Evaluation of Deep Learning Architectures for Complex Immunofluorescence Nuclear Image Segmentation

Florian Kromp, Lukas Fischer, Eva Bozsaky, Inge M. Ambros, Wolfgang Dörr, Klaus Beiske, Peter F. Ambros, Allan Hanbury, and Sabine Taschner-Mandl

***Abstract*— Separating and labeling each nuclear instance (instance-aware segmentation) is the key challenge in nuclear image segmentation. Deep Convolutional Neural Networks have been demonstrated to solve nuclear image segmentation tasks across different imaging modalities, but a systematic comparison on complex immunofluorescence images has not been performed. Deep learning based segmentation requires annotated datasets for training, but annotated fluorescence nuclear image datasets are rare and of limited size and complexity. In this work, we evaluate and compare the segmentation effectiveness of multiple deep learning architectures (U-Net, U-Net ResNet, Cellpose, Mask R-CNN, KG instance segmentation) and two conventional algorithms (Iterative h-min based watershed, Attributed relational graphs) on complex fluorescence nuclear images of various types. We propose and evaluate a novel strategy to create artificial images to extend the training set. Results show that instance-aware segmentation architectures and Cellpose outperform the U-Net architectures and conventional methods on complex images in terms of F1 scores, while the U-Net architectures achieve overall higher mean Dice scores. Training with additional artificially generated images improves recall and F1 scores for complex images, thereby leading to top F1 scores for three out of five sample preparation types. Mask R-CNN trained on artificial images achieves the overall highest F1 score on complex images of similar conditions to the training set images while Cellpose achieves the overall highest F1 score on complex images of new imaging conditions. We provide quantitative results demonstrating that images annotated by under-graduates are sufficient for training instance-aware segmentation architectures to efficiently segment complex fluorescence nuclear images.**

***Index Terms*— Architecture evaluation, artificial images, deep learning, expert-annotated data, nuclear image segmentation.**

Manuscript received August 18, 2020; revised December 5, 2020 and February 13, 2021; accepted March 21, 2021. Date of publication March 30, 2021; date of current version June 30, 2021. This work was supported in part by the Austrian Research Promotion Agency (FFG) COIN "Networks" projects TISQUANT and VISIOMICS, LIQUIDHOPE funded by the Austrian Science Fund (FWF) within the framework of ERA-NET/Transcan-2 Program under Grant I4162 and in part by the Federal Ministry Republic of Austria for Climate Action, Environment, Energy, Mobility, Innovation and Technology (BMK), Federal Ministry Republic of Austria for Digital and Economic Affairs (BMDW), and the Province of Upper Austria in the frame of the COMET Programme managed by FFG and the St. Anna Kinderkrebsforschung. *(Allan Hanbury and Sabine Taschner-Mandl contributed equally to this work.) (Corresponding authors: Allan Hanbury; Sabine Taschner-Mandl.)*

Florian Kromp, Eva Bozsaky, Inge M. Ambros, Peter F. Ambros, and Sabine Taschner-Mandl are with the Tumor Biology Group, Children's Cancer Research Institute, 1090 Vienna, Austria (e-mail: florian.kromp@ccri.at; eva.bozsaky@ccri.at; inge.ambros@ccri.at; sabine.taschner@ccri.at; peter.ambros@ccri.at).

Lukas Fischer is with the Software Competence Center Hagenberg GmbH (SCCH), 4232 Hagenberg, Austria (e-mail: lukas.fischer@scch.at).

Wolfgang Dörr is with the ATRAB-Applied and Translational Radiobiology, Department of Radiation Oncology, Medical University of Vienna, 1090 Vienna, Austria (e-mail: wolfgang.doerr@meduniwien.ac.at).

Klaus Beiske is with the Department of Pathology, Oslo University Hospital, N-0379 Oslo, Norway (e-mail: klaus.beiske@medisin.uio.no).

Allan Hanbury is with the Institute of Information Systems Engineering, TU Wien Informatics, 1040 Vienna, Austria, and also with the Complexity Science Hub, 1080 Vienna, Austria (e-mail: allan.hanbury@tuwien.ac.at).

This article has supplementary downloadable material available at https://doi.org/1010.1109/TMI.2021.3069558, provided by the authors.

Digital Object Identifier 10.1109/TMI.2021.3069558

I. Introduction

MICROSCOPY has become a powerful tool to gain insights into cellular or sub-cellular structures by visualizing cellular compartments such as the nucleus, the cytoplasm, sub-cellular appearance of proteins or DNA elements [1]. By applying automated microscopes and image analysis workflows, quantitative results can be generated at the single cell level. These allow the detection of even subtle biological changes while taking advantage of the statistical power of analyzing thousands of cells. The main sites of operation for quantitative microscopy analysis are pathology departments and diagnostic laboratories. In addition, quantitative microscopy techniques are applied and refined in research laboratories. While pathology departments routinely use Hematoxylin and Eosin (H&E) histological or immunohistochemical (IHC) stainings, research laboratories mainly rely on immunofluorescence (IF) stainings. This is because up to 90 or more (sub-)cellular compartments can be visualized simultaneously using multiplex-IF staining techniques and epifluorescence microscopy. This provides a substantial gain of information compared to the visualization of two to three cellular characteristics when using H&E or IHC stainings and brightfield microscopy. While pathology departments mainly rely on tissue sections to diagnose disease type and grade or stage of cancer [2], [3], research laboratories frequently use additional tissue preparations such as cell lines grown on or cytospinned to microscopy glass slides, cytospin

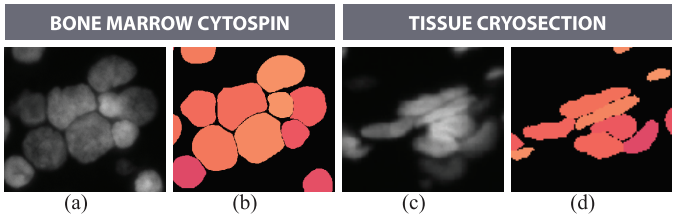

Fig. 1. Examples of nuclear morphologies in various tissue preparations. (a) Neuroblastoma bone marrow cytospin presenting varying nuclear intensity and size. (b) Annotated mask of (a). (c) Ganglioneuroma tissue cryosection presenting overlapping/aggregated nuclei with varying morphology and intensity. (d) Annotated mask of (c).

preparations of bone marrow or tumor touch imprints. Quantitative, microscopy based image analysis workflows generally consist of the following steps: sample preparation, microscopy image acquisition, nuclear and/or cytoplasmic image segmentation, feature extraction and cell population analysis. Each step within such a workflow can impact quantification and thus, interpretation of experiments [4].

A prerequisite that also represents a bottleneck in automated quantitative microscopy is accurate nuclear image segmentation. To generate quantitative results at the single cell level, segmentation algorithms must segment each nucleus instance. Such algorithms are called *instance segmentation* or *instance-aware segmentation* algorithms. Inaccurate image segmentation is frequently caused by tightly aggregated nuclei that cannot be separated by the segmentation algorithm, compromising biological conclusions [5]. Figure 1 shows examples of images highly challenging for human experts as well as for automated nuclear image segmentation methods tasked to separate each nucleus instance. These images are further called *complex images*.

To tackle nuclear image segmentation, deep learning architectures have shown to be capable of detecting nuclear instances while outperforming classical segmentation algorithms [6]. Most prominently, the U-Net architecture and variants thereof are applied, transforming a nuclear image into a probability map indicating the class membership (nucleus or background) for each pixel. Similar to autoencoder networks, the U-Net architecture consists of an encoder CNN (contracting path) coupled with a decoder CNN (expansive path) forming a U-like shape. Additional skip-connections between different layers in the down- and up- sampling part of the network are introduced. This strategy allows to preserve spatial information and image details that would otherwise get lost during the down sampling process. Despite the success of applying the U-Net architectures to nuclear image segmentation, nuclei in tight aggregations sometimes lead to under-segmentation. To overcome this, post-processing steps are commonly applied to a predicted initial nucleus segmentation [7]–[9], relying on morphological or intensity based features derived from segmented objects. These features are further used to identify under-segmented objects and to guide the separation or merging process.

U-Net based nuclear segmentation algorithms frequently fail to segment nuclear instances in IF images because:

- Epifluorescent microscopic images are blurry caused by out-of-focus light, leading to aggregations if nuclei are located in spatially close neighborhoods.
- Nuclei in dense tissue sections, bone marrow cytospins or tumor touch imprints are frequently aggregated and/or overlapped, stem from various cell types showing heterogeneous intensity levels and present arbitrary convex or even concave shapes, complicating the use of morphology-based features in post-processing operations.

Another type of architecture recently applied to tackle nuclear image segmentation are instance-aware segmentation networks. Built upon networks designed to detect object instances, they apply local segmentation after object detection in the coarse region of each located object. High-level feature layers from the object detection part of such networks are reused for segmentation. A prominent example is Mask R-CNN, being among the three best solutions proposed to the Data Science Bowl 2018 contest [10] on nucleus segmentation. By separating object detection from object segmentation, instance-aware segmentation architectures are a good choice to solve nuclear image segmentation in complex nuclear images. However, in contrast to the U-Net architectures, no mechanisms to keep spatial information are applied. Thus masks (nuclear outlines) of predicted instances might be less accurate.

Due to the rapid development of the aforementioned and other deep learning-based methods aimed at nuclear image segmentation, scientists and application specialists working with automated image analysis workflows are faced with an ever-increasing number of publications, challenging the process of selecting the best performing method to process their own dataset. A comparison of deep learning architectures and conventional algorithms on confocal fluorescence nuclear images has been proposed [6] as well as an evaluation of the top-performing models submitted to the Data Science Bowl 2018 contest on segmenting nuclear images of a heterogeneous dataset including fluorescence and brightfield images [10]. However, to the best of our knowledge, there is no work published that systematically evaluates deep learning architectures on instance-aware segmentation of nuclei on complex fluorescent images of various tissue origins, sample preparation types and magnifications, further called *sample conditions*. This might be based on the reason that fluorescent nuclear image datasets that cover a broad range of sample conditions and contain complex images are rare and of limited size. In particular, the annotation of highly complex images such as tissue sections is time consuming and expensive due to the human resources needed. To enable an evaluation of state-of-the-art segmentation methods on such images, we recently published an expert-annotated dataset consisting of fluorescent nuclear images and annotations of various tissue origins and sample preparation types, acquired using multiple modalities, different levels of magnification and signal-to-noise ratio [11], [12].

In this paper, we systematically compare the segmentation effectiveness of two U-Net architectures, a U-Net architecture with modified mask representation (Cellpose) and two instance-aware segmentation architectures utilized to segment cellular nuclei in complex images using the aforementioned

expert-annotated dataset. In addition, we evaluate and compare the deep learning architectures to two conventional methods, namely Iterative h-min based watershed (Iterative h-min) and Attributed Relational Graphs (ARG), designed to segment fluorescence nuclear images. To increase the size of the annotated dataset, we propose a strategy to synthesize artificial images focusing on overlapping nuclei, thereby simulating complex nuclear images. Hou *et al.* generate artificial nuclei on nuclei-free background patches of Hematoxylin&Eosin stained samples and transform them to realistic images using a Generative Adversarial Network (GAN) [13], while Dunn *et al.* [14] place ellipsoids representing nuclei in empty 3D volumes and transform them to realistic volumes using spatially constrained CycleGANs. In contrast to these approaches, we rather create artificial images by modelling a realistic background and by placing cropped and augmented nuclei from real images on this artificially created background, subsequently transforming these patches into natural-looking images using a paired GAN. Thereby, we focus the generation process on overlapping and overlaying nuclei, forcing the deep learning architectures trained on these images to learn how to split aggregated nuclei as occuring in complex images. Moreover, our method allows to generate artificial images using individually augmented nuclei, representing the varying nuclei intensities within fluorescence nuclear images.

The contributions of the paper are as follows:

- We propose a formal description of nuclear image complexity with respect to the challenge to annotate or segment single-nuclei instances.
- We evaluate and compare deep learning architectures and conventional methods prominently used for nuclear image segmentation on complex and heterogeneous fluorescence images with respect to different image complexity levels.
- We demonstrate to which extend investigated deep learning architectures can generalize to new imaging conditions.
- We propose a novel method to extend fluorescence nuclear image training sets by simulating complex fluorescence nuclear images and demonstrate its effectiveness with respect to high complex images and five sample preparation types.
- We provide quantitative results demonstrating to which extent the quality of dataset annotations influence the segmentation effectiveness of each architecture investigated (U-Net, U-Net ResNet, Cellpose, Mask R-CNN, KG instance segmentation, Iterative h-min, ARG). Based on our results, we recommend the combination of *silver-standard* data annotation, artificially synthesized images and an instance-aware segmentation architecture or Cellpose for obtaining the most effective segmentation of complex fluorescence nuclear images.

II. Related Work

The most popular nuclear segmentation algorithms used until deep learning architectures gained importance, further called conventional methods, are based on the watershed algorithm, region growing, level-set or active contour methods (comprehensive overview in [15]). Deep learning has outperformed traditional methods on many tasks including nuclear image segmentation [6]. Thus, we briefly describe deep learning based segmentation approaches targeting bright-field nuclear images, fluorescence nuclear images or both. Moreover, we describe data augmentation strategies and methods generating synthetic images to extend the training sets.

A. Deep Learning Based Methods to Segment Nuclear Images of H&E and IHC Stained Samples

Recent work showed that Deep Convolutional Neural Networks (DCNN) outperformed most standard methods applied in computer vision tasks such as image classification or segmentation by a large margin [16], [17]. The advantage of DCNNs over traditional methods was also demonstrated for nuclear image segmentation [6].

Recently, annotated datasets of H&E and IHC stained samples became publicly available. This led to the development of new deep learning based architectures to segment these challenging nuclear images [18]–[20]. Naylor *et al.* [7] use and compare CNN architectures (FCN, U-Net, Mask R-CNN) for segmenting H&E stained histological slides. In this study the segmentation is presented as a regression task, resulting in the prediction of the distance-transform of the binarized image. Graham *et al.* [21] use a property of histopathology images (rotational symmetry of nuclear objects across images) and employ steerable, rotational filters to reduce the number of network parameters, while maintaining similar segmentation effectiveness.

B. Deep Learning Based Methods to Segment Immunofluorescence Based Nuclear Images

In contrast to annotated H&E or IHC stained nuclear image datasets, only a limited number of annotated fluorescence nuclear images covering a diverse range of preparation types and tissues are publicly available. Most datasets published consist of confocal images of cell line cytospins or cell lines grown on microscopy slide and are thus of low complexity. Alom *et al.* [20] use a Recurrent Residual CNN based U-Net to segment images of the 2018 Data Science Bowl [10] dataset, including, among others, nuclear images of H&E stained tissue sections and IF stained cells grown on microscopy glass slides. The proposed architecture, however, failed to achieve instance-aware segmentation.

Other deep learning based segmentation methods have been tested on datasets with lower segmentation complexity than tissue sections, characterized by the absence of larger nuclear aggregations or overlaps, or operate on 3D image stacks. Caicedo *et al.* [6] compare a U-Net and the DeepCell architecture [22] evaluated on images of the BBBC022 dataset[1] as part of the Broad Bioimage Benchmark Collection, a dataset consisting of nuclear images of cells grown on microscopy slides. Fu *et al.* [23] use the SegNet architecture [24] to segment 3D image stacks of rat kidney tissue. Images of lung carcinoma cells grown on slides and images from the BBBC022 dataset were used to evaluate a FCN network structure by Sadanandan *et al.* [25].

[1] https://data.broadinstitute.org/bbbc/BBBC022/

C. Deep Learning Based Methods Approaching Generalizability

Most recently, methods designed to segment nuclear images across different datasets gained attention. Inspired by the 2018 Kaggle Data Science Bowl contest, researchers aimed to solve a generalized nuclear segmentation problem [10]. The dataset used within the Data Science Bowl consists of H&E stained images, confocal fluorescence images and other light-microscopy images, thus the architectures were aimed at learning underlying characteristics of nuclear objects to be able to detect and segment them. Hollandi *et al.* [26] outperformed all contest submissions by using clustering methods and image style transfer to translate all images of the training set for data augmentation. Despite the success of algorithms proposed to segment images of the Data Science Bowl dataset, images from some types of preparation such as tumor-touch imprints or bone marrow cytospins are missing in this dataset. Another approach proposes an online tool [27] to upload annotated imaging experiments to a cloud server and to fine tune pretrained U-Net models on the given dataset to adapt to the specific type of experiment. Stringer *et al.* [28] convert annotation masks into vectorflow representations that can be learned and predicted by a U-Net-shaped deep neural network called Cellpose. The authors claim that a generalist nuclear and cytoplasm segmentation is achieved without the need to retrain the dataset to segment images of a previously unseen dataset.

A systematic comparison of deep learning architectures on complex fluorescence images has not yet been performed.

D. Data Augmentation and Artificial Image Synthesis

Data augmentation techniques can substantially improve the prediction performance of deep neural networks [29]. On biomedical image segmentation tasks, data augmentation was introduced by Ronneberger *et al.* [30] in 2014. Cui *et al.* demonstrated the benefit of data augmentation in nuclear image segmentation of H&E stained histopathological samples [18]. Moshkov *et al.* [31] recently showed that applying the same data augmentation methods used to extend the training set to images upon inference and calculating a pixel-based majority vote over all augmented images increases the segmentation effectiveness of deep networks.

The standard data augmentation techniques currently applied in deep learning (e.g. flipping, cropping, rotation, elastic deformations, intensity variations, shifting, etc.) do not address the vast number of parameters influencing IF based imaging. Parameters include varying image integration time and varying quality and intensity of a given immuno-staining signal. Weak signals have to be captured with higher integration time to ensure an acceptable dynamic range of the resulting images, leading to an overall increased background intensity and thus a low signal-to-noise ratio. Moreover, the intensity of nuclei can vary within a field of view (FOV), depending on the DNA compaction, cell integrity and proper focus settings during image acquisition.

To allow for a better generalization performance in fluorescence nuclear image segmentation, the use of synthetic datasets was proposed. Russell *et al.* [32] created simulated images by modeling nuclear shape and fluorescence image characteristics by overlaying the image with Gaussian noise and blurring. Hou *et al.* [13] proposed a pipeline using real image patches from histo-pathological images and a neural network called *refiner CNN* to create artificial image patches. Dunn *et al.* [14] use a spatially constrained CycleGAN, trained on sub-volumes of the specific dataset to be segmented, to create synthetic volumes based on ellipsoids placed within empty volumes. Mahmood *et al.* [19] used a dual GAN that learns to transform masks, including polygons, to synthetic histo-pathological patches. Bailo *et al.* [33] use objects of the training set segmentation masks to generate photorealistic images of blood cells to extend the training set.

III. Evaluation of Deep Learning Architectures and Conventional Algorithms

We compare the segmentation effectiveness of five deep learning architectures and two conventional algorithms to segment nuclear images of IF stained samples. The deep learning architectures can be divided into two categories: U-Net architectures (U-Net [30], U-Net with a ResNet34 backbone (U-Net ResNet)), U-Net based on transformed image representation (Cellpose) and instance-aware segmentation architectures (Mask R-CNN [34], KG instance segmentation [35]). The conventional methods investigated were specifically designed to solve image segmentation problems on fluorescence nuclear images and consist of a marker-based approach (Iterative h-min) and a model-based approach (ARG).

1) U-Net: The U-Net architecture is prominently used in nuclear image segmentation [9], [27], [30]. The success of this architecture is based on the fact that accurate segmentation is possible even with small training sets. We use a theano/lasagne based implementation of the U-Net.

2) U-Net ResNet: When substituting the deconvolution part of the U-Net architecture by a deeper network structure and by using pre-trained weights from e.g. the ImageNet dataset [16], one could expect to increase segmentation accuracy as previously demonstrated for segmentation and classification tasks [36], [37]. Therefore, we use a U-Net architecture where the feature encoding part, called the "backbon", was substituted by a ResNet34 [36] architecture, using 34 layers for feature encoding but keeping the skip-connections to ensure spatial resolution for the up-sampling part of the architecture. We use an available keras/tensorflow implementation,[2] but changed the loss function to implement the weighted cross-entropy loss by setting a higher loss to nuclear borders as suggested by the U-Net authors [30].

3) Cellpose: Cellpose uses a U-Net shaped neural network with two modifications as compared to the afore mentioned approaches: image masks are not directly predicted but instead a flow-based representation is predicted. The representation builds upon a heat-diffusion simulation where each pixel within the same object can be assigned to a path converging at the center of the object. In addition to the transformed representation of training images and predictions, test time

[2]https://github.com/qubvel/segmentation_models

enhancements are employed to increase the power of the predictive model. We use an available pytorch implementation.[3]

4) Mask R-CNN: Mask R-CNN was designed to solve instance-aware segmentation. The architecture builds upon the Faster Region-based CNN (R-CNN) approach [38] by predicting object masks in parallel to classification of objects in bounding boxes [34]. Thus, unlike the behaviour of DCNNs, Mask R-CNN does not provide a pixel to pixel mapping but rather splits the problem of image segmentation into region detection and subsequent classification and segmentation of region proposals. By focusing on region proposals using coarse spatial quantization for feature extraction, candidate regions can be extracted with high accuracy. We use an available keras/tensorflow implementation.[4]

5) KG Instance Segmentation: The keypoint graph instance segmentation (KG instance segmentation) network [35] was developed to tackle cell instance segmentation tasks. In contrast to approaches such as the aforementioned Mask R-CNN, which typically utilize anchor box based detectors, keypoint-based detectors in combination with multi-scale feature maps are used. Keypoint detection is applied to find five keypoints per cell. These points are then grouped using a keypoint graph to retrieve cell bounding boxes. A publicly available PyTorch implementation[5] was used in this work.

6) Iterative h-Minima-Based Marker-Controlled Watershed: The nuclear segmentation approach proposed by Koyuncu *et al.* [39] collects a set of watershed candidate markers based on the h-minima transform and on multiple scales. They are iteratively selected based on size constraints and used to apply the watershed transform. The effectiveness of the method was demonstrated on images showing nuclei clumps.

7) Attributed Relational Graphs: In contrast to the iterative h-minima based approach, Arslan *et al.* [40] propose an algorithm for the segmentation of nuclei based on image gradient information. The underlying assumption is that a nucleus is composed of four edges that can be retrieved from the image gradient. To this end, nuclear edges called edge primitives are detected, related to each other to obtain a representation of the nucleus and finally a region growing algorithm is applied, starting from the center of the representation. The approach was demonstrated on dense cell line images.

A. Dataset Description

We use a recently published dataset [11] consisting of 79 images of IF stained nuclei images containing 7813 nuclei in total. The images are from specimens of different diagnosis, namely human ganglioneuroblastoma (GNB) tumors, human neuroblastoma (NB) tumors, Wilms tumor (Wilms) and a human keratinocyte cell line (HaCaT). Among those are GNB, NB and Wilms tissue cryosections, HaCaT cell line cytospin preparations, HaCaT cell line cells grown on slide, NB cell line cytospins, NB bone marrow cytospin preparations and NB touch imprints.

[3]https://github.com/mouseland/cellpose
[4]https://github.com/matterport/Mask_RCNN
[5]https://github.com/yijingru/KG_Instance_Segmentation

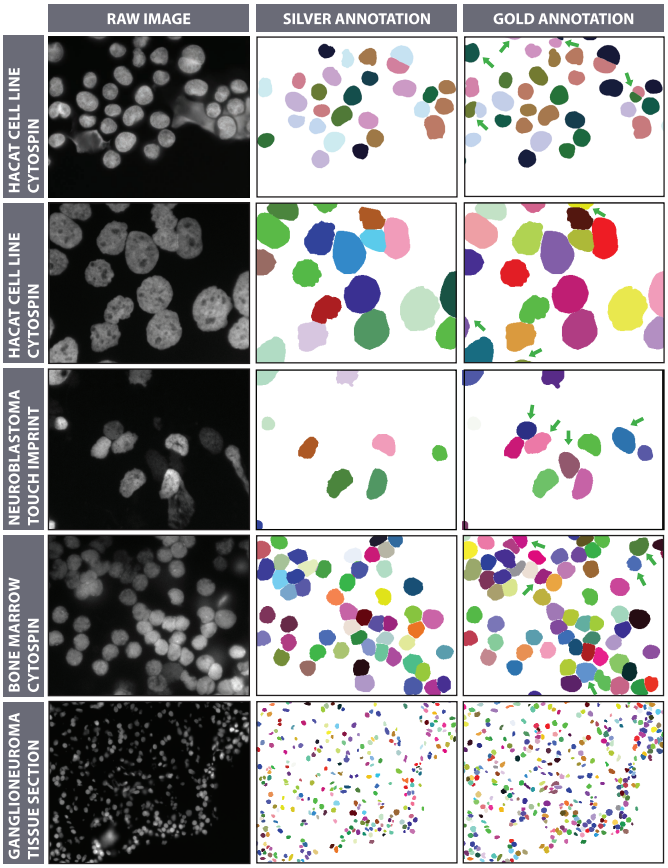

Fig. 2. Examples of all types of preparations/specimen including a comparison between *silver-standard* and *gold-standard* annotations. Green arrows indicate differences between *silver-standard* and *gold-standard* annotations.

The dataset contains accurate nuclear annotations and proposes a split into a training set and a test set. The training set exists as *silver-standard* set and *gold-standard* set, while the test set is only *gold-standard*. The difference between the two annotation standards is given by the level of expertise preparing the annotations (trained under-graduates vs. expert pathologists and biologists), the accuracy of nuclear outlines (contours) and the number of annotated objects (in *gold-standard* annotations, nuclei with weak intensity, nuclei partially present at image borders and the in-focus parts of partially out-of-focus nuclei were annotated, leading to a higher number of annotated nuclei, especially in GNB tissue cryosections). The test set is made up of two sets, one representing images acquired with the same conditions as the training set images, further called *similar test set*, and the other acquired with different conditions, further called *new conditions test set*. The latter allows to evaluate the generalizability of all trained architectures. Both sets form the *cumulative test set*. Examples of all types of preparations/specimens present in the similar test set including a comparison between *silver-standard* and *gold-standard* annotations are given in Figure 2.

B. Image Complexity

The complexity to annotate or segment a fluorescence nuclear image varies between images of different preparation

type, tissue and imaging conditions such as signal-to-noise ratio, sharpness or presence of damaged nuclei. To some extent, it depicts a subjective measure, correlating with the underlying ability to unambiguously separate nuclei from each other, either by experts or by automated methods. There are two ways to estimate image complexity: 1. By classification through experts, i.e. pathologist, biologist, image analysis experts, and 2. by making use of image features potentially representing image complexity. We aimed to address both approaches by calculating the correlation of image-derived features with expert classification.

To generate expert annotations of image complexity, we created a scoring scheme to rate the segmentation complexity of each image of the test set, by the subjective impression of each annotator. We decided to assign each image to one out of three classes: low-, medium- and high complexity level. Four independent experts (two biomedical imaging experts and two biologist experts) scored each image two times (Suppl. Fig. 1a). We then calculated the final complexity annotation by calculating the mean experts' score rounded to the next integer (class level) for each image, resulting in an annotation of one of the three classes for each image.

To analyze whether the generated complexity annotation correlates with image-derived features such as cell density and thus, the latter can be used to represent image complexity, we extracted image features potentially representing nuclear image complexity from all annotated images (Suppl. Tab. I). We then calculated the Spearman's correlation coefficient [41] between each feature extracted and the mean experts' score (Suppl. Fig. 1b). Resulting coefficients show that none of the features highly correlates with mean experts' scores. A medium correlation (correlation coefficient >0.4 and <0.5) is given by the variance of nuclear size (measured by the nuclear area) and the variance of nuclear mean intensities. Thus, the complexity of an image cannot solely be defined based on image-derived features.

Based on the feedback received from expert biologists and pathologists and the image-level features extracted, we conclude that the complexity is influenced by the quality of the image, the signal-to-noise ratio, the number of aggregated/overlapped nuclei, the number of damaged nuclei, the number of out-of-focus nuclei and the homogeneity of nuclear intensity and size. We formally describe the three complexity classes as follows.

- *Low Complexity Level:* Almost no touching nuclei are available in the image. If nuclei touch, they appear with sharp contrast to other nuclear instances and nuclear borders in between the touching nuclei. Only a minor number of burst, damaged or out-of-focus nuclei are present within the image with respect to the total number of nuclei. Nuclear size and morphologies are almost constant, nuclear intensities do not extensively vary.
- *Medium Complexity Level:* Images contain clumps or damaged or out-of-focus nuclei, but in a modest frequency as compared to the total number of nuclei. If images are blurred, nuclei have to appear separated to each other, for a vast majority of nuclei. Borders between nuclei might disappear, but each nucleus has not more than three to four direct neighbors and nuclear instances can be concluded by the shape of single instances within a clump. Nuclear intensities vary, but nuclear sizes are almost homogeneous with only a minor number of nuclei with diverging sizes as compared to the total number of nuclei.
- *High Complexity Level:* These images are characterized by a high number of clumps and/or a high number of burst nuclei and/or a high number of out-of-focus objects or damaged objects and/or a high variation of nuclear size. Nuclei occuring in clumps present varying morphologies and borders can frequently not be determined. Overall, annotating nuclei is highly challenging in these images and uncertainty in annotating at least a modest number of nuclei remains.

Example images for the three levels of complexity are visualized in Suppl. Fig. 2.

C. Artificial Image Generation

We evaluate the influence of adding artificially generated images to the training set on segmentation effectiveness.

The strategy to create artificial images is as follows: we randomly select an annotated natural image and the corresponding annotation mask from one of the datasets with respect to a certain tissue origin/specimen. By dilating the foreground regions of the binarized annotation mask using a circle-shaped structuring element of size 15, pixels with values higher than the mean background intensity, occurring due to blurred nuclei, are included. When inverting the resulting mask, only the region of pixels representing the background signal is covered. We now iteratively sample the intensity values of random pixels from this region and assign them to one of the pixels of an empty image patch. This is done until all pixels of the respective image patch are set. Thus, the background pixel value distribution of the created image patch roughly matches the one of the original image. Subsequently, we use arbitrary nuclei cropped from the annotated image, transform them by rotation, size variation, elastic deformation, intensity variation, blurring, adding Gaussian noise, and combinations of those operations. The transformed, cropped nuclei are then placed at crossing positions of a grid virtually overlaid with the image patch, including a randomly added offset in x- and y- directions. The maximum offset value gives the probability of a nucleus to overlap with a neighbor nucleus placed on the grid. For each crossing position on the grid, we randomly decide if the object shall be placed or not. If a nucleus to be inserted overlaps another nucleus already present, we randomly decide to either replace the existing, overlapped part of the nucleus or to add it to the existing nucleus scaled with a constant between 0 and 1 to imitate overlap. Thus, we can simulate aggregating and overlapping nuclei. The same augmentation and placement is done for cropped nuclear masks, placed on a new image mask patch, except that placed nuclei masks always replace overlapping parts of existing nuclei masks as we do not model masks with fuzzy annotations.

Finally, we obtain a nuclear image and a mask image. The former contains random nuclei augmented with image transformation strategies. The latter contains labeled objects placed on the same positions as the nuclei in the nuclear image. Configuration details can be obtained in Suppl. Tab. II.

Nevertheless, we observed that training DL architectures using such artificially generated images does not lead to better segmentation results. This may be due to the fact that nuclei naturally showing blurred borders in IF images do not show those when cropped, transformed and placed on new image patches, guiding the network to learn features differently from natural image features. To overcome this issue, we trained an image-to-image translation GAN [42] to learn the transformation of artificially generated images into natural-like images. The GAN is trained on pairs of natural and artificial images, where nuclei are cropped from natural image patches and placed at the very same position on a new image patch. By training the network on these paired images, the network is forced to learn the implicit transformation of artificial to natural-like images. The final workflow to create natural-like artificial images is depicted in Figure 3.

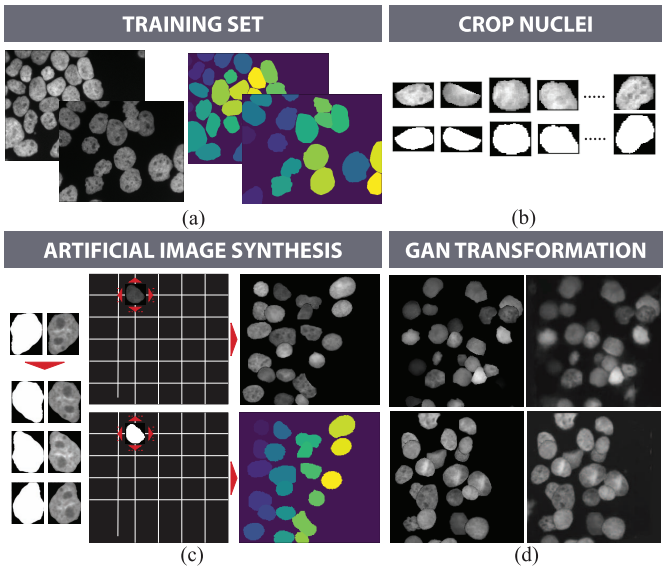

Fig. 3. Generation of natural-like artificial image patches. Nuclei and respective mask objects of (a) a training image patch are (b) collected and cropped from the raw nuclei image respective mask (c). Each nucleus patch is arbitrarily augmented using rotation, intensity variation, elastic deformation, flipping. The same morphological transformations are applied to the mask patches. We create a new image patch and set the background pixels sampled from the original raw image background. Subsequently, we place nuclei at certain positions induced by a grid sized 3×3, 5×5, etc. For each position on the grid, we randomly decide if a nucleus shall be placed there and if so, we add a random offset, where the maximum offset indicates the probability to overlap with neighbor nuclei. The same placement is performed for mask objects on a new mask patch. Finally, a GAN is used to transform the artificial image into a natural-like image (d). This is done for each dataset independently. Training set images were scaled such that all nuclei have the same mean size. First image pair: artificial/natural-like HaCaT image. Second image pair: artificial/natural-like neuroblastoma image.

D. Pipeline for Architecture Comparison

To evaluate the potential of state-of-the-art architectures to segment nuclear images across various tissue origins and sample preparation types with varying levels of image complexity, we set up a pipeline to enable an objective comparison. The code is publicly available.[6] The pipeline, illustrated in Figure 4, operates as follows.

Dataset Split: We use the split proposed with the dataset [11] into training and test set and further split the training set into a training and a validation set, for all of the three different tissue origins present in the training set (HaCaT cell line, NB tumor, GNB tumor) separately. We do not consider the different preparation types applied to the imaged samples for architecture training, but we split the dataset such that at least one image of each sample preparation type present in the training set is contained in the validation and test set. Moreover, the test set consists of additional images acquired using different modalities, signal-to-noise ratios, magnifications and sample specimens. All images of these datasets are further called *natural images*, in contrast to *natural-like artificial images* that result from artificial image synthesis.

Rescaling&Tiling: We use a self-implemented version of the U-Net architecture and third-party generic implementations of the Mask R-CNN, the KG instance segmentation, the U-Net Resnet34 and the Cellpose architectures. There are two ways to use generic implementations of CNNs for a specific dataset: 1. the architectures are modified to the data at hand or 2. the specific dataset is transformed to fit the input layer of the architectures. We decided for the latter to allow straightforward re-use for increased reproducibility. Moreover, this allowed us to use available pre-trained weights.[7] Thus, we transformed the dataset to fit the input layer of the architecture evaluated, for each architecture, and did not modify the generic architectures with respect to the specific dataset used. We set the input layer size of the U-Net architecture to 256×256 while the generic third-party implementations (Mask R-CNN, U-Net ResNet, KG Instance segmentation) expect RGB images resulting in an input layer size of $256 \times 256 \times 3$ or $256 \times 256 \times 2$ (Cellpose). To fit this input sizes, we duplicated or triplicated all images to obtain RGB images.[8]. To prepare the dataset to fit the given input layer dimensions, we apply a tiling strategy to the dataset images as proposed by Ronneberger *et al.* [30] in order to obtain image patches sized 256×256. By using this tiling strategy we observe the following benefits: 1) The training and validation sets are extended due to overlapping tiles. 2) Overlapping tiles prevent artefacts at tile borders when reconstructing final network predictions after network inference on the test set image tiles. Moreover, we rescale the natural images such that nuclei have equal mean size across all images as we want to evaluate the impact of rescaling images on the segmentation effectiveness. We calculated the nuclear area as measure for nucleus size. To rescale images, the mean nuclear size of an image was calculated based on the mean

[6] https://github.com/perlfloccri/NuclearSegmentationPipeline

[7] Pre-trained on the ImageNet dataset (KG instance segmentation, U-Net ResNet.) or the Pascal COCO dataset (Mask R-CNN)

[8] A further minor improvement of segmentation effectiveness might be achieved by customizing publicly available colour image architectures to single-layer inputs and pre-training with grayscale images (e.g. with ImageNet images converted to grayscale) [43]

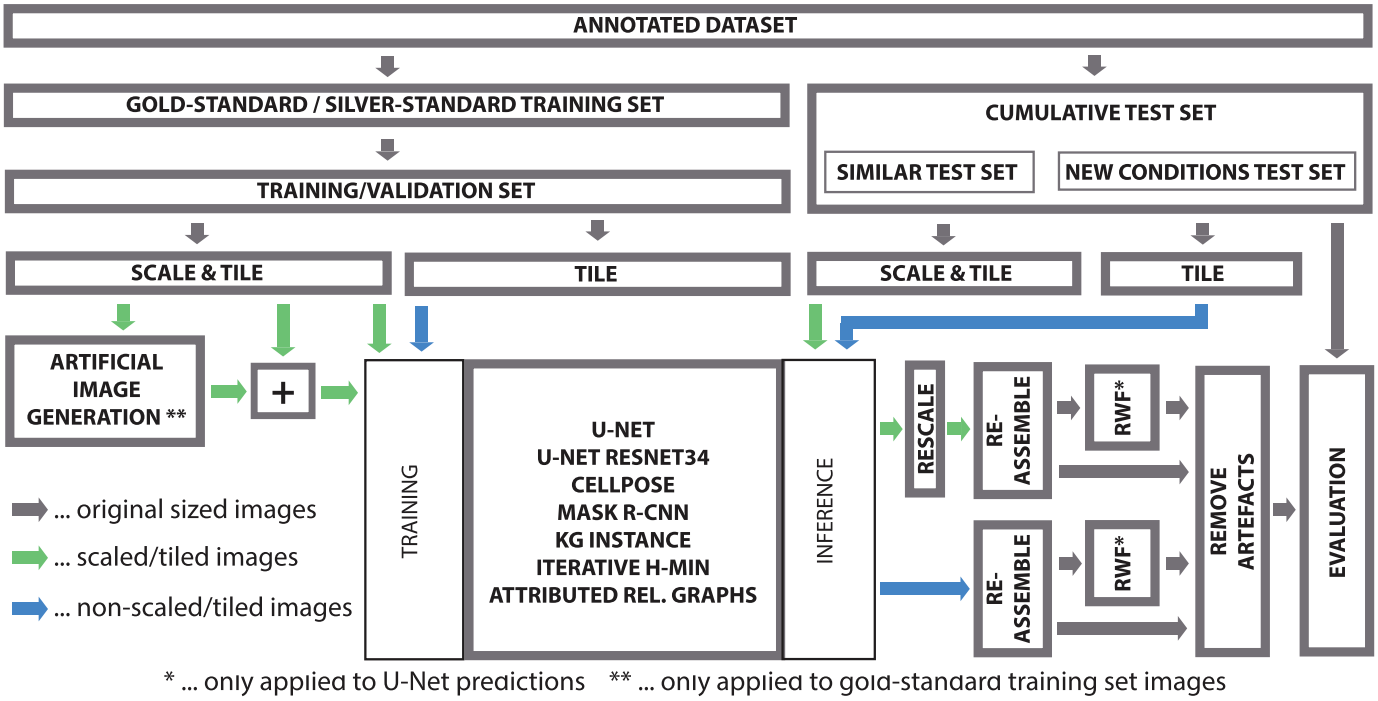

Fig. 4. Pipeline for training and evaluation of deep learning architectures for instance-aware nuclei image segmentation.

size of all nuclear mask objects within the corresponding mask image. Subsequently, all images and masks were resized.

Artificial Image Generation: In addition to the natural images of the training and validation set, we create artificial images of size 256×256 as described in Section III-C and add them to the training set. As we aim to additionally compare the segmentation effectiveness for all architectures between *silver-standard* and *gold-standard* training sets, we apply the same steps except for generating artificial images to the *silver-standard* dataset.

Network Training: We then train all architectures four times using the:

- non-scaled natural images of the *gold-standard* training set
- scaled natural images of the *gold-standard* training set containing nuclei with equal mean size across images
- scaled natural images of the *silver-standard* training set containing nuclei with equal mean size across images
- scaled natural and equally scaled natural-like artificial images of the *gold-standard* training set

Rescaling & Reassembling: After network inference on the test set patches, the patches are reassembled and rescaled to fit the original image size. Thus, prediction results can be compared across all architectures. The output of the tested networks differs: while the U-Net architectures, except for Cellpose result in a probability map, Mask R-CNN and KG instance segmentation inference result in object masks, one mask for each detected object. We threshold the resulting, reassembled probability maps of the U-Net architectures by a value of 0.5 to obtain binary masks and label them to obtain the labeled object masks. For reassembled Mask R-CNN and Cellpose predictions, we label each object and add it to a new image mask to obtain the final labeled object mask.

Recursive Waterflow Post-Processing: We apply the recursive waterflow (RWF) algorithm [44], an algorithm tackling under-segmentation problems, to the U-Net predictions to evaluate the influence of post-processing on the segmentation effectiveness. The only parameter we adapt is related to the mean nuclear size, the other parameters are fixed (see Suppl. Tab. II for details).

Artefact Removal: We apply a common post-processing to all predictions by removing small artefacts that do not fit the size of nuclei, where the threshold was calculated from the respective ground truth mask (size of the smallest nuclear instance). Thus, this post-processing is based on prior knowledge, but is equal for all architectures investigated, is independent of shape or intensity-based features and is not specific to a certain sample preparation type or tissue type.

E. Conventional Segmentation Method Parameter Tuning

Conventional segmentation methods are most frequently parametrized and these have to be tuned to adapt an algorithm to a specific dataset. The two methods evaluated within this work (Iterative h-min, ARG) use four parameters each. We applied a grid search on all possible parameter combinations within a range predefined according to preliminary experiments. To this end, we applied a coarse grid search followed by a fine grid search based on the best coarse score (Suppl. Tab. III). The score to be minimized in order to obtain the best parameter combination for the coarse- and the fine grid search is $F1 - std(f1) + AJI - std(AJI) - |(F1 - AJI)|$, where std is the standard deviation. Thus, for each possible combination we calculate the object-level F1-score and the

Aggregated Jaccard Index (AJI, see Section IV). The best combination maximizes both scores, where the standard deviation (std) of each score is minimal and so is the distance between both scores, ensuring to balance both scores.

F. Data and Code Availability

The annotated dataset used is publicly available at the EMBL BioStudies database, accession number S-BSST265 [11], a detailed description is available [12]. The code used to evaluate the segmentation methods is publicly available.[9]

IV. Evaluation Metrics

Most authors provide object-level as well as pixel-level metrics to evaluate nuclear segmentation methods. While pixel-level metrics penalize deviation from predicted foreground regions to those of ground truth annotations, object-level metrics used in biomedical image segmentation evaluation count and classify predicted objects/instances. To overcome the problem of choosing the right class of metrics, Kumar *et al.* [45] proposed a combined metric, the Aggregated Jaccard Index (AJI), taking object- and pixel-level errors into account. This is achieved by computing an aggregated intersection cardinality numerator and an aggregated union cardinality denominator for all predicted- and ground truth objects. The AJI is prominently used in recent publications [7], [8], [19]. The disadvantage of using this combined metric is that it is not obvious if pixel or object-level errors contribute to a low AJI value.

We provide both types of metrics and report the AJI in addition. Object-level metrics reported are precision (PREC), recall (REC), their harmonic mean, the F1 score (F1), under-segmentation (US) and over-segmentation (OS). Pixel-level metrics presented are mean Dice score (mDICE) and mean Jaccard Index (mJI), calculated on all true positive (TP) objects. The mDICE and mJI are calculated on TP objects only to evaluate the accuracy of nuclear boundaries irrespective of the number and shape of FN or FP objects. We also report the combined metric AJI for completeness as it is prominently used in related publications, but do not focus the discussion on this metric.

We consider a ground truth object to be detected if more than 50% of the ground truth object's pixels are covered by predictions, and assign it as false negative (FN) otherwise. We count a ground truth object as TP if it is overlapped by exactly one predicted object with a JI between ground truth object and predicted object of greater than 0.5. Predicted objects only touching a part of the object would count as false positive (FP). If more than one predicted object overlaps the ground truth object such that the overlapping area covers more than 50% of the predicted object's area, the ground truth object is considered as OS. An object is classified as FP if it overlaps less than 50% with the ground truth.[10] If more than one ground truth object overlaps the predicted object such that the overlapping area covers more than 50% of the ground truth object's area, the ground truth objects overlapped by the prediction are classified as US. We report US (resp. OS) as the ratio of the number of under-segmented nuclei (resp. over-segmented objects) to the number of ground truth objects.

[9]https://github.com/perlfloccri/NuclearSegmentationPipeline

[10]Examples of possible cases of ground truth objects and predictions are illustrated in Suppl. Fig. 3

V. Results

Segmentation effectiveness of deep learning architectures depends on their design and on several additional factors such as the size and quality of the dataset and the architectures' hyperparameters. To systematically evaluate the selected deep learning architectures on small and complex to segment fluorescence nuclear image datasets, we fixed the hyperparameters of all architectures[11] and trained them with the images as described in Section III-D *Pipeline for architecture comparison*. The conventional methods were finetuned as described in Section III-E *Conventional segmentation method parameter tuning*.

Related publications suggest that rescaling images to the same mean nuclear size improves segmentation results if deep learning architectures are applied [26]. In order to identify the most effective strategy for the evaluation of our novel artificial image synthesis approach, we first aimed to confirm this claim. We compared predictions of all deep learning architectures trained on non-scaled vs. scaled images (which resulted in images showing equal mean nuclear sizes). We did not include conventional method methods in this comparison as the test set contains images captured using objectives with magnifications not seen during training. Applying conventional methods tuned on images with magnifications diverging from those of test set images would potentially underestimate their segmentation effectiveness.

As expected, quantitative results show that scaling images to the same mean nuclear size across images results in overall improved REC and PREC scores for all deep learning architectures (Suppl. Fig. 4a). Based on these results, we used scaled images for all subsequently applied experiments.

A. Silver- vs. Gold-Standard Training

We next compared all methods, including conventional ones, on two types of annotated training sets: *silver* and *gold-standard* annotations. While *silver-standard* training sets were generated by trained under-graduate students and contain partially inaccurate annotation masks, *gold-standard* training sets were carefully generated and curated by biology and pathology experts. The process of image annotation is expensive with respect to time and resource requirements, in particular if pathology experts are required. Thus, training with *silver-standard* images created by trained under-graduates could be highly valuable if the results were comparable with results obtained by training with *gold-standard* datasets. To show the impact of annotation standard on the segmentation effectiveness, we trained all deep learning architectures while applying parameter tuning for the conventional methods on both training sets (*silver-standard* and *gold-standard*) and evaluated the prediction results on the cumulative test set.

[11]See Suppl. Table IV

TABLE I

NUCLEI SEGMENTATION METRICS FOR ALL ARCHITECTURES ON THE CUMULATIVE TEST SET, COMPARING DIFFERENT TRAINING CONDITIONS: *silver* VS *gold-standard* TRAINING SETS, SCALING VS. NON-SCALING, POST-PROCESSING VS. NON-POST-PROCESSING AND ARTIFICIAL VS. NON-ARTIFICIAL TRAINING SET IMAGES. THE BEST VALUE PER METRIC IS HIGHLIGHTED

Architecture	FP	TP	FN	US	OS	REC	PREC	F1	mDICE	mJI	AJI
U-Net scaled silver	806	3899	1189	0.096	0.008	0.766	0.829	0.796	0.896	0.818	0.739
U-Net scaled gold	761	4036	1052	0.134	0.010	0.793	0.841	0.817	0.911	0.841	0.758
U-Net non-scaled gold	1050	2888	2200	0.357	0.003	0.568	0.733	0.640	0.896	0.817	0.622
U-Net scaled gold + artificial	928	3747	1341	0.205	0.007	0.736	0.801	0.768	0.910	0.839	0.756
U-Net ResNet scaled silver	806	3498	1590	0.196	0.003	0.688	0.813	0.745	0.908	0.836	0.718
U-Net ResNet scaled gold	607	4034	1054	0.125	0.009	0.793	0.869	0.829	0.915	0.848	**0.779**
U-Net ResNet non-scaled gold	843	2767	2321	0.395	**0.002**	0.544	0.766	0.636	**0.921**	**0.859**	0.649
U-Net ResNet scaled gold + artificial	624	4216	872	0.105	0.013	0.829	0.871	0.849	0.915	0.847	0.775
Cellpose scaled silver	411	3854	1234	0.047	0.006	0.757	0.904	0.824	0.882	0.793	0.766
Cellpose scaled gold	428	4218	870	0.054	0.010	0.829	**0.908**	0.867	0.896	0.816	0.763
Cellpose non-scaled gold	705	3491	1597	0.064	0.009	0.686	0.832	0.752	0.885	0.802	0.743
Cellpose scaled gold + artificial	478	4124	964	0.042	0.015	0.811	0.896	0.851	0.880	0.791	0.768
Mask R-CNN scaled silver	485	4123	965	0.057	0.005	0.810	0.895	0.850	0.860	0.759	0.723
Mask R-CNN scaled gold	458	4295	793	0.062	0.008	0.844	0.904	0.873	0.873	0.778	0.737
Mask R-CNN non-scaled gold	579	3776	1312	0.093	0.003	0.742	0.867	0.800	0.867	0.771	0.761
Mask R-CNN scaled gold + artificial	**546**	4401	687	0.058	0.012	0.865	0.890	**0.877**	0.872	0.776	0.722
KG instance scaled silver	613	4353	735	0.051	0.009	0.856	0.877	0.866	0.872	0.778	0.718
KG instance scaled gold	581	4417	671	0.057	0.010	0.868	0.884	0.876	0.894	0.812	0.740
KG instance non-scaled gold	597	3843	1245	0.124	0.004	0.755	0.866	0.807	0.878	0.790	0.726
KG instance scaled gold + artificial	718	**4441**	**647**	0.062	0.022	**0.873**	0.861	0.867	0.883	0.795	0.703
Iterative h-min scaled silver	1027	3553	1535	0.075	0.013	0.698	0.776	0.735	0.844	0.736	0.544
Iterative h-min scaled gold	1106	3464	1624	0.059	0.014	0.681	0.758	0.717	0.833	0.719	0.535
Attributed relational graphs scaled silver	1621	3906	1182	0.086	0.056	0.768	0.707	0.736	0.881	0.794	0.621
Attributed relational graphs scaled gold	1263	3900	1188	0.114	0.031	0.767	0.755	0.761	0.885	0.799	0.633

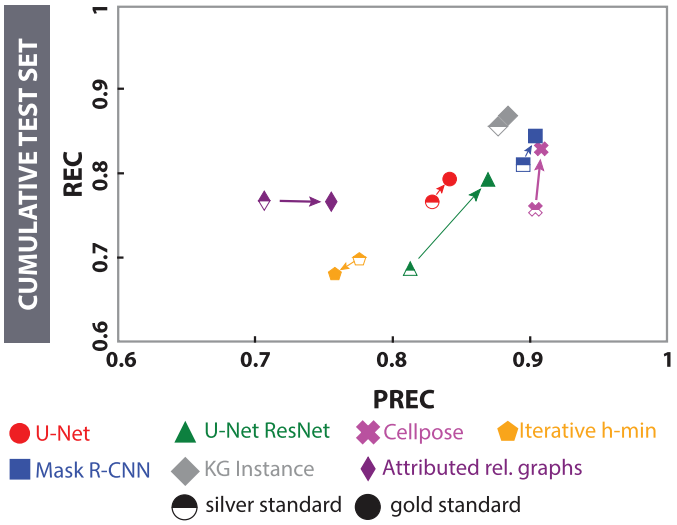

Fig. 5. Performance (precision vs. recall) of deep learning architectures and conventional methods for the silver vs. gold training set on the cumulative test set. Using *gold-standard* training sets, REC and PREC is increased for all deep learning architectures in comparison to *silver-standard* training sets, indicated by the colored arrows connecting silver and gold scores for each architecture. Conventional methods do not profit from parameter finetuning on the *gold-standard* training set.

As expected, PREC and REC increases for all deep learning architectures when trained on *gold-standard* images as compared to *silver-standard* images (see Figure 5). The effect is lowest for the instance-aware segmentation architectures Mask R-CNN and KG instance segmentation while these architectures achieve the highest PREC and REC scores. The conventional algorithms do not profit from *gold-standard* training. The influence of annotation standards on other metrics can be observed in Table I and show that pixel-level scores (mDICE, mJI) are overall only slightly increased when training with *gold-standard* images.

B. Image Complexity

To measure the segmentation effectiveness with respect to the image complexity level and to prove the potential of all methods investigated to generalize to unseen imaging and sampling conditions, we evaluated all methods on the similar and the new conditions test set (Figure 6).

In general, with increasing image complexity evaluation scores decrease (REC, PREC, F1, mDICE, AJI). This holds true for all methods as expected, except for the Iterative h-min method on the similar test set. These method achieves highest REC and F1 scores on high complex images. US scores slightly increase for all methods with rising complexity level, while the overall level is highest for the U-Net, the U-Net ResNet and the ARG method. OS stays constant and is only increased for high complex images of the new conditions test set for the Iterative h-min method. All deep learning architectures achieve comparable scores on the similar and on the new conditions test set. The scores achieved by the conventional methods on high complex images of the new conditions test set (REC,PREC, F1, AJI) are overall decreased as compared to the similar test set.

C. Artificial Data

Annotated fluorescence nuclear image datasets are most frequently rare and of limited size and complexity. To further improve the segmentation effectiveness of the investigated deep learning architectures, we created a strategy to simulate complex nuclear images as discussed in Section III-C. We evaluate the influence of adding artificially generated

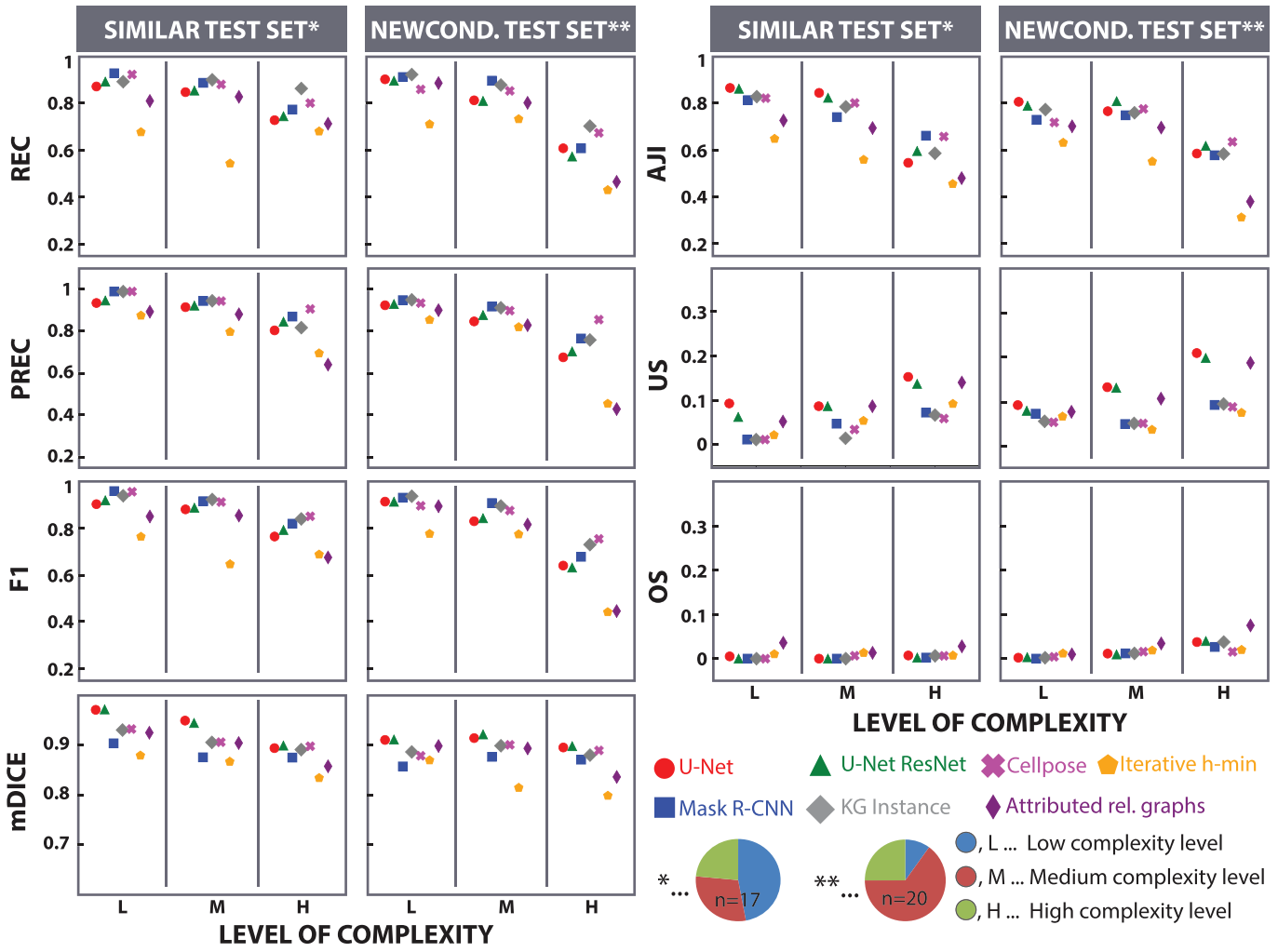

Fig. 6. The influence of segmentation complexity on segmentation effectiveness, evaluated on the similar and the new conditions test set. The pie charts show the distribution of complexity classes within the respective test sets.

images to the training set on the similar and the new conditions test set with respect to the complexity levels (Figure 7a).

Extending the training set by artificially generated images leads to an increased F1 score for the U-Net ResNet and the Mask R-CNN architecture on highly complex images in both test sets. In particular, Mask R-CNN and KG instance, trained on natural and artificial images achieve the overall highest F1 score and REC score, respectively, when compared to all evaluated conditions, and segmentation methods (see Tabl. I). Applying the generated artificial images to the shallow U-Net architecture decreases the F1 score on medium and complex images. The KG instance segmentation and Cellpose architectures segmentation effectiveness is not affected by adding artificial data.

We next evaluated the influence of adding artificial data with respect to the different preparation types (Figure 7b). Adding artificial training data leads to top F1 scores on images of three out of five preparation types: tissue sections, tumor touch imprints (Mask R-CNN) and cell line cytospins (U-Net ResNet). Tissue sections represent high complex images while tumor touch imprints are usually of medium complexity and cell line cytospins result in images of low- and medium complexity. In BM cytospins of the similar test set, where one image of each complexity level is represented, Mask R-CNN and KG instance segmentation F1 scores decrease when adding artificial data, while the F1 score increases for the U-Net and the U-Net ResNet architectures. However, the U-Net ResNet architecture benefits from artificial data for all preparation types. Applied on complex images of the new conditions test set, the KG instance segmentation and Cellpose architectures achieve the highest F1 scores. A comprehensive visualization of results evaluated on the cumulative test set including conventional methods and evaluated with respect to the preparation type is visualized in Suppl. Fig. 4b.

D. Influence of RWF Post-Processing

Recent publications suggest that deep learning architectures achieve improved segmentation results if post-processing strategies are applied [46]–[49]. These can only be used with segmentation architectures generating probability maps as output. As we observed that the U-Net architectures (U-Net, U-Net ResNet) achieved the overall highest mDICE scores on TP objects, we aimed to investigate whether an

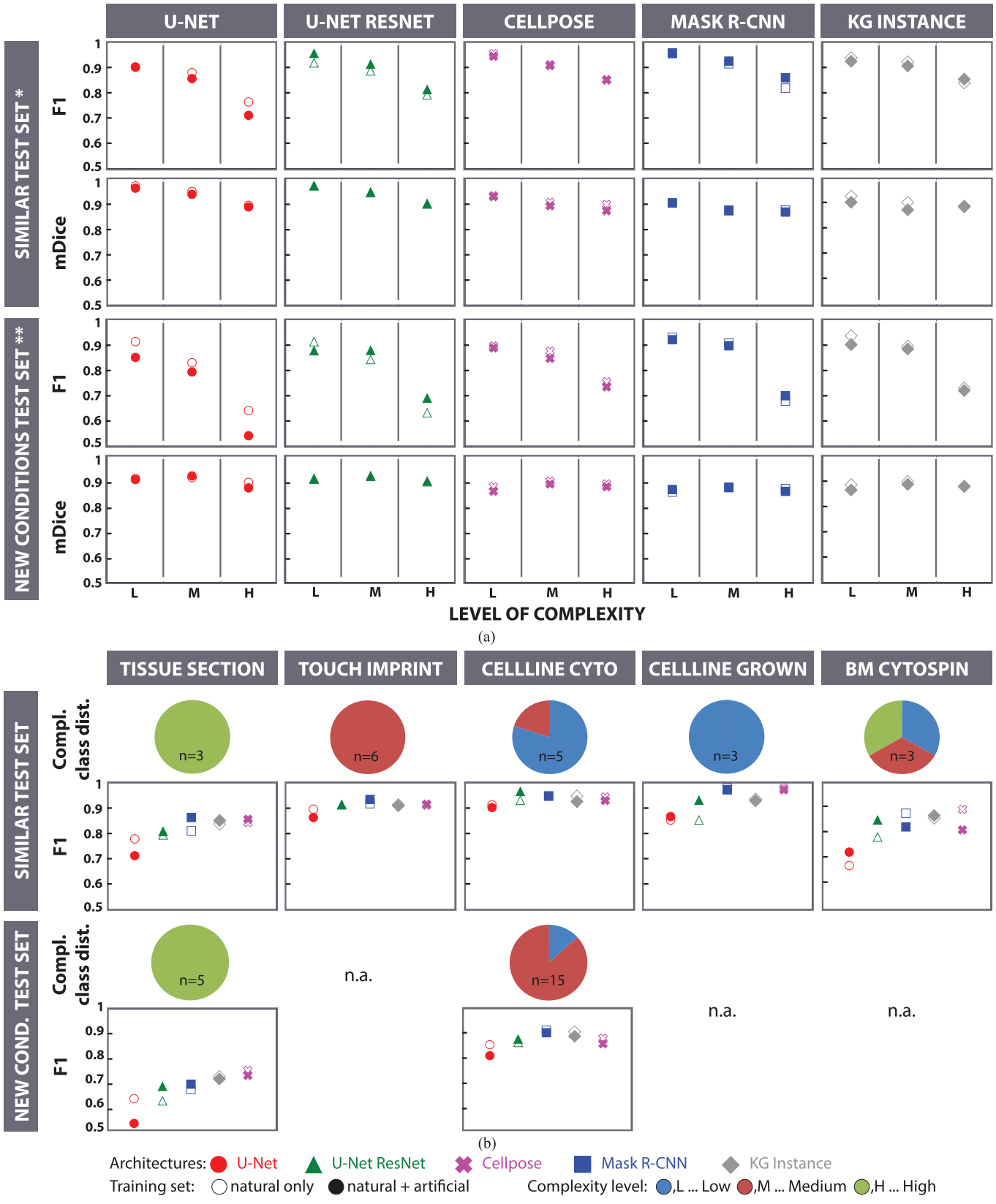

Fig. 7. The influence of extending deep learning training sets with artificial images on segmentation effectiveness with respect to a) segmentation complexity and b) preparation type. Compl. class dist.: Complexity class distribution within the respective test set.

easy-to-apply post-processing method can increase U-Net segmentation effectiveness for complex images, thereby balancing the effort to adapt the segmentation method and the possible improvement of segmentation effectiveness. As our preliminary results indicated that U-Net segmentations are prone to under-segmentation but not to over-segmentation,

TABLE II

MEAN DICE COEFFICIENT OF THE HUMAN EXPERT ANNOTATIONS AND THE ARCHITECTURE'S AND CLASSICAL ALGORITHM'S PREDICTIONS WITH RESPECT TO THE RANDOMLY SELECTED GROUND TRUTH ANNOTATIONS. ANNOT. EXP.: ANNOTATION EXPERT, BIOL. EXP.: EXPERT BIOLOGIST. BOLD VALUES MARK RESULTS COMPARABLE TO HUMAN EXPERT LEVEL

Annotator/Architecture	GNB-I	GNB-II	NB-I	NB-II	NB-III	NB-IV	NC-I	NC-II	NC-III	TS	overall
Biol. exp.	0.853	0.781	0.883	0.932	0.802	0.946	0.969	0.869	0.969	0.895	0.890
Annot. exp.	0.877	0.849	0.896	0.892	0.888	0.957	0.969	0.928	0.973	0.895	0.912
U-Net scaled gold	0.796	0.683	0.865	0.889	0.777	0.876	0.862	**0.907**	0.938	0.662	0.825
U-Net scaled gold + artificial	0.770	0.648	0.841	**0.912**	0.649	0.839	0.887	**0.903**	0.920	0.632	0.800
U-Net scaled gold + RWF	0.833	0.714	0.856	0.804	**0.814**	0.897	0.913	**0.907**	0.906	0.785	0.843
U-Net scaled gold + artificial + RWF	0.822	0.655	0.846	0.816	0.740	0.885	0.882	**0.903**	0.857	0.829	0.823
U-Net ResNet scaled gold	0.800	0.582	0.871	0.838	**0.809**	0.875	0.838	**0.908**	0.937	0.702	0.816
U-Net ResNet scaled gold + artificial	0.814	0.704	0.874	**0.900**	0.749	0.895	0.944	**0.900**	0.915	0.790	0.849
U-Net ResNet scaled gold + RWF	0.816	0.600	0.874	0.773	**0.841**	0.900	0.898	**0.908**	0.958	0.789	0.836
U-Net ResNet scaled gold + artificial + RWF	0.827	0.720	0.874	0.863	0.784	0.913	0.943	**0.900**	0.930	0.863	0.862
Cellpose scaled gold	0.744	0.566	0.859	0.766	0.791	0.870	0.904	**0.889**	0.894	0.760	0.804
Cellpose scaled gold + artificial	0.763	0.615	0.829	0.736	0.726	0.780	0.854	0.842	0.861	0.712	0.772
Mask R-CNN scaled gold	0.707	0.414	0.818	0.820	**0.841**	0.800	0.889	0.866	0.860	0.774	0.779
Mask R-CNN scaled gold + artificial	0.804	0.653	0.821	0.815	**0.836**	0.835	0.892	0.856	0.846	0.753	0.811
KG instance scaled gold	**0.856**	0.778	0.831	0.864	**0.826**	0.867	0.889	**0.883**	0.822	0.727	0.834
KG instance scaled gold + artificial	**0.878**	0.646	0.825	0.857	**0.821**	0.818	0.831	0.865	0.818	0.825	0.818
Iterative h-min scaled gold	0.702	0.485	0.386	0.738	0.708	0.531	0.631	0.765	0.650	0.588	0.619
Attributed relational graphs	0.689	0.551	0.833	0.832	0.746	0.785	0.812	**0.894**	0.736	0.529	0.741

we applied the RWF algorithm, an algorithm specifically designed to tackle under-segmentation problems in cell image segmentation (Suppl. Fig. 5).

The US rate improves for medium and high complex images if RWF post-processing is applied, while the OS rate decreases. Overall, the F1 score is not increased (U-Net) or only slightly increased on complex images (U-Net ResNet) by RWF post-processing. The mDICE score is not affected, while the AJI score overall decreases when RWF post-processing is applied.

E. Comparison to Human Experts

To set a baseline for automated segmentation methods, the dataset used provides single-cell annotations from independent human experts. They were created by randomly selecting 25 nuclei from images and masks of each of the 10 test set classes, further called single-cell ground truth annotations. The selected nuclei were then marked with red crosses on raw images and presented to two independent biomedical imaging experts for annotation. Then, mDICE scores comparing between the experts' annotations and the single-nuclei ground truth annotations were calculated [11] and set the baseline for automated methods to compete with.

We decided to compare all segmentation methods trained or finetuned on the scaled textitgold-standard annotations. To calculate the mDICE scores for each method with respect to the test set classes, we first selected the predicted objects with the highest overlap to the respective objects of the single-cell ground truth annotation. We then calculated the mDICE score between all selected predicted objects and the respective single-cell ground truth annotations for each test set class (Table II).

Overall, the mDICE score baseline on randomly selected nuclei, set by the independent human experts, cannot be achieved by the investigated methods. The KG instance segmentation architecture can achieve human expert level for the GNB-I, NB-III and NC-II class. Human expert level is reached for the NB-III class by all deep learning architectures except for Cellpose. The NC-II class can be segmented by all methods except for Mask R-CNN and Iterative h-min with human level performance, containing cells with low signal-to-noise ratios but almost no touching or damaged cells (low complexity level).

VI. DISCUSSION AND CONCLUSION

Imaging-based microscopy analysis workflows can help biologists to analyze and explore biological specimens by processing images of single or multiple tissue samples, thereby investigating multi-target and multi-scale features such as tissue cell type composition, antibody expression, or DNA level chromosome alterations. Applied to complex nuclear images such as tissue sections, detection of subtle biological effects or generation of reliable quantitative results can only be assured if the segmentation method utilized can provide instance-aware segmentation results.

To this end, we compared and evaluated the segmentation effectiveness of five deep learning architectures and two conventional methods on an expert-annotated nuclear image dataset composed of images from multiple sample preparation types showing a wide range of variations. We investigated the concept of image complexity to evaluate state-of-the-art deep learning architectures and conventional methods with respect to the segmentation challenge these images present. In addition, we evaluated the influence of image rescaling, the quality of groundtruth annotation standard and the influence of U-Net post-processing on segmentation effectiveness. Evaluation was performed on two test sets, a similar test set and a new conditions test set, the latter containing images diverse from training set images.

We show that (i) instance-aware segmentation architectures and Cellpose overall outperform the U-Net architectures and the conventional methods in terms of object-level scores (REC, PREC, F1), while the U-Net architectures achieve highest overall mDICE scores on TP objects. (ii) The conventional method ARG achieves results comparable to U-Net and U-Net ResNet results on low- and medium complex images (PREC, F1). On high complex images, both conventional methods cannot compete with deep learning architectures, with respect to all metrics. (iii) Deep learning architectures better adapt to domain shifts then the conventional methods, as evaluated on the new conditions test set. (iv) Augmenting the training set with artificially generated complex images leads to increased REC and F1 scores for the Mask R-CNN and the U-Net ResNet architectures, thereby resulting in improved segmentation effectiveness. (v) The silver-standard annotated training set, generated by trained under-graduates, in combination with instance-aware segmentation architectures (Mask R-CNN, KG instance segmentation) is sufficient to generate accurate segmentation results. (vi) Combined U-Net based segmentation and RWF post-processing keeps US stable for medium- and high complex images, but does not effectively improve the segmentation results.

When analyzing all methods investigated with respect to the training set applied (*silver-* vs. *gold-standard*), it can be observed that all deep learning architectures benefit from *gold-standard* annotations in terms of REC and PREC scores. Although this was expected, our results demonstrate that the combined use of instance-aware segmentation architectures such as Mask R-CNN or KG instance segmentation and *silver-standard* annotations are sufficient to outperform both U-Net architectures and the conventional methods in terms of PREC and REC. This has major implications for data annotation and thus, for broad use of microscopy image analysis: *silver-standard* data annotation can be performed by trained under-graduate students while *gold-standard* annotations have to be carefully generated and curated by biology and pathology experts in a time consuming manner. As resources are limited in general, *silver-standard* annotations enable a more effective adaption to similar tasks in quantitative microscopy and digital pathology.

When investigating all architectures with respect to image complexity, our results demonstrate that the conventional method ARG can achieve results comparable to U-Net predictions on low- and medium- complex images, but cannot compete with deep learning architectures on complex images. The limited ability to generalize to unseen conditions is based on the low number of parameters used in comparison to deep learning architectures. The latter demonstrated to generalize to previously unseen imaging conditions in terms of signal-to-noise ratio, sample quality, nuclear morphologies, out-of-focus objects and diverging modalities (see [11] for a detailed description of images comprised in the new conditions test set).

Overall, evaluation scores decrease with rising image complexity as expected, but the use of artificially generated images can counteract this tendency. This is not surprising as these images were specifically designed to imitate nuclear aggregations and overlaps. By presenting the images to the network while training, an increased set of complex images is processed as compared to training with natural images only. This especially applies to images of the similar test set where adding artificial images to the training set results in three top F1 scores out of five sample preparation types. This is of high relevance as the ability to separate nuclear instances is crucial especially in dense tissue sections with a high diversity of different cell types. The U-Net ResNet and Mask R-CNN architectures benefit most from artificial images, while the shallow U-Net architecture cannot. As this architecture uses the lowest number of parameters, we speculate that the artificially generated images present a slight domain shift that cannot be compensated with this comparably low number of parameters.

When comparing the overall performance of all deep learning architectures investigated, the instance-aware architectures outperform the U-Net architectures in terms of F1 scores, while the mDICE and mJI scores are overall higher for the U-Net architectures. This is most likely due to the skip-connections used within the U-Net architectures, connecting features from the down-sampling path with the up-sampling path, thereby preserving spatial information and leading to more accurate boundary predictions. An exception is presented by the Cellpose architecture, a U-Net architecture predicting a gradient flow representation of images, achieving the second best overall F1 score, based on the best PREC score. Interestingly, HaCaT cells, grown on microscopoy slides, can be segmented with the highest REC and PREC scores, although these nuclei are more strongly aggregated then nuclei of e.g. HaCaT cells cytospinned to glass slides. We assume that this can be explained by the fact that the former contain nuclei with high texture details, which is known to be of benefit in ImageNet pre-trained CNNs [50].

When comparing all architectures to a baseline set by human experts on randomly selected single-nuclei with respect to 10 test set classes, none of the segmentation methods can achieve overall human expert level. Though, methods can compete with human experts on the NB-II and the NC-II test set (U-Net, U-Net ResNet, KG instance segmentation, Attributed Relational Graphs), containing images of low- and medium complexity level. In addition, the KG instance segmentation, the U-Net and the U-Net ResNet architecture can achieve expert level for the GNB-I or the NB-II classes, if trained with natural and artificial data. The GNB-I class contains high complex images, while the NB-II class consists of medium complex images.

As the application of post-processing to plain deep neural network predictions is frequently used in cell and nuclei segmentation approaches, we decided to investigate RWF post-processing to improve U-Net predictions results, as they achieve the highest mDICE scores on TP objects. Overall, the F1 score is not remarkably improved when RWF post-processing is applied. This can be explained by the fact that RWF post-processing is intended to split aggregated nuclei tackling under-segmentation problems resulting in a lower number of clumps but possibly more over-segmented nuclei.

The use of more sophisticated post-processing strategies such as the application of a region proposal network to predict marker-points for a seeded watershed segmentation [48] might be more promising, but come with the costs of having to modify and tune commonly available and partially pre-trained deep learning architectures.

Based on our comprehensive quantitative evaluation, we recommend *silver-standard* data annotation and the use of artificially generated images to augment small training datasets in combination with instance-aware segmentation algorithms, such as Mask R-CNN, for effective segmentation of complex fluorescence nuclear images. In terms of generalization to previously unseen imaging conditions, the use of Cellpose is suggested. Future approaches to further improve segmentation of complex nuclear images shall combine both, high pixel-level accuracy, currently obtained by the U-Net ResNet architecture, with effective instance-aware segmentation. Given the recent revival and importance of single cell analysis in biomedical research, this work has the potential to improve the accuracy and enable broad applications of microscopy based image analysis workflows on complex images of samples such as tissue sections.

REFERENCES

[1] K. Huang and R. F. Murphy, "From quantitative microscopy to automated image understanding," *J. Biomed. Opt.*, vol. 9, no. 5, pp. 893–912, 2004.

[2] I. M. Ambros *et al.*, "Morphologic features of neuroblastoma (Schwannian stroma-poor tumors) in clinically favorable and unfavorable groups," *Cancer*, vol. 94, no. 5, pp. 1574–1583, Mar. 2002.

[3] M. N. Gurcan, L. E. Boucheron, A. Can, A. Madabhushi, N. M. Rajpoot, and B. Yener, "Histopathological image analysis: A review," *IEEE Rev. Biomed. Eng.*, vol. 2, pp. 147–171, 2009.

[4] O. Ronneberger *et al.*, "Spatial quantitative analysis of fluorescently labeled nuclear structures: Problems, methods, pitfalls," *Chromosome Res.*, vol. 16, no. 3, pp. 523–562, May 2008.

[5] A. A. Hill, P. LaPan, Y. Li, and S. Haney, "Impact of image segmentation on high-content screening data quality for SK-BR-3 cells," *BMC Bioinf.*, vol. 8, no. 1, pp. 1–13, Dec. 2007.

[6] J. C. Caicedo *et al.*, "Evaluation of deep learning strategies for nucleus segmentation in fluorescence images," *Cytometry A*, vol. 95, no. 9, pp. 952–965, Sep. 2019.

[7] P. Naylor, M. Lae, F. Reyal, and T. Walter, "Segmentation of nuclei in histopathology images by deep regression of the distance map," *IEEE Trans. Med. Imag.*, vol. 38, no. 2, pp. 448–459, Feb. 2019.

[8] Z. Zeng, W. Xie, Y. Zhang, and Y. Lu, "RIC-UNet: An improved neural network based on UNet for nuclei segmentation in histology images," *IEEE Access*, vol. 7, pp. 21420–21428, 2019.

[9] A. Mahbod, G. Schaefer, I. Ellinger, R. Ecker, O. Smedby, and C. Wang, "A two-stage U-Net algorithm for segmentation of nuclei in H&E-stained tissues," in *Proc. Eur. Congr. Digit. Pathol.* Cham, Switzerland: Springer, 2019, pp. 75–82, doi: 10.1007/978-3-030-23937-4_2.

[10] J. C. Caicedo *et al.*, "Nucleus segmentation across imaging experiments: The 2018 data science bowl," *Nature Methods*, vol. 16, no. 12, pp. 1247–1253, Dec. 2019.

[11] F. Kromp *et al.*, "An annotated fluorescence image dataset for training nuclear segmentation methods," *BioStudies Database*, 2020. [Online]. Available: https://identifiers.org/biostudies:S-BSST265

[12] F. Kromp *et al.*, "An annotated fluorescence image dataset for training nuclear segmentation methods," *Sci. Data*, vol. 7, no. 1, pp. 1–8, Dec. 2020.

[13] L. Hou, A. Agarwal, D. Samaras, T. M. Kurc, R. R. Gupta, and J. H. Saltz, "Unsupervised histopathology image synthesis," 2017, *arXiv:1712.05021*. [Online]. Available: http://arxiv.org/abs/1712.05021

[14] K. W. Dunn *et al.*, "DeepSynth: Three-dimensional nuclear segmentation of biological images using neural networks trained with synthetic data," *Sci. Rep.*, vol. 9, no. 1, pp. 1–15, Dec. 2019.

[15] F. Xing and L. Yang, "Robust nucleus/cell detection and segmentation in digital pathology and microscopy images: A comprehensive review," *IEEE Rev. Biomed. Eng.*, vol. 9, pp. 234–263, 2016.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NIPS*, vol. 1, 2012, pp. 1097–1105.

[17] W. Zhang *et al.*, "Deep convolutional neural networks for multi-modality isointense infant brain image segmentation," *NeuroImage*, vol. 108, pp. 214–224, Mar. 2015.

[18] Y. Cui, G. Zhang, Z. Liu, Z. Xiong, and J. Hu, "A deep learning algorithm for one-step contour aware nuclei segmentation of histopathology images," *Med. Biol. Eng. Comput.*, vol. 57, no. 9, pp. 2027–2043, Sep. 2019.

[19] F. Mahmood *et al.*, "Deep adversarial training for multi-organ nuclei segmentation in histopathology images," *IEEE Trans. Med. Imag.*, vol. 39, no. 11, pp. 3257–3267, Nov. 2020.

[20] M. Z. Alom, C. Yakopcic, T. M. Taha, and V. K. Asari, "Nuclei segmentation with recurrent residual convolutional neural networks based U-Net (R2U-Net)," in *Proc. NAECON IEEE Nat. Aerosp. Electron. Conf.*, Jul. 2018, pp. 228–233.

[21] S. Graham, D. Epstein, and N. Rajpoot, "Dense steerable filter CNNs for exploiting rotational symmetry in histology images," 2020, *arXiv:2004.03037*. [Online]. Available: http://arxiv.org/abs/2004.03037

[22] D. A. Van Valen *et al.*, "Deep learning automates the quantitative analysis of individual cells in live-cell imaging experiments," *PLoS Comput. Biol.*, vol. 12, no. 11, pp. 1–24, 2016.

[23] C. Fu, D. J. Ho, S. Han, P. Salama, K. W. Dunn, and E. J. Delp, "Nuclei segmentation of fluorescence microscopy images using convolutional neural networks," in *Proc. IEEE 14th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2017, pp. 834–842.

[24] A. Kendall, V. Badrinarayanan, and R. Cipolla, "Bayesian SegNet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding," in *Proc. Brit. Mach. Vis. Conf.*, 2017, pp. 1–12.

[25] S. K. Sadanandan, P. Ranefall, S. Le Guyader, and C. Wählby, "Automated training of deep convolutional neural networks for cell segmentation," *Sci. Rep.*, vol. 7, no. 1, pp. 1–17, Dec. 2017.

[26] R. Hollandi *et al.*, "nucleAIzer: A parameter-free deep learning framework for nucleus segmentation using image style transfer," *Cell Syst.*, vol. 10, no. 5, pp. 453–458, 2020.

[27] T. Falk *et al.*, "U-Net: Deep learning for cell counting, detection, and morphometry," *Nature Methods*, vol. 16, no. 1, pp. 67–70, Jan. 2019.

[28] C. Stringer, T. Wang, M. Michaelos, and M. Pachitariu, "Cellpose: A generalist algorithm for cellular segmentation," *Nature Methods*, vol. 18, no. 1, pp. 100–106, Jan. 2021.

[29] J. Shijie, W. Ping, J. Peiyi, and H. Siping, "Research on data augmentation for image classification based on convolution neural networks," in *Proc. Chin. Automat. Congr.* London, U.K.: Springer, 2017, pp. 4165–4170.

[30] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*, 2015, pp. 234–241.

[31] N. Moshkov, B. Mathe, A. Kertesz-Farkas, R. Hollandi, and P. Horvath, "Test-time augmentation for deep learning-based cell segmentation on microscopy images," *Sci. Rep.*, vol. 10, no. 1, pp. 1–7, Dec. 2020.

[32] R. A. Russell, N. M. Adams, D. A. Stephens, E. Batty, K. Jensen, and P. S. Freemont, "Segmentation of fluorescence microscopy images for quantitative analysis of cell nuclear architecture," *Biophys. J.*, vol. 96, no. 8, pp. 3379–3389, Apr. 2009.

[33] O. Bailo, D. Ham, and Y. M. Shin, "Red blood cell image generation for data augmentation using conditional generative adversarial networks," 2019, *arXiv:1901.06219*. [Online]. Available: http://arxiv.org/abs/1901.06219

[34] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[35] J. Yi *et al.*, "Multi-scale cell instance segmentation with keypoint graph based bounding boxes," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*, 2019, pp. 369–377.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[37] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–14.

[38] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[39] C. F. Koyuncu, E. Akhan, T. Ersahin, R. Cetin-Atalay, and C. Gunduz-Demir, "Iterative h-minima-based marker-controlled watershed for cell nucleus segmentation," *Cytometry A*, vol. 89, no. 4, pp. 338–349, Apr. 2016. [Online]. Available: https://github.com/canfkoyuncu/IterativeHMinima

[40] S. Arslan, T. Ersahin, R. Cetin-Atalay, and C. Gunduz-Demir, "Attributed relational graphs for cell nucleus segmentation in fluorescence microscopy images," *IEEE Trans. Med. Imag.*, vol. 32, no. 6, pp. 1121–1131, Jun. 2013.

[41] C. Spearman, "The proof and measurement of association between two things," *Amer. J. Psychol.*, vol. 100, nos. 3–4, pp. 441–471, 1987.

[42] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," 2016, *arXiv:1611.07004*. [Online]. Available: https://arxiv.org/abs/1611.07004

[43] Y. Xie and D. Richmond, "Pre-training on grayscale ImageNet improves medical image classification," in *Proc. Eur. Conf. Comput. Vis. Workshops*, Sep. 2019, pp. 476–484.

[44] M. Dorfer and J. Mattes, "Recursive water flow: A shape decomposition approach for cell clump splitting," in *Proc. IEEE 13th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2016, pp. 811–815.

[45] N. Kumar, R. Verma, S. Sharma, S. Bhargava, A. Vahadane, and A. Sethi, "A dataset and a technique for generalized nuclear segmentation for computational pathology," *IEEE Trans. Med. Imag.*, vol. 36, no. 7, pp. 1550–1560, Jul. 2017.

[46] N. A. Koohbanani, M. Jahanifar, A. Gooya, and N. Rajpoot, "Nuclear instance segmentation using a proposal-free spatially aware deep learning framework," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*, 2019, pp. 622–630.

[47] M. Abdolhoseini, M. G. Kluge, F. R. Walker, and S. J. Johnson, "Segmentation of heavily clustered nuclei from histopathological images," *Sci. Rep.*, vol. 9, no. 1, pp. 1–13, Dec. 2019.

[48] L. Yang *et al.*, "NuSeT: A deep learning tool for reliably separating and analyzing crowded cells," *PLOS Comput. Biol.*, vol. 16, no. 9, Sep. 2020, Art. no. e1008193.

[49] Y. Al-Kofahi, A. Zaltsman, R. Graves, W. Marshall, and M. Rusu, "A deep learning-based algorithm for 2-D cell segmentation in microscopy images," *BMC Bioinf.*, vol. 19, no. 1, pp. 1–11, Dec. 2018.

[50] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness," 2018, *arXiv:1811.12231*. [Online]. Available: http://arxiv.org/abs/1811.12231