Received XX Month, XXXX; revised XX Month, XXXX; accepted XX Month, XXXX; Date of publication XX Month, XXXX; date of current version XX Month, XXXX.

Digital Object Identifier 10.1109/XXXX.2022.1234567

QANet: A Neural Framework for Quality Assessment of Instance Segmentation in Microscopy Imaging

Amit Aharoni[2], Assaf Arbelle[1], Michael Sidorov[1], Eliav Elul[1] and Tammy Riklin Raviv [1], Senior Member, IEEE

[1]The School of Electrical Engineering, Ben-Gurion University of the Negev
[2]Department of Computer Science, Ben-Gurion University of the Negev

Corresponding author: T. Riklin Raviv (email: rrtammy@ee.bgu.ac.il).

This study was partially supported by the Israel Ministry of Science, Technology and Space (MOST 3-14344 T.R.R.) and The United States - Israel Binational Science Foundation (BSF 2019/135 T.R.R.)

ABSTRACT Reliable instance segmentation of cells in microscopy images is essential for quantitative imaging analysis, yet estimating segmentation quality on unseen data, though fundamental, remains largely overlooked. In practice, performance assessment typically requires manual ground-truth annotations, which are costly and impractical at scale. We address this evaluation gap by introducing QANet, a neural framework for post-hoc quality assessment of instance segmentations. Unlike segmentation models, QANet does not predict masks; instead, it receives an image and a segmentation produced by any method and estimates a quantitative quality score that approximates standard evaluation metrics, without requiring ground-truth annotations at inference time.

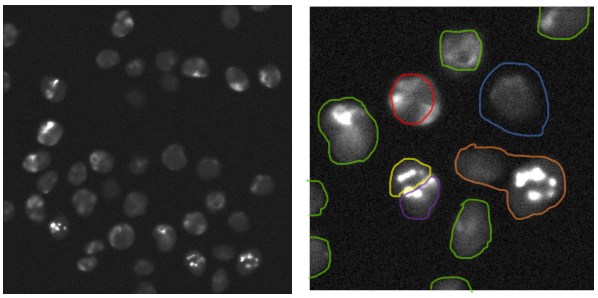
QANet is model-agnostic and formulated as a regression task over image–mask pairs. It is built on the RibCage architecture, which performs multi-scale comparison between image content and segmentation structure, enabling sensitivity to both global shape consistency and fine boundary errors. Training is performed using synthetically perturbed segmentations with known quality scores, enabling supervision across controlled error modes while eliminating the need for large annotated failure datasets. We evaluate QANet on 2D and 3D datasets from the Cell Segmentation Benchmark and demonstrate accurate prediction of both overlap-based measures and boundary-sensitive metrics across multiple segmentation methods. The code is available at https://github.com/amita1996/QANet.

INDEX TERMS Instance segmentation, Microscopy imaging, Quality assessment, Segmentation reliability, Performance prediction

I. INTRODUCTION

IMAGE segmentation is widely used in microscopy image analysis, yet assessing the reliability of produced segmentations on unseen data remains a fundamental limitation of current segmentation pipelines. In practice, segmentation methods are validated on annotated benchmarks, but their performance on user-specific or private datasets is often unknown. Without reliable evaluation, segmentation errors may go unnoticed and compromise downstream analysis.

This challenge is particularly critical in live-cell microscopy, where instance segmentation of individual cells underlies quantitative biological measurements. Even when average benchmark performance is high, specific segmentation failures can invalidate downstream conclusions. Unlike semantic segmentation, which assigns class labels to pixels, instance segmentation requires identifying, delineating, and separating individual objects. This makes the task inherently more sensitive to boundary errors, object proximity, instance topology, and object similarity, and consequently makes reliable evaluation substantially more challenging (Fig. 1). These factors raise a central practical question: how can instance segmentation quality be estimated on user-specific unannotated data?



(a) Raw cell image (b) Erroneous segmentation

FIGURE 1. (a) A raw image of living cells from the Cell Segmentation Benchmark[1]. (b) Zoom-in of the image on the left where the contours were produced by an arbitrary segmentation method. The segmentation errors are color-coded to depict different types of possible errors: Green – accurate segmentation; Red – under-detection; Blue – over-detection; Orange – merged instances; Yellow/Purple – object split.

We address this problem by introducing QANet (Quality Assurance Network), a neural framework for post-hoc estimation of segmentation quality from an image and a predicted mask. Rather than producing segmentations, QANet predicts predefined quality scores by learning a regression mapping over image–mask pairs.

The QANet architecture is inspired by the human ability to evaluate segmentation quality from visual inspection of the raw image. Experts often rely on the alignment of image features, such as edges, with segmentation boundaries. To replicate this behavior, QANet employs the RibCage Network architecture [1], comprising two "ribs" that process the image and the segmentation proposal, and a "spine" that fuses features across multiple levels. This structure enables spatially sensitive, hierarchical comparison between the two inputs. Instance separation is facilitated by encoding segmentations as a trinary function indicating foreground, background, and instance boundaries.

Training is performed using synthetically perturbed segmentation masks derived from ground truth (GT). Controlled non-rigid transformations and morphological operations (erosion, dilation, opening, and closing) simulate diverse segmentation errors while preserving the original image content. This setup enables supervised learning of quality prediction across a range of failure modes.

We evaluate QANet on high-throughput live-cell fluorescence microscopy images from the Cell Segmentation Benchmark (CSB)[1] [2], [3], using segmentations submitted by public challenge participants. Although QANet is trained solely on synthetically perturbed data, its mean predictions closely agree with the official aggregate CSB scores for evaluated segmentation outputs produced by diverse algorithms. For the IoU-based metric, QANet predicts segmentation quality with a maximum relative error of 3.5%. Additional boundary-based metrics, including the Modified Hausdorff Distance (95th percentile), Average Surface Distance (ASD), and Negative Exponential Hausdorff Distance, were evaluated on synthetic data.

The contribution of this work is a post-hoc framework for automated quality assessment of instance segmentation without GT annotations at inference time. Given only an image and its segmentation, QANet predicts a patch-level aggregate instance-segmentation quality score, computed from instance-level contributions rather than semantic foreground/background agreement. It is agnostic to the segmentation algorithm and applies to completed masks from arbitrary sources.

The remainder of the paper is organized as follows: Section II reviews related work. Section III details the QA formulation, RibCage architecture, and training strategy. Section IV presents results and ablation studies. Section V concludes.

II. Related Work

Efforts to improve segmentation reliability have long focused on uncertainty estimation. Classical approaches, such as Markov Chain Monte Carlo (MCMC) sampling [4] and model-dependent uncertainty frameworks [5], generate multiple segmentation hypotheses or confidence maps to characterize prediction variability. In deep learning, Monte Carlo dropout [6] became a standard uncertainty-estimation technique, and uncertainty measures have been shown to correlate with segmentation quality metrics such as Dice [7]. Jungo et al. [8] further analyzed uncertainty estimates for segmentation-error assessment, demonstrating their utility while emphasizing their dependence on the underlying predictive model.

More recently, conformal prediction has been explored for calibrated segmentation quality control. Wundram et al. [9] proposed conformal performance-range prediction for estimating bounds on expected segmentation metrics, while Mossina et al. [10] and Davenport [11] developed conformal confidence sets for semantic and biomedical image segmentation. Despite their theoretical guarantees, uncertainty- and calibration-based approaches typically require calibration data, model outputs, or access to the underlying predictive model.

Other methods assess segmentation quality using model-based or data-assumption-driven criteria. Probabilistic generative models [12], [13] evaluate segmentations under assumptions such as intensity homogeneity and boundary smoothness [14]. While effective in constrained settings, such assumptions may limit applicability across imaging conditions or complex instance interactions.

Reverse Classification Accuracy (RCA) [15], [16] estimates segmentation quality by registering test cases to annotated reference images, and has been used for automated quality control in cardiovascular MR segmentation. However, RCA-based methods require annotated references and assume structural similarity between cases, which may be

[1] http://celltrackingchallenge.net/latest-csb-results/

challenged in dense instance segmentation with high object variability, crowding, and topology changes.

Learning-based real-time quality prediction has also been proposed for cardiovascular MR segmentation [17], but does not directly address dense microscopy instance segmentation, where object separation, boundary interactions, and topology errors are central.

Data-centric work has studied model-agnostic label-quality scoring for detecting annotation errors in segmentation datasets [18], [19]. This setting is complementary to ours: QANet addresses post-hoc quality assessment of predicted instance segmentations, rather than label-error detection in annotated datasets.

Instance segmentation in microscopy is difficult to evaluate due to object similarity, crowding, boundary interactions, and morphological variability. Benchmarks such as the Cell Segmentation Benchmark (CSB) [2], [3] provide standardized evaluation protocols, but their metrics are typically averaged over sequences and do not address the practical need to estimate quality on new, unannotated data.

QANet addresses this setting as learned post-hoc metric prediction from an image–mask pair. It requires the evaluated segmentation mask, but not the segmentation model, logits, uncertainty maps, ensembles, dropout samples, calibration sets, or other model-internal quantities. Training uses synthetically perturbed ground-truth masks rather than outputs of a particular upstream segmenter, making QANet model-agnostic with respect to the segmentation algorithm.

III. Methods

A. Formulation

Our objective is to evaluate the quality of a given instance segmentation. Let $I : \Omega \to \mathbb{R}$ be an image, where Ω is a 2D or 3D image domain. An instance segmentation of an image containing N objects consists of $N + 1$ non-overlapping regions, each assigned a unique label. As in [1], we convert this $N + 1$-label instance map into a trinary segmentation map $\Gamma : \Omega \to L$, with $L = \{0, 1, 2\}$ denoting background, foreground interior, and instance boundaries, respectively. Each connected foreground component represents a single instance, and the boundary class preserves the separation of nearby instances.

Let Γ_{GT} denote the GT segmentation and Γ_{E} the evaluated segmentation of I, both in trinary form. The evaluation criterion can be any predefined segmentation quality metric, either overlap-based, such as IoU [20], or boundary-based, such as the Hausdorff distance [21]. We denote the resulting quality score for a GT–evaluated segmentation pair by $Q(\Gamma_{\mathrm{GT}}, \Gamma_{\mathrm{E}})$. Thus, throughout the paper, "quality metric" refers to the evaluation function Q, "quality score" to its value for a given segmentation pair, and "predicted quality estimate" to QANet's output.

Given only the raw image I and the evaluated segmentation Γ_{E}, QANet predicts $\widehat{Q}_\theta(I, \Gamma_{\mathrm{E}})$, where θ are the network parameters. This scalar is a patch-level aggregate instance-segmentation quality estimate, computed from instance-level contributions within the image patch. A flowchart of the training and inference phases of the proposed QANet framework is presented in Fig. 2.

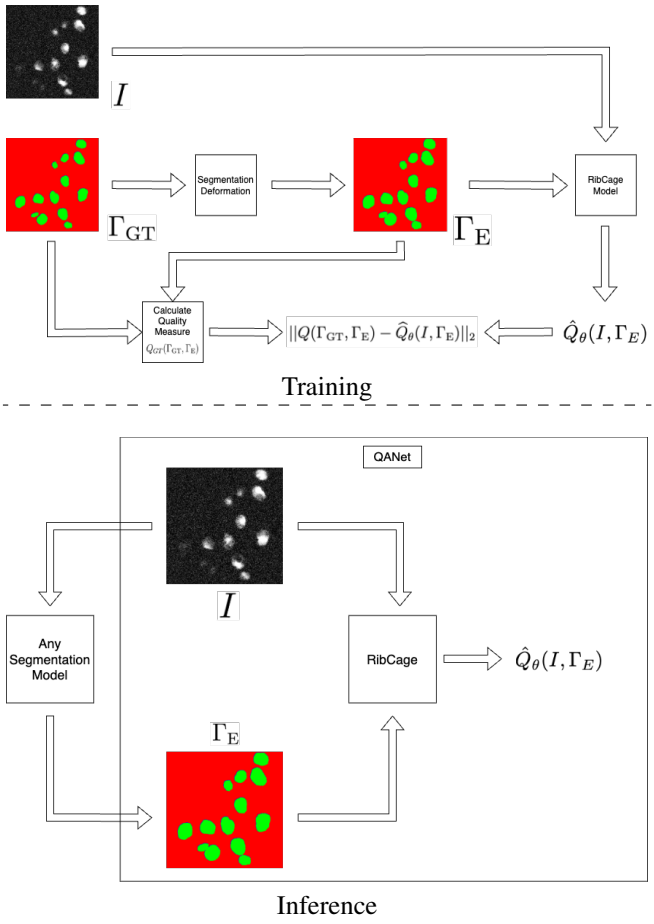


FIGURE 2. QANet flowchart. *Training*—Given GT segmentation masks $\Gamma_{\mathbf{GT}}$ of simulated images I, distorted masks $\Gamma_{\mathbf{E}}$ are generated and the corresponding quality scores $Q_{\mathbf{GT}}(\Gamma_{\mathbf{GT}}, \Gamma_{\mathbf{E}})$ are computed. The network input is the pair $(I, \Gamma_{\mathbf{E}})$, and the loss is $\|\widehat{Q}_\theta(I, \Gamma_{\mathbf{E}}) - Q_{\mathbf{GT}}(\Gamma_{\mathbf{GT}}, \Gamma_{\mathbf{E}})\|_2$. *Inference*—The network input remains $(I, \Gamma_{\mathbf{E}})$, where I may be simulated or real, and $\Gamma_{\mathbf{E}}$ is the segmentation to be evaluated, produced by any segmentation method. QANet evaluates segmentation quality without being tied to the segmentation process itself.

B. RibCage Network Architecture and Loss

QANet is implemented using the RibCage architecture [1], which compares two matched inputs in a spatially sensitive, multi-scale manner. As shown in Fig. 3, the two inputs are the raw image or volume I and the corresponding evaluated trinary segmentation Γ_{E}. The 2D and 3D versions follow the same architectural design, using 2D and 3D convolutional blocks, respectively. The network output is denoted by $\widehat{Q}_\theta(I, \Gamma_{\mathrm{E}})$, where θ are the model parameters.

Let $l \in [1, L]$ denote the index of a RibCage block, where L is the total number of blocks. Let θ^l_S, θ^l_{IM} and θ^l_{SEG} denote the l-th block parameters of the spine, image rib and segmentation rib, respectively. The respective l-th

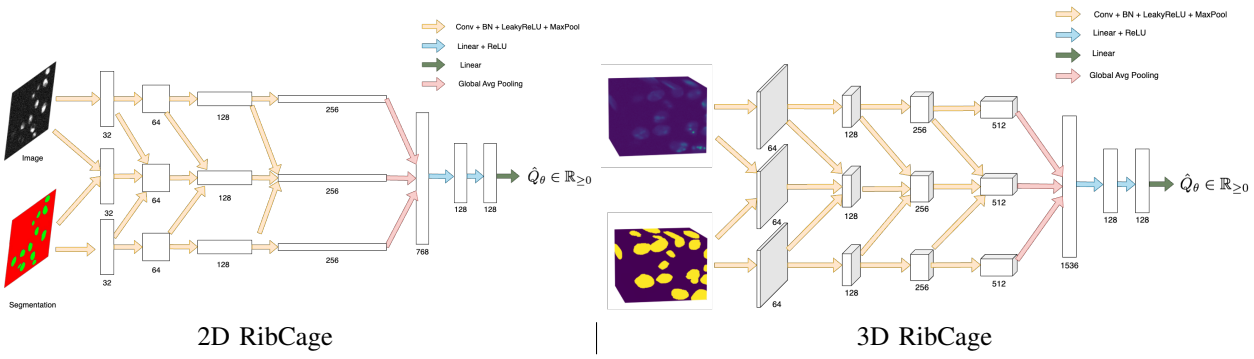


FIGURE 3. **QANet 2D and 3D architectures. QANet uses the RibCage architecture to compare matched image-segmentation pairs. In both 2D and 3D, the top rib receives the raw image/volume I, the bottom rib receives the evaluated trinary segmentation $\Gamma_{\mathbf{E}}$, and the spine fuses their features. The same design is used in 2D and 3D, with 2D/3D convolutional blocks, respectively. Four RibCage blocks are followed by three fully connected layers that output the predicted quality score $\widehat{Q}_\theta(I, \Gamma_{\mathbf{E}})$.**

block outputs, denoted by S^l, R^l_{IM} and R^l_{SEG} are calculated as follows:

$$R^l_{\text{IM}} = f(R^{l-1}_{\text{IM}} * \theta^{l-1}_{\text{IM}}) \tag{1}$$

$$R^l_{\text{SEG}} = f(R^{l-1}_{\text{SEG}} * \theta^{l-1}_{\text{SEG}}) \tag{2}$$

$$S^l = f\left(S^{l-1} * \theta^{l-1}_{S-\text{SP}} \oplus R^{l-1}_{\text{IM}} * \theta^{l-1}_{S-\text{IM}} \oplus R^{l-1}_{\text{SEG}} * \theta^{l-1}_{S-\text{SEG}}\right) \tag{3}$$

where $\theta^{l-1}_S = \{\theta^{l-1}_{S-\text{SP}}, \theta^{l-1}_{S-\text{IM}}, \theta^{l-1}_{S-\text{SEG}}\}$, the symbol $*$ denotes the convolution operation, the symbol $\oplus$ denotes concatenation, and the function $f(\cdot)$ represents the Batch Normalization and the ReLU activation. The initial inputs R^0_{IM} and R^0_{SEG} are the image I and the segmentation Γ_{E}, respectively. The spine input S^0 is set to zero. The outputs of the last block L are then passed to three fully connected (FC) layers resulting in a single scalar $\widehat{Q}_\theta$. QANet is trained to regress a predefined scalar quality measure Q, corresponding to the selected overlap- or boundary-based metric. The loss, denoted by $\mathcal{L}$, is the Mean Squared Error (MSE) between the output of the network and the true measure:

$$\mathcal{L} = ||\widehat{Q}_\theta(I, \Gamma_{\text{E}}) - Q(\Gamma_{\text{GT}}, \Gamma_{\text{E}})||_2 \tag{4}$$

C. Synthesized Segmentations

The QANet input, during both training and inference, consists of image–segmentation pairs. For training, imperfect segmentations are synthesized by deforming GT masks to enable computation of the true quality measure Q. These deformations are designed to mimic realistic segmentation errors, as if produced by an algorithm or untrained annotator, and are generated in two sequential stages: morphological operations (MO) followed by non-rigid perturbations.

Morphological Operations. The morphological operations—erosion, dilation, opening, and closing, simulate under- or over-segmentation. A five-state variable $\mathcal{O}$ is randomly sampled to select one of these operations or the identity. If an operation is chosen, a positive integer σ_{MO} is sampled to define the kernel size. Note that after the MO stage, instances may be completely removed or merged with neighboring ones.

Non-Rigid Perturbations. To further enrich the variability of the segmentation errors, we perform non-rigid perturbations to the MO-processed masks. Specifically, we deformed the image domain of the MO-processed segmentation maps by smoothed, randomly sampled vector fields. In contrast to standard data augmentation, here the deformation is applied only to the segmentation mask, while the corresponding image is kept fixed.

The ten perturbed masks shown in Fig. 4 are examples of final erroneous segmentations produced by this complete two-step pipeline. The average IoU per cell instance is shown below each. The reader is referred to Section IV-A for detailed data generation protocol.

D. Simulated Training Images

The MSE loss is computed between predicted and *true* quality scores, requiring GT segmentation masks alongside the synthesized erroneous ones. Simulated images are generated from known masks, ensuring accurate GT for supervision. In contrast, manual annotations may contain human errors that, if systematic, can bias the loss and impact training [23].

E. Segmentation Quality Measures

The SEG Measure. The SEG measure, introduced in [24] for evaluating cell segmentation in microscopy, extends the IoU to multiple instances. Let K and K' be the number of individual cells in the GT and evaluated segmentation, Γ_{GT} and Γ_{E} respectively. Let $c \in \Gamma_{\text{GT}}$ and $c' \in \Gamma_{\text{E}}$ denote connected components in the GT and evaluated segmentations, respectively. The SEG measure is defined as the IoU of the GT and the evaluated objects, unless their overlap is less than 50%. The mean SEG measure over all the GT objects is formulated as follows:

$$Q(\Gamma_{\text{GT}}, \Gamma_{\text{E}}) = \sum_{c \in \Gamma_{\text{GT}}} \sum_{c' \in \Gamma_{\text{E}}} \begin{cases} \text{IoU}(c, c')/K & \alpha(c, c') > 0.5 \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where $\text{IoU}(c, c') = |c \cap c'|/|c \cup c'|$ and $\alpha(c, c') = |c \cap c'|/|c|$. Note that each connected component $c \in \Gamma_{\text{GT}}$ can match at

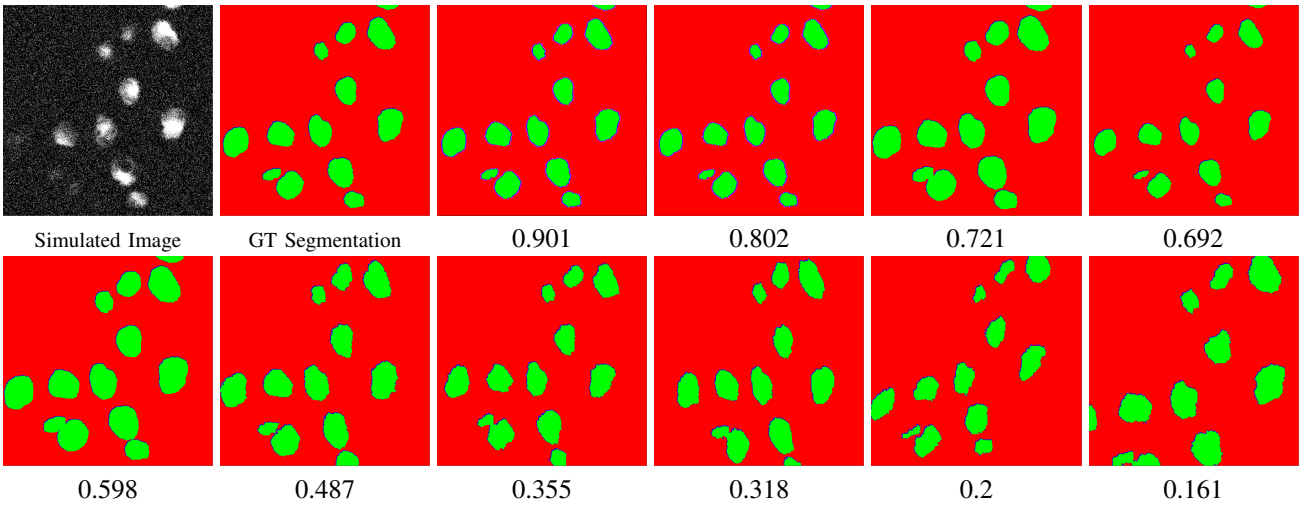


FIGURE 4. Simulated training data. Left to right: a simulated microscopy image using CytoPacq [22], the corresponding GT segmentation, and ten randomly generated erroneous segmentations produced by the complete MO-plus-non-rigid perturbation pipeline described in Section III-C. The average per-instance IoU score is shown below each perturbed segmentation.

most one component $c' \in \Gamma_{\mathrm{E}}$ with over 50% overlap. If no such match exists, the cell is considered undetected and its SEG score is set to zero. The division by K, the number of ground-truth instances in the patch, normalizes the score by the number of objects. Thus, $Q(\Gamma_{GT}, \Gamma_E)$ is the average instance-level SEG contribution over the patch, with values in the range $[0, 1]$.

Boundary-based metrics. To evaluate boundary-sensitive metrics—unlike SEG, which focuses on region overlap—we use variants of the Hausdorff distance. For each matched GT–evaluated instance pair (c, c'), let $X = \{x_i\}_{i=1}^{\chi}$ and $X' = \{x'_j\}_{j=1}^{\chi'}$ denote the corresponding boundary point sets, i.e., boundary pixels in 2D or surface samples in 3D, with cardinalities χ and χ', respectively. The classical Hausdorff distance between X and X' is defined as:

$$D_{\mathrm{H}}(X, X') = \max\left\{ \sup_{\mathbf{x}_i \in X} d(\mathbf{x}_i, X'), \sup_{\mathbf{x}_j \in X'} d(\mathbf{x}_j, X) \right\}, \tag{6}$$

where $d(\mathbf{x}_i, X') = \inf_{\mathbf{x}_j \in X'} d(\mathbf{x}_i, \mathbf{x}_j)$, and similarly $d(\mathbf{x}_j, X) = \inf_{\mathbf{x}_i \in X} d(\mathbf{x}_j, \mathbf{x}_i)$. We use the Euclidean distance for $d(\cdot)$.

Modified Hausdorff Distance (MHD). The Modified Hausdorff Distance (MHD) replaces the suprema in Eq. (6) with summations:

$$D_{\mathrm{MHD}}(X, X') = \max\left\{ \sum_{\mathbf{x}_i \in X} \frac{d(\mathbf{x}_i, X')}{\chi}, \sum_{\mathbf{x}_j \in X'} \frac{d(\mathbf{x}_j, X)}{\chi'} \right\} \tag{7}$$

Modified Hausdorff 95th Percentile. The Modified Hausdorff 95th Percentile Distance (MHD95) is a robust variant of the Hausdorff distance, defined as the 95th percentile of nearest-neighbor boundary distances computed in both directions:

$$D_{\mathrm{H95}}(X, X') = \max\left\{ P_{95}\big(d(X, X')\big), P_{95}\big(d(X', X)\big) \right\}, \tag{8}$$

where $d(X, X') = \{d(x_i, X') : x_i \in X\}$, $d(X', X) = \{d(x'_j, X) : x'_j \in X'\}$, and P_{95} denotes the 95th percentile.

Average Surface Distance (ASD). The Average Surface Distance (ASD), or symmetrized MHD, is a symmetric metric measuring the mean shortest distance between surface points of two sets:

$$D_{\mathrm{ASD}}(X, X') = \frac{\sum_{\mathbf{x}_i \in X} d(\mathbf{x}_i, X') + \sum_{\mathbf{x}_j \in X'} d(\mathbf{x}_j, X)}{\chi + \chi'} \tag{9}$$

Negative Exponential MHD. The HD metric and its MHD variants can take any positive value, where the magnitude depends on the size of the cells and the images. As a normalized alternative, we introduce a novel boundary-based metric, based on the negative exponential function of $D_{\mathrm{MHD}}(X, X')$, defined as follows:

$$Q_{\mathrm{neMH}}(\Gamma_{\mathrm{GT}}, \Gamma_{\mathrm{E}}) = \exp(-D_{\mathrm{MHD}}(X, X')/\tau). \tag{10}$$

Note that when the segmentation mask is optimal and perfectly overlaps the ground-truth segmentation - then $D_{\mathrm{MH}} = 0$ and $Q_{\mathrm{neMH}} = 1$. In general, using the negative exponential function allows us to map the Q_{neMH} values to $(0, 1]$ and the division by the scalar τ stretches them. We set $\tau = 10$.

IV. Experiments

We assess QANet in three complementary settings. First, we evaluate per-example regression accuracy on held-out simulated test data with known ground truth, using several segmentation-quality measures. Second, we use CSB training sequences, where ground-truth segmentations are available, to perform ablation studies of the network architecture and segmentation representation. Third, we apply QANet to CSB test-set image–segmentation pairs produced by public segmentation methods. In this final setting, QANet is trained solely on synthetic CytoPacq data with simulated segmentation perturbations, while the CSB test data are used only for post-hoc evaluation.

A. Synthetic Data.

Images and GT Segmentations. For 2D supervised training data, we used the CytoPacq web service [2] [22] to synthesize 10,000 microscopy images and GT segmentations resembling Fluo-N2DH-SIM+. CytoPacq generates a digital cell phantom, simulates optical image formation, and models detector acquisition effects. Each image was 420×420 pixels and contained 1–60 cells. Fig. 4 shows an example CytoPacq-synthesized image, its GT segmentation, and representative perturbed masks. We used the official CSB Fluo-N3DH-SIM+ dataset to assess the 3D configuration.

Synthesizing Imperfect Segmentations. The five-state morphology variable $\mathcal{O}$ determines the corruption applied to the ground-truth segmentation. If $\mathcal{O}$ is not the identity, we sample $\sigma_{\text{MO}} \sim U(1, 4)$ to determine the MO kernel size. For 2D non-rigid perturbations, we sample a displacement field $[v_x, v_y]$ over the input domain, with $v_x, v_y \sim U(-512, 512)$. To convert the randomly sampled displacement field into a spatially coherent deformation field, avoiding noisy pixel-wise perturbations, we smooth it using a Gaussian kernel with $\sigma_g = 38$. For 3D segmentation maps, we assign random displacements $d \sim U(0, D = 4)$ to a coarse grid of $n = 8$ control points around and inside the image, and interpolate voxel-wise displacements using cubic B-splines, as in [25].

B. CSB Data

CSB Sequences. Examples from CSB datasets used in our experiments are shown in Fig. 5: the simulated Fluo-N2DH-SIM+ dataset in the first row, Fluo-N2DL-HeLa in the second row, and Fluo-N2DH-GOWT1 in the third and fourth rows. For each example, the figure shows the raw image, the corresponding ground-truth instance segmentation, and an automatic segmentation produced by [26]. Fluo-N2DL-HeLa images contain HeLa cells stably expressing H2b-GFP, acquired with an Olympus IX81 microscope at pixel size 0.645×0.645 μm [27]; Fluo-N2DH-GOWT1 images contain GFP-GOWT1 mouse stem cells acquired with a Leica TCS SP5 microscope at pixel size 0.24×0.24 μm [28].

CSB Segmentations. We deployed **seven** instance segmentation methods from the CSB leaderboard, namely, CVUT-CZ [29], BGU-IL(3,4) [30], BGU-IL(5) [26], KTH-SE(1) [31], UNSW-AU [32] and DKFZ-GE [33]. Specifically, we downloaded the codes of these methods from the Cell Segmentation Benchmark website, applied them to the microscopy datasets and used them to generate sets of image-segmentation pairs for testing the QANet.

C. Evaluation for Different Quality Measures

We used the held-out test dataset to evaluate QANet regression accuracy predicting different quality measures for 2D and 3D data. For the 2D experiments, images and GT segmentations were synthesized using CytoPacq as described in Section IV.A. To demonstrate the 3D extension, we used the official CSB Fluo-N3DH-SIM+ dataset. The simulated data were split into training, validation, and held-out synthetic test sets, 70%–15%–15%, respectively. The training set was used to optimize the network parameters, the validation set for model selection and hyperparameter tuning, and the held-out synthetic test set only for the simulated-data evaluation. Fig. 6 shows scatter plots comparing QANet-predicted quality scores with the corresponding ground-truth quality scores on held-out simulated image-segmentation pairs. Each point represents one image patch and its perturbed segmentation. From left to right, the plots show 2D SEG, 3D SEG, MHD95, ASD, and negative exponential MHD regression results, where 3D SEG was evaluated on the Fluo-N3DH-SIM+ data and the other quality measures with CytoPacq-simulated data. The dashed diagonal line indicates perfect agreement between the predicted and true scores. Table 1 reports complementary regression statistics, including Pearson and Spearman correlations, MAE, RMSE, prediction bias, and hit-rate area under curve (AUC) where applicable. Hit rate is defined as the fraction of test examples

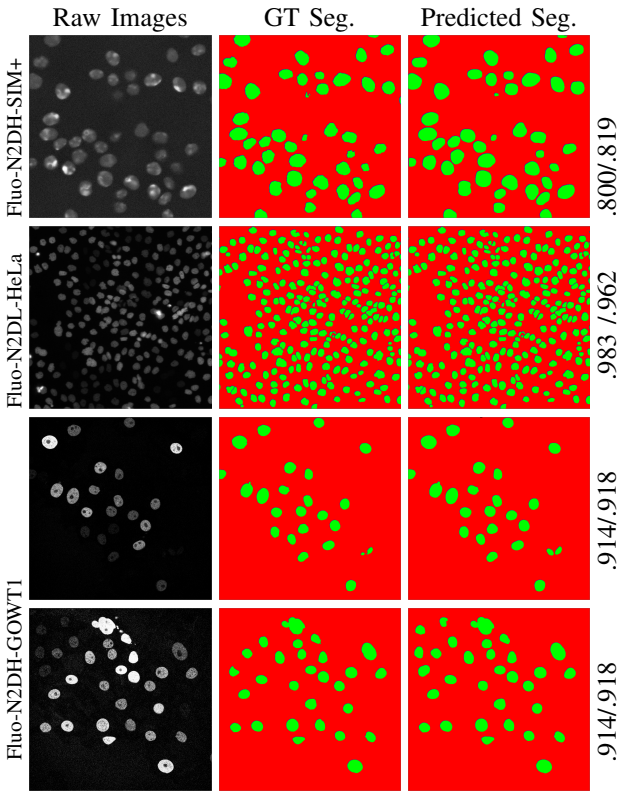


FIGURE 5. Image–segmentation triplets. Representative examples from the Fluo-N2DH-SIM+ dataset (top row), the Fluo-N2DL-HeLa dataset (second row), and the Fluo-N2DH-GOWT1 dataset (third and fourth rows). For each example, the columns show the raw microscopy image, the corresponding ground-truth (GT) segmentation, and the instance segmentation produced by BGU-IL(5) method [26]. The right-most column reports the true average SEG score and the QANet prediction, respectively, in the format GT SEG / predicted SEG. GT segmentations are shown for reference and were used for computing true SEG scores; at inference, QANet receives only the raw image and the predicted segmentation.

[2] https://cbia.fi.muni.cz/simulator

TABLE 1. Regression statistics on held-out simulated test data. Bias is computed as the mean prediction error, $\mathbb{E}[\widehat{Q} - Q]$. AUC denotes hit-rate AUC and is reported only for quality scores normalized to $[0, 1]$. SEG was evaluated in both 2D and 3D settings, whereas the boundary-based metrics were evaluated on the 2D simulated data. Corresponding scatter plots are shown in Fig. 6.

Metric	AUC	Pearson r	Spearman ρ	MAE	RMSE	Bias
2D SEG	0.946	0.901	0.906	0.064	0.085	+0.018
3D SEG	0.934	0.573	0.580	0.066	0.085	+0.009
MHD95	N/A	0.747	0.771	1.252	1.953	−0.196
ASD	N/A	0.765	0.790	1.112	1.573	+0.254
Neg. exp. MHD	0.948	0.784	0.791	0.074	0.098	+0.001

whose absolute prediction error is below a given tolerance; the hit-rate AUC summarizes this curve over tolerances.

D. Ablation on CSB Training Sequences

For the ablation study, we used CSB training sequences for which ground-truth segmentations are available, enabling quantitative evaluation on real microscopy image-segmentation pairs. We examined two aspects of QANet: the network architecture and the segmentation representation, using the CSB SEG measure as the target quality score.

Network architecture. We compared RibCage with Siamese and naive alternatives using comparable numbers of trainable parameters: 2.05M, 2.09M, and 2.08M, respectively. RibCage fuses image and segmentation features throughout the network (Sec. III-B); the Siamese model processes I and Γ_{E} in separate streams and fuses them only before the fully connected layers; and the naive model receives their concatenation as a single input.

We evaluated the three architectures using actual CSB segmentations from CVUT-CZ and KTH-SE(1) on the real Fluo-N2DH-GOWT1 and Fluo-N2DL-HeLa datasets, providing a compact ablation setting with available GT and distinct segmentation outputs. The left panel of Fig. 7 and the first three rows of Table 2 show the hit-rate curves and AUC scores. RibCage outperformed both alternatives across all combinations, indicating higher accuracy at fixed tolerances.

Segmentation representation. We also evaluated the effect of the segmentation representation by comparing binary foreground/background masks with the QANet trinary representation. The trinary representation encodes background, foreground interior, and instance boundaries, preserving object-separation cues that are lost in a binary mask. The corresponding results are reported in Fig. 7 and Table 2.

E. Evaluation on CSB Test Segmentations

Finally, we evaluated QANet as a post-hoc predictor of segmentation quality on completed CSB test-set segmentations. In this experiment, QANet was trained only on synthetic CytoPacq data with simulated segmentation perturbations and then applied to CSB image–segmentation pairs produced by public segmentation methods.

TABLE 2. AUC scores for KTH-SE(1) and CVUT-CZ segmentation predictions on N2DH-GOWT1 and N2DL-HeLa datasets, comparing the RibCage (binary/trinary), Siamese, and Naive architectures. The RibCage with trinary input consistently outperformed the others.

Network Architecture	Segmentation Mode	KTH-SE(1) GOWT1	KTH-SE(1) HeLa	CVUT-CZ GOWT1	CVUT-CZ HeLa
Naive	Trinary	0.890	0.934	0.849	0.914
Siamese	Trinary	0.819	0.745	0.804	0.727
RibCage	Trinary	**0.904**	**0.947**	**0.912**	**0.944**
RibCage	Binary	0.902	0.933	0.908	0.917

We used segmentation outputs from seven CSB methods: CVUT-CZ, BGU-IL(3), BGU-IL(4), BGU-IL(5), KTH-SE(1), UNSW-AU, and DKFZ-GE. These outputs were used only for evaluation/inference, not for training QANet. Each trained QANet model was applied to the corresponding CSB image-segmentation pairs and predicted SEG quality scores.

Unlike the held-out simulated evaluation in Section IV-C, GT segmentations for the official CSB test sets are not publicly available. We therefore compare QANet's predicted mean SEG scores with the official aggregate SEG scores reported by the CSB organizers.

Table 3 summarizes these results across dataset–method pairs and compares QANet with cross-method evaluation, where one segmentation method is used as a surrogate GT for another. This baseline is less accurate than QANet, since no method can be assumed to be consistently superior and similar failure modes may lead different methods to agree on the same errors. The maximum QANet prediction error was 1.01% for Fluo-N2DH-SIM+ and 3.5% for Fluo-N2DH-GOWT1, demonstrating close agreement with the official aggregate scores despite training only on synthetic data.

V. Discussion, Limitations, and Future Work

We presented QANet, a deep learning model for post-hoc assessment of instance segmentation quality from image–mask pairs. In the spirit of Benjamin Disraeli's remark, "it is easier to be critical than to be correct," QANet acts as an independent auditor: it evaluates completed segmentations without producing segmentation masks itself.

Built on the RibCage architecture [1], QANet compares multi-scale image and segmentation features and learns predefined quality measures from synthetically perturbed masks. Experiments on the Cell Segmentation Benchmark show that QANet learns diverse quality metrics, is applicable in both 2D and 3D settings, and closely matches official aggregate SEG scores on CSB test data, with a maximum relative error of 3.5%, outperforming surrogate-GT comparisons.

These results support the intended use of QANet as a patch-level triage tool: QANet ranks segmented microscopy patches by predicted aggregate instance-segmentation quality and flags low-scoring patches for manual inspection. Applying the trained model to one predicted instance at a time, while masking the other predicted instances, could serve only

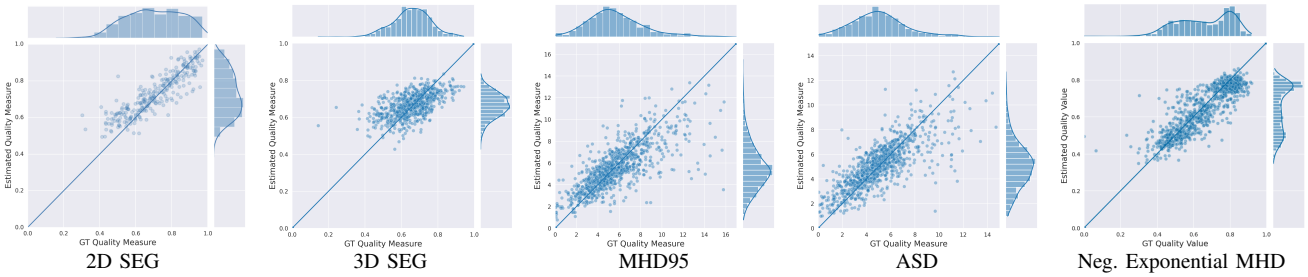


FIGURE 6. **QANet prediction of segmentation-quality measures on simulated test data. Scatter plots of predicted versus target scores where each point represents an image patch. From left to right, the plots show 2D SEG, 3D SEG, MHD95, ASD, and negative exponential MHD regression results. The dashed diagonal line indicates perfect agreement between the predicted and true scores.**

TABLE 3. **Predicted SEG scores for seven CSB methods (BGU-IL(3–5), CVUT-CZ, KTH-SE(1), UNSW-AU, DKFZ-GE) on simulated (Fluo-N2DH-SIM+) and real (Fluo-N2DH-GOWT1) test data. GT instance segmentations were unknown; true aggregated SEG scores were taken from the benchmark site: http://celltrackingchallenge.net/latest-csb-results/. The table shows QANet predictions compared to cross-method evaluations. Absolute relative errors (in brackets) are computed with respect to the official mean SEG scores.**

Dataset	Evaluated Method	CSB SEG	QANet SEG (err%)	Cross-method evaluations					
				Method	SEG (err%)	Method	SEG (err%)	Method	SEG (err%)
Fluo-N2DH-SIM+	CVUT-CZ	.807	**.813** (0.74)	KTH-SE(1)	.769 (4.70)	DKFZ-GE	.779 (3.45)	UNSW-AU	.767(4.91)
Fluo-N2DH-SIM+	KTH-SE(1)	.791	**.799** (1.01)	CVUT-CZ	.772 (2.40)	DKFZ-GE	.745(5.96)	UNSW-AU	.750(5.35)
Fluo-N2DH-SIM+	UNSW-AU	.807	**.809** (0.25)	CVUT-CZ	.838(3.78)	DKFZ-GE	.855(6.00)	KTH-SE(1)	.807(0.05)
Fluo-N2DH-SIM+	DKFZ-GE	.832	**.839** (0.84)	CVUT-CZ	.860(3.36)	KTH-SE(1)	.804(3.34)	UNSW-AU	.913 (9.74)
Fluo-N2DH-GOWT1	BGU-IL(4)	.874	**.887** (1.48)	UNSW-AU	.823(7.21)	BGU-IL(5)	.893(2.23)	KTH-SE(1)	.901(3.11)
Fluo-N2DH-GOWT1	BGU-IL(5)	.920	**.918** (0.21)	UNSW-AU	.883 (4.04)	KTH-SE(1)	.952(3.45)	BGU-IL(4)	.894(2.82)
Fluo-N2DH-GOWT1	UNSW-AU	.933	**.900** (3.50)	KTH-SE(1)	.926(.71)	BGU-IL(5)	.883 (5.36)	BGU-IL(4)	.856(8.26)

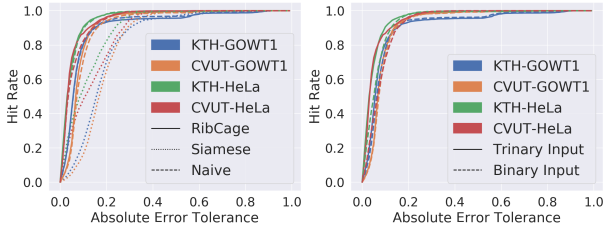


FIGURE 7. **The Hit Rate curves as a function of the SEG prediction tolerance. Left: comparison of the RibCage (solid line), the Siamese (dotted line) and the Naive (dashed line) networks. Right: Results obtained by the RibCage Network with either binary (dashed line) or trinary (solid line) segmentation input. All configurations were tested using two segmentation methods: KTH-SE(1) (blue and green) and CVUT-CZ (orange and red), on two datasets: Fluo-N2DH-GOWT1 (blue and orange) and Fluo-N2DL-HeLa (green and red).**

as a heuristic localization aid, since missed true instances would not be represented in the score.

The synthetic training strategy provides exact ground-truth masks and target quality scores, avoiding biases from imperfect manual annotations or from a specific upstream segmentation model. At the same time, although CytoPacq simulates relevant optical and acquisition effects, synthetic images may not capture the full variability of real microscopy acquisitions. Similarly, the segmentation perturbation model captures common geometric and topological errors, including boundary shifts, missed or removed instances, merges, and over-segmentation/splits, but does not explicitly model all image-dependent failure mechanisms, such as missed dim cells due to weak contrast or texture-driven errors in crowded regions. Future work may extend both the image simulation process and the segmentation perturbation model to better cover diverse acquisition conditions and algorithm-specific errors, as well as high-throughput real 3D microscopy data with complete annotations suitable for quantitative validation.

A related limitation concerns the observability of segmentation errors. In noisy or low-contrast regions, the image evidence may not determine a unique correct boundary, and multiple segmentations may be plausible. QANet should therefore be interpreted as estimating the consistency of a completed segmentation with the learned image statistics and annotation convention, rather than as recovering an absolute unobservable boundary. In ambiguous cases, low predicted quality should serve as a signal for manual inspection rather than as automatic correction or definitive error localization.

By decoupling quality assessment from the segmentation process itself, QANet provides a general tool for post-hoc evaluation. Such capability may support benchmarking on private data, data selection in active learning pipelines, and automated quality control in large-scale image analysis workflows.

REFERENCES

[1] A. Arbelle and T. Riklin Raviv, "Microscopy cell segmentation via adversarial neural networks," *IEEE ISBI*, pp. 645–648, 2018.

[2] V. Ulman, M. Maška, K. Magnusson *et al.*, "An objective comparison of cell-tracking algorithms," *Nature methods*, vol. 14, no. 12, p. 1141, 2017.

[3] M. Maška, V. Ulman, P. Delgado-Rodriguez, E. Gómez-de Mariscal, T. Nečasová, F. A. Guerrero Peña, T. I. Ren, E. M. Meyerowitz, T. Scherr, K. Löffler *et al.*, "The cell tracking challenge: 10 years of objective benchmarking," *Nature Methods*, pp. 1–11, 2023.

[4] A. C. Fan, J. W. Fisher, W. M. Wells, J. J. Levitt, and A. S. Willsky, "MCMC curve sampling for image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2007, pp. 477–485.

[5] T. Hershkovitch and T. Riklin Raviv, "Model-dependent uncertainty estimation of medical image segmentation," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, 2018, pp. 1373–1376.

[6] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*, 2016, pp. 1050–1059.

[7] A. G. Roy, S. Conjeti, N. Navab, C. Wachinger, A. D. N. Initiative *et al.*, "Bayesian QuickNAT: model uncertainty in deep whole-brain segmentation for structure-wise quality control," *NeuroImage*, vol. 195, pp. 11–22, 2019.

[8] A. Jungo, F. Balsiger, and M. Reyes, "Analyzing the quality and challenges of uncertainty estimations for brain tumor segmentation," *Frontiers in Neuroscience*, vol. 14, p. 282, 2020.

[9] A. M. Wundram, P. Fischer, M. Muehlebach, L. M. Koch, and C. F. Baumgartner, "Conformal performance range prediction for segmentation output quality control," *arXiv preprint arXiv:2407.13307*, 2024.

[10] L. Mossina, J. Dalmau, and L. Andéol, "Conformal semantic image segmentation: Post-hoc quantification of predictive uncertainty," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2024.

[11] S. Davenport, "Conformal confidence sets for biomedical image segmentation," *arXiv preprint arXiv:2410.03406*, 2024.

[12] B. Audelan and H. Delingette, "Unsupervised quality control of image segmentation based on bayesian learning," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 21–29.

[13] ——, "Unsupervised quality control of segmentations based on a smoothness and intensity probabilistic model," *Medical Image Analysis*, vol. 68, p. 101895, 2021.

[14] D. Mumford and J. Shah, "Optimal approximations by piecewise smooth functions and associated variational problems," *Communications on Pure and Applied Mathematics*, vol. 42, pp. 577–684, 1989.

[15] V. Valindria, I. Lavdas, W. Bai, K. Kamnitsas, E. Aboagye, A. Rockall, D. Rueckert, and B. Glocker, "Reverse classification accuracy: predicting segmentation performance in the absence of ground truth," *IEEE transactions on medical imaging*, vol. 36, no. 8, pp. 1597–1606, 2017.

[16] R. Robinson, V. Valindria, W. Bai, O. Oktay, B. Kainz, H. Suzuki, M. M. Sanghvi, N. Aung, J. M. Paiva, F. Zemrak *et al.*, "Automated quality control in image segmentation: application to the uk biobank cardiovascular magnetic resonance imaging study," *Journal of Cardiovascular Magnetic Resonance*, vol. 21, no. 1, pp. 1–14, 2019.

[17] R. Robinson, O. Oktay, W. Bai, V. Valindria, M. M. Sanghvi, N. Aung, J. Paiva, F. Zemrak, K. Fung, E. Lukaschuk *et al.*, "Real-time prediction of segmentation quality," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 578–585.

[18] V. Lad and J. Mueller, "Estimating label quality and errors in semantic segmentation data via any model," in *ICML Workshop on Data-centric Machine Learning Research*, 2023.

[19] J. Kuan and J. Mueller, "Model-agnostic label quality scoring to detect real-world label errors," in *ICML DataPerf Workshop*, 2022.

[20] P. Jaccard, "Étude comparative de la distribution florale dans une portion des alpes et des jura," *Bull Soc Vaudoise Sci Nat*, vol. 37, pp. 547–579, 1901.

[21] M.-P. Dubuisson and A. K. Jain, "A modified hausdorff distance for object matching," in *Pattern Recognition, 1994. Vol. 1-Conference A: Computer Vision & Image Processing., Proceedings of the 12th IAPR International Conference on*, vol. 1. IEEE, 1994, pp. 566–568.

[22] D. Wiesner, D. Svoboda, M. Maška, and M. Kozubek, "Cytopacq: a web-interface for simulating multi-dimensional cell imaging," *Bioinformatics*, 2019.

[23] O. Shwartzman, H. Gazit, I. Shelef, and T. Riklin-Raviv, "The worrisome impact of an inter-rater bias on neural network training," in *The 5th International Conference on Medical Imaging and Computer-Aided Diagnosis (MICAD)*, 2024.

[24] M. Maška, V. Ulman, D. Svoboda *et al.*, "A benchmark for comparison of cell tracking algorithms," *Bioinformatics*, vol. 30, no. 11, pp. 1609–1617, 2014.

[25] D. Rueckert, L. Sonoda, C. Hayes, D. Hill, M. Leach, and D. Hawkes, "Nonrigid registration using free-form deformations: application to breast MR images," *IEEE transactions on medical imaging*, vol. 18, no. 8, pp. 712–721, 1999.

[26] A. Arbelle, S. Cohen, and T. Riklin Raviv, "Dual-Task ConvLSTM-UNet for instance segmentation of weakly annotated microscopy videos," *IEEE Transactions on Medical Imaging (TMI)*, vol. 41, no. 8, pp. 1948–1960, 2022.

[27] B. Neumann, T. Walter, J.-K. Hériché, J. Bulkescher, H. Erfle, C. Conrad, P. Rogers, I. Poser, M. Held, U. Liebel *et al.*, "Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes," *Nature*, vol. 464, no. 7289, p. 721, 2010.

[28] E. Bártová, G. Šustáčková, L. Stixová, S. Kozubek, S. Legartová, and V. Foltánková, "Recruitment of oct4 protein to uv-damaged chromatin in embryonic stem cells," *PLoS One*, vol. 6, no. 12, p. e27281, 2011.

[29] T. Sixta, "Probabilistic models for joint segmentation, detection and tracking," Ph.D. dissertation, Czech Technical University in Prague, 2019.

[30] A. Arbelle and T. Riklin Raviv, "Microscopy cell segmentation via convolutional LSTM networks," pp. 1008–1012, 2019.

[31] K. E. G. Magnusson, "Segmentation and tracking of cells and particles in time-lapse microscopy," Ph.D. dissertation, KTH Royal Institute of Technology, 2016.

[32] Y. Zhu and E. Meijering, "Automatic improvement of deep learning-based cell segmentation in time-lapse microscopy by neural architecture search," *Bioinformatics*, vol. 37, no. 24, p. 4844–4850, 2021.

[33] F. Isensee, P. Jaeger, S. Kohl, J. Petersen, and K. Maier-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, pp. 203–211, 2021.