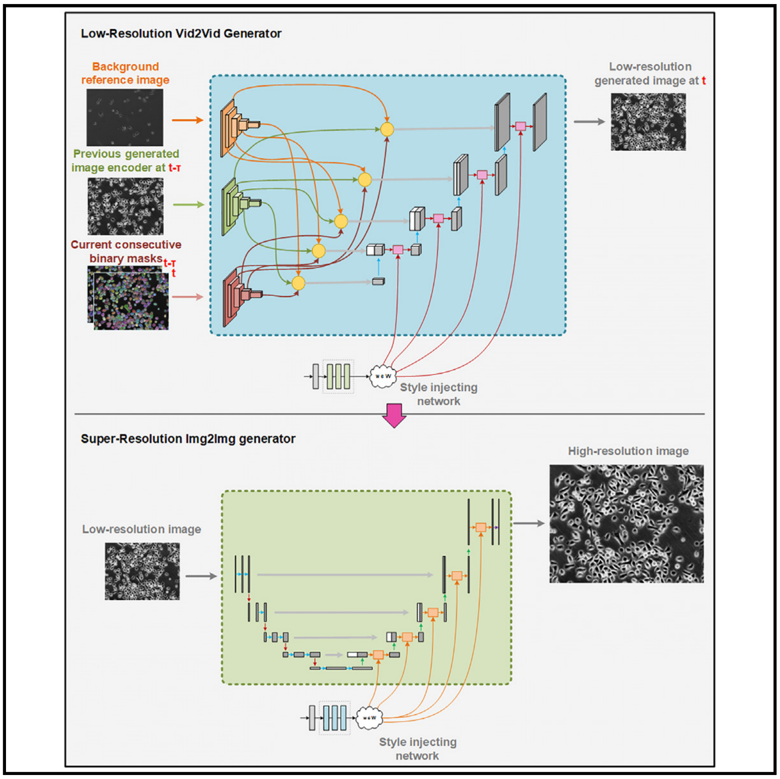
iScience

Article

Enhanced cell tracking using a GAN-based super-resolution video-to-video time-lapse microscopy generative model

Graphical abstract


Low-Resolution Vid2Vid Generator
Background reference image
Previous generated image encoder at t-T
Current consecutive binary masks t-T t
Low-resolution generated image at t
Style injecting network
Super-Resolution Img2Img generator
Low-resolution image
High-resolution image
Style injecting network

Authors

Abolfazl Zargari, Najmeh Mashhadi,
S. Ali Shariati

Correspondence

abzargar@ucsc.edu (A.Z.),
alish@ucsc.edu (S.A.S.)

In brief

Biotechnology; Applied computing; Computer modeling; Automation

Highlights

- tGAN produces realistic annotated microscopy videos with high temporal consistency
- tGAN dual-resolution design balances computational efficiency and high-resolution detail
- tGAN improves performance of single-cell tracking models with limited annotated data
- tGAN extrapolates to diverse scenarios, including high-density cell arrangements

Zargari et al., 2025, iScience *28*, 112225
April 18, 2025 © 2025 The Author(s). Published by Elsevier Inc.
https://doi.org/10.1016/j.isci.2025.112225

Article

Enhanced cell tracking using a GAN-based super-resolution video-to-video time-lapse microscopy generative model

Abolfazl Zargari,[1,*] Najmeh Mashhadi,[2] and S. Ali Shariati[3,4,5,6,*]
[1]Department of Electrical and Computer Engineering, University of California, Santa Cruz, Santa Cruz, CA, USA
[2]Department of Computer Science and Engineering, University of California, Santa Cruz, Santa Cruz, CA, USA
[3]Department of Biomolecular Engineering, University of California, Santa Cruz, Santa Cruz, CA, USA
[4]Institute for The Biology of Stem Cells, University of California, Santa Cruz, Santa Cruz, CA, USA
[5]Genomics Institute, University of California, Santa Cruz, Santa Cruz, CA, USA
[6]Lead contact
*Correspondence: abzargar@ucsc.edu (A.Z.), alish@ucsc.edu (S.A.S.)

https://doi.org/10.1016/j.isci.2025.112225

SUMMARY

Cells are among the most dynamic entities, constantly undergoing processes like growth, division, movement, and interaction with their environment and other cells. Time-lapse microscopy is central to capturing these dynamic behaviors, providing detailed spatiotemporal information at single-cell resolution in real time. Although deep learning has transformed cell segmentation, cell tracking remains challenging due to limited annotated time-lapse data. To address this, we introduce tGAN, a generative adversarial network (GAN)-based time-lapse microscopy generator that enhances the quality and diversity of synthetic annotated time-lapse microscopy data. Featuring a dual-resolution architecture, tGAN accurately captures both low- and high-resolution cellular details essential for accurate tracking. Our results show that tGAN generates high-quality, realistic annotated time-lapse videos with high temporal consistency and fine details. Importantly, annotated videos generated by tGAN enhance the performance of recent cell tracking models, reducing reliance on manual annotations. tGAN enhances deep learning's impact on bioimage analysis, enabling more generalizable cell tracking models.

INTRODUCTION

Time-lapse microscopy, a crucial tool in cell biology, captures the dynamics of cellular processes with high temporal resolution at single-cell level. However, the analysis of time-lapse images poses significant challenges, particularly in tracking cells over time.[1] A key hurdle in this process is the lack of annotated datasets, which are vital for training deep learning models to accurately segment and track individual cells over time.[2] Annotated time-lapse microscopy images are indispensable for understanding cellular mechanisms and responses, yet creating such datasets is often labor intensive and requires expert knowledge.

Recent advancements in deep learning, particularly in generative adversarial networks (GANs),[3] offer an unprecedented opportunity for the generation of realistic synthetic data.[4,5] GAN models, known for their ability to generate highly realistic synthetic data, have found applications across a wide range of disciplines, from art creation to medical imaging.[6–8] GAN models provide innovative solutions to simulate realistic biological data, which is essential for training and testing analytical models.[9] Among these applications, the generation of synthetic cell images from time-lapse microscopy using deep learning models, especially GANs, stands out as a particularly promising area.[10] These models offer new avenues for research in cell biology, enabling the creation of detailed and accurate representations of cellular processes.

To address the paucity of annotated time-lapse microscopy images, we introduce a video-to-video generative approach designed to produce annotated time-lapse microscopy videos. Our time-lapse generative model, termed tGAN, converts binary mask image sequences into corresponding high-resolution, synthetic images of cells. In addition to the generation of realistic time-lapse videos of cells, tGAN can be used to train cell tracking models with limited annotated datasets. The model offers a significant advancement in the automated generation of annotated time-lapse microscopy datasets, facilitating more efficient and accurate analysis of cellular dynamics.

RESULTS

Designing a novel GAN-based super-resolution video-to-video generator

The architectural design of our video-to-video generative model is central to its performance. Our design is characterized by its two-part structure, encompassing both video-to-video

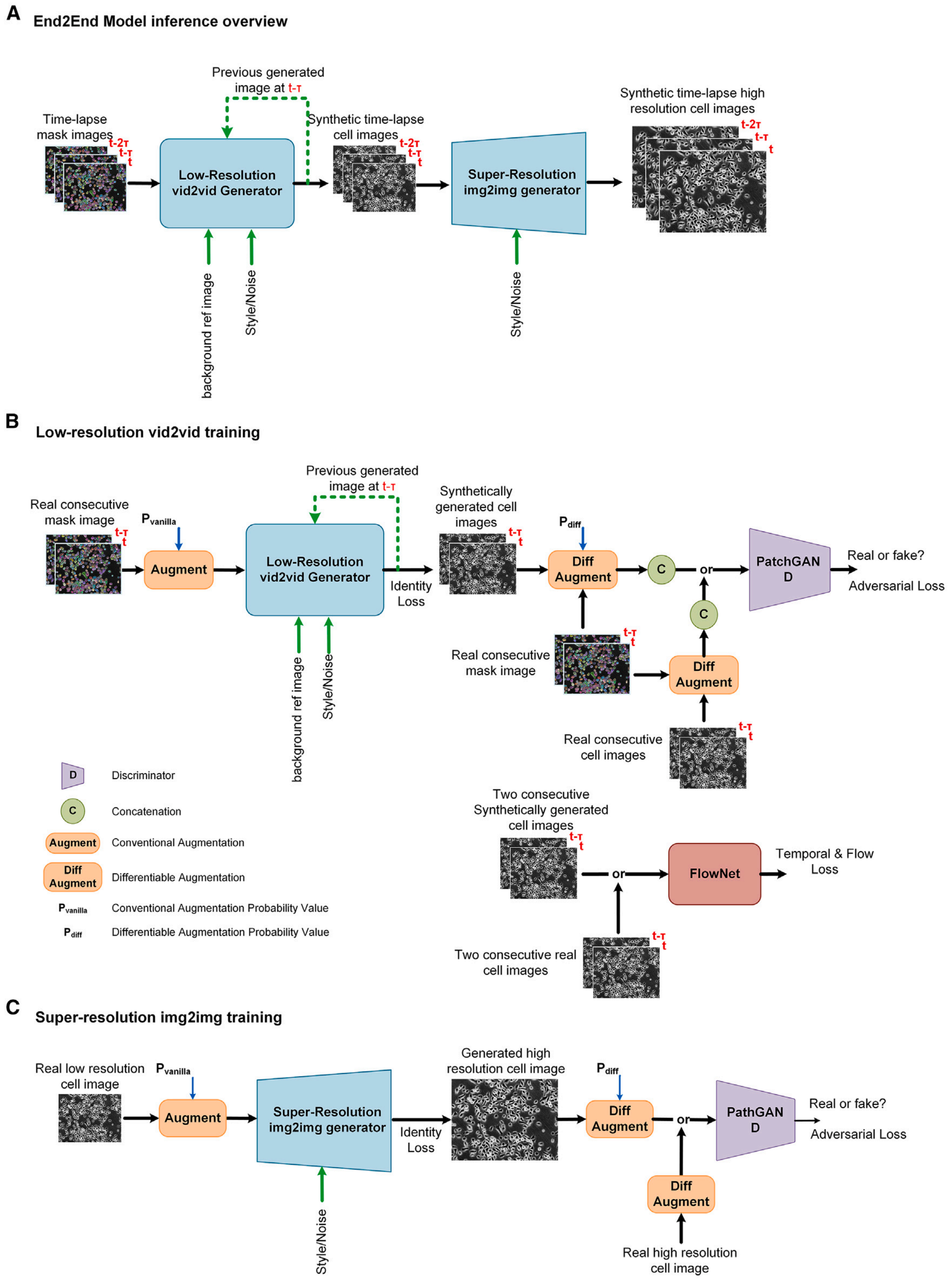

A
End2End Model inference overview
Previous generated image at t-τ
Time-lapse mask images
t-2τ
t-τ
t
Low-Resolution vid2vid Generator
Synthetic time-lapse cell images
t-2τ
t-τ
t
Super-Resolution img2img generator
Synthetic time-lapse high resolution cell images
t-2τ
t-τ
t
background ref image
Style/Noise
Style/Noise
B
Low-resolution vid2vid training
Real consecutive mask image
t-τ
t
P_vanilla
Augment
Previous generated image at t-τ
Low-Resolution vid2vid Generator
Identity Loss
Synthetically generated cell images
t-τ
t
P_diff
Diff Augment
C
or
PatchGAN D
Real or fake?
Adversarial Loss
C
Real consecutive mask image
t-τ
t
Diff Augment
Real consecutive cell images
t-τ
t
background ref image
Style/Noise
D
Discriminator
C
Concatenation
Augment
Conventional Augmentation
Diff Augment
Differentiable Augmentation
P_vanilla
Conventional Augmentation Probability Value
P_diff
Differentiable Augmentation Probability Value
Two consecutive Synthetically generated cell images
t-τ
t
or
FlowNet
Temporal & Flow Loss
Two consecutive real cell images
t-τ
t
C
Super-resolution img2img training
Real low resolution cell image
P_vanilla
Augment
Super-Resolution img2img generator
Identity Loss
Generated high resolution cell image
P_diff
Diff Augment
or
PathGAN D
Real or fake?
Adversarial Loss
Diff Augment
Real high resolution cell image
Style/Noise

Figure 1. Overview of the proposed GAN-based generator

(A) Illustration of the two-part inference architecture: This includes both the low-resolution video-to-video and super-resolution image-to-image models, where the initial phase generates low-resolution time-lapse microscopy image sequences to capture essential cellular dynamics, followed by the super-resolution phase that refines these sequences into more detailed and high-quality images.

(legend continued on next page)

low-resolution and image-to-image super-resolution models (Figure 1A). While the low-resolution model adequately captures the dynamics of time-lapse microscopy videos, our independent super-resolution model further enhances the quality of synthetic images by adding morphological details to single cells. This two-part structure is designed to optimize computational efficiency and precision. Since the low-resolution model training process involves multiple subunits such as discriminators, flow networks, and the video-to-video generator (Figure 1B), integrating all these elements at a high resolution would significantly increase computational costs and could potentially degrade the performance and training efficiency of the generator. Therefore, we initially trained the video-to-video generator in low resolution, achieving higher accuracy with less computational overhead. Subsequently, we trained a GAN-based super-resolution model (Figure 1C) to refine the video-to-video model outputs to high resolution, enhancing the quality and adding additional morphological details to cells for applications requiring high fidelity. This sequential approach not only ensures detailed, high-quality outputs but also maintains manageable computational demands, facilitating more efficient processing and superior performance in scenarios that require detailed, high-resolution images.

The low-resolution video-to-video generator, inspired by 2D-UNET architecture,[11] accepts multiple inputs, including current n (for example, two) consecutive mask frames, a previously generated cell frame (at t-T), and a consistent reference background image, ensuring context-aware and consistent background generation (Figures 1B and S2). One of the distinctive features of this phase is the inclusion of a reference background image. Incorporating a reference background image in our low-resolution GAN model enhances contextual accuracy to ensure consistency in background features. By increasing the variability of background visual features in synthetic samples, tGAN can accommodate various real-world scenarios like variable background noises, debris, or optical artifacts, which are relatively common in microscopy images. This approach enriches the realism and detail of the generated images, which is crucial for precise analysis in applications such as time-lapse microscopy, where background features can change over time. The integration of attention layers allows for adaptively integrating visual features from three model inputs. The inclusion of style and noise injection into the decoding path adds variability and realism to the generated images, as we have validated in our prior research.[12]

Transitioning to the super-resolution phase, our image-to-image generator employs an enhanced UNET-based architecture, incorporating style and noise injection for refining textural details and aesthetic elements, adding finer details and thus resulting in higher quality and more detailed images (Figures 1C and S3). While, in our experiments, we targeted a high resolution of 512 × 768 for the super-resolution model output, our model is adaptable and can be easily configured to achieve higher resolutions by adjusting its hyperparameters. The discriminators, designed with a PatchGAN architecture[13] and enhanced with a linear attention layer,[14] effectively distinguish fine details in images (Figures 1B and S4). This enhances the model's accuracy in differentiating between real and synthetic images. In the training process of the video-to-video generator, alongside the GAN and discriminator models, we concurrently trained and used a FlowNet (Figures 1B and S5) designed to calculate and integrate flow loss into the training regime. This FlowNet plays a crucial role in determining the optical flow loss,[15] comparing motion between consecutive frames in both real and generated sequences. This is essential for preserving the dynamic nature of cell movements in time-lapse microscopy. The optical flow loss computed by our FlowNet ensures that the temporal coherence and motion consistency of generated images align closely with actual microscopy sequences. However, it is not incorporated in the super-resolution model, which concentrates on image-to-image translation rather than the generation of video sequences, thereby making flow consistency less relevant in that context.

We selected loss functions, including temporal consistency and perceptual losses, which ensure temporal coherence and visual similarity between generated and real images (Equations 1, 2, and 3). Throughout the training process, we also employed augmentation strategies, including video-level augmentations for general model training, as well as video-level differentiable augmentations[16] for the discriminators (Figures S6A and S6B). These augmentations enhance the model's robustness and prevent overfitting, especially when the training dataset size is limited, as we have shown previously.[12]

Our tGAN model leverages real binary masks to generate synthetic time-lapse cell images, using these real masks of specific cell types to create synthetic videos as augmented data. It uses the real masks of specific cell types to generate synthetic time-lapse videos that serve as augmented versions of real data. In this case, both real and synthetic sequences share the same binary masks but differ in their visual features. To increase the diversity of real binary masks and, consequently, the diversity of the corresponding synthetic sequences, one can enhance existing real masks without generating entirely new ones. This can be achieved by applying conventional augmentations such as rotations, sequence splitting, and morphological operations (e.g., erosion and dilation) to individual cell regions within frames. These targeted modifications will increase the diversity and robustness of the dataset, strengthening model training on limited annotated data without requiring a fully trainable model for mask generation.

Benchmarking tGAN performance for generation of annotated video frames

To evaluate the performance of our model, we compared our model with the vid2vid model,[17] a notable example among state-of-the-art video-to-video generative models (Table 1). The vid2vid framework is particularly relevant for our comparative analysis as it is among very few publicly available generative models capable of converting mask scene sequences into

(B) Low-resolution video-to-video training process: showcasing the sequence-based approach and the integration of various inputs and deep learning components.

(C) Super-resolution image-to-image training process: highlighting the approach and models used for refining textural details and enhancing aesthetic elements, thereby producing scientifically accurate and visually high-resolution images.

Table 1. Comparative assessment of model performance, tGAN vs. vid2vid

Method	FVD (↓)	SSIM (↑)	PSNR (↑)	LPIPS (↓)
Mouse embryonic stem cells[21]				
vid2vid [17]	44.18	0.81	25.31	0.36
tGAN (ours)	8.83	0.95	32.58	0.19
Bronchial epithelial cells[21]				
vid2vid [17]	47.32	0.83	26.61	0.35
tGAN (ours)	17.21	0.90	28.42	0.25
Mouse C2C12 muscle progenitor cells[21]				
vid2vid [17]	12.72	0.73	19.19	0.36
tGAN (ours)	14.93	0.89	23.14	0.21
PhC-C2DH-U373[22]				
vid2vid [17]	8.86	0.88	25.91	0.25
tGAN (ours)	6.43	0.91	26.58	0.18
PhC-C2DL-PSC[22]				
vid2vid [17]	127.58	0.51	10.03	0.39
tGAN (ours)	98.27	0.84	15.68	0.13

This table presents a comparison of our model's performance across five different cell-type image sequences against the vid2vid model, measuring different video similarity metrics.

realistic scene sequences and generating high-resolution video frames. We used the structural similarity index (SSIM) and peak signal-to-noise ratio (PSNR) to compare image quality and similarity to real frames between the vid2vid and the models.[18] Additionally, we utilized the Frechet video distance (FVD)[19] metric with a pre-trained Inception_v3 to measure the distributional similarity between generated and real images, providing insights into the perceptual quality of tGAN's outputs. To evaluate the temporal coherence of generated video sequences, we adopted specific metrics designed to assess the smoothness and continuity of changes across frames. Finally, the perceptual image patch similarity (LPIPS) metric[20] was used to gauge the perceptual resemblance of generated images to real ones, ensuring that tGAN's outputs align closely with human visual judgment. Together, these metrics provided a robust framework for evaluating and benchmarking our model for video-to-video generation tasks. As presented in Table 1, our tGAN model obtained better performance across almost all the metrics and five cell types when tested on the DeepSea[21] and Cell Tracking Challenge[22] dataset time-lapse image sequences. Figure 2 showcases examples of two consecutive frames generated by our tGAN for each of the three DeepSea cell types, which are compared with the corresponding outputs from the vid2vid model. tGAN-generated images successfully capture realistic details of the cell bodies, including features like the nucleus, as well as the nuances of the background, demonstrating the model's effectiveness in rendering intricate biological structures. Figure S1 showcases an example of performance of our tGAN output with the vid2vid model, specifically focusing on the ability to replicate fine visual features of single stem cells. This comparison highlights the intricacies and effectiveness of each model in capturing detailed cellular characteristics. In Figure 3, we also present the capability of our proposed approach in generating cell image sequences against a variety of backgrounds given two reference background images. As shown, these backgrounds are precisely referenced from the reference background image used in the low-resolution video-to-video model, highlighting our method's adaptability in replicating diverse cell environmental settings.

To test the stability of tGAN over time, we analyzed the FVD score against the length of the tGAN-generated video in frames (Figure 4). We observed that the FVD scores exhibited only a slight fluctuation, approximately 1 FVD unit, when comparing videos ranging from 10 to 30 frames in length. This consistency in FVD scores, regardless of video length, underscores our tGAN model's stability and reliability in generating high-quality video sequences over varying durations.

We noticed that the tGAN training dataset (DeepSea and Cell Tracking Challenge datasets) predominantly contained low- and mid-density cell image sequences, which led us to investigate if our model could generate synthetic high-density cell videos of cells, a type not seen during training. This aspect is crucial as manual annotation of high-density microscopy videos is laborious and prone to errors. Therefore, we developed an algorithm specifically for creating synthetic high-density and colony-like time-lapse binary mask images for this experiment (as described in the STAR Methods section), which is comparatively straightforward using conventional image processing techniques. These were then used as inputs for our tGAN generator during testing. We demonstrated that tGAN could successfully extrapolate from low- and mid- to high-density images, validating its ability to produce a broad spectrum of realistic cell images and highlighting its potential in various applications (Figure 5).

Together, these findings demonstrate that tGAN can generate videos of cells with morphological details and temporal coherence similar to that of real videos of cells and across different annotated time-lapse microscopy datasets, further emphasizing the robustness and versatility of our approach.

Evaluating synthetic data from tGAN with cell tracking models

To assess the practical utility of our tGAN-generated synthetic data, we conducted an evaluation using two established cell tracking models in an inference-only setup. The primary goal of this assessment was to determine how well the synthetic time-lapse videos generated by tGAN align with real annotated microscopy data in supporting cell tracking tasks.

This assessment is conducted using the DeepSea cell tracker[21] and Bayesian tracker (btrack),[23] comparing the results to real annotated time-lapse videos from the DeepSea dataset. We chose btrack as a baseline due to its robust performance in various cell tracking tasks, particularly in challenging microscopy datasets with complex cell behaviors. Unlike deep learning-based models, btrack employs a Bayesian framework to probabilistically estimate cell trajectories by incorporating prior knowledge of cell dynamics, division rates, and lineage probabilities. This makes it highly effective in dense and noisy microscopy data, where large annotated datasets are scarce. The evaluation leverages various object-tracking metrics that we described in our previous research,[21] allowing for a detailed comparison of

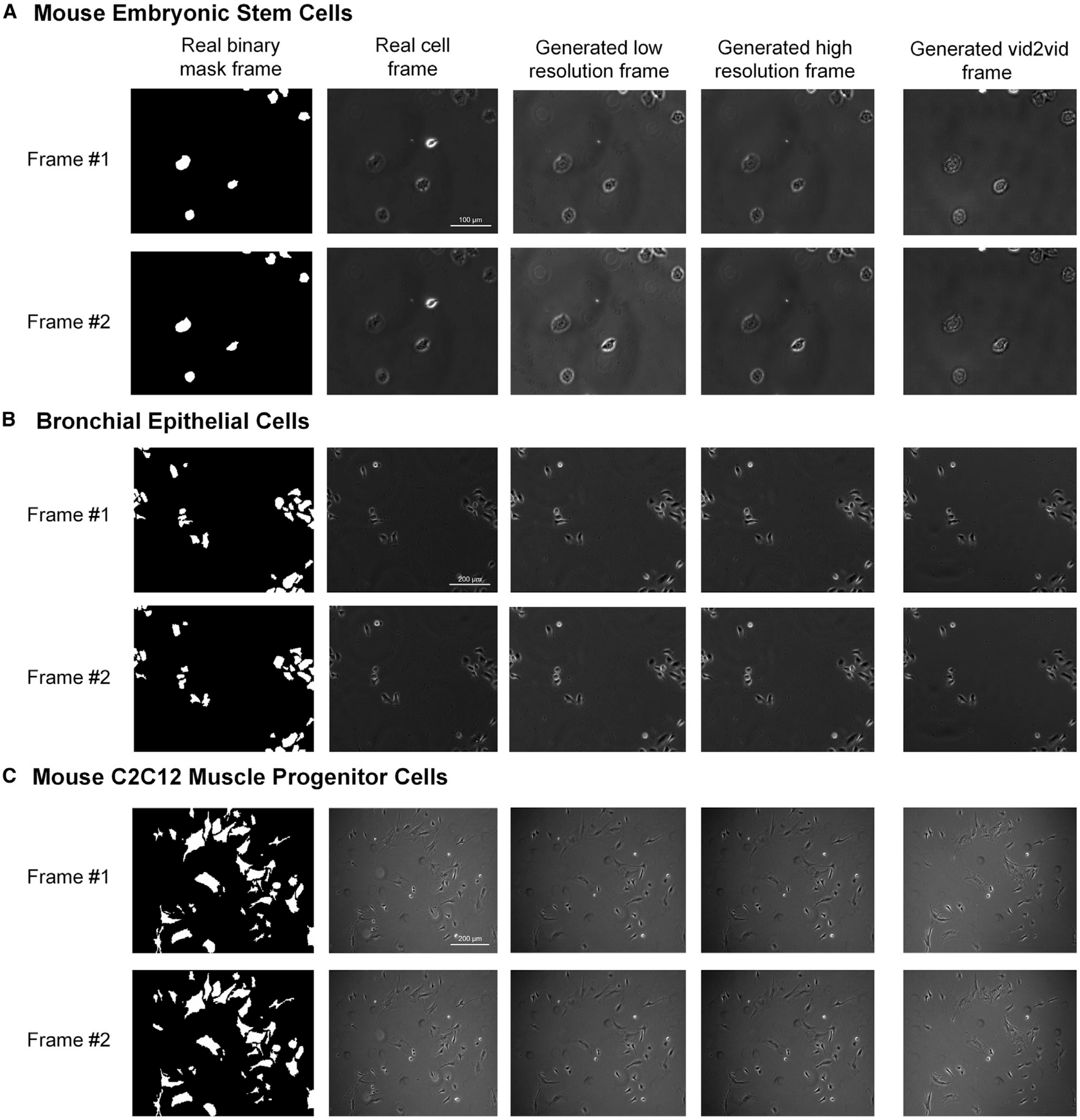

A Mouse Embryonic Stem Cells
Real binary mask frame
Real cell frame
Generated low resolution frame
Generated high resolution frame
Generated vid2vid frame
Frame #1
Frame #2
100 μm
B Bronchial Epithelial Cells
Frame #1
Frame #2
200 μm
C Mouse C2C12 Muscle Progenitor Cells
Frame #1
Frame #2
200 μm

Figure 2. Synthetic cell images, tGAN vs. vid2vid

The examples of two consecutive synthetic cell images generated by our proposed tGAN model compared to the vid2vid model outputs. (A) Mouse embryonic stem cells. (B) Bronchial epithelial cells. (C) Mouse C2C12 muscle progenitor cells.

each cell tracking model's performance. This evaluation focuses on how closely our model's tracking scores align with the ground truth annotated cell image sequences and align with our primary objective, which is to develop a GAN model capable of producing realistic annotated time-lapse microscopy images, addressing the scarcity of annotated data for training the sequence-based deep learning models, such as cell trackers.

As shown in Tables 2 and 3, the DeepSea tracker showed overall slightly better results than btrack in all cases, likely because the model is already trained on the DeepSea dataset samples. More importantly, our tGAN model shows better and closer results to real annotated time-lapse microscopy cell images compared to the vid2vid model. This confirms that the video frame sequences generated by our

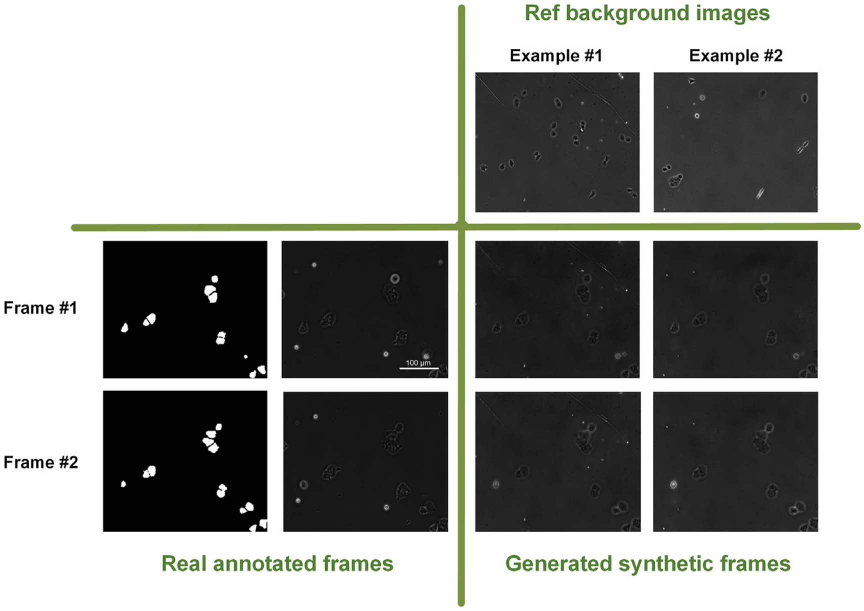


Figure 3. Varying reference backgrounds
Examples of applying different reference background images for the same sample.

tGAN model possess more realistic static and dynamic structures across frames, further validating the effectiveness of our approach in producing high-quality synthetic imagery.

tGAN-generated videos enhance performance of cell tracking models

The scarcity of annotated data for training sequence-based deep learning models, such as cell trackers, limits the application of deep learning methods for microscopy analysis. We aimed to determine whether videos generated by tGAN could enhance performance of cell tracking models, thereby addressing the shortage of annotated time-lapse videos for training cell tracking models.

To investigate the potential of synthetic data in enhancing cell tracking performance, we designed an additional set of experiments to evaluate the impact of using tGAN-generated videos as supplementary training data. We used two cell tracking models, the DeepSea[21] and TrackAstra,[24] focusing on three training scenarios: first, training on only 25% of the real dataset (limited real training data); second, using the entire real dataset (full real training data); and third, training the models on a combination of the same limited real dataset from the first scenario and the synthetic data generated by tGAN (limited real training data + synthetic training data), where tGAN itself was trained exclusively on the same limited real dataset, resulting in a total sample count matching the full real dataset scenario. To ensure robustness, each experiment was repeated five times with different random splits of the limited dataset, following a 5-fold cross-validation approach. The models were then tested on unseen real test data to assess their generalization capabilities.

As outlined in Table 4, augmenting the limited dataset with tGAN-generated data significantly improved tracking performance Multiple Object Tracking Accuracy (MOTA) score compared to using limited real data alone. Moreover, the results approached those obtained by training on the full real annotated dataset. Overall, the DeepSea tracker performed better than TrackAstra across all scenarios, likely due to prior optimization of DeepSea on the dataset. These findings demonstrate the effectiveness of tGAN in generating synthetic data that can enhance training when real data are scarce, reducing the dependency on labor-intensive annotations while maintaining robust tracking accuracy.

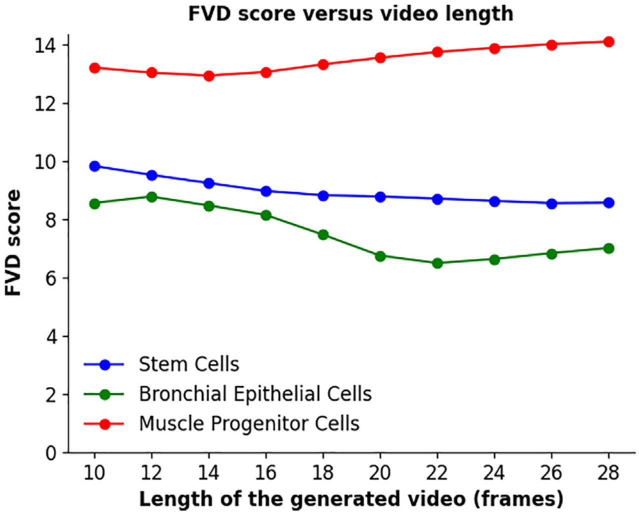


Figure 4. FVD variation with video length
Frechet video distance (FVD) scores across different video lengths for three DeepSea cell-type time-lapse video frames.

Ablation studies on model components

To further validate the effectiveness of the architectural components in our tGAN model, we conducted a series of ablation experiments. The goal of these experiments was to evaluate the impact of key design elements on the performance of our model, particularly in enhancing cell tracking accuracy and maintaining the fidelity of generated time-lapse microscopy sequences. For this analysis, we focused on the mouse embryonic stem cells (mESCs) from our DeepSea dataset, assessing the effect of each component using various quantitative metrics, including MOTA, SSIM, PSNR, and FVD. mESCs dataset includes enough training samples that represent a challenging yet well-annotated dataset that includes diverse cell behaviors, such as dynamic cell movements, proliferation, and interactions. This complexity

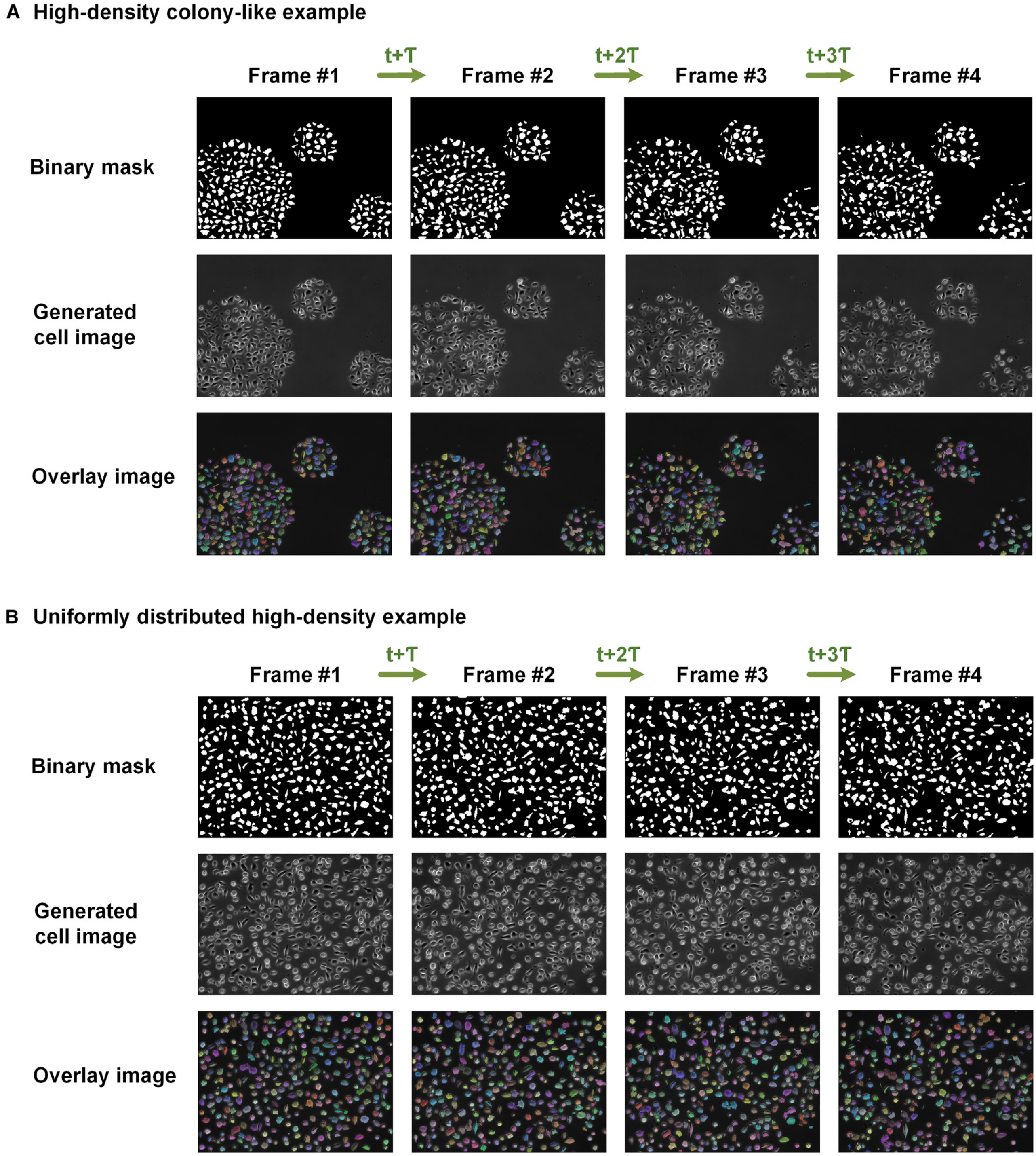


Figure 5. Synthetic colony-like and high-density cell images

The examples of producing synthetic colony-like (A) and high-density (B) time-lapse cell video frames generated by tGAN from synthetic binary masks.

makes it an ideal candidate for assessing the impact of various architectural components on the performance of our tGAN model.

The results, summarized in Table 5, demonstrate that FlowNet loss is the most critical component for maintaining the temporal coherence of generated sequences, as evidenced by a significant increase in FVD scores (from 8.83 to 12.27) and a substantial decrease in MOTA (from 0.93 to 0.86) when it is removed. This indicates that FlowNet loss plays a crucial role in preserving smooth temporal transitions, which are essential for accurate cell tracking. In contrast, removing the reference background

Table 2. Evaluation of synthetic time-lapse sequences with DeepSea cell tracking

Time-lapse test sequences	MOTA (↑)	MT (↑)	ML (↓)	Precision (↑)	Recall (↑)
Mouse embryonic stem cells[21]					
Real	0.90	0.82	0.02	0.96	0.94
Synthetic vid2vid	0.73	0.57	0.09	0.94	0.81
Synthetic tGAN	0.93	0.90	0.01	0.97	0.98
Bronchial epithelial cells[21]					
Real	0.93	0.94	0.02	0.96	0.98
Synthetic vid2vid	0.88	0.88	0.03	0.89	0.90
Synthetic tGAN	0.91	0.92	0.03	0.95	0.97
Mouse C2C12 muscle progenitor cells[21]					
Real	0.80	0.64	0.03	0.93	0.89
Synthetic vid2vid	0.52	0.23	0.25	0.74	0.64
Synthetic tGAN	0.76	0.60	0.06	0.94	0.85

Quality evaluation of synthetically generated time-lapse microscopy sequences using DeepSea Cell tracking model, measuring different object-tracking metrics.

primarily affects visual consistency, resulting in a noticeable drop in SSIM (from 0.95 to 0.87) and PSNR (from 32.58 to 28.90), with a comparatively smaller impact on MOTA (from 0.93 to 0.90). These results confirm that, while the reference background enhances image quality, it is less critical for tracking performance compared to maintaining temporal coherence through FlowNet loss.

DISCUSSION

A foundational aspect of developing deep learning models, particularly in the field of biomedical imaging, is the availability of a large and diverse annotated dataset. However, the creation of such datasets for microscopy images is often hindered by the laborious and time-consuming nature of manual annotation. Addressing this bottleneck, our study introduced tGAN, a GAN-based super-resolution video generator, designed to generate synthetic microscopy videos to address the scarcity of annotated live microscopy datasets, a field that demands high accuracy and detail in image processing. tGAN effectively circumvents the need for extensive manual annotation, generating realistic and diverse sets of synthetic annotated time-lapse cell images. Our study introduces a new approach to enhance the breadth of available training data for live microscopy to significantly boost the performance of segmentation and tracking models, particularly in scenarios with limited annotated datasets.

Table 3. Evaluation of synthetic time-lapse sequences with btrack

Test set	MOTA (↑)	MT (↑)	ML (↓)	Precision (↑)	Recall (↑)
Mouse embryonic stem cells[21]					
Real	0.85	0.80	0.02	0.93	0.94
Synthetic vid2vid	0.73	0.51	0.18	0.86	0.84
Synthetic tGAN	0.92	0.90	0.01	0.96	0.97
Bronchial epithelial cells[21]					
Real	0.84	0.75	0.25	0.87	0.98
Synthetic vid2vid	0.78	0.52	0.12	0.79	0.95
Synthetic tGAN	0.83	0.73	0.27	0.86	0.96
Mouse C2C12 muscle progenitor cells[21]					
Real	0.80	0.62	0.04	0.93	0.88
Synthetic vid2vid	0.53	0.23	0.25	0.74	0.63
Synthetic tGAN	0.75	0.56	0.09	0.93	0.81

Quality evaluation of synthetically generated time-lapse microscopy sequences using btrack model, measuring different object-tracking metrics.

The model's two-part structure adeptly handles both low-resolution and high-resolution image generation, a design choice that has proven valuable for tGAN to be computationally efficient. The integration of style and noise injection enhances the realism and morphological details of the images generated by tGAN. In addition, the inclusion of a reference background image ensures the adaptability of tGAN to diverse time-lapse scenarios with different background characteristics. We included different cell types in our study to demonstrate the performance of tGAN in a variety of challenging cell imaging scenarios. Our model's superior performance in generating high-quality, realistic time-lapse microscopy videos, as evidenced by its outperformance of state-of-the-art models, underscores its effectiveness in handling complex microscopy images. This success is quantitatively supported by comprehensive metrics, which collectively affirm the model's superiority in image quality, temporal coherence, and perceptual accuracy. Notably, our model's ability to generate annotated time-lapse microscopy images to enhance the performance of tracking models marks a useful and novel advancement in the field. tGAN is also able to generate synthetic video cells with high density, which can address challenges associated with manual annotation of high-density cell videos for the development of segmentation and tracking models. By generating synthetic high-density cell binary masks and using them as the model inputs in the test phase, we demonstrated the model's capacity to extend its application beyond the conditions experienced during training. This extrapolation is a practical benefit that can help reduce the time and effort required for manual annotation. Additionally, our experiments showed that using tGAN-generated videos as supplementary training data for cell tracking models improved tracking performance, even with limited real data. The augmented datasets achieved results close to those obtained with full real datasets, highlighting tGAN's potential to reduce the reliance on extensive manual annotations while maintaining robust tracking accuracy.

In conclusion, our tGAN model stands as a valuable tool with novel applications in microscopy imaging, capable of generating high-quality synthetic annotated time-lapse cell images across a spectrum of densities and cell types.

Limitations of the study

While our tGAN model shows promise in generating high-quality synthetic annotated time-lapse cell images, it has several

Table 4. Comparison of MOTA scores with real data and tGAN-augmented data

Tracking model	Limited real training data	Full real training data	Limited real training data + synthetic training data
Mouse embryonic stem cells[21]			
TrackAstra	0.79 ± 0.02	0.85 ± 0.01	0.86 ± 0.01
DeepSea	0.87 ± 0.03	0.94 ± 0.02	0.94 ± 0.02
Bronchial epithelial cells[21]			
TrackAstra	0.86 ± 0.01	0.89 ± 0.01	0.88 ± 0.01
DeepSea	0.90 ± 0.01	0.93 ± 0.01	0.93 ± 0.01
Mouse C2C12 muscle progenitor cells[21]			
TrackAstra	0.73 ± 0.01	0.80 ± 0.01	0.77 ± 0.01
DeepSea	0.81 ± 0.02	0.86 ± 0.01	0.84 ± 0.02

Comparative tracking performance (MOTA scores) using limited real data, full real data, and synthetic data augmentation with tGAN.

limitations. First, the model is trained and evaluated on a limited number of datasets (DeepSea and Cell Tracking Challenge) and may not generalize to all cell types or imaging conditions. Second, tGAN currently relies on real binary masks to guide the generation process, making it less applicable for scenarios where annotated masks are sparse or unavailable. Third, although we demonstrated the model's ability to extrapolate to higher cell densities, further investigations are required to confirm its robustness across extremely dense or highly heterogeneous cellular contexts. Additionally, our two-stage training approach, involving super-resolution, increases computational overhead, which may hinder large-scale or real-time applications. Future work could focus on improving the mask generation process, extending the model to more diverse imaging modalities and optimizing computational efficiency to address these constraints.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to the lead contact, Ali Shariati (alish@ucsc.edu).

Materials availability

All unique/stable materials generated in this study are available from the lead contact upon reasonable request with a completed Materials Transfer Agreement.

Data and code availability

- The image datasets utilized in our study are publicly available via a link on our GitHub repository, https://github.com/abzargar/tGAN, facilitating easy access for replication and further research efforts.
- The Python scripts encompassing the methodologies we developed are publicly accessible for download at our GitHub repository: https://github.com/abzargar/tGAN.
- Any additional information required to reanalyze the data reported in this work paper is available from the lead contact upon reasonable requests.

Table 5. Ablation metrics for mESCs, removing architectural components

Component removed	MOTA (↑)	SSIM (↑)	PSNR (↑)	FVD (↓)
Baseline (full model)	0.93	0.95	32.58	8.83
No reference background	0.90	0.87	28.90	10.50
No style and noise injection	0.89	0.89	29.50	10.45
No PatchGAN attention	0.91	0.92	31.10	9.75
No FlowNet loss	0.86	0.91	30.75	12.27

Comparative metrics for mouse embryonic stem cells with ablation of architectural components.

ACKNOWLEDGMENTS

We would like to thank the members of the Shariati lab for their feedback on the manuscript. We also would like to thank Ryan Shariati for his critical discussion on the potential application of synthetic live microscopy images. We are also thankful to the IBSC and Genomics Institute leadership for their generous support in providing facilities and computational resources to complete this study. Dr. Shariati is supported by the NIGMS/NIH through a Pathway to Independence Award K99GM126027/ R00GM126027 and Maximizing Investigator Award (R35GM147395), UCSC Large COR award and a start-up package from the University of California, Santa Cruz (S.A.S). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or the IBSC. Technical support from Benjamin Abrams, UCSC Life Sciences Microscopy Center, NIH S10 Grant 1S10OD23528-01, RRID: SCR_021135.

AUTHOR CONTRIBUTIONS

A.Z.K. and S.A.S. conceptualized the project and wrote the manuscript. Both A.Z.K. and N.M developed the related Python scripts and prepared the manuscript materials.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - ○ Cell lines
  - ○ Animal models
  - ○ Human participants
  - ○ Plant models
  - ○ Microbial models
- METHOD DETAILS
  - ○ Datasets
  - ○ Low-resolution video-to-video generative model overview

- ○ Super-resolution image-to-image generative model overview
- ○ Discriminator architecture overview
- ○ FlowNet architecture overview
- ○ Augmentation pipeline for training
- ○ Synthetic high-density time-lapse binary mask sequence generation
- ○ Loss functions and their rationale
- ● QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2025.112225.

Received: August 21, 2024
Revised: December 11, 2024
Accepted: March 12, 2025
Published: March 17, 2025

REFERENCES

1. Wong, C., Chen, A.A., Behr, B., and Shen, S. (2013). Time-lapse microscopy and image analysis in basic and clinical embryo development research. Reprod. Biomed. Online *26*, 120–129. https://doi.org/10.1016/j.rbmo.2012.11.003.
2. Chai, B., Efstathiou, C., Yue, H., and Draviam, V.M. (2024). Opportunities and challenges for deep learning in cell dynamics research. Trends Cell Biol. *34*, 955–967. https://doi.org/10.1016/j.tcb.2023.10.010.
3. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2 (MIT Press).
4. Karthika, S., and Durgadevi, M. (2021). Generative Adversarial Network (GAN): A General Review on Different Variants of GAN and Applications. In 6th International Conference on Communication and Electronics Systems (ICCES), Coimbatre, India, 8-10 July 2021, pp. 1–8. https://doi.org/10.1109/ICCES51350.2021.9489160.
5. Zhu, J.Y., Park, T., Isola, P., and Efros, A.A. (2017). Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks. 22-29 Oct. 2017, pp. 2242–2251. https://doi.org/10.48550/arXiv.1703.10593.
6. Shahriar, S. (2022). GAN computers generate arts? A survey on visual arts, music, and literary text generation using generative adversarial network. Displays *73*, 102237. https://doi.org/10.1016/j.displa.2022.102237.
7. Tripathi, S., Augustin, A.I., Dunlop, A., Sukumaran, R., Dheer, S., Zavalny, A., Haslam, O., Austin, T., Donchez, J., Tripathi, P.K., and Kim, E. (2022). Recent advances and application of generative adversarial networks in drug discovery, development, and targeting. Artif. Intell. Life Sci. *2*, 100045. https://doi.org/10.1016/j.ailsci.2022.100045.
8. Yi, X., Walia, E., and Babyn, P. (2019). Generative adversarial network in medical imaging: A review. Med. Image Anal. *58*, 101552. https://doi.org/10.1016/j.media.2019.101552.
9. Kazeminia, S., Baur, C., Kuijper, A., van Ginneken, B., Navab, N., Albarqouni, S., and Mukhopadhyay, A. (2020). GANs for medical image analysis. Artif. Intell. Med. *109*, 101938. https://doi.org/10.1016/j.artmed.2020.101938.
10. Lambard, G., Yamazaki, K., and Demura, M. (2023). Generation of highly realistic microstructural images of alloys from limited data with a style-based generative adversarial network. Sci. Rep. *13*, 566. https://doi.org/10.1038/s41598-023-27574-8.
11. Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. Preprint at arXiv. https://doi.org/10.48550/arXiv.1505.04597.
12. Zargari, A., Topacio, B.R., Mashhadi, N., and Shariati, S.A. (2024). Enhanced cell segmentation with limited training datasets using cycle generative adversarial networks. iScience *27*, 109740. https://doi.org/10.1016/j.isci.2024.109740.
13. Isola, P., Zhu, J.Y., Zhou, T., and Efros, A.A. (2017). Image-to-Image Translation with Conditional Adversarial Networks. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21-26 July 2017, pp. 5967–5976. https://doi.org/10.1109/CVPR.2017.632.
14. Jetley, S., Lord, N.A., Lee, N., and Torr, P.H.S. (2018). Learn To Pay Attention. Preprint at arXiv. https://doi.org/10.48550/arXiv.1804.02391.
15. Fischer, P., Dosovitskiy, A., Ilg, E., Häusser, P., Hazırbaş, C., Golkov, V., van der Smagt, P., Cremers, D., and Brox, T. (2015). FlowNet: Learning Optical Flow with Convolutional Networks. Preprint at arXiv. https://doi.org/10.48550/arXiv.1504.06852.
16. Zhao, S., Liu, Z., Lin, J., Zhu, J.-Y., and Han, S. (2020). Differentiable Augmentation for Data-Efficient GAN Training. Preprint at arXiv. https://doi.org/10.48550/arXiv.2006.10738.
17. Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Liu, G., Tao, A., Kautz, J., and Catanzaro, B. (2018). Video-to-Video Synthesis. Preprint at arXiv. https://doi.org/10.48550/arXiv.1808.06601.
18. Horé, A., and Ziou, D. (2010). Image Quality Metrics: PSNR vs. SSIM. In 2010 20th International Conference on Pattern Recognition, Istanbul, Turkey, 23-26 Aug. 2010, pp. 2366–2369. https://doi.org/10.1109/ICPR.2010.579.
19. Thomas, U., Sjoerd van, S., Karol, K., Raphal, M., Marcin, M., and Sylvain, G. (2019). FVD : A new Metric for Video Generation. ICLR Workshop DeepGenStruct. https://openreview.net/forum?id=rylgEULtdN.
20. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., and Wang, O. (2018). The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. Preprint at arXiv. https://doi.org/10.48550/arXiv.1801.03924.
21. Zargari, A., Lodewijk, G.A., Mashhadi, N., Cook, N., Neudorf, C.W., Araghbidikashani, K., Hays, R., Kozuki, S., Rubio, S., Hrabeta-Robinson, E., et al. (2023). DeepSea is an efficient deep-learning model for single-cell segmentation and tracking in time-lapse microscopy. Cell Rep. Methods *3*, 100500. https://doi.org/10.1016/j.crmeth.2023.100500.
22. Maška, M., Ulman, V., Delgado-Rodriguez, P., Gómez-de-Mariscal, E., Nečasová, T., Guerrero Peña, F.A., Ren, T.I., Meyerowitz, E.M., Scherr, T., Löffler, K., et al. (2023). The Cell Tracking Challenge: 10 years of objective benchmarking. Nat. Methods *20*, 1010–1020. https://doi.org/10.1038/s41592-023-01879-y.
23. Ulicna, K., Vallardi, G., Charras, G., and Lowe, A.R. (2021). Automated Deep Lineage Tree Analysis Using a Bayesian Single Cell Tracking Approach. Front. Comput. Sci. *3*, 734559. https://doi.org/10.3389/fcomp.2021.734559.
24. Gallusser, B., and Weigert, M. (2024). Trackastra: Transformer-based cell tracking for live-cell microscopy. Preprint at arXiv. https://doi.org/10.48550/arXiv.2405.15700.
25. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. (2019). Analyzing and Improving the Image Quality of StyleGAN. Preprint at arXiv. https://doi.org/10.48550/arXiv.1912.04958.
26. Johnson, J., Alahi, A., and Fei-Fei, L. (2016). Perceptual Losses for Real-Time Style Transfer and Super-Resolution. Preprint at arXiv. https://doi.org/10.48550/arXiv.1603.08155.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
Vid2vid	Wang et al.[17]	https://github.com/NVIDIA/vid2vid
DeepSea	Zargari et al.[21]	https://deepseas.org/
Cell Tracking Challenge	Maska et al.[22]	https://celltrackingchallenge.net/
Btrack	Ulicna et al.[23]	https://github.com/quantumjot/btrack
TrackAstra	Gallusser et al.[24]	https://github.com/weigertlab/trackastra
tGAN	This paper	https://github.com/abzargar/tGAN Zenodo https://doi.org/10.5281/zenodo.14911707

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Cell lines

We used published microscopy data of to develop our tGAN model. All the dataset are described in the method details section.

Animal models

We didn't use animal models in this study.

Human participants

We didn't have human participants in this study.

Plant models

We didn't use plant models in this study.

Microbial models

We didn't use microbial models in this study.

METHOD DETAILS

Datasets

In our study, we utilized two distinct training datasets from time-lapse microscopy, each representing different cell types, to ensure the robustness and generalizability of our proposed generative model across various biological contexts. These datasets are 1) Our recently published annotated dataset of phase-contrast images from the DeepSea collection,[21] which includes a large set of accurately annotated phase-contrast time-lapse microscopy images of three cell types: Mouse Embryonic Stem Cells, Bronchial Epithelial Cells, and Mouse C2C12 Muscle Progenitor Cells. 2) The Cell Tracking Challenge dataset (CTC),[22] a repository of 2D and 3D time-lapse sequences of fluorescent images featuring different cell types, such as PSC and U373 cells.

The selection of datasets was driven by the need to evaluate the robustness and adaptability of the tGAN model across diverse microscopy scenarios. The DeepSea dataset includes well-annotated phase-contrast time-lapse microscopy images, providing a solid foundation for testing our model's ability to generate realistic synthetic sequences. In contrast, the Cell Tracking Challenge (CTC) dataset serves as a benchmark in the field, offering challenging microscopy videos across various imaging modalities (e.g., fluorescence, phase-contrast) and cell types (e.g., epithelial cells, stem cells, muscle progenitor cells). For our experiments, we specifically chose cell types from the CTC dataset that had a sufficient number of samples. This selection ensured robust training and evaluation, allowing us to effectively assess the impact of tGAN-generated synthetic data on enhancing cell tracking models.

Low-resolution video-to-video generative model overview

As illustrated in Figures 1B and S2, our low-resolution model is a sequence-based architecture inspired by the 2D-UNET design, which is widely used for image segmentation tasks.[11] This architecture was selected for its ability to capture detailed spatial features while preserving efficient processing speeds, which is crucial for generating time-lapse microscopy sequences. The model processes three distinct types of input: the current consecutive mask images (e.g., two consecutive mask images), the previously generated synthetic cell image, and a consistent background image that serves as a reference for background visual features. The mask sequences act as the driving force, guiding the synthesis of cell images that are subsequently superimposed onto the reference background. This approach ensures realistic, context-aware generation of cell images, preserving background consistency across sequences.

Central to our model's architecture is the incorporation of specialized attention layers. These layers are strategically positioned to merge features from the three encoded inputs adaptively at multiple down-sampling stages. The attention mechanism was incorporated to enhance feature relevance by directing focus toward the most critical features in each input, which we observed improves the detail and consistency of generated images. This suggests that attention layers play a key role in achieving spatial fidelity and relevance in synthetic images.

In addition to structured feature integration, our model applies style and noise injection techniques in the decoding pathway, inspired by principles from neural style transfer.[25] The style and noise injections introduce controlled variability and texture that closely match the biological detail required for realistic microscopy images. As demonstrated in our previous research,[12] adding style and noise variability into the decoding path of UNET leads to visually richer synthetic images with enhanced texture. This approach was selected after initial experiments showed that excluding these injections led to synthetic outputs that lacked textural diversity and biological detail. The model's ability to process and integrate multiple input types, combined with the strategic use of attention mechanisms and style injections, positions it as a novel approach in synthetic image generation for time-lapse microscopy.

Super-resolution image-to-image generative model overview

Following the generation of low-resolution synthetic cell image sequences, our approach employs a super-resolution generative model to further refine and enhance these synthetic images (Figures 1C and S3). This super-resolution model, similar to its low-resolution counterpart, is built on a UNET-based architecture[11] but with additional enhancements. This model similarly integrates style and noise injection techniques in its decoding path, which is influenced by StyleGAN principles.[25] The style and noise injection significantly enhance the textural details and stylistic elements, contributing to the generation of more realistic and aesthetically consistent images, as demonstrated in our previous research.[12] The architecture features a sequence of encoder blocks that increase feature map depth, capturing intricate cell details. This is followed by a bottleneck process that prepares these features for nuanced reconstruction. In the decoder stages, the model combines upsampled features with style and noise information, progressively enhancing image resolution and quality. An additional upsampling layer in the decoder further increases the output resolution to 512x768, which can be adjusted to higher resolutions by modifying the model parameters.

Discriminator architecture overview

The discriminator plays a pivotal role in distinguishing between real and generated images. We have chosen the PatchGAN architecture for the discriminator of both low-resolution and high-resolution training, enhanced with the addition of a linear attention layer in its early layers (Figures 1 and S4). This architecture and the inclusion of specific components are deliberate choices aimed at optimizing the discriminator's performance. PatchGAN is known for its effectiveness in distinguishing fine details in images, making it a good choice for our purposes.[13] Unlike traditional discriminators that classify an entire image as real or fake, PatchGAN focuses on classifying smaller patches of the image. This approach is particularly beneficial for our model as it ensures that the generated images not only look realistic on a macro scale but also maintain high fidelity in finer details. In the measurement of our low-resolution discriminator loss, we employed a specific approach to enhance the model's discriminative capability. At each training step, we concatenated the last n (e.g., two) real and synthetic (fake) frames along with their corresponding last two real binary mask frames. This concatenation provides the discriminator with a more comprehensive context, allowing it to assess not just the individual frames but also their temporal consistency and alignment with the binary masks. This technique is particularly effective in reinforcing the discriminator's ability to discern subtle differences between real and generated image sequences, thereby sharpening the adversarial dynamic of the model.

The integration of a linear attention layer in the early layers of the discriminator is a beneficial approach. Attention mechanisms have gained popularity in various deep-learning applications for their ability to enhance model performance by focusing on relevant features while ignoring irrelevant ones.[14] The linear attention layer in our model allows the discriminator to prioritize certain aspects of the image, such as specific textures or patterns that are crucial for making accurate classifications. This focused approach improves the model's efficiency and accuracy in distinguishing real images from synthetic ones.

FlowNet architecture overview

In the training process of the low-resolution video-to-video generative model, we simultaneously trained a FlowNET model (Figures 1B and S5), a specialized component designed to calculate and integrate flow loss. This flow loss is crucial for accurately simulating the dynamics of cell movements, thereby enhancing the temporal consistency across video frames.[15] We incorporated this additional loss metric to optimize the generative model's weight updates, specifically aiming to maintain temporal coherence in the synthesized video sequences. This strategy ensures that the generated sequences not only mirror real-world temporal dynamics but also enhance the realism and scientific applicability of the generated images.

The optical flow loss, calculated by our FlowNET, is pivotal in maintaining the integrity of temporal dynamics across generated frames, aligning these dynamics closely with real-world observations to enhance the realism and scientific utility of the synthetic images. It's important to note, however, that the application of flow loss is selectively applied; it is not utilized in the high-resolution model, which focuses primarily on enhancing image detail through image-to-image translation, thus reducing the need for temporal consistency in that specific context.

Augmentation pipeline for training

We employed a robust augmentation pipeline to enhance the training of both the generators and discriminators. This process involves applying a series of video-level augmentation functions to each training sample, designed to introduce variability and improve the model's generalization capabilities. As illustrated in Figure S6A, we applied some mostly used conventional image augmentation functions, including random adjustments in histogram equalization, sharpness, brightness, and contrast, as well as horizontal and vertical flips, cropping, saturation modifications, and the addition of Gaussian noise and blur. The training algorithm executes a sequence of the provided augmentation functions for each cell and mask video pair with a pre-defined probability value 'p_vanilla'. In the requested augmentation pipeline, each function is randomly chosen with a consistent probability of 50% and is also applied in a randomized sequence.

In the training of GAN models, the limitations of conventional augmentation become apparent, particularly in its inability to significantly diversify the generated images when the training dataset size is limited. To address this challenge, the concept of differentiable augmentation[16] proves to be invaluable. This approach has also been validated in our previous research.[12] Differentiable augmentation applies the same random augmentations to both real and fake samples in a way that is differentiable with respect to the model parameters. This approach encourages the discriminator to mitigate overfitting and improve training stability, making it particularly beneficial for GANs trained with limited data, thus causing the generator to produce more diverse images, thereby improving the overall image generation performance. In the training process, we ran five distinct differentiable video-level augmentation functions, such as random contrast, brightness, cutout, translation, and saturation (Figure S6B). The application of each augmentation is controlled by a predetermined probability variable, 'p_diff'. To promote unbiased representation and randomness in the training data, these augmentations are applied in a random sequence. Each function has an equal chance of being selected, set at a 50% probability.

Synthetic high-density time-lapse binary mask sequence generation

To generate synthetic high-density binary mask sequences, we employed a multi-step process that leverages conventional image processing techniques and augmentation strategies, rather than using a trainable deep learning model. The process is detailed as follows:

1. Extracting Real Cell Masks: We began by extracting binary cell masks from an annotated dataset. These real masks were preprocessed to enhance their diversity by applying basic augmentations, such as random rotations, horizontal and vertical flips. This preprocessing step ensured that the extracted cell regions could later be used to generate synthetic masks with varying visual features.
2. Generating the Initial High-Density Mask: The next step involved generating an initial high-density binary mask. To achieve this, multiple extracted cell regions were combined and placed on a blank binary canvas. During this process, additional augmentations (e.g., scaling, rotations, and morphological operations like erosion and dilation) were applied to each cell region to introduce further diversity. The placement of cells was performed randomly, guided by a weighted probability distribution to simulate either flat or colony-like patterns. This allowed us to generate masks with different cell densities, testing the model's ability to generalize to high-density inputs. The resulting high-density mask was saved as the initial frame of the time-lapse sequence.
3. Creating the Synthetic Time-Lapse Sequence: Once the initial high-density mask was created, it served as the starting point for generating a synthetic time-lapse sequence. To simulate realistic cell movements over time, we applied slight shifts, rotations, and morphological changes to each cell region across multiple frames. Using connected component labeling, each cell region was identified, and random transformations were applied to simulate natural cell movements. These transformations included random shifts and rotations to simulate movement, morphological augmentations (e.g., erosion or dilation) to alter the shape of individual cells, and collision avoidance checks to ensure that cells did not overlap with one another in subsequent frames.

This iterative process was repeated for each frame, with each new frame generated based on the previous one, thereby creating a realistic sequence of cell dynamics. The final output was a sequence of frames that captured cell movements over time.

Loss functions and their rationale

We employed a combination of other critical loss functions (Equations 1, 2, and 3) to guide the models effectively during the training process. These include perceptual (VGG) loss,[26] L1 loss, and discriminator loss, each serving a specific purpose and contributing to the overall performance and accuracy of the model. The perceptual (VGG) loss and the L1 loss, respectively, ensure perceptual and pixel-wise similarity between the generated and real images. These losses focus on high-level features and granular accuracy. The discriminator loss, Mean Squared Error (MSE), plays a dual role: for the generator, it gauges effectiveness in deceiving the discriminator, and for the discriminator, it evaluates its capacity to differentiate real from synthetic images. This adversarial loss is key in driving the generative process toward producing images that closely mimic real ones.

We also, in the training process, assigned a specific weight (w1-w3) for each of these loss functions aiming to ensure a balanced contribution during the optimization process. This weighting is crucial as it fine-tunes the impact of each loss function according to its relevance and importance in the image generation task. It is important to note that we do not employ flow loss in the training process of the super-resolution model, as the FlowNET, which calculates flow loss, is not utilized in this phase of the training. This decision is based on the super-resolution model's focus on image-to-image translation rather than temporal video sequence generation.

$$Low\_Res\_Gen\_loss = D\_MSE + w1 \times Flow\_Loss + w2 \times L1\_Loss + w3 \times VGG\_Loss \quad \text{(Equation 1)}$$

$$High\_Res\_Gen\_loss = D\_Fake\_MSE + w2 \times L1\_Loss + w3 \times VGG\_Loss \quad \text{(Equation 2)}$$

$$D\_loss = 0.5 \times D\_Real\_MSE + 0.5 \times D\_Fake\_MSE \quad \text{(Equation 3)}$$

QUANTIFICATION AND STATISTICAL ANALYSIS

To verify the reproducibility of our findings, we repeated some training and testing experiments using the cross-validation method. We selected five random subsets for training, validation, and testing from the generated dataset and reported the average performance metrics.