MACHINE LEARNING
Science and Technology

PURPOSE-LED PUBLISHING™ Part of the MACHINE LEARNING Series

PAPER • OPEN ACCESS

Automated detection of potential artifacts in machine learning based bio-image segmentation

To cite this article: Saiyam B Jain *et al* 2025 *Mach. Learn.: Sci. Technol.* **6** 045029

IOP Publishing Mach. Learn.: Sci. Technol. 6 (2025) 045029 https://doi.org/10.1088/2632-2153/ae1546

MACHINE LEARNING
Science and Technology

PAPER

Automated detection of potential artifacts in machine learning based bio-image segmentation

Saiyam B Jain[1], **Zongru Shao**[2] and **Michael Hecht**[3,4,5,*]

[1] Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
[2] Silicon Austria Labs (SAL), Linz, Austria
[3] Center for Advanced Systems Understanding (CASUS), Görlitz, Germany
[4] Helmholtz-Zentrum Dresden-Rossendorf e.V. (HZDR), Dresden, Germany
[5] University Wrocław, Mathematical Institute, Wrocław, Poland
[*] Author to whom any correspondence should be addressed.

E-mail: m.hecht@hzdr.de, saiyam.jain@kit.edu and zongru.shao@silicon-austria.com

Keywords: artifacts, bio-image, segmentation

OPEN ACCESS

RECEIVED
20 February 2025

REVISED
29 August 2025

ACCEPTED FOR PUBLICATION
20 October 2025

PUBLISHED
31 October 2025

Abstract

Image segmentation algorithms, while powerful, are inherently prone to artifacts, making perfect segmentation theoretically and practically impossible. We propose an automated artifact identification scheme for posterior rapid manual re-correction to address this challenge. Hereby, our contribution is twofold: We extend our previous work, delivering polynomial defenses (PDs). These defenses mimic noise distributions that significantly improve segmentation quality when removed from the training images. In practice, we perturb unseen images with PDs and demonstrate that the resulting segmentation differences achieve promising precision in artifact detection compared to traditional Gaussian and Poisson noise perturbations. This automated guidance is our essential contribution. Beyond improving the reliability of image-processing outputs, our approach provides a valuable tool for enhancing manually segmented training datasets. Hereby, the automated guidance massively decreases manual cross-checking time.

1. Introduction

In the field of *image processing*, a non-exhaustive list of machine learning (ML) based tasks includes *image classification, image generation, object detection, pattern recognition, and image segmentation* (Gonzales and Wintz 1987, Chakravorty 2018). The primary challenge of all approaches addressing these tasks is to ensure stability when solving the task-specific underlying *inverse problem*. While this article focuses on *image segmentation*, we briefly outline this subject before discussing stability.

1.1. ML image segmentation

Image segmentation is partitioning a digital image into one or more regions, also known as image segments or image objects (Shapiro *et al* 2001, Gonzalez 2009). It appears as a crucial task in bio-medicine (Ronneberger *et al* 2015, Fisch *et al* 2020, Yakimovich *et al* 2020, Andriasyan *et al* 2021, Galimov and Yakimovich 2022) and beyond.

Classic approaches are given by thresholding (Batenburg and Sijbers 2008), clustering (Shapiro *et al* 2001), histogram-based methods (Shapiro *et al* 2001), compression-based methods (Mobahi *et al* 2011), edge detection (Kimmel 2003), region-growing methods (Nock and Nielsen 2004), partial differential equation-based methods (Caselles *et al* 1997), and graph partitioning methods (Grady 2006).

The center of our empirical investigations is a U-NET based ML architecture termed STARDIST, proposed by Schmidt *et al* (2018a). STARDIST is trained on manually annotated fluorescence microscopy images (Booz 2018, Schmidt *et al* 2018a). While closely packed cells are prone to causing instabilities, resulting in segmentation errors given by merged cell boundaries, STARDIST proposes a star-convex polygon method of cell localization, explicitly addressing this problem. However, segmentation artifacts cannot be avoided completely, reflecting the inherent instability.

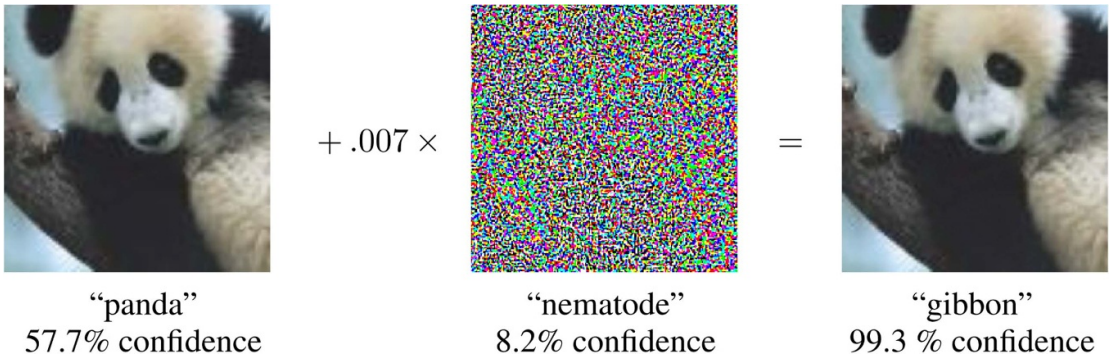

Figure 1. An example of an adversarial attack on GoogLeNet (Szegedy *et al* 2014) causing the classification to change even with a higher confidence level Reproduced with permission from Goodfellow *et al* (2015).

The existence and study of adversarial attacks, which we outline below, provide one perspective on the instability phenomenon.

1.2. Adversarial attacks

The forte of an adversarial attack is that it appears as noise that humans cannot perceive. However, the attacked version appears radically different from the ML model. A demonstration of an adversarial attack is shown in figure 1 by adding an imperceptible perturbation that fools the GoogLeNet (Szegedy *et al* 2014) into misclassifying the image.

Though adversarial attacks were initially considered for classification tasks, they are extended to the various domains of ML, including the initially mentioned ML image processing tasks (Papernot *et al* 2017, Balda *et al* 2020, Chen *et al* 2020, Lord *et al* 2022).

Apart from their specific representation, one may ask: ‘*How can the existence of adversarial example attacks be explained ?*’ (Balda *et al* 2020). Although there are numerous hypothetical and theoretical explanations, such as the *linear hypothesis* (Balda *et al* 2020), we consider the perspective given by the study of *ill-posed inverse problems* to be the most promising for a formal and deeper treatment.

1.3. The inherent instability of ill-posed inverse problems

The *instability phenomenon of ill-posed inverse problems* states that, in general, one cannot guarantee solutions to inverse problems to be stable. The phenomenon was explicitly studied by Szegedy-Goodfellow, Huang, Hansen *et al* (Szegedy*et al* 2014, Huang *et al* 2018, Antun *et al* 2021), with deeper treatments and discussions in Antun *et al* (2020), Gottschling *et al* (2020). The insights translate to the fact that, in general, the *local Lipschitz constant*

$$L_\varepsilon(\Gamma, y) = \sup_{0<\|y'-y\|<\varepsilon} \frac{\|\Gamma(y') - \Gamma(y)\|_X}{\|y' - y\|_Y}, \quad \varepsilon > 0$$

of the reconstruction map $\Gamma : Y \to X$, modelled on metric spaces X, Y with metrics $\|\cdot\|_X, \|\cdot\|_Y$, might be unbounded. Consequently, small noise perturbations $\|y' - y\|_Y \approx 0$ of the input can result in large differences in the reconstruction $\|\Gamma(y') - \Gamma(y)\|_X \gg 0$. This fact can only be avoided if an additional control on the null space, ‘*the troublesome kernel*’ (Gottschling *et al* 2020) of the forward model $\Psi : X \to Y$, with $\Gamma \circ \Psi = \mathrm{id}$ is given. However, such direct control can only be given in rare situations where the null space of Ψ is explicitly known.

Increasing the reliability of ML image segmentation, even in the absence of this knowledge, is the essence of our contribution.

1.4. Contribution—automated artifact detection by polynomial adversarial defenses

Given that it is theoretically and practically impossible to guarantee a perfect image segmentation (and other image-processing outputs), we propose to identify potential artifacts automatically and subsequently correct the detected (potential) wrong segmentations manually, but very quickly, thanks to the prior automated guidance.

To mark potential artifacts we generate bounding boxes B_i, $i = 1, \ldots, n \in \mathbb{N}$ (based on OPENCV) for each of the n segmented cells, including the minimum and maximum coordinates of each segment. Hereby, we choose the features top-left corner $(x_{i,1}, y_{i,1})$, center $(x_{i,0}, y_{i,0})$, and bottom-right corner

$(x_{i,2}, y_{i,2})$ to define each bounding box B_i uniquely. The center feature stabilizes the distance between the bounding boxes:

$$\operatorname{dist}\left(B_i, B_j\right) = \|X_i - X_j\|_2\,, \tag{1}$$

defined as the Euclidean distance of the features, $X_k = (x_{k,1}, y_{k,1}, x_{k,0}, y_{k,0}, x_{k,2}, y_{k,2})$, $k = i, j$. Given these ingredients, our contributions can be summarized as:

C1) We extend and deepen our former work (Jain *et al* 2022, Volpe *et al* 2023), generating polynomial defenses (PDs). The PDs mimic noise distributions that, when removed from the images, significantly increase the segmentation quality, validated on the training dataset.

C2) We perturb an unseen (test) image by the PDs and measure the distances $\operatorname{dist}(B_i, B_i')$ of the original and resulting bounding boxes. In case the distance is large, we flag the box B_i as a potential segmentation artifact. In comparison to classic Gaussian and Poisson noise perturbations, in the case of StarDist and its dataset (Booz 2018, Schmidt *et al* 2018a), we demonstrate this PD-based artifact detection to be much more precise, enabling quick manual re-adjustment of the segmentation output.

We want to stress that our contribution **C2**) not only enables to increase in the reliability of the StarDist output but can be used to cross-check the given manually segmented training dataset. Though this work is a proof of concept in its current state, it may serve as a baseline for extensions to other image-processing tasks, datasets, and models.

2. Image segmentation

We formalize the image segmentation problem: We model images $\mathrm{im} \in \mathbb{R}^{n\times n}$ as a pixel-value-vector on an equidistant pixel grid $G = \{k/(n-1) : k = 0, \ldots, n-1\}^2$ of resolution $n \times n$. We assume *noisy images* $\mathrm{im} \in \mathbb{R}^{n\times n}$ to be given as an affine perturbation of a ground-truth image $\mathrm{gt} \in \mathbb{R}^{n\times n}$ by a *noise distribution* $\eta \in \mathbb{R}^{n\times n}$,

$$\mathrm{im} = \mathrm{gt} + \eta, \quad \mathrm{im}, \mathrm{gt}, \eta \in \mathbb{R}^{n\times n}. \tag{2}$$

Definition 1. Given a set of images $\mathrm{IM} = \{im_i : i = 1, \ldots, N\} \subseteq \mathbb{R}^{n\times n}$, $M \in \mathbb{N}$ of resolution $n \times n$, $n \in \mathbb{N}$. We call $\Phi : \mathbb{R}^{n\times n} \to \mathbb{N}^{n\times n}$ an *image segmentation map* (ISM) assigning labels $k \in \mathbb{N}$ to each pixel, whereas a zero label $k = 0$ indicates an empty pixel (the absence of an object), and call the resulting integer vector

$$\mathrm{sm}_i = \Phi\left(\mathrm{im}_i\right) \in \mathbb{N}^{n\times n}, \forall\, i = 1, \ldots, N \tag{3}$$

segmentation masks (SMs).

Computing an ISM states the *segmentation problem*. In practice, ISMs can only be approximated. Measuring the ISM quality rests on the ground truth SM and is usually computed by the *Intersection-over-Union* (IoU) metric (Padilla *et al* 2020), with respect to the label $k \in \mathbb{N}$:

$$\mathrm{IoU}\left(\mathrm{sm}, \Phi\left(\mathrm{im}\right), k\right) = \frac{\left|\mathrm{sm} \cap_k \Phi\left(\mathrm{im}\right)\right|}{\left|\mathrm{sm} \cup_k \Phi\left(\mathrm{im}\right)\right|} \in [0, 1]\,, \tag{4}$$

where $\mathrm{sm} \cap_k \Phi(\mathrm{im}) = \{g \in G : \mathrm{sm}(g) = \Phi(g)\} = k$, $\mathrm{sm} \cup_k \Phi(\mathrm{im}) = \{g \in G : \mathrm{sm}(g) = k \text{ or } \Phi(g) = k\}$, see figure 2 for an illustration.

Two edge cases might appear: Either there is no k-label overlap of sm, $\Phi(\mathrm{im})$ and the IoU vanishes, or sm, $\Phi(\mathrm{im})$ coincide with respect to the label $k \in \mathbb{N}$ and the IoU reaches 1, which motivates the following notion.

Definition 2 (false positives and negatives). We fix a label $k \in \mathbb{N}$, introduce a threshold $0 < \tau < 1$, and an *object mask* $\mathrm{om} \in \{0, k\}^{n\times n}$. If $\mathrm{IoU}(\mathrm{om}, \Phi(\mathrm{im}), k) \geqslant \tau$ the segmented instance $\Phi(\mathrm{im}) \cap \{0, k\}^{n\times n}$ is set as *true positive* (TP) with respect to om. Vice versa, if $\mathrm{IoU}(\mathrm{om}, \Phi(\mathrm{im}), k) < \tau$ then $\Phi(\mathrm{im}) \cap \{0, k\}^{n\times n}$ is set as *false positive* (FP) and the object mask om is set as *false negative* (FN) (Mechea 2019, Padilla *et al* 2020), see figure 2 for an example.

Given several object segments $\mathrm{OM} = \{\mathrm{om}_i \in \{0, k_i\}^{n\times n} : k_i \in \mathbb{N}, i = 1, \ldots, M\}$, $M \in \mathbb{N}$ counting TPs, FPs, FNs yields the commonly used measurements *precision* and *recall*:

$$\mathrm{precision} = \frac{TP}{TP + FP}\,, \quad \mathrm{recall} = \frac{TP}{TP + FN}\,. \tag{5}$$

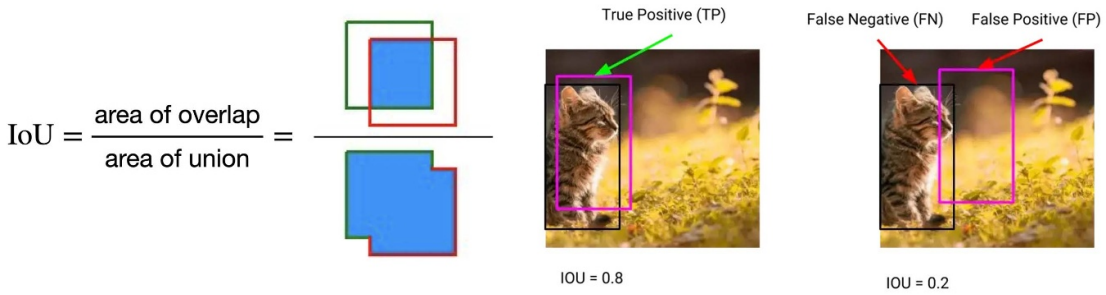

Figure 2. Illustration of IoU (left). The green box represents the object mask om and the red box $\Phi(\mathrm{im})$. An example of a true positive segment (TP), false negative segment (FN), and false positive segment (FP) (right).

There is a trade-off between precision and recall appearing for the edge cases $FP \ll FN$, $FP \gg FN$, which are balanced by introducing the $\mathrm{f1}_{\mathrm{score}}$ (Taha and Hanbury 2015) or *Dice Coefficient* (Manning 2008) as the harmonic mean:

$$\mathrm{f1}_{\mathrm{score}} = 2\frac{\mathrm{precision}\cdot\mathrm{recall}}{\mathrm{precision}+\mathrm{recall}} = \frac{TP}{TP+\frac{1}{2}(FP+FN)}. \tag{6}$$

As long as $TP>0$, $\mathrm{f1}_{\mathrm{score}} \approx 1$ indicates high precision and recall, whereas small $\mathrm{f1}_{\mathrm{score}} \approx 0$ appears for either of them being small. Similarly,

$$\mathrm{accuracy} = \frac{TP}{TP+FP+FN} \tag{7}$$

or the *Panoptic Quality* (PQ) (Taha and Hanbury 2015, Kirillov *et al* 2019)

$$PQ = \frac{\sum_{(\Phi(\mathrm{im}),\mathrm{om})\text{ is a }TP}\mathrm{IoU}(\mathrm{om},\Phi(\mathrm{im}),k)}{TP+\frac{1}{2}FP+\frac{1}{2}FN}.$$

They are used for measuring the quality of segmentation tasks.

We use these classic ingredients for defining *adversarial defenses*:

Definition 3 (adversarial defences of ISMs). Given a set of noisy images $\mathrm{IM}=\{\mathrm{im}_i=\mathrm{gt}_i+\eta_i\in\mathbb{R}^{n\times n}: i=1,\ldots,N\}$ and an ISM $\Phi:\mathbb{R}^{n\times n}\to\mathbb{N}^{n\times n}$. We fix an image $\mathrm{im}=\mathrm{gt}+\eta\in\mathrm{IM}$, a label $k\in\mathbb{N}$, a threshold $0<\tau<1$, and an object mask $\mathrm{om}\in\{0,k\}^{n\times n}$. Let $\varepsilon\in\mathbb{R}^{n\times n}$, with $|\varepsilon|\leqslant|\eta|$ and

$$\mathrm{IoU}(\mathrm{om},\Phi(\mathrm{im}+\varepsilon),k) > \mathrm{IoU}(\mathrm{om},\Phi(\mathrm{im}),k)+\tau \tag{8}$$

then we call ε an adversarial defence (AD) of Φ with respect to $(\mathrm{om},\mathrm{im})$.

In a nutshell, ADs are distributions on the level of noise that significantly increase the segmentation quality. How to compute ADs is the focus of the next section.

3. Polynomial parametrization of adversarial defences

Based on our prior work (Hecht *et al* 2017, 2018, 2020, 2025, Schmidt *et al* 2018a), addressing polynomial interpolation and regression, we propose parametrizing ADs by bivariate polynomials. That is, we consider the multi-index set $A_{n,p}=\{\alpha\in\mathbb{N}^2:\|\alpha\|_p\leqslant n\}$, where $\|\cdot\|_p$ denotes the l_p-norm, $p\geqslant 1$. The sets $A_{n,p}$ generalize the concept of 1D polynomial degree $n\in\mathbb{N}$ to polynomial l_p-degree in 2D. We denote with

$$\Pi_A = \mathrm{span}\{x^\alpha = x_1^{\alpha_1}\cdot x_2^{\alpha_2}\}_{\alpha\in A}, A=A_{n,p}$$

the polynomial space generated by the canonical monomials with exponents $\alpha\in A$. It is a classic fact (Trefethen 2019) that the *Chebyshev–Lobatto nodes* $\mathrm{Cheb}_n=\{\cos(k\pi/n):0\leqslant k\leqslant n\}$ are (asymptotically) optimal for polynomial interpolation, avoiding *Runge's over-fitting phenomenon*. Consequently, we here consider the 2D-Chebyshev-grids

$$P_A=\left\{p_\alpha=(p_{\alpha_1},p_{\alpha_2})\in[-1,1]^2: p_{\alpha_1},p_{\alpha_2}\in\mathrm{Cheb}_n^{\leq_{\mathrm{Leja}}},\alpha\in A\right\},A=A_{n,p},$$

generated by ordering the Chebyshev–Lobatto nodes $\mathrm{Cheb}_n^{\leq_{\mathrm{Leja}}}$ with respect to the *Leja ordering* (Leja 1957). The 2D-Chebyshev-grids maintain the approximation power of the 1D-version (Hecht *et al* 2020, 2025), providing uniform and exponentially fast approximations of regular functions $f : [-1,1]^2 \to \mathbb{R}$ (Hecht *et al* 2020, 2025). In particular, choosing Euclidean degree $p = 2$ results in a pivotal choice (Trefethen 2017, Bos and Levenberg 2018, Hecht *et al* 2020, 2025, Veettil *et al* 2022). We incorporate these insights into the following definitions.

Definition 4 (Newton Polynomials). Given P_A, $A = A_{n,p}$, $n \in \mathbb{N}$, $p \geqslant 1$ the *bivariate Newton polynomials* are defined by

$$N_\alpha(x) = \prod_{i=1}^{2} \prod_{j=0}^{\alpha_i - 1} \left(x_i - p_{j,i}\right), \quad \alpha \in A, \tag{9}$$

generalizing the classic notion of 1D Newton polynomials.

Definition 5 (Lagrange polynomials). Given P_A, $A = A_{n,p}$, $n \in \mathbb{N}$, $p \geqslant 1$ the Lagrange polynomials $L_\alpha \in \Pi_A$ are defined by the requirement

$$L_\alpha(p_\beta) = \delta_{\alpha,\beta}, \tag{10}$$

where $\delta_{\cdot,\cdot}$ denotes the Kronecker delta. The Lagrange polynomials form a basis of Π_A and can be expressed as analytic expressions in Newton-form (Hecht *et al* 2020, 2025)

$$L_\alpha(x) = \sum_{\beta \in A} c_{\alpha,\beta} N_\beta(x). \tag{11}$$

This enables the evaluation of $L_\alpha(y) \in \mathbb{R}$ in any argument $y \in \mathbb{R}^2$ efficiently and numerically stable. Documentation of implementation and further details are provided in Wicaksono *et al* (2023).

The notion allows us to set up numerically stable regression schemes (Veettil *et al* 2022), which we introduce here in a simplified version matching our needs.

Definition 6 (Regression matrices). Let $G = \{k/(n-1) : k = 0, \ldots, n-1\}^2$ be an equidistant pixel grid of resolution $n \times n$, $n \in \mathbb{N}$. We consider the multi-index sets $B = A_{n,\infty}$, $A = A_{k,2}$, $k \leqslant n \in \mathbb{N}$ and define the regression matrix

$$R_{B,A} = \left(L_\alpha(g_\beta)\right)_{\beta \in B, \alpha \in A} \in \mathbb{R}^{|B| \times |A|}, \quad g_\beta = (g_{\beta_1}, g_{\beta_2}) \in G, \tag{12}$$

where L_α are the Lagrange polynomials with respect to the 2D-Chebyshev-grid.

Given a polynomial $Q = \sum_{\alpha \in A} c_\alpha L_\alpha$ with coefficients $C = (c_\alpha)_{\alpha \in A} \in \mathbb{R}^{|A|}$ in Lagrange form, its values $D = (Q(p_\beta))_{\beta \in B}$ on the equidistant pixel grid G are given by $D = R_{B,A} C$. Here, we assume that $A = A_{k,2}$, $k \in \mathbb{N}$ is chosen such that $|A| \ll |B| = |G|$, and regression is applied in the over-determined case. For function values $F = (f(p_\beta))_{\beta \in B}$, $f : \Omega \to \mathbb{R}$ a least square solution minimizing

$$C = \mathrm{argmin}_{X \in \mathbb{R}^{|A|}} \|F - R_{B,A} X\|^2$$

yields a polynomial fit Q_C of f, delivering close polynomial approximations for a vast class of regular functions f (Dian *et al* 2020).

Here, we focus on the converse approach: We seek to generate function values $F = R_{B,A} C$ that mimic noise distributions $F \approx \varepsilon$ resulting in ADs, definition 3. The computation of the polynomial coefficients $C \in \mathbb{R}^{|A|}$, identifying ε, rests on the following loss formulation.

Definition 7 (polynomial loss for PDs). Consider polynomials $Q_C = \sum_{\alpha \in A} c_\alpha L_\alpha \in \Pi_A$, with coefficients $C = (c_\alpha)_{\alpha \in A} \in \mathbb{R}^{|A|}$, definition 5. Let further $\mathrm{im} \in \mathbb{R}^{n \times n}$ be an image of resolution $n \times n$, $n \in \mathbb{N}$, $R_{B,A} \in \mathbb{R}^{|B| \times |A|}$ be the regression matrix, and Φ be an ISM. For the ground truth segmentation $\mathrm{gm} \in \mathbb{N}^{n \times n}$, we consider the *calibration loss*

$$\mathcal{L}(\mathrm{gm}, X) = -\mathrm{f1}_{\mathrm{score}}(\mathrm{gm}, \Phi(\mathrm{im} - Q_X)) \tag{13}$$

and call $Q_C \in \Pi_A$, with $C = \mathrm{argmin}_{X \in \mathbb{R}^{|A|}} \mathcal{L}_S(\mathrm{gm}, X)$ a *polynomial adversarial defence* (PD) of Φ whenever Q_C matches definition 3. The notions translate when exchanging the $\mathrm{f1}_{\mathrm{score}}$ metric with the aforementioned alternatives.

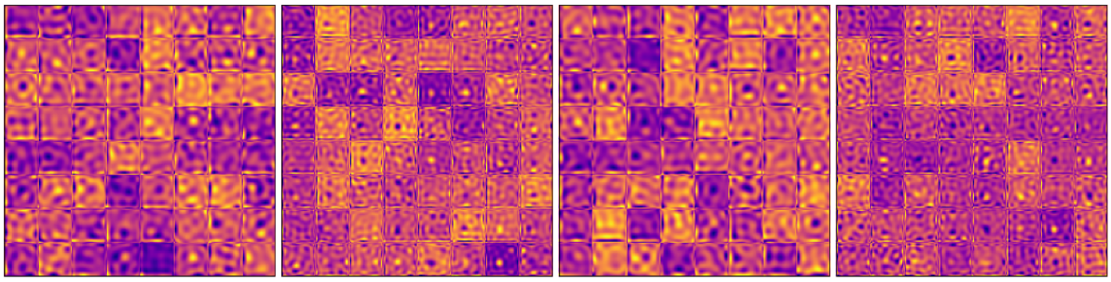

Figure 3. Example of PDs on a pixel grid G of resolution 256×256 generated for 8×8 subdomains and polynomials of l_2-degree $n = 13$.

For suitable polynomial degrees, e.g. $n = 1, \ldots, 15$, we have $|C| \ll |B| = |G|$. Thus, the optimization problem in equation (13) is of much lower dimension than when directly formulated for the pixel grid G. As we demonstrate in section 4, this makes the PDs computable with reasonable effort when decomposing the pixel grid G into suitable sub-domains, as illustrated in figure 3.

4. Experiments

We have designed representative experiments to validate and demonstrate our theoretical findings. Execution and evaluation were done for PYTHON VERSION 3.10 implementations on an HPC WITH NVIDIA COMPUTE UNIFIED DEVICE ARCHITECTURE (CUDA, version 11.2 and above). The PD generation follows definition 7 based on MINTERPY (Wicaksono *et al* 2023) and HYPPOPY (André and Wanner 2019). A repository including implementations, datasets, and all benchmarks is available at https://doi.org/10.5281/zenodo.15 901 527.

For the following experiments, the pre-trained STARDIST segmentation architecture by Schmidt *et al* (2018a, 2018b)was trained on manually annotated real fluorescence microscopy images of cell nuclei from the 2018 Data Science Bowl (Booz 2018, Schmidt *et al* 2018a). This dataset contains 450 training images and 50 test images of varying resolutions. For simplicity, we cropped all the images to a resolution of 256×256.

Experiment 1 (Calibrating STARDIST by PDs). From the total 450 training images of resolution 256×256 each, we found $N = 280$ images with $\mathrm{f1}_{\mathrm{score}}(\mathrm{StarDist}(\mathrm{im})) < 0.9$ based on an IoU-threshold $\tau = 0.75$, equation (6). We generated 280 PDs subdivided into 4×4 or 8×8sub-domains, with a polynomial of Euclidean degree $A_{n,p}$, $n = 9$, $p = 2$ for each subdomain, one for each of these images, optimizing the calibration loss definition 7. The generated PDs have very low intensity, not damaging the structural properties of the images, see table 1.

Figure 4 shows the impact of the PD calibration, significantly improving the STARDIST segmentation accuracy, demonstrating the PDs to be capable of preventing false positive or negative segmentation artifacts. How to extend the PD calibration to unseen images is outlined below.

4.1. Identifying potential false positives (FPs) or false negatives (FNs)

Algorithm 1 formalizes the outline process of detecting false positives or false negatives. Essentially, we draw bounding boxes B_i, $i = 1, \ldots, n \in \mathbb{N}$ for each of the identified cells of the STARDIST segmentation and B_i', $i = 1, \ldots, n' \in \mathbb{N}$ for each of the identified cells of a PD perturbed STARDIST segmentation. We compute the distance of the corresponding bounding boxes, according to equation (1) and flag the original box B_i as potential FP if the distance $\mathrm{dist}(B_i, B_i') > \tau$ is larger than a predetermined distance threshold τ. The procedure is repeated for all of the $N = 280$ PDs chosen to calibrate the unseen image. If the original bounding box B_i is zero (no cell), then B_i' is flagged as a potential FN. As multiple false negative flags may occur for the same cell, overlapped FPs within an overlap ratio R are merged, indicating a region of false negative segments.

The *distance threshold* $\tau \in [7, 35]$, the *percentage of the PD pertubation* $X\% \in [0\%, 100\%]$, and the *overlap ratio* $R \in [1, 3]$ are prior determined by hyperparameter optimization on the validation images with respect to the $\mathrm{f1}_{\mathrm{score}}$, equation (6) in the range specified above, resulting in $\tau = 15$, $X\% = 70\%$, and $R = 1.7$.

Table 1. Average and maximum pixel intensities for 280 PDs generated and the pixel intensities for their corresponding images from Booz (2018), Schmidt *et al* (2018a). Reproduced from Booz (2018). CC0.

	Raw image pixel intensity	Normalized image	PD intensity pixel intensity	PD percentage
Average	14.43	0.09	0.005	5.56%
Maximum	255	1	0.25	25%

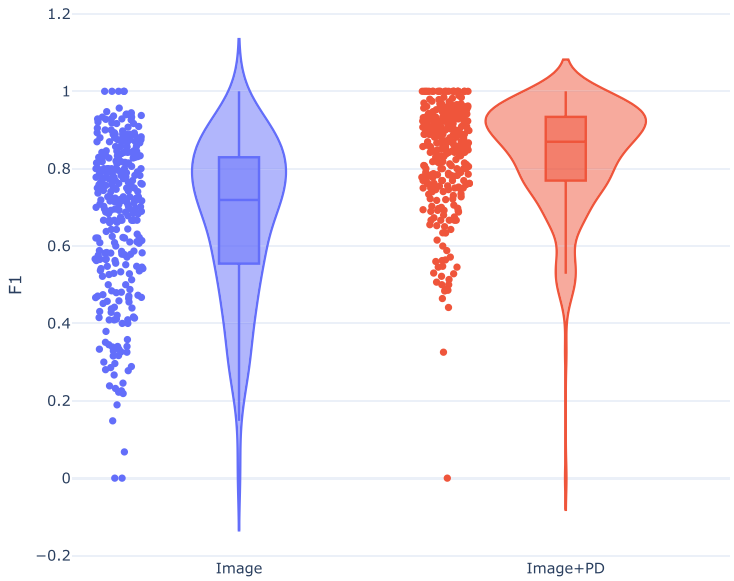

Figure 4. f_1 score comparison of the StarDist segmentation of the 280 training images without and by applying the corresponding PD.

Algorithm 1. Detecting False Negative and False Positive Segmentations.

Input: StarDist-Model, Image, Mask, GroundTruth
Input: PolynomialDefenses, $X\%$, τ, R
StarDist-Prediction ← StarDist-Model.predict(Image)
StarDist-Cells ← BoundingBoxes(StarDist-Prediction)
FalsePositives ← List()
FalseNegatives ← List()
for Defense in PolynomialDefenses **do**
    DefenseImage ← Image—Defense∗X%
    DefensePrediction ← StarDist-Model.predict(DefenseImage)
    DefenseCells ← BoundingBoxes(DefensePrediction)
    CellDistances ← PairwiseDistances(StarDist-Cells, DefenseCells)
    for Distance, Cell **in** CellDistances **do**
        if Distance $> \tau$ **then**
            FalsePositives.update(Cell)
        end if
    end for
    for Distance, Cell **in** CellDistances.Transpose **do**
        if Distance $> \tau$ **and** StarDist-Prediction $==0$ **then**
            FalseNegatives.update(Cell)
        end if
    end for
    FalseNegatives ← **Merge**(Overlapped Cells)
end for

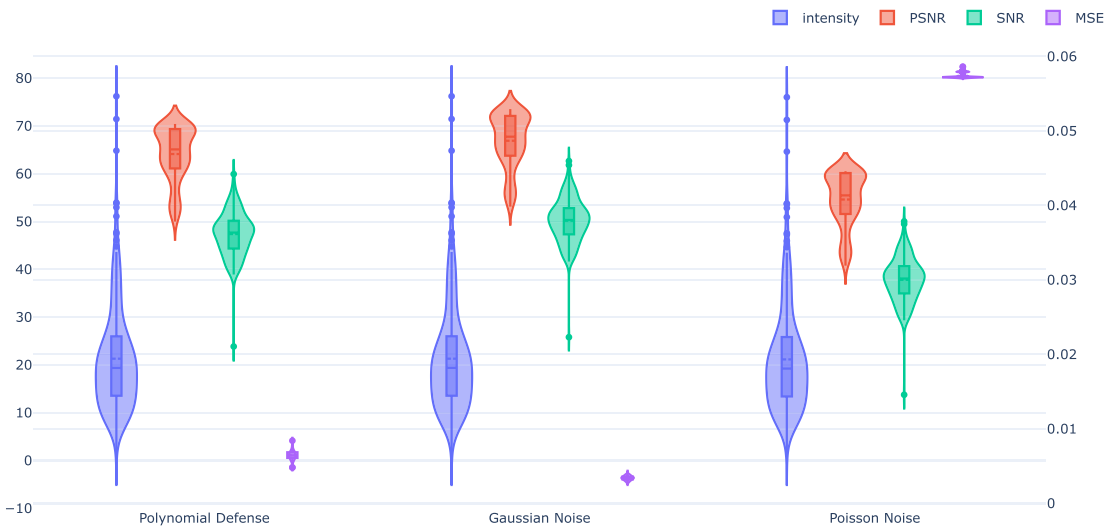

Figure 5. Comparison of PDs, Gaussian, and Poisson noises. The left *y*-axis, *Intensity*, denotes the averaged pixel intensity per image, *PSNR* and *SNR* are in the log scale. *MSE* is shown on the right *y*-axis. The corresponding means are reported in table 2.

Table 2. Mean of PDs, Gaussian, and Poisson noises w.r.t. pixel intensity, PSNR, SNR, and MSE compared to the original image.

Metric	Polynomial Defense	Gaussian Noise	Poisson Noise
intensity	21.31	21.31	21.16
PSNR (dB)	64.13	66.91	54.62
SNR (dB)	47.39	50.17	37.81
MSE	0.01	0.00	0.06

Experiment 2. The 280 PDs from experiment 1 have a controlled pixel intensity as shown in figure 5, with their mean values shown in table 2. The unseen dataset contains 80 images of resolution 256×256, equally divided into a validation set and a test set. We execute algorithm 1 to detect potential false segmentations.

Figures 6 and 7 show examples of identified potential FPs and FNs based on the PDs. We recognize a very precise detection of critical segmentations, that thanks to this guidance can be quickly, manually re-adjusted.

Of course, the question arises whether the PDs allow tighter identification of potential segmentation artifacts as classic noise perturbations. This question is investigated below.

4.2. Precision of the PD based detection

The precision of PD artifact detection is compared with classic Gaussian and Poisson noise perturbations.

Experiment 3. We keep the dataset from experiment 2 and compare the PD artifact detection with $N = 280$ randomly generated Gaussian and Poisson noise distributions, whose hyperparameter settings are optimized as for the PDs.

Figure 8 plots the intensity, PSNR, SNR, and MSE of the three defences in total, showing the defence distributions to be of similar impact. Figures 8–11 show case the prediction of all three resulting artifact detections. We observe that the PD artifact detections are much more accurate than the detections that Gaussian and Poisson defenses achieve. More precisely:

A total of 80 test images were evaluated, of which 50 images achieved an $\mathrm{f1}_{\mathrm{score}}$ greater than 0.95, indicating that they were already well-segmented by StarDist. For these well-segmented images, applying our proposed PD-based detection method resulted in approximately 15 correctly identified FP or FN cells in total. Apart from 5 incorrect FPs or FNs predictions, this leads to a $\mathrm{f1}_{\mathrm{score}}$ improvement of 1.0 for these images.

In contrast, the Gaussian and Poisson defenses were unable to identify any of the FPs or FNs and predicted at least 2 incorrect FPs or FNs per image.

For the remaining 30 images with $\mathrm{f1}_{\mathrm{score}}$ below 0.95, the PD-based detection method identified 65 correct FPs or FNs cells and 25 incorrect FPs or FNs cells in total, averaging to 2 correct and less than 1

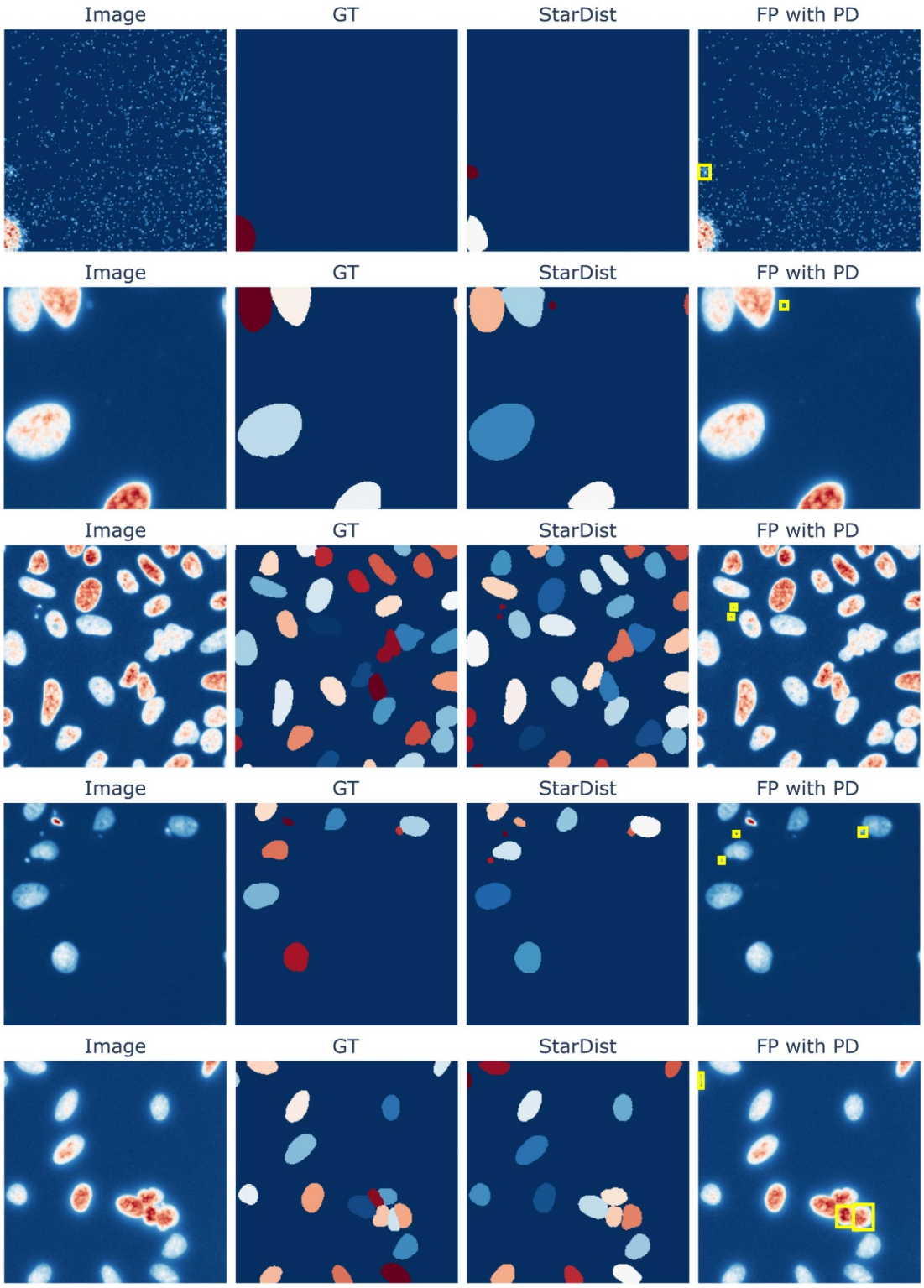

Figure 6. Potential false positives (FP) identified by PDs.

incorrect predictions per image. The Gaussian and Poisson defenses applied to these 30 images averaged to more than 5 incorrect predictions and fewer than 1 correct detection per image.

4.3. Generalization across datasets and models

To evaluate the robustness and generalizability of the present PD-approach, we extend the experiments in section 4 to a different dataset and a distinct segmentation model. Hereby, we use the pre-trained StarDist architecture (Booz 2018, Schmidt *et al* 2018a) labeled as model - '`2D_versatile_he`', which

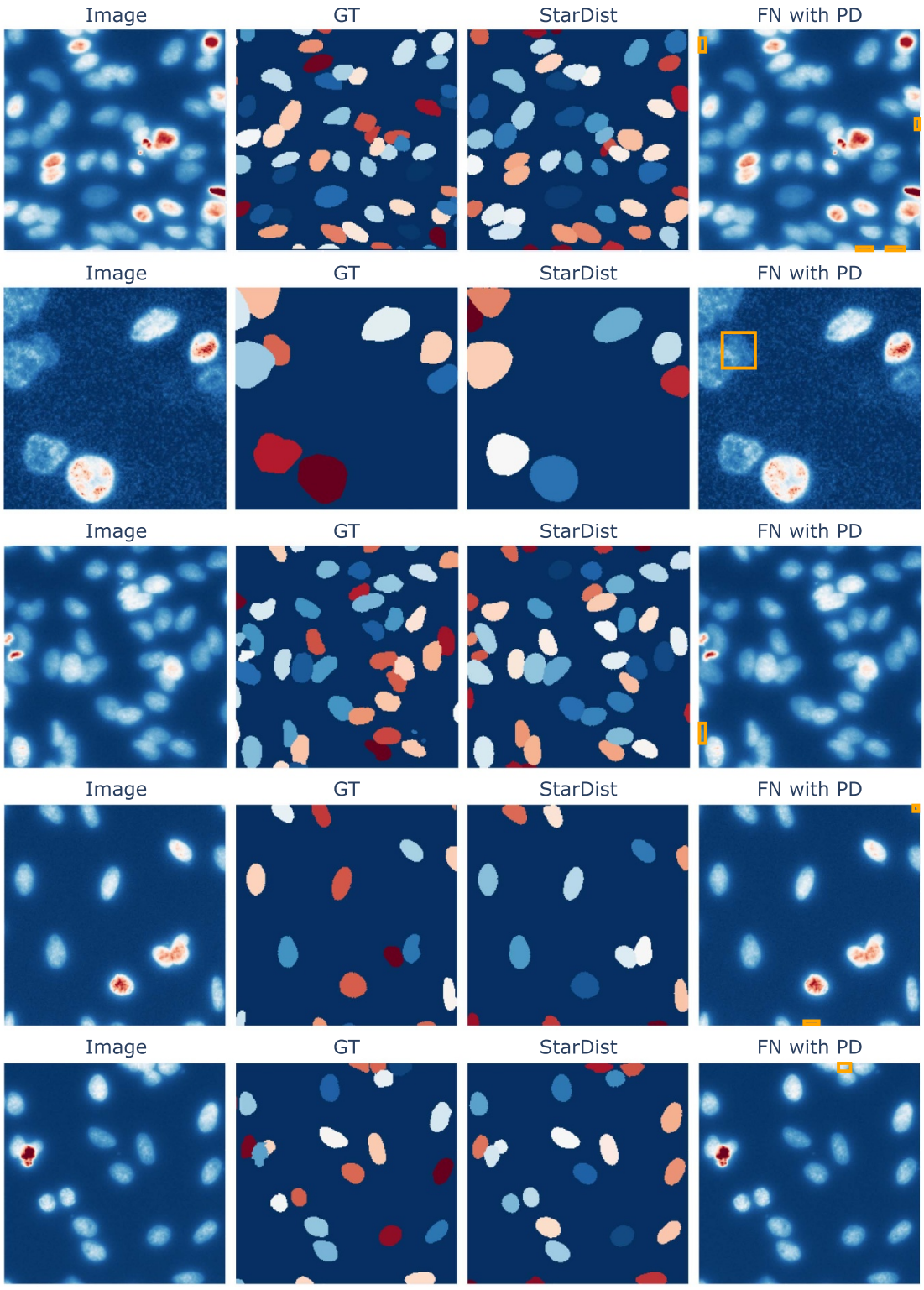

Figure 7. Potential false negatives (FP) identified by PDs.

was trained on three-channel (RGB) brightfield H&E (Hematoxylin and Eosin) stained images. In contrast, the StarDist model used in section 4 was trained on single-channel fluorescent nuclei images from the 2018 Data Science Bowl (Booz 2018, Schmidt *et al* 2018a) dataset.

For this analysis, we employed the MoNuSeg dataset (Kumar *et al* 2017), which contains H&E-stained tissue images of tumors from various organs, collected from multiple hospitals. The training set comprises 30 images of size 1026×1026, with approximately 22.000 manually annotated nuclear boundaries.

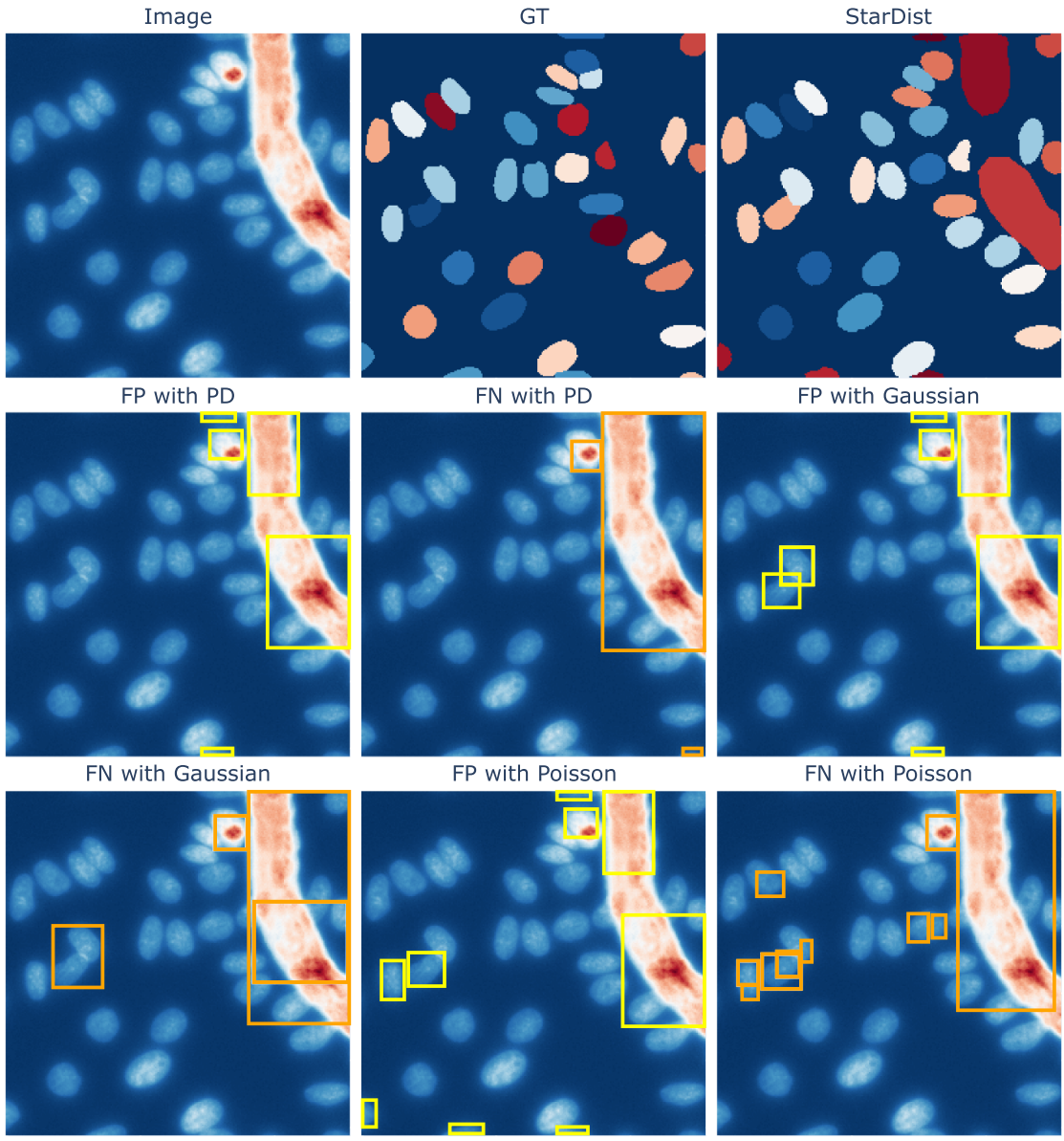

Figure 8. PD, Gaussian, and Poisson defences.

Hyperparameter tuning was performed using HYPPOPY (André and Wanner 2019). For computational efficiency, the dataset was split into 20% validation and 80% test subsets. In the ranges $\tau \in [7, 35]$, $X\% \in [0\%, 100\%]$, $R \in [1, 3]$ the optimal values found are $\tau = 24$, $X = 60\%$, and $R = 2.5$.

Experiment 4. We cropped the images into 9 non-overlapping sub-images of size 256×256, yielding a total of 270 images. The same 280 PDs generated in experiment 1 were used to detect potential FPs and FNs in this extended analysis using algorithm 1. Since the images in the MoNuSeg dataset are RGB, each PD was applied channel-wise, unlike the single-channel processing used for the 2018 Data Science Bowl dataset.

Figures 12 and 13 illustrate examples of identified FPs and FNs on the test images. Similar to previous results, the PD-based detection method successfully highlights segmentation inconsistencies. Compared to Gaussian and Poisson noise perturbations, the PD-based detections remain notably more precise. See figures 14–18 for comparison plots.

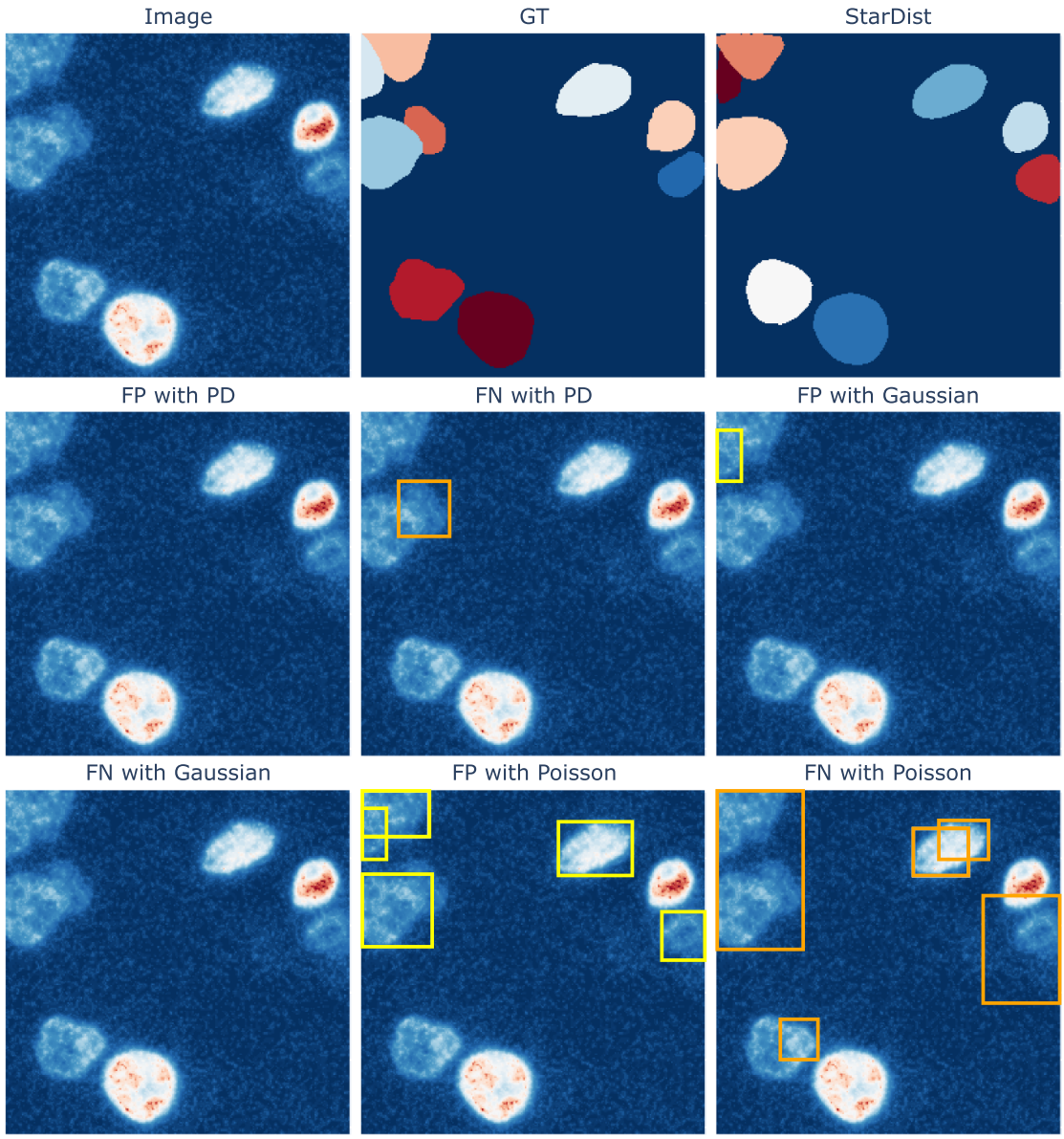

Figure 9. PD, Gaussian, and Poisson defences.

To quantify performance, 45 test images were manually evaluated. The proposed PD-based method correctly identified 215 FP or FN cells, with 120 incorrect identifications-yielding an average of 5 correct and fewer than 3 incorrect predictions per image.

Experiment 5. We adapt experiment 3 for randomly generating Gaussian and Poisson noise distributions, whose hyperparameter settings are optimized as done for the PDs.

The Gaussian and Poisson perturbations produced over 6 incorrect and fewer than 2 correct detections per image. This evaluation was conducted manually, as automated comparison against ground truth masks proved infeasible due to annotation ambiguities. Notably, several apparent inaccuracies were observed in the ground truth masks themselves, suggesting that the proposed method could also serve to enhance the quality of manually labeled datasets.

While the rate of incorrect detections was slightly higher for the MoNuSeg dataset compared to the DSB2018 dataset, the experiment demonstrates the applicability of our approach to diverse data sources and imaging modalities. These findings provide strong proof of concept of the proposed PD-based detection mechanism and algorithm 1, being robust tools for identifying segmentation artifacts in varied nuclei segmentation tasks.

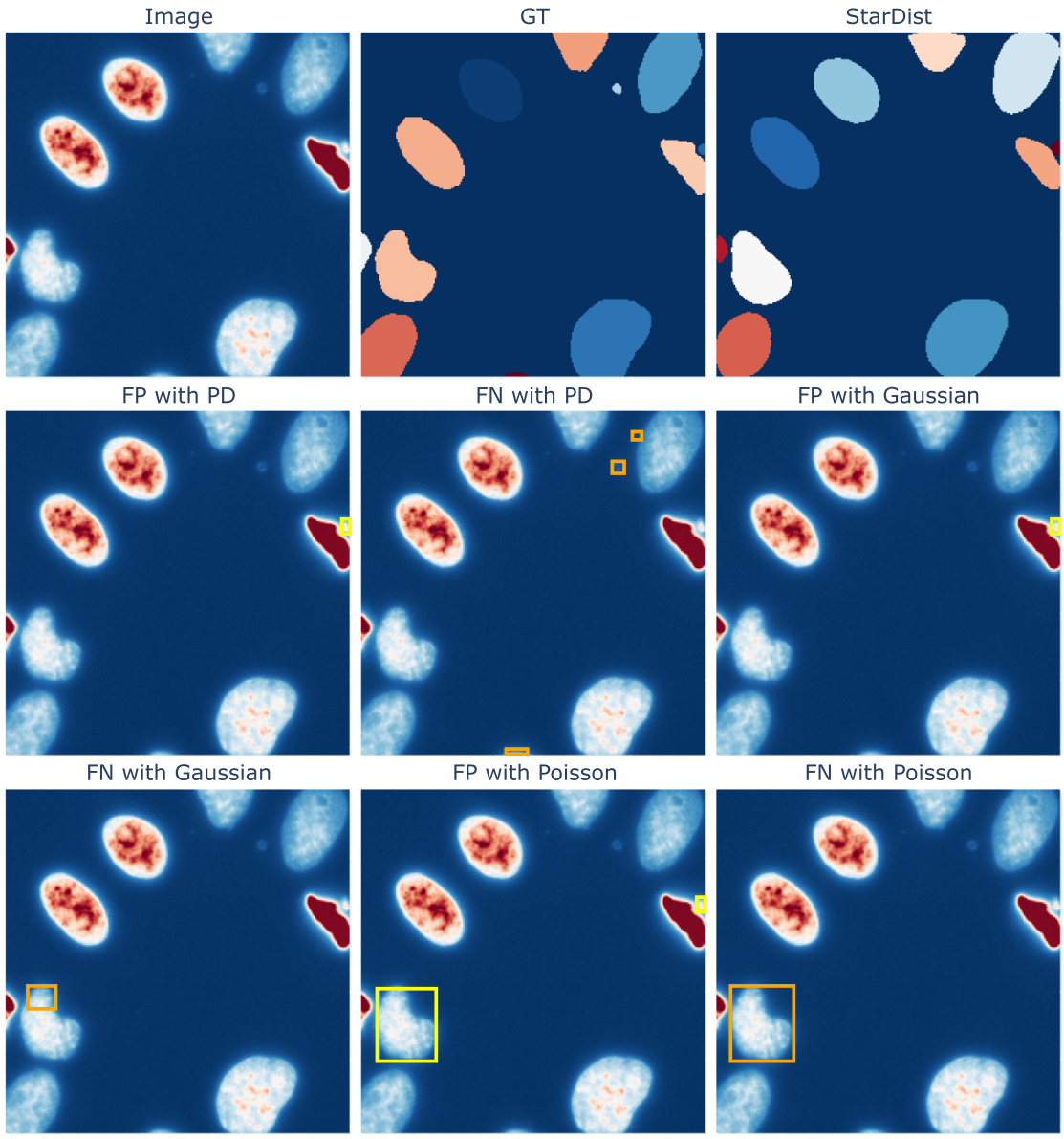

Figure 10. PD, Gaussian, and Poisson defences.

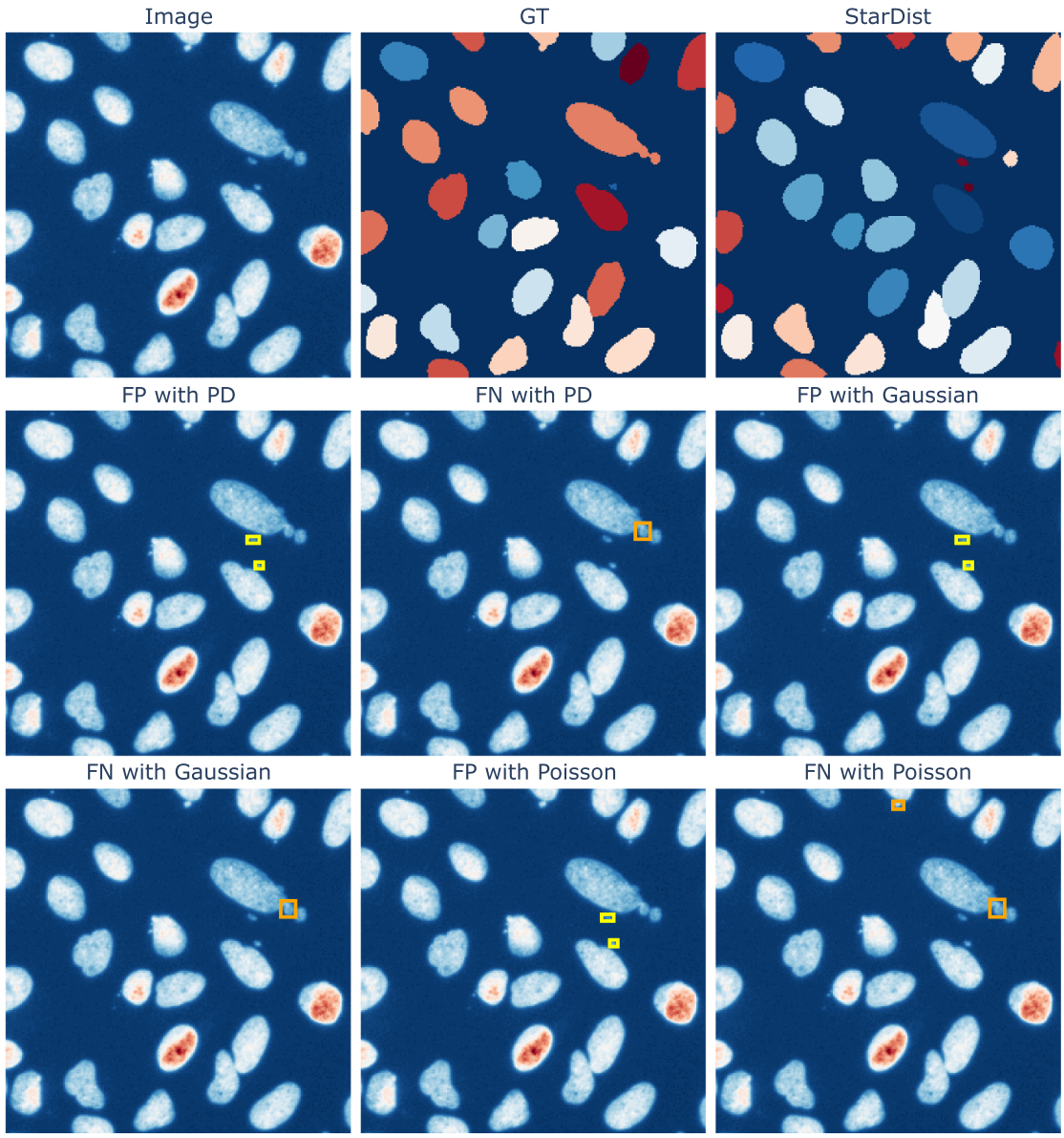

Figure 11. PD, Gaussian, and Poisson defences.

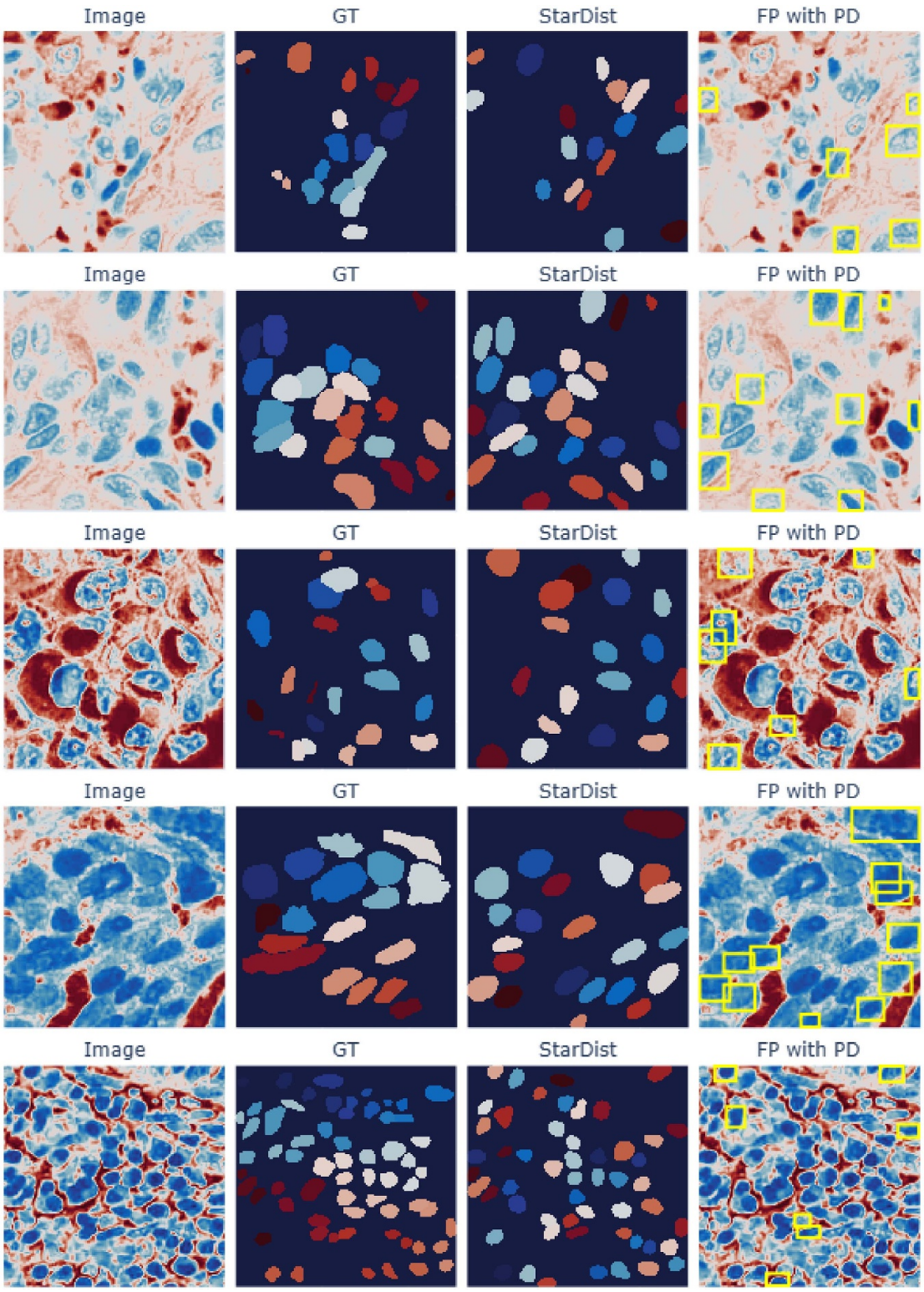

Figure 12. Potential false positives (FP) identified by PDs.

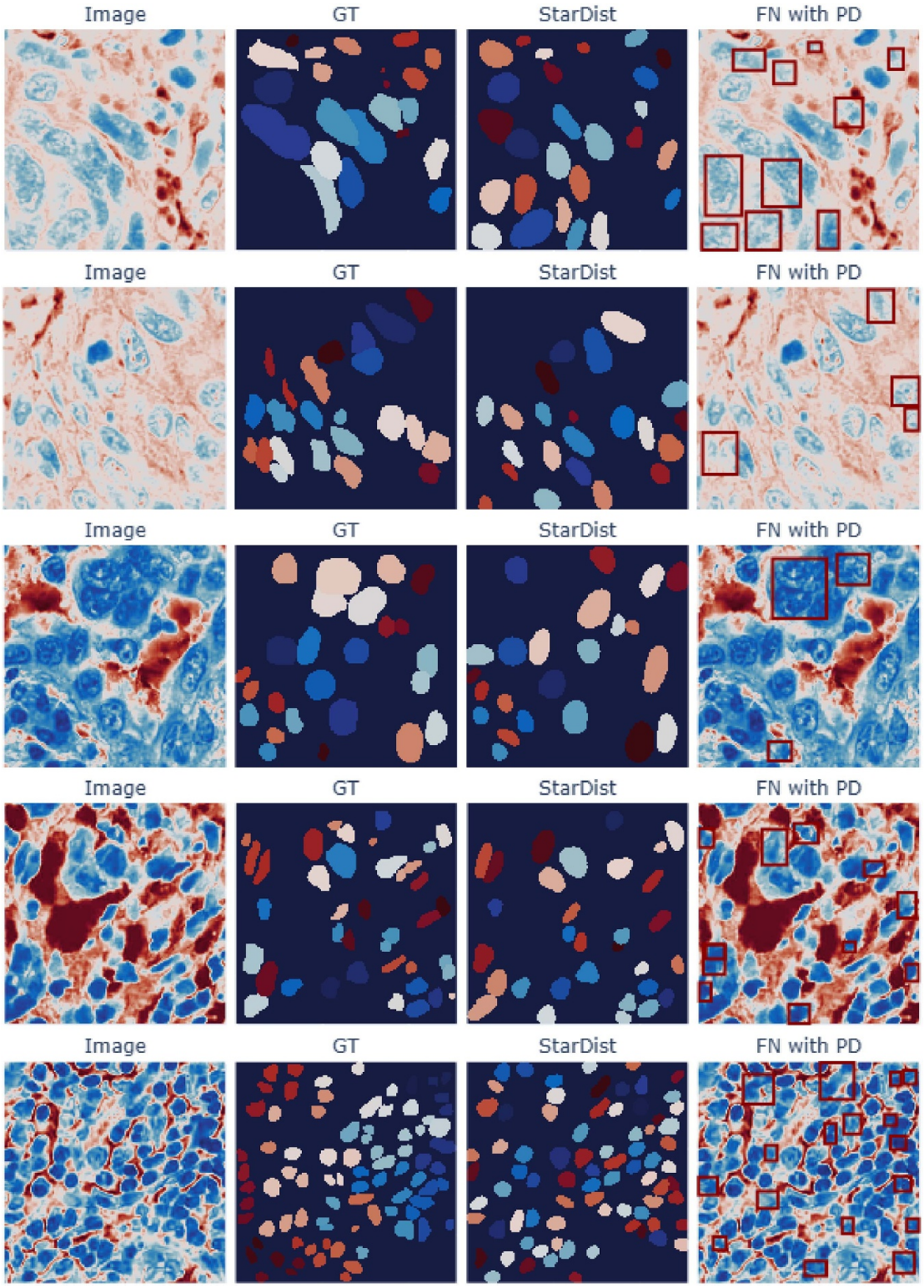

Figure 13. Potential false negatives (FN) identified by PDs.

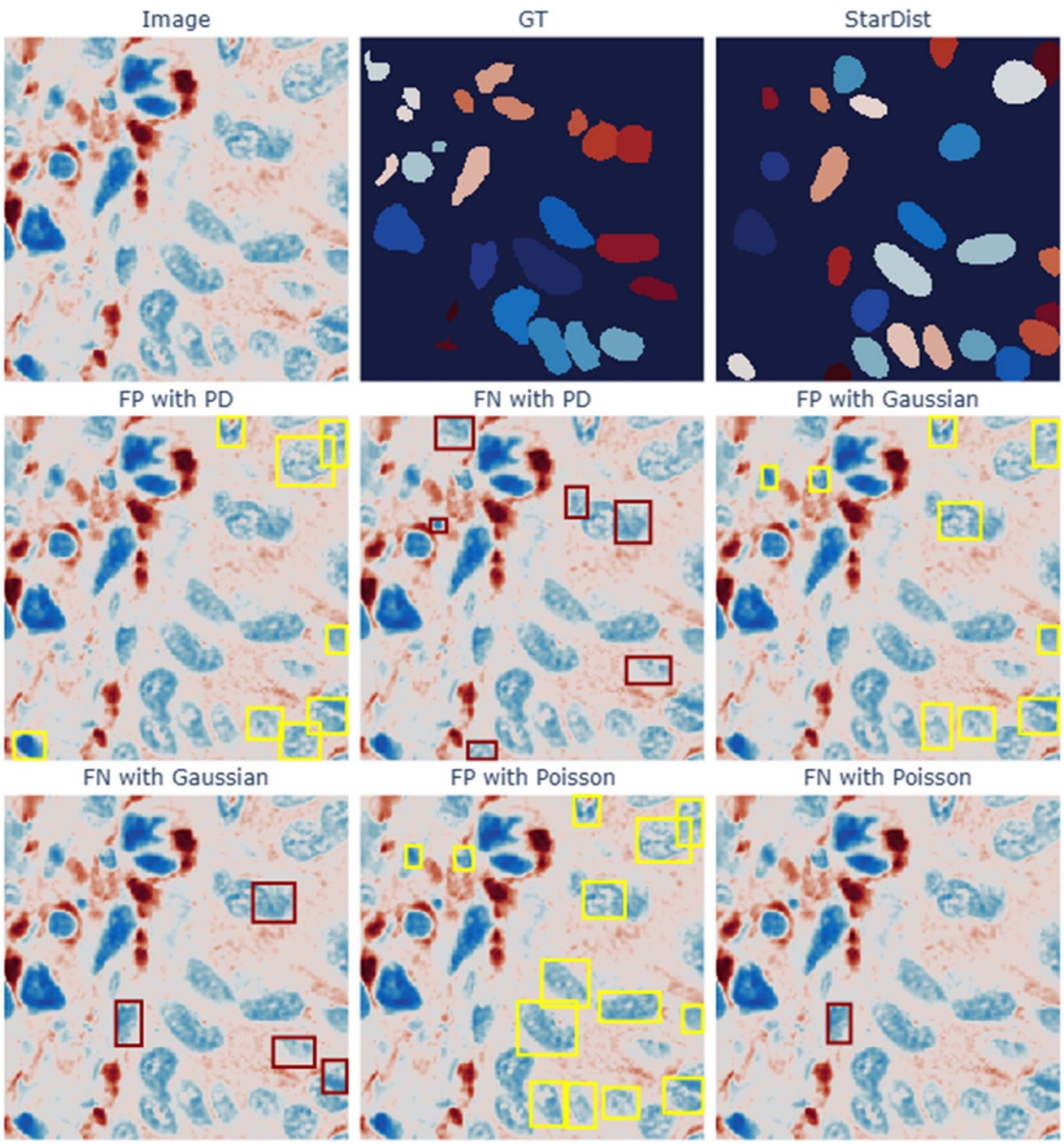

Figure 14. PD, Gaussian, and Poisson defences.

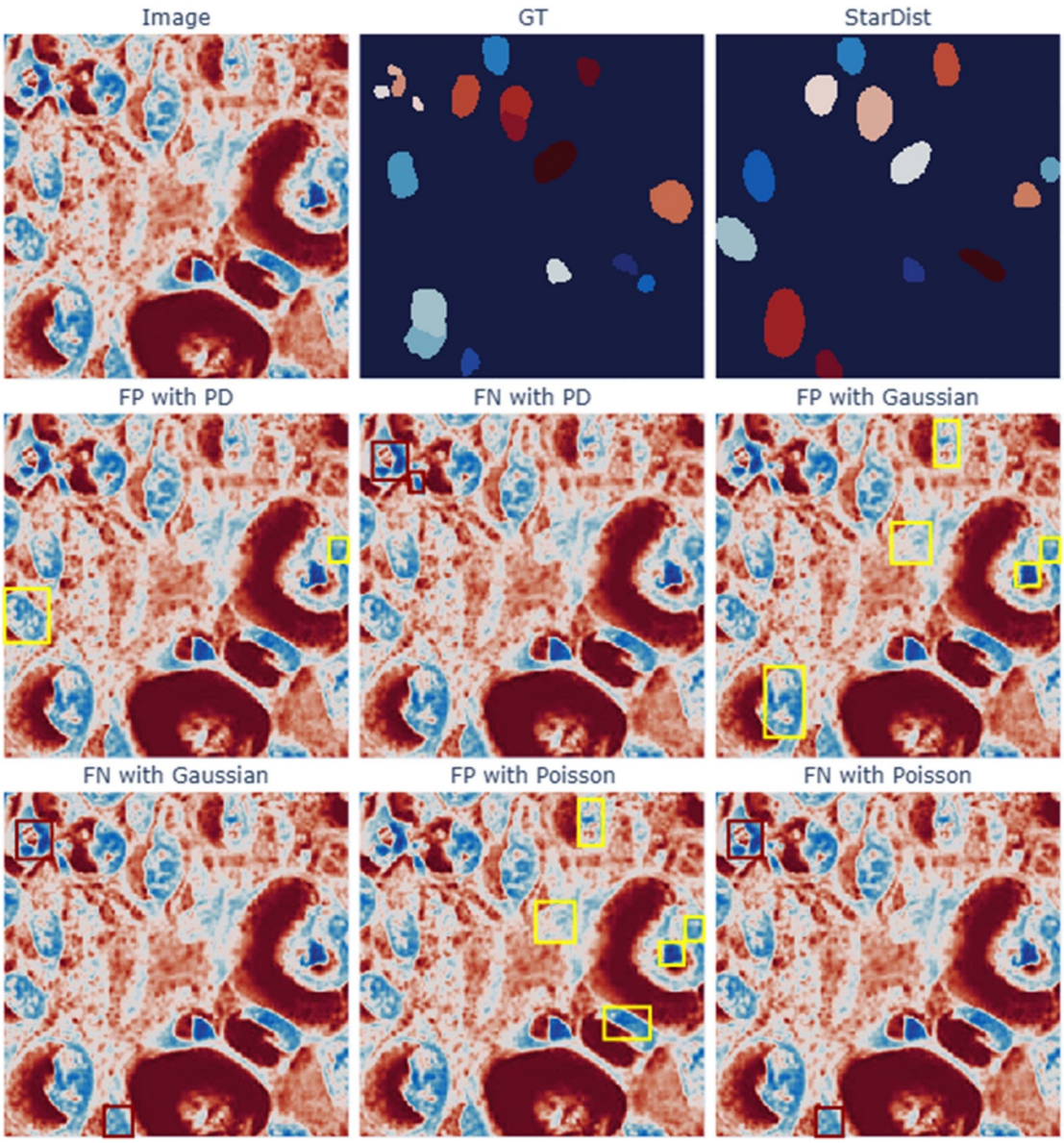

Figure 15. PD, Gaussian, and Poisson defences.

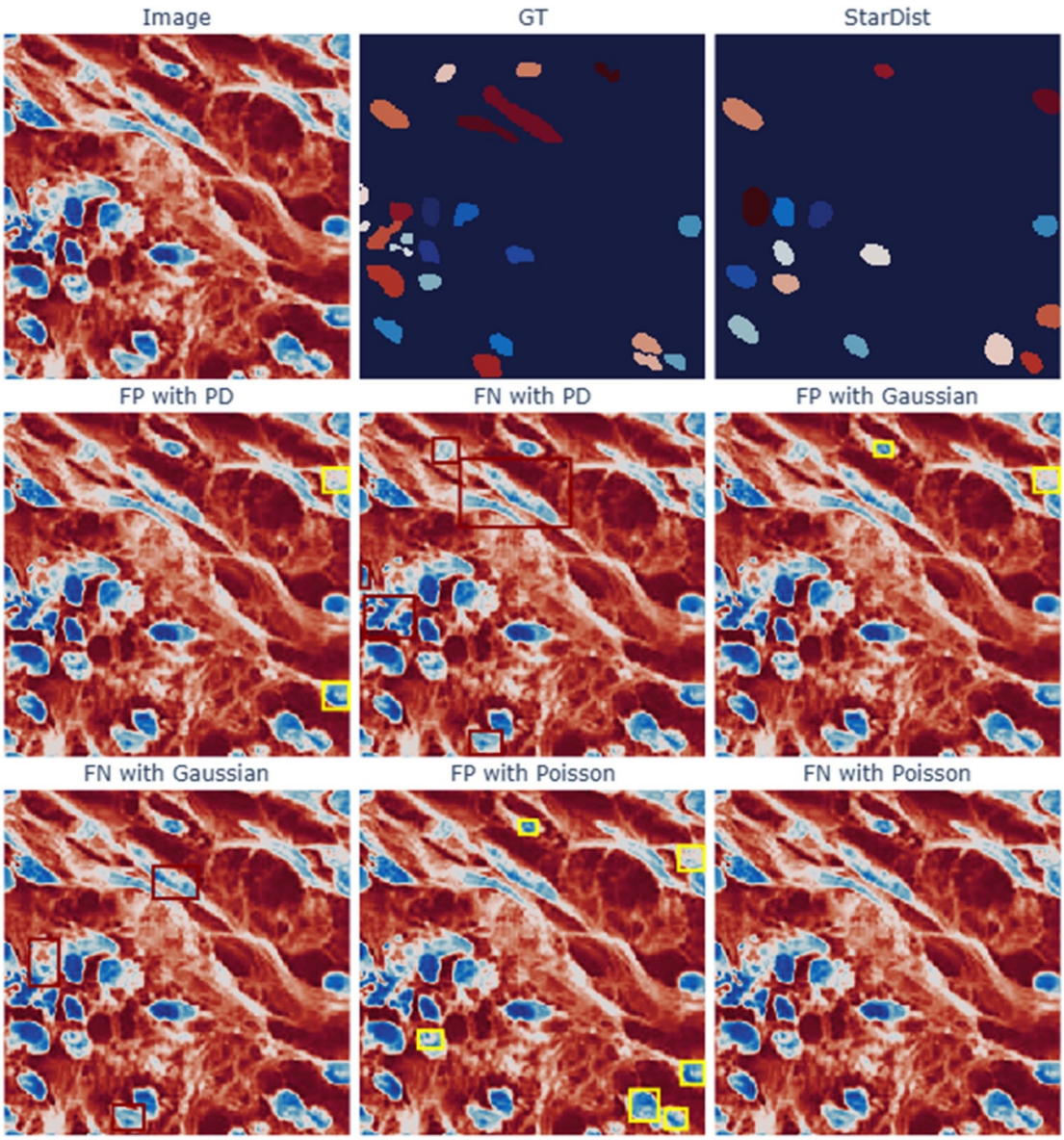

Figure 16. PD, Gaussian, and Poisson defences.

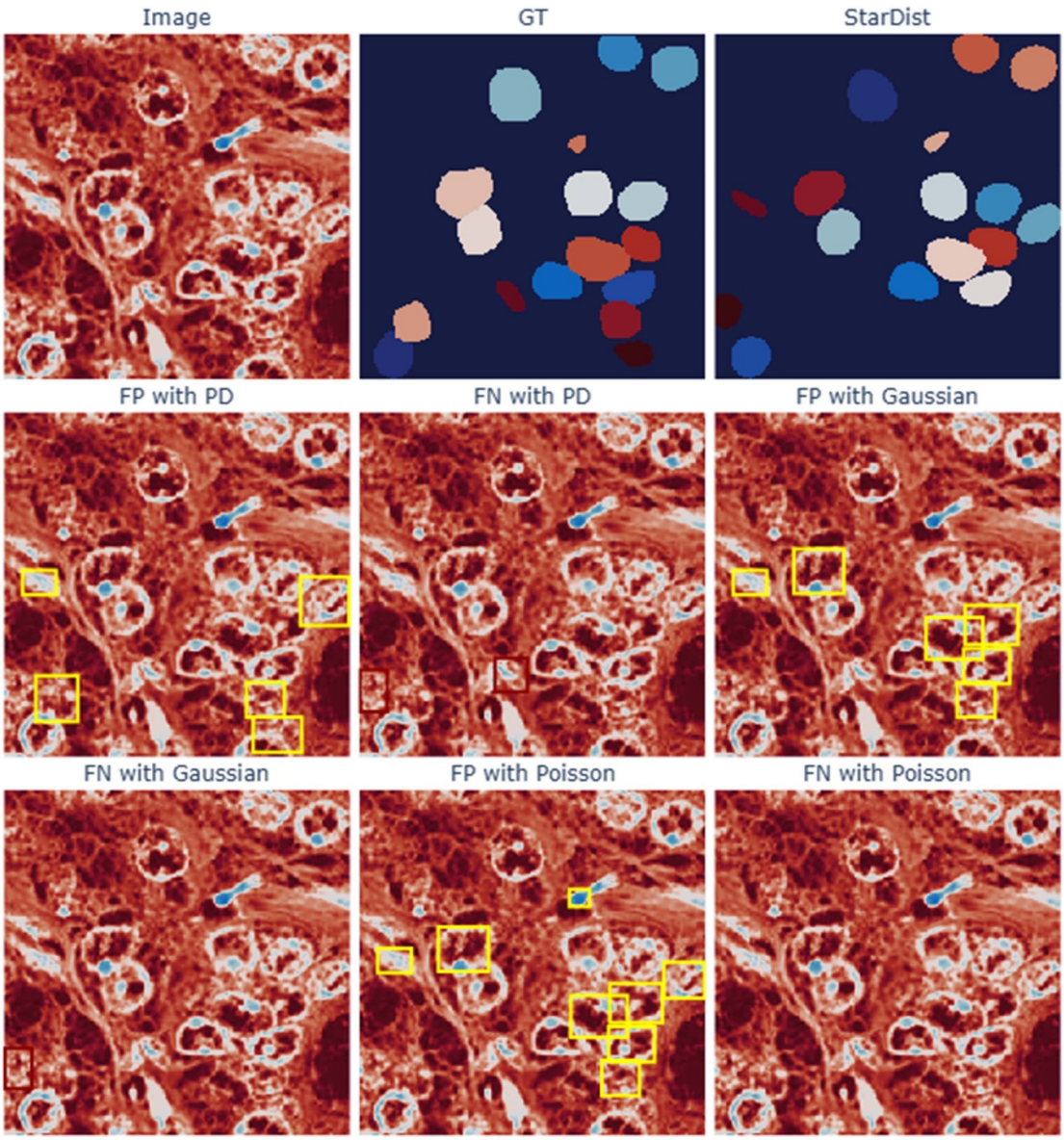

Figure 17. PD, Gaussian, and Poisson defences.

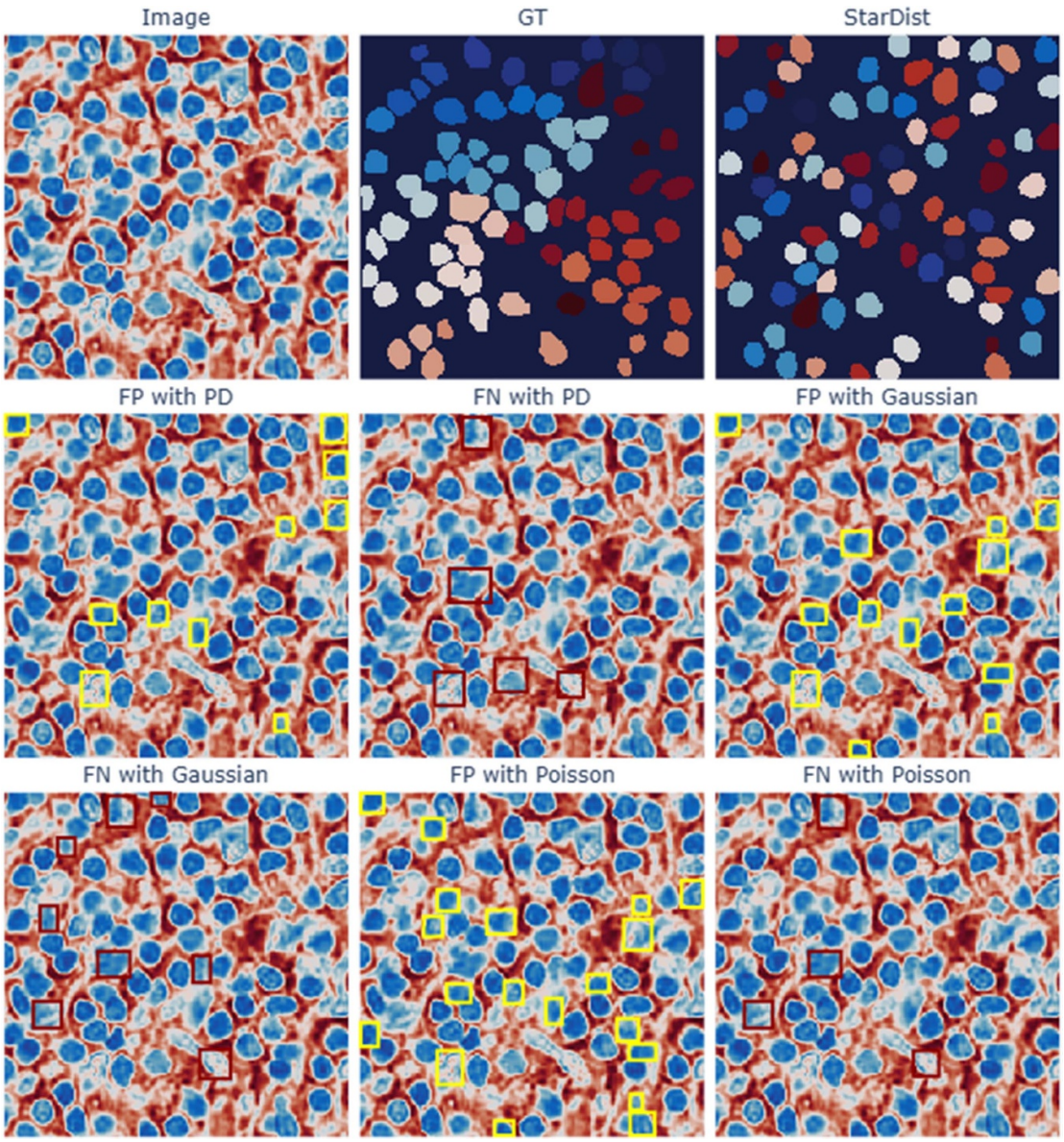

Figure 18. PD, Gaussian, and Poisson defences.

5. Conclusion

We theoretically discussed the inherent instability phenomenon of image segmentation tasks and beyond. Consequently, we proposed to generate PDs, mimicking noise distributions that defend the (StarDist) segmentation from delivering segmentation artifacts. We empirically demonstrated the efficacy of the PDs on the training data set and, especially, when using them as perturbations for detecting potential wrong segments for unseen (test) images. In comparison to classic Gaussian and Poisson noise perturbations, the performance of the PD-based artifact detection suggests to be much more precise, enabling quick manual re-adjustment of the segmentation output and cross-checking of the manually segmented training dataset.

The performance of the PD-detection on the StarDist dataset (Booz 2018, Schmidt *et al* 2018a) promises broad capability of the approach, defending ML and classic image-processing tools from false predictions. Generating *universal* PDs that generalize across more datasets and models is the aim of our future work.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://doi.org/10.5281/zenodo.15 901 527.

ORCID iDs

Saiyam B Jain 0009-0003-7649-4836
Michael Hecht 0000-0001-9214-8253

References

André P and Wanner S 2019 hypopy - A hyper-parameter-toolbox (available at: https://github.com/MIC-DKFZ/Hyppopy)
Andriasyan V, Yakimovich A, Petkidis A, Georgi F, Witte R, Puntener D and Greber U F 2021 Microscopy deep learning predicts virus infections and reveals mechanics of lytic-infected cells *Iscience* **24** 102543
Antun V, Gottschling N M, Hansen A C and Adcock B 2021 Deep learning in scientific computing: understanding the instability mystery *SIAM NEWS MARCH* **54** (available at: www.siam.org/publications/siam-news/articles/deep-learning-in-scientific-computing-understanding-the-instability-mystery/)
Antun V, Renna F, Poon C, Adcock B and Hansen A C 2020 On instabilities of deep learning in image reconstruction and the potential costs of AI *Proc. Natl Acad. Sci.* **117** 30088–95
Balda E R, Behboodi A and Mathar R 2020 Adversarial examples in deep neural networks: an overview *Deep Learning: Algorithms and Applications* pp 31–65
Batenburg K J and Sijbers J 2008 Optimal threshold selection for tomogram segmentation by projection distance minimization *IEEE Trans. Med. Imaging* **28** 676–86
Booz Allen H 2018 Dataset DSB2018 (available at: www.kaggle.com/competitions/data-science-bowl-2018)
Bos L and Levenberg N 2018 Bernstein-Walsh theory associated to convex bodies and applications to multivariate approximation theory *Comput. Methods Funct. Theory* **18** 361–88
Caselles V, Kimmel R and Sapiro G 1997 Geodesic active contours *Int. J. Comput. Vis.* **22** 61–79
Chakravorty P 2018 What is a signal?[lecture notes] *IEEE Signal Process. Mag.* **35** 175–7
Chen J, Jordan M I and Wainwright M J 2020 Hopskipjumpattack: a query-efficient decision-based attack *2020 IEEE Symp. on Security and Privacy (sp)* (IEEE) pp 1277–94
Dian R, Li S and Kang X 2020 Regularizing hyperspectral and multispectral image fusion by CNN denoiser *IEEE Trans. Neural Netw. Learn. Syst.* **32** 1124–35
Fisch D, Yakimovich A, Clough B, Mercer J and Frickel E-M 2020 Image-based quantitation of host cell–toxoplasma gondii interplay using hrman: a host response to microbe analysis pipeline *Toxoplasma Gondii: Methods and Protocols* pp 411–33
Galimov E and Yakimovich A 2022 A tandem segmentation-classification approach for the localization of morphological predictors of c. elegans lifespan and motility *Aging* **14** 1665
Gonzales R C and Wintz P 1987 *Digital Image Processing* (Addison-Wesley Longman Publishing Co., Inc.)
Gonzalez R C 2009 *Digital Image Processing* (Pearson Education India)
Goodfellow I J, Shlens J and Szegedy C 2015 Explaining and harnessing adversarial examples (arXiv:1412.6572)
Gottschling N M, Antun V, Adcock B and Hansen A C 2020 The troublesome kernel: why deep learning for inverse problems is typically unstable (arXiv:2001.01258)
Grady L 2006 Random walks for image segmentation *IEEE Trans. Pattern Anal. Mach. Intell.* **28** 1768–83
Hecht M, Cheeseman B L, Hoffmann K B and Sbalzarini I F 2017 A quadratic-time algorithm for general multivariate polynomial interpolation (arXiv:1710.10846)
Hecht M, Gonciarz K, Michelfeit J, Sivkin V and Sbalzarini I F 2020 Multivariate interpolation in unisolvent nodes–lifting the curse of dimensionality (arXiv:2010.10824)
Hecht M, Hoffmann K B, Cheeseman B L and Sbalzarini I F 2018 Multivariate Newton interpolation (arXiv:1812.04256)
Hecht M, Hofmann P A, Wicaksono D, Acosta U H, Gonciarz K, Kissinger J, Sivkin V and Sbalzarini I F 2025 Multivariate Newton interpolation in downward closed spaces reaches the optimal geometric approximation rates for Bos–Levenberg–Trefethen functions (arXiv:2504.17899)

Hecht M and Sbalzarini I F 2018 Fast interpolation and Fourier transform in high-dimensional spaces *Intelligent Computing. Proc. 2018 IEEE Computing Conf.* (*Advances in Intelligent Systems and Computing*) vol 2,857, ed K Arai, S Kapoor and R Bhatia (Springer) pp 53–75

Huang Y, Würfl T, Breininger K, Liu L, Lauritsch G and Maier A 2018 Some investigations on robustness of deep learning in limited angle tomography *Int. Conf. on Medical Image Computing and Computer-Assisted Intervention* (Springer) pp 145–53

Jain S B, Zongru S, Veettil S K and Hecht M 2022 Adversarial attacks for machine learning denoisers and how to resist them *Emerging Topics in Artificial Intelligence (ETAI) 2022* vol 12204 (SPIE) p 1220402

Kimmel R 2003 Fast edge integration *Geometric Level Set Methods in Imaging, Vision and Graphics*' (Springer) pp 59–77

Kirillov A, He K, Girshick R, Rother C and Dollár P 2019 Panoptic segmentation *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition* pp 9404–13

Kumar N, Verma R, Sharma S, Bhargava S, Vahadane A and Sethi A 2017 A dataset and a technique for generalized nuclear segmentation for computational pathology *IEEE Trans. Med. Imaging* **36** 1550–60

Leja F 1957 Sur certaines suites liées aux ensembles plans et leur application à la représentation conforme *Ann. Polon. Math.* **1** 8–13

Lord N A, Mueller R and Bertinetto L 2022 Attacking deep networks with surrogate-based adversarial black-box methods is easy (arXiv:2203.08725)

Manning C D 2008 *Introduction to Information Retrieval* (Syngress Publishing)

Mechea D 2019 Panoptic segmentation – the panoptic quality metric (available at: https://medium.com/@danielmechea/panoptic-segmentation-the-panoptic-quality-metric-d69a6c3ace30)

Mobahi H, Rao S R, Yang A Y, Sastry S S and Ma Y 2011 Segmentation of natural images by texture and boundary compression *Int. J. Comput. Vis.* **95** 86–98

Nock R and Nielsen F 2004 Statistical region merging *IEEE Trans. Pattern Anal. Mach. Intell.* **26** 1452–8

Padilla R, Netto S L and Da Silva E A 2020 A survey on performance metrics for object-detection algorithms *2020 Int. Conf. on Systems, Signals and Image Processing (IWSSIP)* (IEEE) pp 237–42

Papernot N, McDaniel P, Goodfellow I, Jha S, Celik Z B and Swami A 2017 Practical black-box attacks against machine learning *Proc. of the 2017 ACM on Asia Conf. on Computer and Communications Security* pp 506–19

Ronneberger O, Fischer P and Brox T 2015 U-net: convolutional networks for biomedical image segmentation *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th Int. Conf., (Munich, Germany, 5 October–9 October 2015 Proc., Part III)* vol 18 (Springer) pp 234–41

Schmidt U, Weigert M, Broaddus C and Myers G 2018a Cell detection with star-convex polygons *Medical Image Computing and Computer Assisted Intervention - MICCAI 2018 - 21st Int. Conf., (Granada, Spain, 16 September–20 September 2018, Proc., Part II)* pp 265–73

Schmidt U, Weigert M, Broaddus C and Myers G 2018b StarDist - cell detection with star-convex polygons (available at: https://github.com/stardist/stardist)

Shapiro L G *et al* 2001 *Computer Vision* vol 3 (Prentice Hall)

Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V and Rabinovich A 2014 Going deeper with convolutions *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* pp 1–9

Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I and Fergus R 2014 Intriguing properties of neural networks *2nd Int. Conf. on Learning Representations, ICLR 2014*

Taha A A and Hanbury A 2015 Metrics for evaluating 3d medical image segmentation: analysis, selection and tool *BMC Med. Imaging* **15** 1–28

Trefethen L N 2017 Multivariate polynomial approximation in the hypercube *Proc. Am. Math. Soc.* **145** 4837–44

Trefethen L N 2019 *Approximation Theory and Approximation Practice* vol 164 (SIAM)

Veettil S K T, Zheng Y, Acosta U H, Wicaksono D and Hecht M 2022 Multivariate polynomial regression of Euclidean degree extends the stability for fast approximations of trefethen functions (arXiv:2212.11706)

Volpe G et al 2023 Roadmap on deep learning for microscopy (arXiv:2303.03793)

Wicaksono D C, Hernandez Acosta U, Thekke Veettil S K, Michelfeit J and Hecht M 2023 Minterpy - multivariate polynomial interpolation (version 0.2.0-alpha) *Rodare: GitHub* (available at: https://github.com/casus/minterpy/)

Yakimovich A, Huttunen M, Samolej J, Clough B, Yoshida N, Mostowy S, Frickel E-M and Mercer J 2020 Mimicry embedding facilitates advanced neural network training for image-based pathogen detection *Msphere* **5** e00836–20