REPORT

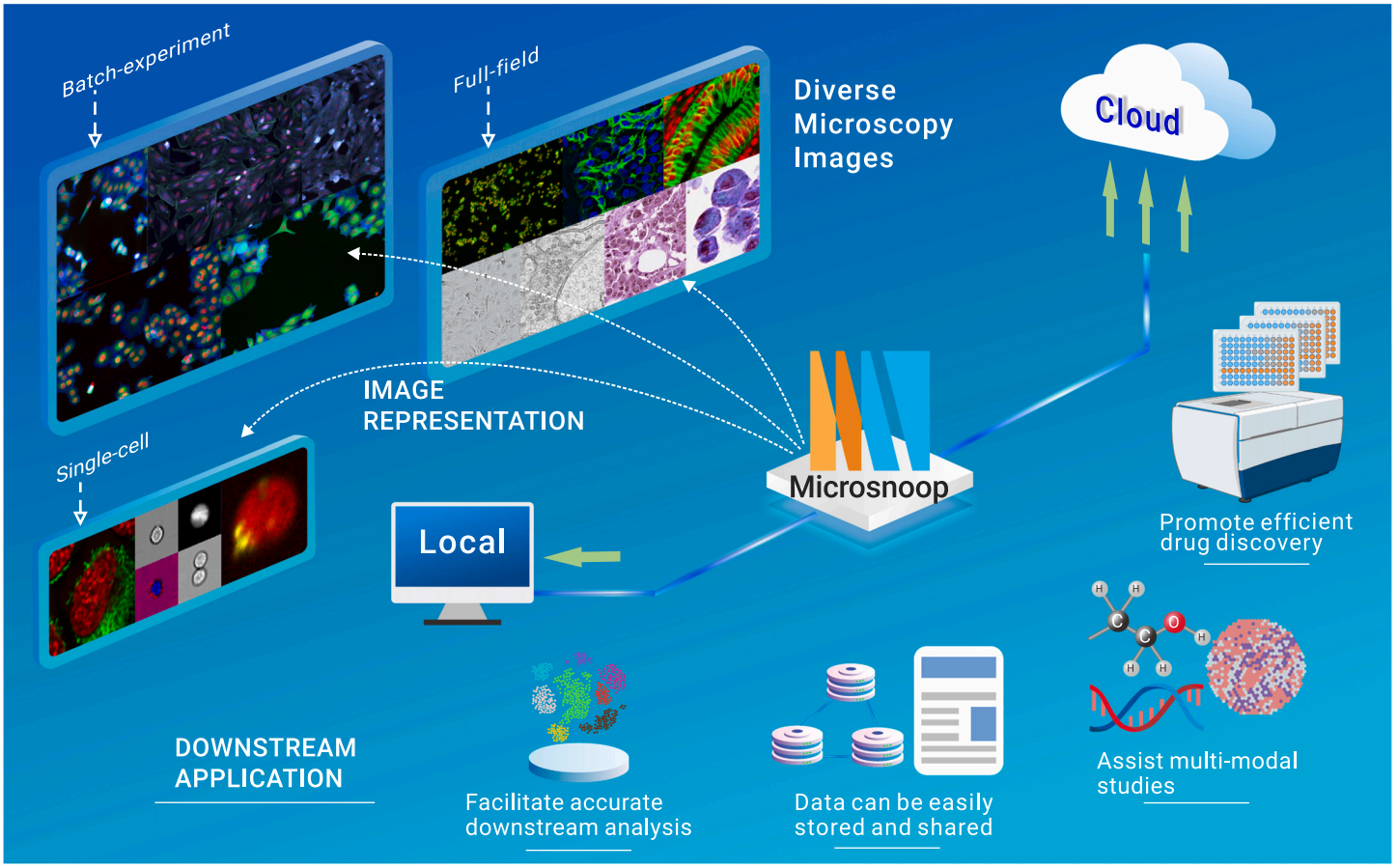
Microsnoop: A generalist tool for microscopy image representation

Dejin Xun,[1] Rui Wang,[2,*] Xingcai Zhang,[3,*] and Yi Wang[1,4,5,*]

*Correspondence: ruiwang@zju.edu.cn (R.W.); xingcai@seas.harvard.edu (X.Z.); zjuwangyi@zju.edu.cn (Y.W.)

Received: May 31, 2023; Accepted: November 17, 2023; Published Online: January 2, 2024; https://doi.org/10.1016/j.xinn.2023.100541

GRAPHICAL ABSTRACT



PUBLIC SUMMARY

- Microsnoop is a deep learning tool for profiling heterogeneous microscopy images.
- Microsnoop provides generalist pipelines for processing various types of images.
- Microsnoop achieves cutting-edge microscopy image representation ability with great potential for expansion.
- Microsnoop is highly scalable for studies from small scale to high throughput.

REPORT

Microsnoop: A generalist tool for microscopy image representation

Dejin Xun,[1] Rui Wang,[2,*] Xingcai Zhang,[3,*] and Yi Wang[1,4,5,*]

[1]Pharmaceutical Informatics Institute, College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, China
[2]State Key Lab of Computer-Aided Design & Computer Graphics, Zhejiang University, Hangzhou 310058, China
[3]John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA
[4]Innovation Institute for Artificial Intelligence in Medicine of Zhejiang University, Hangzhou 310018, China
[5]National Key Laboratory of Chinese Medicine Modernization, Innovation Center of Yangtze River Delta, Zhejiang University, Jiaxing 314100, China
*Correspondence: ruiwang@zju.edu.cn (R.W.); xingcai@seas.harvard.edu (X.Z.); zjuwangyi@zju.edu.cn (Y.W.)

Received: May 31, 2023; Accepted: November 17, 2023; Published Online: January 2, 2024; https://doi.org/10.1016/j.xinn.2023.100541

Citation: Xun D., Wang R., Zhang X., et al., (2024). Microsnoop: A generalist tool for microscopy image representation. The Innovation **5(1)**, 100541.

Accurate profiling of microscopy images from small scale to high throughput is an essential procedure in basic and applied biological research. Here, we present Microsnoop, a novel deep learning–based representation tool trained on large-scale microscopy images using masked self-supervised learning. Microsnoop can process various complex and heterogeneous images, and we classified images into three categories: single-cell, full-field, and batch-experiment images. Our benchmark study on 10 high-quality evaluation datasets, containing over 2,230,000 images, demonstrated Microsnoop's robust and state-of-the-art microscopy image representation ability, surpassing existing generalist and even several custom algorithms. Microsnoop can be integrated with other pipelines to perform tasks such as superresolution histopathology image and multimodal analysis. Furthermore, Microsnoop can be adapted to various hardware and can be easily deployed on local or cloud computing platforms. We will regularly retrain and reevaluate the model using community-contributed data to consistently improve Microsnoop.

INTRODUCTION

Automatic quantitative profiling of microscopy images has become increasingly ubiquitous in various aspects of biological research, from small-scale investigations to high-throughput experiments.[1] The analysis of visual phenotypes, which involves profiling intricate information according to images, is useful in diverse areas of biology,[2] including protein localization,[3] cell-cycle stage classification,[4] mechanisms of action prediction,[5] and high-content drug discovery.[6] In addition, the emergence of spatial omics has led to new requirements for the quantification of microscopy images. For example, spatial proteomics methods can image more than 50 disease-related proteins in a single tissue slice,[7] whereas spatial transcriptomics methods enable the simultaneous acquisition of image data and transcriptional profiles.[8] These developments demonstrate the need for a high-performance, generalist representation tool that can effectively handle heterogeneous microscopy images.

The traditional approach for profiling microscopy images involves extracting predefined morphological features, such as intensity, shape, texture, granularity, and colocalization.[9,10] However, these methods have several limitations, including low computational efficiency, potential information loss, and sensitivity to image quality.[11] To address these issues, learning-based feature extraction methods have been developed, with recent advancements in computer vision and deep learning. These representation learning techniques involve pretraining models with pretext tasks and using part of the network as a feature extractor for downstream analysis.

These methods can be divided into two categories: task-oriented custom methods and generalist methods. Task-oriented methods[4,12–15] are designed specifically for particular biological research, such as cell-cycle stage prediction, and are generally pretrained with data from the same source. In contrast, generalist methods can be applied to many image types, and the training data are usually not specific to any particular biological problem. One of the most widely used generalist methods involves using models trained on ImageNet[16] (a natural image classification task), which has also been used in recent multimodal research.[17] However, the extent to which the feature extraction patterns learned from natural images can capture the subtle phenotypes of microscopy images has not been fully validated by comparative research. To better match the feature domain to downstream microscopy image profiling tasks, the CytoImageNet[18] study was conducted, in which image representations were learned on a microscopy image classification task (890,000 images, 894 classes). Although this approach demonstrated performance comparable to that of ImageNet, it still relied on a supervised learning approach, which can be labor intensive, prone to biases from semantic annotations, and potentially increase the difficulty of achieving good representation performance.

Self-supervised representation learning methods allow models to learn directly from pixels without relying on predefined semantic annotations. This approach involves transforming the original images and training the model to learn the mapping between the transformed and original images. Various transformation methods have been used, such as direct copying,[19] partial channel drop,[20] and image masking,[21] with masked visual representation learning being particularly popular in natural image studies.[22–24]

Recently, several studies have been reported about self-supervised learning techniques developed on specific microscopy image datasets. Pandey et al.[25] showed the effectiveness of a colorization pretext task pretrained on an electron microscopy image dataset. The GAN-DL study presented a generative self-supervised learning method that can learn efficient image representation based on Cell Painting high-content screening data.[26] The Cytoself approach[19] exhibited good performance with self-supervised protein localization profiling and clustering. Furthermore, recent advances in generalist cell segmentation algorithms[27–29] have demonstrated that heterogeneous microscopy images can be effectively handled by a single model. Despite this exciting progress, the complexity and diversity of microscopy images pose significant challenges in the development of generalist tools for microscopy image profiling, including handling images with varying resolutions and channel numbers (such as 1, 2, 3, 5, or 56),[3,4,7,28,30] joint representation learning for multiple image styles, processing various image types, and addressing technical variations in high-content experiments that may introduce batch effects in the feature space.[31,32]

The development of a high-performance, generalist image representation tool is important for microscopy image analysis. In addition to facilitating accurate downstream analyses, such a tool would enable unsupervised analysis for identifying new phenotypes. Moreover, generalist tools can facilitate the separation of feature extraction and downstream analysis steps, allowing downstream process to be performed on computers with limited computing power. Then, image representations that are much smaller than the original images could be easily stored and transferred. In addition, secondary analyses, such as the creation of large image databases or joint use with other data representations, can be performed.

This study presents Microsnoop, a generalist tool for creating microscopy image representations based on masked self-supervised learning. The developed tool can handle heterogeneous images and includes a task distribution module to assist users with varying computing capabilities. We constructed effective processing pipelines for three different image categories (single-cell, full-field, and batch-experiment images). We evaluated the performance of Microsnoop using 10 evaluation datasets from various biological studies and compared the performance with that of generalist and custom algorithms. The findings demonstrate Microsnoop's robust and excellent feature extraction capabilities and potential for analyzing superresolution histopathology images and multimodal biological data. The tool is freely available at https://github.com/cellimnet/microsnoop-publish.

RESULTS

The design of a generalist representation tool

In this study, we developed a generalist tool called Microsnoop for creating microscopy image representations. Because large and diverse datasets are beneficial for training generalist models, we collected and curated 10,458 high-quality microscopy images from various sources published by the cell segmentation

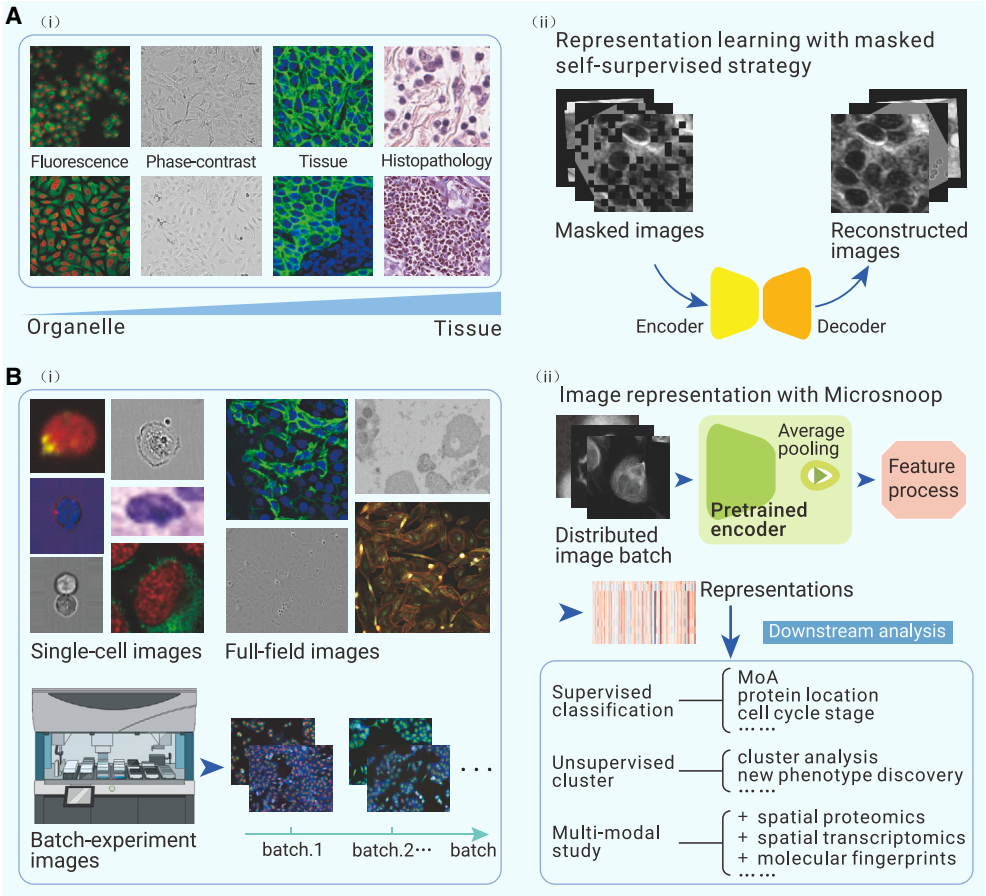


Figure 1. Design of Microsnoop for microscopy image representation (A) Schematic of the learning process. (i) Examples of the 4 main category images are shown. The image contents range from cellular organelles to tissues. (ii) A masked self-supervised learning strategy was used, and only images were needed for training without additional manual annotation. One-channel masked images were set as the input, and the encoder-decoder was used to reconstruct the original images. (B) At test time. (i) Example images from various downstream tasks are shown, with different resolutions, numbers of channels, and image types. These microscopy images are divided into 3 categories to ensure the broad coverage of image profiling needs. (ii) Application of Microsnoop. First, images are managed by an in-built task distribution module (see Figure 3A), which generates 1 batch of 1-channel images for feature extraction. Each batch of images is fed into the pretrained encoder, and the output smallest convolutional maps are processed by average pooling. Then, all of the extracted embeddings are processed according to different profiling tasks (detailed in sections about profiling single-cell, full-field, and batch-experiment images). The potential downstream analyses of our generalist representation tool are shown.

community.[27–29,33–35] These images were acquired using different technologies and have various resolutions and channel numbers, ranging from cellular organelles to tissues. The four main types of images are fluorescence, phase-contrast, tissue, and histopathology images (Figure 1A(i); Table S1). To account for the variability in the number of image channels, we configured the input to the neural network as one-channel images. We organized the training set into a one-channel data pool, from which images are sampled, augmented, and transformed for each training batch (see materials and methods). In terms of the network architecture design, we used a convolutional neural network (CNN)-based[36] architecture, despite the growing interest in transformer-based architectures[37] for natural image analysis. This choice was motivated by the superior performance of the CNN architecture observed in our preliminary evaluations

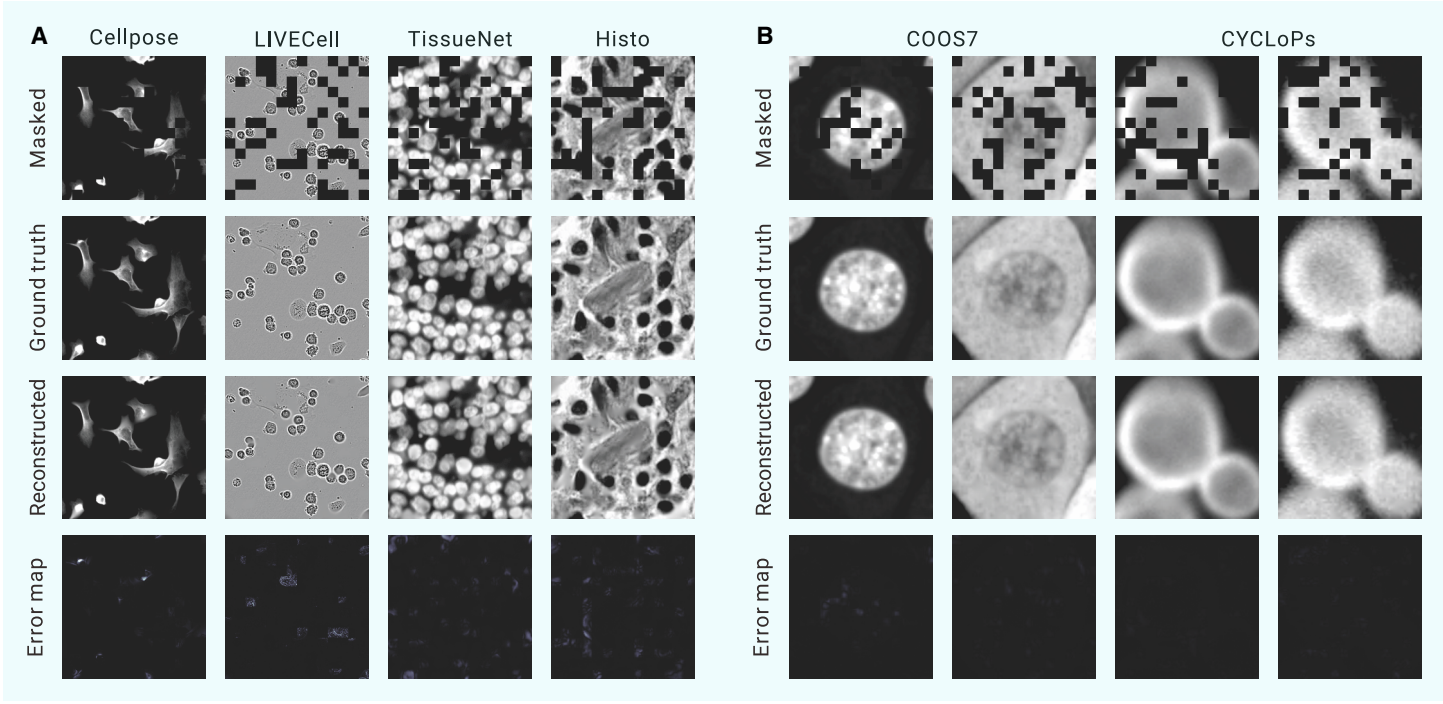


Figure 2. Reconstruction results with Microsnoop (A) Example results for images in the test set (validation set for the Histo subset because it has no test set) from 4 training subsets (Cellpose, LIVECell, TissueNet, and Histo), with a masking ratio of 25% applied on inputs. One representative image is selected for each image type. (B) Example results for single-cell images from evaluated data, with a masking ratio of 25% applied on inputs. The left 2 columns are from COOS7, and the right 2 columns are from CYCLoPs. Two representative images (different imaging channels of the same cell) are selected for each dataset. Example results for other evaluated datasets are shown in Figure S5. The error map shows the difference between the ground truth and reconstructed images.

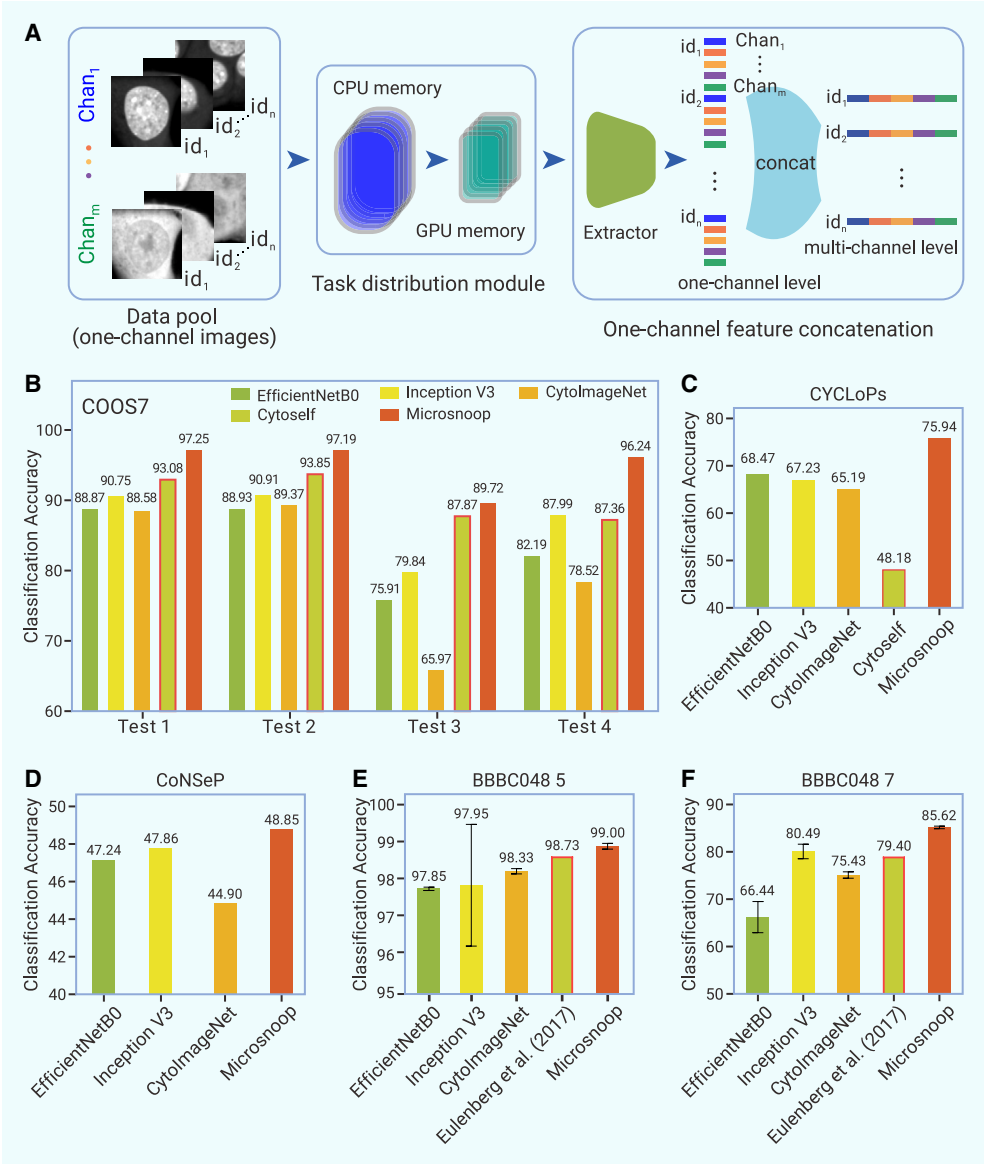


Figure 3. Profiling with Microsnoop on single-cell images (A) Pipeline. Every channel in the single-cell image is processed independently, and the 1-channel level embeddings are concatenated to obtain multi-channel level image representations. A task distribution module is used to prevent memory overflow. The extractor denotes the pretrained encoder combined with the average pooling layer shown in Figure 1B(ii). (B–F) Benchmarks. (B) Benchmark on COOS7, containing 4 separate test sets. (C) Benchmark on CYCLoPs. (D) Benchmark on CoNSeP. (E and F) Benchmarks on BBBC048, with 2 different classification tasks. The performance of each custom method is highlighted with a red border. Error bars represent the mean ± SD based on the 5-fold cross-validation results. The results of other metrics are shown in Table S3.

(Figure S1). This performance disparity may be attributed to the difference in the amount of training data. Typically, pretraining a vision transformer architecture[38] requires considerable data, with over 1 million or even 1 billion images used in natural image studies.[21] However, our microscopy image dataset included fewer training data, which may not have been sufficient to adequately train the transformer-based architecture.

We used a masked self-supervised learning strategy to train the network, with a randomly selected percentage of image patches masked and used as inputs. The network was then tasked with reconstructing the original, unmasked images. During training, masked images were encoded as high-level features through four consecutive downsampling steps, and image reconstruction was performed with a mirror-symmetric upsampling strategy (Figure 1A(ii)). The learning process was guided by minimizing the self-supervised loss function (see materials and methods), which allows the model to learn useful features to recover the masked parts of the images based on the information present in the remaining parts. This is a challenging task that necessitates a comprehensive understanding beyond simple low-level image statistics. We investigated the optimal mask ratio for learning features from microscopy images and found that a 25% mask was optimal for this task (Figure S2). We also conducted an experiment to explore the impact of the training data scale. The results show that the subset Cellpose provides a good starting point, and the model performance improves as more datasets are added (Figure S3).

During the test, our tool was not focused on a particular downstream profile task. After comprehensively analyzing various styles of microscopy images, we classified them into three categories: single-cell images, full-field images, and batch-experiment images (Figure 1B(i)). To account for the variability in device performance, we developed a task distribution module that operates on a batch-by-batch basis to create image representations (Figure 1B(ii)). To accommodate the three different image categories, we describe the corresponding feature processing pipelines in detail in the following sections.

Diversified evaluation datasets

In prior studies, researchers primarily investigated a limited number of specific datasets.[5,39–41] In our work, to more comprehensively evaluate our generalist tool, we collected and curated 10 evaluation datasets, including commonly used datasets and some novel additions, with over 2,230,000 images (Figures S4A–S4J, see materials and methods). These images show diverse characteristics, including various resolutions, image types, numbers of channels, and biological applications (Table S2). We used t-distributed stochastic neighbor embedding[42] to visualize the representations of the evaluation images and pretrained images, and noted that in the embedding space, images were generally placed according to the dataset categories, with little overlap in the distributions of the evaluation and training data (Figure S4K). Among the 10 evaluation datasets, 4 included single-cell images, 4 included full-field images, and 2 included batch-experiment images. To test the representation performance of the models on fluorescence images, including bright-field channels, we used COOS7 Test 1–4,[41] CYCLoPs,[3] BBBC048,[4] BBBC014, BBBC021[43] and RxRx19a.[26] RxRx19a, consisting of 5-channel Cell Painting[30] images, can also be used to test the ability of the model to represent high-dimensional data. The LIVECell Test[28] and TissueNet Test[29] datasets were designed to evaluate the representation performance of a model on phase-contrast and tissue images, respectively. To assess the ability of the model to handle challenging histopathology images, we used the CoNSeP[44] dataset. Furthermore, we used the CEM500K[25] dataset to explore model performance on electron microscopy images, which greatly differed from our training data.

Microsnoop accurately reconstructs the masked input images

To qualitatively evaluate model performance in the reconstruction task, we show examples of each image type (Figure 2A). When the 25% masked image was input into the pretrained network, the network produced a reconstructed image that closely resembled the original image, with only some of the detailed textures lost. This level of detail recovery is not easily achievable by humans. The reconstruction results of single-cell images from the evaluation datasets were even more impressive, with the reconstructed images nearly indistinguishable from the original images (Figures 2B and S5). The model performed better on single-cell images than full-field images, which can be attributed to cellular heterogeneity, and results in diverse cell phenotypes.

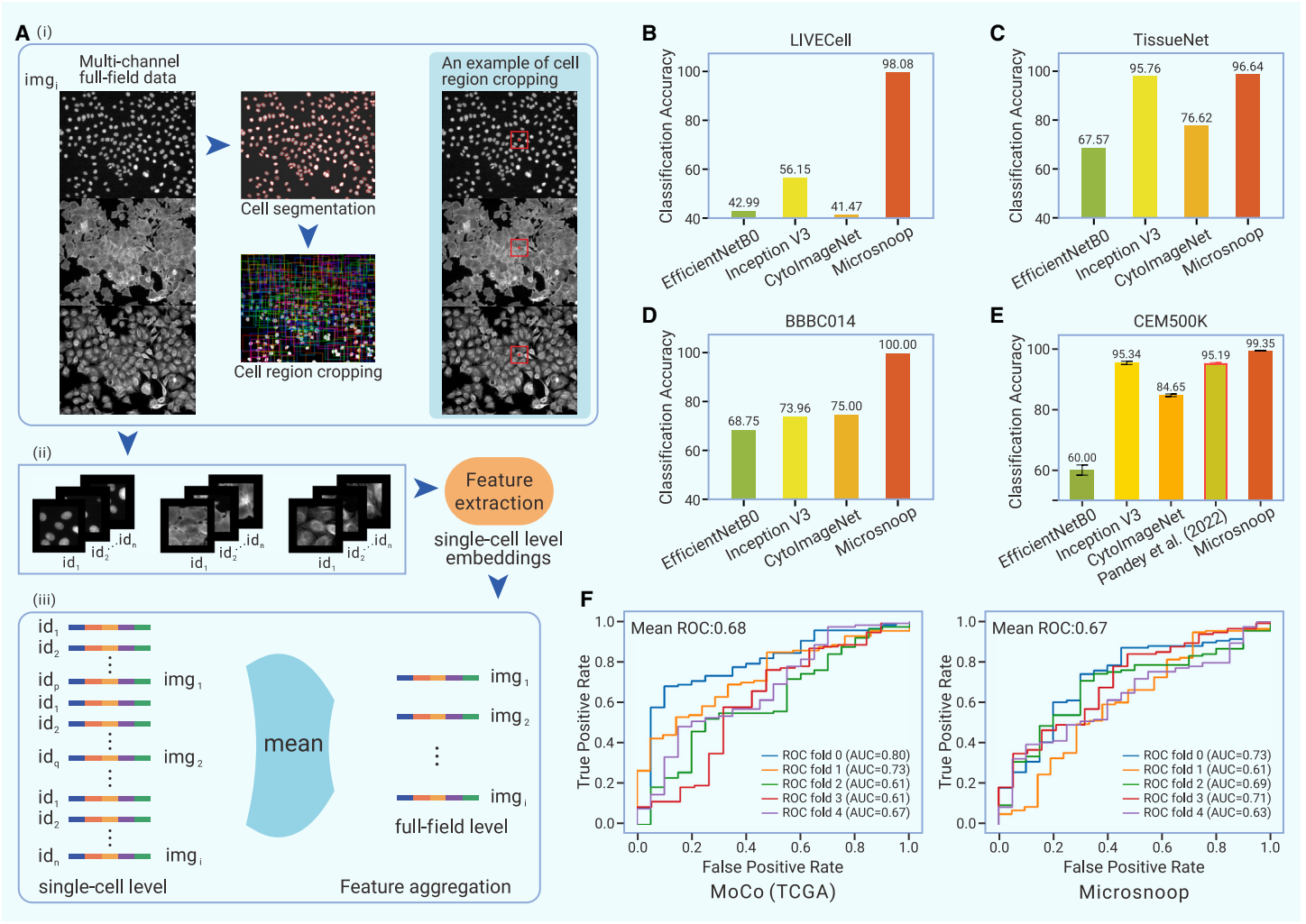


Figure 4. Profiling with Microsnoop on full-field images (A) Pipeline. (i) The cell segmentation algorithm is applied to the easiest channel (e.g., the nucleus channel) in the multichannel full-field image, and then the cell region for each single cell is computed and cropped. (ii and iii) Multichannel single-cell images are processed as in Figure 3A (ii), and (iii) the output single-cell level embeddings are aggregated to obtain the full-field level image representations. (B) Benchmark on LIVECell. (C) Benchmark on TissueNet. (D) Benchmark on BBBC014. (E) Benchmark on CEM500K. The performance of each custom method is highlighted with a red border. Error bars represent the mean ± SD based on the 5-fold cross-validation results. The results of other metrics are shown in Table S3. (F) Joint use of Microsnoop with the Lazard et al.[50] pipeline for analyzing TCGA dataset. Receiver operating characteristic curves are shown.

Compared to reconstructing cells in full-field images, reconstructing cells in single-cell images is simpler because relevant information can be obtained more easily. According to the outstanding performance achieved in the pretext task, we performed quantitative evaluation experiments profiling single-cell, full-field, and batch-experiment images, which are detailed in the following sections.

Microsnoop profile of single-cell images with one-channel feature concatenation

Single-cell images can be produced directly by an imaging instrument such as an imaging flow cytometer[45] or obtained through cell segmentation processing on full-field images. To accommodate the variable number of channels, we devised a one-channel feature concatenation strategy (Figure 3A). Each channel in the multichannel image is processed independently by Microsnoop, and the resulting embeddings are concatenated in an orderly manner. This concatenation strategy was developed based on our evaluations of five feature aggregation methods (Figure S6; see materials and methods). To prevent confusion during processing, a unique index is assigned to each image when multiple images are being processed. To address potential memory overflow issues when processing large batches of data, we established a task distribution module. This module efficiently manages image pathways and distributes images for processing, reads the images in the central processing unit (CPU) and transfers them to the graphics processing unit (GPU) as needed. The user can optimize performance by adjusting parameters according to the available memory capacity of the CPU and GPU. Furthermore, our system has a distributed design that can support multiple GPUs and can thus address increasing data demands.

In our benchmark study, we included three previously developed generalist methods for comparison: EfficientNetB0,[46] Inception V3,[47] and CytoImageNet[18] (see materials and methods). To comprehensively evaluate the performance of Microsnoop, we compared the performance of Microsnoop with that of several custom models based on datasets for which these models show good performance, such as comparing Microsnoop to Cytoself (pretrained on a dataset containing 1,100,253 cropped images of 1,311 endogenously labeled proteins from the OpenCell database) using protein localization datasets. We did not use any data from the evaluated dataset during the feature extraction process, and our evaluation using four different metrics demonstrated the outstanding performance of Microsnoop (Table S3), which consistently outperformed all of the other methods, including the custom methods (Figures 3B–3F).

Microsnoop profile of full-field images with cell region cropping

Full-field images are a common format directly obtained by most microscopes. Cell segmentation is usually the first step in phenotype profiling due to the inherent heterogeneity of cells. Although various generalist segmentation algorithms[27–29] and fine-tuning strategies[48,49] have been developed, segmentation errors may still occur. For instance, in large drug screening experiments, cell body images may include several phenotypes, and segmentation algorithms may perform well with some phenotypes but poorly with others (Figure S7A), leading to unpredictable effects in downstream analyses. To mitigate these

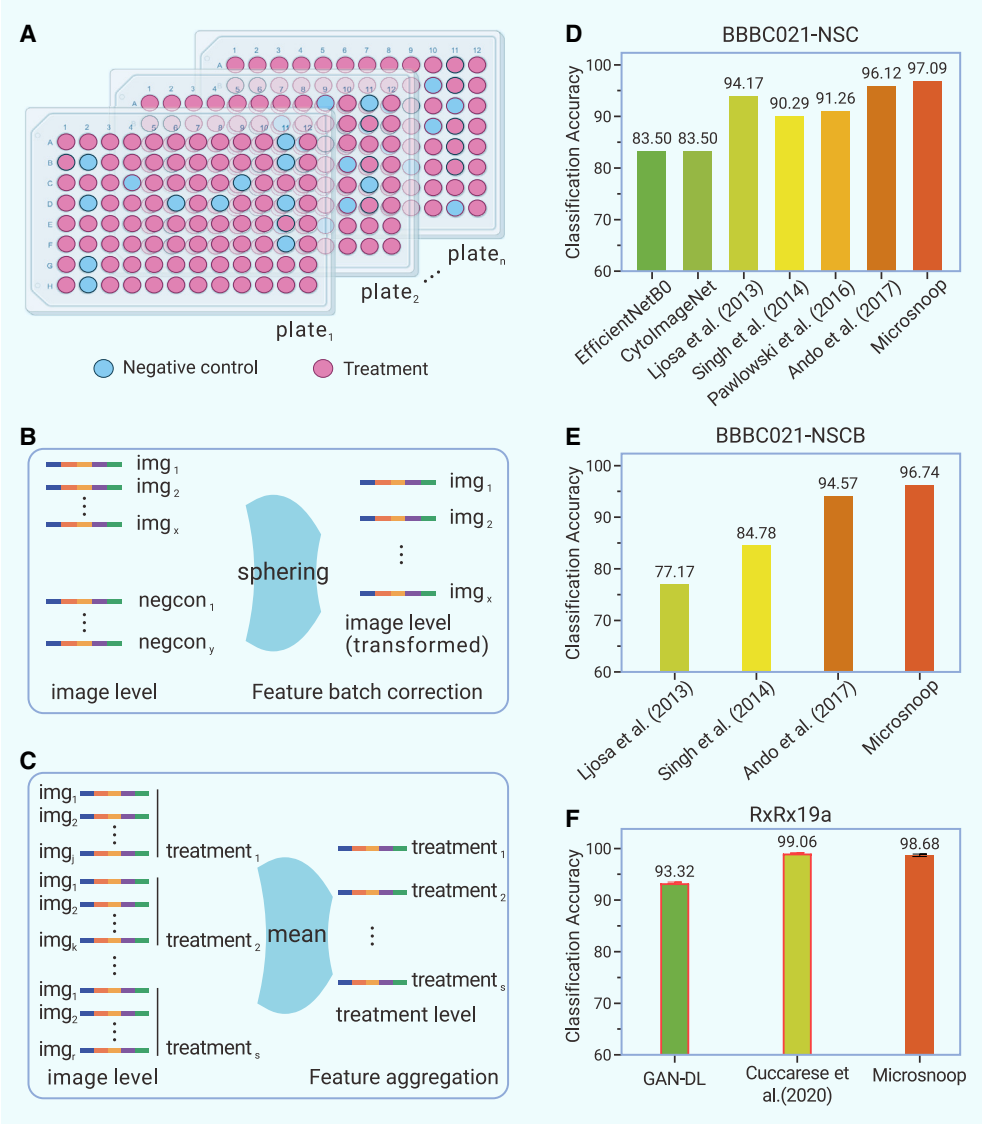


Figure 5. Profiling with Microsnoop on batch-experiment images (A) Schematic of multiwell plates in a drug screening experiment containing negative control wells and different treatment wells set in each plate. (B) Batch correction on image level representations. (C) Feature aggregation on image level embeddings to obtain treatment-level image representations. (D and E) Benchmark on BBBC021, with Not-Same-Compound (NSC) metric for (D) and Not-Same-Compound-or-Batch (NSCB) metric for (E). (F) Benchmark on RxRx19a. The performance of each custom method is highlighted with a red border. Error bars represent the mean ± SD based on the 5-fold cross-validation results. The results of other metrics are shown in Table S3.

issues, we introduced a cell region cropping strategy, in which the segmentation algorithm is applied only on the easiest channel, such as the nucleus channel, which presents more robust segmentation results (Figure S7B). Cell regions are computed and cropped on the segmentation masks and rescale constant (Figure 4A(i); see materials and methods). Then, Microsnoop extracts features from the cropped single-cell images, as described above (Figure 4A(ii)). Finally, the resulting single-cell level embeddings are aggregated by computing their mean to obtain the full-field level representations (Figure 4A(iii)).

We evaluated the representation ability of Microsnoop on three full-field image classification tasks. The results showed that Microsnoop outperformed the other methods, with a 41.93% improvement based on the LIVECell test dataset (Figures 4B–4D). Furthermore, Microsnoop showed robustness with various image styles, with accuracies of 98.08%, 96.64%, and 100.00% based on the LIVECell Test, TissueNet Test, and BBBC014 datasets, respectively.

Two other full-field image profile modes and the robustness of the cell region cropping mode

In addition to the cell region cropping mode, we provided two alternative modes for processing full-field datasets: rescaling and tile mode (Figures S8A–S8C; see materials and methods). We evaluated the performance of these three processing modes, including different rescale constants for the cell region cropping mode, on the full-field and batch-experiment datasets (Figures S8D–S8G). The rescaling and tile modes outperformed the single-cell mode on the LIVECell and TissueNet tests; however, both modes displayed significantly reduced performance on the BBBC021 dataset. The rescaling mode may have underperformed because it discards high-resolution phenotypic information during the rescaling process. The reduced performance with the tile mode may have occurred because important subtle phenotype variations present in certain regions in the full-field images are averaged out. In contrast, the cell region cropping mode displayed robust performance across a range of parameters on all four datasets.

Nevertheless, these two modes require less time and memory than the single-cell mode (Figures S8H and S8I). In addition, these modes enable the application of our tool in scenarios in which segmentation is unavailable or not considered. For the CEM500K dataset, segmentation is challenging, but the rescaling mode still works. Our model exhibited the best performance, even when compared to the Pandey et al.[25] model, which was pretrained on a subset of this dataset (Figure 4E). To explore whether our tool can be applied to analyze superresolution histopathology images, The Cancer Genome Atlas (TCGA) dataset[50] was used. We integrated Microsnoop into the Lazard et al.[50] pipeline by replacing their custom MoCo representation method. Areas of interest can be effectively tiled in this pipeline (Figure S9), and the image tiles can be directly embedded with the rescaling mode of Microsnoop without segmentation. Our results show that Microsnoop exhibits competitive results with MoCo (TCGA), which was pretrained on 5,300,000 histopathology image tiles of size 224 × 224 tiled from the evaluation dataset (Figure 4F). For datasets containing a large number of images, such as the RxRx19a dataset (1,527,600 images of size 1,024 × 1,024), cell segmentation can further increase the number of images by hundreds of times. This increase results in significant economic and time costs for computations, making it challenging to use the segmentation-based mode. In this scenario, the tile mode can be applied because it balances cost-effectiveness with reducing information loss.

Microsnoop profile of batch-experiment images with sphering batch correction

Batch effects can be introduced into single-cell or full-field data due to technical variability, which can affect downstream analysis[31,32,39,40] (Figure 5A). To address this issue, we used a sphering batch correction method.[51] This assumes that the large variations observed in negative controls are associated with confounders, and any variation that is not observed in controls is associated with phenotypes. The sphere transformation method aims to separate phenotypic variation from confounders. In our image representation pipeline for batch-experiment images, Microsnoop first extracts features from images, and the resulting image-level representations are corrected by the sphering transformation strategy (Figure 5B). Finally, the image-level representations are aggregated to treatment-level representations by computing their mean (Figure 5C).

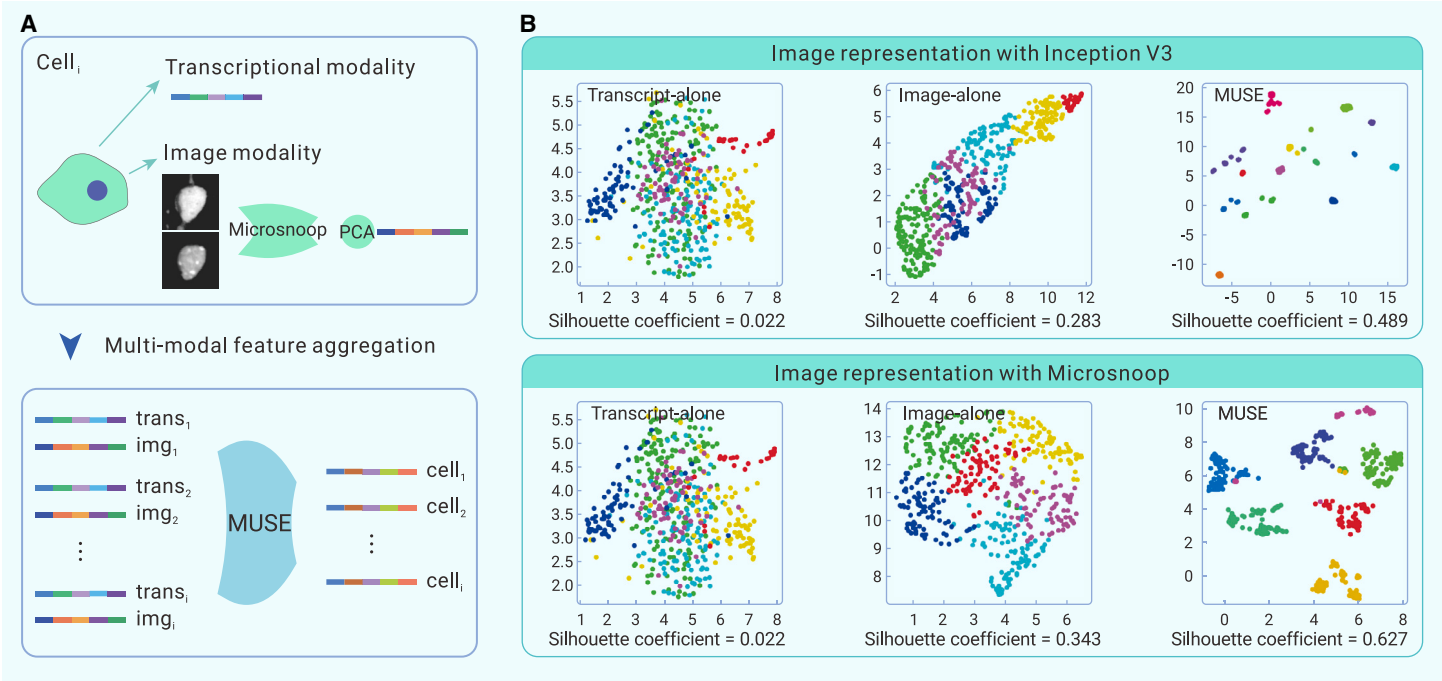


Figure 6. Joint use of Microsnoop and MUSE (A) Pipeline. Image modality data are first processed by Microsnoop, and then principal component analysis is performed on the output representations to reduce feature dimensionality. Finally, 2 modality representations are combined by MUSE. (B) Uniform manifold approximation and projection visualization of different modality latent spaces on seqFISH+ using 2 image representation methods. Silhouette score was used to quantify the separateness of clusters.

First, we evaluated the representation ability of Microsnoop on the classic BBBC021 dataset, including previously reported results in the comparisons. We did not use any data from this dataset during the feature extraction process, but Microsnoop still achieved state-of-the-art performance even compared with the methods exclusively studied on it (Figures 5D and 5E). For the 5-channel Cell Painting[30] dataset RxRx19a, our method outperforms another self-supervised method, GAN-DL,[26] which was pretrained on the whole RxRx19a dataset and achieves comparable performance to a supervised learning approach[6] that was pretrained on a very large annotated Cell Painting dataset (containing 125,510 images), RxRx1[52] (Figure 5F).

Microsnoop improves the performance of the multimodal structured embedding algorithm

A recent study of the multimodal structured embedding algorithm (MUSE[17]) showed that this model obtains impressive performance in integrative spatial analyses of image and transcriptional data. The authors conducted simulation experiments to assess the performance of MUSE with degraded transcriptional data. Here, we focused on the impact of image feature quality, and the simulation experiment results showed that as the quality of the image representations improves, the performance of MUSE significantly improves (Figure S10). Next, we tested Microsnoop on the real-world dataset seqFISH+[8] and compared the performance with that of the representation method used in the original paper. We combined Microsnoop and MUSE (Figure 6A) and found that higher quality image representations lead to greater improvements in MUSE performance (Figure 6B).

DISCUSSION

The accurate analysis of heterogeneous microscopy images, a critical aspect of both fundamental and applied biological research, is highly valued by the microscopy image analysis community.[53,54] In this study, we present Microsnoop, a generalist deep learning tool for microscopy image profiling. Our proposed tool offers promising advancements in this field. Microsnoop was trained on large-scale high-quality data using a masked self-supervised pretext task, and the model learned valuable features for generalist image representations. Our tool is flexible, with an efficient task distribution module and custom pipelines for three image categories, and can meet various user needs. A one-channel feature concatenation strategy was proposed for adapting to varying channel numbers. For full-field images, we provided three analysis modes. The cell region cropping-based single-cell profile mode shows more robust performance, and the rescaling and tile modes can cover segmentation-independent profiling scenarios. In addition, Microsnoop can mitigate batch effects in batch-experiment images with a sphering transformation strategy. Our benchmark results demonstrate the excellent microscopy image representation ability of Microsnoop without using any new data for fine-tuning. By integrating Microsnoop with an exceptional pipeline, superresolution histopathology images can be analyzed. Furthermore, the enhanced representation of unimodal image data leads to significant improvements in the performance of multimodal algorithms.

In our methodology experiments, we found that a mask ratio of 25% is optimal for microscopy images, which is significantly lower than the 75% mask that has been found to be optimal for natural images.[21] The difference is due primarily to the smaller size and varied content of microscopy images, which may result in lost information if too much reference information is masked. In the CytoImageNet study, the authors attempted to develop a microscopy image classification task to mimic the success of ImageNet. However, unlike natural images, it is difficult to obtain and determine class labels for microscopy images. Therefore, they assigned weak labels to images based on associated metadata. Although CytoImageNet and Microsnoop both use microscopy images for pretraining, a more effective pretext task seems to be more beneficial for microscopy image representation learning. Compared to Cytoself, a custom self-supervised representation method for protein subcellular location images, our model demonstrates stronger generalizability with COOS7 and CYCLoPs, despite not being specifically trained for protein localization tasks. Our method is unique in that it does not require domain-specific knowledge and was developed to create generalist image representations. Our benchmark study showed that a single network can handle heterogeneous microscopy images, which is consistent with results in the related domain of cell segmentation.[27] Furthermore, our pretext task was trained using the same network structure as Cellpose. In the future, additional systematic research could be conducted to investigate the effectiveness of transfer learning methods based on our pretrained model for handling other tasks, such as segmentation and tracking. Moreover, this is reminiscent of the recent success of large pretrained language models in the field of natural language processing.[55–57] With continued advancements in computer vision and models for microscopy image representations and other image processing tasks such as cell segmentation, it may be possible to merge these models into a single, unified model in the future.

Although Microsnoop is a powerful tool, there are several areas for improvement. For example, further evaluation is needed to determine the efficacy of our approach in one-channel feature concatenation and feature aggregation

with three-dimensional and time-series imaging datasets in comparison to training a network to directly extract spatial or temporal information. To enhance the capabilities of Microsnoop, future work could include exploring alternative architectures, incorporating additional self-supervised pretext tasks for multitask learning, using cross-channel correlation information, and refining the single-cell level feature aggregation methods. Moreover, the current training images are still smaller than natural images, and more training data combined with the transformer architecture could be studied to improve performance. The investigation into the influence of the training set demonstrates the capacity of our model for ongoing development. Notably, a considerable number of datasets that we evaluated were not involved in pretraining. We intend to regularly retrain the model using community-contributed data to consistently improve Microsnoop. Continual learning-based strategies[58,59] can be explored to prevent catastrophic forgetting. Furthermore, although Microsnoop can be used for data storage and sharing, sharing private data via embeddings still poses risks.[60] Thus, more specialized research on the safety of embeddings should be conducted. Finally, deploying our model on mobile devices to aid rapid detection could be a valuable application scenario.[61]

Overall, we developed an impressive, generalist tool for microscopy image representation. We anticipate its positive impact on the microscopy image analysis community, facilitating new phenotype discovery, data sharing, and the establishment of large image databases. Furthermore, we envision that Microsnoop can be used effectively in multimodal studies, such as combining molecular and image representations for mechanism of action prediction,[62,63] exploring the relationship between gene expression and image representations for drug discovery[64] and other broad applications.[65,66]

MATERIALS AND METHODS

See supplemental information for details.

DATA AND CODE AVAILABILITY

The links to download the raw data of the training set and evaluation datasets are provided in Tables S1 and S2. The evaluation results of four different metrics (accuracy, Matthews correlation coefficient, balanced accuracy and F1 score) are provided in Table S3. The per-class accuracy of Microsnoop for each evaluation is provided in Table S4. The new evaluation datasets generated by this study are available on figshare: https://doi.org/10.6084/m9.figshare.22197607.

The TCGA dataset is available at https://portal.gdc.cancer.gov/.

seqFISH+ mouse cortex dataset: Transcript data were downloaded from https://github.com/CaiGroup/seqFISH-PLUS. Image data were provided by L. Cai, the corresponding author of the seqFISH+ paper.[8]

The source code for Microsnoop, including a detailed tutorial, is available on GitHub (https://github.com/cellimnet/microsnoop-publish). A configured Amazon Machine Image is available for quickly and conveniently deploying Microsnoop for microscopy image analysis.

All of the data in this study are available from the corresponding author upon reasonable request.

REFERENCES

1. Caicedo, J.C., Singh, S., and Carpenter, A.E. (2016). Applications in image-based profiling of perturbations. Curr. Opin. Biotechnol. ***39***, 134–142.
2. Pratapa, A., Doron, M., and Caicedo, J.C. (2021). Image-based cell phenotyping with deep learning. Curr. Opin. Chem. Biol. ***65***, 9–17.
3. Lu, A.X., Chong, Y.T., Hsu, I.S., et al. (2018). Integrating images from multiple microscopy screens reveals diverse patterns of change in the subcellular localization of proteins. Elife ***7***, e31872.
4. Eulenberg, P., Köhler, N., Blasi, T., et al. (2017). Reconstructing cell cycle and disease progression using deep learning. Nat. Commun. ***8***, 463.
5. Pawlowski, N., Caicedo, J.C., Singh, S., et al. (2016). Automating morphological profiling with generic deep convolutional networks. Preprint at. bioRxiv. http://biorxiv.org/lookup/doi/10.1101/085118.
6. Cuccarese, M.F., Earnshaw, B.A., Heiser, K., et al. (2020). Functional immune mapping with deep-learning enabled phenomics applied to immunomodulatory and COVID-19 drug discovery. Preprint at. bioRxiv. http://biorxiv.org/lookup/doi/10.1101/2020.08.02.233064.
7. Schürch, C.M., Bhate, S.S., Barlow, G.L., et al. (2020). Coordinated cellular neighborhoods orchestrate antitumoral immunity at the colorectal cancer invasive front. Cell ***182***, 1341–1359.e19.
8. Eng, C.-H.L., Lawson, M., Zhu, Q., et al. (2019). Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. Nature ***568***, 235–239.
9. Carpenter, A.E., Jones, T.R., Lamprecht, M.R., et al. (2006). CellProfiler: image analysis software for identifying and quantifying cell phenotypes. Genome Biol. ***7***, R100.
10. Pau, G., Fuchs, F., Sklyar, O., et al. (2010). EBImage–an R package for image processing with applications to cellular phenotypes. Bioinformatics ***26***, 979–981.
11. Singh, S., Bray, M.-A., Jones, T.R., et al. (2014). Pipeline for illumination correction of images for high-throughput microscopy. J. Microsc. ***256***, 231–236.
12. Caicedo, J.C., McQuin, C., Goodman, A., et al. (2018). Weakly supervised learning of single-cell feature embeddings. In Proc. IEEE Conference on Computer Vision and Pattern Recognition, *2018*, pp. 9309–9318.
13. Lu, A.X., Kraus, O.Z., Cooper, S., et al. (2019). Learning unsupervised feature representations for single cell microscopy images with paired cell inpainting. PLoS Comput. Biol. ***15***, e1007348.
14. Adnan, M., Kalra, S., and Tizhoosh, H.R. (2020). Representation learning of histopathology images using graph neural networks. In Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 988–989.
15. Perakis, A., Gorji, A., Jain, S., et al. (2021). Contrastive learning of single-cell phenotypic representations for treatment classification. In Machine Learning in Medical Imaging, *12966*, C. Lian, X. Cao, I. Rekik, X. Xu, and P. Yan, eds., pp. 565–575.
16. Russakovsky, O., Deng, J., Su, H., et al. (2015). ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. ***115***, 211–252.
17. Bao, F., Deng, Y., Wan, S., et al. (2022). Integrative spatial analysis of cell morphologies and transcriptional states with MUSE. Nat. Biotechnol. ***40***, 1200–1209.
18. Hua, S.B.Z., Lu, A.X., and Moses, A.M. (2021). CytoImageNet: a large-scale pretraining dataset for bioimage transfer learning. In Proc. Advances in Neural Information Processing Systems.
19. Kobayashi, H., Cheveralls, K.C., Leonetti, M.D., et al. (2022). Self-supervised deep learning encodes high-resolution features of protein subcellular localization. Nat. Methods ***19***, 995–1003.
20. Wong, D.R., Conrad, J., Johnson, N., et al. (2022). Trans-channel fluorescence learning improves high-content screening for Alzheimer's disease therapeutics. Nat. Mach. Intell. ***4***, 583–595.
21. He, K., Chen, X., Xie, S., et al. (2022). Masked autoencoders are scalable vision learners. In Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 16000–16009.
22. Liu, X., Zhou, J., Kong, T., et al. (2022). Exploring target representations for masked autoencoders. Preprint at. arXiv. https://arxiv.org/abs/2209.03917.
23. Li, Z., Chen, Z., Yang, F., et al. (2021). MST: masked self-supervised transformer for visual representation. In SAVE Proc. Advances in Neural Information Processing Systems, p. 35.
24. Wei, C., Fan, H., Xie, S., et al. (2022). Masked feature prediction for self-supervised visual pre-training. In Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 14668–14678.
25. Pandey, V., Brune, C., and Strisciuglio, N. (2022). Self-supervised learning through colorization for microscopy images. In Image Analysis and Processing - ICIAP 2022, S. Sclaroff, C. Distante, and M. Leo, et al., eds., pp. 621–632.
26. Mascolini, A., Cardamone, D., Ponzio, F., et al. (2022). Exploiting generative self-supervised learning for the assessment of biological images with lack of annotations. BMC Bioinf. ***23***, 295.
27. Stringer, C., Wang, T., Michaelos, M., et al. (2021). Cellpose: a generalist algorithm for cellular segmentation. Nat. Methods ***18***, 100–106.
28. Edlund, C., Jackson, T.R., Khalid, N., et al. (2021). LIVECell-a large-scale dataset for label-free live cell segmentation. Nat. Methods ***18***, 1038–1045.
29. Greenwald, N.F., Miller, G., Moen, E., et al. (2022). Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning. Nat. Biotechnol. ***40***, 555–565.
30. Bray, M.-A., Singh, S., Han, H., et al. (2016). Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. Nat. Protoc. ***11***, 1757–1774.
31. Leek, J.T., Scharpf, R.B., Bravo, H.C., et al. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. Nat. Rev. Genet. ***11***, 733–739.
32. Lin, A., and Lu, A.X. (2022). Incorporating knowledge of plates in batch normalization improves generalization of deep learning for microscopy images. In Proc. International Conference on Machine Learning, pp. 74–93.
33. Kumar, N., Verma, R., Anand, D., et al. (2020). A multi-organ nucleus segmentation challenge. IEEE Trans. Med. Imag. ***39***, 1380–1391.
34. Verma, R., Kumar, N., Patil, A., et al. (2021). MoNuSAC2020: a multi-organ nuclei segmentation and classification challenge. IEEE Trans. Med. Imag. ***40***, 3413–3423.
35. Amgad, M., Atteya, L.A., Hussein, H., et al. (2022). NuCLS: a scalable crowdsourcing, deep learning approach and dataset for nucleus classification, localization and segmentation. GigaScience ***11***, giac037.
36. Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: convolutional networks for biomedical image segmentation. In Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 234–241.
37. Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. In Proc. Advances in Neural Information Processing Systems, *30*.
38. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. (2021). An image is worth 16x16 words: transformers for image recognition at scale. In International Conference on Learning Representations.
39. Ando, D.M., McLean, C.Y., and Berndl, M. (2017). Improving phenotypic measurements in high-content imaging screens. Preprint at. bioRxiv. http://biorxiv.org/lookup/doi/10.1101/161422.
40. Bray, M.-A., Weck, A.D., Durand, E., et al. (2022). High-content cellular screen image analysis benchmark study. Preprint at. bioRxiv. https://www.biorxiv.org/content/10.1101/2022.05.15.491989v1.abstract.

41. Lu, A., Lu, A., Schormann, W., et al. (2019). The Cells Out of Sample (COOS) dataset and benchmarks for measuring out-of-sample generalization of image classifiers. In Proc. Advances in Neural Information Processing Systems, p. 32.
42. Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. J. Mach. Learn. Res. ***9***, 2579–2605.
43. Caie, P.D., Walls, R.E., Ingleston-Orme, A., et al. (2010). High-content phenotypic profiling of drug response signatures across distinct cancer cells. Mol. Cancer Therapeut. ***9***, 1913–1926.
44. Graham, S., Vu, Q.D., Raza, S.E.A., et al. (2019). Hover-Net: simultaneous segmentation and classification of nuclei in multi-tissue histology images. Med. Image Anal. ***58***, 101563.
45. Schraivogel, D., Kuhn, T.M., Rauscher, B., et al. (2022). High-speed fluorescence image-enabled cell sorting. Science ***375***, 315–320.
46. Tan, M., and Le, Q.V. (2019). EfficientNet: rethinking model scaling for convolutional neural networks. In Proc. International Conference on Machine Learning, pp. 6105–6114.
47. Szegedy, C., Vanhoucke, V., Ioffe, S., et al. (2016). Rethinking the Inception architecture for computer vision. In Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826.
48. Xun, D., Chen, D., Zhou, Y., et al. (2022). Scellseg: a style-aware deep learning tool for adaptive cell instance segmentation by contrastive fine-tuning. iScience ***25***, 105506.
49. Pachitariu, M., and Stringer, C. (2022). Cellpose 2.0: how to train your own model. Nat. Methods ***19***, 1634–1641.
50. Lazard, T., Bataillon, G., Naylor, P., et al. (2022). Deep learning identifies morphological patterns of homologous recombination deficiency in luminal breast cancers from whole slide images. Cell Rep. Med. ***3***, 100872.
51. Moshkov, N., Bornholdt, M., Benoit, S., et al. (2022). Learning representations for image-based profiling of perturbations. Preprint at. bioRxiv. http://biorxiv.org/lookup/doi/10.1101/2022.08.12.503783.
52. Taylor, J., Earnshaw, B., Mabey, B., et al. (2019). RxRx1: an image set for cellular morphological variation across many experimental batches. In 7th International Conference on Learning Representations.
53. Caicedo, J.C., Cooper, S., Heigwer, F., et al. (2017). Data-analysis strategies for image-based cell profiling. Nat. Methods ***14***, 849–863.
54. Chandrasekaran, S.N., Ceulemans, H., Boyd, J.D., et al. (2021). Image-based profiling for drug discovery: due for a machine-learning upgrade? Nat. Rev. Drug Discov. ***20***, 145–159.
55. Devlin, J., Chang, M.-W., Lee, K., et al. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. Preprint at. arXiv. https://arxiv.org/abs/1810.04805.
56. Brown, T.B., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. In Proc. Advances in Neural Information Processing Systems, p. 33.
57. Min, B., Ross, H., Sulem, E., et al. (2021). Recent advances in natural language processing via large pre-trained language models: a survey. Preprint at. arXiv. http://arxiv.org/abs/2111.01243.
58. Davari, M., Asadi, N., Mudur, S., et al. (2022). Probing representation forgetting in supervised and unsupervised continual learning. In Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 16691–16700.
59. Mundt, M., Hong, Y., Pliushch, I., et al. (2023). A wholistic view of continual learning with deep neural networks: forgotten lessons and the bridge to active and open world learning. Neural Network. ***160***, 306–336.
60. Abdalla, M., Abdalla, M., Hirst, G., et al. (2020). Exploring the privacy-preserving properties of word embeddings: algorithmic validation study. J. Med. Internet Res. ***22***, e18055.
61. Wang, B., Li, Y., Zhou, M., et al. (2023). Smartphone-based platforms implementing microfluidic detection with image-based artificial intelligence. Nat. Commun. ***14***, 1341.
62. Sanchez-Fernandez, A., Rumetshofer, E., and Hochreiter, S. (2022). Contrastive learning of image- and structure- based representations in drug discovery. In International Conference on Learning Representations.
63. Tian, G., Harrison, P.J., Sreenivasan, A.P., et al. (2022). Combining molecular and cell painting image data for mechanism of action prediction. Preprint at. bioRxiv. http://biorxiv.org/lookup/doi/10.1101/2022.10.04.510834.
64. Haghighi, M., Caicedo, J.C., Cimini, B.A., et al. (2022). High-dimensional gene expression and morphology profiles of cells across 28,000 genetic and chemical perturbations. Nat. Methods ***19***, 1550–1557.
65. Liu, L., Bi, M., Wang, Y., et al. (2021). Artificial intelligence-powered microfluidics for nanomedicine and materials synthesis. Nanoscale ***13***, 19352–19366.
66. Wang, X., Xie, P., Chen, B., et al. (2022). Chip-based high-dimensional optical neural network. Nano-Micro Lett. ***14***, 221.

ACKNOWLEDGMENTS

This study was supported by the National Key R&D Program of China (2021YFC1712905), the National Natural Science Foundation of China (nos. 82173941 and 61872319), and the Key R&D Program of Zhejiang Province (no. 2023C01039). Y.W. was supported by the Innovation Team and Talents Cultivation Program of National Administration of Traditional Chinese Medicine (no. ZYYCXTD-D-202002) and the Fundamental Research Funds for the Central Universities (no. 226-2023-00114). We thank L. Cai at the California Institute of Technology for providing the seqFISH+ image data. We thank T. Walter for providing the pretrained MoCo model on the TCGA dataset. We thank W.K. Wang and L. Sun at Amazon Web Services China for their indispensable support in terms of computing resources and technology. We are grateful for the support from the ZJU PII-Molecular Devices Joint Laboratory and support from the "Medicine + X" interdisciplinary Center of Zhejiang University.

AUTHOR CONTRIBUTIONS

Y.W., X.Z., and R.W. supervised the study. D.X. acquired the data, established the pipelines, conducted the experiments, and performed the data analysis. D.X., Y.W., X.Z., and R.W. wrote the manuscript.

DECLARATION OF INTERESTS

A patent covering all of the main aspects of the use of Microsnoop as a tool for microscopy image representation has been filed by Zhejiang University (CNIPA application no. 202310155113.8). The application is currently pending. D.X., R.W., and Y.W., as employees and student of Zhejiang University, are named as co-inventors on the patent application. The remaining authors declare no competing interests.

SUPPLEMENTAL INFORMATION

It can be found online at https://doi.org/10.1016/j.xinn.2023.100541.

LEAD CONTACT WEBSITE

https://person.zju.edu.cn/en/yiwang_en.