nature methods

Analysis

https://doi.org/10.1038/s41592-024-02233-6

The multimodality cell segmentation challenge: toward universal solutions

Received: 27 July 2023

Accepted: 4 March 2024

Published online: 26 March 2024

Check for updates

Jun Ma[1,2,3], Ronald Xie[1,3,4], Shamini Ayyadhury[5,6,42], Cheng Ge[7,42], Anubha Gupta[8,42], Ritu Gupta[9,42], Song Gu[10,42], Yao Zhang[11,42], Gihun Lee[12], Joonkee Kim[12], Wei Lou[13,14], Haofeng Li[13], Eric Upschulte[15], Timo Dickscheid[15,16], José Guilherme de Almeida[17,18], Yixin Wang[19], Lin Han[20], Xin Yang[21], Marco Labagnara[22], Vojislav Gligorovski[22], Maxime Scheder[22], Sahand Jamal Rahi[22], Carly Kempster[23], Alice Pollitt[23], Leon Espinosa[24], Tâm Mignot[24], Jan Moritz Middeke[25,26], Jan-Niklas Eckardt[25,26], Wangkai Li[27], Zhaoyang Li[28], Xiaochen Cai[29], Bizhe Bai[30], Noah F. Greenwald[31], David Van Valen[32,33], Erin Weisbart[34], Beth A. Cimini[34], Trevor Cheung[1,35], Oscar Brück[36,37], Gary D. Bader[4,5,6,38,39,40] & Bo Wang[1,2,3,38,41]

Cell segmentation is a critical step for quantitative single-cell analysis in microscopy images. Existing cell segmentation methods are often tailored to specific modalities or require manual interventions to specify hyper-parameters in different experimental settings. Here, we present a multimodality cell segmentation benchmark, comprising more than 1,500 labeled images derived from more than 50 diverse biological experiments. The top participants developed a Transformer-based deep-learning algorithm that not only exceeds existing methods but can also be applied to diverse microscopy images across imaging platforms and tissue types without manual parameter adjustments. This benchmark and the improved algorithm offer promising avenues for more accurate and versatile cell analysis in microscopy imaging.

Cell segmentation is a fundamental task that is universally required for biological image analysis across a large number of different experimental settings and imaging modalities. For example, in multiplexed fluorescence image-based cancer microenvironment analysis, cell segmentation is the prerequisite for the identification of tumor subtypes, composition and organization, which can lead to important biological insights[1–3]; however, the development of a universal and automatic cell segmentation technique continues to pose major challenges due to the extensive diversity observed in microscopy images. This diversity arises from variations in cell origins, microscopy types, staining techniques and cell morphologies. Recent advances[4] have successfully demonstrated the feasibility of automatic and precise cellular segmentation for specific microscopy image types and cell types, such as fluorescence and mass spectrometry images[5,6], differential interference contrast images of platelets[7], bacteria images[8] and yeast images[9,10], but the selection of appropriate segmentation models remains a nontrivial task for nonexpert users in conventional biology laboratories.

Efforts have been made toward the development of generalized cell segmentation algorithms[8,11]; however, these algorithms were primarily trained using datasets consisting of gray-scale images and two-channel fluorescent images, lacking the necessary diversity to ensure robust generalization across a wide range of imaging modalities. For example, segmentation models have struggled to perform effectively on RGB images, such as bone-marrow aspirate slides stained with Jenner–Giemsa. Furthermore, these models often require manual selection of both the model type and the specific image channel to be segmented, posing challenges for biologists with limited computational expertise. In addition to directly training general cell segmentation

A full list of affiliations appears at the end of the paper. e-mail: bowang@vectorinstitute.ai

models with large-scale labeled datasets, transfer-learning-based algorithms are a complementary branch toward universal solutions, allowing biologists to rapidly train customized models on their own microscopy images. A prime example is Cellpose 2.0 (ref. 12), which demonstrates the efficacy of adapting a pretrained model to new images. Remarkably, it only requires 500–1,000 user-annotated image patches to achieve performance on par with models trained on thousands of image patches.

Biomedical image data science competitions have emerged as an effective way to accelerate the development of cutting-edge algorithms. Several successful competitions have been specifically organized for microscopy image analysis, such as the Cell Tracking Challenge (CTC)[13,14], the Data Science Bowl (DSB) Challenge[15] and the Colon Nuclei Identification and Counting (CoNIC) Challenge[16]. These competitions have played a crucial role in expediting the adoption of modern machine learning and deep-learning algorithms in biomedical image analysis; however, it is worth noting that these challenges have primarily focused on a limited subset of microscopy image types. For example, the CTC primarily concentrated on label-free images, thereby excluding stained images such as multiplexed immunofluorescent images. Similarly, the DSB Challenge emphasized nucleus segmentation in fluorescent and histology images while disregarding phase-contrast (PC) and differential interference contrast (DIC) images. The segmentation task in the CoNIC Challenge is also limited to nucleus segmentation in hematoxylin and eosin-stained images. Consequently, the algorithms developed through these competitions are often tailored to handle only specific types of microscopy images, limiting their generalizability. Moreover, the evaluation metrics used in these challenges predominantly prioritize segmentation accuracy, while neglecting algorithm efficiency. As a result, the pursuit of higher accuracy scores often leads to the adoption of computationally demanding approaches. For instance, the CTC top-performing algorithms[14] employed customized models for each dataset in the cell segmentation task, while the DSB winning algorithm[15] used an ensemble of 32 models. Such resource-intensive strategies hinder the wide deployment of these algorithms in biology practice.

To overcome the aforementioned limitations and foster the development of universal and efficient cell segmentation methods for microscopy images, we took the initiative to organize a global challenge at the Conference on Neural Information Processing Systems (NeurIPS). As one of the largest international conferences in the field of artificial intelligence (AI), NeurIPS provided an ideal platform for this endeavor. Participants in the challenge were provided with a diverse training set and a separate tuning set to develop and refine their cell segmentation algorithms. During the testing phase, participants were required to package their algorithms as Docker containers, enabling the challenge organizers to evaluate them on a carefully curated holdout testing set on the same computing platform. Notably, the holdout testing set incorporated images from new biological experiments, aiming to assess the algorithms' ability to generalize effectively to previously unseen data. Additionally, the testing set included two whole-slide images (WSIs), serving as a means to evaluate the algorithms' suitability for handling large-scale images. Different from existing challenges that focused on specific microscopy image types, this initiative represents the first instance where cell segmentation algorithms were challenged to efficiently handle a broad spectrum of microscopy images with one single model and generalize to new images without manual intervention.

Results

Challenge design: toward universal and efficient cell segmentation algorithms

The primary objective of this challenge was to benchmark universal algorithms capable of accurately segmenting cells from a wide range of microscopy images obtained from various imaging platforms and tissue types, without requiring additional parameter tuning (Fig. 1a). The algorithms were expected to operate in a fully automatic manner, generating cell instance masks where each cell is assigned a unique label. The challenge comprised two phases (Fig. 1b). In the development phase, participants were provided with a dataset consisting of 1,000 microscopy images, each accompanied by annotated cell masks. Recognizing the potential benefits of leveraging unlabeled data to enhance model performance[17], we also made an additional set of 1,725 unlabeled images available for participants to utilize. Participants were given the flexibility to decide whether to incorporate this unlabeled dataset into their algorithms. This setup aligns with real-world scenarios encountered in biological research, where only a limited number of labeled images are typically available alongside a wealth of unlabeled images.

To facilitate timely model validation, a separate tuning set containing 101 images was provided to participants, but the corresponding annotations were not disclosed. Instead, we established an online evaluation platform, enabling participants to upload their segmentation results and receive evaluation scores. These scores were made publicly available on a leaderboard, enabling direct comparisons among participants and their algorithms throughout the development phase.

In the subsequent testing phase, the top 30 teams, as ranked on the public tuning set leaderboard, were invited to make the testing submission. The testing set remained hidden from participants, aiming to avoid potential label leaking and cheating. To ensure standardized evaluation, participants were required to package and submit their algorithms as Docker containers. Challenge organizers ran the submitted Docker containers on the holdout testing set comprising 422 microscopy images. Out of the 30 invited top teams, 28 teams made successful submissions, whereas one team did not submit and another team submitted after the deadline, making their submission ineligible for the final ranking. To ensure a fair comparison, we executed the Docker containers sequentially on the same workstation. The running time for each image was recorded, alongside the corresponding segmentation accuracy score. Both of them were used for the final ranking and subsequent analysis of the algorithms (Methods).

Challenge data: a large and diverse multimodality microscopy image dataset

Data diversity plays a pivotal role in constructing generalist microscopy image segmentation models[18]. In this challenge, we incorporated the diversity of microscopy images from four dimensions: cell origins, staining methods, microscope types and cell morphologies (Fig. 2a). First, the origin of cells in microscopy images varies substantially, as they can derive from diverse tissues or exist within cell cultures under various conditions. This introduces considerable variability, as cells within tissues tend to be densely packed and spatially organized, whereas cells in culture are often sparsely distributed and randomly positioned. Second, the choice of staining methods, such as Jenner–Giemsa in brightfield microscopy or the utilization of specific antibodies in fluorescent microscopy, further contributes to the diversity by highlighting different cellular structures or proteins. Third, the use of different microscope types, such as brightfield, fluorescent, PC and DIC, introduces substantial differences in image characteristics, textures and associated artifacts. Fourth, cell morphologies exhibit substantial variations across different cell types. While most cells tend to have a round shape, certain cells may display elongated or irregular shapes.

We curated a diverse microscopy image dataset by collecting images (and annotations if available) from over 20 biology laboratories, including more than 50 different biological experiments (Supplementary Tables 1–3). This comprehensive dataset encompassed four common microscopy image modalities: brightfield, fluorescent, PC and DIC. The challenge garnered a large number of interest and participation, attracting over 400 participants from 37 different countries, reflecting the global reach and impact of the challenge (Fig. 2b). The training set contained a total of 1,000 images, with 300 images each in

https://doi.org/10.1038/s41592-024-02233-6

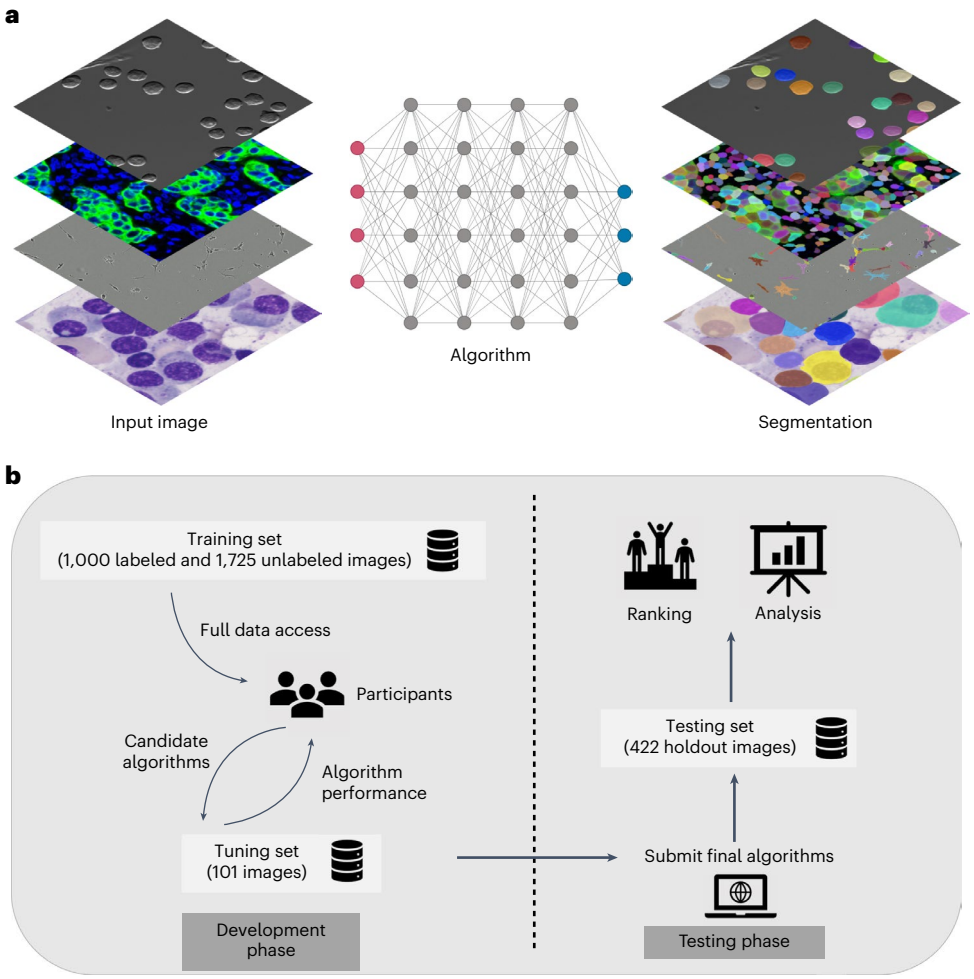


Fig. 1 | Overview of the challenge task and pipeline. a, The challenge aimed to facilitate the development of universal cell segmentation algorithms that can segment a wide range of microscopy images without manual intervention. **b**, The challenge contained two phases. During the development phase, participants developed automatic segmentation algorithms based on 1,000 labeled images and 1,725 unlabeled images. The algorithms could be evaluated on a tuning set with 101 images and the online evaluation platform automatically returned back the quantitative performance. During the testing phase, each team could submit one algorithm via the Docker container as the final solution, which was independently evaluated on the holdout testing set with 422 images to obtain ranking results.

the brightfield and fluorescent categories and 200 images each in the PC and DIC categories (Fig. 2c). The annotated dataset contained 12,702 cells in brightfield images, 130,194 cells in fluorescent images, 9,504 cells in PC images and 16,091 cells in DIC images (Fig. 2d). Notably, the higher cell count in fluorescent images compared to other modalities can be attributed to the denser distribution of cells observed in the collected fluorescent images.

Figure 2e shows four microscopy images randomly selected from each modality in the training set and testing set. To assess the algorithm's generalization capabilities, all testing images were sourced from new biological experiments, including some that featured previously unseen tissues or cell types not present in the training set. The testing set consisted of 120 brightfield images, 122 fluorescent images, 120 PC images and 60 DIC images (Fig. 2f). These quantities were determined based on the available images collected for the challenge. The number of cells in the testing set was comparable to or greater than that of the training set (Fig. 2g). Additionally, the fluorescent image subset of the testing set included two WSIs, which served the purpose of evaluating the algorithms' ability in handling large-scale imaging datasets.

In comparison to previous datasets utilized in cell segmentation challenges[14,19] and nucleus segmentation challenges[15,16], our dataset exhibits substantially enhanced diversity and encompasses a larger number of labeled cells. This extensive dataset serves as a fertile ground for fostering the development of advanced cell segmentation algorithms, enabling researchers to explore and innovate in the field.

Algorithm overview: the Transformer-based algorithm achieved superior performance

All algorithms in this challenge employed deep-learning-based approaches, a prevailing trend considering the remarkable performance achieved in various specific cell segmentation tasks[5,10,14], as well as in recent generalist cell segmentation algorithms[8,12]. Existing algorithms predominantly relied on convolutional neural networks (CNNs) such as U-Net[20] and DeepLab[21]; however, it is worth noting that these CNN-based cell segmentation models exhibited limited generalization capability when confronted with the task of segmenting diverse images without additional human intervention, such as manual selection of channels or model fine-tuning, as demonstrated in the following sections.

In contrast, Transformers[22], a new type of deep-learning network integrating attention mechanisms for feature extraction, have exhibited robust performance and generalization capabilities across various computer vision tasks[23,24]; however, the potential of Transformers in biological image analysis remains relatively unexplored[18]. Distinguished from existing benchmarks[14,15], our challenge provided

https://doi.org/10.1038/s41592-024-02233-6

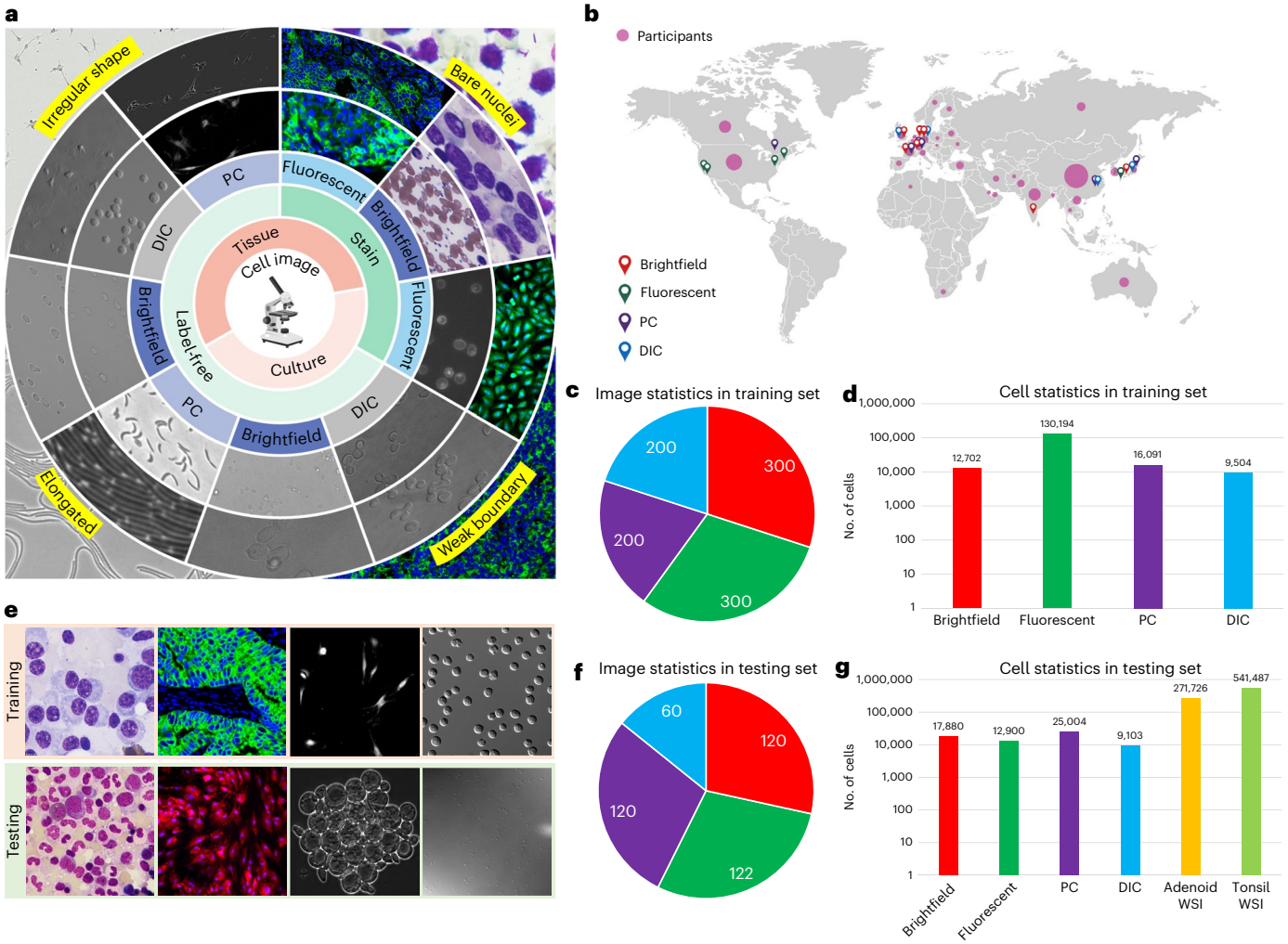


Fig. 2 | Dataset overview. a, The challenge provides a diverse microscopy image dataset that includes tissue cells, cultured cells, label-free cells, stained cells and different microscopes (brightfield, fluorescent, PC and DIC). **b**, The geographical distribution of data sources and challenge participants. The red, green, purple and blue address icons denote the countries or regions where the brightfield, fluorescent, PC and DIC image datasets were from, respectively. The size of the pink circle in each country is proportional to the number of participants from the corresponding country. **c**, The number of images in the training set. **d**, The number of labeled cells in the training set. **e**, Randomly selected examples (from left to right, brightfield, fluorescent, PC and DIC images) from the training set (first row) and testing set (second row). **f**, The number of images in the testing set. **g**, The number of cells in the testing set. There are two fluorescent WSIs in the testing set.

a significantly larger and more diverse microscopy image dataset. Leveraging this unprecedented dataset and a meticulously designed benchmark, Transformer-based deep-learning models emerged as exceptional algorithms and achieved notably superior performance.

Best-performing algorithm. Lee et al.[25] (T1-osilab) proposed a Transformer-based framework to harmonize model-centric and data-centric approaches. The model architecture used SegFormer[26] and the multiscale attention network[27] as the encoder and decoder. The SegFormer encoder was a hierarchical Transformer, enabling the extraction of both coarse and fine-grained features. The decoder contained position-wise attention blocks and multiscale fusion attention blocks for feature map fusion. The model output comprised two separate heads for cell recognition and distinction, originally proposed elsewhere[11]. The model underwent a two-step training process. It was first pretrained on public microscopy images and then fine-tuned on the challenge dataset with cell-aware data augmentation. Additionally, cell memory reply[28], concatenating the images from the pretraining and fine-tuning datasets in each mini-batch, was used to avoid catastrophic forgetting during fine-tuning (Methods).

Second-best-performing algorithm. Lou et al.[29] (T2-sribdmed) first divided the images into four distinct categories based on low-level image features (for example, intensities) in an unsupervised way. Then, class-wise cell segmentation models were trained for each category. The model employed U-Net-like architecture where ConvNeXT[30] was used as the building blocks. To address the diverse cell morphologies, two distinct decoder heads were employed. One decoder predicted the cell distance map and semantic map, effectively segmenting round-shaped cells, while the other decoder predicted the cell gradient map to handle cells with irregular shapes. The training process involved pretraining the model on the entire dataset, followed by fine-tuning on each of the four categories, resulting in the creation of four models. During inference, the image was initially classified into one of the four categories and subsequently, the corresponding model was used to perform the segmentation process (Methods).

Third-best-performing algorithm. Upschulte et al.[31] (T3-cells) designed an uncertainty-aware contour proposal network, employing ResNeXt-101 (ref. 32) to extract multiscale features from images, which were then processed through four decoder heads. A classification head

https://doi.org/10.1038/s41592-024-02233-6

Table 1 | Characteristics of the top three best-performing algorithms in preprocessing, data augmentation, network architecture and post-processing

Team	Preprocessing		Data augmentation				Network architecture		Post-processing
	IN	PS	IS	ED	UD	Others	Encoder backbone	Decoder heads	
T1-osilab[25]	✓	✓	✓	✓	–	Cell-wise intensity perturbation; boundary exclusion; oversample minor modality	SegFormer	Cell probability head; gradient fields regression head	Gradient tracking; exclude small cells; fill holes; TTA;
T2-sribdmed[29]	✓	✓	✓	✓	✓	–	ConvNeXt	Cell probability head; radial distance head; gradient fields regression head	NMS; watershed
T3-cells[31]	✓	✓	✓	✓	✓	Cell-aware rescaling	ResNeXt-101	Classification head; contour regression head; local refinement regression head; boundary uncertainty estimation head	NMS; convert contours to masks; region growing

IN, intensity normalization; PS, patch sampling; IS, intensity and spatial data augmentation; ED, external dataset; UD, unlabeled data; NMS, nonmaximum suppression; TTA, test-time augmentation. – indicates not used.

identified potential cell locations, while a contour regression head predicted sparse cell contours. To further improve accuracy, a refinement regression head was employed to revise the pixels within the cell contour. In addition, they incorporated an uncertainty head to estimate prediction confidence, which played a crucial role in the nonmaximum suppression post-processing. This incorporation of uncertainty information effectively facilitated the removal of redundant contour proposals and enhanced segmentation accuracy (Methods).

Other strategies. Table 1 summarizes the strategies employed by the top three teams. Given the considerable variation in image intensity and size across different modalities, all teams adopted intensity normalization techniques (for example, scaling the intensity to [0, 255]) during preprocessing and opted for patch-based sampling for model training. To enhance the model generalization ability and mitigate the risk of overfitting, diverse data augmentation methods were utilized. In addition to using external datasets, teams T2 and T3 leveraged the unlabeled data for model pretraining. Despite the common adoption of an encoder–decoder framework to construct networks, the top teams showcased variations in their choice of backbone networks and decoder heads. Consequently, the corresponding post-processing methods exhibited diversity.

Next, we present the quantitative results of the 28 algorithms on the holdout testing set. Figure 3a (Supplementary Table 4) shows a comparative view of F1 scores across 28 algorithms on the testing set. The scores are presented in the form of a dot and box plot, offering insights into both their central tendency and dispersion of the scores. The top three algorithms surpass other algorithms by a clear margin, resulting in median F1 scores of 89.7% (interquartile range (IQR) 84.1–94.8%), 84.5% (IQR 70.6–92.3%) and 84.4% (IQR 77.4–91.1%), respectively. Of particular note is the performance of the winning algorithm (T1-osilab). It stands apart not merely for its superior median F1 score, but also for the reduced number of outliers in its score distribution, suggesting a heightened level of robustness in its performance.

The bubble plot (Fig. 3b) presents the median F1 score, running time and the maximum GPU memory consumption of 28 algorithms, which can provide insights into the tradeoff between algorithm accuracy and efficiency. Most algorithms optimized the efficiency, enabling them to finish the inference within 13 s. It is essential to mention that this time metric also included the Docker starting time, hence the actual inference time is considerably shorter. For instance, the best-performing algorithm (T1-osilab) achieved an inference time of approximately 2 s for an image size of 1,000 × 1,000. Additionally, the median maximum GPU memory consumption was 3,099 MB (approximately $500), suggesting that these algorithms are affordable for practical deployment. This favorable combination of accuracy and efficiency makes them well suited for real-world applications in biological image analysis.

We also performed a statistical significance analysis for the 28 algorithms (Fig. 3c). Each team was compared to the other teams based on the one-sided Wilcoxon signed-rank test. Yellow shading indicates that the F1 scores of the algorithm on the x axis are significantly superior ($P < 0.05$) to those from the algorithm on the y axis, whereas blue shading indicates no significant difference between the two algorithms. The winning algorithm is significantly better than all the others. The second algorithm and the third algorithm obtain comparable performances with no significant differences, but they are significantly superior to other teams.

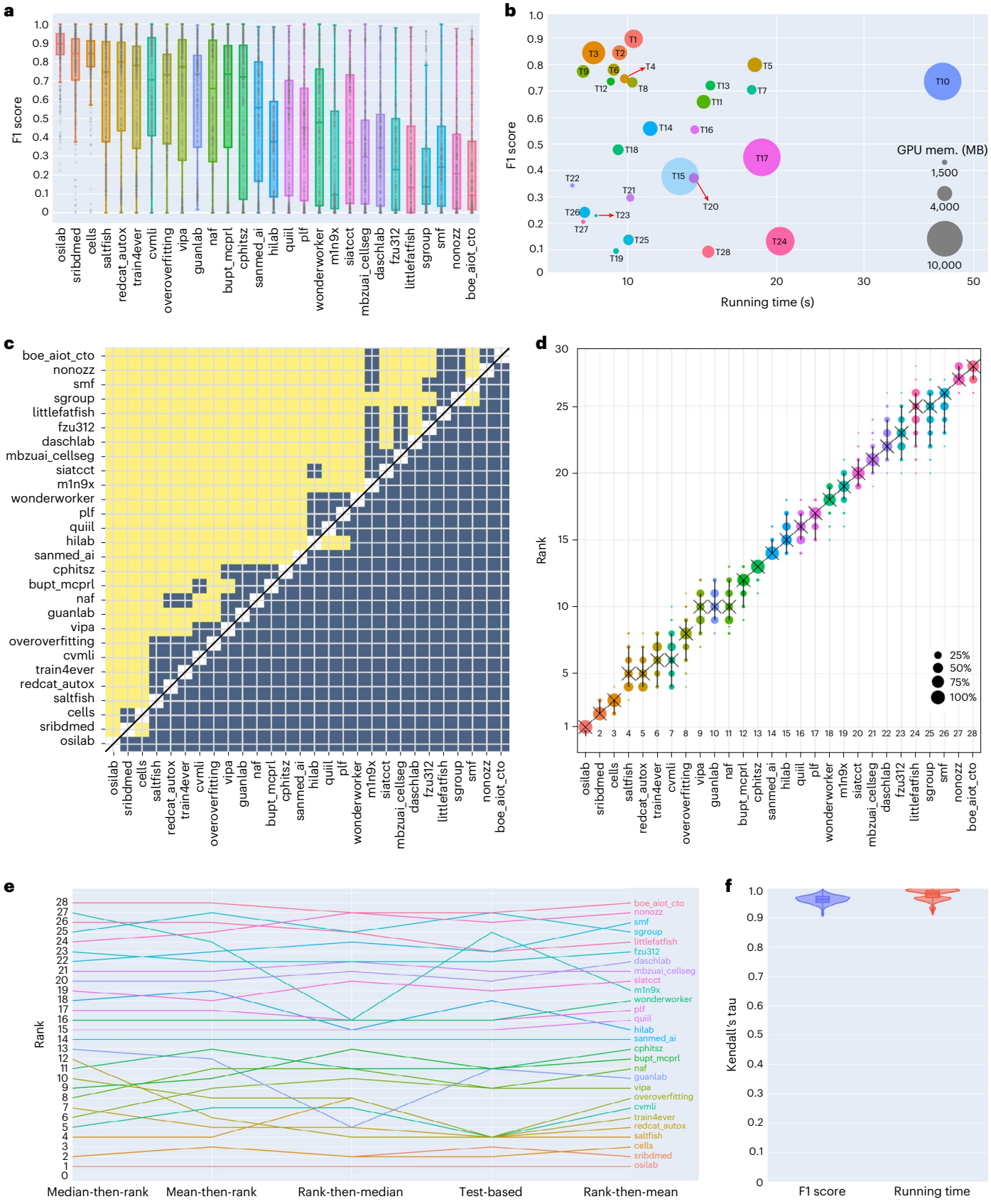
Furthermore, we analyzed the ranking stability based on bootstrap sampling (1,000 times). The results are visualized by blob plot (Fig. 3d). The blob area is proportional to the relative frequency of achieved ranks across the bootstrap samples and the median rank for each algorithm is indicated by a black cross. The winning algorithm has a blob area of 100%, indicating that it outperforms other algorithms in all the bootstrap samples. The second and third-best-performing algorithms still obtain better rank than other algorithms with a clear gap, whereas the second-best-performing algorithm has a lower median rank than the third-best-performing algorithm. Moreover, we compared the ranks of the 28 algorithms based on different ranking schemes (Fig. 3e): median-then-rank, mean-then-rank, rank-then-median, statistical significance test-based ranking and rank-then-mean (Methods). The winning algorithm consistently ranks first place across all the ranking schemes, whereas most of the other teams have fluctuations in their rank.

Finally, we analyzed the ranking stability of the employed metrics. The ranking list based on the full testing set is pairwise compared to the ranking lists based on the individual sample in the 1,000 bootstrap samples. Kendall's tau correlation is computed as a quantitative metric (Fig. 3f and Extended Data Fig. 2). It can be seen that Kendall's τ scores are very close to 1 for both F1 scores and running time, indicating a high degree of ranking agreement. Additionally, the compact distributions of these scores further confirm the stability of the ranking results with respect to sampling variability. These findings provide robust evidence that the obtained rankings are highly consistent and reliable across different samples.

The best-performing algorithms outperform state-of-the-art cell segmentation algorithms

To demonstrate the advancement of the winning algorithm beyond the state of the art (SOTA) in cell segmentation, we conducted a comparative analysis involving the top three algorithms from our challenge, the leading algorithm KIT-GE[33] from the CTC cell segmentation task and

https://doi.org/10.1038/s41592-024-02233-6



two widely recognized pretrained generalist models, Cellpose[11] and Omnipose[8]. The comparisons also included model variants of Cellpose and Omnipose that were trained from scratch on our challenge dataset. The aim was to determine whether the performance improvement mainly resulted from the training set. Recognizing the importance of transfer-learning-based algorithms in achieving a universal solution, we further collected a new external testing set with 157 diverse yeast and bacteria cell images (Methods and Supplementary Table 4)

Fig. 3 | Evaluation results of 28 algorithms on the holdout testing set. a, Dot and box plot of the F1 scores on the testing set ($n = 422$ independent images). The box plots display descriptive statistics across all testing cases, with the median value represented by the horizontal line within the box, the lower and upper quartiles delineating the borders of the box and the vertical black lines indicating 1.5 × IQR. **b**, The top algorithms achieve a good tradeoff between segmentation accuracy (*y* axis) and efficiency (*x* axis). The circle size is proportional to GPU memory consumption. **c**, Pairwise significant test results (one-sided Wilcoxon signed-rank test) show that the winning algorithm is significantly better than the other algorithms. **d**, Blob plot for visualizing ranking stability based on bootstrap sampling. The median area of each blob is proportional to the relative frequency of achieved ranks across 1,000 bootstrap samples. The median rank for each algorithm is indicated by a black cross. The 95% bootstrap intervals across bootstrap samples are indicated by black lines. **e**, The winning algorithm holds the first place across five different ranking schemes. **f**, High Kendall's tau scores indicate that the ranking results are stable. The violin plot shows descriptive statistics with the median value represented by the horizontal solid line within the box, the mean value represented by the horizontal dashed line, the lower and upper quartiles delineating the borders of the box and the vertical black lines indicating 1.5 × IQR.

to thoroughly compare the top three algorithms to the fine-tuned Cellpose (Cellpose 2.0, ref. 12) and Omnipose models. For both Cellpose and Omnipose, we used the 'cyto2' model checkpoints, recognized for their exceptional generalizability, as the pretrained model and the foundation for further fine-tuning.

Figure 4a (Supplementary Table 6) illustrates the F1 scores of these eight methods on the testing set, revealing that the top three best-performing algorithms achieved significantly higher accuracy than the existing SOTA algorithms. Specifically, The T1 algorithms achieved a median F1 score of 89.7% (IQR 36.7–82.4%), surpassing the KIT-GE, Cellpose-pretrain, Cellpose-scratch, Omnipose-pretrain and Omnipose-scratch by 49.9%, 24.4%, 35.4%, 58.9% and 48.7%, respectively.

Figure 4b presents the results on the brightfield images, where the top two best-performing algorithms remained at the forefront, achieving median F1 scores of 91.4% (IQR 88.0–94.9%) and 91.0% (IQR 86.1–93.8%), respectively. The Cellpose-finetune model exhibited comparable performance to the third-best-performing algorithm, achieving a significant improvement of 16.6% in median F1 score over the Cellpose-pretrain model, as anticipated due to its training on the challenge dataset. Figure 4c shows the results on the fluorescent images, where the third-best-performing algorithm outperformed others with a median F1 score of 80.8% (IQR 71.6–91.5%), followed by the best-performing algorithm and the Cellpose-pretrain model; however, the F1 score of Cellpose-scratch and Omnipose-scratch declined substantially by 44.1% and 7.4%, respectively. This decrease can be attributed to the testing images being from new cell types not present in the training set.

In Fig. 4d, the results for PC images demonstrated that the top three best-performing algorithms maintained their superiority in this category, achieving median F1 scores of 93.6% (IQR 87.9–96.4%), 88.8% (77.7–96.4%) and 90.3% (84.3–95.0%), respectively. The CTC Challenge's top-performing segmentation algorithm, KIT-GE, excelled in PC images due to its design for label-free images and the relatively simple segmentation of round-shaped cells. Figure 4e shows the results on DIC images with the top three best-performing algorithms once again achieving the highest performance, achieving median F1 scores of 86.8% (IQR 83.5–88.0%), 75.0% (IQR 68.9–78.1%) and 80.3% (77.2–85.1%), respectively. While Omnipose-scratch yielded the best performance among the SOTA methods with a median F1 score of 43.4% (IQR 33.1–60.9%), it still fell significantly behind the top three best-performing methods. Conversely, KIT-GE and Cellpose struggled in this category, because the DIC testing images were from new biological experiments and exhibited very low contrast.

We further visualized segmentation examples of the seven algorithms to gain insights into their characteristics (Fig. 4f and Extended Data Fig. 1). The top three best-performing algorithms demonstrated relatively robust results, with the best-performing algorithm (T1-osilab) displaying exceptional accuracy across diverse microscope types, cell types and image contrasts. Notably, KIT-GE exhibited better performance on PC images than stained images, as it was designed based on a label-free challenge dataset. Nevertheless, KIT-GE struggled to segment other images from new biological experiments, indicating limited generalization ability in this context. The Cellpose models outperformed Omnipose models on most images, except for DIC images featuring numerous small objects with low contrasts. Additionally, the Cellpose-scratch model surpassed Cellpose-pretrain on brightfield images, exhibiting fewer segmentation errors; however, its performance decreased on other modalities that contained previously unseen images, leading to an increased number of missed cells in the segmentation results.

Finally, we conducted a post-challenge analysis by evaluating the top three algorithms, KIT-GE and three variants each of Cellpose and Omnipose models (pretrained, trained from scratch and fine-tuned by transfer learning) using a new testing set comprising unseen images (Methods). As shown in Fig. 4g (Supplementary Table 7 and Fig. 2), the top three algorithms outperformed others, achieving median F1 scores of 95.0% (IQR 93.0–97.9%), 88.7% (IQR 72.1–93.9%) and 93.3% (IQR 83.7–96.7%), respectively. Notably, the fine-tuned Cellpose and Omnipose models surpassed their scratch-trained counterparts by 12.2% and 11.7%, respectively, demonstrating the value of previously learned features in new learning contexts; however, their performances were still lower than the original pretrained models. This discrepancy is largely attributed to the testing images originating from new sources, leading to a case of catastrophic forgetting during the fine-tuning process, a common phenomenon in transfer learning[34,35].

Discussion

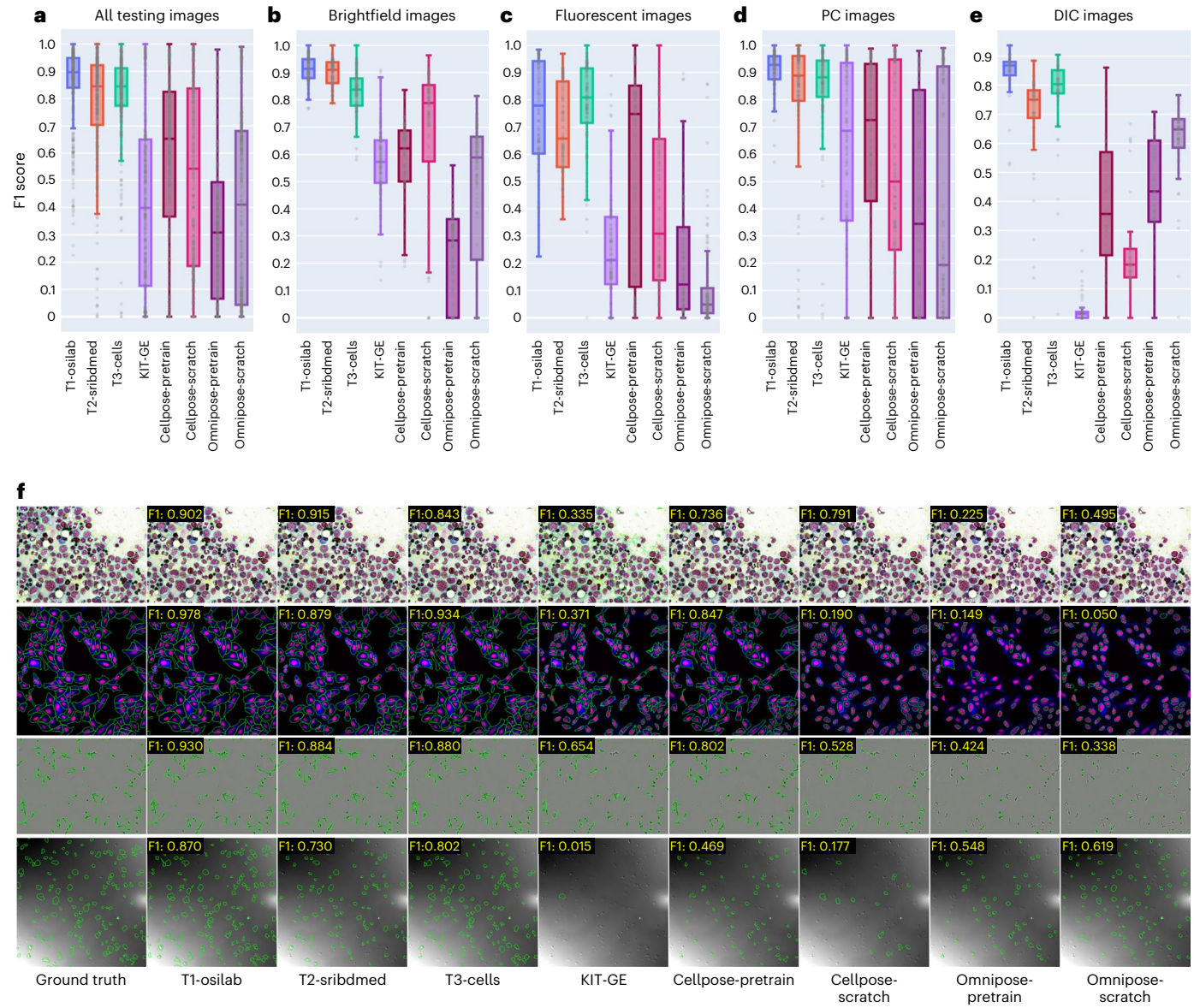
The primary and arguably most notable observation in this challenge is the unequivocal superiority of the Transformer-based algorithm, which exhibited significantly enhanced performance compared to existing SOTA cell segmentation algorithms. Transformers offer several unique advantages compared to CNNs. First, Transformers[22] use self-attention mechanisms that can capture global context and

Fig. 4 | Quantitative and qualitative comparison between the top three algorithms and SOTA generalist cell segmentation algorithms: KIT-GE (top solution in the segmentation benchmark of the CTC), Cellpose, Omnipose and their variants under different training strategies. a–**e**, Dot and box plot of the F1 scores on the whole testing test ($n = 422$ independent images) (**a**); brightfield images ($n = 120$ independent images) (**b**); fluorescent images ($n = 122$ independent images) (**c**); PC images ($n = 120$ independent images) (**d**); and DIC images ($n = 60$ independent images) (**e**). The box plots display descriptive statistics with the median value represented by the horizontal line within the box, the lower and upper quartiles delineating the borders of the box and the vertical black lines indicating 1.5 × IQR. **f**, Example segmentation results of the four microscopy image modalities, brightfield, fluorescent, PC and DIC images (from top to bottom). **g**, Quantitative comparison on the post-challenge testing set ($n = 157$). The box plot shows descriptive statistics across the post-challenge testing cases, with the median value represented by the horizontal line within the box, the lower and upper quartiles delineating the borders of the box and the vertical black lines indicating 1.5 × IQR. Cellpose-pretrain, Cellpose pretrained model ('cyto2'); Cellpose-scratch, Cellpose model trained from scratch on the challenge dataset; Cellpose-fine-tune, Cellpose fine-tuned model on the challenge dataset; Omnipose-pretrain, Omnipose pretrained model ('cyto2'); Omnipose-scratch, Omnipose model trained from scratch on the challenge dataset; Omnipose-finetune, Omnipose fine-tuned model on the challenge dataset.

https://doi.org/10.1038/s41592-024-02233-6

long-range dependencies in images, whereas CNNs usually process local image patches. Second, Transformers have a larger model capacity than CNNs[23], enabling them to learn intricate patterns and model nuanced relationships in images, which are essential for accurate cell segmentation. Third, Transformers excel in transfer-learning settings, allowing the model to pretrain on large datasets and subsequently fine-tune specific downstream tasks or new datasets with limited annotations. Notably, this effective strategy was also successfully adopted by the winning algorithm.

The winning algorithm demonstrated a remarkable level of superiority compared to the leading algorithm from the CTC Challenge, even after the latter was retrained on our dataset. This notable improvement can be attributed to the unparalleled diversity of the dataset. Unlike the CTC Challenge dataset, which only comprised label-free images,



our challenge dataset encompassed both labeled and label-free images. Furthermore, our challenge focused on universal segmentation algorithms, whereas the top-performing teams of the CTC Challenge developed tailored models for each dataset[36,37]. This fundamental difference in strategy likely contributed to the substantial performance gap.

In addition to the Transformer-based architecture, we also identified several useful strategies for achieving top performance. First, different from the typical detection-then-segmentation paradigm[38], multihead outputs were employed by most of the top algorithms[25,29,31], which converted the instance segmentation task into distance map regression tasks and a cell foreground semantic segmentation task, followed by post-processing to merge the output as instance labels. This approach was conclusively demonstrated to be superior to the conventional detection-then-segmentation paradigm in this challenge. Another crucial aspect was the adoption of diverse and robust data augmentation techniques, which is important to improve the model generalization ability and reduce overfitting. In addition to commonly used global intensity augmentations (for example, scaling, noise addition and blurring) and spatial augmentations (for example, rotation, zooming and flipping), participants introduced innovative augmentation methods. For example, Lee et al.[25] employed cell-wise random perturbations in image intensity, whereas Li et al.[39] used Mosaic data augmentation[40], enabling the model to learn object identification at varying scales. Moreover, employing efficient backbone networks, such as SegFormer[26] and ConvNext[30], offered a favorable accuracy–efficiency tradeoff. The winning algorithm also demonstrated that the slide-window-based method was an efficient strategy for scalable inference (Supplementary Table 8). Specifically, the input image was partitioned into multiple smaller patches and their predictions were subsequently stitched together to form the final label map. This method proved particularly crucial for whole-slide image segmentation, considering the inherent limitations of RAM and GPU memory in real practice.

Additionally, all the top three teams explored the potential of leveraging the unlabeled images to improve the segmentation performance. Specifically, T1-osilab[25] employed consistency regularization[41] to match the algorithm's predictions on the clean and degraded unlabeled images and introduced an additional head module to reconstruct the unlabeled images[42]. Both T1-osilab[25] and T2-sribdmed[29] investigated pseudo-label learning, generating pseudo-labels for unlabeled images using trained models, followed by training the network with both pseudo-labels and ground-truth annotations. T3-cells[31] implemented the uncertainty-aware Listen2Student mechanism[43] to train a student network with low-uncertainty pseudo-labels; however, despite these joint efforts, none of the employed methods demonstrated a notable enhancement in segmentation performance. Thus, it remains an open question how to effectively use unlabeled data to boost cell segmentation performance.

Furthermore, we made a noteworthy observation concerning the commonly employed transfer-learning algorithm, which exhibited a phenomenon known as catastrophic forgetting[35]. The original Cellpose and Omnipose generalist models, pretrained on a diverse array of microscopy images, demonstrated the ability to generalize to a portion of the unseen testing images; however, their fine-tuned counterparts, exhibited a notable performance degradation, as they could only segment images present in the training set, while losing previously learned capability to handle unseen images. The winning algorithm addressed this issue by implementing a simple yet effective strategy known as cell memory replay[28], aiming to relearn the existing data during fine-tuning. More specifically, the fine-tuning procedure involved combining images from both the existing dataset and the new dataset as a mini-batch for training the model, allowing the algorithm to retain its competence in handling both known and new images.

To promote the widespread applicability of the new SOTA algorithms, all top-performing teams have made their algorithms publicly available on GitHub, complete with comprehensive preprocessing, training and testing code; however, a critical challenge remains in bridging the gap between these advanced algorithms and their seamless integration into daily biological practice, as it often demands a basic level of computational expertise to apply these algorithms to new images successfully. To bridge this gap, we invited the top three best-performing teams to integrate their algorithms into Napari[44], an open-source interface specifically designed for user-friendly biological image visualization and analysis. In this way, users gain convenient access to these high-performing algorithms, enabling them to effortlessly apply the segmentation techniques to their own images without necessitating additional coding. Furthermore, to facilitate even greater accessibility and ease of use, the algorithm Docker containers were thoughtfully released. This strategic move empowers users to perform batch image segmentation with utmost simplicity, as a single-line command suffices to initiate the process.

These new cell segmentation algorithms have many potential applications in various biological tasks. For example, cell segmentation in mass cytometry imaging, as demonstrated by Jackson et al.[1], was pivotal in characterizing cellular phenotypes in breast tumor tissues. These phenotypes, aligning closely with pathologist-assigned tumor grades, revealed complex multicellular structures. Similarly, cell segmentation played a crucial role in quantifying molecules at a single-cell level, as seen in the work of Capolupo et al.[2], leading to the discovery of new regulatory mechanisms in dermal fibroblasts. Additionally, the application of cell segmentation in disease progression studies, such as those by Risom et al.[45], has been instrumental in characterizing cancer microenvironments in multiplexed ion beam imaging by time of flight of tissue microarrays.

This work has certain limitations. While the challenge dataset was indeed diverse, it was confined to two-dimensional (2D) microscopy images due to the available datasets; however, three-dimensional (3D) microscopy images are becoming increasingly prevalent[46], which pose new segmentation challenges, such as the large-scale volume and anisotropic resolutions. Additionally, while the integration of napari cell segmentation interfaces has improved accessibility for biologists, these algorithms currently do not support interactive feedback from users. Furthermore, the scope of the challenge was restricted to segmentation tasks, omitting classification tasks. Future endeavors should aim to broaden the benchmark to include more complex 3D images, coupled with classification tasks. There is also a compelling need to develop a biologist-in-the-loop system, enabling more effective collaboration between algorithms and human experts.

In conclusion, the challenge results present a successful proof of concept of generalist cell segmentation algorithms, benefiting from the collective expertise of both biological imaging and machine learning experts. The Transformer-based algorithm surpassed previous SOTA methods by a large margin, which can efficiently generate accurate cell contours on a wide range of microscopy images without user intervention. Furthermore, the top algorithms have been made open-source and seamlessly integrated into user-friendly interfaces. This integration holds great potential for accelerating microscopy image analysis throughput and fostering new discoveries in quantitative biological research. We aim to establish this challenge as a sustainable benchmark platform and we enthusiastically welcome contributions of various new data to expand data diversity, paving the way for continuous advancement in this vital field.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-024-02233-6.

References

1. Jackson, H. W. et al. The single-cell pathology landscape of breast cancer. *Nature* **578**, 615–620 (2020).
2. Capolupo, L. et al. Sphingolipids control dermal fibroblast heterogeneity. *Science* **376**, eabh1623 (2022).
3. Lin, J.-R. et al. Multiplexed 3d atlas of state transitions and immune interaction in colorectal cancer. *Cell* **186**, 363–381 (2023).
4. Hollandi, R. et al. Nucleus segmentation: towards automated solutions. *Trends Cell Biol.* **32**, 295–310 (2022).
5. Greenwald, N. F. et al. Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning. *Nat. Biotechnol.* **40**, 555–565 (2021).
6. Lee, M. Y. et al. Cellseg: a robust, pre-trained nucleus segmentation and pixel quantification software for highly multiplexed fluorescence images. *BMC Bioinform.* **23**, 1–17 (2022).
7. Kempster, C. et al. Fully automated platelet differential interference contrast image analysis via deep learning. *Sci. Rep.* **12**, 1–13 (2022).
8. Cutler, K. J. et al. Omnipose: a high-precision morphology-independent solution for bacterial cell segmentation. *Nat. Meth.* **19**, 1438–1448 (2022).
9. Bunk, D. et al. Yeastmate: neural network-assisted segmentation of mating and budding events in *Saccharomyces cerevisiae*. *Bioinformatics* **38**, 2667–2669 (2022).
10. Dietler, N. et al. A convolutional neural network segments yeast microscopy images with high accuracy. *Nat. Commun.* **11**, 1–8 (2020).
11. Stringer, C., Wang, T., Michaelos, M. & Pachitariu, M. Cellpose: a generalist algorithm for cellular segmentation. *Nat. Meth.* **18**, 100–106 (2021).
12. Pachitariu, M. & Stringer, C. Cellpose 2.0: how to train your own model. *Nat. Meth.* **19**, 1634–1641 (2022).
13. Ulman, V. et al. An objective comparison of cell-tracking algorithms. *Nat. Meth.* **14**, 1141–1152 (2017).
14. Maška, M. et al. The cell tracking challenge: 10 years of objective benchmarking. *Nat. Meth.* **20**, 1010–1020 (2023).
15. Caicedo, J. C. et al. Nucleus segmentation across imaging experiments: the 2018 data science bowl. *Nat. Meth.* **16**, 1247–1253 (2019).
16. Graham, S. et al. CoNIC challenge: pushing the frontiers of nuclear detection, segmentation, classification and counting. *Med. Image Anal.* **92**, 103047 (2024).
17. Tajbakhsh, N. et al. Embracing imperfect datasets: a review of deep learning solutions for medical image segmentation. *Med. Image Anal.* **63**, 101693 (2020).
18. Ma, J. & Wang, B. Towards foundation models of biological image segmentation. *Nat. Meth.* **20**, 953–955 (2023).
19. Gupta, A. et al. Segpc-2021: a challenge & dataset on segmentation of multiple myeloma plasma cells from microscopic images. *Med. Image Anal.* **83**, 102677 (2023).
20. Falk, T. et al. U-net: deep learning for cell counting, detection, and morphometry. *Nat. Meth.* **16**, 67–70 (2019).
21. Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K. & Yuille, A. L. Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **40**, 834–848 (2017).
22. Vaswani, A. et al. Attention is all you need. in *Advances in Neural Information Processing Systems*, vol. 30 (NeurIPS, 2017).
23. Dosovitskiy, A. et al. An image is worth 16×16 words: Transformers for image recognition at scale. in *International Conference on Learning Representations* (ICLR, 2021).
24. Ma, J. et al. Segment anything in medical images. *Nat. Commun.* **15**, 654 (2024).
25. Lee, G., Kim, S., Kim, J. & Yun, S.-Y. Mediar: harmony of data-centric and model-centric for multi-modality microscopy. in *Proceedings of The Cell Segmentation Challenge in Multi-modality High-Resolution Microscopy Images*, vol. 212, pages 1–16 (2023).
26. Xie, E. et al. Segformer: simple and efficient design for semantic segmentation with transformers. in *Advances in Neural Information Processing Systems*, vol. 34 (NeurIPS, 2021).
27. Fan, T., Wang, G., Li, Y. & Wang, H. Ma-net: a multi-scale attention network for liver and tumor segmentation. *IEEE Access* **8**, 179656–179665 (2020).
28. Chaudhry, A., Gordo, A., Dokania, P., Torr, P. & Lopez-Paz, D. Using hindsight to anchor past knowledge in continual learning. in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pages 6993–7001 (AAAI, 2021).
29. Lou, W. et al. Multi-stream cell segmentation with low-level cues for multi-modality images. *Proc. Mach. Learn. Res.* **212**, 1–10 (2023).
30. Liu, Z. et al. A convnet for the 2020s. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11976–11986 (IEEE, 2022).
31. Upschulte, E., Harmeling, S., Amunts, K. & Dickscheid, T. Uncertainty-aware contour proposal networks for cell segmentation in multi-modality high-resolution microscopy images. *Proc. Mach. Learn Res.* **212**, 1–12 (2023).
32. Xie, S., Girshick, R., Dollár, P., Tu, Z. & He, K. Aggregated residual transformations for deep neural networks. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1492–1500 (IEEE, 2017).
33. Scherr, T., Löffler, K., Böhland, M. & Mikut, R. Cell segmentation and tracking using cnn-based distance predictions and a graph-based matching strategy. *PLoS ONE* **15**, e0243219 (2020).
34. Parisi, G. I., Kemker, R., Part, J. L., Kanan, C. & Wermter, S. Continual lifelong learning with neural networks: a review. *Neural Netw.* **113**, 54–71 (2019).
35. De Lange, M. et al. A continual learning survey: defying forgetting in classification tasks. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 3366–3385 (2021).
36. Pena, F. A. G. et al. J regularization improves imbalanced multiclass segmentation. in *IEEE 17th International Symposium on Biomedical Imaging*, 1–5 (IEEE, 2020).
37. Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J. & Maier-Hein, K. H. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. Meth.* **18**, 203–211 (2021).
38. He, K., Gkioxari, G., Dollár, P. & Girshick, R. Mask r-cnn. in *Proceedings of the IEEE International Conference on Computer Vision*, 2961–2969 (IEEE, 2017).
39. Wangkai, L. et al. Maunet: modality-aware anti-ambiguity u-net for multi-modality cell segmentation. *Proc. Mach. Learn Res.* **212**, 1–12 (2023).
40. Bochkovskiy, A., Wang, C.-Y. & Liao, H.-Y. M. Yolov4: optimal speed and accuracy of object detection. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2004.10934 (2020).
41. Jeong, J., Lee, S., Kim, J. & Kwak, N. Consistency-based semi-supervised learning for object detection. in *Advances in Neural Information Processing Systems*, vol. 32 (NeurIPS, 2019).
42. Chen, S., Bortsova, G., Juárez, A.G.-U., Van Tulder, G. & De Bruijne, M. Multi-task attention-based semi-supervised learning for medical image segmentation. in *Medical Image Computing and Computer Assisted Intervention*, 457–465 (MICCAI, 2019).
43. Liu, Y.-C., Ma, C.-Y. & Kira, Z. Unbiased teacher v2: semi-supervised object detection for anchor-free and anchor-based detectors. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9819–9828 (IEEE, 2022).
44. Sofroniew, N. et al. napari: a multi-dimensional image viewer for Python. *Zenodo* https://zenodo.org/10.5281/zenodo.3555620 (2022).

45. Risom, T. et al. Transition to invasive breast cancer is associated with progressive changes in the structure and composition of tumor stroma. *Cell* **185**, 299–310 (2022).
46. Fu, S. et al. Field-dependent deep learning enables high-throughput whole-cell 3D super-resolution imaging. *Nat. Meth.* **20**, 459–468 (2023).

[1]Peter Munk Cardiac Centre, University Health Network, Toronto, Ontario, Canada. [2]Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Ontario, Canada. [3]Vector Institute, Toronto, Ontario, Canada. [4]Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada. [5]Donnelly Centre, University of Toronto, Toronto, Ontario, Canada. [6]Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario, Canada. [7]School of Medicine and Pharmacy, Ocean University of China, Qingdao, China. [8]Department of Electronics and Communications Engineering, Indraprastha Institute of Information Technology Delhi (IIITD), New Delhi, India. [9]Laboratory Oncology Unit, Dr. BRAIRCH, All India Institute of Medical Sciences, New Delhi, India. [10]Department of Image Reconstruction, Nanjing Anke Medical Technology Co., Nanjing, China. [11]Shanghai Artificial Intelligence Laboratory, Shanghai, China. [12]Graduate School of AI, KAIST, Seoul, South Korea. [13]Shenzhen Research Institute of Big Data, Shenzhen, China. [14]Chinese University of Hong Kong (Shenzhen), Shenzhen, China. [15]Institute of Neuroscience and Medicine (INM-1) and Helmholtz AI, Research Center Jülich, Jülich, Germany. [16]Faculty of Mathematics and Natural Sciences - Institute of Computer Science, Heinrich Heine University Düsseldorf, Düsseldorf, Germany. [17]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Hinxton, UK. [18]Champalimaud Foundation - Centre for the Unknown, Lisbon, Portugal. [19]Department of Bioengineering, Stanford University, Palo Alto, CA, USA. [20]Tandon School of Engineering, New York University, New York, NY, USA. [21]School of Biomedical Engineering, Health Science Center, Shenzhen University, Shenzhen, China. [22]Laboratory of the Physics of Biological Systems, Institute of Physics, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland. [23]School of Biological Sciences, University of Reading, Reading, UK. [24]Laboratoire de Chimie Bactérienne, CNRS-Université Aix-Marseille UMR, Institut de Microbiologie de la Méditerranée, Marseille, France. [25]Department of Internal Medicine I, University Hospital Dresden, Technical University Dresden, Dresden, Germany. [26]Else Kroener Fresenius Center for Digital Health, Technical University Dresden, Dresden, Germany. [27]Department of Automation, University of Science and Technology of China, Hefei, China. [28]Institute of Advanced Technology, University of Science and Technology of China, Hefei, China. [29]Department of Computer Science and Technology, Nanjing University, Nanjing, China. [30]School of EECS, The University of Queensland, Brisbane, Queensland, Australia. [31]School of Medicine, Stanford University, Palo Alto, CA, USA. [32]Division of Computing and Mathematical Science, Caltech, Pasadena, CA, USA. [33]Howard Hughes Medical Institute, Chevy Chase, MD, USA. [34]Imaging Platform, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [35]Department of Computer Science, University of Waterloo, Waterloo, Ontario, Canada. [36]Hematoscope Laboratory, Comprehensive Cancer Center & Center of Diagnostics, Helsinki University Hospital, Helsinki, Finland. [37]Department of Oncology, University of Helsinki, Helsinki, Finland. [38]Department of Computer Science, University of Toronto, Toronto, Ontario, Canada. [39]Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, Ontario, Canada. [40]CIFAR Multiscale Human Program, CIFAR, Toronto, Ontario, Canada. [41]UHN AI Hub, University Health Network, Toronto, Ontario, Canada. [42]These authors contributed equally: Shamini Ayyadhury, Cheng Ge, Anubha Gupta, Ritu Gupta, Song Gu, Yao Zhang.
✉e-mail: bowang@vectorinstitute.ai

Methods

Challenge organization

This challenge was preregistered in the Thirty-sixth Conference on Neural Information Processing Systems (NeurIPS) (https://neurips.cc/Conferences/2022/CompetitionTrack) following a peer review process conducted by the competition program committee comprising experts in both machine learning and challenges organization. The challenge was officially launched on 15 June 2022 and ran for 139 days until 31 October 2022, marking the testing submission deadline. Throughout the development phase, participants were given the opportunity to submit their tuning set segmentation results on the challenge platform and obtain corresponding F1 scores. Moreover, to minimize entry barriers, we supplied a U-Net-based baseline model and offered a step-by-step tutorial to assist participants in becoming familiar with model training, inference and submission. Additionally, we furnished guidelines and suggestions on SOTA cell segmentation methods[5,8,11,12], empowering participants to surpass the baseline and achieve higher levels of performance.

Dataset curation and preprocessing

The images were curated from multiple laboratories, each specializing in different cell types and modalities. Five labeled datasets were obtained from publicly available datasets with proper license permits or approval from the respective authors. Two public datasets contained a subset of labeled images and we augmented them with complete cell annotations for the previously unlabeled images. The annotation process was conducted in the original data center by our author team. Eight public datasets lacked annotations and we generated cell annotations for the challenge. All testing images were newly acquired by ourselves for this challenge. It should be noted that the testing images remained unavailable to the public during the challenge, avoiding potential data or label leakage. In addition to the primary testing set, we further collected a new batch of yeast (for example, *clb1-6* Δ conditional mutants and pseudohyphal cells) and bacteria images (for example, *Myxococcus xanthus* and *Ruminiclostridium cellulolyticum*) for post-challenge analysis, aiming to evaluate the algorithms' accuracy and robustness when applied to unseen cells exhibiting a diverse range of appearances and morphologies. Detailed sources can be found in Supplementary Tables 1–4. We also presented the distribution of image size in Extended Data Fig. 3. The majority of the images are microscope patches, but several WSIs are also provided in each set. The training (labeled) set, tuning set and testing set exhibit a similar median image size, approximately one million pixels (1,000 × 1,000).

The original image formats included png, bmp, jpg, tif, tiff, npy and npz files. The npy and npz formats, which are not typical image formats, were converted to the widely used png format. All other image formats were retained as they were, to accommodate the diverse array of image formats that the developed algorithms might encounter. External datasets and pretrained models were allowed, but participants were asked to post the corresponding links to the competition forum and we also maintained a document of external datasets on the challenge homepage to make sure that these external datasets were available to all participants.

For the labeled dataset, all cells were annotated in each image, with the exception of red blood cells in blood and bone-marrow slides, as biomedical researchers predominantly centered on the stained leukocytes. The annotation team consisted of two biologists with 10 years of experience, responsible for ensuring compliance with annotation requirements. In cases where data contributors provided cell annotations, the annotations were thoroughly checked and revised as needed. For contributors who provided unlabeled images, publicly available specialist models[5,7,9,10] were initially employed to generate predictions. The resulting segmentation outcomes were subsequently subjected to manual revision by the biologists. Additionally, to maintain the quality and reliability of the dataset, each image–annotation pair underwent stringent quality control. Images with fewer than five cells were excluded from the dataset and cells containing fewer than 15 pixels were also removed. In total, we created more than 900,000 new cell annotations for the challenge, which is significantly larger than the provided new annotations in the recent CoNIC Challenge[16] and CTC[14].

Top three best-performing algorithms

Best-performing algorithm. Lee et al.[25] (T1-osilab) incorporated model-centric and data-centric approaches to learn generalizable representations for heterogeneous microscopy image modalities and achieved a good tradeoff between model accuracy and efficiency. From the model-centric perspective, the framework adopted a typical encoder–decoder architecture to extract hierarchical features and integrate them through skip connections. Concretely, SegFormer[26] served as the encoder, while MA-Net[27] was employed as the decoder, utilizing the Mish[47] activation function. The network jointly predicted cell probability maps and regressed cell-wise vertical and horizontal gradient flows, followed by a gradient tracking post-processing to separate touched cells, which was originally proposed in Cellpose[11]. From the data-centric perspective, they tailored two cell-aware augmentations to extensively enrich the diversity of the dataset and combined them with commonly used intensity and spatial augmentation methods to improve model generalization. Specifically, image intensities were randomized in a cell-wise manner and cell boundary pixels were excluded to separate the crowded cells. Moreover, a two-phase pretraining and fine-tuning pipeline was used to retrain the knowledge from external datasets, including TissueNet[5], Omnipose[8], Cellpose[11] and LiveCell[48]. Furthermore, to address minor modalities, they were selected through unsupervised clustering with the latent embedding and subsequently over-sampled during training, thereby aiming to enhance the performance of these less-represented modalities.

The model inputs were three-channel images. The overall loss function was the combination of binary cross-entropy loss and mean squared error (MSE) loss. The inference process relied on a sliding window strategy, a highly efficient approach for processing WSIs. During the merging of predictions from these small window patches, an importance map was generated and applied to the predictions, thereby preventing the recognition of the same cells at the patch boundary as multiple cells. The comprehensive integration of these approaches resulted in exceptional performance, effectively handling diverse microscopy image modalities with high accuracy.

Second-best-performing algorithm. Lou et al.[29] (T2-sribdmed) designed a classification-and-segmentation framework that first classified the input image into one group and then performed cell segmentation with a model trained for that group. The classification pipeline consisted of three steps. First, it employed a segmentation model trained on labeled images to generate pseudo-labels for unlabeled images. Second, the images were classified into four groups based on image intensities. Specifically, the first class included all single-channel images. The three-channel RGB images were converted to hue, saturation and value color space. Within this transformed domain, images exhibiting a mean saturation (S) greater than 0.1 and a mean value (V) falling within the range [0.1, 0.6] were assigned to the second class. The remaining images with cell areas larger than 8,000 pixels were classified as the third class, while others were designated as the fourth class. Finally, a ResNet18 (ref. 49) was trained for automated group classification. The segmentation network followed a design of U-Net-like architecture, where the encoder was ConvNeXt[30]. Motivated by the observation that most cells in the first and second classes were roundish, a decoder with star-convex polygon-based cell representation[50] was integrated with the encoder for cell instance segmentation, termed as ConvNeXt-Stardist. NMS[51] was employed in the post-processing to remove duplicated predictions. For the third and fourth classes,

the prediction head in HoverNet[52] was adopted as the decoder, termed as ConvNeXt-Hover. The marker-based watershed algorithm was applied in the post-processing phase to separate touching cells.

There were four segmentation models in total trained for four image groups respectively. ConvNeXt-Stardist was trained with a combination of cross-entropy loss, Dice loss and MAE loss and ConvNeXt-Hover was trained with a combination of cross-entropy loss, Dice loss, MSE loss and mean squared gradient error loss. Both ConvNeXt-Stardist and ConvNeXt-Hover were pretrained on all images and fine-tuned on the images from the corresponding group. The model inputs were three-channel images. During inference, the input image was first classified into certain groups by the classification model and then processed by the segmentation model trained for the group.

Third-best-performing algorithm. Upschulte et al.[31] (T3-cells) proposed a contour proposal network (CPN)[53], which treated instance segmentation as a sparse detection problem by regressing object contours anchored at pixel locations. This enabled the model to handle multiple objects assigned to the same pixel and recover partially superimposed objects accurately. The shape-focused nature of contour representation learning also facilitated the development of inductive shape priors, potentially improving robustness in challenging conditions.

The CPN utilized the ResNeXt backbone network[32] to extract multiscale feature maps, a regression head to generate candidate contour representations for each pixel and a classification head to determine whether an object was present or not at these locations. A proposal sampling stage extracted a sparse list of contour representations, which were transformed into the pixel domain using differentiable Fourier transformation to encode contour information in the frequency domain[54]. The precision of the contours was further improved by using a displacement field generated by an additional regression head. In addition to the original CPN, this work introduced dedicated supervision for boundaries and proposed an extra branch to estimate localization uncertainty for boundaries. The multitask training objective was defined by a combination of the average absolute difference loss for contour regression, the generalized intersection over union (IoU) loss for boundary localization[55], the absolute L1 distance for local refinement[53], the distance loss for frequency regularization[53], the binary cross-entropy loss for classification and the negative power log-likelihood loss for uncertainty estimation[56].

The uncertainty-aware Listen2Student mechanism[43] was applied to incorporate unlabeled examples during training, where a teacher model generated bounding boxes as pseudo-labels to supervise the student model. The model inputs were three-channel images. For post-processing, the Vanilla NMS relying solely on the classification score might not reliably indicate the proposal's quality. To address this issue, the approach proposed in[56] was employed to incorporate uncertainty estimations into the NMS selection process. The object contours were transformed into segmentation masks through rasterization and region filling. A region-growing technique[57] was further adopted for overlapping regions.

Existing SOTA cell segmentation algorithms

The following methods were designed for gray and two-channel microscopy images, whereas the challenge dataset was curated for developing universal algorithms that were agnostic to different image channel formats. Thus, we preprocessed the challenge images to gray images using the 'skimage.color.rgb2gray' function.

Cellpose[11] represents an important advancement in the field of general cellular segmentation algorithms. It used U-Net[58] to predict horizontal and vertical gradient maps of cell instances and a foreground binary mask. After that, individual cells are segmented by grouping the pixels that point to the same center point in the gradient maps. This unique design allows it to be capable of processing a wide variety of cell morphologies in a unified framework. In the comparative studies, we used the most generalizable 'cyto2' model as the pretrained model, which was trained on the Cellpose dataset and user-submitted images.

Omnipose[8] was an extension of Cellpose, aiming to handle very elongated cells, especially bacterial cells. The network architecture backbone was still U-Net but the model had four heads to predict four components: two gradient flows, a distance transform map and a boundary map. We also chose the 'cyto2' model to infer the challenge testing images.

Cellpose 2.0 (ref. 12) further introduced a transfer-learning-based method, an important branch toward general cell segmentation solutions, to quickly adapt the pretrained models to new microscopy images. With a human-in-the-loop pipeline, users can train customized cellular segmentation models by fine-tuning pretrained Cellpose models with only 100–200 annotated regions of interest. The network architecture in Cellpose 2.0 was the same as the Cellpose model.

KIT-GE[33] trained a U-Net model to predict cell distance and neighbor distance, followed by watershed post-processing. Compared to the original U-Net[58], the maximum pooling layers were replaced with 2D convolutional layers with stride 2 and batch normalization layers were added after the convolutional layers.

Evaluation metrics

This challenge focused on two key metrics: segmentation accuracy and efficiency. While segmentation accuracy is a fundamental metric in cell segmentation, we included efficiency in the evaluation to account for its significance during model deployment. If the challenge metrics only considered the algorithm accuracy, participants may solely prioritize it by employing the ensemble of multiple models[15]; however, such solutions may not be practical in real-world scenarios, particularly for biologists who typically have limited computational resources. Recognizing this, we incorporated efficiency as an evaluation metric to guide participants in considering the tradeoff between model accuracy and efficiency.

Segmentation accuracy metric: F1 score. Cell segmentation is a typical instance segmentation task. We employed the widely used F1 score to evaluate the segmentation results[5,59,60]. Specifically, each predicted cell mask is matched to the most similar ground-truth mask based on the predefined IoU threshold (0.5). A predicted cell mask is classified as correct segmentation as long as its IoU is over the predefined IoU threshold. A higher threshold requires a larger overlap between the predicted cell mask and the ground-truth mask and a commonly used threshold is 0.5. Then, all the cells can be divided into three categories, including true positives (TPs), false positives (FPs) and false negatives (FNs). TP denotes correctly segmented cells, FP denotes wrongly segmented cells and FN denotes missed cells in the segmentation mask. After that, we can compute the precision and recall, which are defined by $\text{precision} = \frac{\text{TP}}{\text{TP+FP}}$ and $\text{recall} = \frac{\text{TP}}{\text{TP+FN}}$, respectively. The F1 score can be interpreted as a harmonic mean of the precision and recall, which is defined by

$$\text{F1} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}.$$

As the cells located in the boundaries are usually incomplete and have low practical value in various downstream analysis tasks, we removed these cells when computing the metrics.

Segmentation efficiency metric: running time. All the submitted Docker containers were run on the same desktop workstation with a 12-core CPU, 32 GB RAM and one NVIDIA 2080Ti GPU. To obtain the running time T for each image, the testing images were segmented one by one. To compensate for the Docker container startup time, we gave a time tolerance for the running time. Specifically, the time tolerance was 10 s if the image size (height (H) × width (W)) was

no more than 1,000,000. If the image size was more than 1,000,000, the time tolerance was (H × W)/1,000,000 × 10 s. This time tolerance was determined by the open-source U-Net baseline.

Ranking scheme: rank-then-aggregate

Both F1 score and running time were used for ranking; however, the two metrics cannot be directly fused because they have different dimensions. Thus, we used a 'rank-then-aggregate' scheme for ranking, including the following three steps:

- Step 1. Computing the two metrics for each testing case and each team;
- Step 2. Ranking teams for each of the n testing cases such that each team obtains $n \times 2$ rankings;
- Step 3. Computing ranking scores for all teams by averaging all these rankings and then normalizing them by the number of teams. The final rank will be determined by the mean ranking scores.

In addition to the employed rank-then-aggregate scheme, several other strategies can be used to obtain a ranking, but these may lead to different orderings of algorithms and thus different winners[61]. A typical ranking scheme was 'aggregate-then-rank': computing mean scores across all testing cases for each team and then using this aggregation to rank each team. One can also use test-based procedures for ranking. Specifically, each pair of algorithms are compared by a statistical hypothesis tests. The ranking is then performed according to the resulting relations or according to the number of significant one-sided test results. In the latter case, if algorithms have the same number of significant test results, then they obtain the same rank. For analysis purposes, we computed the ranks of the 28 algorithms based on five different ranking schemes: mean-then-rank, median-then-rank, rank-then-mean, rank-then-median and statistical significance test-based ranking.

Notably, for a transparent challenge, the evaluation code and ranking scheme were publicly available at the beginning of the challenge. For comparative analysis, we applied different ranking schemes to the 28 algorithms, including rank-then-mean, rank-then-median, median-then-rank, mean-then-rank and test-based rank. Most algorithms had fluctuations under different ranking schemes but the winning algorithm consistently held the first place.

Ranking stability and statistical analysis

Ranking stability is an important factor for robust challenge results[62]. Thus, we applied bootstrapping and computed Kendall's τ (ref. 63) to quantitatively analyze the variability of our ranking scheme. Specifically, we first extracted 1,000 bootstrap samples from the international validation set and computed the ranks again for each bootstrap sample. Then, the ranking agreement was quantified by Kendall's τ. Kendall's τ computes the number of pairwise concordances and discordances between ranking lists. Its value ranges $[-1, 1]$ where -1 and 1 denote inverted and identical order, respectively. A stable ranking scheme should have a high Kendall's τ value that is close to 1. To compare the performance of different algorithms, we performed a Wilcoxon signed-rank test because it is a paired comparison. Results were considered statistically significant if the P value was less than 0.05.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The dataset is available on the challenge website at https://neurips22-cellseg.grand-challenge.org/. It is also available on Zenodo at https://zenodo.org/records/10719375 (ref. 64). Source data are provided with this paper.

Code availability

The top ten teams have made their code publicly available at https://neurips22-cellseg.grand-challenge.org/awards/. They are also available on Zenodo at https://zenodo.org/records/10718351.

References

47. Misra, D. Mish: a self regularized non-monotonic activation function. in *British Machine Vision Conference* (2020).
48. Edlund, C. et al. Livecell—a large-scale dataset for label-free live cell segmentation. *Nat. Meth.* **18**, 1038–1045 (2021).
49. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778 (IEEE, 2016).
50. Schmidt, U., Weigert, M., Broaddus, C. & Myers, G. Cell detection with star-convex polygons. in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, 265–273 (MICCAI, 2018).
51. Ren, S., He, K., Girshick, R. & Sun, J. Faster r-cnn: towards real-time object detection with region proposal networks. in *Advances in Neural Information Processing Systems*, vol. 28 (NeurIPS, 2015).
52. Graham, S. et al. Hover-net: simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Med. Image Anal.* **58**, 101563 (2019).
53. Upschulte, E., Harmeling, S., Amunts, K. & Dickscheid, T. Contour proposal networks for biomedical instance segmentation. *Med. Image Anal.* **77**, 102371 (2022).
54. Kuhl, F. P. & Giardina, C. R. Elliptic fourier features of a closed contour. *Comput.Graph. Image Process.* **18**, 236–258 (1982).
55. Rezatofighi, H. et al. Generalized intersection over union: a metric and a loss for bounding box regression. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2019).
56. Lee, Y. et al. Localization uncertainty estimation for anchor-free object detection. in *Computer Vision – ECCV 2022 Workshops*, 27–42 (ECCV, 2023).
57. Adams, R. & Bischof, L. Seeded region growing. *IEEE Trans. Pattern Anal. Mach. Intell.* **16**, 641–647 (1994).
58. Ronneberger, O., Fischer, P. & Brox, T. U-net: convolutional networks for biomedical image segmentation. in *International Conference on Medical Image Computing and Computer-assisted Intervention*, 234–241 (MICCAI, 2015).
59. Maier-Hein, L. et al. Metrics reloaded: recommendations for image analysis validation. *Nat. Meth.* https://doi.org/10.1038/s41592-023-02151-z (2024).
60. Hirling, D. et al. Segmentation metric misinterpretations in bioimage analysis. *Nat. Meth.* https://doi.org/10.1038/s41592-023-01942-8 (2023).
61. Maier-Hein, L. et al. Why rankings of biomedical image analysis competitions should be interpreted with care. *Nat. Commun.* **9**, 5217 (2018).
62. Wiesenfarth, M. et al. Methods and open-source toolkit for analyzing and visualizing challenge results. *Sci. Rep.* **11**, 1–15 (2021).
63. Kendall, M. G. A new measure of rank correlation. *Biometrika* **30**, 81–93 (1938).
64. Ma, J. et al. NeurIPS 2022 Cell Segmentation Competition Dataset. in *Thirty-sixth Conference on Neural Information Processing Systems (NeurIPS) Zenodo* https://doi.org/10.5281/zenodo.10719375 (2024).

Acknowledgements

This work was supported by the Natural Sciences and Engineering Research Council of Canada (RGPIN-2020-06189 and DGECR-2020-00294) and CIFAR AI Chair programs. This research was enabled, in part, by computing resources provided by the Digital

Research Alliance of Canada. We thank P. Byrne, M. Kost-Alimova, S. Singh and A.E. Carpenter for contributing U2OS and adipocyte images. We thank A.J. Radtke and R. Germain for contributing adenoid and tonsil whole-slide fluorescent images. We thank S. Banerjee for providing multiple myeloma plasma cell annotations in stained brightfield images. The platelet DIC images collected by C. Kempster and A. Pollitt were supported by the British Heart Foundation/NC3Rs (NC/S001441/1) grant. A.G. thanks the Department of Science and Technology, Government of India for the SERB-POWER fellowship (grant no. SPF/2021/000209) and the Infosys Centre for AI, IIIT-Delhi for the financial support to run this challenge. M.L., V.G., M.S. and S.J.R. were supported by SNSF grants CRSK-3_190526 and 310030_204938 awarded to S.J.R. E.U. and T.D. received funding from Priority Program 2041 (SPP 2041) 'Computational Connectomics' of the German Research Foundation and the Helmholtz Association's Initiative and Networking Fund through the Helmholtz International BigBrain Analytics and Learning Laboratory under the Helmholtz International Laboratory grant agreement InterLabs-0015. The authors gratefully acknowledge the computing time granted through JARA on the supercomputer JURECA at Forschungszentrum Jülich. We also thank the grand-challenge platform for hosting the competition.

Author contributions

J.M. conceived and designed the analysis, collected and cleaned the data, contributed analysis tools, managed challenge registration and evaluation, performed the analysis, wrote the initial manuscript and revised the manuscript; R.X. conceived and designed the analysis, managed challenge registration and evaluation and revised the manuscript. S.A., C.G., A.G., R.G., S.G. and Y.Z. conceived and designed the analysis, cleaned data, contributed labeled images, managed challenge registration and evaluation, performed the analysis and revised the manuscript. G.L., J.K., W.L., H.L., E.U. and T.D. participated in the challenge, developed the top-three algorithms and made the code publicly available. J.G.A., Y.W., L.H. and X.Y. cleaned data, contributed labeled images and managed challenge registration and evaluation. M.L., V.G., M.S., S.J.R., C.K., A.P., L.E., T.M. J.M.M. and J.-N.E., contributed new labeled data in the competition. W.L., Z.L., X.C. and B.B. participated in the challenge, developed algorithms and made the code publicly available. N.F.G., D.V.V., E.W., B.A.C. and O.B. contributed public or unlabeled data to the competition. T.C. managed the challenge registration and evaluation. G.D.B. and B.W. conceived and designed the analysis and wrote and revised the manuscript.

Competing interests

S.G. is employed by Nanjing Anke Medical Technology Co. J.M.M. and J.-N.E. are co-owners of Cancilico. D.V.V. is a co-founder and chief scientist of Barrier Biosciences and holds equity in the company. O.B. declares the following competing financial interests: consultancy fees from Novartis, Sanofi and Amgen, outside the submitted work; research grants from Pfizer and Gilead Sciences, outside the submitted work; and stock ownership (Hematoscope) outside the submitted work. All other authors declare no competing interests.

Additional information

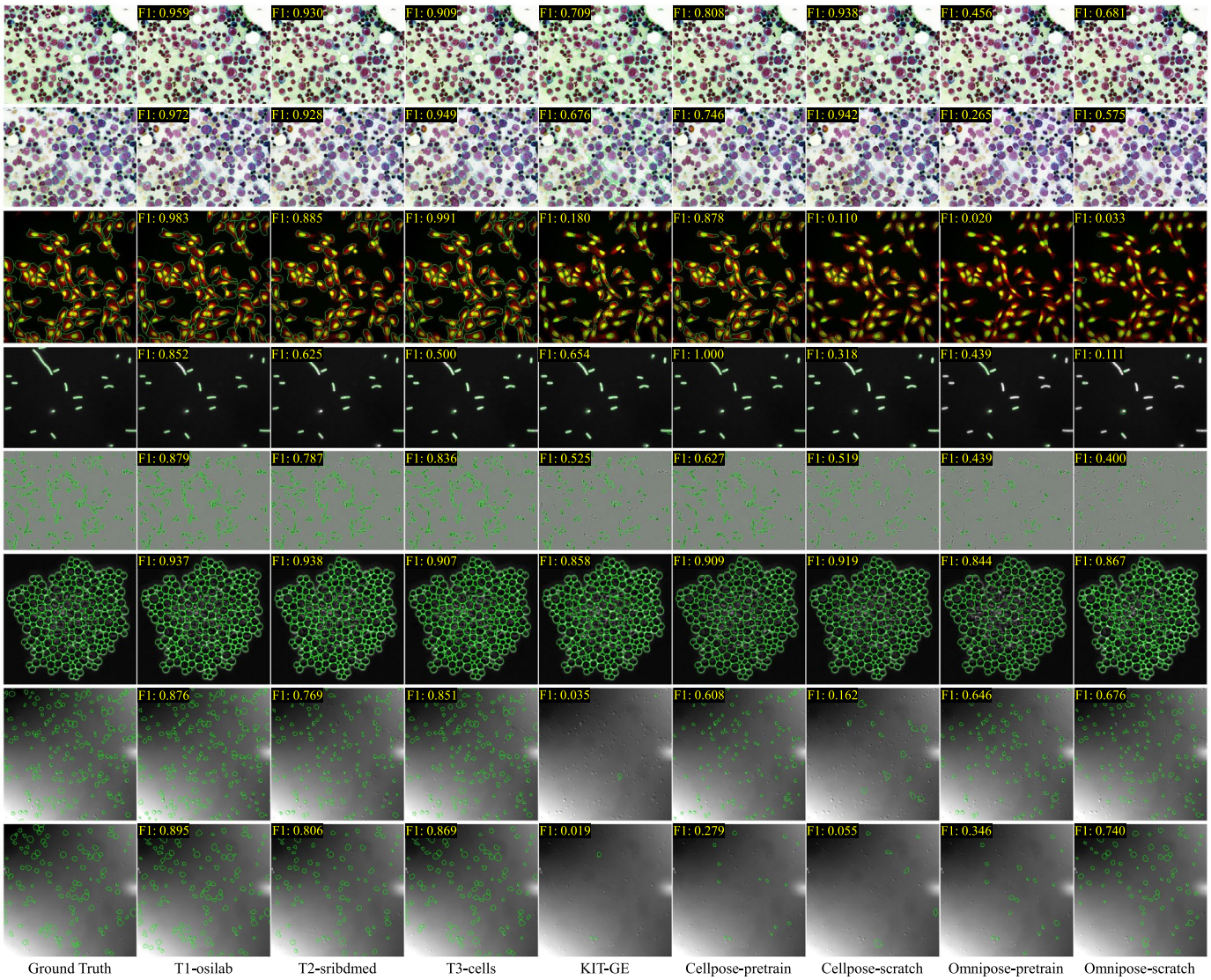
Extended data is available for this paper at https://doi.org/10.1038/s41592-024-02233-6.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-024-02233-6.

Correspondence and requests for materials should be addressed to Bo Wang.

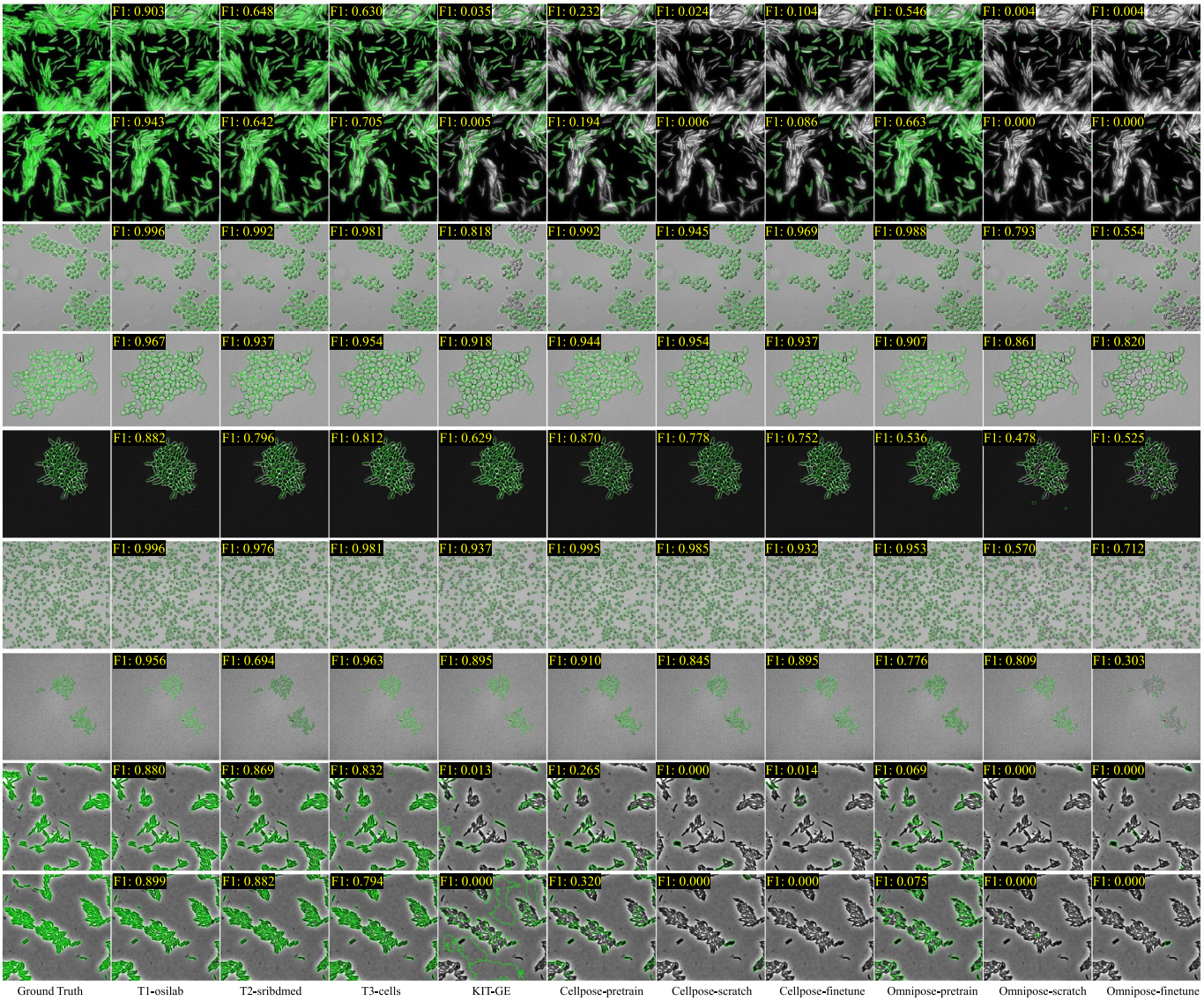
Peer review information *Nature Methods* thanks Yi Wang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Rita Strack, in collaboration with the *Nature Methods* team.

https://doi.org/10.1038/s41592-024-02233-6



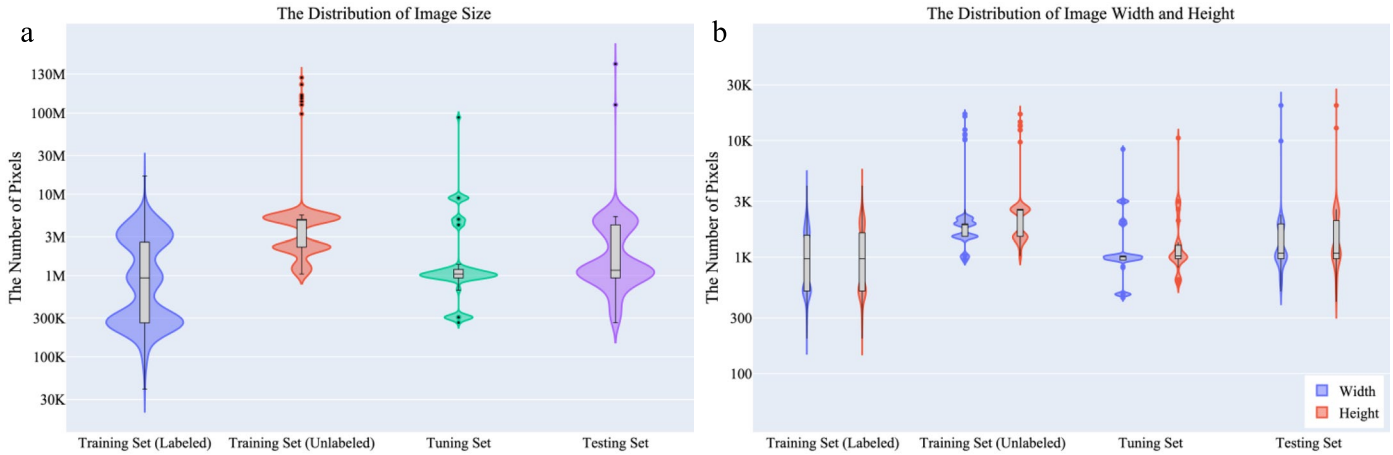
Extended Data Fig. 1 | Example segmentation results for four microscopy image modalities. Brightfield (the 1st-2nd rows), fluorescent (the 3rd-4th rows), phase-contrast (the 5th-6th rows), and DIC images (the 7th-8th) rows. Cellpose-pretrain: Cellpose pretrained model ('cyto2'). Cellpose-scratch: Cellpose model trained from scratch on the challenge dataset. Omnipose-pretrain: Omnipose pretrained model ('cyto2'). Omnipose-scratch: Omnipose model trained from scratch on the challenge dataset.

https://doi.org/10.1038/s41592-024-02233-6



Extended Data Fig. 2 | Example segmentation results for the post-challenge testing images. Cellpose-pretrain: Cellpose pretrained model ('cyto2'). Cellpose-scratch: Cellpose model trained from scratch on the challenge dataset. Cellpose-finetune: Cellpose fine-tuned model on the challenge dataset. Omnipose-pretrain: Omnipose pretrained model ('cyto2'). Omnipose-scratch: Omnipose model trained from scratch on the challenge dataset. Omnipose-finetune: Omnipose fine-tuned model on the challenge dataset.

https://doi.org/10.1038/s41592-024-02233-6



Extended Data Fig. 3 | Statistics of image size. a, Distribution of image size across training (labeled (n=1000 independent images) and unlabeled (n=1725 independent images)), tuning (n=101 independent images), and testing sets (n=422 independent images). **b**, Distribution of image width and height across training (labeled (n=1000 independent images) and unlabeled (n=1725 independent images)), tuning (n=101 independent images), and testing sets (n=422 independent images).

nature portfolio

Corresponding author(s): Bo Wang

Last updated by author(s): Nov 27, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	N/A No special software was used for data collection.
Data analysis	All the code to generate the results in this work has been publicly available at https://github.com/JunMa11/NeurIPS-CellSeg and https://neurips22-cellseg.grand-challenge.org/awards/. Analysis of the challenge results was performed by Python 3.10 and Challenge R 1.0.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Data has been publicly avaiable on the challenge website https://neurips22-cellseg.grand-challenge.org/.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	N/A This manuscript reports on procedures and software methods used to analyze images, not the actual use of those methods. There were no formal hypothesis either expected results in advance, therefore we collected the possible highest sample size regarding to the capacities.
Data exclusions	Images with less than five cells were excluded from the dataset, and cells containing fewer than 15 pixels were also removed.
Replication	N/A (see comment about Sample size) We did not perform any replication to have the groups independent as much as possible.
Randomization	N/A (see comment about Sample size) There was no randomization as we have no concerns about biased results due to chosen techniques.
Blinding	N/A (see comment about Sample size) There was no blinding as the participants chose the technique by their own to get the best results.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging