scientific reports

Check for updates

OPEN

ArtSeg—Artifact segmentation and removal in brightfield cell microscopy images without manual pixel-level annotations

Mohammed A. S. Ali[1,5], Kaspar Hollo[1,5], Tõnis Laasfeld[2,5], Jane Torp[2], Maris-Johanna Tahk[2], Ago Rinken[2], Kaupo Palo[3], Leopold Parts[1,4]✉ & Dmytro Fishman[1]✉

Brightfield cell microscopy is a foundational tool in life sciences. The acquired images are prone to contain visual artifacts that hinder downstream analysis, and automatically removing them is therefore of great practical interest. Deep convolutional neural networks are state-of-the-art for image segmentation, but require pixel-level annotations, which are time-consuming to produce. Here, we propose ScoreCAM-U-Net, a pipeline to segment artifactual regions in brightfield images with limited user input. The model is trained using only image-level labels, so the process is faster by orders of magnitude compared to pixel-level annotation, but without substantially sacrificing the segmentation performance. We confirm that artifacts indeed exist with different shapes and sizes in three different brightfield microscopy image datasets, and distort downstream analyses such as nuclei segmentation, morphometry and fluorescence intensity quantification. We then demonstrate that our automated artifact removal ameliorates this problem. Such rapid cleaning of acquired images using the power of deep learning models is likely to become a standard step for all large scale microscopy experiments.

Advanced microscopes extract rich visual information from biological samples at scales from individual atoms to cells and tissues. Among the different imaging modalities, brightfield illumination with transmitted light is the simplest to acquire while avoiding damaging the sample[1]. The usefulness of this technology has led to its widespread adoption[2–4], and thereby to a dramatic increase in the volumes of microscopy data. However, the automated analysis techniques required to extract information at scale are often hindered by the artifacts present in the images[5,6]. Detecting and neutralizing the impact of such problematic image areas would provide more accurate results from experiments[3], making artifact segmentation an important, albeit overlooked, research area in cell biology and beyond[7,8].

While any signal that deviates from the reflection of expectation can be considered artifactual[9], the common source of artifacts in cell microscopy is the introduction of foreign objects during sample preparation. These include dust, fragments of dead cells, bacterial contamination, reagent impurities, defects on the light path, etc. We focus on detecting these low-level anomalies[8,10] in brightfield microscopy and use the term *artifact* with this meaning. Manually identifying all the affected images or image regions is a time-consuming solution to this problem[11,12]. A common alternative approach for large datasets is computer-aided delineation and removal of the artifacts, but two complexities make this task challenging. First, artifacts appear stochastically in microscopy images leading to sparse data. Second, artifact characteristics, such as morphology and texture, are often very heterogeneous and hence are challenging to define.This means, it is unfeasible to comprehensively collect representative examples of all possible artifact types, which renders computational modeling difficult.

Deep learning has emerged as the favored solution to artifact detection[7,8]. While strongly supervised convolutional neural networks (CNN) such as U-Net[13–17] are state-of-the-art for most computer vision tasks, they cannot overcome some challenges that artifact detection brings[7]. A major bottleneck for the strongly supervised

[1]Department of Computer Science, University of Tartu, Narva mnt 18, 51009 Tartu, Estonia. [2]Institute of Chemistry, University of Tartu, Ravila 14a, 50411 Tartu, Estonia. [3]PerkinElmer Cellular Technologies Germany GmbH, Hamburg, Germany. [4]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton CB10 1SA, Cambridgeshire, UK. [5]These authors contributed equally: Mohammed A. S. Ali, Kaspar Hollo and Tõnis Laasfeld. ✉email: leopold.parts@ut.ee; dmytro.fishman@ut.ee

deep learning methods is their requirement of pixel-level annotation, which is time-consuming, and requires substantial expertise. As an alternative, weakly supervised techniques such as ScoreCAM[18], which involve only image-level labeling, greatly reduce the time needed to prepare the dataset. In particular, generative autoencoder-based models[19–24] are trained to reconstruct artifact-free images and report artifacts on test images as areas with large reconstruction error. Alternatively, one-class classification approaches[25–27] train a classifier on artifact-free images and report artifacts as images with a low probability of belonging to this clean class. Combining the performance advantages of the strongly supervised methods and the convenience of image-level annotations would therefore be of great practical interest and impact.

In this work, we make the following key contributions: (1) we empirically confirm the prevalence of artifacts in brightfield microscopy images; (2) test a range of existing approaches in domains other than microscopy for the artifact segmentation task, and find none of them is accurate for practical use; (3) combine the merits of weakly and strongly supervised methods for artifact segmentation from brightfield cell microscopy images using only image-level annotations. To our knowledge, this is the first attempt to segment artifacts in microscopy images in a weakly supervised way. We introduce ScoreCAM-U-Net, a model that combines the informative pixel-level[28] and cheap-to-generate image-level[18] annotation schemes, and accurately detects artifacts in held-out samples. As training is performed using only image-level labels, generating training data is orders of magnitude cheaper, but without substantially sacrificing performance compared to pixel-level data. (4) We study the impact of removing artifacts on different downstream applications. We demonstrate that artifacts in microscopy images confound downstream analyses such as nuclei segmentation or quantification of ligand binding, and that ScoreCAM-U-Net successfully overcomes these problems.

Methods

To delineate artifacts from brightfield microscopy images, we introduce ScoreCAM-U-Net, a method that uses image-level annotations as input for training, and produces artifact segmentations as an output. We compare the performance of our pipeline with a strongly supervised counterpart trained on pixel-level annotations as well as with state-of-art models that are trained using image-level labeling on three different datasets.

Datasets. We chose three datasets for this study to cover multiple common variables in experimental design to better assess the generalizability of the results. Overall, the datasets cover nine different cell lines, fixed and live cells, two different plate formats and two microscopes. The datasets provenances have been described previously[3,4,29,30] and we briefly describe their most important properties here.

Seven cell lines dataset. Seven types of cells including human cells from breast cancer (MCF7), fibrosarcoma (HT1080), cervical cancer (HeLa), hepatocellular carcinoma (HepG2), alveolar basal epithelial (A549), dog cells from kidney tissue (MDCK), and mouse embryonic fibroblast cells (NIH3T3) were seeded in Collagen type 1-coated CellCarrier-384 Ultra Microplates (PerkinElmer, Waltham, MA; cat. 6057700). The cells were stained with 10 µg/ml Hoechst 33342 (Thermo Fisher, Waltham, MA; cat. H3570) and fixed in formaldehyde (Sigma, St. Louis, MO; cat. 252549). A 20 × water immersion objective was used to acquire images on an Opera Phenix high-content screening system (PerkinElmer) in confocal mode. Nine fields of view were acquired from each well with a total of 3024 images of size 1080 × 1080 px (1 px = 0.59 µm) with 350 cells in each field of view on average. All fields of view were imaged in fluorescent and brightfield modalities, with one modality acquired first on all wells and then the second. This dataset is referred to as "seven cell lines" in the further text.

LNCaP dataset. The cells of human prostate adenocarcinoma (LNCaP, from ATCC) were seeded in a CellCarrier-384 Ultra Microplate (PerkinElmer), fixed in formaldehyde, and stained using DRAQ5 fluor (Abcam, Cambridge, United Kingdom) to tag nuclear DNA. A 20 × objective was used to acquire images on a CellVoyager 7000 (Yokogawa, Tokyo, Japan) instrument in confocal mode to acquire fluorescence and brightfield images of size 2556 × 2156 pixels (1 pixel = 0.325 µm) with 681 cell in each field of view on average. Similar to the seven cell lines dataset, one modality was acquired on all wells before moving on to the second modality.

ArtSeg-CHO-M4R dataset. The imaging was performed as described previously[29]. Briefly, live CHO-K1-hM_4R cells were seeded with a density of 25 000 cells per well into µ-Plate 96 Well Black plate (Ibidi) 5–7 h before the imaging to allow attachment. All the experiments were performed in the cell culture medium DMEM/F-12 with 9% FBS (Sigma), antibiotic antimycotic solution (100 U/ml penicillin, 0.1 mg/ml streptomycin, 0.25 µg/ml amphotericin B, Sigma) and 750 µg/ml of selection antibiotic geneticin (G418, Capricorn Scientific). The final volume in the well was 200 µl. All imaging experiments were carried out at 37 °C in the 5% CO_2 atmosphere. The images were captured with Cytation 5 Imaging Multi-Mode Reader (BioTek, Bad Friedrichshall, Germany). Images were obtained using a LUCPLFLN 20 × objective lens with working-distance of 6.6 mm, and numerical aperture of 0.45 (Olympus), using LED excitation source with 531(40) nm filter and captured with 593(40) nm emission filter. The field of view size was 1224 × 904 pixels (1 pixel = 0.323 µm). For a single field of view, a brightfield image was obtained first, which was immediately followed by fluorescence image acquisition. These steps were repeated for four fields of view in each well. In all experiments, a constant concentration of 2 nM UR-CG072[31], a TAMRA labeled fluorescence ligand was used to visualize cells expressing muscarinic M4 receptors in the fluorescence channel. In concentration–response experiments atropine, arecholine (Sigma), UNSW-MK259[32] and UR-SK75[33] were used. UNSW-MK259, UR-SK75 and UR-CG072 were kindly provided by Dr. Max Keller from the University of Regensburg. The ArtSeg-CHO-M4R dataset is made freely available for public use.

Artifact annotation. The seven cell lines and LNCaP data were inspected and 11.4% and 6.5% of the samples were found to have artifacts, 344/3024 and 51/784 fields-of-view respectively. The same number of fields-of-view from each dataset were randomly sampled to be used as training images without artifacts. At the same time, 99.2% of samples in the ArtSeg-CHO-M4R dataset (1171/1181) were found to have artifacts. The clean images for this dataset were generated as described below.

For all three datasets, pixel-level ground truth masks of artifacts were generated by manual annotation. All annotators had prior training in bioimage analysis, microscopy and cell biology. For seven cell lines and LNCaP datasets, the artifacts were annotated as polygons using VGG image annotator[34] and for ArtSeg-CHO-M4R dataset, as freehand annotations with the MembraneTools module of Aparecium software[35]. For all datasets, the artifact pixels were annotated while keeping the number of background pixels annotated as artifacts as low as possible. Due to the fuzzy nature of the artifacts borders, including some background pixels during annotation was inevitable. We nevertheless tried to emulate the typical human annotator and keep this number as low as possible, while maintaining a reasonable speed of the annotation process.

For the ArtSeg-CHO-M4R dataset, the artifact annotations contain a considerable number of background pixels in some images as it speeds up annotation and better reflects the annotation process in real-world conditions.

To obtain the weak labels for the seven cell lines and LNCaP datasets, the images were classified to be either clean or artifact-containing by manual inspection by the annotator. An image was considered clean if no artifacts were observed. For the ArtSeg-CHO-M4R dataset, as the vast majority of images contain at least one artifact, the clean images were generated by replacing the pixel values of manually annotated artifacts with the values of the corresponding pixels in the estimated background image. The background is estimated by fitting the original image with a two-dimensional second order polynomial function[36]. To simulate imaging noise, a zero-centered noise profile of the background pixels is added to the estimated background. Testing the clean images using trained model shows that no artifacts could be detected from the resulting images. Moreover, the modified areas were also not visually detectable by human experts (Supplementary Fig. S1).

ScoreCAM-U-Net for artifact segmentation. Our weakly supervised artifact segmentation pipeline combines the ScoreCAM model[18] that highlights areas of the image most useful for differentiating between clean and artifact-containing images with U-Net model[4] that directly classifies pixels into categories. We call this pipeline "ScoreCAM-U-Net " (Fig. 1, Appendix A, and supplementary Table S1).

ScoreCAM[18] is a technique used to explain predictions made by deep learning methods, mostly applied to models that perform image classification. ScoreCAM analyzes both the model output and the corresponding image, and highlights parts of the image that had a large impact for the particular prediction. It proceeds in four steps. First, visual representations (activation maps) of the last convolutional layer are extracted from an image classification model (ResNet[37] in our implementation). Next, each activation map is upscaled to match the size of the input image, normalized to a range between 0 and 1, and projected onto a copy of the input image via multiplication, producing a projected input image. Then, the classification model (ResNet in our implementation) uses projected inputs to calculate the probability of the input image belonging to each class. Finally, all activation maps are summed, each multiplied by the corresponding class-largest probability and passed through the ReLU[38] activation function to generate the final output (Supplementary Fig. S2). Unlike other competitors that rely on gradients, ScoreCAM uses the largest class probability to obtain the resulting map. It has been empirically shown that this feature makes ScoreCAM less noisy and therefore more useful in practice[18].

The strongly supervised U-Net[13] model has already been successfully adapted for brightfield nuclei segmentation and its architecture is described in detail in the corresponding paper[4] (Supplementary Fig. S2). The architecture consists of an encoder and a decoder connected by a bottleneck, and skip links which pass the signal from the encoder to the decoder. We used an encoder consisting of 15 convolutional layers that use convolutional filters of size 3×3 and a rectified linear unit (ReLU)[4,38] activation function. After every third layer, there is a 2×2 max-pooling layer and a skip connection to the decoder. Symmetrically, the decoder has 15 convolutional layers with ReLU activation functions. After every third convolutional layer, there is an upsampling layer that upscales its input height and width by a factor of 2. Finally, the bottleneck after the encoder has three convolutional layers. There are 64 filters in each convolutional layer in the encoder, decoder, and bottleneck.

Model training and evaluation. *Training.* To train the ScoreCAM-U-Net model, the ResNet50[37] classification model in the ScoreCAM[18] framework was first trained to classify clean and artifact-containing images. The model is trained on the seven cell lines dataset using 482 images for training, 101 for validation and 104 for testing; on ArtSeg-CHO-M4R, using 1386 images for training, 404 for validation, and 572 for testing; and on LNCaP, using 70 images for training, 16 images for validation and 16 images for testing. The test set in ArtSeg-CHO-M4R dataset was chosen such that ten concentration–response curves with multiple competitive ligands could be obtained. The Adam optimizer[39] was used to optimize binary cross-entropy loss for 150 epochs. The initial learning rate (0.002) was reduced by a factor of 10 when the validation loss did not improve for 10 consecutive epochs.

The output of ScoreCAM was binarized with the threshold of 0.05 and used as pseudo-labels for the U-Net model, which was subsequently trained to segment the artifacts using the same datasets' splits and training procedure. The ScoreCAM binarization threshold was selected to maximize the pixel-wise IoU of the validation set. precisely, ScoreCAM-U-Net was trained using different thresholds for binarizing the results of ScoreCAM before using them as pseudo-labels for the U-Net model. The results then were evaluated using the validation set and the threshold with the best IoU score was selected (Supplementary Fig. S3). All the experiments were conducted using a Tesla V100- PCIE-32 GB Graphics Processing Unit.

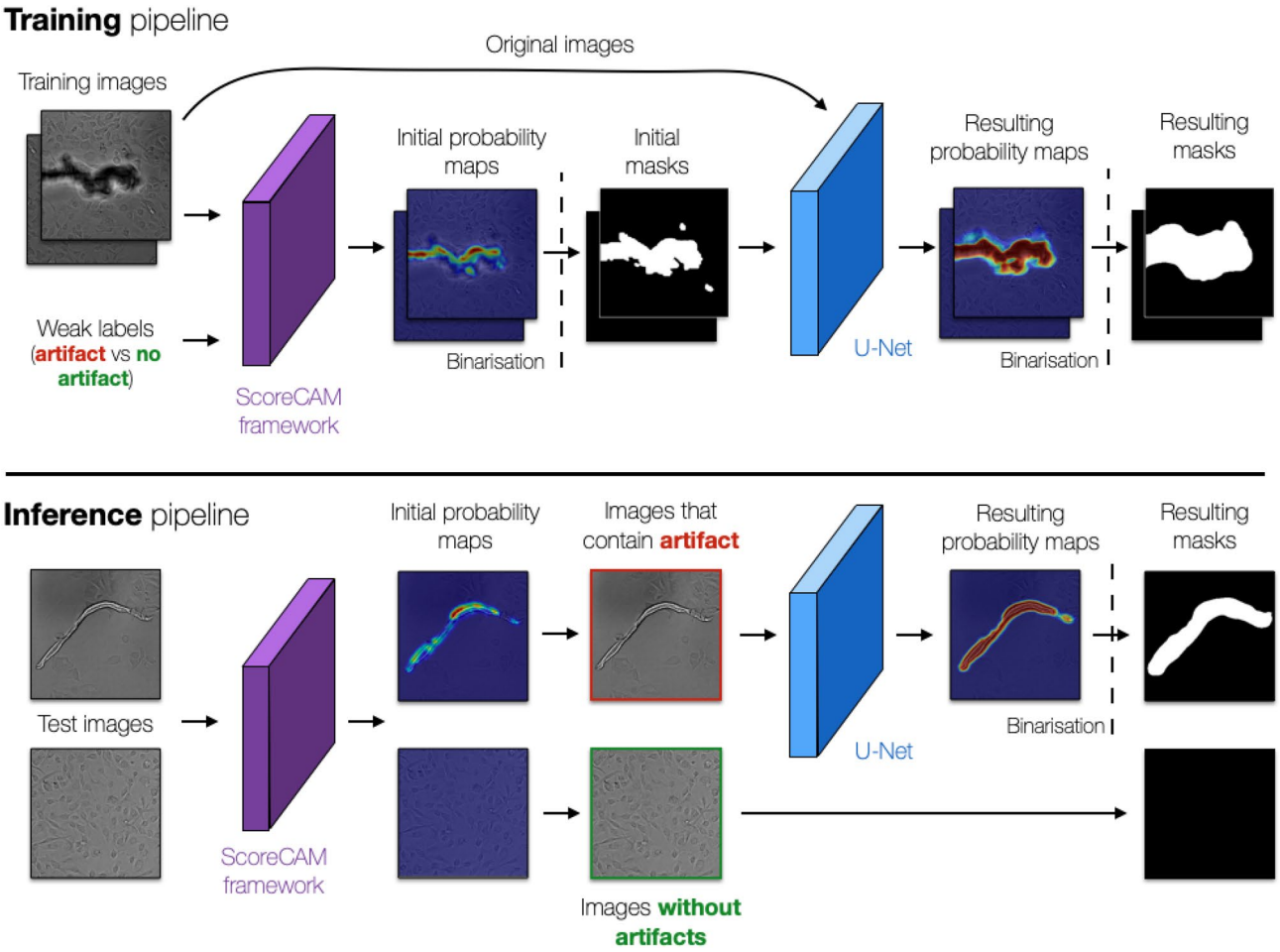


Figure 1. Artifact segmentation pipeline—ScoreCAM-U-Net. During training (top), ScoreCAM[18] (purple) is used to generate pixel-level probability maps of artifacts and the corresponding binary masks that are used to train the U-Net[4] segmentation model (blue). During the inference (bottom), the trained U-Net (blue) is used to segment artifacts from the images that were deemed to contain artifacts (image with red borders) by the ScoreCAM (purple). Vertical dashed lines: binarization of pixel probability maps values.

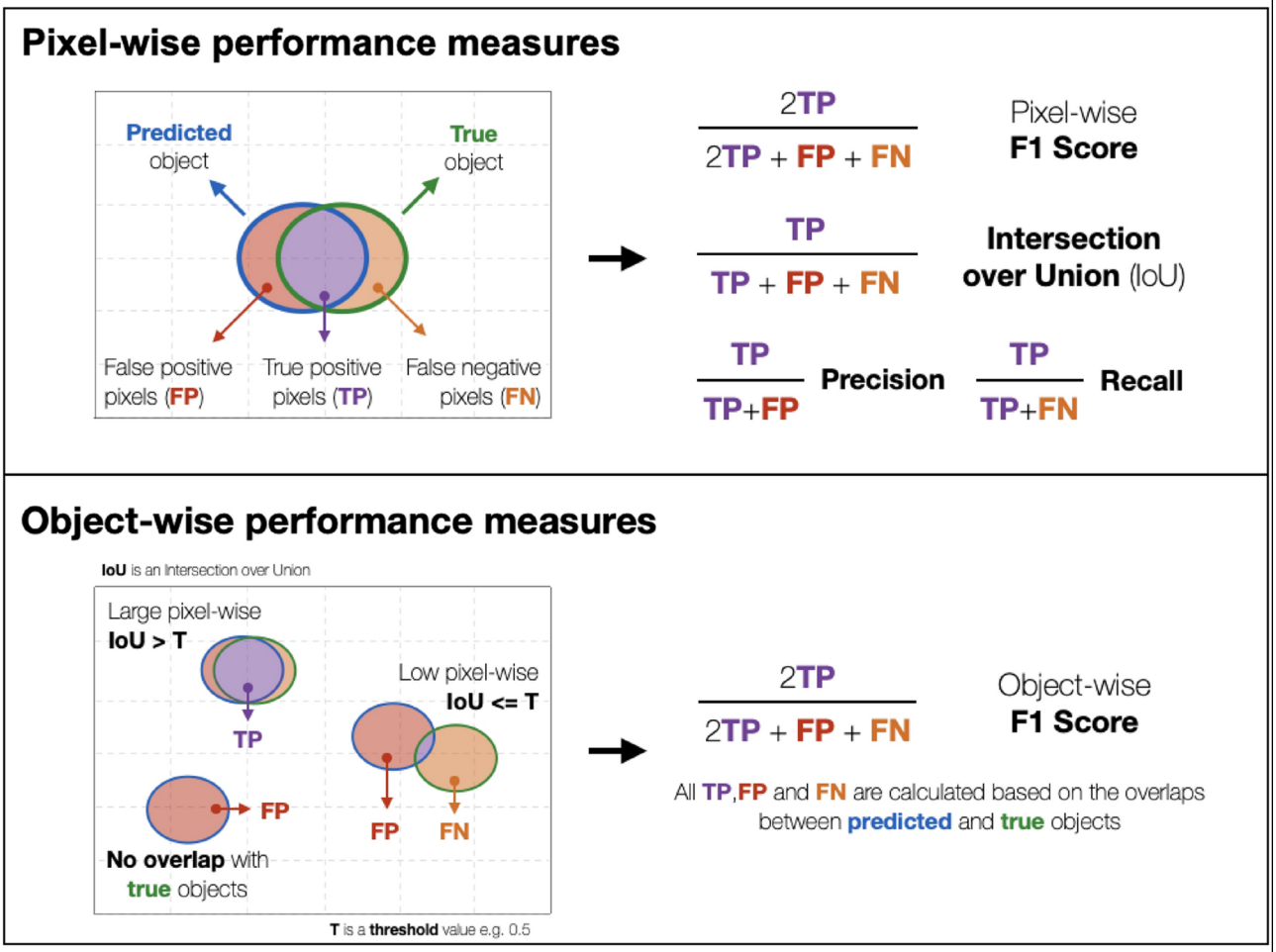
Comparison with other methods. We compared the segmentation results obtained from the ScoreCAM-U-Net to a number of alternative solutions. As ScoreCAM-U-Net is a combination of ScoreCAM and U-Net models, we first compared our performance to each of these models separately. We expected a strongly supervised U-Net model trained on pixel-level annotations to show better performance than its weakly supervised counterparts. We also compared the proposed approach to the current state-of-the-art algorithms used to detect anomalies using image labels in domains other than microscopy: Patch Support Vector Data Description (PatchSVDD)[26,27], Patch Distribution Modeling (PaDiM)[27], and an autoencoder-based method (AE)[24]. All model architectures, training parameters and training processes are adopted here as defined in the original papers[24,26,27].

PaDiM and PatchSVDD are both embedding similarity-based methods that use convolutional neural network-based approaches (encoders) that learn robust and short representations (embeddings) from patches of clean images. During the inference, the encoders are used to extract embeddings from test image patches and compare them to the embeddings extracted from the clean images based on a similarity metric. The main difference between these two methods is in the similarity metric employed to compare the embeddings as well as in the way the embeddings are constructed. PaDiM applies the Mahalanobis distance metric[40] and constructs an embedding by combining the features of multiple encoder layers, whereas PatchSVDD uses the Euclidean distance metric and constructs the representation from the features of a single encoder layer. Based on the embedding comparisons, each test image patch is assigned a similarity score in which a low similarity score indicates the presence of artifacts. The final segmentation of each test image is constructed after the similarity scores of these patches are distributed to their pixels and the corresponding patch segmentations are merged together.

The AE method also utilizes a convolutional neural network based approach (an encoder-decoder network architecture) that first learns representations of the clean input images (using the encoder) and then to reconstruct the original clean input images from the learned representations (via the decoder). During inference, the trained model is expected to fail to reconstruct the artifactual areas of the test images as the network has only acquired rich representations of clean images. Therefore, artifacts manifest themselves in areas with a high pixel-wise difference between the input image and its reconstructed counterpart.

We measure the ability of the models to correctly identify the presence of an artifact in the image using the F1 score which is the harmonic mean of precision and recall. We also assess the segmentation performance via calculating pixel-wise precision, recall, F1, and the intersection over union (Box 1).

Box 1: Performance measures.



Post-processing. We first binarized the probability maps produced by the models at cutoffs of 0.75 for AE, 0.3 for PaDIM, 0.0005 for PatchSVDD, 0.001 for ScoreCAM, 0.001 for ScoreCAM-U-Net and 0.45 for U-Net. These cutoffs were selected to maximize pixel-wise IoU (Box 1) performance on validation data. We then filtered out objects smaller than 1000, 500, and 500 pixels in the seven cell lines, the ArtSeg-CHO-M4R, and the LNCaP datasets respectively using *remove_small_objects* function from the skimage package[41]. The sizes of the filtered-out objects were selected to maximize the pixel-wise IoU of the majority number of models, and different sizes do not drastically change the performance of the models (Supplementary Tables S2, S3, and S4). We recommend using expert knowledge to select the size of objects to filter out.

Measuring impact of artifacts and artifact removal on the downstream analyses. To evaluate the utility of removing artifacts in microscopy experiments, we focused on two common types of downstream analyses: nuclei segmentation and effective concentration estimation from concentration–response assays. The former is a standard step in the majority of cell microscopy workflows while the latter is an example of a commonly used pipeline where cell segmentation is used for image intensity quantification which is followed up by regression analysis.

Nuclei segmentation. In order to assess how nuclei segmentation accuracy inside the artifactual regions compares to artifact-free areas, we evaluated the performance of nuclei segmentation in the seven cell lines dataset inside and outside the artifactual areas. To detect and segment the nuclei from the brightfield images we used an existing PPU-Net[3] model. The training, ground truth preparation, and post-processing steps for this model are described in the original publication[3]. We calculated segmentation pixel-wise F1 and object-wise F1 scores

(Box 1), following previously described approaches[3,42], and morphological properties (size and solidity) of the resulting nuclei.

Ligand affinity estimation. In downstream analysis of pharmacological experiments, the cell bodies are segmented from brightfield images using a U-Net-based deep learning model[29], and the cell fluorescence intensities are quantified from a parallel fluorescence channel based on the segmentation. The fluorescence intensities of cells depend on the strength of interaction (affinity) between the protein and the interaction partner (ligand) as well as the ligand concentration. The strength of protein–ligand interaction is determined using regression analysis of competitive ligand concentrations and the well average fluorescence intensity information from up to 64 individual images.

We studied the impact of artifacts and artifact removal on the determination of receptor-ligand interaction affinity. For that, in each of the ten individual concentration–response experiments, the cells were detected from brightfield images using a previously developed U-Net based segmentation model with an F1-score of 0.89[29]. The artifactual areas determined manually or with ScoreCAM-U-Net were removed from the analysis. For experimental control, the analysis was also carried out without any artifact removal. The average intensity of the detected cell pixels as well as the average intensity of the background were determined from the aligned red fluorescence protein filter (excitation: 531(40) nm, emission: 593(40) nm) fluorescence images made in parallel with the brightfield images. The values were averaged for all images from the same well. For each well, to find the specific fluorescence intensity of bound fluorescence ligand the difference between cellular and background fluorescence intensities was calculated. LogIC50 values corresponding to half maximal displacement of the fluorescence ligand were obtained via nonlinear regression analysis. For that, the fluorescence intensity dependence on the competitive ligand concentration was fitted with the Hill equation using GraphPad Prism 5.0 and "log(inhibitor) vs. response" nonlinear regression model which is equivalent to the logistic regression.

Concentration–response experiments serve as a good example for image analysis pipelines that rely on image intensity calculation and regression in the downstream analysis. For quantifying the quality of the full pipeline, we chose the absolute difference between the $LogIC_{50}$ values calculated from manual artifact removal and the alternative option. The difference of $LogIC_{50}$ values describes how accurate pharmacological parameters can be obtained with and without anomaly removal. We also used the R^2 value of the Hill equation fit as a metric, which reflects the overall agreement between the experiment and the model. Finally, we chose the Pearson's correlation coefficient r between predicted fluorescence intensity values using manual artifact removal and the alternative method, which allows isolating the effect of artifacts on the signal directly without the influence of other sources of uncertainty.

Results

To develop and test a weakly supervised method for artifact segmentation: we confirmed that artifacts exist and are prevalent in brightfield microscopy images; annotated artifacts in three datasets; tested models for finding them automatically; and evaluated the impact of removal on downstream analysis results.

Artifacts in brightfield images are prevalent and diverse. The artifacts in the seven cell lines dataset range from very big (e.g. a clump of detached cells covering 49% of the image pixels) to tiny ones only a few pixels in size. The average annotated artifact size in this dataset is 4,417 pixels, which is larger than a typical nucleus in this dataset, and 16% of images had at least 10% of their area covered by artifacts. The artifacts in the seven cell lines dataset were heterogeneous in their size and morphological properties (Fig. 2, Supplementary Fig. S4).

In the LNCaP dataset, we annotated 60 objects that affected 6.5% of the images. The sizes of artifacts range from big (e.g. a hair covering 10% of the image pixels) to small, which covers only 0.07% of the pixels, with the average artifact being 75,933 pixels (Supplementary Fig. S4).

In the ArtSeg-CHO-M4R dataset, almost all images had artifacts, with a total of 13,713 artifact objects in 1,171 affected images. Again, the largest object covered a large part of the image (e.g. 63% as a clump of detached cells), while the smallest one was a few pixels in size (Supplementary Fig. S4). An average artifact in this dataset had an area of 3,450 pixels, or 0.31% of image size.

Artifacts can be accurately detected with weak supervision. Next, we compared different approaches for artifact detection and segmentation qualitatively and quantitatively (Fig. 3A,B; Supplementary Fig. S5). We first evaluated the ability of the models to detect artifacts in the images. As ScoreCAM-U-Net and ScoreCAM both use the same ResNet classification backend, their detection performance is the same, with both models achieving image classification F1 scores of 93.2%, 93.7% and 90% in seven cell lines, LNCaP and ArtSeg-CHO-M4R datasets respectively (Fig. 3B, Supplementary Table S5). Other methods were less accurate, with the only exception of U-Net outperforming ScoreCAM-based models in the LNCaP dataset (99.4% F1 score for U-Net over 93.7% for ScoreCAM-U-Net; Fig. 3B).

We then assessed the models' performance in segmenting artifacts. ScoreCAM-U-Net outperforms the other non-strongly supervised models by achieving the highest area under the precision-recall curve, as well as the largest average object intersection over union on seven cell lines and LNCAP datasets (Fig. 3B). There was no dominant weakly supervised model in the ArtSeg-CHO-M4R dataset. Compared to the strongly supervised U-Net model, ScoreCAM-U-Net got the second-highest IoU performance in the seven cell lines (49.5 ScoreCAM-U-Net vs 72.9 U-Net) and the LNCaP (39.9 ScoreCAM-U-Net vs 65.74 U-Net) datasets (Supplementary Table S5).

Although the strongly supervised approach outperformed weakly supervised methods, it took substantial time to prepare the pixel-level annotations required for the U-Net model compared to weak labeling. On average, an expert spent 279 s to produce pixel-level annotation for a single microscopy image, while it took them

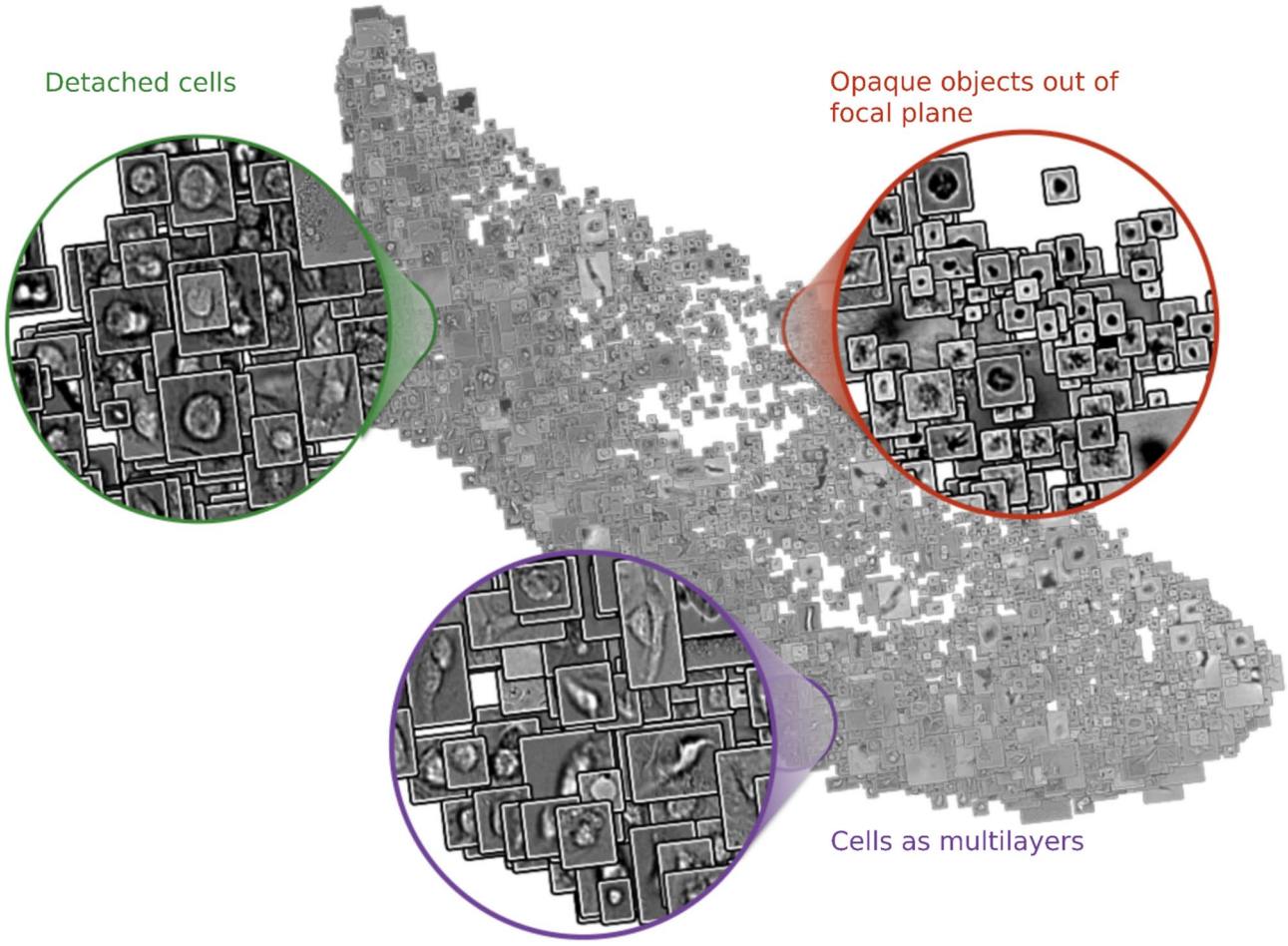


Figure 2. Artifacts are heterogeneous, and range in shapes and sizes. A UMAP projection of all artifacts from the seven cell lines dataset. The inputs to the UMAP are the pixels of each patch that contains an artifact and the outputs are the first two features in the UMAP embeddings of each patch. We then used these two features respectively as 'x' and 'y' values to plot the corresponding input patch in 2D space.

only 2 s to point out if a given image contained an artifact. Hence, weakly supervised methods consume about two orders of magnitude less of expert time for the given case. Therefore, when making a choice of method for dealing with artifacts, it is reasonable to take into account the dataset size and the amount of time needed to produce relevant annotations. For complex projects which require large training datasets for model development, generating precise pixel-level labels is prohibitively time-consuming, and hence weakly supervised approaches like ScoreCAM-U-Net are the best automated option available.

Weakly supervised artifact removal improves downstream analysis. After establishing the quality of the proposed ScoreCAM-U-Net method for artifact detection and segmentation, we evaluated the impact of using it for cleaning images on two downstream applications.

Removing artifacts improves quality of nuclei segmentation. As artifacts distort pixels that otherwise represent nuclei (Fig. 4), we observed substantial degradation in nuclei segmentation performance due to artifacts. The pixel-wise F1 score decreased from 0.89 in artifact-free to 0.60 in artifactual regions; and the object-wise F1 score decreased from 0.65 in artifact-free to 0.28 in artifactual regions (Fig. 5). This had a direct impact on naive analyses that do not differentiate between artifactual and clean regions, reducing segmentation accuracy (0.87 pixel-wise F1, 0.61 object-wise F1; Fig. 5). Importantly, automatically removing artifacts using ScoreCAM-U-Net has the same impact as manual removal, improving the segmentation performance to near-optimal 0.89 and 0.64 pixel-wise and object-wise F1 scores (Fig. 5).

We next considered nuclear size and morphology metrics with and without artifact correction. Nuclei in areas containing artifacts show different morphological properties with nuclei solidity of 0.92 and size of 213 pixels while the same properties are 0.95 and 400 pixels respectively in the artifact-free regions (Supplementary Fig. S6). In concordance with the segmentation results, automatically removing artifacts using ScoreCAM-U-Net recovers the expected nuclei size and solidity of 397 pixels and 0.95 respectively for artifact-free areas, again performing close to the gold standard of manual removal (Supplementary Fig. S6). These results demonstrate that automatic removal can overcome the detrimental effect of artifacts with quality close to manual filtering.

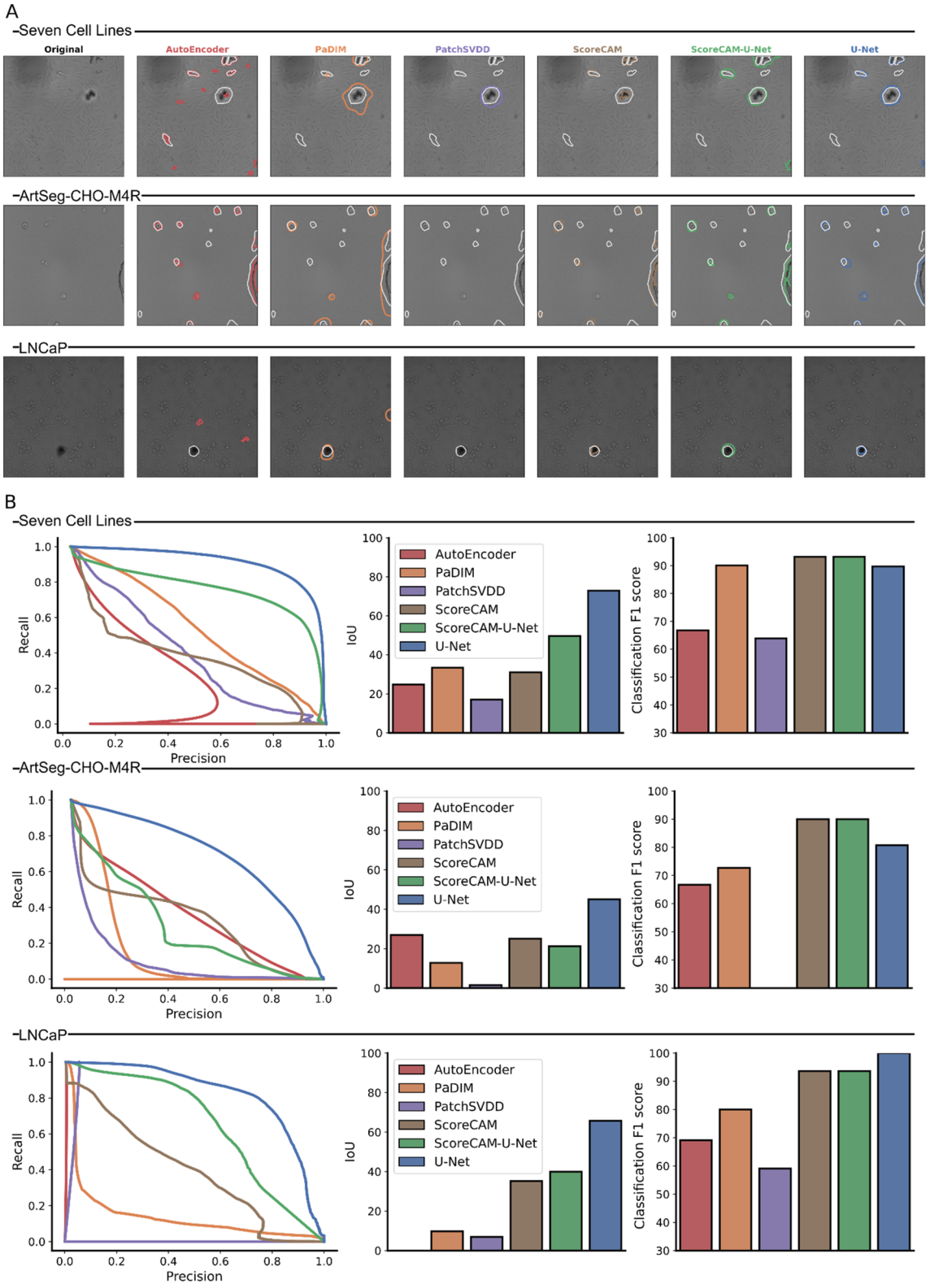


Figure 3. Artifact segmentation and image-level classification results for all models (colors) in seven cell lines, LNCaP, and ArtSeg-CHO-M4R datasets. (**A**) Examples of brightfield images and the corresponding artifact segmentation of all models (columns, colors) and datasets (rows; separated by lines and dataset names). White contour: expert annotated artifact boundaries; colored contours: artifact segmentation boundaries of the corresponding model. (**B**) Different performance metrics for all models (colors) and datasets (rows). Left column: artifact segmentation precision (x-axis) and recall (y-axis) of artifact detection at different thresholds (points along the curve) for all models and datasets. Middle column: artifact segmentation pixel-wise IoU (y-axis) for all models and all datasets. Right column: image-level classification F1 score (y-axis) for all models (x-axis) and datasets (rows).

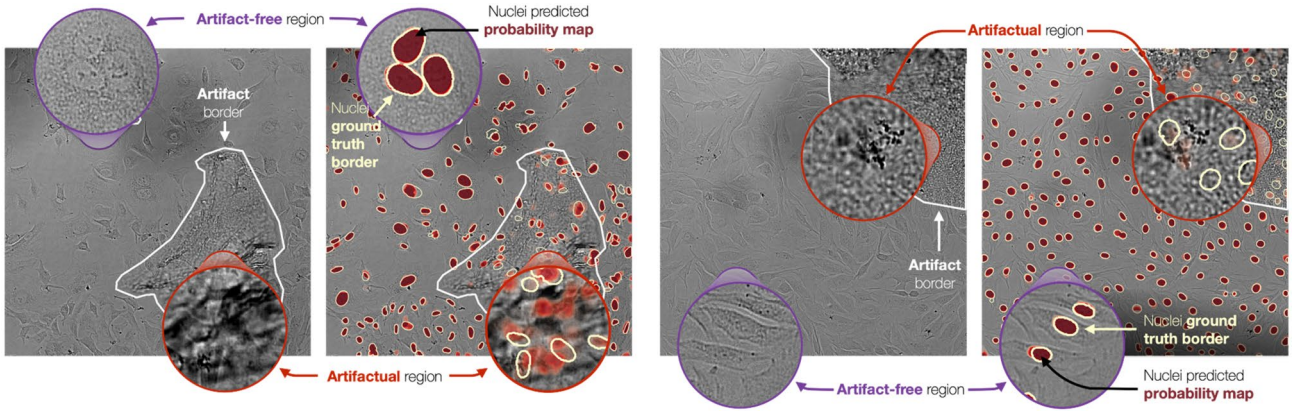


Figure 4. Visual impact of artifacts on nuclei segmentation. Two pairs of brightfield images with corresponding nuclei segmentation (in dark red) overlaid. Zoomed-in purple circles represent examples of artifact-free areas and artifactual areas (light red). White contours: artifact borders; yellow-ish white contours: nucleus ground truth borders; arrows and text: guides to corresponding regions and elements.

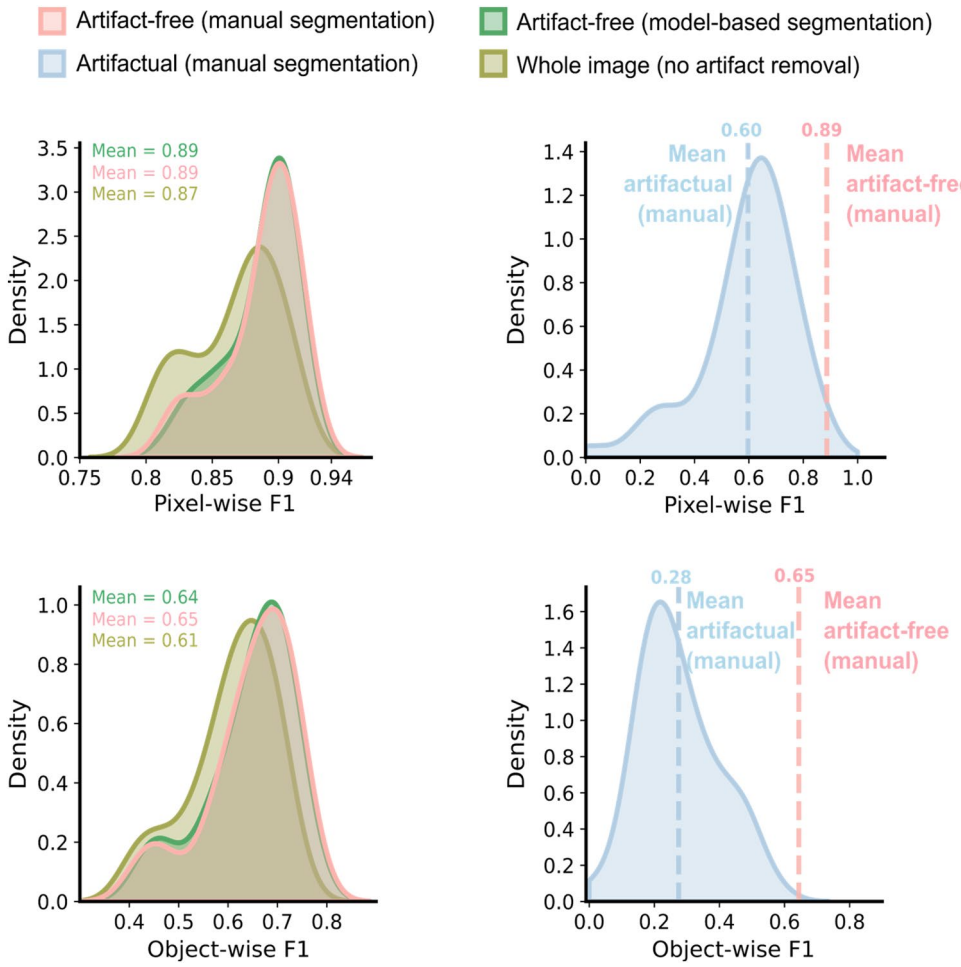


Figure 5. Impact of artifacts and artifact removal on downstream analyses. Density (y-axis) of image-average nucleus segmentation pixel-wise F1 (top, x-axis) and object-wise F1 (bottom, x-axis) in the seven cell lines dataset for different areas of the image (colors). Pink: area in the images manually annotated as not artifacts; blue: area in the images manually annotated as artifacts; green: area in the images automatically annotated as not artifacts by ScoreCAM-U-Net; yellow: all image area. Dashed lines: mean pixel-wise F1 and object-wise F1 of segmented nuclei in the artifactual and artifact-free regions (different colors).

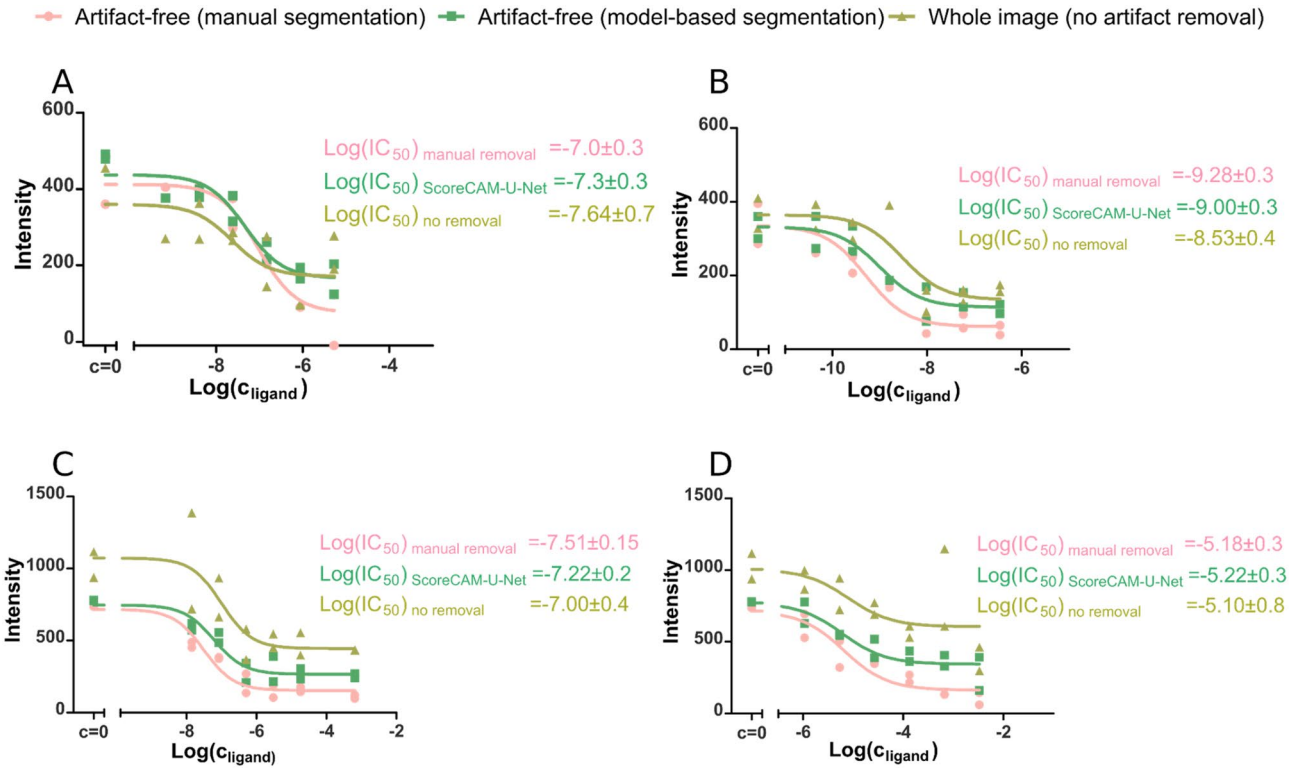


Figure 6. Cell fluorescence intensity dependence on M4 receptor ligand concentration determined with live-cell fluorescence microscopy at the presence of 2 nM UR-CG072. Displacement curves of three different ligands are shown: pirenzepine (**A**,**C**), atropine (**B**) and arecholine (**D**). Three different artifact removal methods at the image analysis stage are compared (colors): manual artifact segmentation, ScoreCAM-U-Net segmentation and no artifact removal. For each combination of ligand and artifact removal method a regression analysis is performed with Hill equation (Hill coefficient fixed at − 1) with the best fits shown as continuous lines. For each displacement curve, the Log(IC_{50}) ± SD is presented, where SD represents the standard deviation estimation of Log(IC_{50}). Each displacement curve was measured in duplicates with each data point representing the average fluorescence intensity of cells in each well.

Removing artifacts improves pharmacological parameter estimates. Cell segmentation in images is a commonly used process to determine biochemical or pharmacological parameters from microscopy experiments. Common examples of such experiments include quantifying image intensity of the segmented areas. This can be followed up by a test of significance or regression analysis to determine biochemical parameters like half-life of a reaction or the half maximal effective concentration of a substance. We analyzed how presence of artifacts affects the quality of a microscopy image-based analysis used to determine ligand affinity to M_4 muscarinic receptors using the ArtSeg-CHO-M4R dataset. Manual anomaly removal has a clear effect on both the plateau locations and the estimated Log(IC_{50}) values (Fig. 6, Supplementary Fig. S7). The mean absolute difference between Log(IC_{50}) calculated with manual artifact removal and no artifact removal is 0.29 units, equivalent to a two-fold error in dose. In contrast, after automatic artifact removal with ScoreCAM-U-Net, the Log(IC_{50}) difference from manual anomaly removal was reduced to just 0.16 units, which is similar to the standard deviation of 0.11 observed between biological replicate experiments. The model fit explained 0.89, 0.86, and 0.74 of the data variation for manual anomaly removal, ScoreCAM-U-Net anomaly removal, and no anomaly removal respectively. Finally, the Pearson's correlation coefficient of well-average fluorescence intensities between manual anomaly removal and ScoreCAM-U-Net anomaly removal is 0.98 while the correlation between manual anomaly removal and no anomaly removal is only 0.93. Overall, removing artifacts leads to an increase in replicate correlation, which itself results in reduction in estimate uncertainty. The estimated ligand affinity better reflects the values established from manually cleaned images. This confirms that artifact removal leads to considerable improvement of downstream regression or statistical analysis which relies on image intensity quantification.

Discussion

We proposed ScoreCAM-U-Net, a deep learning model for identifying artifacts in brightfield microscopy images that combines the benefits of weakly supervised learning which does not require delineating objects, and strongly supervised learning that provides pixel-level resolution. ScoreCAM-U-Net outperforms the other non fully supervised models in segmenting the artifacts from the images. Moreover, ScoreCAM-based methods outperform the others in detecting artifactual images with the only exception of U-Net in the LNCaP dataset. This is due to the relatively small size of the test set in the LNCaP dataset (16 images), so missing only one image could cause this difference in classification accuracy. Inspecting the results, we found that the scoreCAM-U-Net constantly

classified one of the clean images as artifactual while U-Net missed it only once. As the training of the ScoreCAM-U-Net is performed using only image-level labels, generating training data is orders of magnitude faster, but without substantially sacrificing performance compared to pixel-level annotation. To our knowledge, this is the first attempt to automatically detect artifacts in large sets of brightfield microscopy images.

Several factors can contribute to dissimilar model performance in different datasets, including but not limited to dataset size, nature of artifacts, average artifacts size, etc. We speculate that the relatively inferior performance of the models on the ArtSeg-CHO-M4R dataset compared to the other two datasets could be attributed to the more blurry nature of the artifact borders and the relatively small average artifact size ("Artifacts in brightfield images are prevalent and diverse"). On the other hand, the moderate fuzziness of the borders in both seven cell lines and the LNCaP datasets may contribute to the better performance than the ArtSeg-CHO-M4R; having seven cell lines superior as it has more representative samples in the training set.

Due to the blurry nature of the borders of the artifact objects, a mismatch between predicted and annotated pixels has appeared which led to an accuracy gap. However, the impact of artifact removal on downstream tasks, rather than classification and segmentation performance, is arguably the most relevant metric for practical application ("Weakly supervised artifact removal improves downstream analysis"). Our results demonstrate that artifacts have an adverse impact on nuclei segmentation and that detection and measurement of nuclei are improved when removing such artifacts. We showed that this impact manifests in both quantitative segmentation metrics such as pixel-wise and object-wise F1 score, as well as morphological properties of the nuclei like solidity and size, which are central for cytometry applications. Almost all study designs that use large-scale cell microscopy and image quantification-based readout would benefit from our model.

One important application of cell microscopy is intensity quantification for studying the localization and co-localization of fluorescently labeled molecules. To exemplify this type of analysis, we studied how artifact removal affects the calculation of drug-receptor binding affinities based on live-cell fluorescence and brightfield microscopy. After artifact removal with ScoreCAM-U-Net, the estimated ligand affinities are in better agreement with the values established from manually cleaned images. The model-based estimates also reduce linear regression uncertainty and result variability of independent experiments, indicating a combination of better fit of the theoretical model and improved reproducibility of the measurements. Thus, artifact removal improves image intensity quantification independent of the nature of statistical analysis applied downstream.

Our ScoreCAM-U-Net method establishes the utility of automatically segmenting artifacts from brightfield microscopy images. The key advantage of our approach is its scalability, such that clean images can be obtained for screening campaigns that would be prohibitively expensive to process manually. For example, a screening experiment generating 100,000 images could take around 388 h of continuous work by an expert to delineate artifacts from only 5% of them.

The limitation of our approach is an inability to differentiate different types of artifacts. For example, the current model would not tell if an image contains an artifact of cell debris and the other contains bacterial contamination. A natural extension can build on our approach to train a model that can differentiate between different types of artifacts. Other extensions can use the power of deep learning for other imaging modalities, such as histopathology, as well as to further reduce annotation time. We envision that ultimately, all common artifacts will be automatically segmented and optionally removed at the time of acquisition with no input needed by the operator. Moreover, we believe that the encouraging results presented in this work will motivate the use of weakly supervised segmentation methods such ScoreCAM-U-Net in other areas where pixel-level annotations are prohibitively expensive or time-consuming to acquire, i.e. medicine.

Data availability

The ArtSeg-CHO-M4R dataset is publicly available at https://datadoi.ee/handle/33/433; with https://doi.org/10.23673/re-307.

Received: 10 February 2022; Accepted: 10 June 2022
Published online: 06 July 2022

References

1. Wang, G. & Fang, N. Detecting and tracking nonfluorescent nanoparticle probes in live cells. *Methods Enzymol.* **504**, 83–108 (2012).
2. Salem, D. *et al.* YeastNet: Deep-learning-enabled accurate segmentation of budding yeast cells in bright-field microscopy. *Appl. Sci.* **11**, 2692 (2021).
3. Ali, M. A. S. *et al.* Evaluating very deep convolutional neural networks for nucleus segmentation from brightfield cell microscopy images. *SLAS Discov* **26**, 1125–1137 (2021).
4. Fishman, D. *et al.* Practical segmentation of nuclei in brightfield cell images with neural networks trained on fluorescently labelled samples. *J. Microsc.* https://doi.org/10.1111/jmi.13038 (2021).
5. Ayache, J., Beaunier, L., Boumendil, J., Ehret, G. & Laub, D. Artifacts in transmission electron microscopy. In *Sample Preparation Handbook for Transmission Electron Microscopy* 125–170 (2010). https://doi.org/10.1007/978-0-387-98182-6_6.
6. Ellis, E. A., Ann Ellis, E. & Cohen-Gould, L. Recognizing and preventing artifacts in microscopy: A roundtable discussion. *Microsc. Microanal. 22*, 2074–2075 (2016).
7. Pang, G., Shen, C., Cao, L. & van den Hengel, A. Deep learning for anomaly detection: A review. *arXiv [cs.LG]* (2020).
8. Ruff, L. *et al.* A Unifying review of deep and shallow anomaly detection. *arXiv [cs.LG]* (2020).
9. Hawkins, D. M. *Identification of Outliers* (Springer, 1980).
10. Ahmed, F. & Courville, A. Detecting semantic anomalies. *AAAI* **34**, 3154–3162 (2020).
11. Chen, S. *et al.* Avoiding artefacts during electron microscopy of silver nanomaterials exposed to biological environments. *J. Microsc.* **261**, 157–166 (2016).
12. Whelan, D. R. & Bell, T. D. M. Image artifacts in single molecule localization microscopy: why optimization of sample preparation protocols matters. *Sci. Rep.* **5**, 7924 (2015).

13. Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* 234–241 (Springer International Publishing, 2015).
14. Falk, T. *et al.* U-Net: deep learning for cell counting, detection, and morphometry. *Nat. Methods* **16**, 67–70 (2019).
15. Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K. & Yuille, A. L. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **40**, 834–848 (2018).
16. Redmon, J., Divvala, S., Girshick, R. & Farhadi, A. You only look once: Unified, real-time object detection. in *Proceedings of the IEEE conference on computer vision and pattern recognition* 779–788 (cv-foundation.org, 2016).
17. Tan, M. & Le, Q. V. EfficientNetV2: Smaller models and faster training. *arXiv [cs.CV]* (2021).
18. Wang, H., Wang, Z., Du, M. & Yang, F. Score-CAM: Score-weighted visual explanations for convolutional neural networks. *Proceedings of the* (2020).
19. Chen, X. & Konukoglu, E. Unsupervised Detection of Lesions in Brain MRI using constrained adversarial auto-encoders. *arXiv [cs.CV]* (2018).
20. Zhou, C. & Paffenroth, R. C. Anomaly Detection with Robust Deep Autoencoders. in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 665–674 (Association for Computing Machinery, 2017).
21. Abati, D., Porrello, A., Calderara, S. & Cucchiara, R. Latent space autoregression for novelty detection. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 481–490 (openaccess.thecvf.com, 2019).
22. Huang, C. *et al.* Attribute restoration framework for anomaly detection. *arXiv [cs.CV]* (2019).
23. Gong, D. *et al.* Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. in *Proceedings of the IEEE/CVF International Conference on Computer Vision* 1705–1714 (openaccess.thecvf.com, 2019).
24. Bergmann, P., Löwe, S., Fauser, M., Sattlegger, D. & Steger, C. Improving unsupervised defect segmentation by applying structural similarity to autoencoders. *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications* (2019). https://doi.org/10.5220/0007364503720380.
25. Khan, S. S. & Madden, M. G. One-class classification: Taxonomy of study and review of techniques. *Knowl. Eng. Rev.* **29**, 345–374 (2014).
26. Yi, J. & Yoon, S. Patch SVDD: Patch-Level SVDD for Anomaly Detection and Segmentation. *Comput. Vis. ACCV 2020* 375–390 (2021). https://doi.org/10.1007/978-3-030-69544-6_23.
27. Defard, T., Setkov, A., Loesch, A. & Audigier, R. PaDiM: A patch distribution modeling framework for anomaly detection and localization. *Pattern Recognition. ICPR International Workshops and Challenges* 475–489 (2021). https://doi.org/10.1007/978-3-030-68799-1_35.
28. Fishman, D. *et al.* Practical segmentation of nuclei in brightfield cell images with neural networks trained on fluorescently labelled samples. *J. Microsc.* **284**, 12–24 (2021).
29. Tahk, M.-J. *et al.* Live-cell microscopy or fluorescence anisotropy with budded baculoviruses—which way to go with measuring ligand binding to M4 muscarinic receptors? (2021). https://doi.org/10.1101/2021.12.22.473643.
30. Tahk, M.-J. *et al.* UT-GPCR001 microscopy of ligand binding to M4 muscarinic receptor in live CHO-K1-hM4 cells. *Live-cell microscopy or fluorescence anisotropy with budded baculoviruses-which way to go with measuring ligand binding to M4 muscarinic receptors?* (2022). https://doi.org/10.23673/re-306.
31. Gruber, C. G. *et al.* Differently fluorescence-labelled dibenzodiazepinone-type muscarinic acetylcholine receptor ligands with high MR affinity. *RSC Med Chem* **11**, 823–832 (2020).
32. Keller, M. *et al.* M2 Subtype preferring dibenzodiazepinone-type muscarinic receptor ligands: Effect of chemical homo-dimerization on orthosteric (and allosteric?) binding. *Bioorg. Med. Chem.* **23**, 3970–3990 (2015).
33. She, X. *et al.* Heterodimerization of dibenzodiazepinone-type muscarinic acetylcholine receptor ligands leads to increased M2R affinity and selectivity. *ACS Omega* **2**, 6741–6754 (2017).
34. Dutta, A. & Zisserman, A. The VIA Annotation Software for Images, Audio and Video. in *Proceedings of the 27th ACM International Conference on Multimedia* 2276–2279 (Association for Computing Machinery, 2019).
35. Allikalt, A., Laasfeld, T., Ilisson, M., Kopanchuk, S. & Rinken, A. Quantitative analysis of fluorescent ligand binding to dopamine D3 receptors using live-cell microscopy. *FEBS J.* **288**, 1514–1532 (2021).
36. Lanza, A., Tombari, F. & Di Stefano, L. Accurate and Efficient Background Subtraction by Monotonic Second-Degree Polynomial Fitting. in *2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance* 376–383 (ieeexplore.ieee.org, 2010).
37. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. in *Proceedings of the IEEE conference on computer vision and pattern recognition* 770–778 (openaccess.thecvf.com, 2016).
38. Nair, V. & Hinton, G. E. Rectified linear units improve restricted Boltzmann machines. *ICML* (2010).
39. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv [cs.LG]* (2014).
40. Mahalanobis, P. C. On the generalized distance in statistics. in (National Institute of Science of India, 1936).
41. van der Walt, S. *et al.* scikit-image: Image processing in Python. *PeerJ* **2**, e453 (2014).
42. Caicedo, J. C. *et al.* Evaluation of deep learning strategies for nucleus segmentation in fluorescence images. *Cytom. A* **95**, 952–965 (2019).

Acknowledgements

We thank the High-Performance Computing center at the University of Tartu, Institute of Computer Science for providing the computational resources. We thank Dr. Max Keller from the University of Regensburg for providing ligands UNSW-MK259, UR-SK75 and UR-CG072. We thank Erika Charola for proofreading and language editing; Nurlan Kerimov for helping with the project setups; and Martin Türk for helping with the visuals.

Author contributions

Designed project: D.F., L.P., M.A., K.H., T.L. Performed experiments: K.H., M.A., T.L. Performed analysis: M.A., K.H., T.L., D.F. Supervised study: D.F., L.P., K.P., A.R. Prepared data: M.T., J.T., T.L., K.H. Wrote paper: M.A., T.L., L.P., D.F. with input from all authors.

Funding

M.A. and K.H. were supported by the Estonian Research Council (PRG1095, PSG59, IUT34-4) and the Estonian Centre of Excellence in IT (EXCITE) (TK148). T.L. and M.T. were supported by the University of Tartu ASTRA Project PER ASPERA, financed by the European Regional Development Fund, Estonian Research Council grant (PSG230), and COST action CA 18133 ERNEST. J.T. was supported by the Estonian Research Council grant (PSG230) and COST action CA 18133 ERNEST. A.R. was supported by COST action CA 18133 ERNEST. L.P. was supported by Wellcome (206194), the Estonian Research Council (IUT34-4), and the Estonian Centre of Excellence in IT (EXCITE) (TK148). K.P. was supported by PerkinElmer Cellular Technologies. D.F. was

supported by Estonian Research Council grants (PRG1095, PSG59,IUT34-4, and ERA-NET TRANSCAN-2 (BioEndoCar)); Project No 2014-2020.4.01.16-0271, ELIXIR and the European Regional Development Fund through EXCITE Center of Excellence.

Competing interests

KP was an employee of PerkinElmer while performing this work. DF was an employee of Better Medicine OÜ while performing this work. All other authors declare no conflict of interest.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-14703-y.

Correspondence and requests for materials should be addressed to L.P. or D.F.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.