Building Robust Vision Encoders for Cross-Dataset Evaluation in Immunofluorescent Microscopy

Umar Marikkar[1] Syed Sameed Husain[2] Muhammad Awais[1,2] Sara Atito[2]
[1]Surrey Institute for People-Centred AI [2]CVSSP
University of Surrey, UK
{u.marikkar, sameed.husain, muhammad.awais, sara.atito}@surrey.ac.uk

Abstract

Immunofluorescence (IF) images reveal detailed information about structures and functions at the subcellular level. However, unlike RGB images, IF datasets pose challenges for deep learning models due to their inconsistencies in channel count and configuration, stemming from varying staining protocols across laboratories and studies. Although existing approaches build channel-adaptive models for training, they do not perform evaluations across IF datasets with unseen channel configurations. To address this, we first introduce a biologically informed view of cellular image channels by grouping them into either context or concept, where we treat the context channels as a reference for the concept channels in the image. We leverage this view to propose Channel Conditioned Cell Representations (C3R), a framework that learns representations that transfers well to both in-distribution (ID) and out-of-distribution (OOD) datasets which contain same and different channel configurations, respectively. C3R is a two-fold framework comprising a channel-adaptive encoder architecture and a masked knowledge distillation training strategy, both built around the context-concept principle. We find that C3R outperforms existing benchmarks on both ID and OOD tasks, while yielding state-of-the-art results on frozen encoder evaluation on the CHAMMI benchmark. Our method opens a new pathway for cross-dataset generalization between IF datasets, with no need for retraining on unseen channel configurations.

1. Introduction

Immunofluorescence (IF) images reveal subcellular patterns that provide rich information about cell structure and function, enabling machine learning methods to perform tasks such as disease prediction, subcellular localization, and prediction of drug perturbation outcomes. Recently, self-supervised learning (SSL) on IF datasets followed by evaluation on in-distribution (ID) and out-of-distribution (OOD) tasks within or across IF datasets has become the standard paradigm for assessing vision foundation models in this domain [1–5].

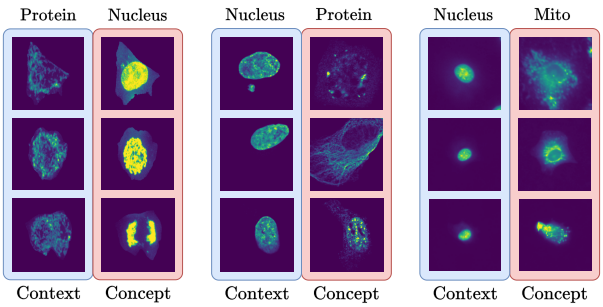

Figure 1. Intrinsic separation of channels in IF datasets (left: WTC-11, middle: HPA, right: CP). Context channels (blue) provide stable structural reference, while concept channels (red) carry dataset-specific semantic variation. Channel assignments may vary by dataset. As shown here, Nucleus is a concept channel in WTC-11 (as the primary focus is on cell-cycle stage), but a context channel in HPA (as the focus is on protein localization). Each row shows an example from a distinct training label from the CHAMMI benchmark dataset.

However, learning strong image representations via SSL that are transferable across IF datasets remains challenging. Different staining protocols produce different channel configurations across datasets, while typical image encoders assume a fixed number of channels. This makes training and evaluating models across multiple channel configurations non-trivial. Prior studies address multi-channel training across datasets [6–8], but provide no mechanism for OOD evaluation on unseen channel configurations, limiting their evaluations to configurations observed during training.

To address this, we identify an intrinsic separation of channels in IF datasets that reflects laboratory-specific experimental design choices [9–11]. We find that certain 'context' channels provide a structural reference of the cell, and are visually consistent across images in a dataset. In contrast,

'concept' channels, capture experiment-specific semantic information, and are meant to be interpreted in relation to the context channels (see Fig. 1). This assignment of context and concept channels is dataset-dependent, and is a result of the experimental protocols used by experts when interpreting IF images.

Therefore, IF images can be viewed as two groups of channels, with one group carrying contextual information and carrying concept. We refer to this as the 'context-concept principle' in IF channels, and take a step towards leveraging it to learn high-quality, transferable representations across IF datasets. To this end, we propose **Channel-Conditioned-Cell Representations (C3R)**, a two-fold framework that combines (i) an architectural design with strong representational capacity and support for OOD evaluation, and (ii) a pre-training method that further strengthens IF image representations.

We begin from the observation that context and concept channel groups carry semantically distinct information. Based on this, we propose the Context-Concept Encoder (CCE), a network architecture that acts as an information handler for distinct context and concept groups. The CCE model learns separate group wise intermediate features to capture within-group structure, and then integrates them to learn a unified final representation. During inference, unobserved channels are manually assigned to either group and processed without additional training. We find that the group-wise information handling in CCE yields strong representations over baseline methods, while the group assignment using CCE makes systematic evaluation possible on unseen channel configurations.

Next, we further leverage the structural coherence of context channels and the dependence of concept channels on context. We propose Masked Context Distillation (MCD), a momentum-based pre-training strategy that enhances representations for IF images. MCD introduces a distillation loss that regulates how much context the concept channels access when forming the overall representation. This encourages the model to encode relevant conceptual information from concept channels to contribute effectively to the overall representation even with limited context, yielding stronger, concept-driven IF image representations.

Together, CCE and MCD form the C3R framework, enabling both systematic OOD evaluation and improved representation learning across diverse IF datasets. We evaluate C3R on downstream tasks on HPA (ID, with same channel configurations) [9] and JUMP-CP (OOD, with novel channel configurations) [12] datasets, where C3R outperforms existing ID and OOD baselines. We also evaluate C3R on the CHAMMI benchmark on a frozen encoder evaluation setting [6]. In addition to the original CHAMMI benchmark that evaluates generalization to novel semantic differences within IF datasets, our implementation evaluates generalization to unobserved channels across IF datasets. C3R achieves state-of-the-art performance on CHAMMI-FE, surpassing existing channel-adaptive and channel-agnostic methods.

2. Related work

Representation learning for IF images. Representation learning using SSL for cell images ranges from fully unsupervised to weakly supervised approaches. While some focus on scaling [13–16], we focus on architectural and SSL design choices. In fully unsupervised methods, discriminative approaches such as SimCLR, DINO, and iBOT [17–19] have been shown to outperform reconstruction-based methods [20, 21] in IF datasets. As such, DINO4Cells [3] shows that DINO learns unbiased morphology features, learning useful cell structure without labels. In addition to fully unsupervised SSL, weak supervision introduces biological signals to further improve representations. Examples of these signals include protein identity, screen-derived labels, or RNA-seq data [22–24]. SubCell [4] uses a supervised contrastive loss with antibody-stained cells, achieving strong results on ID and OOD datasets. However, they re-train models to match OOD datasets, which limits direct transfer to unobserved channels. We use DINO4Cells [3] and SubCell [4] as benchmarks, since they are trained and tested on HPA [9] and JUMP-CP [12], but require retraining to handle OOD channels. In contrast, our models are evaluated without retraining.

The context-concept principle in IF images. IF datasets consistently separate channels into those that provide structural context and those that capture task-specific concepts. In the HPA dataset, each sample includes the protein of interest alongside three reference markers (Nucleus, Microtubules, and Endoplasmic Reticulum (ER)) which serve as structural landmarks [9]. Similarly, in JUMP-CP, the DNA (Nucleus) and ER channels act as positional references while the remaining channels (RNA, AGP, Mitochondria) capture perturbation-induced morphological phenotypes [12]. In WTC-11, the Nucleus provides the primary concept, while and Membrane and Protein channels contribute auxiliary information (context) [11]. In CM4AI-Bridge2AI [25], Proteins of interest are labelled with specific antibodies, while cells are co-stained with DAPI (Nuclei), anti-tubulin (Microtubules), and ER as structural/positional markers. These examples show that certain channels act as stable context for cellular localization, while others provide variable, concept-driven information relevant to the biological task. In summary, the context-concept principle reflects how IF experiments are commonly designed and interpreted: (i) landmark/structural (context) channels that provide spatial reference for segmentation and registration; and (ii) readout (concept) channels whose signal is expected to vary with

the biological factor of interest (antibody target, compound/-MoA, genotype, pathogen).

Vision transformers for multi-channel imaging. Handling multi-channel data has previously been studied outside biology, for example in climate modelling [26, 27]. For IF datasets, the CHAMMI benchmark [6] evaluates cross-dataset multi-channel training with over 220,000 single-cell images from three sources. However, the CHAMMI benchmark does not evaluate generalization to unobserved channels. Architectures such as ChAdaViT [28], ChannelViT [7], and DiChaViT [8] tokenize channels separately, assign channel-specific parameters and use sampling or regularization to promote channel and token diversity. However, these methods are tied to Channels observed during training. Similar to ChannelViT, CellRep [16] dynamically generates channel-specific tokens via convolutional layers, but is pre-trained on cell painting assays. SubCell [4] shows that models pre-trained on feature rich datasets like HPA perform similarly or better on cell painting downstream tasks compared to models pre-trained on cellpainting assays. Therefore, we use ChannelViT and DiChaViT as our channel-token specific architectures and pre-train them on HPA. Both ChannelViT and DiChaViT scale poorly with the number of channels due to long token sequence lengths. As an alternative, single-channel approaches [14, 20, 21, 29] process each channel independently during pre-training. Although these models can be directly evaluated on unseen datasets without further training, they fail to capture inter-channel dependencies. Nonetheless, we include a single-channel approach in our evaluation for completeness.

3. Methodology

In this section, we introduce C3R, our proposed method for learning cell-level features conditioned on the biologically informed context-concept principle in IF image channels. C3R includes two main components: the Context-Concept Encoder (CCE) and the Masked Context Distillation (MCD) training strategy. The two components, when combined, encourages distinct feature learning for each context and concept group, while encouraging the context channels to act as a reference for the concept. We implement C3R using a Vision Transformer (ViT) backbone and use iBOT with the SubCell antibody loss as our base pre-training pipeline [4, 19, 30]. The C3R pre-training code is available at https://anonymous.4open.science/r/C3R-5015.

To describe CCE and MCD, we first define the input x as a multi-channel IF image, viewed as two groups of images, where C_1, C_2 are number of channels in the context and concept groups, and (h, w) are the spatial dimensions. The group assignments for each dataset can be found in the supplementary material.

$$x = [x_{c1},\, x_{c2}]\,, x_{c1} \in \mathbb{R}^{C_1 \times h \times w}, x_{c2} \in \mathbb{R}^{C_2 \times h \times w}. \quad (1)$$

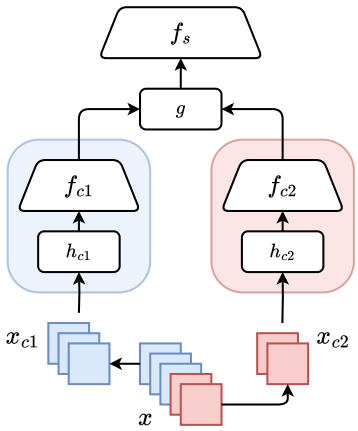

Figure 2. Context-Concept Encoder: The input channels are separated into context and concept, where each group is processed independently through their respective h_c and f_c layers. The two group-wise representations are then combined and passed through joint encoder layers f_s.

3.1. Context-Concept Encoder (CCE)

We design CCE to first construct separate intermediate representations for context and concept channel groups, then combine them to model inter-group dependencies and form a global representation. Leveraging the inherent distinction between these groups, CCE encodes the context and concept groups independently up to a certain depth. Since the context-concept distinction is consistent across IF datasets, the model serves as an information handler of these groups in an OOD setting with novel channels.

The CCE architecture comprises group-wise convolutional stems and lightweight encoder layers for processing context and concept-specific inputs (h_{c1}, h_{c2} for stems; f_{c1}, f_{c2} for encoders), a combiner function g for merging the group-wise representations, and joint encoder layers f_s for further joint processing, as illustrated in Fig. 2.

A high-level overview is provided here, while full architectural specifications of all components (h_c, f_c, g, f_s) are detailed in the supplementary material.

Given the input $x = [x_{c1}, x_{c2}]$, channels in x_{c1} and x_{c2} are tokenized by group-specific stems:

$$\tilde{x}^i_{c1} = h_{c1}(x^i_{c1}) \in \mathbb{R}^{N \times d}, \quad i = 1, \ldots, C_1, \quad (2)$$

$$\tilde{x}^j_{c2} = h_{c2}(x^j_{c2}) \in \mathbb{R}^{N \times d}, \quad j = 1, \ldots, C_2, \quad (3)$$

where d is the output dimensionality of h_{c1} and h_{c2}, and N is the number of tokens. Each tokenized channel

is then encoded independently via group-specific branched lightweight encoders:

$$\hat{x}^i_{c1} = f_{c1}(\tilde{x}^i_{c1}) \in \mathbb{R}^{N\times d}, \quad i = 1, \ldots, C_1, \tag{4}$$

$$\hat{x}^j_{c2} = f_{c2}(\tilde{x}^j_{c2}) \in \mathbb{R}^{N\times d}, \quad j = 1, \ldots, C_2. \tag{5}$$

Unlike ChannelViT and its variants [7, 8, 28], the individual channels in context and concept groups are independently passed through the encoders f_{c1} and f_{c2}. This results in linear computational complexity with channel count, as opposed to quadratic complexity in previous methods.

The intermediate group-wise encoded outputs are then combined by a function g (see supplementary material for implementation details), before passing through joint encoder layers to obtain the final output y,

$$\hat{x} = g\left(\{\hat{x}^i_{c1}\}_{i=1}^{C_1}, \{\hat{x}^j_{c2}\}_{j=1}^{C_2}\right) \in \mathbb{R}^{N\times D}, \tag{6}$$

$$y = f_s\left(\hat{x}\right) \in \mathbb{R}^{D}. \tag{7}$$

Typically, $D = 2d$, where D is the embedding dimension of the baseline encoder, and the layer depths of f_c and f_s are set to match the overall parameter count of baselines. This ensures fair comparison by parameter and FLOP count versus comparative methods. However, in contrast to the comparative methods, the distinction of groups learned by CCE to yield the final representation y is directly transferable to other datasets with unobserved channels.

3.2. Masked context distillation (MCD)

We propose MCD as a pre-training strategy that governs the interaction between context and concept channels during training. In the CCE architecture, the context channels appear in its complete form as a reference to the concept channels, at the point of the aggregator $g(\cdot)$. Based on the structural coherence of context channels across samples as observed in Fig. 1, using MCD, we encourage the model to use limited intermediate context representations to act as a reference to the intermediate concept representations, building robust global representations.

The core pretraining pipeline of our work is based on iBOT [19], which contains masked patch prediction and student-teacher distillation. However, unlike existing distillation methods in DINO and iBOT, the student signal in MCD differs from the teacher signal not only by cropping and augmentation, but also through a novel context channel sampling strategy that allows for conditioning of the concept channel through limited context. Fig. 3 shows the schematic of the distillation strategy in MCD.

We define the student and teacher networks as $\mathcal{S}$ and $\mathcal{T}$, and each network contains a projection head defined as p_s and p_t, respectively. Given the input $x = [x_{c1}, x_{c2}]$, we draw two augmented views $u = [u_{c1}, u_{c2}]$ and $v = [v_{c1}, v_{c2}]$.

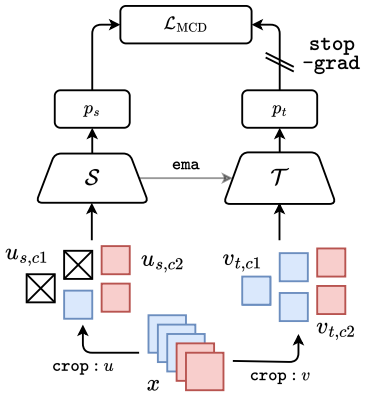

Figure 3. Masked Context Distillation: During training, the student encoder $\mathcal{S}$ randomly samples a subset of context channels prior to the forward pass, while the teacher encoder $\mathcal{T}$ passes the full set of context channels. The loss is computed between the context-masked student representation and the dense teacher representation.

Assuming crop u is passed to the student network and v to the teacher, we randomly sample without replacement c channels from the context group u_{c1}, where $1 \leq c \leq C_1$. Conversely, when u is passed to the teacher and v to the student, sampling is instead applied to the context group v_{c1}. Given student crop u and teacher crop v, the student and teacher inputs are then defined as,

$$u_{s,c1-\texttt{smp}} = \texttt{sample}(u_{c1}, c) \in \mathbb{R}^{c\times h\times w}, \tag{8}$$

$$u_s = [u_{c1-\texttt{smp}}, u_{c2}], \quad v_t = [v_{c1}, v_{c2}], \tag{9}$$

where $1 \leq c \leq C_1$. u_s and v_t are then passed through student and teacher encoders $\mathcal{S}$ and $\mathcal{T}$ to obtain feature representations, and then passed through respective projection heads p_s and p_t to obtain z_s and z_t. The Masked Context Distillation loss then is computed as the KL-divergence between them.

$$y_s = \mathcal{S}(u_s), \quad y_t = \mathcal{T}(v_t), \tag{10}$$

$$z_s = p_s(y_s), \quad z_t = p_t(y_t) \tag{11}$$

$$\mathcal{L}_{\text{MCD}} = \mathcal{L}_{\text{KL}}(z_t \,\|\, z_s) = \sum_{k=1}^{K} z_t(k)\, \log\frac{z_t(k)}{z_s(k)}. \tag{12}$$

where K is the number of classes in the softmax output space of z_s and z_t. The overall loss is formulated as,

$$\mathcal{L} = \mathcal{L}_{\text{MCD}} + \mathcal{L}_{\text{iBOT-MIM}} + \mathcal{L}_{\text{SubCell-WSL}}, \tag{13}$$

where $\mathcal{L}_{\text{iBOT-MIM}}$ is the iBOT masked image modelling loss [19], and $\mathcal{L}_{\text{SubCell-WSL}}$ is the weakly supervised Subcell antibody loss [4]. All comparative methods and CCE without MCD have been trained using the same losses, but with the original KL-Divergence loss in iBOT in place of the MCD loss.

Table 1. Overview of downstream tasks and their data distributions. "OOD w.r.t channels" implies that the adaptation/evaluation is carried out on a dataset different to the source dataset (HPA), hence a different channel configuration in the evaluation dataset. "OOD w.r.t task" implies that within the evaluation procedure, the training (i.e.: linear classifier or proxy labelling, not direct retrieval), is performed on a different subset that is OOD by a different metric (such as cell line, or plate profile) other than channel configuration, within the main evaluation dataset.

Evaluation	**Feature adaptation**	**Distribution**	
		w.r.t channels	**w.r.t task**
HPA-Loc	3-layer MLP for multi-label classification	ID: HPA	OOD: Novel antibodies
JUMP-Ret	Direct retrieval on fixed normalized features	OOD: JUMP-CP	OOD: No training data
JUMP-Trex	4-layer MLP for compound and MoA prediction	OOD: JUMP-CP	OOD: Novel plates
CHAMMI-FE	2-layer MLP. All ID tasks are trained jointly with learnable proxies. OOD tasks are solved via retrieval with the learned proxies following the original CHAMMI benchmark.	OOD: WTC-11	T1-ID: Known labels T2-OOD: Novel labels
		ID: HPA	T1-ID: Known cell lines T2-OOD: Unseen cell lines T3-OOD: Novel labels
		OOD: JUMP-CP	T1-ID: Known plates T2-OOD: Novel plates T3-OOD: New CP dataset T4-OOD: Novel labels

The pseudocode for MCD can be found in the supplementary material. By the separate-and-joint processing of context and concept channels through CCE, and the conditioned learning of the context-concept relationship through MCD, our method yields strong image representations and generalize well to unseen IF datasets.

4. Experiments

4.1. Implementation details

Pre-training and evaluation. All models (except SubCell and DINO4Cells for which we use their pre-trained checkpoints) are pre-trained on the HPA training set ($\approx 800k$ images) using iBOT [19] with the SubCell antibody loss [4]. Adding to comparative methods, we train three baselines: **Base-HPA** (4-channel HPA), **Base-CP** (3-channel HPA, excluding Microtubules), and **Base-SC** (single-channel HPA, one channel per forward pass). ID evaluation (**HPA-Loc**) is performed on the HPA test set for 19,31-class protein localization, reported in mAP.

Channel-wise OOD evaluation (cross-dataset evaluation with fixed, precomputed features on unseen channels) is conducted on JUMP-CP and the CHAMMI benchmark dataset (containing images from WTC-11, HPA, and JUMP). On JUMP-CP, we test two settings with fixed features: (i) **JUMP-Ret**: zero-shot compound replicate retrieval (matching drug perturbations between cells); and (ii) **JUMP-TRex**: task-specific feature adaptation, following TRex [31], where linear classifiers are trained with task-specific losses on fixed IF features for compound replicate matching and mechanism-of-action (MoA) prediction (a task that yields random performance in zero-shot). Comparative method SubCell [4] re-trains a 3-channel model on HPA to match the JUMP-CP dataset, and replicates the protein channel thrice to match the 5 channels in JUMP. DINO4Cells [3] trains the model on a 5-channel cell painting dataset to evaluate on JUMP-CP.

On the CHAMMI benchmark dataset, we introduce a frozen encoder evaluation (**CHAMMI-FE**) of the original benchmark designed to test generalization to unseen channels. As in the original benchmark, joint training is performed on three IF datasets and evaluated across nine tasks. However, CHAMMI-FE trains a 2-layer MLP on fixed image representations. This prevents encoder or stem network fine-tuning like in CHAMMI, thereby preventing the HPA-pretrained encoder from adapting weights to new channel configurations, and as a result emulating a setting where encoders are exposed to unseen channel configurations.

We summarize the downstream tasks in Tab. 1. Each task is defined by the feature adaptation strategy, the dataset used, and the distribution shift with respect to channels or tasks.

Comparisons. We pre-train C3R with ViT-S and ViT-B (with adjusted layer depths) and compare against Base-HPA, Base-CP, Base-SC, DINO4Cells (D4C) [3], SubCell [4], ChannelViT [7], and DiChaViT [8]. We use official repositories and model checkpoints when available and otherwise re-implement methods under our setting. We set the depths of f_c and f_s to be 2 and 11 for C3R, and the choice of depths is investigated in **??**. Further implementation details of all methods can be found in the supplementary material.

4.2. Benchmark results

HPA-Loc and JUMP-Ret. Tab. 2 shows the results on the HPA protein localization and JUMP-CP zero-shot compound retrieval benchmarks. From Tab. 2, we find that C3R outperforms all methods except Base-CP with ViT-S, which is pre-trained on HPA with a channel configuration matched to JUMP-CP. Without any target-specific pre-training, C3R achieves nearly the same performance as Base-CP in this setting, and surpasses all other methods across both ViT-S and ViT-B.

Table 2. Results on HPA-Loc and JUMP-Ret (OOD by channels). *: re-trained from scratch to match the JUMP-CP channel configuration. //: unsuitable for evaluation due to channel mismatch.

Enc.	Method	HPA-Loc		JUMP-Ret	
		mAP-31	mAP-19	mAP	kNN
ViT-S	Base-HPA	0.505	0.686	//	//
	Base-CP*	//	//	**0.355**	0.507
	Base-SC	0.380	0.528	0.327	0.457
	ChannelViT	0.438	0.602	0.345	0.503
	DiChaViT	0.429	0.590	0.343	0.494
	C3R	**0.536**	**0.722**	0.354	**0.518**
ViT-B	Base-HPA	0.515	0.698	//	//
	Base-CP*	//	//	0.355	0.513
	Base-SC	0.385	0.528	0.339	0.473
	D4C*	0.508	0.683	0.339	0.509
	SubCell*	0.519	0.695	0.350	0.514
	C3R	**0.548**	**0.737**	**0.363**	**0.530**

CHAMMI-FE. Tab. 3 summarizes the CHAMMI-FE benchmarks. The results for CHAMMI-FE containing all tasks can be found in the supplementary material. From Tab. 3, we observe significant improvements on CHAMMI-FE and HPA-Loc, while on JUMP-Ret, C3R matches or exceeds Base-CP and substantially outperforms the remaining methods.

Table 3. Results on CHAMMI-FE. We report the ID and OOD average scores, and overall CHAMMI Performance Score (CPS).

Encoder	Method	ID	OOD	CPS
ViT-S	Base-SC	0.812	0.427	0.474
	ChannelViT	0.852	0.423	0.472
	DiChaViT	0.812	0.410	0.459
	C3R	**0.861**	**0.485**	**0.543**
ViT-B	Base-SC	0.797	0.423	0.459
	C3R	**0.873**	**0.468**	**0.522**

JUMP-TRex. Tab. 4 presents results on JUMP-CP using TRex with ViT-B trained on A549 and U2OS cell lines separately, and then on both cell lines jointly. C3R outperforms all methods on all experiments on compound matching, while it outperforms all methods on 2/3 experiments on MoA identification. It is worth noting that Base-CP, and SubCell are pre-trained to match the target channel configuration, and DINO4Cells is pre-trained on a cell painting assay. C3R on the other hand is pre-trained on the original 4-channel HPA dataset. While zero-shot retrieval capabilities (i.e.: JUMP-Ret) are valuable for broad applicability without retraining, C3R's advantages also extend to scenarios where features are explicitly adapted to subsets of the target dataset via probing.

Table 4. Results on JUMP-TRex, trained on cell lines A549 and U2OS on ViT-B. *: re-trained to match target channel configuration.

Method	A549		U2OS		A549+U2OS	
	Cpd	MoA	Cpd	MoA	Cpd	MoA
Base-CP*	0.427	0.306	0.335	0.276	0.321	0.281
Base-SC	0.381	0.234	0.276	0.211	0.268	0.231
D4C*	0.439	**0.348**	0.314	0.287	0.310	0.329
SubCell*	0.424	0.291	0.334	0.260	0.323	0.267
C3R	**0.444**	0.346	**0.345**	**0.298**	**0.325**	**0.341**

4.3. Analysis

The contribution of each component of C3R. Tab. 5 shows the incremental contribution of each component in C3R over the baseline ViT. We first replace the standard convolutional stem of the ViT with per-group stems h_{c1}, h_{c2}. Here, the combiner g from Sec. 3 is applied directly on the outputs of h_c. We observe that h_{c1}, h_{c2} alone improves ID performance on HPA. However, when applied to JUMP-CP, it outperforms Base-SC, ChannelViT, and DiChaViT (methods with workarounds for unseen channels, results shown in Tab. 2), but falls short of our OOD baseline set by re-training an encoder to match JUMP-CP's channel configuration (Base-CP). We suspect that while per-group convolutional stems aid low-level feature extraction, they fail to learn a context-concept distinction, limiting transfer to the OOD dataset. Introducing f_{c1}, f_{c2} before g, thereby processing input groups independently for several layers, significantly improves OOD performance, matching or surpassing Base-CP, suggesting this distinction is learned in these branched encoder layers. We experimentally validate this later in this section.

Applying masked context distillation (MCD), where some context channels are dropped during training and the student network benefits from the teacher's full context channel representation, yields significant improvements in ID performance. However, we find that the ID performance

Table 5. The effect of each component in C3R. *: pre-trained to match the JUMP-CP channel configuration. h_c: grouped stems, f_c: branched encoders, 'MCD': context distillation. //: unsuitable for evaluation due to channel mismatch.

Enc.	Method	HPA-Loc		JUMP-Ret	
		mAP-31	mAP-19	mAP	kNN
ViT-S	Base-HPA	0.505	0.686	//	//
	Base-CP*	//	//	0.355	0.507
	Base-SC	0.380	0.528	0.327	0.457
	h_c	0.523	0.700	0.347	0.510
	$+f_c$	0.520	0.705	0.351	**0.529**
	+MCD	**0.535**	**0.725**	**0.354**	0.518
ViT-B	Base-HPA	0.515	0.698	//	//
	Base-CP*	//	//	0.355	0.513
	Base-SC	0.385	0.528	0.339	0.473
	h_c	0.529	0.710	0.344	0.508
	$+f_c$	0.531	0.716	0.358	**0.532**
	+MCD	**0.548**	**0.737**	**0.363**	0.530

Table 6. Effect of MCD on JUMP-TRex and CHAMMI-FE.

(a) JUMP-TRex

MCD	A549		U2OS		A549+U2OS	
	Cpd	MoA	Cpd	MoA	Cpd	MoA
✗	0.441	**0.348**	0.338	0.291	0.323	0.335
✓	**0.444**	0.346	**0.345**	**0.298**	**0.325**	**0.341**

(b) CHAMMI-FE.

MCD	ID	OOD	CPS
✗	0.634	0.379	0.338
✓	**0.861**	**0.484**	**0.543**

boost obtained by MCD does not translate to the JUMP-Ret task. Nevertheless, we find improvements with MCD in both CHAMMI-FE and JUMP-TRex (see Tab. 6). We suspect that some learnable adaptation is necessary to extract strong representations from these fixed features.

Effect of MCD on CHAMMI-FE and JUMP-TRex. Tab. 6 shows the effect of MCD on JUMP-CP evaluation using TRex pipeline, where we observe a marginal improvement with MCD on 5 out of 6 tasks. Moreover, Tab. 6 shows a significant improvement with MCD on the CHAMMI-FE benchmark.

Overall, including the results from Tab. 5, we find that MCD improves performance on 3 out of 4 evaluations (HPA-ID, CHAMMI-FE, JUMP-CP TRex) while matching the no MCD variant on a single evaluation (JUMP-CP retrieval). However, it should be noted that all evaluations where MCD performs better contain some learnable component such as MLPs, while the JUMP-CP retrieval task does not. We aim to investigate the performance of pre-training strategies on pure zero-shot tasks without any dataset-specific adaptation in future work.

Experimental validation of group-specific learning. To further validate the hypothesis that the branched encoders $fc1$ and f_{c2} encode distinct context-concept information, we deliberately flip group assignments during OOD evaluation. A significant drop in performance would suggest that f_{c1}, f_{c2} learn group-specific distinct information which is passed on to the OOD dataset. Conversely, if the performance remains similar, it would indicate that f_{c1}, f_{c2} are group-agnostic, implying that the performance increase is simply a result of channel separation during training. The results in Fig. 4 support the hypothesis, as flipping consistently reduces performance. Interestingly, the performance drop grows with deeper f_{c1} and f_{c2} layers, indicating that separating the branches over more layers strengthens the distinction between groups.

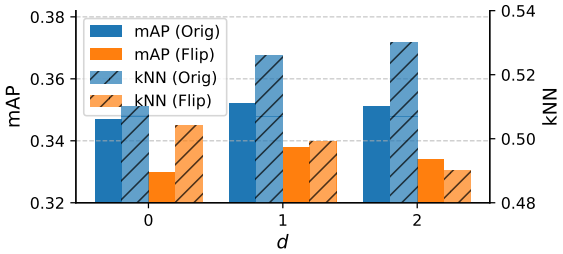

Figure 4. Effects of group switching assignments on JUMP-Ret. d: layers per each f_c. When $d = 0$, groups are combined directly after h_c stems.

We also validate the context-concept principle on the CHAMMI-FE benchmark, by switching the channel assignments for the WTC-11 dataset. In the true setting, the Nucleus is passed as a concept channel, and Protein + Membrane as context. This is due to cell cycle stage being the main downstream focus of the WTC dataset. In the flipped setting, Nucleus acts as context while Protein + Membrane act as concept. We find that in the original setting, the WTC-11 average F1 score is 0.551 while when flipped, it reduces to 0.536. We find this to be an interesting result, as C3R is pre-trained on HPA, and there the Nucleus acts as context and Protein acts as concept. This means the concept branch has never witnessed a Nucleus channel during pre-training. This provides more confidence that the concept branch indeed encodes conceptual information from channels.

Modes of channel sampling for MCD. Our hypothesis for MCD lies in encouraging the concept channels of the model to contribute to the overall image representation with

Table 7. Choice of channel dropping for MCD. Using ViT-S. $\mathcal{S}, \mathcal{T}$: student, teacher.

Drop	HPA-Loc		JUMP-Ret	
	mAP-31	mAP-19	mAP	kNN
ChannelViT	0.438	0.602	0.345	0.503
DiChaViT	0.429	0.590	0.343	0.494
None	0.519	0.702	0.351	**0.529**
$\mathcal{S}$ only	**0.536**	0.722	**0.354**	0.518
$\mathcal{S}+\mathcal{T}$	0.533	**0.724**	0.347	0.504

limited context. To validate this, we run experiments to identify the networks ($\mathcal{S}$ or $\mathcal{T}$) on which the channels need to be dropped in order to yield better representations. From Tab. 7, we observe that ensuring the concept channels are preserved during the forward pass results in better performance (None, $\mathcal{S}$ only, $\mathcal{S}+\mathcal{T}$), compared to the context-concept agnostic sampling strategies such as ChannelViT and DiChaViT. We also find that over no sampling at all, sampling the context under either setting ($\mathcal{S}$ only or $\mathcal{S}+\mathcal{T}$) yields better ID performance.

This observation aligns with our motivation of using MCD, where a masked context generally outperforms a non-masked context channel set. However, we observe a drop in average performance when the teacher context channels are masked ($\mathcal{S}+\mathcal{T}$). We attribute this to the limited learning capability caused by the weak teacher signal, as seen in existing SSL methods where masked teachers such as SdAE [32] under perform in linear evaluation tasks in comparison to methods with full teacher representations.

Effect of context channel sampling rates for MCD. We explore the optimal sampling rate for the context channels when training with MCD. Specifically, we run fixed sampling rates where we keep $c=1$ and $c=2$ channels, and a variable sampling rate where we randomly keep $c \in \{1,2,3\}$ context channels, sampled uniformly at random. The total number of context channels in the HPA dataset is 3 (Microtubules, ER and Nucleus).

Table 8. Effect of context channel sampling rates in during MCD training. All experiments are performed on ViT-S.

c	HPA-Loc		JUMP-Ret	
	mAP-31	mAP-19	mAP	kNN
3 (all)	0.516	0.702	0.351	**0.529**
2	0.531	0.708	**0.356**	0.521
1	0.513	0.700	0.336	0.499
$\{1,2,3\}$	**0.536**	**0.722**	0.354	0.518

Tab. 8 shows the effect of context channel sampling rates, where we find that a random sampling rate $c=\{1,2,3\}$ yields better overall performance vs. fixed sampling. Consistently sparse sampling at $c=1$ exhibits the lowest performance, possibly because the distillation task becomes too difficult. In contrast, dropping only a single channel at $c=2$ yields similar OOD metrics but lower ID metrics. This result, along with the higher metrics obtained with variable sampling rates, suggests that the student $\mathcal{S}$ can still benefit from sparse context with $c=1$, provided it receives enough support from other training iterations where more channels are preserved (e.g., $c=2$ or $c=3$).

5. Broader Impacts and limitations

Broader Impact. IF imaging plays a critical role in clinical diagnostics and biomedical research, providing insights into cellular morphology, protein localization, and disease progression at the subcellular level. However, the integration of deep learning methods towards these diagnostics has been limited by the lack of generalizable models that can operate reliably across heterogeneous datasets, where retraining models for every lab, institution, or imaging protocol is impractical and costly. Without dataset-specific adaptation or re-training, C3R has the potential to accelerate the integration of deep learning into diagnostic pipelines, reducing the time and cost of biomarker discovery, drug response prediction, and personalized treatment planning.

Limitations. Based on the context-concept principle, the ability to perform OOD evaluation on a target dataset depends on the principle being valid for the OOD dataset i.e: we find a natural separation of channels into context and concept in the target dataset. In general IF datasets (HPA [9], JUMP [12], WTC-11 [11], OpenCell [33], Bridge2AI [25]) we find this assumption to be true, as the datasets have been created with the intention of detecting variations in channel intensities that then correspond to specific phenotypes. Due to how the imaging experiments are carried out, IF datasets carry the principle of context and concept. However, C3R at its current state cannot be applied to IF datasets that may not follow this principle.

Also, like SubCell, DINO4Cells, ChannelViT and DiChaViT, we train and evaluate CCE on ViTs. We aim to adapt CCE towards building a general architectural framework beyond ViTs.

6. Conclusion

We introduce C3R, a two-fold architectural and pre-training framework that builds strong IF image representations that transfer well to datasets with unobserved channel configurations. We build C3R based on the context-concept principle of IF images, which we validate experimentally and show that this principle can be modeled and transferred across IF datasets. We show that C3R significantly outperforms

existing methods in both ID and OOD tasks, and matches dataset-targeted OOD training and evaluation strategies without any re-training. Overall, this work offers a new perspective on leveraging this principle in IF datasets and opens a pathway for cross-dataset generalization without requiring dataset-specific adaptation or retraining at the image-level.

References

[1] Mark-Anthony Bray, Shantanu Singh, Han Han, Chadwick T Davis, Blake Borgeson, Cathy Hartland, Maria Kost-Alimova, Sigrun M Gustafsdottir, Christopher C Gibson, and Anne E Carpenter. Cell painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nature protocols*, 11(9):1757–1774, 2016. 1

[2] Juan C Caicedo, John Arevalo, Federica Piccioni, Mark-Anthony Bray, Cathy L Hartland, Xiaoyun Wu, Angela N Brooks, Alice H Berger, Jesse S Boehm, Anne E Carpenter, et al. Cell painting predicts impact of lung cancer variants. *Molecular biology of the cell*, 33, 2022.

[3] Michael Doron, Théo Moutakanni, Zitong S Chen, Nikita Moshkov, Mathilde Caron, Hugo Touvron, Piotr Bojanowski, Wolfgang M Pernice, and Juan C Caicedo. Unbiased single-cell morphology with self-supervised vision transformers. *bioRxiv*, 2023. 2, 5

[4] Ankit Gupta, Zoe Wefers, Konstantin Kahnert, Jan N Hansen, Will Leineweber, Anthony Cesnik, Dan Lu, Ulrika Axelsson, Frederic Ballllosera Navarro, Theofanis Karaletsos, et al. Subcell: Vision foundation models for microscopy capture single-cell biology. *bioRxiv*, pages 2024–12, 2024. 2, 3, 4, 5

[5] Vladislav Kim, Nikolaos Adaloglou, Marc Osterland, Flavio M Morelli, Marah Halawa, Tim König, David Gnutt, and Paula A Marin Zapata. Self-supervision advances morphological profiling by unlocking powerful image representations. *Scientific Reports*, 15(1):4876, 2025. 1

[6] Zitong Sam Chen, Chau Pham, Siqi Wang, Michael Doron, Nikita Moshkov, Bryan Plummer, and Juan C Caicedo. Chammi: A benchmark for channel-adaptive models in microscopy imaging. *Advances in Neural Information Processing Systems*, 36:19700–19713, 2023. 1, 2, 3

[7] Yujia Bao, Srinivasan Sivanandan, and Theofanis Karaletsos. Channel vision transformers: An image is worth 1 x 16 x 16 words. In *The Twelfth International Conference on Learning Representations*, 2024. 3, 4, 5

[8] Chau Pham and Bryan Plummer. Enhancing feature diversity boosts channel-adaptive vision transformers. *Advances in Neural Information Processing Systems*, 37:89782–89805, 2024. 1, 3, 4, 5

[9] Peter J Thul, Lovisa Åkesson, Mikaela Wiking, Diana Mahdessian, Aikaterini Geladaki, Hammou Ait Blal, Tove Alm, Anna Asplund, Lars Björk, Lisa M Breckels, et al. A subcellular map of the human proteome. *Science*, 356(6340): eaal3321, 2017. 1, 2, 8

[10] Srinivas Niranj Chandrasekaran, Beth A Cimini, Amy Goodale, Lisa Miller, Maria Kost-Alimova, Nasim Jamali, John G Doench, Briana Fritchman, Adam Skepner, Michelle Melanson, et al. Three million images and morphological profiles of cells treated with matched chemical and genetic perturbations. *Nature Methods*, 21(6):1114–1121, 2024.

[11] Matheus P Viana, Jianxu Chen, Theo A Knijnenburg, Ritvik Vasan, Calysta Yan, Joy E Arakaki, Matte Bailey, Ben Berry, Antoine Borensztejn, Eva M Brown, et al. Integrated intracellular organization and its variations in human ips cells. *Nature*, 613(7943):345–354, 2023. 1, 2, 8

[12] Srinivas Niranj Chandrasekaran, Jeanelle Ackerman, Eric Alix, D Michael Ando, John Arevalo, Melissa Bennion, Nicolas Boisseau, Adriana Borowa, Justin D Boyd, Laurent Brino, et al. Jump cell painting dataset: morphological impact of 136,000 chemical and genetic perturbations. *BioRxiv*, pages 2023–03, 2023. 2, 8

[13] Kian Kenyon-Dean, Zitong Jerry Wang, John Urbanik, Konstantin Donhauser, Jason Hartford, Saber Saberian, Nil Sahin, Ihab Bendidi, Safiye Celik, Marta Fay, et al. Vitally consistent: Scaling biological representation learning for cell microscopy. *arXiv preprint arXiv:2411.02572*, 2024. 2

[14] Alice V. De Lorenci, Seung Eun Yi, Théo Moutakanni, Piotr Bojanowski, camille couprie, Juan C Caicedo, and Wolfgang Maximilian Anton Pernice. Scaling channel-invariant self-supervised learning, 2025. 3

[15] Oren Kraus, Kian Kenyon-Dean, Saber Saberian, Maryam Fallah, Peter McLean, Jess Leung, Vasudev Sharma, Ayla Khan, Jia Balakrishnan, Safiye 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*, pages 11757–11768, 2024.

[16] Lawrence Phillips. Cellrep: A multichannel image representation learning model. In *1st CVPR Workshop on Computer Vision For Drug Discovery (CVDD): Where are we and What is Beyond?*, 2025. URL https://openreview.net/forum?id=vis0FIYD12. 2, 3

[17] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020. 2

[18] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[19] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021. 2, 3, 4, 5

[20] Dejin Xun, Rui Wang, Xingcai Zhang, and Yi Wang. Microsnoop: A generalist tool for microscopy image representation. *The Innovation*, 5(1):100541, 2024. ISSN 2666-6758. doi: https://doi.org/10.1016/j.xinn.2023.100541. 2, 3

[21] Rashmi Sreeramachandra Murthy, Shobana V Stassen, Dickson MD Siu, Michelle CK Lo, Gwinky GK Yip, and Kevin K Tsia. Generalizable morphological profiling of cells by interpretable unsupervised learning. *bioRxiv*, pages 2024–09, 2024. 2, 3

[22] Hirofumi Kobayashi, Keith C Cheveralls, Manuel D Leonetti, and Loic A Royer. Self-supervised deep learning encodes high-resolution features of protein subcellular localization. *Nature methods*, 19(8):995–1003, 2022. 2

[23] Heming Yao, Phil Hanslovsky, Jan-Christian Huetter, Burkhard Hoeckendorf, and David Richmond. Weakly supervised set-consistency learning improves morphological profiling of single-cell images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6978–6987, 2024.

[24] Xiaohang Fu, Yingxin Lin, David M Lin, Daniel Mechtersheimer, Chuhan Wang, Farhan Ameen, Shila Ghazanfar, Ellis Patrick, Jinman Kim, and Jean YH Yang. Bidcell: Biologically-informed self-supervised learning for segmentation of subcellular spatial transcriptomics data. *Nature Communications*, 15(1):509, 2024. 2

[25] Timothy Clark, Jillian Mohan, Leah Schaffer, Kirsten Obernier, Sadnan Al Manir, Christopher P Churas, Amir Dailamy, Yesh Doctor, Antoine Forget, Jan Niklas Hansen, et al. Cell maps for artificial intelligence: Ai-ready maps of human cell architecture from disease-relevant cell lines. *bioRxiv*, 2024. 2, 8

[26] Pedro Herruzo, Aleksandra Gruca, Llorenç Lliso, Xavier Calbet, Pilar Rípodas, Sepp Hochreiter, Michael Kopp, and David P. Kreil. High-resolution multi-channel weather forecasting – first insights on transfer learning from the weather4cast competitions 2021. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 5750–5757, 2021. 3

[27] Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K Gupta, and Aditya Grover. Climax: A foundation model for weather and climate. *arXiv preprint arXiv:2301.10343*, 2023. 3

[28] Nicolas Bourriez, Ihab Bendidi, Ethan Cohen, Gabriel Watkinson, Maxime Sanchez, Guillaume Bollot, and Auguste 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*, pages 11556–11565, 2024. 3, 4

[29] Wenyi Lian, Joakim Lindblad, Patrick Micke, and Nataša Sladoje. Isolated channel vision transformers: From single-channel pretraining to multi-channel finetuning. *arXiv preprint arXiv:2503.09826*, 2025. 3

[30] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 3

[31] Syed Sameed Husain, Jan Bober, Amaia Irizar, and Miroslaw Bober. Bridging self-supervision and mechanism of action discovery in morphological profiling. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4278–4285, 2025. 5

[32] Yabo Chen, Yuchen Liu, Dongsheng Jiang, Xiaopeng Zhang, Wenrui Dai, Hongkai Xiong, and Qi Tian. Sdae: Self-distillated masked autoencoder. In *European conference on computer vision*, pages 108–124. Springer, 2022. 8

[33] Nathan H Cho, Keith C Cheveralls, Andreas-David Brunner, Kibeom Kim, André C Michaelis, Preethi Raghavan, Hirofumi Kobayashi, Laura Savy, Jason Y Li, Hera Canaj, et al. Opencell: Endogenous tagging for the cartography of human cellular organization. *Science*, 375(6585):eabi6983, 2022. 8