cellular component, such as the nucleus, mitochondria, or cytoskeleton. Because each channel provides a unique layer of biological information, the resulting data is highly heterogeneous among channels [4]. This creates a significant industrial challenge, as research pipelines often evolve through experiments that use varying numbers of channels or involve the introduction of novel fluorescent markers. Consequently, there calls for a flexible, channel-agnostic vision model.

The rise of Vision Transformers (ViTs) [5] offers a promising foundation for analyzing these complex cell images, yet their existing application introduces its own set of challenges. Given the heterogeneous nature of channels, the standard ViTs on cell images tokenize each channel independently [6–9]. Existing methods then typically employ a single, global self-attention mechanism that processes the concatenated tokens from all channels. To distinguish between tokens originating from different channels within this shared computational space, a learnable channel embedding is added to each token [6, 10]. Although this enables correct modeling of cross-channel interactions, it introduces a critical limitation: the model becomes channel-specific, hard-coded to the initial channel configuration and unable to adapt to new experimental setups with different channels.

Furthermore, this reliance on a single global attention mechanism creates a computational bottleneck. By concatenating tokens from all channels, the input sequence length becomes dramatically

[*] Research supporting this publication conducted while authors were employed at Insitro.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The 3rd Workshop on Imageomics: Discovering Biological Knowledge from Images Using AI.

inflated—for example, a 224×224 image with eight channels contains eight times more tokens than a natural RGB image of the same resolution. This exacerbates the quadratic computational complexity inherent to the self-attention mechanism, leading to their slow speed [6].

To address these limitations, we introduce ChannelSFormer, a novel Transformer architecture which decomposes of the standard, monolithic self-attention into two subsequent attentions: a channel-wise attention that learns interactions across different channels, and a spatial-wise attention that learns interactions across spatial locations within each channel. This decoupled design directly tackles the aforementioned limitations by treating channels as a distinct dimension, thereby eliminating the need for fixed channel embeddings and alleviating the quadratic scaling bottleneck. We also use per-channel class (CLS) tokens for different channels, which are duplicated from a single CLS token, to allow it lean channel specific information and maintain a channel agnostic model. This not only improve the performance but also allow the model to output per-channel feature tokens along with a global image-wise feature token, allowing more flexibility for the downstream tasks. We evaluated the proposed method on the public Jump-CP [11] dataset and shows that it surpass SOTA methods by 4.12% to 7.58% in accuracy and is 27% to 281% faster.

2 Related work

The success of ViTs [5] on standard RGB images has inspired their variant on handling multi-channel cell images. DepthwiseViT [8] draws inspiration from depthwise convolutions, processing each channel separately before aggregating features into a new representation for the main ViT backbone. ChannelViT [6] tokenizes each channel with a shared linear projection and adds a learnable channel embedding to preserve channel-specific information within a global attention mechanism. Similarly, ChAda-ViT [9] padded images to the same number of channels and used a global attention cross various channels. CA-MAE [7] also utilize the global attention but did not use the channel embedding and thus is channel agnostic. To combat redundancy and improve robustness, some methods employ sophisticated channel sampling techniques. ChannelViT [6] introduces Hierarchical Channel Sampling (HCS) as a regularization method, and DiChaViT [10] proposes Diverse Channel Sampling (DCS) to actively select less similar channels during training. However, a shared limitation persists, as most architectures depend on a global self-attention mechanism that necessitates learnable channel embeddings, thus preventing true channel agnosticism; even when these embeddings are removed for flexibility, as in CA-MAE [7], performance degrades significantly as the global attention struggles to differentiate channel-specific features, as we show in the result section.

3 Method

We present **ChannelSFormer**, a Vision Transformer architecture designed to be efficient, effective, and agnostic to the number of input channels in multi-channel microscopy images.

Overall architecture. As illustrated in Fig. 1, the input to our model is a multi-channel image $X \in \mathbb{R}^{C\times H\times W}$, where C is the number of channels, and H, W are the height, width. Each channel is processed independently by a shared patch embedding layer, which divides the channel into a sequence of N flattened 2D patches, each of size $P \times P$. These patches are then mapped to a D dimensional embedding space. A standard learnable position embedding is added to the patch tokens for each channel to retain positional information. These tokens along with the class (CLS) tokens, are fed into a series of L ChannelSFormer encoders, which features our spatial-channel attention mechanism. The final image token is obtained by aggregating the per-channel CLS tokens through a multi-head latent query attention mechanism. This approach allows the model to output both channel-specific and global image-wise representations.

Per-channel CLS token. A unique aspect of ChannelSFormer is the class (CLS) token. Instead of a single CLS token, we duplicate it for each of the C channels, creating a set of per-channel CLS tokens $\{\mathrm{CLS}_1, \ldots, \mathrm{CLS}_C\}$. The per-channel CLS tokens are the **same** before the first ChannelSFormer encoder. Then each per-channel CLS tokens goes through the spatial-channel attention on different tokens, they become **different** and learn channel specific information.

Spatial-channel attention. Inspired by [12], each ChannelSFormer encoder is the same as the standard ViT encoder but decompose the the multi-head self attention (MSA) into two MHAs along the spatial and channel dimensions respectively. Given the input sequence of patch embeddings $z^{(l-1)} \in \mathbb{R}^{C\times(N+1)\times D}$ from the previous layer, we first apply the **spatial-wise attention**, which

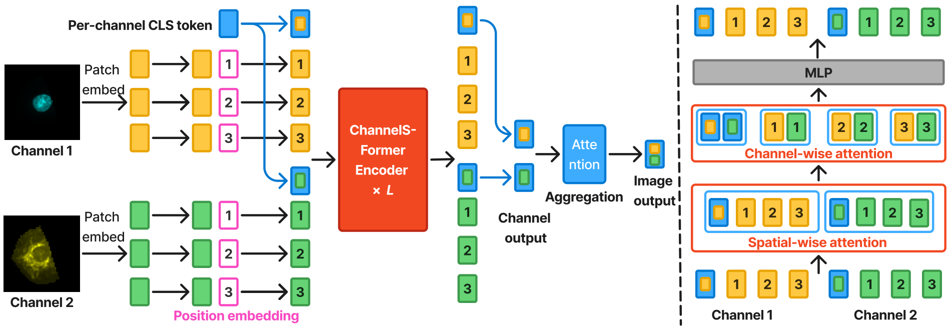

Figure 1: **Left**: The overall architecture of ChannelSFormer. Each channel of an input image is independently processed by a shared patch embedding layer, and position embeddings are added. The resulting tokens, along with per-channel class (CLS) tokens for each channel which are duplicated from a signle CLS token, are fed into a series of L ChannelSFormer encoders. Finally, the per-channel CLS tokens are aggregated via an attention mechanism to produce an image-level feature representation. **Right**: Architecture of the ChannelSFormer encoder. Instead of a single global attention block, our design first applies spatial-wise attention independently within each channel. It then performs channel-wise attention independently for each spatial location to exchange information across channels. Residual connections and layer normalization are omitted for clarity.

conducts self-attention independently within each channel. This operation is performed in parallel for each channel $c \in \{1, \dots, C\}$, ensuring it is channel-independent. The formulation is:

$$z'_c = \text{SAttn}(\mathcal{LN}(z^{(l-1)})_c) + z_c^{(l-1)} \quad \text{SAttn}(x_c) = \text{SM}\left((x_c W_Q)(x_c W_K)^T / \sqrt{D}\right)(x_c W_V) \quad (1)$$

where SM represents the softmax function, $z_c^{(l-1)} \in \mathbb{R}^{(N+1)\times D}$ represents the tokens for a single channel c, subindex c represents channel index, $\mathcal{LN}$ denotes Layer Normalization and the weight matrices $W_Q, W_K, W_V \in \mathbb{R}^{D\times D}$ are shared across all channels. The outputs for each channel are concatenated into $z' \in \mathbb{R}^{C\times(N+1)\times D}$. Next, to enable information flow across channels, we apply **channel-wise attention**. The intermediate representation z' is reshaped so that attention is computed across C channel dimensions for each spatial location $s \in \{1, \dots, N+1\}$. This models the interaction of different channels for each specific patch region. Formally,

$$z''_s = \text{CAttn}(\mathcal{LN}(z')_s) + z'_s \qquad \text{CAttn}(x_s) = \text{SM}\left((x_s W'_Q)(x_s W'_K)^T / \sqrt{D}\right)(x_s W'_V) \quad (2)$$

where subindex s represents spatial index and the weight matrices $W'_Q, W'_K, W'_V \in \mathbb{R}^{D\times D}$ are shared across all spatial locations. Finally, the output from the attention modules is passed through the MLP block to produce the final output for the layer: $z^{(l)} = \text{MLP}(\mathcal{L}_{\mathcal{N}}(z'')) + z''$. By decoupling a global attention into two attention mechanisms, ChannelSFormer remains fully agnostic to the number of channels while reducing computational complexity.

4 Experiments

To evaluate ChannelSFormer, we conducted experiments on the JUMP-Cell Painting (JUMP-CP) dataset [11, 13], a large-scale public collection of cell images used for phenotypic profiling. We focused on classifying the compound treatment applied to the cells. We also use ImageNet-1k [14] for ablation analysis. Our evaluations were based on the Top-1 accuracy.

Comparison with the SOTA method. We evaluated our models on the JUMP-CP classification task, comparing their accuracy against a SOTA ChannelViT [6] baseline. As shown in Tab. 1, ChannelSFormer consistently and significantly outperforms the ChannelViT baseline across both "Tiny" and "Small" scales. Notably, ChannelSFormer-Tiny achieves a massive 7.58% improvement over its ChannelViT-Tiny counterpart with a just a slight increase in parameters (7M vs. 5M). ChannelSFormer-Tiny also outperforms ChannelViT-Small by 3.96% with just one third the number of parameters (7M vs. 22M), supporting use of our architectural design for efficient handling of multi-channel Cell Painting images.

Table 1: Comparison of the top-1 accuracy on the Jump-CP dataset.

Model	#Params	Acc. (%)
ChannelViT-Tiny	5M	69.48
ChannelSFormer-Tiny	7M	**77.06**
ChannelViT-Small	22M	73.10
ChannelSFormer-Small	29M	**77.22**

Table 2: Throughput comparison between ChannelViT and ChannelSFormer.

Model	Throughput (img/s)	
	Patch size 16	Patch size 8
ChannelViT-Tiny	187.71	13.23
ChannelSFormer-Tiny	**248.11**	**50.46**
ChannelViT-Small	84.10	6.59
ChannelSFormer-Small	**106.70**	**22.07**

Speed analysis. ChannelSFormer is not only accurate but also fast. ChannelSFormer is 27% to 32% faster than the ChannelViT with a patch size of 16 on 8 channel 224x224 images. When using a patch size of 8, the number of tokens quadruples and ChannelSFormer is nearly three times faster. This efficiency boost is achieved by the decomposition of global self-attention into channel and spatial components. This alleviates the quadratic complexity of MHA and makes it ideal for token-heavy, multi-channel Cell Painting images.

Ablation on the channel embedding. As shown in Tab. 3, removing the channel embedding in the ChannelViT, similar to [7], leads to a substantial performance drop of 8.51% on JUMP-CP. This is because ChannelViT's global self-attention mechanism relies heavily on these embeddings to distinguish between tokens originating from different channels. In contrast, the reliance on channel embeddings is greatly reduced in ChannelSFormer, as shown by the diminished impact of removing channel embeddings. This trend holds true on the ImageNet dataset as well.

Ablation on the per-channel CLS token. Compared with an "averaged CLS token" baseline (as used in TimeSFormer [12] in the video domain, where a single averaged CLS token is caclculated from all frames/channels), our per-channel CLS token achieves a significant (3.17%) boost in accuracy on the JUMP-CP dataset. Interestingly, use per-channel CLS token dropped accuracy by 1.31% on the ImageNet-1k. This contrasting behavior may be explained by fundamental differences between Cell Painting and natural images. The RGB channel of natural images often contain correlated information, where an averaged representation can be robust. Conversely, the channels in Cell Painting images carry heterogeneous information, where averaging leads to a loss of channel-specific details, thus degrading performance. This finding strongly supports the use of a channel-differentiated architecture for Cell Painting image analysis.

Ablation on the channel-wise attention. Finally, the critical importance of our channel-wise attention is confirmed by the "Spatial attention only" experiment, where we conduct channel independent attention and use an attention in the last layer to aggregate channel information, which utilized by CellCLIP [15]. Tab. 3 shows that the performance of this model plummets to 44.44%, indicating that interaction among channels should not be ignored.

Table 3: Results of the ablation studies on the Jump-CP and ImageNet-1k datasets.

Model	Jump-CP Acc. (%)	ImageNet Acc. (%)
ChannelSFormer-Tiny	**77.06**	74.26
w. channel embedding	**78.31**	74.26
w/o per-channel CLS token (TimeSFormer [12])	73.89	**75.57**
ChannelViT-Tiny	69.48	74.31
w/o channel embedding (CA-MAE [7])	60.97	74.03
Spatial attention only (CellCLIP [15])	44.44	/

5 Conclusion

In this work, we introduce ChannelSFormer, a novel Vision Transformer architecture designed for multi-channel Cell Painting images. By decoupling the self-attention mechanism into separate channel-wise and spatial-wise attention, our model effectively processes heterogeneous information from different fluorescent dyes while remaining flexible on channels. The utilization of per-channel CLS token not only improves the performance but also allows more flexibility on downstream tasks. Our experiments demonstrate that ChannelSFormer not only achieves state-of-the-art accuracy on the JUMP-CP dataset but also offers significant improvements in computational efficiency. This work represents a significant step towards more flexible and scalable models for high-content imaging analysis, paving the way for more efficient pipelines in drug discovery and biological research.

References

[1] L. Von Chamier, R. F. Laine, J. Jukkala, C. Spahn, D. Krentzel, E. Nehme, M. Lerche, S. Hernández-Pérez, P. K. Mattila, E. Karinou, *et al.*, "Democratising deep learning for microscopy with zerocostdl4mic," *Nature communications*, vol. 12, no. 1, p. 2276, 2021.

[2] F. Vincent, A. Nueda, J. Lee, M. Schenone, M. Prunotto, and M. Mercola, "Phenotypic drug discovery: recent successes, lessons learned and new directions," *Nature Reviews Drug Discovery*, vol. 21, no. 12, pp. 899–914, 2022.

[3] M. Boutros, F. Heigwer, and C. Laufer, "Microscopy-based high-content screening," *Cell*, vol. 163, no. 6, pp. 1314–1325, 2015.

[4] M.-A. Bray, S. Singh, H. Han, C. T. Davis, B. Borgeson, C. Hartland, M. Kost-Alimova, S. M. Gustafsdottir, C. C. Gibson, and A. E. Carpenter, "Cell painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes," *Nature protocols*, vol. 11, no. 9, pp. 1757–1774, 2016.

[5] 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.

[6] Y. Bao, S. Sivanandan, and T. Karaletsos, "Channel vision transformers: an image is worth 1 x 16 x 16 words," *arXiv preprint arXiv:2309.16108*, 2023.

[7] O. Kraus, K. Kenyon-Dean, S. Saberian, M. Fallah, P. McLean, J. Leung, V. Sharma, A. Khan, J. Balakrishnan, S. Celik, *et al.*, "Masked autoencoders for microscopy are scalable learners of cellular biology," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11757–11768, 2024.

[8] Z. S. Chen, C. Pham, S. Wang, M. Doron, N. Moshkov, B. Plummer, and J. C. Caicedo, "Chammi: A benchmark for channel-adaptive models in microscopy imaging," *Advances in Neural Information Processing Systems*, vol. 36, pp. 19700–19713, 2023.

[9] N. Bourriez, I. Bendidi, E. Cohen, G. Watkinson, M. Sanchez, G. Bollot, and A. Genovesio, "Chada-vit: Channel adaptive attention for joint representation learning of heterogeneous microscopy images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11556–11565, 2024.

[10] C. Pham and B. Plummer, "Enhancing feature diversity boosts channel-adaptive vision transformers," *Advances in Neural Information Processing Systems*, vol. 37, pp. 89782–89805, 2024.

[11] S. N. Chandrasekaran, B. A. Cimini, A. Goodale, L. Miller, M. Kost-Alimova, N. Jamali, J. G. Doench, B. Fritchman, A. Skepner, M. Melanson, *et al.*, "Three million images and morphological profiles of cells treated with matched chemical and genetic perturbations," *Nature Methods*, vol. 21, no. 6, pp. 1114–1121, 2024.

[12] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?," in *Icml*, vol. 2, p. 4, 2021.

[13] Y. Bao and T. Karaletsos, "Contextual vision transformers for robust representation learning," *arXiv preprint arXiv:2305.19402*, 2023.

[14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009.

[15] M. Lu, E. Weinberger, C. Kim, and S.-I. Lee, "Cellclip–learning perturbation effects in cell painting via text-guided contrastive learning," *arXiv preprint arXiv:2506.06290*, 2025.

Training hyper-parameters All models on the JUMP-CP dataset are trained from scratch for 100 epochs using the AdamW optimizer, a cosine learning rate schedule with 10 epoch of warmup, the initial learning rate was set to 5e-4 with a batch size of 256. Random resize cropping, horizontal and vertical flipping, are applied during training.

Multi-head latent query attention mechanism The attention we used at the end of the network is the multi-head latent query attention, it aggregates the per-channel class tokens at the end of the network CLS^L into a image-wise token z_{img}:

$$z_{img} = SM(\frac{qK^T(CLS^L)}{\sqrt{D}}V(CLS^L)) \tag{3}$$

where q is a single learnable token, similar to the class token, d is the dimension of tokens and $K(\cdot), V(\cdot)$ are learnable linear functions. Unlike the self attention mechanism, this attention only has one query and thus has linear time complexity.