Spherical Harmonics for Annotation-Free and Shape-Constrained 3D Cell Segmentation

Anonymous BIC@ECCV submission

Paper ID 13

Abstract. Recent microscopy imaging techniques allow to precisely analyze cell morphology in 3D image data. The vast amount of image data renders manual investigations a challenging or even infeasible task and varying image quality limits the applicability of automated methods. The resulting lack of annotated data not only impedes the successful training procedure of machine learning approaches, but also complicates the comparability of publicly available methods. This can be compensated by image synthesis approaches, which require an exact modeling of the underlying cellular structures. The spherical structure of cells in the 3D space can be realistically modeled with spherical harmonics, which offer a way to directly constrain segmentations to represent seamless spherical shapes. This work proposes how this representation of spherical objects can be utilized to model and to predict cellular structures in 3D microscopy image data. We incorporate those descriptors into an image synthesis pipeline to generate synthetic 3D image data that can be utilized to replace manually obtained ground truth annotations for training of automated segmentation approaches. Segmentations obtained with fully-synthetic data are shown to be similar to segmentations obtained with manually annotated data. Moreover, we propose a network architecture, which is designed to predict spherical harmonic coefficients to obtain immanently shape-constrained segmentations and compare the results to the results obtained by established methods used in the field.

Keywords: 3D Cell Analysis, Spherical Harmonics, Annotation-Free, Shape-Constrained

1 Introduction

The continuous development of microscopy imaging techniques, allows to better understand developmental processes at the cellular level. Particularly 3D imaging techniques provide powerful insights, but create vast amounts of data that have to be analyzed. Thus, manual investigations are a tedious or even infeasible task. The amount of annotation effort even increases, when considering densely packed scenes as in the case of 3D confocal microscopy images of fluorescently labeled cell membranes used for detailed morphological analysis. Precise manual segmentation of this complex membrane network is often additionally complicated by low image quality and proximity of cells. To this end, automated cell

segmentation has been addressed with different machine learning-based algorithms.

Recent techniques to perform cell instance segmentation in 3D microscopy image data range from pixel-wise classification methods to direct shape predictions. Methods working in a pixel-wise manner use convolutional neural networks as pre-processing step [9] or post-processing step [20] to improve classical watershed-based segmentation or perform segmentations with region proposal networks [24]. Those techniques are reported to accurately distinguish separate cell instances, but the generated segmentations are still prone to having fragmented or unnatural shapes. To circumvent this problem of noisy segmentations, shapes can be partly constrained by directly predicting global shape representations, as done in [21], rather than performing segmentation on the pixel-level.

However, all machine learning-based approaches need annotated data for training. Although efforts have been made to reduce the amount of training data needed to perform accurate cell segmentation [11, 24], there still is a high demand of data when it comes to generalizing algorithms. But even with small amounts of annotations needed, the manual effort remains expensive. Consequently, data synthesis approaches became more popular, which range from physically inspired simulations of cell dynamics [17] to deep learning-based synthesis [22]. Deep learning-based synthesis of microscopy images showing cell membranes instead of nuclei, has recently been shown to work for 2D images [8].

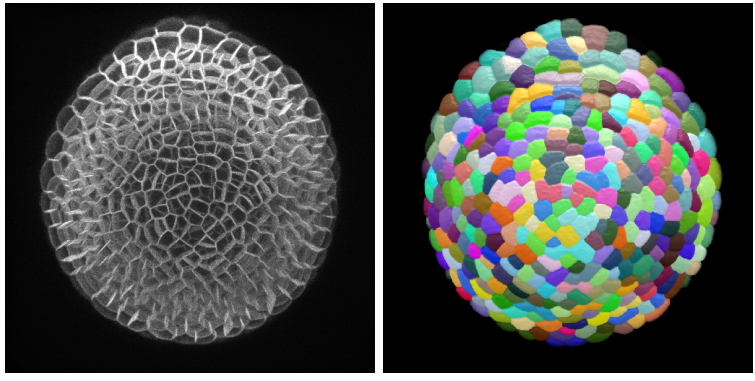


Fig. 1: Maximum intensity projection of a 3D microscopy image (left) and a 3D volume rendering of an instance label mask (right) of *A. thaliana* from the public data set published in [23].

Although previous approaches present techniques to avoid noisy segmentations and generate synthetic microscopy image data, they still need many parameters or do not work for large image data that has to be divided into patches. In this work, we propose how spherical harmonic descriptors [16] can be incorporated into different parts of deep learning-based approaches for synthesis and segmentation of cells in 3D microscopy image data. These descriptors have

been shown to be suitable for a meaningful representation of 3D shapes [5, 6], which suggests an incorporation into current deep-learning approaches. Demonstrations are made on the example of fluorescently labeled cell membranes in *A. thaliana* [23], whose dense cell populations are challenging to segment (Fig. 1). The proposed work demonstrates, (1) how large 3D microscopy image data can be synthesized in a patch-based manner by incorporating positional information and by tweaking patch merging. Furthermore, we introduce (2) a parameter-efficient way to describe and immanently constrain 3D object shapes, which can be used for prediction as well as generation of cell shapes. The resulting pipeline (3) allows performing cell instance segmentation without the need of any manual annotations, while still being on a par with the results obtained with manually annotated ground truth data.

2 Method

The proposed pipeline can be divided into two parts, first the image synthesis and second the cell segmentation. For both parts spherical harmonic (SH) [16] basis functions are employed to describe the spherical shape of a whole specimen as well as the shape of individual cells. Using SH, every spherical shape $\mathbf{S}$ can be decomposed using R different basis functions Y_j weighted by scalars c_j. An individual spherical shape is defined by

$$\mathbf{S} = \sum_{j=1}^{R} c_j \cdot Y_j, \tag{1}$$

where

$$Y_j = Y_l^m(\theta, \phi) = \sqrt{\frac{2l+1}{4\pi} \cdot \frac{(l-m)!}{(l+m)!}} \cdot P_l^m(\cos\theta) e^{i \cdot m \cdot \phi}. \tag{2}$$

Here, $P_l^m(\cos\theta)$ describes the Legendre polynomials of degree m and order l, with $l \geq 0$ and $-l \leq m \leq l$, while higher orders encode the higher frequency components of the sphere. Consequently, the first coefficient, *i.e.*, the first basis function represents a perfect sphere. For each order l we use all available degrees m, which allows to calculate the total number of basis functions up to order l by

$$R = 1 + l + \sum_{i=0}^{l} 2i. \tag{3}$$

Thus, for each of our approaches we fix the order l and describe a spherical shape by determining the weights c_j for each corresponding spherical harmonic Y_j, with $j \in (0, R)$. Parameters θ and ϕ denote the spherical angular coordinates and their quantity determines how detailed each shape can be represented. Instead of sampling those angular orientations from a fixed uniform grid, we select them following the concept of electrostatic repulsion [13]. This ensures an optimal sampling of each shape given a fixed number of orientations.

In the following we explain how this representation of spherical shapes can be used to synthesize and to predict the shape of cellular structures in 3D microscopy images.

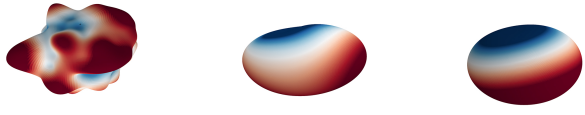
2.1 Patch-based Synthesis of Microscopy Images

For the generation of microscopy images of fluorescently labeled cell membranes, we follow a top-down approach, which extends the work done in [8] to work for 3D data and starts by generating annotations first and proceeds with generating the corresponding simulated microscopy images. This offers the advantage to constrain the morphology of the generated images by the generated masks and results in the creation of a data set with error-free annotations, independent from image quality. As shown in [8], the generation of realistic annotations is essential to leverage the generation of realistic synthetic data sets, that allow annotation-free training of segmentation approaches. In order to adjust the proposed synthesis pipeline to work for 3D images and in order to create more realistic annotations, extensive adaptions have to be made.

Foreground Generation For the determination of foreground regions, which delineate the specimens shape, spherical harmonics are utilized. Since the objects of interest of the considered data set [23] are spherically shaped (Fig. 1), an exponentially decreasing power spectrum of spherical coefficients are a good approximation to create realistic foreground regions. To model various shape alterations, each coefficient is initialized by a random value drawn from a standard normal distribution and weighted by the exponentially decreasing power spectrum for each harmonic degree m, which leads to a description of harmonic coefficients by the following formula:

$$c_l^m = w(l,m) \cdot e^{-\gamma m}, \tag{4}$$

with $w(l,m)$ constituting the random initialization assigned to the coefficient for degree m and order l. γ controls the smoothness of the resulting shape, whose effect is shown in Figure 2.



(a) $\gamma = 0.5$ (b) $\gamma = 2$ (c) $\gamma = 8$

Fig. 2: Overview of possible spherical shapes initialized with harmonic coefficients generated by Equation 4 for different values of γ. Colors indicate the positive (red) or negative (blue) deviation from a perfect sphere.

Cell Generation Within the foreground region, random points are placed following a uniform distribution to indicate cell centroids. To prevent areas of

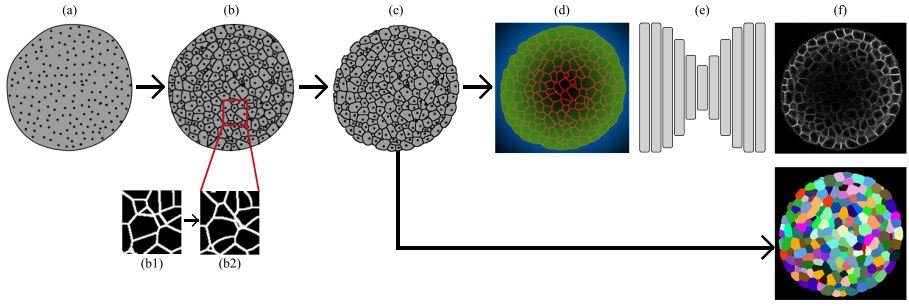


Fig. 3: Qualitative 2D illustration of the 3D data generation pipeline, showing the foreground generation with random cell centroid placement (a), Voronoi-like tessellation (b) and morphological enhancement (c). Initially straight membranes (b1) are bent to imitate more natural shapes (b2). The generated membrane mesh and positional information (d) are utilized for the GAN-based (e) patch-wise translation to the image space (f, top). The final color-coded multi-instance segmentation (f, bottom) is directly obtained from the generated mask (c).

unnaturally low or high density of cells, an k-means clustering using a k obtained from real cell density refines the positions of cells and reduces the overall cell count to n_{cells} (Fig. 3, a). Each voxel within the foreground region is then assigned to the closest cell centroid, which creates separate cell instances as a Voronoi tessellation (Fig. 3, b). Since the resulting segments show unnaturally straight edges (Fig. 3, b1), an additional weighting is applied to the distance calculation, which is adapted from [3] and bends the planes between different segments (Fig. 3, b2). A random weight r_i drawn from a standard uniform distribution is assigned to centroid i and changes the distance calculation of each voxel $\mathbf{x}$ to cell centroid $\mathbf{x}_i$ to

$$d_i(\mathbf{x}) = \frac{||\mathbf{x} - \mathbf{x}_i||_2}{r_i}. \tag{5}$$

To further enhance the outer shape of the foreground region and simulate cell bumps, morphological opening is applied as a final step of the mask generation (Fig. 3, c). Note that the opening operation is constrained to the outer layer of the foreground region to prevent inner holes from being created. In principle, the generation of single cell instances within the foreground region can also be done utilizing SH, similar to the foreground generation. However, due to the seamless membrane mesh in the considered data set, we rely on a tweaked Voronoi tessellation concept.

Image Translation For translation of the generated membrane mask into a realistic microscopy image, the cycleGAN approach from [8] is replaced by a plain GAN approach (Fig. 3, e) [10], which was observed to generate images of identical quality. As generator network a slightly modified U-Net architecture [7] is used

and a PatchGAN [12] acts as a discriminator. Those slight modifications of the U-Net affect the transposed convolution layers within the decoder part, since checkerboard artifacts were observed in the generated images. These artifacts are eliminated by replacing all transposed convolutions by the pixel-shuffle technique [1], which proposes to perform low-resolution convolutions first, followed by a periodic shuffling of the resulting convolution outputs to assemble the higher resolution.

Due to the size of the 3D data, images can no longer be processed as a whole and have to be divided into patches. In the patch-based processing of images, global illumination characteristics, such as depth-dependent fluorescence intensity degradation, are almost completely neglected by the generator network. This leads to illumination artifacts in the resulting full-size image, which highly impede the quality of the generated data. To compensate for this, positional information are supplied as additional channels to the input of the generator network. This positional information is encoded as two distance maps, encoding the distance from the specimen's outer boundary to the specimen's center and to the background region, respectively (Fig. 3, d).

When merging patches back to a full-size image (Fig. 3, f), intensity inconsistencies were observed at patch transitions. Those inconsistencies are prevented by incorporating a 3D variant of the weighted fading technique proposed in [2], which uses a distance weighted overlapping of patches. The weight of a pixel at position $\mathbf{x} = (x, y, z)^T$ within the patch is calculated by

$$w = \frac{1}{n} \cdot \min(|x - x_c|, |y - y_c|, |z - z_c|), \tag{6}$$

with $(x_c, y_c, z_c)^T$ denoting the patch center and n being a normalization factor calculated from overlap and patch size to preserve the intensity range after merging. The complete pipeline is illustrated in Figure 3.

2.2 Shape-Constrained Cell Segmentation

Pixel-wise segmentation approaches like the watershed algorithm are often prone to generating fragmented or noisy segmentations (Fig. 4), which deteriorates the accuracy of morphological quantification for biological assessments. Instead

Fig. 4: 2D crops of possible fragmented cell instances obtained from the method proposed in [9].

of following a pixel-level approach, one could consider to sample the shape of objects, which allows to easily regularize the predicted shapes to be connected.

However, the more complex object shapes become, the more sampling points are needed for an accurate representation. The number of necessary sample points easily exceeds multiple hundred, which is even more significant in the 3D space. To reduce the number of required points, the location of sampling points could be optimized or a compressed encoded representation could be used. For the 3D space, spherical harmonics serve as such a compressed and parameter-efficient encoding for star-convex volumes. To this end, we demonstrate how deep neural networks can be adapted to predict the position and the shape encoding of objects.

To transform between volumetric pixel-wise segmentations and spherical harmonic representations with a minimum loss of information, the segmentation is initially oversampled by considering 5000 angular orientations (θ, ϕ). From the obtained sampling points, the harmonic coefficients can be accurately calculated. To reverse the encoding to a pixel-wise segmentation, we again use the same sampling pattern and subsequently apply a Delaunay triangulation.

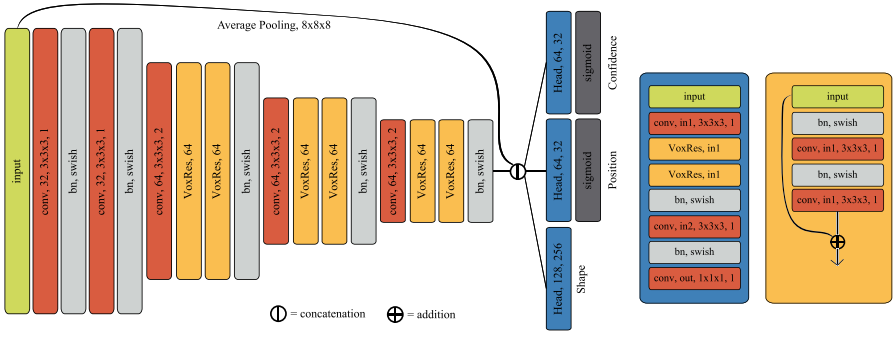


Fig. 5: Illustration of the shape prediction network, which comprises a shared encoder part [4] splitting into three different head modules (blue) for prediction of detection confidence, position regression and shape encoding, respectively. The main building block is presented by the VoxRes module (orange) [4]. Each convolution block performs 3D convolutions, *bn* expresses batch-normalization and Swish [18] is implemented as activation function.

The proposed network is inspired by the work done in [19] and [14] and follows the principle of a one-shot detector. The architecture comprises a residual-based encoder part, which splits into separate heads for prediction of detection confidence, position regression and shape encoding (Fig. 5). The main building block is given by the VoxRes module, proposed in [4]. Considering, that the encoder part of the network down-samples the image into a smaller representation, each voxel of the encoder output represents an area in the original image, which can thus be thought of as a division into a grid. In each grid segment the prediction of n_{det} objects is allowed, each of them being constituted by the

parameter tuple $(p, \delta x, \delta y, \delta z, c_1, c_2, ..., c_R)$, with p indicating the detection confidence, $(\delta x, \delta y, \delta z)$ denoting the positional offset within the grid segment and $c_{1...R}$ representing the spherical coefficients. Each of those three compartments is predicted by an individual head module and concatenated to a tuple afterwards. The positional offsets $(\delta x, \delta y, \delta z)$ range between 0 and 1 to indicate to which region of the segment the detection should be shifted. Thus, each detected center position is bounded to a specific segment with the center of a segment given by $(0.5, 0.5, 0.5)$, whereas the detected shape can exceed the segment boundaries. The number of the overall required output channels is determined by the number of allowed detections per grid segment n_{det} and the number of coefficients R, which results in a total of $(4 + R) \cdot n_{det}$ outputs for each segment. For prediction of confidence and position regression, a sigmoid activation function is used to generate the final output. Values for spherical coefficients, however, are not bounded, to which end no final activation function was applied and the raw output of the last convolution generates the final output.

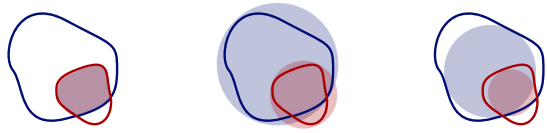
During training, each head is assigned a different loss function, which are ultimately combined to form the whole training criterion. The confidence prediction is evaluated using a mean squared logarithmic error (MSLE) and a weighting function to compensate for class imbalances, as done in [9]. Predictions of positions and harmonic coefficients are assessed using a masked L1 loss. Segments not containing any cells in the ground truth mask do not contribute to the calculation of the positional loss and the shape loss. Those segments only contribute to the confidence loss, since positions and shape descriptors are undefined.

Although we follow a grid-based detection, which limits the number of predictions per region, there still are redundant detections that have to be merged by a sequence of steps. Initially, the confidence of each prediction is weighted by the positional distance to the patch borders to account for possible truncated shape predictions at patch borders. The weight is degraded towards the patch border by a tangens hyperbolicus as stated in the following equation:

$$w_{conf} = \tanh\left(\frac{\min(x_{pred}, s - x_{pred})}{s} \cdot \alpha\right), \tag{7}$$

with x_{pred} denoting the predicted position within the patch, s denoting the patch size and constant α controlling the steepness of the weight decay towards the patch border. To reduce computational effort, redundancy of predictions is reduced within each patch by the application of hierarchical agglomerative clustering, using Ward's linkage method. A distance threshold $d_{cluster}$ identifies, which detections should be ultimately merged to form clusters. Since this step is only applied to remove unambiguous redundancies, the threshold $d_{cluster}$ is chosen to be small compared to the cell sizes. The resulting position and shape of each cluster are calculated from a weighted average among each prediction within the clusters. The weight is determined by the corresponding predicted confidences, *i.e.*, the higher the confidence, the more influence on the resulting position and shape. All detections within a cluster are discarded afterwards and only the cluster average is kept. After all patches are concatenated to form the

full-size image, a non-maximum-suppression (NMS) is applied to remove objects whose overlap exceeds a threshold $t_{overlap}$. For calculation of the overlap the intersection over union (IoU) is typically used, which, however, does not take the different sizes of objects into account. This potentially leads to small IoU values even if a small object lies almost completely within a larger object, as depicted in Fig. 6 (a). To this end, the overlap is set in relation to the volume of the smallest object, which is referred to as intersection over smallest volume (IoSV) in the following. Furthermore, the exact pixel-wise calculation of the overlap is computationally heavy and infeasible if many overlaps in a 3D image have to be calculated. Therefore, the overlap is approximated by averaging the overlaps of the minimum enclosing (Fig. 6, b) and maximum enclosed spheres (Fig. 6, c) of both segmentations, respectively [21]. The intersection I of partially



(a) Intersection (b) Enclosing (c) Enclosed

Fig. 6: 2D representation of the intersection of two spherical objects (a). The overlap is approximated by averaging the intersections of the corresponding minimum enclosing (b) and maximum enclosed (c) spheres [21].

overlapping spheres can be efficiently calculated from the distance d between the centroids of both spheres and their radii r_1 and r_2:

$$I = \frac{\pi}{12d}(r_1 + r_2 - d)^2(d^2 + 2d(r_1 + r_2) - 3(r_1 - r_2)^2). \tag{8}$$

3 Experiments and Results

For the evaluation of the proposed pipeline, three assessments were made. First, the quality and plausibility of generated images were evaluated by comparison to the original data. With different experiments, the influences of the label generation and the domain translation were assessed individually. Second, the segmentation accuracy of the proposed network was determined and compared to the results obtained by other methods used in the field. As a third assessment, the applicability of the generated data as a training set was demonstrated.

All experiments were carried out using the publicly available data set from [23], comprising 125 3D images showing fluorescently labeled cell membranes of *Arabidopsis thaliana* and corresponding manually corrected watershed-based instance segmentations. Segmentation of fluorescently labeled membranes is a problem present in many biological screens, with *Arabidopsis* being one of the

most commonly used plant model organism. Moreover, animal models like *mouse*, *zebrafish* or *fruit fly* exhibit similar signal patterns and shapes that can be modeled with the proposed methods alike, *i.e.*, the proposed method is extendable to these models as well.

3.1 Quality of the Generated Images

The data generation pipeline has two sources of potential impairments, namely the label generation and the translation from the label space into the image space. To determine the individual influences of both parts on the quality of the generated data, two additional data sets were generated:

- A manually corrected data set [23].
- A semi-synthetic data set by translating manually annotated masks to the image domain using the GAN approach described in Section 2.1.
- A fully-synthetic data set by generating masks as proposed in Section 2.1 and creating corresponding synthetic images utilizing the GAN.

Note that, for both synthetic cases the training of the generative networks was carried out with unpaired data, which allowed to use the semi-synthetic images to solely measure the impairments contributed by the domain translation.

For the initialization of spherical harmonic coefficients with Equation 4, γ has been empirically set to 6 to best fit the shape of the real foreground regions. The number of randomly placed cells n_{cells} was adapted to roughly match the real density of cells within the foreground region, which was determined to be one cell per $20 \times 20 \times 20$ pixel region. Translation of the generated masks to the image domain was done in a patch-based manner, utilizing a patch-size of $128 \times 128 \times 128$ pixel.

To assess the quality of the generated semi-synthetic images (16-bit), the structural similarity (SSIM), the zero-normalized correlation coefficient (ZNCC), the peak signal-to-noise ratio (PSNR) and the normalized root-mean-squared-error (NRMSE) are measured. Quality of generated data is still hard to measure by statictical metrics and not feasible in case of an unpaired setup, like for the fully-synthetic data set. Instead, since the synthesis is performed to create data sets that can be used as a training set, we show how well the fully-synthetic data can be used to train the proposed pipeline and carry out different experiments to show to which extend they can be used as a replacement for real data (see Section 3.3). Obtained metrics for the semi-synthetic data set are shown in Table 1 and qualitative results for all data sets are illustrated in Fig. 7.

SSIM	ZNCC	PSNR	NRMSE
0.77 ± 0.05	0.70 ± 0.07	23.97 ± 1.93	1.11 ± 4.64

Table 1: Evaluation metrics calculated for the semi-synthetic images generated from manual annotations.

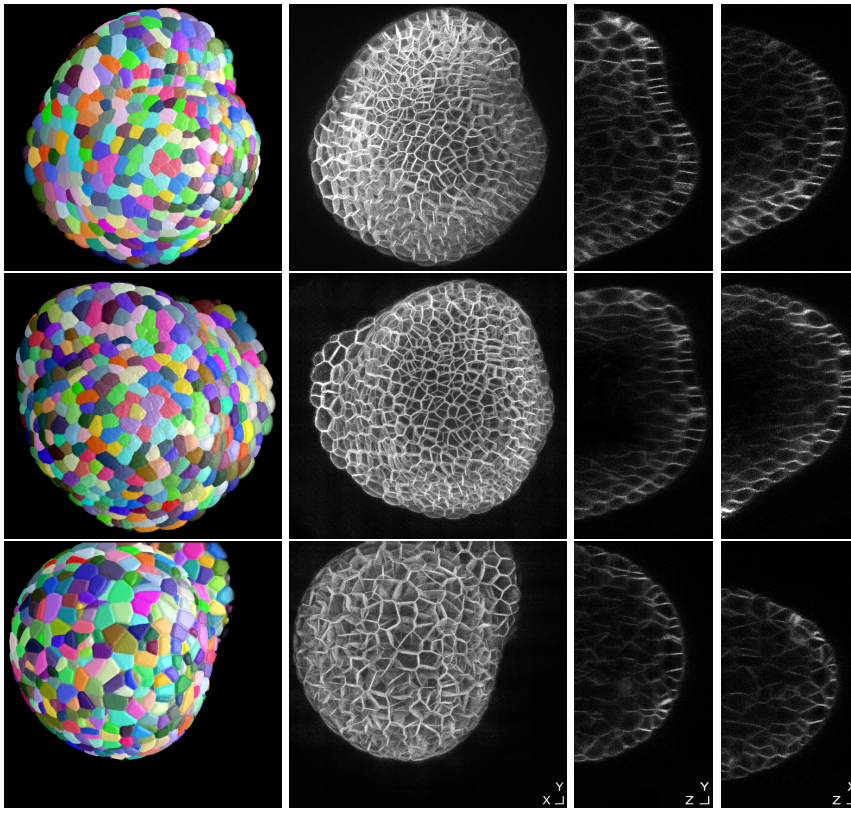


Fig. 7: Exemplary data from the original data set (top), the semi-synthetic data set (middle) and the fully-synthetic data set (bottom). For those data sets, 3D renderings of the color-coded instance masks (far left), maximum intensity projections of the 3D images (middle left), slices from the yz-plane (middle right) and slices from the xz-plane (far right) are displayed.

3.2 Segmentation Accuracy

For cell shape prediction, the harmonic order l and, thereby, the number of harmonic coefficients R has to be chosen, which also defines how precise high-frequency information can be represented. Therefore, the shape encoding is lossy and naturally lowers the maximum segmentation scores that can be obtained. To measure to which extend the number of descriptors impedes the accuracy, the IoU between manual instance segmentations and corresponding reconstructed segmentations was computed using different quantities of harmonic coefficients. Outcomes have been aggregated over a total of 5000 individual cell instances and resulted in the scores shown in Figure 8 (left).

Segmentation results obtained by the proposed network were compared to the results obtained with two other methods established in the field, namely

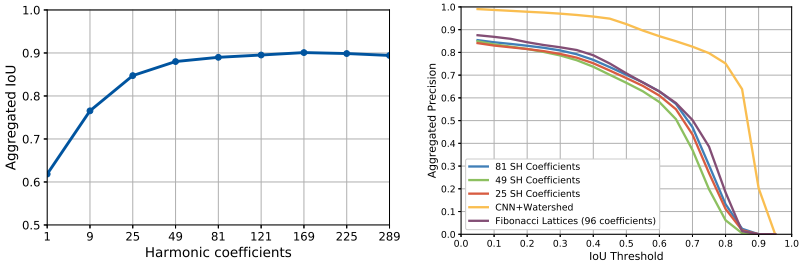


Fig. 8: Aggregated IoU over a total of 5000 individual cells, utilizing different quantities of harmonic coefficients (left) and aggregated accuracy of each considered instance segmentation algorithm (right). IoU thresholds range from 0.05 to 0.95 and determine, which cells were correctly segmented to formulate the overall accuracy.

a convolutional neural network (CNN) enhanced watershed segmentation [9] and a U-Net-based detection network, which encodes cell shapes by Fibonacci lattices [21]. For the proposed pipeline, we empirically decided to allow $n_{det} = 2$ detections per segment and use $R = [25, 49, 81]$ harmonic coefficients for different experiments, which resulted in a total of $[58, 106, 170]$ output parameters per segment. Furthermore, the distance threshold $d_{cluster}$ to form initial clusters was set to a Euclidean distance of 10 pixel and the NMS threshold was set to $t_{overlap} = 0.3$. Utilizing manually annotated data, the IoU was calculated as a metric to measure which cells were correctly segmented. For different IoU thresholds, the number of true (TP) and false (FP) predictions and the number of missed cells (FN) could be determined, which led to a formulation of an overall accuracy by

$$acc = \frac{TP}{TP + FP + FN}. \tag{9}$$

The accuracy curves for each algorithm are plotted in Figure 8 (right) and qualitative results are illustrated in Figure 9.

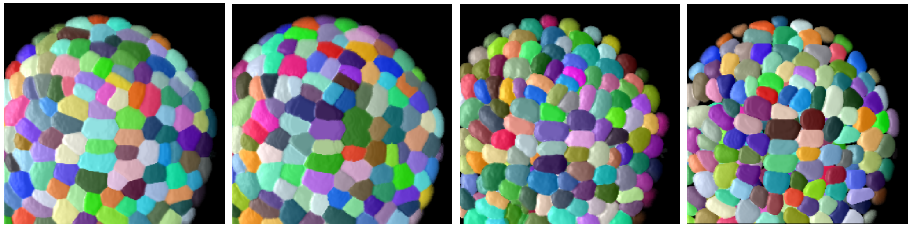


Fig. 9: Instance segmentation masks showing the manually annotated ground truth [23] (left), CNN+watershed results [9] (middle left), Fibonacci lattice predictions [21] (middle right) and the SH predictions using 81 coefficients (right).

3.3 Annotation-Free Training

To evaluate how well the simulated data can be utilized as a training set and how well it can substitute manual annotated data, different experiments were performed. In addition to training and testing on manually annotated data, the proposed pipeline and the approach from [9] were trained on synthetic data and evaluated on manually annotated data. Furthermore the pipeline was trained on manual ground truth data and tested on simulated data. As a metric the same formulation as described in Section 3.2 was applied, which led to the accuracy curves plotted in Figure 10.

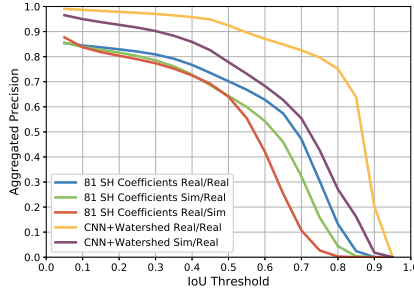


Fig. 10: Segmentation accuracy of the proposed pipeline, using different train/test constellations of simulated and manually annotated data.

4 Discussion

Obtained results for the structural quality of the semi-synthetic data (Table 1 and Fig. 7, middle row) indicated, that after translation into the image domain, positions of membranes remained colocalized, as also shown by the relatively high SSIM and ZNCC values. By introducing a positional encoding, the generator network was able to consider global illumination characteristics, which are accountable for the largest intensity deviations between real data and simulated data, leading to poor PSNR and NRMSE scores. These resulting intensity differences impaired the obtained similarity measurements, but did not necessarily lead to poor synthesized images and rather imitated the strong position-dependent variations in the data. Since the global intensity deviations and the areas exhibiting low intensity and low contrast are most challenging for segmentation approaches, the availability of these characteristics in a fully-annotated data set is of high value for other organisms with sparse or no annotations available. This claim is supported by the experimental outcomes, when replacing manually annotated data sets with fully synthesized data sets (Fig. 10). The outcomes of these different experiments show that the generation pipeline can

be used to diminish the need of manual annotations for acquisition of training data, with a trade-off of a minor loss in precision, considering that no manual interactions are required.

Modelling shapes with SH comes with a trade-off between a shape-constraint, parameter-efficient description of 3D shapes and a loss of high-frequency components of the shape, as shown in Figure 8. This loss of high-frequency segmentation components can be directly observed, when comparing to the results obtained with a watershed-based approach. However, for morphological assessments of cell shapes, consistent and realistic segmentations are important, which makes the shape-constrained encodings a good alternative. Also note, that the ground truth was constructed by manually correcting segmentations obtained from a watershed-based algorithm [23], leading to a bias of the reported segmentation results obtained with the CNN+watershed method. The same trade-off could be shown in [21] for Fibonacci latices and is visible in Figure 8 (right). Within the examined quantity range of harmonic coefficients, only modest changes in segmentation accuracy could be observed, which shows the advantage of SH over Fibonacci lattices. Since cells are naturally represented by spherical shapes and one single SH descriptor can already represent a perfectly spherical object, 25 descriptors are sufficient to represent 3D cell shapes. Furthermore, the encoded representation directly provides information about the rough size of objects (first coefficient), the roughness of the objects surface (distribution of coefficient values) and it can be used as a feature vector describing the cell.

5 Conclusion

This work proposed how spherical harmonic coefficients can be incorporated into different parts of 3D microscopy image data processing approaches. Modelling entire specimen with SH coefficients provided an efficient way to generate various parametrized 3D shapes for data generation. By modelling membrane structures and enhancing morphological details, a tweaked image synthesis pipeline generated data, which in turn can be used to dispense manual annotation efforts for training data acquisition. Using a CNN to predict the harmonic shape representation of single cell instances has been shown to produce segmentation results that can be represented by only a small amount of descriptors. This representation prevents the prediction of unnaturally degenerated segmentations, benefiting biological assessments by only a small accuracy trade-off. In general, SH serve as an efficient and promising alternative for incorporating natural shape constraints into deep learning-based segmentation and synthesis approaches. Future work includes the extension of the parametrized shape generation, to enable the synthesis of arbitrary kinds of specimens and to further enhance the membrane morphology and cell size distribution to better match biological constraints. Furthermore, segmentation results could be improved by incorporating spherical convolutions [15] into the shape prediction head of the network or by using the SH predictions as enhanced seeds for watershed-based segmentation to retain both, the shape-contraints and the high-frequency shape information.

References

1. Aitken, A., Ledig, C., Theis, L., Caballero, J., Wang, Z., Shi, W.: Checkerboard Artifact Free Sub-Pixel Convolution: A Note on Sub-Pixel Convolution, Resize Convolution and Convolution Resize. arXiv:1707.02937 (2017)
2. de Bel, T., Hermsen, M., Kers, J., van der Laak, J., Litjens, G.: Stain-Transforming Cycle-Consistent Generative Adversarial Networks for Improved Segmentation of Renal Histopathology. In: International Conference on Medical Imaging with Deep Learning. vol. 102, pp. 151–163 (2018)
3. Bock, M., Tyagi, A.K., Kreft, J.U., Alt, W.: Generalized Voronoi Tessellation as a Model of Two-Dimensional Cell Tissue Dynamics. Bulletin of Mathematical Biology **72**(7), 1696–1731 (2010)
4. Chen, H., Dou, Q., Yu, L., Heng, P.A.: VoxResNet: Deep Voxelwise Residual Networks for Volumetric Brain Segmentation. arXiv:1608.05895 (2016)
5. Christel Ducroz, J.C.O.M., Dufour, A.: Spherical Harmonics based Extraction and Annotation of Cell Shape in 3D Time-Lapse Microscopy Sequences. In: 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society. pp. 6619–6622 (2011)
6. Christel Ducroz, J.C.O.M., Dufour, A.: Characterization of Cell Shape and Deformation in 3D using Spherical Harmonics. In: 2012 9th IEEE International Symposium on Biomedical Imaging (ISBI). pp. 848–851 (2012)
7. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 424–432 (2016)
8. Eschweiler, D., Klose, T., Müller-Fouarge, F.N., Kopaczka, M., Stegmaier, J.: Towards Annotation-Free Segmentation of Fluorescently Labeled Cell Membranes in Confocal Microscopy Images. In: International Workshop on Simulation and Synthesis in Medical Imaging. pp. 81–89 (2019)
9. Eschweiler, D., Spina, T.V., Choudhury, R.C., Meyerowitz, E., Cunha, A., Stegmaier, J.: CNN-based Preprocessing to Optimize Watershed-based Cell Segmentation in 3D Confocal Microscopy Images. In: International Symposium on Biomedical Imaging. pp. 223–227 (2019)
10. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative Adversarial Nets. In: Advances in Neural Information Processing Systems. pp. 2672–2680 (2014)
11. Guerrero-Peña, F.A., Fernandez, P.D.M., Ren, T.I., Cunha, A.: A Weakly Supervised Method for Instance Segmentation of Biological Cells. In: Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data, pp. 216–224 (2019)
12. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-Image Translation with Conditional Adversarial Networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1125–1134 (2017)
13. Jones, D.K., Horsfield, M.A., Simmons, A.: Optimal Strategies for Measuring Diffusion in Anisotropic Systems by Magnetic Resonance Imaging. Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine **42**, 515–525 (1999)
14. Khosravan, N., Bagci, U.: S4ND: Single-Shot Single-Scale Lung Nodule Detection. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 794–802 (2018)

15. Koppers, S., Merhof, D.: DELIMIT PyTorch - An Extension for Deep Learning in Diffusion Imaging. arXiv:1808.01517 (2018)
16. Müller, C.: Spherical Harmonics, vol. 17. Springer (2006)
17. Peterlík, I., Svoboda, D., Ulman, V., Sorokin, D.V., Maška, M.: Model-Based Generation of Synthetic 3D Time-Lapse Sequences of Multiple Mutually Interacting Motile Cells with Filopodia. In: International Workshop on Simulation and Synthesis in Medical Imaging. pp. 71–79 (2018)
18. Ramachandran, P., Zoph, B., Le, Q.V.: Swish: a self-gated activation function. arXiv:1710.05941 (2017)
19. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You Only Look Once: Unified, Real-Time Object Detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 779–788 (2016)
20. Stegmaier, J., Spina, T.V., Falcão, A.X., Bartschat, A., Mikut, R., Meyerowitz, E., Cunha, A.: Cell Segmentation in 3D Confocal Images Using Supervoxel Merge-Forests with CNN-based Hypothesis Selection. In: International Symposium on Biomedical Imaging. pp. 382–386 (2018)
21. Weigert, M., Schmidt, U., Haase, R., Sugawara, K., Myers, G.: Star-Convex Polyhedra for 3D Object Detection and Segmentation in Microscopy. arXiv:1908.03636 (2019)
22. Wiesner, D., Nečasová, T., Svoboda, D.: On Generative Modeling of Cell Shape Using 3D GANs. In: International Conference on Image Analysis and Processing. pp. 672–682 (2019)
23. Willis, L., Refahi, Y., Wightman, R., Landrein, B., Teles, J., Huang, K.C., Meyerowitz, E.M., Jönsson, H.: Cell Size and Growth Regulation in the Arabidopsis Thaliana Apical Stem Cell Niche. Proceedings of the National Academy of Sciences **113**(51), E8238–E8246 (2016)
24. Zhao, Z., Yang, L., Zheng, H., Guldner, I.H., Zhang, S., Chen, D.Z.: Deep Learning Based Instance Segmentation in 3D Biomedical Images Using Weak Annotation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 352–360 (2018)