A framework for building robust deep-learning models against out-of-focus artifact in whole-slide images

Qinle Ba[1], and Jim F. Martin[1]

[1]Digital Pathology, Roche Tissue Diagnostics, Santa Clara, CA, USA. E-mail: qinle.ba@roche.com

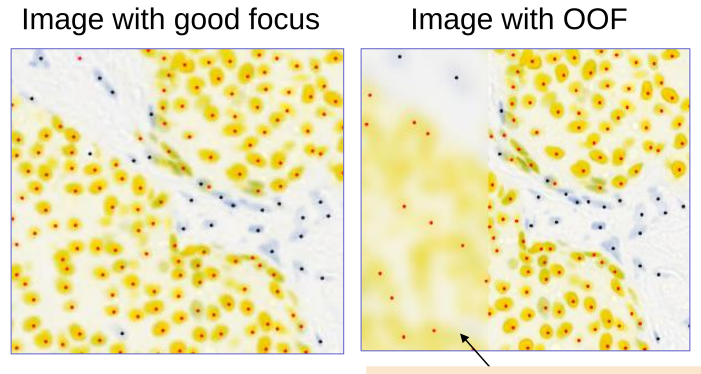
Overview

During tissue processing and slide scanning, artifacts such as tissue folds or coverslip defect can be easily introduced. These artifacts may cause a region of the scan to be out-of-focus (OOF), which in turn adversely impacts the performance of computational algorithms. A typical strategy to avoid such model prediction errors is to use a manual procedure to identify the artifact regions so they can be excluded from digital pathology (DP) analysis. However, such subjectivity not only causes inconsistent quality control (QC) results across specimens and analyzers, but also leads to a mismatch between analyzers' perception of blurriness and the blur levels that can considerably degrade DP algorithm performance.

To address this issue, we developed a computational framework for building robust DP algorithms against OOF artifact for cell phenotype classification.

Challenge of automated DP algorithms in OOF regions

Example: Cell detection and phenotype classification in IHC

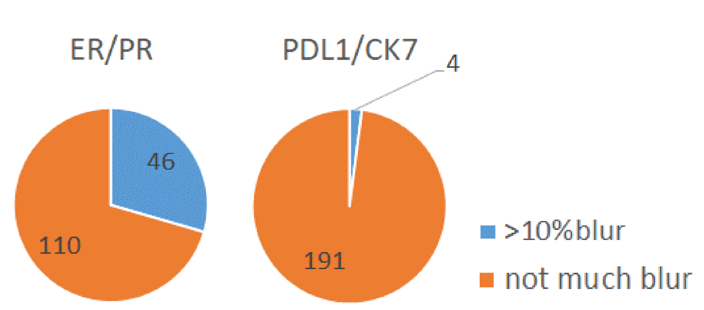


Dabsyl Estrogen (ER) IHC
Red dot: ER positive
Black dot: ER negative

Compromised algorithm performance

The prevalence of whole-slide images with OOF in two duplex IHC cohorts

WSI blur analysis of ER/PR IHC (breast cancer) & PDL1/CK7 IHC (lung cancer)



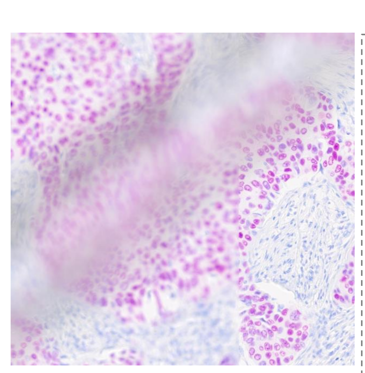
With a gradient-based blur detector, identified the number of slides with considerable (>10% tissue regions) OOF tissue regions

Current solution: reliable for algorithm but inconvenient for users

Manual Slide QC → Cell classification → Phenotype scoring

Deficiencies of the current solution

- Manually exclude these regions: **time-consuming** and **inconvenient**
- Qualitative and **subjective** determination for OOF threshold



Is this image analyzable?

- Marker specific decision from pathologists: 60-70% analyzable
- Non-pathologist or DP algorithm may identify ~30% analyzable

- Inconsistent QC results across specimens and analyzers
- Mismatch between subjective QC criteria vs. blur levels that causes considerable errors for analysis

Methods

Proposed solutions for building robust DP algorithms against OOF artifacts

Automatic slide QC → Cell classification → Phenotype scoring

Data augmentation

Automated QC for high-levels of OOF:
- Better user interaction
- Higher confidence in diagnosis
- Higher analysis efficiency

Data augmentation for low-levels of OOF:
- More robust model performance in blurry regions below the blur threshold for QC

Results

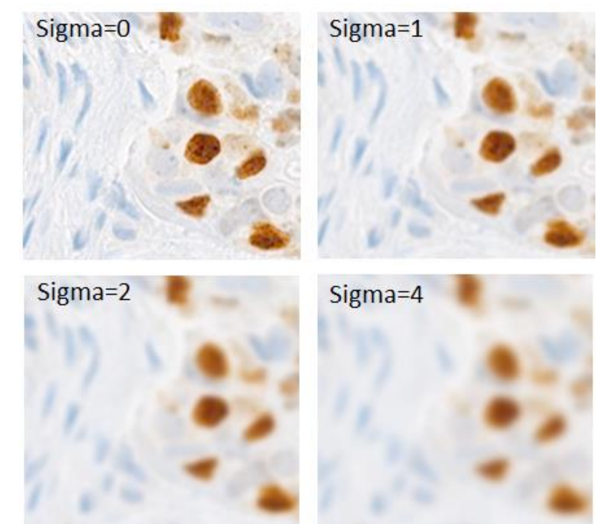
Automated whole-slide-image (WSI) OOF detection with deep-learning (DL)

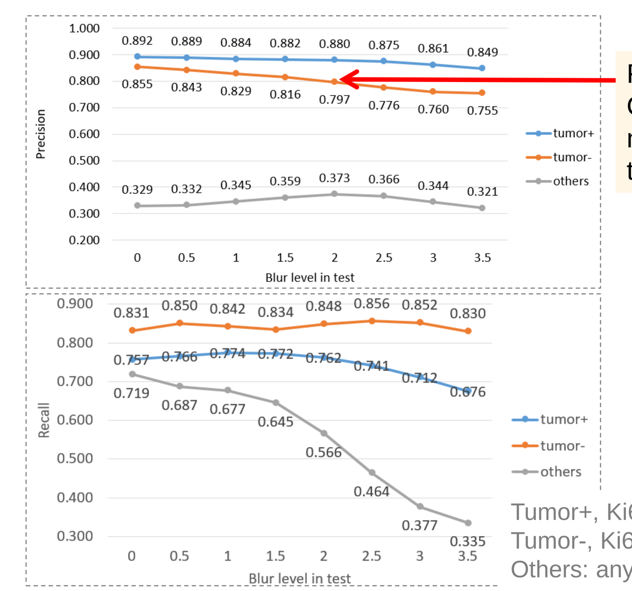
A framework for ground-truth (GT) generation to detect OOF image regions

Question: What degree of blur is acceptable?

- **Proposed solution 1: Find the blur level(s) that cause issues for <u>specific</u> downstream algorithms**
  How? <u>Model robustness analysis:</u> Evaluate performance of trained models on blurry images
  Experiment: Ki67 cell detection and phenotype classification

Artificially blurred images (Ki67-DAB IHC)





For an 80% cutoff: QC for any regions more blurry than this blurriness level

Tumor+, Ki67-positive tumor cells
Tumor-, Ki67-negative tumor cells
Others: any other cells

- **Proposed solution 2: Leverage the volume-scan features of Roche DP scanners**

Determine DP algorithm of interest → Use Solution 1 + volume scans to identify a blur range for QC → Generate GT → <u>Optional</u>: modify GT to reach readers' consensus

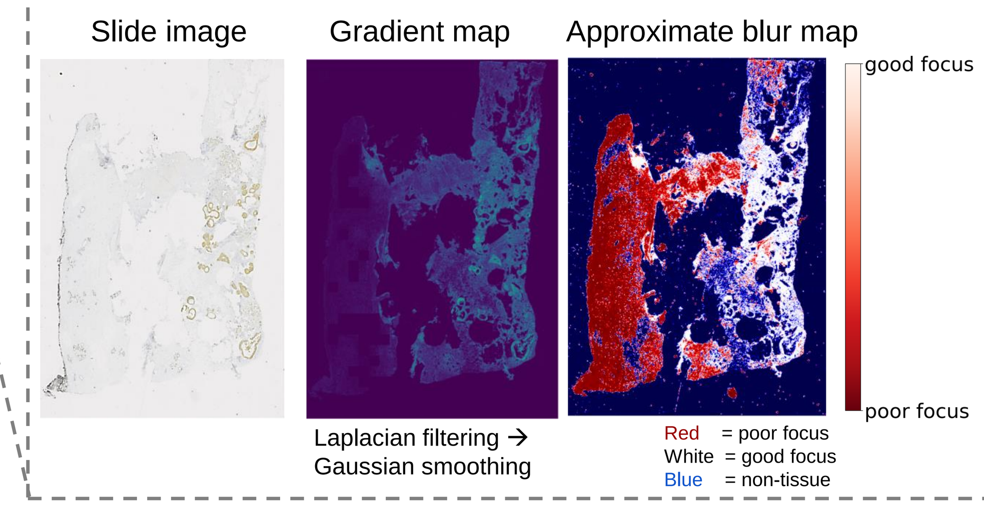
- **Volume-scan mode**:
  Blurry ← Non-blurry at the nominal focal plane → Blurry

- Given <u>a fixed assay and a developed DP algorithm</u>:
  → Rescan GT slides in "volume scan" mode
    ◆ Non-nominal-focus scan planes with 1 micron intervals
  → Determine where the DP algorithm accuracy becomes inadequate.
  → Retrain (transfer learning) this task's DL architecture to detect the blurriness beyond accuracy tolerance for that DP algorithm.
- Development time for such a custom blur-detection model could be as short as **six weeks** after the DP algorithm is ready.

Workflow of accelerated GT generation

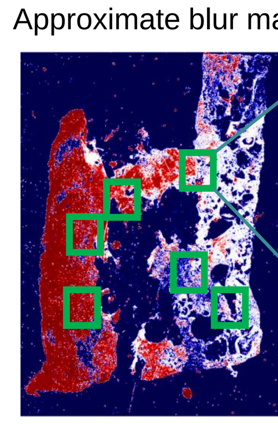
Step 1: Approximate blur detection with image processing → Step 2: Automated selection of tiles with blurs → Step 3: DL-based semi-automated correction of GT masks

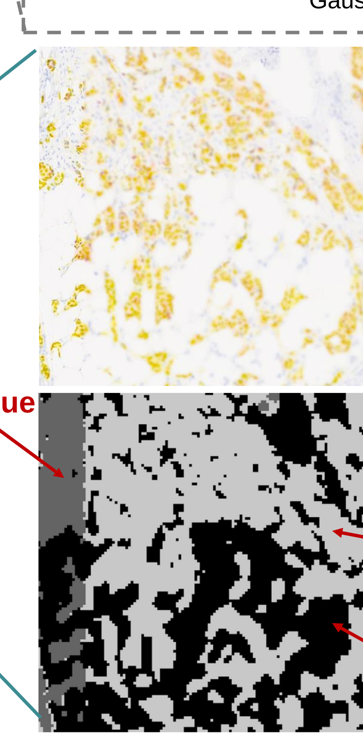
Step 1:
Gradient-based blur detector identified the slides with considerable (>10% tissue regions) OOF tissue regions. Combined with a tissue mask, we identified approximate blurry regions within the tissue region.



Step 2:
Image tiles with various amounts of blur were automatically collected along with the corresponding approximate GT.

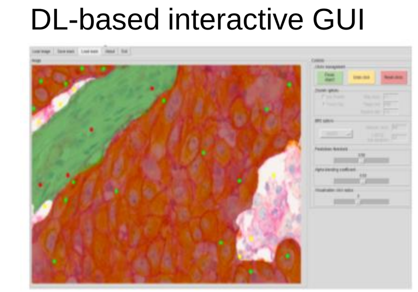
Approximate blur map





Step 3:
Approximate GT blur masks were corrected using DL-based interactive GUI (*Ref2*) to generate masks

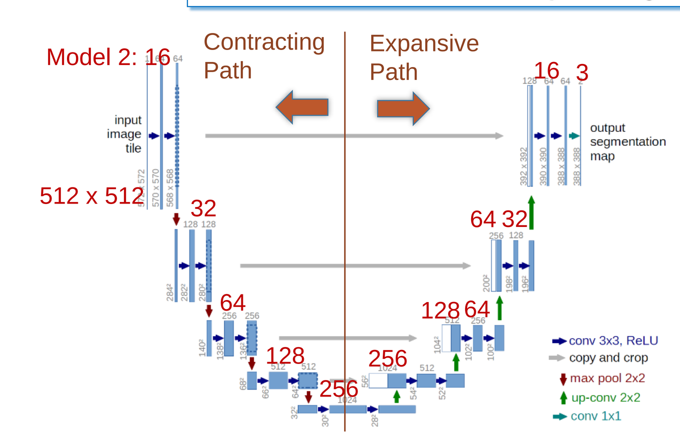
DL-based interactive GUI



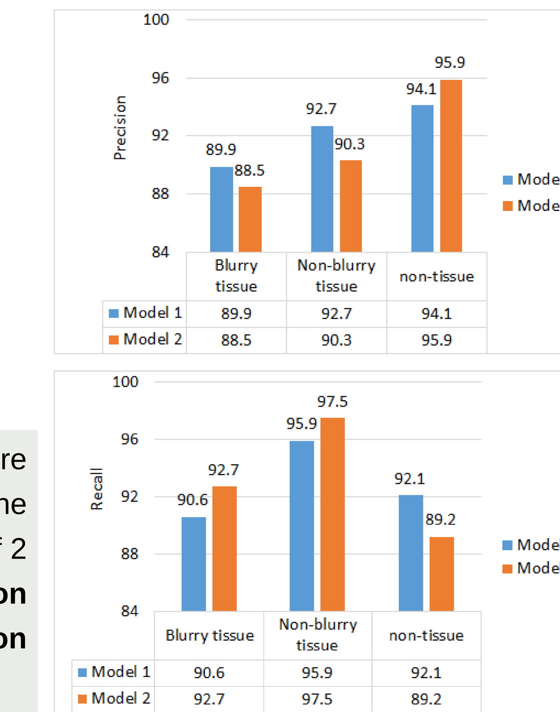
Results

Automated WSI OOF detection with deep-learning (DL)

OOF identification by image segmentation: Assay-agnostic and generalizable



Compared two modified UNet models, where we reduced the number of channels in the intermediate convolutional layer by a factor of 2 and 4, resulting in **Model 1 (7.76 million parameters)** and **Model 2 (1.94 million parameters)**, respectively.

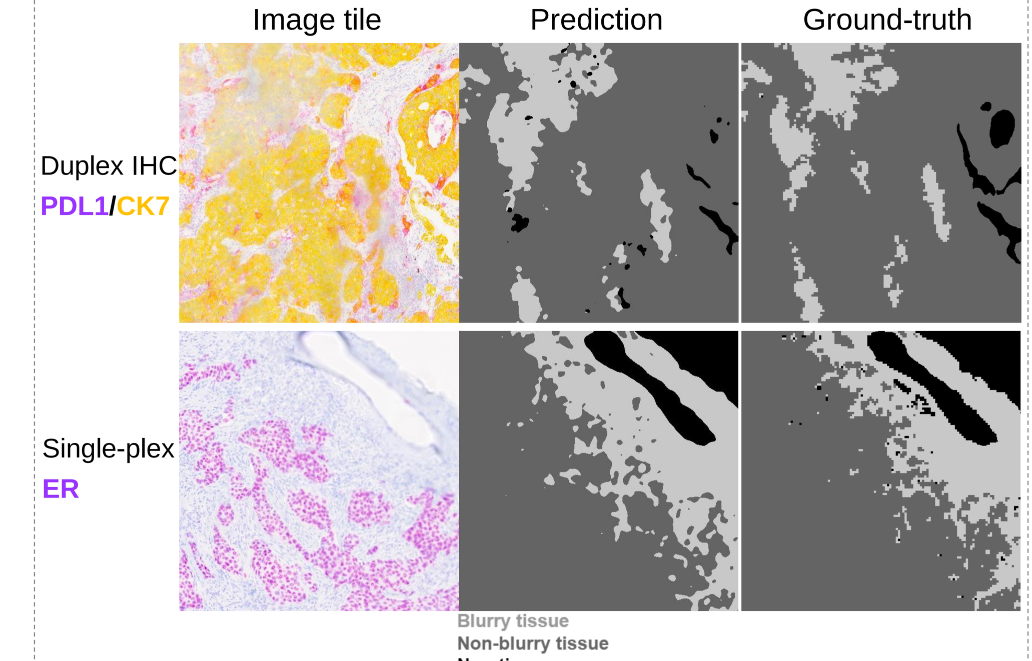


Precision	Blurry tissue	Non-blurry tissue	non-tissue
Model 1	89.9	92.7	94.1
Model 2	88.5	90.3	95.9

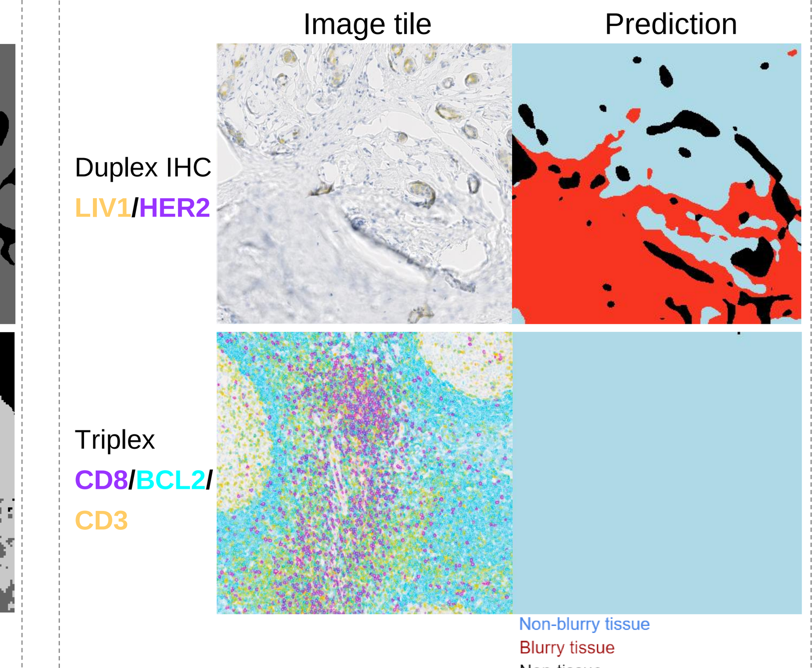
Recall	Blurry tissue	Non-blurry tissue	non-tissue
Model 1	90.6	95.9	92.1
Model 2	92.7	97.5	89.2

- To maximize generalizability to unseen assays, model inputs are **gray-scale** image tile
- To emphasize the gradient features that indicates image edges, **gradient map** is concatenated to input image as model input.
- Training data:
  - Single-plex & duplex from two assays: PDL1/CK7, ER/PR
  - Chromogens: Dabsyl, Tamra
  - Data split:
    - Train: 462 tiles
    - Validation: 246 tiles
    - Held-out test: 270 tiles
    - Additional testing:100 WSIs

Quantitative assessment

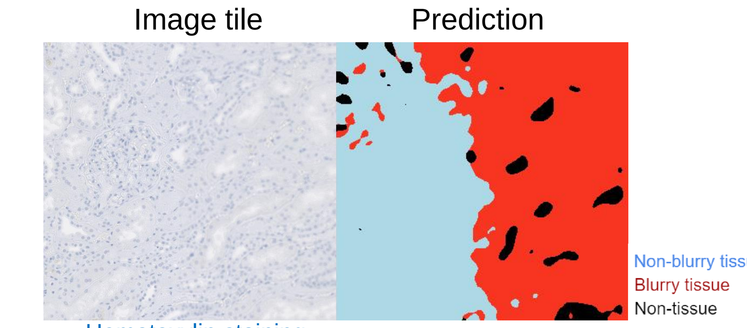


Generalizability to unseen assays

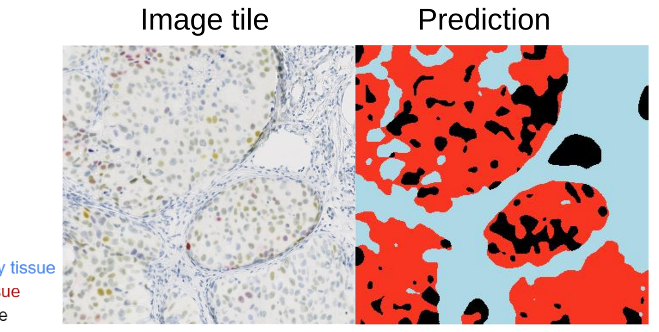


Generalizability to unseen tissue types

Train: breast & lung samples --> Test: kidney samples



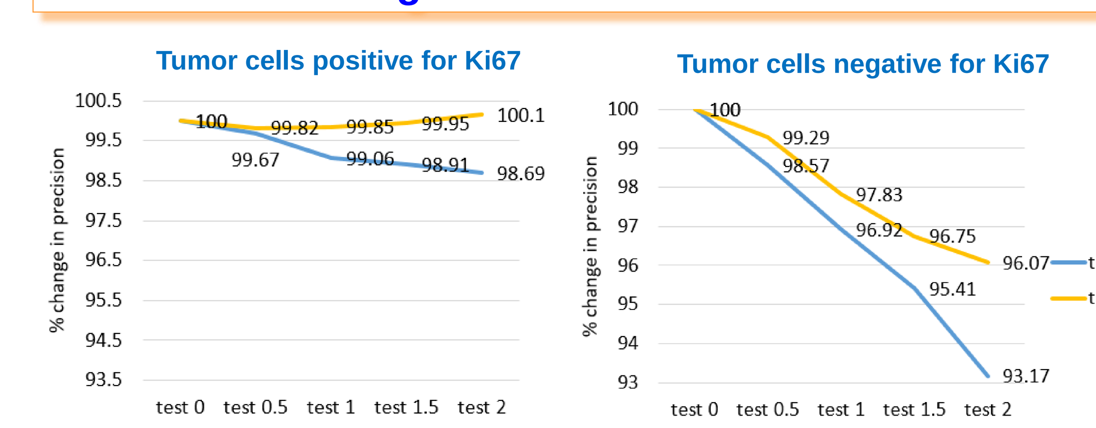
Example ambiguous case: calling for a more controlled approach for determining QC criteria



Need the proposed Solution 2!

Scoring % of blurry tissue
Prediction: 55%
Pathologist: 0%
Reader 1: 40%
Reader 2: 0%
Reader 3: 60%

Robust models against low to medium levels of blurriness below QC threshold



- Train 0: training with no data augmentation for blurriness
- Train 1.5: training with data augmentation (randomly add blur with a sigma from [0,1] for Gaussian filtering)
- Test 0: test with original images
- Test 0.5 (1,1.5,2): test with artificially blurred images of various sigma levels

- **Approach:**
  - **Data augmentation** for model training by adding low to medium levels of blurriness to training images.
- **Experiment**:
  - Phenotype classification and cell detection in Ki67-DAB IHC images
  - Train: augmented image data
  - Test: artificially blurred images

References

1. O. Ronneberger, P. Fischer, and T. Brox. "U-net: Convolutional networks for biomedical image segmentation." International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015.1
2. Sofiiuk, Konstantin, Ilia A. Petrov, and Anton Konushin. "Reviving Iterative Training with Mask Guidance for Interac-tive Segmentation." arXiv preprint arXiv:2102.06583 (2021).

Acknowledgements

We thank members from the Roche DP Imaging team for insightful discussions and suggestions, especially Raghavan Venugopal, Xingwei Wang, Xiangxue Wang, Yao Nie and Christoph Guetter. We thank multiplex IHC project team from Roche Assay and Reagent Development for providing image data and insightful discussions.