SynCellFactory: Generative Data Augmentation for Cell Tracking

Moritz Sturm, Lorenzo Cerrone, and Fred A. Hamprecht

IWR, Heidelberg University, 69120 Heidelberg, Germany
moritz.sturm.aca@gmail.com, lorenzo.cerrone@uzh.ch,
fred.hamprecht@iwr.uni-heidelberg.de

Abstract. Cell tracking remains a pivotal yet challenging task in biomedical research. The full potential of deep learning for this purpose is often untapped due to the limited availability of comprehensive and varied training data sets. In this paper, we present SynCellFactory, a generative method for cell video augmentation. At the heart of SynCellFactory lies the ControlNet architecture, which has been fine-tuned to synthesize cell imagery with photorealistic accuracy in style and motion patterns. This technique enables the creation of synthetic, annotated cell videos that mirror the complexity of authentic microscopy time-lapses. Our experiments demonstrate that SynCellFactory boosts the performance of well-established deep learning models for cell tracking, particularly when original training data is sparse.

Keywords: Cell Tracking · Generative Data Augmentation · Microscopy Time-lapses

1 Introduction

Digital time-lapse microscopy allows for large-scale observations of cells over time, providing a deeper understanding of cellular processes [4, 11, 2]. However, to fully harness the potential of time-lapse imaging, automated cell tracking approaches are needed, which can provide a quantitative analysis of cell behavior for vast amounts of data.

Cell tracking is characterized by challenges such as variable image contrast, intricate behaviors such as cell division and, in some examples, indistinguishability of cells. Recent advancements in computer vision have shown that neural networks are highly effective in multi-object tracking tasks [8, 15, 6]. However, their application in cell tracking remains limited and exploratory, primarily due to the scarcity of annotated cell tracking data [13].

In recent years, although medium-scale annotated tracking data sets in the order of several thousand timeframes have become available [21, 27, 13], they are limited to specific cell styles.

To address these challenges, this paper introduces *SynCellFactory*, a generative data augmentation strategy specifically designed for cell tracking. We embrace the power of conditioned 2D diffusion models [19, 28] to generate high-quality, fully annotated synthetic cell videos that mimic the appearance and behavior of real cell data sets.

Utilizing as little as one annotated cell video for training, *SynCellFactory* can generate an extensive library of annotated videos in a consistent style, effectively augmenting the available training data. This advancement holds the potential to transform cell tracking by enabling the application of sophisticated deep learning models previously hindered by data scarcity. Our research focuses on evaluating the impact of this data augmentation on the performance of state-of-the-art cell tracking models. The results indicate that this novel approach addresses the data scarcity issue and enhances tracking accuracy. We aim to open the door for large-scale deep learning architectures in cell tracking.

Previous work already demonstrated the effectiveness of generative data augmentation using diffusion models in terms of achievable image classification accuracy [1, 9, 25, 20]. These advances also start to impact the medical imaging domain, where they show great promise [24, 18, 5, 17] and have already proven effective when used to train deep learning models [3, 7]. Closely related to our work, [22] generates an entire video at once using a 3D diffusion model guided by optical flow; crucially, their approach does not produce tracking pseudo ground truth labels, while *SynCellFactory* does. In [10], the authors propose a model based on CycleGAN [29] which is capable of generating raw data as well as lineage and segmentation pseudo ground truth. However, it relies on simulated ground truth segmentation masks to condition the data generation.

To summarize, our contributions are:

- *SynCellFactory*, a generative data augmentation pipeline for cell tracking.
- The proposed pipeline demonstrates out-of-the-box robust results on a wide variety of data sets without complex hyperparameter tuning or domain-specific knowledge.
- Empirical proof that the proposed data augmentation strategy can further enhance accuracy of a leading deep learning cell tracking method.

2 Method

SynCellFactory operates on the principle of decoupling cell dynamics from their appearance. Our model comprises two principal components:

1. A simple 2D motion model that simulates the spatial distribution and dynamics of cell cultures. It uses statistical parameters derived from real data, enabling *SynCellFactory* to produce coherent and physically plausible time-lapses. This simulation approach enhances the diversity and realism of the generated data.
2. Two distinct ControlNets, each with a specific function, are trained for photorealistic rendering. The first, CN-Pos, is adept at inpainting cells at accurate spatial positions. The second, CN-Mov, focuses on the temporal displacement of individual cells across consecutive frames, ensuring temporal

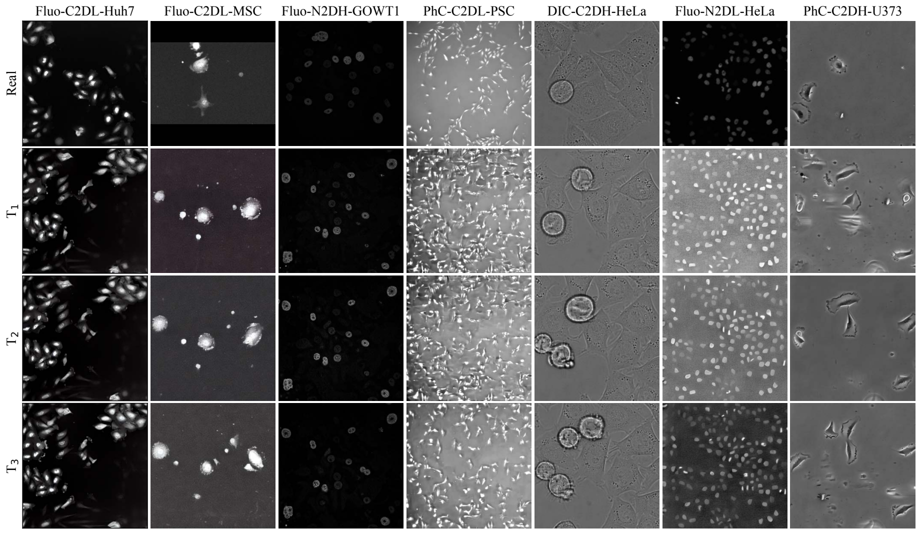

Fig. 1. Showcase of real (top row) and synthetic (other rows) images generated using *SynCellFactory*. Generated videos are provided in the supplementary materials. The training data sets are a subset of the 2D Cell Tracking Challenge [26, 13] and provide a broad spectrum of cell lines and microscopy modalities.

consistency. These ControlNets show efficient training capabilities, making them well suited for augmenting data sets in cell tracking applications.

In the following sections, we will detail the methodology of our model and introduce an automated protocol for training *SynCellFactory* on new data sets.

2.1 Motion Model

The goal of our motion model is to generate plausible spatial cell configuration and displacement. Our engine represents cells as 2D disks. The motion model initializes a population of cells with randomly sampled positions and sizes drawn from the cell area distribution of the training data. This population dynamically evolves following a stochastic Brownian motion, guided by statistics extracted from annotated real cell videos, which we model using a gamma distribution. Collision detection and resolution occur when two cells overlap. Using a hard sphere model, positions are adjusted with a repulsion vector until overlaps are resolved. By manipulating variables such as the number of cells, their movement speed, and the frequency of cell splitting events, we can tailor the complexity of the tracking task.

2.2 ControlNet

ControlNet [28] is a popular architecture for enabling the conditioning of text-to-image generative models. Our ControlNet uses the standard pre-trained Stable

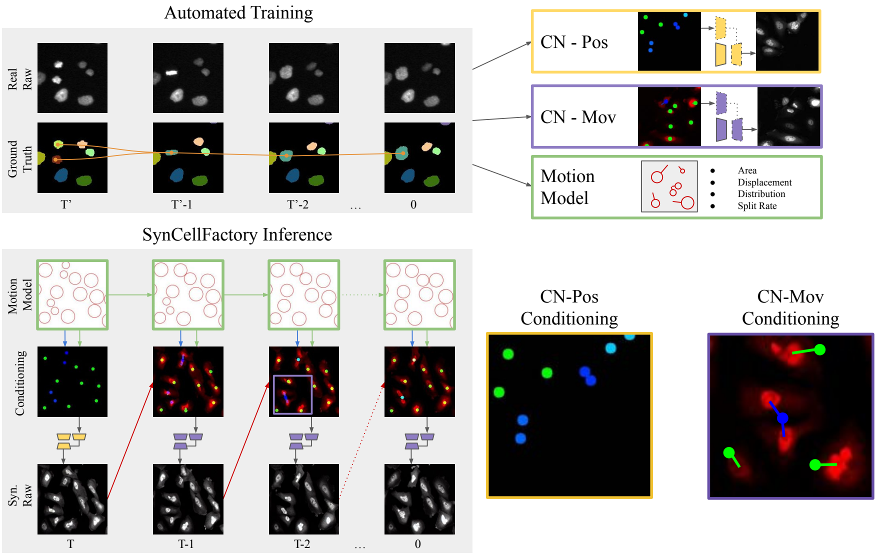

Fig. 2. *SynCellFactory* is a data augmentation pipeline designed to create unlimited high-quality synthetic raw video data and corresponding pseudo ground truth. It trains three key components using a small, and possibly sparsely labeled data set: Positional ControlNet (CN-Pos), Movement ControlNet (CN-Mov), and a 2D movement engine for realistic simulations. The process initiates in reverse, with the motion model generating a conditioning image at time T for CN-Pos. This image illustrates the expected centers of cells using colored dots, where each color signifies a specific cell state in the mitotic cycle: green during the interphase and blue during the different phases of cell division. CN-Pos then employs this information to generate a realistic frame for time T. Subsequently, CN-Mov assumes the role of producing the next frame $T-1$, using as conditioning an RGB image that combines the previously generated frame (in the red channel) with the projected positions and movement patterns (in green and blue channels). Derived from the motion model, these patterns represent each cell's trajectory from its current to its anticipated next position as a line connecting the two. By iteratively applying CN-Mov, *SynCellFactory* can efficiently produce time-lapse sequences of any desired length, suitable for training deep learning pipelines in cell tracking.

Diffusion v1.5 backbone trained on natural images, fine-tuned to the appearance of biological images. Training a ControlNet for latent text-to-image diffusion models uses triplets $(c_{\text{txt}}, c_{\text{img}}, i_{\text{tgt}})$. The text conditioning for our experiments was fixed to the prompt "cell, microscopy, image". In the following sections, we describe the specific details and c_{img} conditioning the CN-Pos and CN-Move models.

Positional ControlNet The positional ControlNet CN-Pos is tasked with drawing realistic looking cells conditioned on a given position. This model is used to generate the last frame of our synthetic videos and is the pre-trained backbone of the CN-Move model. At train time, the position map is constructed by computing the center coordinates for each cell given the corresponding detection ground truth. Each detection is represented in the conditioning image c_{img} as a disk with a fixed radius of $r = \frac{\sqrt{A_c/\pi}}{4}$, where A_c is the data set average cell area in pixels. The disks are colored according to the stage in the cell cycle, changing colors through the phases of Mitosis and reverting post-cytokinesis. At inference time, the conditioning image c_{img} is derived by the state simulated by the motion model, converting the center of simulated cells into color-coded disks.

Movement ControlNet CN-Mov is tasked with predicting frame $t-1$ conditioned on the frame at time t, the position map at time $t-1$, and the displacement vectors. Displacement vectors describe the movement of the cells between frames and are encoded as lines connecting the center position of cells.

2.3 SynCellFactory Inference

During the sampling process, we apply the trained ControlNets on the output of our motion module to generate realistic-looking videos (see Fig. 2). The inference process is initiated with CN-Pos generating frame t. Subsequently, CN-Mov iteratively takes the generated frame t and, in conjunction with the motion model, samples a new frame $t-1$. This iterative process continues until the desired video length of 12 frames is reached, which is double the minimum mitosis cycle duration of tested datasets.

2.4 Segmentation Pseudo Ground Truth

SynCellFactory produces only raw video, detection, and lineage ground truth. To address the challenge of not producing instance segmentation ground truth, we rely on Cellpose [23, 16], a renowned deep learning framework for cell segmentation, to create pseudo-ground truth segmentation. We fine-tuned a pre-trained model from Cellpose for 100 epochs using ground truth segmentation masks from our training data.

We integrated the motion model into the segmentation process to improve accuracy. When a ground truth detection is present without a corresponding segmentation mask, we generate one by drawing a circle with a radius $r = \sqrt{A_c/\pi}$, providing an approximate mask for cells that are challenging to segment. In cases where a segmentation mask does not overlap with any part of the generated detection ground truth, it is removed.

2.5 Automated Training

To reduce the domain-specific expertise required to train *SynCellFactory*, we propose a fully automated pipeline for training the ControlNets and the sampling from the motion module.

The only required hyperparameters that have to be manually specified to produce our synthetic videos are: the number of videos to be generated, the number of frames per video, and the characteristic length of the mitosis cycle.

Other parameters, such as movement statistics for the motion model, are automatically inferred from the raw data and ground truth annotations.

3 Experiments and Results

3.1 Data sets

To validate the proposed *SynCellFactory*, we use the publicly available Cell Tracking Challenge (CTC) [26, 13] [1] data sets.

For our experiments, we focus on the seven 2D data sets enumerated in Fig. 1. Each data set consists of two timelapses with full tracking annotations of ground truth and only partial hand-curated segmentation. The tracking ground truth includes detection and identity masks for each frame and the corresponding cell lineages. The number of frames per video ranges from 30 (Fluo-C2DL-Huh7) to 300 (PhC-C2DL-PSC).

Our experiments used a single time-lapse for the training and validation and one for testing tracking accuracy. Sample still frames from the mentioned data sets can be found in Fig. 1.

3.2 Experimental Setup

The training of CN-Pos and CN-Move follows the standard procedure as presented in [28]; the only major difference is that we also fine-tune the stable diffusion UNet decoder block. Since we could only access a single annotated timelaps for training our ControlNet, we relied on data augmentation. In particular, we found random cropping (h/2 x w/2) of the images and random 90 degree rotations beneficial. To evaluate the usefulness of *SynCellFactory* as a data augmentation strategy, we used it in conjunction with the state-of-the-art deep learning model EmbedTrack [12]. EmbedTrack uses CNNs to predict cell segmentation and tracking jointly and already incorporates standard data augmentation techniques during training. The generative data augmentation by *SynCellFactory* can therefore be seen as additional data augmentation.

Tracking Metric To evaluate the tracking performance of our trained models, we use the official tracking accuracy measure (TRA) provided by the Cell Tracking Challenge [14]. The TRA score is based on the concept of Acyclic Oriented Graph Matching AOGM. The TRA score is defined as $\mathrm{TRA} = 1 - \frac{\min(\mathrm{AOGM}, \mathrm{AOGM}_0)}{\mathrm{AOGM}_0}$, where AOGM represents the weighted sum of operations required to build the prediction graph and AOGM_0 represents the weighted sum of operations required to build the ground truth graph. Higher TRA scores indicate better tracking performance.

[1] http://celltrackingchallenge.net/2d-datasets/

Table 1. Tracking Accuracy Measure TRA (higher is better) obtained with and without our *SynCellFactory*. The proposed data augmentation increases tracking accuracy for all but one of the CTC tested data sets. The α mixing coefficient has been set to the optimal value for each data set. Error bars indicate the standard deviation over three runs. The number of tracking predictions underlying the TRA score are reported in supplementary material Tab. 1.

data set	W/o *SynCellFactory*	With *SynCellFactory*	α
Fluo-C2DL-Huh7	0.960 ± 0.002	**0.966** ± 0.003	0.66
Fluo-C2DL-MSC	0.624 ± 0.060	**0.685** ± 0.060	0.50
DIC-C2DH-HeLa	0.968 ± 0.001	**0.974** ± 0.001	0.80
Fluo-N2DH-GOWT1	0.980 ± 0.003	**0.989** ± 0.002	0.48
Fluo-N2DL-HeLa	0.939 ± 0.002	**0.981** ± 0.002	0.87
PhC-C2DL-PSC	0.958 ± 0.001	**0.960** ± 0.001	0.20
PhC-C2DH-U373	**0.938** ± 0.008	0.935 ± 0.007	0.87

Table 2. Comparison of the official CTC results. We improved the performance of EmbedTrack by using our data generation for all three data sets. Originally, EmbedTrack was inapplicable to Fluo-C2DL-Huh7 due to a shortage of segmentation masks for training. The generated data sets provided a sufficient amount of segmentation masks, enabling EmbedTrack to track Fluo-C2DL-Huh7.

data set	EmbedTrack [12]	Ours
Fluo-C2DL-Huh7	-	**0.920**
Fluo-C2DL-MSC	0.693	**0.703**
DIC-C2DH-HeLa	0.934	**0.943**

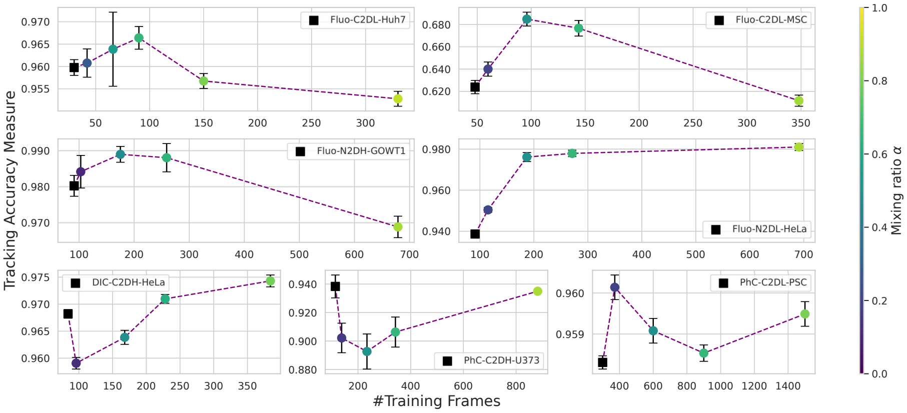

Fig. 3. Quantitative results according to the Tracking Accuracy Measure TRA (higher is better). We trained the EmbedTrack model without data augmentation (black square) and with different real and synthetic training data mixing ratios α. Here, one can observe that although *SynCellFactory* augmentation positively impacts the TRA score in all but one of the tested data sets, the correct choice of α is critical for the model we benchmarked. Error bars indicate the standard deviation over three runs.

3.3 Quantitative results

In all but one of the test data sets, our *SynCellFactory* data augmentation improved the tracking quality as measured by the TRA score; the results can be found in Tab. 1.

As shown in previous work using generative models for data augmentation [9, 25], the ratio between synthetic and real data $\alpha = \frac{\#\text{syn-frames}}{\#\text{syn-frames}+\#\text{real-frames}}$ in the training set is crucial. This is also true for the proposed methods; in Fig. 3, we show the tracking accuracy achieved using different mixing ratios α. Our experiments showed two behaviors between fluorescence microscopy data sets (denominated as Fluo-*) and all other microscopy modalities. In fluorescence datasets, TRA scores typically increased with the mixing ratio up to an optimal $0.5 < \alpha < 0.7$, then dropped, except in one dataset where improvement continued steadily. In Phase Contrast (PhC-*) and Differential Interference Contrast (DIC-*) experiments, most datasets initially showed a performance drop at low α, followed by consistent improvement at high $\alpha \sim 0.8$, with one exhibiting an initial increase, a subsequent drop, and then improvement.

CTC Results In addition to our standard experimental setup, we tested our strategy on the official cell tracking challenge evaluation data set. The CTC organizers evaluate the submissions on a private ground truth. We submitted the results for three data sets. Here, we trained *SynCellFactory* and the EmbedTrack model on all available training data and using the optimal mixing ratio α. The results are presented in Tab. 2 and show an improvement on all three data sets compared to those in [12].

4 Conclusion

Despite the positive results, *SynCellFactory* is not without limitations.

The current motion module in *SynCellFactory* is simplistic, focusing on basic cell movements. Future iterations of *SynCellFactory* could benefit from a more sophisticated motion module that can accurately model a broader range of biological behaviors and interactions. Addressing this limitation would enhance the model's utility in more complex biological environments. *SynCellFactory* can sample videos of arbitrary length, but there is a noticeable drop in quality for extended sequences with more than 30 frames. Future developments could focus on maintaining consistent quality across varying video lengths.

In conclusion, *SynCellFactory* represents a significant step forward in the field of biological data augmentation. Its ability to generate realistic and diverse data sets holds great potential for advancing deep learning-based cell tracking pipelines.

Code Availability. The code is publicly available at:
https://github.com/sciai-lab/SynCellFactory

Acknowledgments. This work is supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Projektnummer 240245660 - SFB 1129 and under Germany's Excellence Strategy EXC-2181/1 - 390900948 (the Heidelberg STRUCTURES Excellence Cluster).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Azizi, S., Kornblith, S., Saharia, C., Norouzi, M., Fleet, D.J.: Synthetic data from diffusion models improves imagenet classification. Transactions on Machine Learning Research (2023)
2. Burke, R.T., Orth, J.D.: Through the looking glass: Time-lapse microscopy and longitudinal tracking of single cells to study anti-cancer therapeutics. Journal of visualized experiments: JoVE (2016)
3. Chen, R.J., Lu, M.Y., Chen, T.Y., Williamson, D.F.K., Mahmood, F.: Synthetic data in machine learning for medicine and healthcare. Nature Biomedical Engineering (2021)
4. Collins, J.L., van Knippenberg, B., Ding, K., Kofman, A.V.: Time-lapse microscopy. In: Cell Culture, chap. 3. IntechOpen (2018)
5. Dorjsembe, Z., Pao, H.K., Odonchimed, S., Xiao, F.: Conditional diffusion models for semantic 3d medical image synthesis. Authorea Preprints (2023)
6. Du, Y., Zhao, Z., Song, Y., Zhao, Y., Su, F., Gong, T., Meng, H.: Strongsort: Make deepsort great again (2023)
7. Fernandez, V., Pinaya, W.H.L., Borges, P., Tudosiu, P.D., Graham, M.S., Vercauteren, T., Cardoso, M.J.: Can segmentation models be trained with fully synthetically generated data? In: International Workshop on Simulation and Synthesis in Medical Imaging. pp. 79–90. Springer (2022)
8. Hassan, S., Mujtaba, G., Rajput, A., Fatima, N.: Multi-object tracking: a systematic literature review. Multimedia Tools and Applications (2023)
9. He, R., Sun, S., Yu, X., Xue, C., Zhang, W., Torr, P., Bai, S., QI, X.: IS SYNTHETIC DATA FROM GENERATIVE MODELS READY FOR IMAGE RECOGNITION? In: The Eleventh International Conference on Learning Representations (2023)
10. Liu, Q., Gaeta, I.M., Zhao, M., Deng, R., Jha, A., Millis, B.A., Mahadevan-Jansen, A., Tyska, M.J., Huo, Y.: Asist: annotation-free synthetic instance segmentation and tracking by adversarial simulations. Computers in biology and medicine (2021)
11. Loewke, K.E., Pera, R.A.R.: The Role of Time-Lapse Microscopy in Stem Cell Research and Therapy, pp. 181–191 (2011)
12. Löffler, K., Mikut, R.: Embedtrack—simultaneous cell segmentation and tracking through learning offsets and clustering bandwidths. IEEE Access **10**, 77147–77157 (2022). https://doi.org/10.1109/ACCESS.2022.3192880
13. Maska, M., Ulman, V., Delgado-Rodriguez, P., Gomez-de Mariscal, E., Necasova, T., et al.: The cell tracking challenge: 10 years of objective benchmarking. Nature Methods (2023)
14. Matula, P., Maška, M., Sorokin, D.V., Matula, P., Ortiz-de Solórzano, C., Kozubek, M.: Cell tracking accuracy measurement based on comparison of acyclic oriented graphs. PLOS ONE (2015)

15. Meinhardt, T., Kirillov, A., Leal-Taixe, L., Feichtenhofer, C.: Trackformer: Multi-object tracking with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2022)
16. Pachitariu, M., Stringer, C.: Cellpose 2.0: how to train your own model. Nature Methods (2022)
17. Pinaya, W.H., Graham, M.S., Kerfoot, E., Tudosiu, P.D., Dafflon, J., Fernandez, V., Sanchez, P., Wolleb, J., da Costa, P.F., Patel, A., et al.: Generative ai for medical imaging: extending the monai framework. arXiv preprint arXiv:2307.15208 (2023)
18. Pinaya, W.H., Tudosiu, P.D., Dafflon, J., Da Costa, P.F., Fernandez, V., Nachev, P., Ourselin, S., Cardoso, M.J.: Brain imaging generation with latent diffusion models. In: MICCAI Workshop on Deep Generative Models (2022)
19. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
20. Sariyildiz, M.B., Alahari, K., Larlus, D., Kalantidis, Y.: Fake it till you make it: Learning transferable representations from synthetic imagenet clones. In: CVPR 2023–IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
21. Schwartz, M.S., Moen, E., Miller, G., Dougherty, T., Borba, E., Ding, R., Graf, W., Pao, E., Valen, D.V.: Caliban: Accurate cell tracking and lineage construction in live-cell imaging experiments with deep learning. bioRxiv (2023)
22. Serna-Aguilera, M., Luu, K., Harris, N., Zou, M.: Neural cell video synthesis via optical-flow diffusion (2022)
23. Stringer, C., Wang, T., Michaelos, M., Pachitariu, M.: Cellpose: a generalist algorithm for cellular segmentation. Nature Methods (2021)
24. Sun, L., Chen, J., Xu, Y., Gong, M., Yu, K., Batmanghelich, K.: Hierarchical amortized gan for 3d high resolution medical image synthesis. IEEE journal of biomedical and health informatics (2022)
25. Trabucco, B., Doherty, K., Gurinas, M.A., Salakhutdinov, R.: Effective data augmentation with diffusion models. In: The Twelfth International Conference on Learning Representations (2024)
26. Ulman, V., Maška, M., Magnusson, K.E.G., Ronneberger, O., Haubold, C., Harder, N., Matula, P., Matula, P., Svoboda, D., Radojevic, M., Smal, I., Rohr, K., Jaldén, J., Blau, H.M., Dzyubachyk, O., Lelieveldt, B., Xiao, P., Li, Y., Cho, S.Y., Dufour, A.C., Olivo-Marin, J.C., Reyes-Aldasoro, C.C., Solis-Lemus, J.A., Bensch, R., Brox, T., Stegmaier, J., Mikut, R., Wolf, S., Hamprecht, F.A., Esteves, T., Quelhas, P., Demirel, O., Malmström, L., Jug, F., Tomancak, P., Meijering, E., Muñoz-Barrutia, A., Kozubek, M., Ortiz-de Solorzano, C.: An objective comparison of cell-tracking algorithms. Nature Methods (2017)
27. Zargari, A., Lodewijk, G.A., Mashhadi, N., Cook, N., Neudorf, C.W., Araghbidikashani, K., Hays, R., Kozuki, S., Rubio, S., Hrabeta-Robinson, E., Brooks, A., Hinck, L., Shariati, S.A.: Deepsea is an efficient deep-learning model for single-cell segmentation and tracking in time-lapse microscopy. Cell Reports Methods (2023)
28. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023)
29. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision (2017)