CELL PHANTOM VIDEO GENERATION IN ELLIPTICAL FOURIER DESCRIPTOR DOMAIN

Francesco Benedetto *Roberto Basla* *Luca Magri* *Giacomo Boracchi*

Department of Electronics Information and Bioengineering, Politecnico di Milano, Italy

ABSTRACT

Training Deep Neural Networks for tracking individual cells in biomedical videos requires a large amount of annotated data. The annotation of videos for cell tracking is very time consuming and often requires domain expertise; this explains the limited availability of public annotated data to address important medical problems like tissue repair or cancer treatment. Generating synthetic videos along with their Ground Truth annotations is a promising solution that relies, as a foundational first step, on the synthesis of single cell annotations (or *phantoms*). Phantoms need to be time consistent, as they have to replicate biological processes that are specific to the cell types. In this work, we propose a novel framework for generating videos of cell phantoms in the Elliptical Fourier Descriptors (EFDs) domain, a compact and geometrically interpretable representation for 2D closed contours. We represent the cell phantom evolution as a multivariate time series of EFD coefficients, introducing a strong prior for cell morphology and enabling the efficient generation of sequences that evolve coherently in time. Our experimental validation proves that modelling the temporal evolution in EFD space enables the generation of biologically plausible phantom videos. Our method can be used in generative pipelines for synthesizing annotated data for cell tracking, thus strongly mitigating the annotation effort for creating new datasets. Our code is available for download here: https://github.com/FrancescoBenedetto99/efd-cell-video-gen.

***Index Terms*—** Cell Phantom, Cell Tracking, Biomedical Video Generation

1 Introduction

Cell tracking is the problem of analysing videos of living cells (Fig. 1a) and segment and track each cell instance along the frames [1] (Fig. 1b). State-of-the-art solutions in cell tracking are represented by Deep Learning (DL) models [2], which are becoming fundamental tools to study the properties of living cells and to understand tissue development and repair, cancer treatment and drug development [3]. Unfortunately, training DL models for cell tracking requires large datasets annotated by domain experts, which are very time consuming to prepare. Moreover, the variability in cell types (or *cell lines*) and tissue preparation procedures for microscopy imaging may require annotating videos for each setting. An appealing perspective to train cell tracking models consists in synthetically generating custom annotated datasets, which is nowadays possible given the advances in data generation. In this work, we address the problem of generating Ground Truth (GT) videos of binary masks for single cells (also called *cell phantoms*, Fig. 1c) in cell tracking settings.

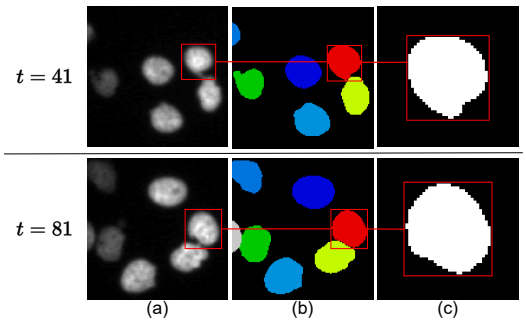

Fig. 1: Frames from the Fluo-N2DL-HeLa dataset [6] at different time steps t. (a) Histological video, (b) cell tracking Ground Truth, and (c) a cell phantom. The bounding box highlights the same cell at different times $t = 41$, $t = 81$.

Research on the generation of synthetic annotated datasets for cell segmentation and tracking follows two main paradigms: *end-to-end* and *pipeline-based*. End-to-end approaches use Deep generative models (e.g., GANs or diffusion models) to directly synthesize complete microscopy images and videos with their annotations in a single step [4]. Although achieving high visual realism, these data-driven, black-box methods require large annotated datasets themselves and offer limited control over individual cell properties (morphology, positioning, movements and texture), losing control on the generation process. On the other hand, pipeline-based approaches explicitly decompose generation into multiple stages, allowing the injection of domain expertise in the generation process. Pipelines generating images are typically made of three steps: (1) phantom generation, (2) spatial positioning, and (3) texture image synthesis, enabling better control over each generation aspect [5]. In videos, a straightforward extension of these pipelines for the accounting of cell evolution and movement is (1) *phantom videos generation*, (2) *composition, movement and interaction modelling*, and (3) *video texture synthesis*. This paper proposes a method to address the essential first step.

So far, both end-to-end and pipeline-based approaches focused primarily on the generation of realistic textures for images and videos, often paying less attention to phantom generation. However, phantoms determine the quality of video-level annotations and consequently of the generated dataset. Phantoms must capture the macroscopic properties of the cell, namely the shape and size, and need to be morphologically realistic. Thus, phantoms should consist of a single, closed-contour and *simply connected* component (i.e., without disjoint parts nor holes), and should evolve coherently over time. The mainstream approaches [7, 8, 9] use phantoms generated by a Statistical Shape Model (SSM). In particular, CellCycleGAN [7] models a Markov process over the cell life cycle stages (that go from birth to division, or mitosis) to define more precisely the morphology, but it requires phantoms to be additionally annotated with

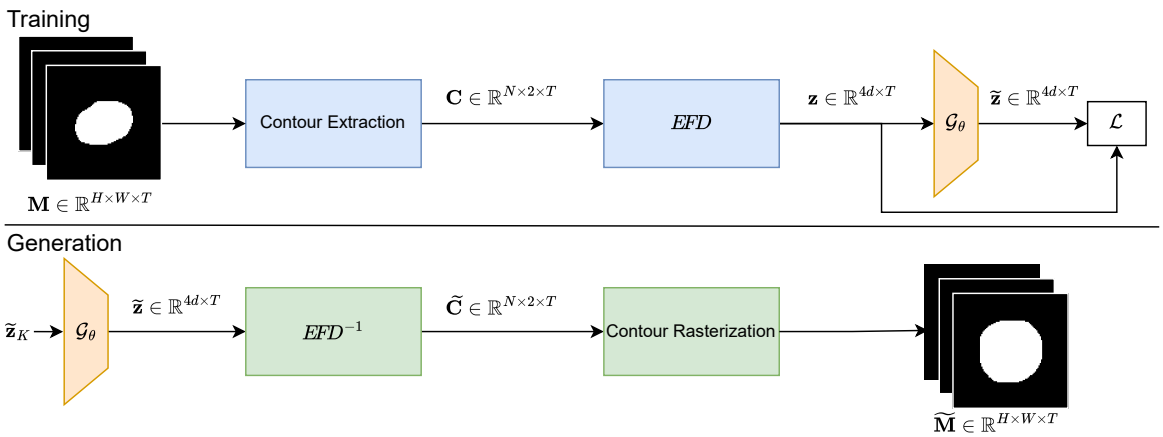

Fig. 2: Our pipeline for phantom video generation. (Top) During training we extract a contour video $\mathbf{C}$ from each binary phantom video $\mathbf{M}$, then encode $\mathbf{C}$ in the EFD domain to obtain the time series $\mathbf{z}$. We train a generative model $\mathcal{G}_\theta$ minimizing the loss $\mathcal{L}$ between real EFD time series $\mathbf{z}$ and generated time series $\widetilde{\mathbf{z}}$ to learn the temporal evolution of cell phantoms in EFD space. (Bottom) During generation, we unconditionally sample synthetic EFD time series $\widetilde{\mathbf{z}}$ from the trained model starting from a noise vector $\widetilde{\mathbf{z}}_K$. We decode $\widetilde{\mathbf{z}}$ into a synthetic contour $\widetilde{\mathbf{C}}$ which, once rasterized, corresponds to a synthetic mask video $\widetilde{\mathbf{M}}$.

their cell state. Another limitation of [7] is the description of the coordinates of the phantom's contours with a Principal Component Analysis (PCA), often resulting in a completely data-driven process that is constrained by the dataset and lacks an interpretable geometric structure. Our work addresses the limitations of SSM-based phantom generation by proposing a generative framework that learns to synthesize videos of individual cell phantoms according to the cellular dynamics. Our method generates realistic phantom videos, enabling the following stages of a cell tracking dataset generation pipeline to model cell positioning and texture synthesis.

In particular, we generate phantom videos in the domain of Elliptical Fourier Descriptors (EFDs), which provide a compact representation of cell phantoms. We encode each video frame in few coefficients representing ellipses that approximate the cell's contour. Therefore, we obtain a collection of multivariate time series in the compact EFD space describing the cellular evolution that we use for training Diffusion-TS [10], a Diffusion Model specific for multivariate time series. Diffusion-TS explicitly disentangles the trend and seasonality components of a time series, learning the long-term deformations of each cell phantom and its higher-frequency shape variations. We then unconditionally synthesize EFD time series that, once reconstructed through the inverse EFD transform, correspond to realistic-looking cell phantom evolutions in pixel space. Indeed, the EFD representation has great advantages for cell phantom modelling: it naturally enforces geometrical constraints such as the presence of only one simply connected component, allows the representation of phantoms using few parameters, and enables an effective generative model to operate on simpler signals. To the best of our knowledge, our work is the first to apply EFDs in video generation settings, with prior research limited to static phantom generation in 2D or 3D [11, 12].

In our experiments, we verify that our approach enjoys two fundamental properties: morphological statistical consistency and temporal coherence. We show that phantom videos generated in EFD space closely follow the evolution of morphological descriptors of real data in time. We compare our phantoms against the SSM generative method in CellCycleGAN, highlighting more realistic morphological properties in both a quantitative evaluation and a qualitative comparison over a dataset used in both works. Additionally, we generate phantom videos starting from the Fluo-N2DL-HeLa dataset of the Cell Tracking Challenge [2] (CTC Dataset), assessing the generalization power of our method over differently-evolving cell lines. The morphological comparison with real data suggests that our phantom generation method can be integrated in a pipeline-based framework for generating diverse cell tracking datasets.

2 Problem Formulation

We focus on phantom video generation, i.e., the generation of the binary mask of a single, simply connected component representing the cell's morphology in time. We assume that a real dataset D_{real} of cell phantom videos $\mathbf{M}$ is provided:

$$D_{\text{real}} = \{\mathbf{M}^{(i)}\}_{i=1}^{S_{\text{real}}}, \tag{1}$$

where each video $\mathbf{M}^{(i)}$ is a sequence of 2D binary images containing a single phantom each:

$$\mathbf{M}^{(i)} = (M_t^{(i)})_{t=1}^{T_i}, \quad M_t^{(i)} \in \{0,1\}^{H\times W}, \tag{2}$$

where H and W are the spatial dimensions, and $\mathbf{M}^{(i)}$ is a tuple of binary masks with T_i frames. We assume a limited number of S_{real} samples is available as in realistic biomedical scenarios. Our objective is to generate a synthetic dataset:

$$D_{\text{synth}} = \{\widetilde{\mathbf{M}}^{(j)}\}_{j=1}^{S_{\text{synth}}}, \quad \widetilde{\mathbf{M}}^{(j)} = (\widetilde{M}_t^{(j)})_{t=1}^{T_j}, \tag{3}$$

with $\widetilde{M}_t^{(j)} \in \{0,1\}^{H\times W}$, such that $S_{\text{synth}} \gg S_{\text{real}}$. Additionally, phantom variations across frames should be coherent with the morphology of the specific cell line in D_{real} and the temporal interval between two consecutive frames. For ease of notation, in the following we omit the dataset indices (i and j).

3 Method

As illustrated in Fig. 2, our method is organized in two main phases: model training and synthetic dataset generation. For training, we first extract the contour from each frame of a phantom video and encode it in EFD space, using this representation to train a generative model that learns the phantom evolution in time. During generation, we unconditionally generate new EFD time series to be converted first into contours and, finally, in phantom videos.

3.1 EFD Encoding of Phantom Videos

We map each mask video $\mathbf{M}$ containing a single cell phantom in a multivariate time series $\mathbf{z} \in \mathbb{R}^{4d \times T}$ where each frame is encoded using d ellipses that approximate the cell contour (Fig. 3). To this purpose, we first apply a contour extraction operator $\mathcal{C}$ that, for each frame at time steps $t \in [1, \dots, T]$, *(i)* extracts the contour pixels that represent a closed curve, and *(ii)* applies geodesic uniform resampling via arc-length parametrization for obtaining a better spectral representation [12]. This process maps each binary frame M_t to the set of (x, y) coordinates of the N contour points of the shape:

$$\mathcal{C} : \{0,1\}^{H \times W} \to \mathbb{R}^{N \times 2}. \tag{4}$$

Thus, we obtain the contour video

$$\mathbf{C} = (\mathcal{C}(M_t))_{t=1}^{T}, \qquad \forall M_t \in \mathbf{M}, \tag{5}$$

where $\mathbf{C} \in \mathbb{R}^{N \times 2 \times T}$. We take $N = 128$ to select enough details for contour sampling.

Cell contours $\mathbf{C}$ are closed loops with the points' coordinates that are continuous, periodic functions $x(s)$ and $y(s)$ of the normalized arc-length parameter $s \in [0, 2\pi]$. This periodicity allows us to decompose the shape using the Elliptical Fourier Transform *EFD* [13]. For each harmonic order n, the specific frequency component of $x(s)$ (weighted by coefficients $\mathrm{a}_n, \mathrm{b}_n$) and $y(s)$ (weighted by $\mathrm{c}_n, \mathrm{d}_n$) are combined. While these components represent simple 1D oscillations individually, their superposition in the 2D plane traces a unique ellipse. The full contour is approximated by summing the centroid (A_0, C_0) and all the harmonic ellipses up to a truncation order d which represents the number of selected ellipses:

$$\begin{bmatrix} x(s) \\ y(s) \end{bmatrix} = \underbrace{\begin{bmatrix} A_0 \\ C_0 \end{bmatrix}}_{\text{Centroid}} + \sum_{n=1}^{d} \underbrace{\begin{bmatrix} \mathrm{a}_n \cos(ns) + \mathrm{b}_n \sin(ns) \\ \mathrm{c}_n \cos(ns) + \mathrm{d}_n \sin(ns) \end{bmatrix}}_{n\text{-th harmonic ellipse}}. \tag{6}$$

Fig. 3 shows this EFD decomposition: the sum in (6) corresponds to a chain of epicycles where the center of the harmonic ellipse e_n of order n travels along the perimeter of the ellipse e_{n-1} (see red box). The first harmonic captures the overall shape and orientation (best-fitting ellipse e_1), while higher-order harmonics progressively add finer morphological details. In our encoding, we translate the center of e_1 to the origin, namely we set $A_0 = C_0 = 0$ for invariance to the cell position. Further details on EFD transforms can be found in [13].

We encode the phantom contour at frame t as a spectral vector $z_t \in \mathbb{R}^{4d}$ by concatenating the EFD coefficients:

$$z_t = [\mathrm{a}_1, \mathrm{b}_1, \mathrm{c}_1, \mathrm{d}_1, \dots, \mathrm{a}_d, \mathrm{b}_d, \mathrm{c}_d, \mathrm{d}_d]^\top. \tag{7}$$

The entire video evolution is thus mapped to the multivariate time series $\mathbf{z} = [z_1, \dots, z_T] \in \mathbb{R}^{4d \times T}$. We normalize values in $\mathbf{z}$ between 0 and 1 for easier training of the generative model. We also determine the number of ellipses d via Power Spectral Density (PSD) analysis on the training set, selecting the minimum d that retains 99.99% of the cumulative power, yielding $d = 9$.

This EFD transform offers two key advantages. First, it reduces the dimensionality of each frame by mapping the $H \times W$ pixel grid into a compact descriptor space of size $4d$ for a total of $4 \cdot 9 = 36$ parameters while a 128×128 mask corresponds to 16384 pixels. Second, it enforces a topological prior: unlike raw segmentation masks which may contain holes or disconnected artifacts due to noise, the truncated Fourier reconstruction is guaranteed to be smooth and yield a closed boundary with a simply connected phantom. The encoding process is illustrated in Fig. 4 (top), where the contour extraction and *EFD* decomposition yield a multivariate time series $\mathbf{z} \in \mathbb{R}^{4d \times T}$ corresponding to the phantom video.

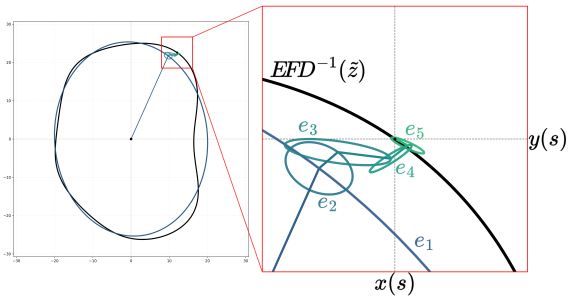

Fig. 3: The *EFD* transform parametrizes the points of each phantom's contour as a function of $s \in [0, 2\pi]$, representing it through the coordinates $x(s)$ and $y(s)$. Then, it traces a chain of ellipses $(e_1 \dots e_5)$ that add increasing detail to the reconstructed contour (in black). The contour is centered in $(0, 0)$ for invariance with respect to the position.

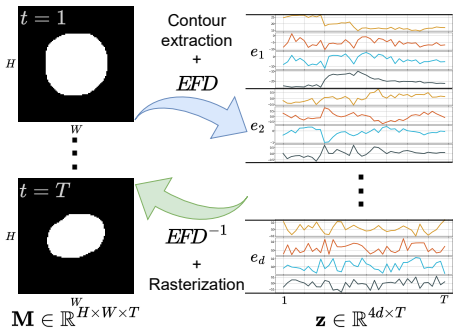

Fig. 4: We encode each mask video $\mathbf{M}$ into a time series representation $\mathbf{z}$ by first extracting each frame's contour video $\mathbf{C}$ and then encoding it with the *EFD* transformation. The multivariate time series $\mathbf{z}$ encodes d ellipses $(e_1 \dots e_d)$, each represented by four coefficients for each time step t. The inverse transformation EFD^{-1}, coupled with the contour rasterization, reconstructs the video.

3.2 Modelling EFD Multivariate Time Series

To generate realistic phantom evolutions, we model the probability distribution of time series in EFD space. We employ the Diffusion-TS framework [10], a state-of-the-art Diffusion Model designed for time series that is particularly effective as it decouples the trend (or long-term variation) and seasonality (or periodic fluctuations) components. As per the Diffusion Model paradigm, the training involves a forward diffusion process that gradually adds Gaussian noise to a real EFD multivariate time series $\mathbf{z}_0 \in \mathbb{R}^{4d \times T}$ over K diffusion steps. At step k, the noisy signal $\mathbf{z}_k$ is defined as:

$$\mathbf{z}_k = \sqrt{\bar{\alpha}_k}\mathbf{z}_0 + \sqrt{1 - \bar{\alpha}_k}\epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}) \tag{8}$$

where $\bar{\alpha}_k$ follows a cosine noise schedule to preserve signal structure until later steps, and $\mathbf{z}_k$ has the same dimensionality as $\mathbf{z}_0$.

The generative model $\mathcal{G}_\theta$ learns to reverse the diffusion process through a Transformer encoder-decoder architecture but, unlike Diffusion Models for images that at iteration k predict the noise ϵ to be removed, Diffusion-TS directly predicts the clean signal $\hat{\mathbf{z}}_0$ from the noisy input $\widetilde{\mathbf{z}}_k$. For the next diffusion step, the model computes the signal $\widetilde{\mathbf{z}}_{k-1}$ as a weighted linear combination of $\hat{\mathbf{z}}_0$, $\widetilde{\mathbf{z}}_k$ and a Gaussian noise term for stochasticity. Moreover, the model ensures high fidelity in the spectral domain by optimizing a hybrid loss function $\mathcal{L}$. This loss promotes the disentangling of temporal dynamics by combining a Mean Squared Error in the time domain with a loss term based on the Fast Fourier Transform $\mathcal{F}$ applied to the signal as per the original work [10]:

Table 1: Evaluation datasets information. We evaluate the properties of the generated phantoms against the test set.

Dataset	$H \times W$	T	S_{real}	#Train	#Test
CellCycleGAN dataset	96×96	40	326	258	68
CTC dataset	128×128	50	2954	2652	302

$$\mathcal{L} = \mathbb{E}_{k,\mathbf{z}_0}\left[\lambda_1 ||\mathbf{z}_0 - \hat{\mathbf{z}}_0||^2 + \lambda_2 ||\mathcal{F}(\mathbf{z}_0) - \mathcal{F}(\hat{\mathbf{z}}_0)||^2\right], \quad (9)$$

where $\hat{\mathbf{z}}_0 = \mathcal{G}_\theta(\mathbf{z}_k, k)$ is the reconstructed time series.

3.3 Synthetic Video Generation

During inference (Fig. 2, bottom), generation starts by sampling a noise seed $\widetilde{\mathbf{z}}_K \in \mathbb{R}^{4d \times T}$ representing a sequence of pure Gaussian noise $\widetilde{\mathbf{z}}_K \sim \mathcal{N}(0, \mathbf{I})$. Then, $\mathcal{G}_\theta$ iteratively denoises $\widetilde{\mathbf{z}}_K$ for $K = 200$ steps, obtaining a synthetic EFD time series $\widetilde{\mathbf{z}}$. We then apply the inverse transformation EFD^{-1} to map the generated time series $\widetilde{\mathbf{z}}$ into the contour coordinates $\widetilde{\mathbf{C}} \in \mathbb{R}^{N \times 2 \times T}$, which are finally rasterized into binary mask videos $\widetilde{\mathbf{M}} \in \{0,1\}^{H \times W \times T}$ (Fig. 4, bottom).

4 Experiments

We assess the realism of our generated phantoms by comparing their morphological properties against real phantom videos. We use the Compyda tool [14] to quantitatively compare real and synthetic datasets.

4.1 Datasets and Preprocessing

In order to evaluate the generation power of our method, we generate phantoms starting from two different datasets described in Table 1.

CellCycleGAN dataset This dataset [15] contains annotated videos of cell evolutions with both the phantom and the cell cycle stage. All phantom videos in the dataset are synchronized over a common cell state. However, during and after mitotic events, the annotation follows only the daughter cell closest to the parent's centroid, discarding the other. Consequently, each video $\mathbf{M}$ represents a continuous evolution of a cell phantom which remains a single simply connected component throughout the $T = 40$ frames. In Fig. 6 (Real) and 7 (Test set), videos synchronized over cell states exhibit an area drop due to mitotic events. Since the original data consists of centered cell phantom videos, these data perfectly align with our problem formulation, allowing modelling the entire cell cycle without explicitly handling the division of cells.

We compare our solution against CellCycleGAN [7], which is trained on this dataset, enabling a fair comparison. Since the dataset groups cell phantom videos in 7 GT histological sequences, we use 5 for training and 2 for testing corresponding to a 80/20 train-test split.

CTC dataset The Fluo-N2DL-HeLa dataset [6] from the Cell Tracking Challenge contains annotated biomedical videos (see Fig. 1a and 1b) with 92 frames. We extract cell phantoms from the given *Silver Truth* provided in the CTC and filter out incomplete instances like cells touching the image border by more than 10 pixels (deemed partially out of frame) or having a high overlap with other cells. This processing removes incomplete cells which we identify by computing the solidity $\mathtt{s}$ and the overlap ratio $\mathtt{o}_{i,i'}$ between two instances i and i' as:

$$\mathtt{s}_i = \frac{|M_i|}{|\text{ConvHull}(M_i)|}, \quad \mathtt{o}_{i,i'} = \frac{|\text{ConvHull}(M_i) \cap M_{i'}|}{|\text{ConvHull}(M_i)|}. \quad (10)$$

We discard cells for having high overlap if $\mathtt{s}_i < 0.969$ and $\mathtt{o}_{i,i'} > 0.017$. We manually set these thresholds to retain well-segmented cells while avoiding excessive data reduction. Since this outlier removal process may interrupt the cell's temporal evolution, we extract uninterrupted blocks of 50 frames for each phantom using a sliding window approach. From the original 924 phantom videos of varying lengths, this extraction yields 2954 videos with $T = 50$ (see Table 1). Differently from the CellCycleGAN dataset, phantoms do not continue after mitosis and cell states are not synchronized. Therefore, we model the mitotic cycle only up to the cell division.

The CTC dataset also includes lineage information, i.e., the tracking GT that includes relationships between parent and daughter cells in the video. To prevent data leakage, we account for lineages in the train-test split assigning entire lineages to one split. This process guarantees that no cell track appears in both train and test since all descendants of a cell belong to the same split as their ancestors.

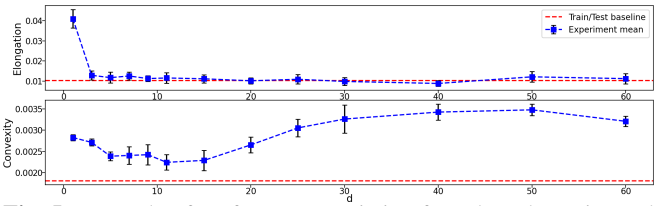

Fig. 5: Example of performance variation for selected metrics and values of $d \in [1, 60]$. Providing enough harmonics, our method reaches a plateau of performance close to the real data baseline in the elongation case (top), while convexity reports a minimum close to the value of $d = 9$ found by our PSD analysis (Section 3.1).

4.2 Evaluation Metrics

We compare generated and real datasets using the Compyda framework [14], which computes morphological features over time from phantom videos and compares the distance between their time series (DIFF) also after Dynamic Time Warping (DTW) alignment [16]. DIFF provides an absolute measure of dissimilarity between the two time series over time and evaluates whether the method successfully generates samples within the correct range of values. However, it does not capture whether the features follow the same temporal pattern. The DTW metric measures whether the morphological feature signals exhibits similar dynamics, regardless of frame-by-frame similarity. Since phantom videos can start from different phases of the cells' life cycles, DTW is a more realistic figure of merit.

We select a subset of features computed by Compyda capturing relevant morphological aspects that need to be modelled by a generative process. These features correspond to the phantom's *area* (in pixels), *roundness* (how much it resembles a circle, between 0 and 1), and *elongation* (the ratio between the major and the minor axes). We additionally report *convexity* as the ratio between the phantom's area and the area of its convex hull.

4.3 Comparative Results

We compare our phantom video generation framework against the SSM method used in CellCycleGAN. To ensure a fair comparison, we use the same preprocessing and generative pipeline of [7], with the train/test split from Section 4.1. Since this pipeline generates all cell tracking GTs as a single video, we extract each phantom in the same format as the original dataset (Section 4.1). Additionally, we compare the training set and the test set using the DIFF and DTW metrics. This real-against-real comparison represents a reference for the metrics (*Train* in Table 2). Since the diffusion process operates in the low-dimensional EFD space (36×40 parameters) rather than pixel space ($96 \times 96 \times 40$), our generation is highly efficient. It

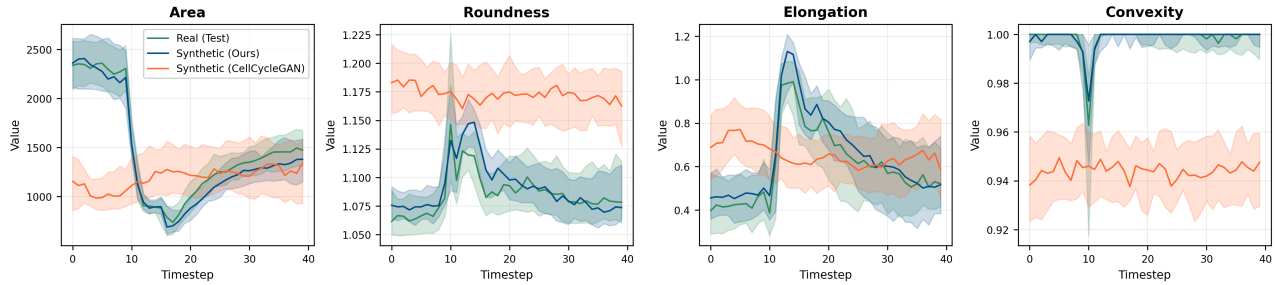

Fig. 6: Temporal evolution of selected morphological features on the CellCycleGAN dataset. Solid lines represent population means over synchronized phantom videos, while shaded regions show interquartile ranges. We report the real test set evolution (green), the phantoms generated by our method (blue) and generated by the SSM (orange). The sudden drop in the area metric for the test set corresponds to mitotic events.

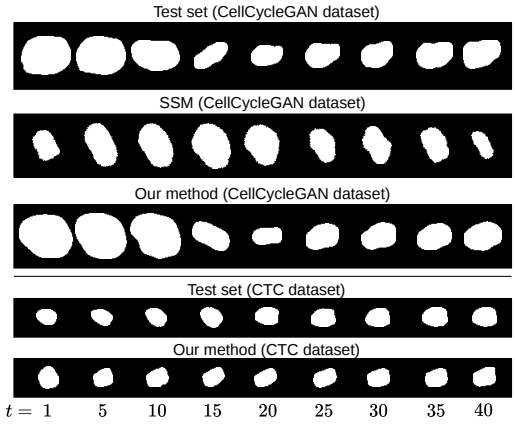

Fig. 7: Real and synthetic cell phantom videos at intervals of 5 frames. In the CellCycleGAN dataset, mitoses happen after $t = 10$. Our method replicates the cell cycle dynamics as modelled in the training set and generates more regular and realistic phantoms than CellCycleGAN's SSM [7]. We can also replicate the phantom evolution in the CTC dataset. In both cases, thanks to the *EFD* encoding phase, our contours are smooth.

takes approximately 19 minutes to train Diffusion-TS and 0.09s per phantom video generation (inference plus EFD-to-video conversion time) on a desktop computer equipped with an Intel i9 13900k CPU and a nVidia RTX 3070 GPU. The generation time is in line with the SSM method (0.08s per phantom video on the same hardware setup).

Our approach achieves higher performance over all metrics computed by Compyda. Table 2 shows how our synthetic phantom videos exhibit morphological features that are closer to the *Train* baseline with respect to the SSM method. We replicate the generation and evaluation 10 times and report in Table 2 the mean metric value and the standard deviation of the mean. We additionally perform a t-test between the set of metrics computed over our synthetic dataset and the one generated by the SSM, always obtaining a p-value lower than 1E-10. Notably, in the CTC dataset we achieve an elongation DIFF and DTW lower than the real-against-real reference. This indicates that our method can generalize well enough over the morphological properties and avoid overfitting. Fig. 7 shows frames from the test set, from CellCycleGAN's SSM and from our method, highlighting a higher variability of videos generated by our framework. We report frames from the CTC dataset as well, showing that generated phantoms are coherent with the test set evolution in both cases. The contours of phantoms generated

Table 2: We report DIFF and DTW over selected morphological features for each model against the test set (lower is better). We report the metrics computed between the real train and the test sets (*Train*), showing that our method generates phantoms that are morphologically close to the reference for both datasets. We report the mean and the standard deviation over experiments replicated 10 times for both our method and the SSM baseline, reporting better metrics in all cases. Rows marked with (*) correspond to a t-test yielding a p-value < 1E-10, showing that the difference between the distributions of our method and SSM is statistically significant.

Feature	**Metric**	CellCycleGAN dataset			CTC dataset	
		Train	**SSM**	**Ours**	***Train***	**Ours**
Area	DIFF ↓	71.17	479.91 ± 10.66	$\mathbf{71.48 \pm 5.26}$*	47.25	52.40
	DTW ↓	32.79	248.28 ± 7.17	$\mathbf{34.61 \pm 2.91}$*	23.49	28.45
Roundness	DIFF ↓	0.04	0.12 ± 0.00	$\mathbf{0.01 \pm 0.00}$*	0.00	0.00
	DTW ↓	0.01	0.09 ± 0.00	$\mathbf{0.01 \pm 0.00}$*	0.00	0.00
Elongation	DIFF ↓	0.13	0.58 ± 0.02	$\mathbf{0.15 \pm 0.01}$*	0.18	0.11
	DTW ↓	0.07	0.28 ± 0.01	$\mathbf{0.08 \pm 0.00}$*	0.08	0.03
Convexity	DIFF ↓	0.0	0.06 ± 0.00	$\mathbf{0.00 \pm 0.00}$*	0.00	0.00
	DTW ↓	0.0	0.06 ± 0.00	$\mathbf{0.00 \pm 0.00}$*	0.00	0.00

by our method are also smoother as a result of the EFD encoding. We also performed an ablation study repeating 10 times generation and dataset evaluation by changing the number of preserved EFD coefficients $d \in [1, 60]$, showing that few harmonics are enough to capture the cell dynamics, and that the use of a large d can be detrimental to some features (Fig. 5). Finally, Fig. 6 shows that our phantom videos follow closely on the morphological evolution of real data, as the sudden drop in area and convexity, together with the peaks of roundness and elongation, highlight a mitotic event around time step $t = 10$.

5 Conclusions and Future Work

We presented a parametric generative framework for cell phantom video generation using Elliptical Fourier Descriptors and Diffusion-TS. Results demonstrate that our framework effectively learns both morphological properties and temporal dynamics patterns of cells, mimicking the distribution of real data. Our work can be integrated into generative pipelines for cell tracking datasets that include composition, movement simulation and texture synthesis. Future work aims at extending our framework to explicitly handle mitotic events (generating both daughter cells during cell divisions) and at generating 3D phantom videos using Spherical Harmonics Decomposition [17].

6 References

[1] Anirban Chakraborty and Amit Roy-Chowdhury, "A conditional random field model for tracking in densely packed cell structures," in *2014 IEEE International Conference on Image Processing (ICIP)*, Paris, France, Oct. 2014, pp. 451–455, IEEE.

[2] Martin Maška, Vladimír Ulman, Pablo Delgado-Rodriguez, Estibaliz Gómez-de Mariscal, Tereza Nečasová, Fidel A. Guerrero Peña, Tsang Ing Ren, Elliot M. Meyerowitz, Tim Scherr, Katharina Löffler, Ralf Mikut, Tianqi Guo, Yin Wang, Jan P. Allebach, Rina Bao, Noor M. Al-Shakarji, Gani Rahmon, Imad Eddine Toubal, Kannappan Palaniappan, Filip Lux, Petr Matula, Ko Sugawara, Klas E. G. Magnusson, Layton Aho, Andrew R. Cohen, Assaf Arbelle, Tal Ben-Haim, Tammy Riklin Raviv, Fabian Isensee, Paul F. Jäger, Klaus H. Maier-Hein, Yanming Zhu, Cristina Ederra, Ainhoa Urbiola, Erik Meijering, Alexandre Cunha, Arrate Muñoz-Barrutia, Michal Kozubek, and Carlos Ortiz-de Solórzano, "The Cell Tracking Challenge: 10 years of objective benchmarking," *Nature Methods*, vol. 20, no. 7, pp. 1010–1020, July 2023.

[3] Reza Yazdi and Hassan Khotanlou, "A survey on automated cell tracking: challenges and solutions," *Multimedia Tools and Applications*, vol. 83, no. 34, pp. 81511–81547, Mar. 2024.

[4] P. Dimitrakopoulos, G. Sfikas, and C. Nikou, "ISING-GAN: Annotated Data Augmentation with a Spatially Constrained Generative Adversarial Network," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, Iowa City, IA, USA, Apr. 2020, pp. 1600–1603, IEEE.

[5] Roberto Basla, Loris Giulivi, Luca Magri, and Giacomo Boracchi, "An Expert-Driven Data Generation Pipeline for Histological Images," in *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, Athens, Greece, May 2024, pp. 1–5, IEEE.

[6] Beate Neumann, Thomas Walter, Jean-Karim Hériché, Jutta Bulkescher, Holger Erfle, Christian Conrad, Phill Rogers, Ina Poser, Michael Held, Urban Liebel, Cihan Cetin, Frank Sieckmann, Gregoire Pau, Rolf Kabbe, Annelie Wünsche, Venkata Satagopam, Michael H. A. Schmitz, Catherine Chapuis, Daniel W. Gerlich, Reinhard Schneider, Roland Eils, Wolfgang Huber, Jan-Michael Peters, Anthony A. Hyman, Richard Durbin, Rainer Pepperkok, and Jan Ellenberg, "Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes," *Nature*, vol. 464, no. 7289, pp. 721–727, Apr. 2010.

[7] Dennis Bahr, Dennis Eschweiler, Anuk Bhattacharyya, Daniel Moreno-Andres, Wolfram Antonin, and Johannes Stegmaier, "Cellcyclegan: Spatiotemporal Microscopy Image Synthesis Of Cell Populations Using Statistical Shape Models And Conditional Gans," in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, Nice, France, Apr. 2021, pp. 15–19, IEEE.

[8] Dennis Eschweiler, Malte Rethwisch, Mareike Jarchow, Simon Koppers, and Johannes Stegmaier, "3D fluorescence microscopy data synthesis for segmentation and benchmarking," *PLOS ONE*, vol. 16, no. 12, pp. e0260509, 2021.

[9] Rüveyda Yilmaz, Dennis Eschweiler, and Johannes Stegmaier, "Annotated Biomedical Video Generation Using Denoising Diffusion Probabilistic Models and Flow Fields," in *Simulation and Synthesis in Medical Imaging*, Virginia Fernandez, Jelmer M. Wolterink, David Wiesner, Samuel Remedios, Lianrui Zuo, and Adrià Casamitjana, Eds., vol. 15187, pp. 197–207. Springer Nature Switzerland, Cham, 2025, Series Title: Lecture Notes in Computer Science.

[10] Xinyu Yuan and Yan Qiao, "Diffusion-TS: Interpretable Diffusion for General Time Series Generation," in *International Conference on Learning Representations*, 2024, pp. 41582–41610.

[11] Marin Scalbert, Florent Couzinie-Devy, and Riadh Fezzani, "Generic Isolated Cell Image Generator," *Cytometry Part A*, vol. 95, no. 11, pp. 1198–1206, 2019, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cyto.a.23899.

[12] Khaled Al-Thelaya, Marco Agus, Nauman Ullah Gilal, Yin Yang, Giovanni Pintore, Enrico Gobbetti, Corrado Calí, Pierre J. Magistretti, William Mifsud, and Jens Schneider, "InShaDe: Invariant Shape Descriptors for visual 2D and 3D cellular and nuclear shape analysis and classification," *Computers & Graphics*, vol. 98, pp. 105–125, Aug. 2021.

[13] Frank P Kuhl and Charles R Giardina, "Elliptic Fourier features of a closed contour," *Computer Graphics and Image Processing*, vol. 18, no. 3, pp. 236–258, Mar. 1982.

[14] Tereza Nečasová, Daniel Múčka, and David Svoboda, "COMPYDA: An Online Tool for Verifying the Similarity of Image Datasets," in *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, Athens, Greece, May 2024, pp. 1–5, IEEE.

[15] Qing Zhong, Alberto Giovanni Busetto, Juan P Fededa, Joachim M Buhmann, and Daniel W Gerlich, "Unsupervised modeling of cell morphology dynamics for time-lapse microscopy," *Nature Methods*, vol. 9, no. 7, pp. 711–713, July 2012.

[16] John Paparrizos, Haojun Li, Fan Yang, Kaize Wu, Jens E. d'Hondt, and Odysseas Papapetrou, "A Survey on Time-Series Distance Measures," Dec. 2024, arXiv:2412.20574 [cs].

[17] Ignacio Roa, Jorge César Brändle de Motta, and Alexandre Poux, "Spherical harmonics decomposition based on Cartesian level-set field," Technical Report hal-04004128, CNRS, Normandy Univ., UNIROUEN, INSA Rouen, CORIA, 2023.