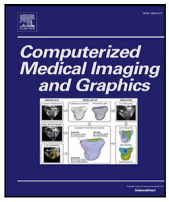
Computerized Medical Imaging and Graphics 112 (2024) 102328

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

journal homepage: www.elsevier.com/locate/compmedimag



Uninformed Teacher-Student for hard-samples distillation in weakly supervised mitosis localization

Claudio Fernandez-Martín [a,*], Julio Silva-Rodriguez [d], Umay Kiraz [b,c], Sandra Morales [a], Emiel A.M. Janssen [b,c], Valery Naranjo [a]

[a] Instituto Universitario de Investigación en Tecnología Centrada en el Ser Humano, HUMAN-tech, Universitat Politècnica de València, Valencia, Spain
[b] Department of Chemistry, Bioscience and Environmental Engineering, University of Stavanger, Stavanger, Norway
[c] Department of Pathology, Stavanger University Hospital, Stavanger, Norway
[d] ÉTS Montréal, Montréal, Québec, Canada

ARTICLE INFO

Dataset link: https://tupac.grand-challenge.org/, https://mitos-atypia-14.grand-challenge.org/, https://imig.science/midog2021/, https://github.com/DeepMicroscopy/

Keywords:
Mitosis detection
MAI estimation
Weakly supervised learning
Hard samples distillation

ABSTRACT

Background and Objective: Mitotic activity is a crucial biomarker for diagnosing and predicting outcomes for different types of cancers, particularly breast cancer. However, manual mitosis counting is challenging and time-consuming for pathologists, with moderate reproducibility due to biopsy slide size, low mitotic cell density, and pattern heterogeneity. In recent years, deep learning methods based on convolutional neural networks (CNNs) have been proposed to address these limitations. Nonetheless, these methods have been hampered by the available data labels, which usually consist only of the centroids of mitosis, and by the incoming noise from annotated hard negatives. As a result, complex algorithms with multiple stages are often required to refine the labels at the pixel level and reduce the number of false positives.

Methods: This article presents a novel weakly supervised approach for mitosis detection that utilizes only image-level labels on histological hematoxylin and eosin (H&E) images, avoiding the need for complex labeling scenarios. Also, an Uninformed Teacher-Student (UTS) pipeline is introduced to detect and distill hard samples by comparing weakly supervised localizations and the annotated centroids, using strong augmentations to enhance uncertainty. Additionally, an automatic proliferation score is proposed that mimicks the pathologist-annotated mitotic activity index (MAI). The proposed approach is evaluated on three publicly available datasets for mitosis detection on breast histology samples, and two datasets for mitotic activity counting in whole-slide images.

Results: The proposed framework achieves competitive performance with relevant prior literature in all the datasets used for evaluation without explicitly using the mitosis location information during training. This approach challenges previous methods that rely on strong mitosis location information and multiple stages to refine false positives. Furthermore, the proposed pipeline for hard-sample distillation demonstrates promising dataset-specific improvements. Concretely, when the annotation has not been thoroughly refined by multiple pathologists, the UTS model offers improvements of up to $\sim 4\%$ in mitosis localization, thanks to the detection and distillation of uncertain cases. Concerning the mitosis counting task, the proposed automatic proliferation score shows a moderate positive correlation with the MAI annotated by pathologists at the biopsy level on two external datasets.

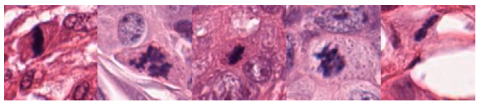
Conclusions: The proposed Uninformed Teacher-Student pipeline leverages strong augmentations to distill uncertain samples and measure dissimilarities between predicted and annotated mitosis. Results demonstrate the feasibility of the weakly supervised approach and highlight its potential as an objective evaluation tool for tumor proliferation.

1. Introduction

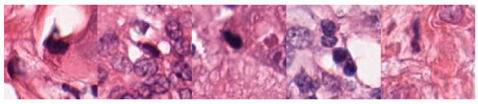
Mitosis counting is a crucial task in histopathological clinical practice. Particularly for breast cancer, the mitotic activity index (MAI) is considered one of the strongest prognostic factors. However, the process of mitosis counting is laborious and time-consuming, as pathologists must recognize and manually count mitotic figures in 1.59 mm^2

* Corresponding author.
E-mail address: clferma1@htech.upv.es (C. Fernandez-Martín).

https://doi.org/10.1016/j.compmedimag.2024.102328
Received 12 July 2023; Received in revised form 2 November 2023; Accepted 12 December 2023
Available online 4 January 2024



(a) Mitotic cells



(b) Non-mitotic cells

Fig. 1. Visual illustration of the morphological heterogeneity and the challenge of differentiating patterns between mitotic and non-mitotic cells, extracted from TUPAC16 (Veta et al., 2019).

on Hematoxylin and Eosin (H&E) slides under a microscope (Baak et al., 2005). The low occurrence of mitotic figures and the large size of the slides make this task particularly challenging. Moreover, the high variability in patterns and the similarity between mitotic and non-mitotic cells (as shown in Fig. 1) contribute to the moderate reproducibility of this task among clinical experts (Elmore et al., 2015).

In recent years, deep learning algorithms have emerged as a promising approach to address the challenge of mitosis localization and counting, offering objectivity and reproducibility. Convolutional neural networks (CNNs) have achieved remarkable results in the field of computational histopathology on various supervised learning applications, such as cell or nuclei segmentation, tissue classification, tumour detection or disease and prognosis prediction (Srinidhi et al., 2021). However, they require a significant amount of accurately labeled data to perform well. In the case of mitosis localization, labeling is a tedious and time-consuming task that usually requires the consensus of several pathologists. Reference datasets for mitosis localization commonly contain centroid-based labels (Veta et al., 2019) or inexact pixel-level annotations (Ludovic et al., 2013), since delineating individual mitotic cells at the pixel level is unfeasible. Consequently, previous works to automate the mitosis localization process have struggled to match the available labels to the use of segmentation or object detection CNNs, which are commonly used in localization tasks.

Common solutions based on pseudo-labeling strategies have emerged in the last few years to supplement the limited data. However, these may introduce noise and magnify model uncertainty, exacerbating the problem of hard negative cases (i.e., mitotic figures that are difficult to differentiate from other non-mitotic nuclei due to their morphologic similarities). Other previous works have attempted to address this issue by using multiple inference stages to force correct predictions, which can lead to suboptimal performance. Contrary to this line of work, this article proposes to make use of the inherent spatial localization capacity of CNNs in image-level classification tasks (Oquab et al., 2015) without the need to resort to an exact localization of the mitotic cell inside the region of interest. This weakly supervised approach provides higher flexibility and versatility, as image labels do not depend on detailed pixel-level annotations. This allows the proposed approach to be applied to a broader range of histological images, tumor types and databases while reducing the annotation effort by pathologists.

Despite the promising results previously achieved on weakly supervised mitosis localization (Fernandez-Martín et al., 2022), which enabled the use of image-level labels instead of centroid labels, the available information regarding the centroids of the annotated mitoses has yet to be fully exploited. For instance, the proposed model may produce a correct positive classification while erroneously locating a nucleus during training. Prior research has proposed addressing this challenge by incorporating prior knowledge in the form of constraint formulations (Jia et al., 2017; Silva-Rodrguez et al., 2022; Pathak et al., 2015). One example is the use of centroid coordinates constraints, which have been successfully applied to prostate MRI segmentation (Kervadec et al., 2022). Building on this prior work, the weakly supervised formulation might be extended to integrate the approximate centroid supervision. Nevertheless, and contrary to other works, we propose to integrate this knowledge as a proxy for hard sample detection, instead of as an implicit guidance during training. Concretely, we hypothesize that incorrect localizations during training can be used to identify hard negatives and noisy-labeled mitosis (i.e., mitotic figures where there is a lack of clear agreement between pathologists that translates to limitations or inaccuracies in the labeling data adding noise to the dataset). Thus, a reliable dataset is distilled for training a noise-free model by detecting the uncertainty associated with the predictions in such images. The motivation for using only the noise-free or clean cases (i.e., cases where there is an agreement between the predicted mitotic nuclei by the model and the annotated one by the pathologists) is supported by previous literature (Zhang et al., 2020a; Bernhardt et al., 2022), which has empirically studied the advantages of using only a subset of clean labeled samples over the whole dataset.

The main contributions of the study are summarized as follows:

- A deep learning strategy for weakly supervised segmentation of mitotic figures on H&E patches using image-level labels is proposed. Specifically, the approach is based on the maximum aggregation of instance-level predictions during training.
- A novel Uninformed Teacher-Student (UTS) formulation is introduced to distill hard samples based on disagreements between located and annotated mitosis on the training subset. Particularly, this involves noise integration through image augmentations to enhance uncertainty during training.
- Comprehensive experiments are conducted on three open-access datasets, demonstrating competitive performance using a single-phase pipeline that does not require exact localization information for training. Additionally, the robustness and generalization capabilities of the approach are validated on two external datasets, with various image acquisition systems, for mitotic activity counting at the biopsy level.

This journal paper presents a substantial extension of the conference work presented in Fernandez-Martín et al. (2022). First, the literature survey is expanded in terms of previous work on mitosis detection and uncertainty estimation methods, considerably increasing the umbrella of reference methods. Regarding the methodology, the benefits of the proposed weakly supervised formulation are further explored. The authors investigate how to incorporate location information in the form of constraint formulations into the weakly supervised strategy. Based on observations that forcing exact mitosis location hinders model performance, employing the disagreements between detected and annotated mitoses as a proxy for hard-sample distillation is proposed. Lastly, comprehensive experiments are incorporated to empirically validate the method's generalization capability. Additional datasets for mitosis detection and mitotic activity counting, together with new ablation experiments, are included in this study.

2. Related work

2.1. Mitosis detection

The development of support systems for detecting and counting mitotic figures has been one of the major topics of interest in digital and computational pathology during the last decade. The proposed methods have been highly dependent on the publicly released datasets and their associated annotations, such as MITOS12 (Ludovic et al., 2013), its extension to MITOS14 (Roux et al., 2014), TUPAC16 (Veta et al., 2019) or the most recent MIDOG21 (Aubreville et al., 2023). Some works have attempted to solve this challenge using classic texture or color handcrafted features extraction and machine learning classifiers (Saha et al.,

2018; Maroof et al., 2020; Sigirci et al., 2022). Nevertheless, the main core of the literature has shifted to training deep learning models via convolutional neural networks (CNNs) due to their remarkable results since the first mitosis detection challenges (Ciresan et al., 2013b,a).

Deep learning-driven methods can be differentiated into two categories, according to the used target region size: (i) cell-level patch-wise classification and (ii) image-level object detection. The firsts use small patches that contain single cells using a sliding window to train a CNN (Ciresan et al., 2013a; Lafarge et al., 2017; Zerhouni et al., 2017; Paeng et al., 2017; Akram et al., 2018; Chen et al., 2016; Sabeena Beevi et al., 2019; Hwang et al., 2020). In the second approach, images covering a larger tissue area containing multiple nuclei are used. Thus, the model is trained under an object detection paradigm that locates the mitotic figures in the image. Initially, the pixel-level annotations included in the MITOS12 challenge enabled direct training of pixel-level segmentation models (Li et al., 2018; Mahmood et al., 2020; Lei et al., 2021). Nevertheless, since labeling each mitotic figure at the pixel level is a cumbersome process, subsequent datasets released the location of the mitotic cells in the form of centroid-based labels. In this scenario, pseudo-labeling strategies were employed to train object detection algorithms. These strategies usually include utilizing pre-trained networks on the MITOS12 dataset (Sebai et al., 2020; Sohail et al., 2021; Lu, 2021), models trained for nuclei segmentation such as Hover-Net (Wang et al., 2022), and unsupervised approaches based on prior knowledge about the blue-ratio of the nuclei color distribution (Wahab et al., 2019). Additionally, recent methods have attempted to use only the available centroid information in a weakly supervised manner through the use of concentric losses (Li et al., 2019) or by adapting classical region-proposal networks to use anchor centroids instead of bounding boxes (Wollmann and Rohr, 2021).

Despite yielding promising results in mitosis detection, there are still open challenges for deep learning-based methods. Concretely, three main topics of interest are: (i) how to exploit weak, centroid-based annotations, (ii) how to deal with noisy annotations and hard negative mitosis, and (iii) how to develop robust models that can generalize to inter-center variability.

Regarding (i), centroid-based annotations have overtaken pixel-level labeling due to clinicians' convenience of preparing databases at reasonable times. Still, standard CNN methods for object detection are not prepared for this type of reference. As already mentioned, prior works have attempted to assign pseudo-labels at the pixel level. Nonetheless, the noise introduced in this assignment might be detrimental to the convergence of the model, thus worsening its performance. Concerning (ii), uncertainty in mitosis annotations and hard negative predictions are known problems in mitosis detection. Although experts agree moderately well on biopsy-level mitotic activity quantification, they usually present large variability in locating individual mitotic figures (Veta et al., 2016). To address this problem, the MITOS14 and MIDOG21 challenges label mitosis through majority voting by three experts, while TUPAC16 requires consensus from two different pathologists. Because of the similarity between mitotic figures and other nuclei (see Fig. 1) and the uncertainty introduced in the annotation, many previous methods have reported problems with false positive predictions. To alleviate this issue, recent works have proposed two-stage pipelines (Li et al., 2018; Tellez et al., 2018; Wahab et al., 2019; Sohail et al., 2021), in which a second CNN is trained to refine the predictions on hard-negative cases from the first stage. Hard negative cases are retrieved based on the predicted score of the first network (Tellez et al., 2018), or false positive predictions (Li et al., 2018). Other approaches have recurred to training networks on immunohistochemistry images like PHH3 and using image registration for selecting possible candidates at inference (Tellez et al., 2018). However, regardless of the computational inefficiency of these approaches during inference, empirical evidence presented on this essay suggests that forcing the model to produce certain predictions in hard cases, in which visual features may not correspond to the annotation due to uncertainty, may cause detriments to the model optimization. Finally, the inter-center adaptation (iii) plays a crucial role in model generalization, which may cause a significant performance drop according to Aubreville et al. (2021). Although this topic falls out of the scope of this paper, it is worth highlighting the efforts presented in previous literature to improve such generalization using adversarial training (Lafarge et al., 2017; Sebai et al., 2020), stain normalization (Zerhouni et al., 2017; Sebai et al., 2020) or stain augmentation (Tellez et al., 2018; Sebai et al., 2020; Wang et al., 2022), which are the main focus of the MIDOG21 challenge (Aubreville et al., 2023). In this work, standard stain normalization methods are employed, which have shown satisfactory results on histological images in the past (Zerhouni et al., 2017; Silva-Rodríguez et al., 2020; Sebai et al., 2020). Furthermore, to streamline the workflow, minimize costs, and reduce complexity, the analysis presented in this paper are carried out on H&E-stained images, excluding other staining techniques such as PHH3 or Ki-67. This approach enables to concentrate efforts on the inherent information provided by H&E staining and avoids the additional expenses associated with IHC staining, which can be costly.

In this work, the focus relies on training mitosis detection models that can effectively utilize centroid-based annotations in a flexible manner. Unlike previous literature, the proposed approach avoids assigning any pixel-level pseudo-label. Instead, the mitosis detection model is trained in a weakly supervised manner, relying solely on image-level labels (i). This work also distinguishes itself from previous literature in the way uncertainty of hard cases is addressed. Instead of employing two-stage pipelines during inference, the uncertainty is distilled by detecting dissimilarities between the annotated and located figures within a weakly supervised framework. Subsequently, a Student model is trained using only images with correctly localized mitoses, avoiding the approach of coercing the model to produce the desired output on hard samples (ii).

2.2. Weakly supervised segmentation

Weakly supervised segmentation (WSS) aims to leverage pixel-level localization using global (a.k.a image-level) labels during training. According to Ilse et al. (2018), WSS methods use fully convolutional CNNs with an aggregation function that merges all the spatial information into one value that serves as global prediction (Silva-Rodríguez et al., 2021). This output is then used to compute the loss function and drives the network optimization. Different strategies include aggregating spatial features (embedding-based) or pixel-level predictions (instance-based). Finally, the probability maps before the aggregation operation are used as segmentation predictions. Lately, these segmentation maps are refined to incorporate self-supervised learning pipelines (Wang et al., 2020) or uncertainty proxies (Belharbi et al., 2021), among others.

2.3. Uncertainty estimation

In the medical image analysis domain, labeling datasets might become a challenging task due to the high level of expertise required, known inter-observed variability on different tasks (Arvaniti et al., 2018; Veta et al., 2016), and the visual similarity between different categories, such as mitotic figures (see Fig. 1). This might produce noisy annotations, which harm model performance and generalization (Arpit et al., 2017; Zhang et al., 2016). Uncertainty estimation aims to improve the understanding of deep learning models' performance and generalization capabilities by quantifying the quality of predicted labels.

Previous works have shown that quantifying the uncertainty of the model outputs makes it possible to highlight possible noisy labels and refine the model training (Bernhardt et al., 2022). Uncertainty-aware pipelines present two stages: uncertainty estimation and model refinement. Various techniques have been proposed for estimating

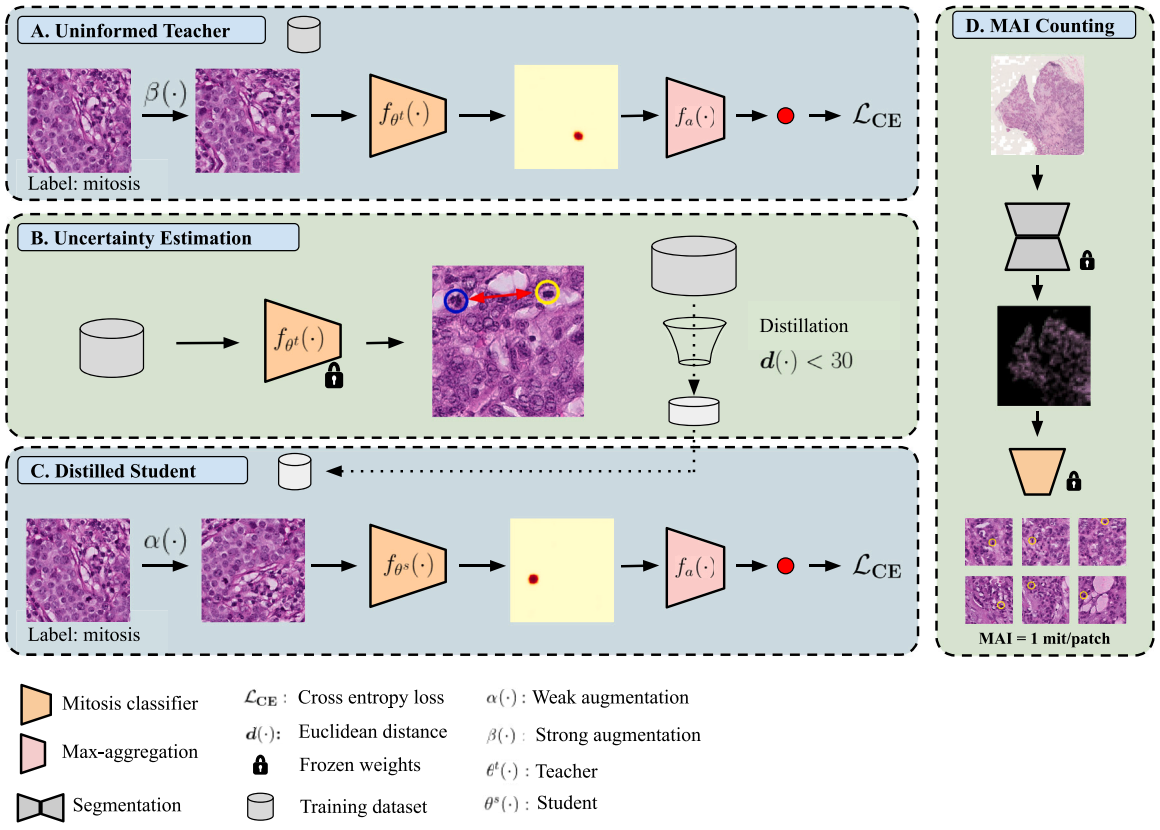


Fig. 2. Method overview. This work offers a novel strategy for weakly supervised mitosis detection that eliminates the need for assigning pixel-level labels to centroid-based annotations. Specifically, a Convolutional Neural Network (CNN) is trained to predict pixel-level scores, which are subsequently aggregated using a global max-pooling operation to optimize the network via standard cross-entropy loss at the image level (refer to Eq. (1)). Additionally, the method introduces a mechanism – Uninformed Teacher-Student (UTS) – to reduce false positive detection by removing hard cases. The proposed approach comprises four stages: **A. Uninformed Teacher** First, a classifier is trained using strong augmentations as a proxy for uncertainty enhancement. **B. Uncertainty detection.** Then, dissimilarities are checked between the annotated and the localized mitosis by the Teacher model at correctly classified patches on the training set. Thus, hard cases are identified using the Euclidean distance between centroids, which are further removed (see Eqs. (4) and (5)). **C. Distilled Student.** A Student model is trained on the distilled dataset, utilizing soft augmentations. **D. MAI counting** Finally, the mitosis detection model is employed to estimate the Mitotic Activity Index (MAI) on external datasets, focusing on tumor regions delineated using a UNet model.

uncertainty in deep learning models. These include measuring model robustness to image augmentation through Monte Carlo dropout (Ju et al., 2022), utilizing curriculum learning (Guo et al., 2018), co-teaching (Han et al., 2018), or using entropy sorting (Bernhardt et al., 2022). The estimated uncertainty can then be utilized for sample rejection during inference (Ghesu et al., 2021), refinement of a Student model through label weighting based on certainty (Ju et al., 2022), co-teaching based only on clean labels (Han et al., 2018), or even re-labeling cases with high uncertainty (Bernhardt et al., 2022).

This work aims to detect noisy and hard cases for mitosis localization by exploiting the uncertainty of the model predictions. To this end, leveraging a weakly supervised strategy for mitosis detection is proposed. Consequently, even if the network can effectively memorize the image-level labels, it remains uninformed about the exact object location, and thus its errors serve as indicators of model uncertainty. Subsequently, drawing inspiration from prior research (Zhang et al., 2020a; Bernhardt et al., 2022), a distillation process is introduced to generate a set of clean samples by excluding challenging cases, which is then used to train a Student model.

3. Methods

An overview of the proposed method is depicted in Fig. 2. In the subsequent subsections, the problem formulation and each component proposed will be described.

3.1. Weakly supervised mitosis localization

Problem formulation. In the paradigm of weakly supervised segmentation (WSS), the training set is composed of images $\{x_n\}_{n=1}^N$, whose binary label $\{Y_n^k\}_{n=1}^N$, such that, $Y_n^k = \{0,1\}$ is known and defines if a category k is present within the image. Also, each positive image has pixel-level labels $y_{n,i}$ for each i pixel in the image, but they remain unknown during training. Further, Y_n^k is denoted as Y_n for simplicity since one unique class is considered, and n is assumed as the image index.

Instance-based WSS. This work aims to train a CNN capable of locating positive mitosis during inference while being trained only with image-level labels. To do so, an instance-based weakly supervised learning strategy is used. Let us denote a CNN model, $f_\theta(\cdot) : \mathcal{X} \rightarrow \mathcal{H}^K$, parameterized by $\boldsymbol{\theta}$, which processes instances $x \in \mathcal{X}$ to output sigmoid-activated instance-level probabilities, h_i, such that $h_i \in [0,1]$. Also, a parameter-free aggregation function is employed, $f_a(\cdot)$, in charge of combining the pixel-level scores into one global output H, such that $H = f_a(f_\theta(x))$. Then, the optimization of $\boldsymbol{\theta}$ is driven by the minimization of cross entropy loss between reference and predicted image-level score.

$$\mathcal{L}_{ce} = Y log(H) + (1-Y) log(1-H) \tag{1}$$

In this work, it is proposed to use the maximum operation as an aggregation function, $f_a(\cdot)$. Although this aggregation only backpropagates gradients through the maximum-activated spatial regions, this

effect produces that only very discriminative cells will be classified as mitosis, which avoids false positive predictions.

Inference. During inference, pixel-level predictions are inferred using the pixel-level predictions given by the trained CNN, $h_i = f_\theta(x)$. The probability maps are resized to the original image dimensions by bilinear interpolation. Then, sigmoid scores are converted to a binary mask by applying a threshold to the probability maps. Concretely, the threshold is obtained from the operative point of the ROC curve between image-level predictions and references. Finally, a centroid is assigned to each element in the mask to be located as a mitosis.

3.2. Centroid constraint localization

Given a sample with one annotated mitosis, x, and its corresponding pixel-level scores, $h_i = f_\theta(x)$, the centroid coordinates of the predicted map can be computed as:

$$\widehat{C}_p = \frac{\sum(\sum_p h_i \; c_i)}{\sum h_i} \tag{2}$$

where $c_i \in \mathbb{N}^2$ is a 2D grid with the pixel coordinates for both directions in the image, p, for each pixel in the image, i. For example, the value ci for the first pixel - $i = 0$ - would be (0, 0), or (width/2, high/2) for the pixel in the center of the image.

Then, the model is optimized to minimize the euclidean distance, $d(\cdot)$, between the predicted and annotated centroid, C, as a constraint to the global optimization in Eq. (1), such as:

$$\mathcal{L} = \mathcal{L}_{ce} + \beta_{loc} \; d(C, \widehat{C}) \tag{3}$$

where β weights the relative importance of the constraint during training.

As discussed in the experimental section of the article (see Section 5.2), this formulation achieves better localization of mitoses during training. Surprisingly, however, it also produces a detriment in the results obtained in the validation set. Based on these findings, authors argue that a weakly supervised formulation allows more flexibility during training to recover relevant mitotic figures. In addition, it is proposed to take advantage of this phenomenon to locate hard samples by quantifying the disagreement between the annotated and the located mitosis.

3.3. UTS: Uninformed teacher - student

Hard mitosis distillation. Formally, let us define the training dataset D, as the union between positive images ($Y = 1$) and negative samples ($Y = 0$), such that $D = D^+ \cup D^-$. The proposed framework for uncertainty distillation is composed of the following stages:

1. **Uninformed Teacher training**. A Teacher model, θ^t, is trained on the whole training dataset, D, following the weakly supervised mitosis detection formulation by minimizing Eq. (1).
2. **Hard samples distillation**. From the positive labeled samples, $D^+ = \{x_m\}_{m=1}^M$, the Teacher network is used to predict the pixel-level segmentation maps, such that $h_{i,m} = f_{\theta^t}(x_m)$. It is worth mentioning that multiple mitoses can be predicted and annotated for each image. Then, using the formulation in Eq. (2), the predicted centroids, $\widehat{C}_{p,m,j}$, are computed for each object, j, in each predicted mask, m. Finally, the predicted locations are compared pairwise with each annotated mitosis, l, by using the euclidean distance such that:

$$\{e_{m,jl}\}_{jl=1}^{jl=j\times l} = \{\{d(C_{p,m,l}, \widehat{C}_{p,m,j})\}_{j=1}^{J}\}_{l=1}^{L} \tag{4}$$

   Finally, a subset of clean samples is obtained by retaining the predictions that contain at least one true positive prediction, i.e. a minimum distance $\tau = 30$ pixels, following previous literature (Sohail et al., 2021).

$$D_s^+ = \{x_m \text{ iff } \exists e_m < \tau\}_{m=1}^M \tag{5}$$

3. **Uninformed Student training**. A Student model, θ^s, is trained on the distilled dataset, $D_s = D^- \cup D_s^+$, following the same training procedure as the Teacher model.

Hereafter, authors refer to this formulation as Uninformed Teacher-Sudent (UTS) training.

Noise integration. One common way to reveal uncertain predictions on previous literature is through noise incorporation, in the form of dropout (Leibig et al., 2017; Ju et al., 2022) or image augmentations (Zhang et al., 2020a). As stated previously, this study aims to distill only clean samples for the training of the Student model. For this purpose, it is assumed that by incorporating strong image transformations, the memorability of the model during training is hindered, and thus localized mitoses will only coincide with those annotated in clean cases. A set of weak transformations, $\alpha(\cdot)$, is defined for this purpose, which includes random flips, rotations, and mirroring operations. Additionally, a set of strong augmentations, $\beta(\cdot)$, is defined, which includes optical and grid distortions. It is worth mentioning that the latter transformations affect the cell morphology, while the former only change the orientation of the histological structures. Finally, the Teacher model is trained using strong augmentations, $\beta(\cdot)$, while the Student model is trained on the clean subset samples, as indicated previously, using weak augmentations, $\alpha(\cdot)$.

3.4. MAI counting at whole-slide image level

Mitosis counting is a commonly used technique in pathological analysis to determine the proliferation rate (MAI) of cells in a tumor tissue sample and to estimate the associated prognosis. In the following, the proposed mitosis localization methods are extended to estimate a mitosis activity index from whole slide images (WSIs).

Tumor tissue segmentation. WSIs are gigapixel images that present a wide tissue heterogeneity. First, the authors propose to train a tissue segmentation model, ϕ_{tissue}, on patches x, to obtain the relevant tumor regions. Hereafter, this module is referred to as the tissue segmentation model, TSM. Concretely, a UNet model is used to segment patches at pixel level into $P = 5$ relevant tissue categories: tumor, stroma, inflammation, necrosis, or other. Thus, softmax pixel-level probabilities are obtained such that $q_{\Omega,p} = \phi_{tissue}(x)$, where Ω indicates the spatial domain. Secondly, the tissue proportion in each patch is computed by averaging pixel-level probabilities over Ω, and a patch is considered relevant for mitosis counting if the tumor percentage is above a certain threshold, τ_{tumor}, which is empirically fixed.

Mitotic counting. Finally, the mitosis detection model, θ^s, is applied to the extracted tumor patches that cover, at least, the same area used by the pathologists for performing the MAI counting. For a given WSI, the proposed MAI is the average number of mitoses detected per patch. Note that while the proposed MAI is calculated from the number of mitoses detected per patch over the whole tumor tissue in the entire WSI, it should be kept in mind that the MAI score assigned by pathologists is typically based on counting mitoses only within the tumor hotspot (i.e., the tumor area with the highest proliferation). As a result, the patch MAI may be more diluted since it includes more tumor tissue area that may contain fewer mitoses than the tumor hotspot.

4. Experimental setting

4.1. Datasets

The experiments described in this work were carried out using five different datasets. These datasets were selected to evaluate the main tasks developed in this work: mitosis localization and slide-level mitotic activity counting. The following is a description of them.

4.1.1. Mitosis localization

To evaluate the performance of the proposed Teacher-Student on mitosis localization, three popular open-access datasets of breast histology tissue regions with curated pixel-level labels of mitotic figures were used: TUPAC16, MITOS14, and MIDOG21.

TUPAC16-auxiliary. The auxiliary dataset of the 2016 TUmor Proliferation Assessment Challenge (TUPAC16) (Veta et al., 2019) is composed of 73 breast cancer whole-slide images from two different institutions. In particular, the auxiliary mitosis dataset contains 1552 processed regions of interest at 40× magnification, with centroid-labeled mitosis by consensus of expert pathologists.

MITOS14. The MITOS-ATYPIA 2014 challenge of ICPR (MITOS14) (Roux et al., 2014) gathers 1200 high-power fields (HPFs) from 11 different breast biopsies, using two different digitalization devices, at 40× magnification. For the training dataset, coordinates of annotated mitosis based on the agreement of 3 pathologists are available.

MIDOG21. The Mitosis domain generalization in histopathology images challenge (MIDOG21) (Aubreville et al., 2023) introduced a collection of breast histology tissue regions from 6 different digitalization scanners (denominated A to F subsets), at 40× magnification. The training subset is composed of 200 tissue regions (from centers A to D), from which 150 samples (centers A to C) were carefully analyzed by three expert pathologists, who identified the existing mitotic figures. Concretely, the dataset contains up to 1721 annotated mitotic figures. In addition, this dataset includes a collection of 2714 hard negative samples. Those are look-alike mitotic figures that were discarded by the consensus of the three pathologists.

4.1.2. MAI counting at the WSI-level

A publicly available dataset (CCMCT-MEL) and a private dataset (SUH) were used to validate the capacity of the proposed pipeline for mitosis localization for mimicking the pathologist's activity mitotic counting (MAI) at the slide level. Thus, the datasets were selected due to the availability of entire whole slide images.

CCMCT-MEL. The canine cutaneous mast cell tumor manual expert labeled dataset (CCMCT-MEL) (Bertram et al., 2019) constitutes a comprehensive collection of microscopy annotations. It includes 32 whole slide images (WSIs) scanned at 40× magnification, representing both low-grade and high-grade cases, each meticulously annotated by two expert veterinary pathologists, focusing on mitotic figures, and providing additional annotations for neoplastic mast cells, inflammatory granulocytes, and mitotic figure-like objects. This dataset comprises 238,340 annotations, with 42,465 dedicated to mitotic figures.

SUH. The SUH dataset is a private collection of 260 surgical specimens of triple-negative breast cancer (TNBC) obtained from the Stavanger University Hospital in Norway between 1978 and 2004. These samples were carefully selected by an experienced breast pathologist who chose the most representative slide for each case. The slides were then scanned at 40× magnification using three different scanners: the Hamamatsu NanoZoomer S60, Hamamatsu NanoZoomer 2.0HT, and Leica Aperio AT2, located in Stavanger (Norway), Atlanta (USA), and Oslo (Norway), respectively. Note that this dataset presents a wide heterogeneity of potential image acquisition systems to obtain whole-slide images. The same pathologist also performed the MAI scoring on the most active tumor area of the H&E slides, using consecutive high-power fields (HPFs) representing a total area of 1.59 mm^2 (Baak et al., 2005).

4.2. Data partitions and preparation

4.2.1. Mitosis localization

In this work, mitosis localization is evaluated on datasets inherited from previous open-access contests. Although these competitions provide an exceptional platform to compare different strategies, the test subset is hidden from the participants. Thus, it is possible to evaluate these methods during the contest, but not after some time due to lack of maintenance, as is the case with the databases used. Therefore, a partitioning of the training cases into three subsets: training, validation, and testing, has been carried out for each dataset, to promote better comparison between methods in the future.

TUPAC16-auxiliary. The dataset is divided into patient-level training, validation, and testing cohorts in a similar fashion to prior literature (Li et al., 2019). For all partitions, subsets contain samples from both centers. Concretely, cases 30, 37, 44, 51, 58, 65, and 72 were used for validation, while cases 31, 38, 45, 52, 59, 66, and 73 were used for testing.

MITOS14. Samples A03 and H03 were used for validation, while A04 and H04 were incorporated into the test subset. Note that for a fair evaluation of the proposed methods, cases from both scanners are used for validation and testing.

MIDOG21. The samples from scanners A, B, and C are used in this work. Following the strategy pursued in the other databases, the partition is performed to ensure that all training, validation, and testing subsets contain samples from each scanner. Concretely, cases 41 to 45, 91 to 95, and 141 to 145 are used for validation, and cases 46 to 50, 96 to 100, and 146 to 150 are incorporated into the testing subset. The rest of the cases are used for the training subset. In addition to the dataset containing the curated labeled mitosis, the annotated hard negatives (HN) are used to challenge the proposed approach in noisy scenarios. The combined dataset is referred to as MIDOG21_{HN}.

4.2.2. MAI counting at the WSI-level

CCMCT-MEL. Although this dataset is validated in the context of proliferation counting at the biopsy level, explicit measures are not present for each whole slide image. To alleviate this issue, the mitotic activity index is estimated from a hotspot in the slide. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) (Pedregosa et al., 2011) is utilized to identify the mitotic hotspot area within the biopsy. The initial epsilon parameter (e.g., the maximum distance between two samples for them to be considered neighbors) is set to half the diagonal length of the region of interest (ROI), and one-sixth of the total annotated mitoses in the WSI is established as the minimum number of samples per cluster. These parameters are adjusted to ensure the identification of at least two clusters when multiple mitoses are annotated within the WSI. Subsequently, after identifying the most populated mitoses cluster (the hotspot), the centroid coordinates for this area are determined. An ROI measuring $1.59mm^2$ is then extracted, centered around this hotspot's centroid. Finally, the mitotic figures within the hotspot ROI are identified. This enables the comparison of the annotated Mitotic Activity Index (MAI) against the predicted MAI generated by the model.

SUH. The entire dataset is used as external testing, thus no partition is required in this case.

4.3. Metrics

Standard metrics for mitosis localization evaluation are employed. First, the model is optimized using only global image-level labels by means of the accuracy, AUC, and F1-score. Then, the comparison with state-of-the-art methods on mitosis detection is assessed using the standard criteria of mitosis detection contests (Sohail et al., 2021). At 40× magnification, detected mitosis is considered true if it is located at most 30 pixels from an annotated mitosis. Under this criteria, precision, recall, and F1-score are computed. Regarding MAI counting evaluation, Spearman and Pearson correlations are used as figures for merit to compare the automatically estimated proliferation score with the one manually annotated by the pathologist, following previous works (Tellez et al., 2018).

Table 1
Performance comparison of the proposed model with existing methods on TUPAC16-auxiliary dataset. The presented figures of merit are F1-score, precision, and recall, obtained for mitosis detection. The best result for each metric is highlighted in bold, and the proposed methods are emphasized in gray. (See Nateghi et al. (2021) and Rehman et al. (2022)).

Method	F1 score	Recall	Precision	Multiple phases	Location supervision	External data	External validation
DeepConsensus - Wollmann and Rohr (2021)	0.470	–	–		×		
Tellez et al. (2018)	0.480	–	–	×	×	×	×
Lafarge et al. (2017)	0.620	–	–		×		
Akram et al. (2018)[b]	0.640	0.671	0.613	×	×	×	
Mahmood et al. (2020)	0.642	0.642	0.641	×	×	×	×
Zerhouni et al. (2017)[b]	0.648	0.623	0.675	×	×	×	
Paeng et al. (2017)	0.652	–	–	×	×		
SegMitos - Li et al. (2019)	0.669	0.700	0.640		×		
TL-Mit-Seg - Wahab et al. (2019)	0.713	0.660	0.770	×	×		
UTS - Teacher (ours) (Fernandez-Martín et al., 2022)	0.729	0.720	0.739				
Nateghi et al. (2021)	0.738	0.714	0.764	×	×		
FMDet - Wang et al. (2022)	0.745	**0.801**	0.697		×	×	×
Sohail et al. (2021)	0.750	0.760	0.710	×	×	×	
UTS - Student (ours)	0.767	0.716	**0.828**		×		
Rehman et al. (2022)[a]	**0.783**	–	–		×	×	

[a] Ensemble of hand-crafted features models.
[b] Shows the methods evaluated on the unavailable hidden test set of the original TUPAC16 challenge.

4.4. Implementation details

4.4.1. Image pre-processing and normalization

Following relevant literature in Li et al. (2019), patches of size 500 pixels are extracted from the regions of interest for computational efficiency during training and inference. To decrease the effect of inter-center stain variability, patches from the different datasets are color-normalized using the stain normalization method of Macenko et al. (2009) to the same reference space, represented by a selected region containing representative tumor tissue.[1]

4.4.2. Mitosis localization

The proposed Uninformed Teacher-Student (UTS) is trained using ResNet-18 (He et al., 2016) convolutional blocks as a backbone. Concretely, the first 3 blocks pre-trained on ImageNet (Deng et al., 2009) are used as feature extractors and are then retrained for the mitosis detection task. This architecture was trained during 40 epochs to optimize Eq. (1) using a batch size of 32 images and a learning rate of 0.0001. In order to deal with class imbalance, the images are sampled homogeneously according to their class in each epoch. Strong augmentations were used as detailed in Section 3.3 to train the Teacher model over the whole dataset. Then, a clean training dataset is distilled via hard samples detection on located mitosis as indicated in Eq. (5). Finally, a Student model is trained following the same training procedure as the Teacher model but using the distilled dataset and changing the augmentation to weak transforms. During the training of both models, performance metrics on the validation dataset are monitored, and the best model in terms of image-level classification is saved for testing. The code will be publicly available on https://github.com/cvblab/Mitosis-UTS.

4.4.3. MAI counting at the WSI-level

The proposed automatic MAI counting is obtained from whole-slide images as specified in Section 3.4. Non-overlapping patches of size 512 at 20× magnification are extracted from whole slide images, and the tissue segmentation model from López-Pérez et al. (2023) is used to obtain the tissue proportion. Then, the threshold of tumor proportion is empirically fixed at $\tau_{tumor} = 40\%$ to get relevant patches. Also, an exclusion criterion on a minimum tumor area of 1.59 mm^2 is included to use, at least, the same tumor tissue as pathologists use in clinical practice (Baak et al., 2005). Therefore, the minimum number of patches representing the same tumor area used by pathologists will depend on the pixel resolution of the different scanners. Finally, the MAI is computed using the Student model for mitosis detection.

5. Results

5.1. Comparison to literature

5.1.1. Mitosis localization

Performance on TUPAC16. The proposed mitosis detection methods are trained using the TUPAC16 training partition. The quantitative results obtained by the proposed UTS method for mitosis localization on the test cohort are presented in Table 1 for both Teacher and Student models. Results reported in previous literature on the TUPAC16 dataset are also included. In addition, qualitative visualizations of the model performance are shown in Fig. 3. The proposed Teacher weakly supervised method reaches an F1-score value of 0.729, comparable to prior literature using deep learning methods, but without requiring access to any supervision regarding the exact location of the mitosis in the image. It should be noted that, in addition, the best previous methods use additional training data and require multiple stages of label refinement. In contrast, the proposed method uses only one training cycle. Moreover, the proposed approach obtains the best precision on mitosis localization that only uses one training phase. Once the proposed hard samples mining is introduced and the Student model is trained, the obtained results reach an F1-score value of 0.767, which increases the performance over the Teacher model by nearly 4%. In addition, the Student model is competitive with the best previous results reported on the TUPAC16 dataset that rely on deep learning, showing a promising precision of 0.828. This is because the UTS Student model produces a low number of false positives, which is one common limitation of mitosis detection algorithms. These results support the claim that training the model with a clean dataset improves the generalization of the model due to the decrease in noise. This is a paradigm shift with respect to how previous methods deal with difficult negative cases, which usually employ a second model to strengthen the classification.

Generalization to other datasets: MITOS14. In order to validate the generalization capabilities of the proposed method, the UTS models trained on TUPAC16 dataset are directly used for inference on the test subset of MITOS14 (UTS w/o FT). Also, the UTS pipeline is trained from scratch on the training subset of MITOS14 (UTS w/ FT). The results obtained, together with the results reported in previous literature, are presented in Table 2. In addition, qualitative visualizations

[1] The reference image is available on https://github.com/cvblab/Mitosis-UTS/tree/main/local_data/color_norm/.

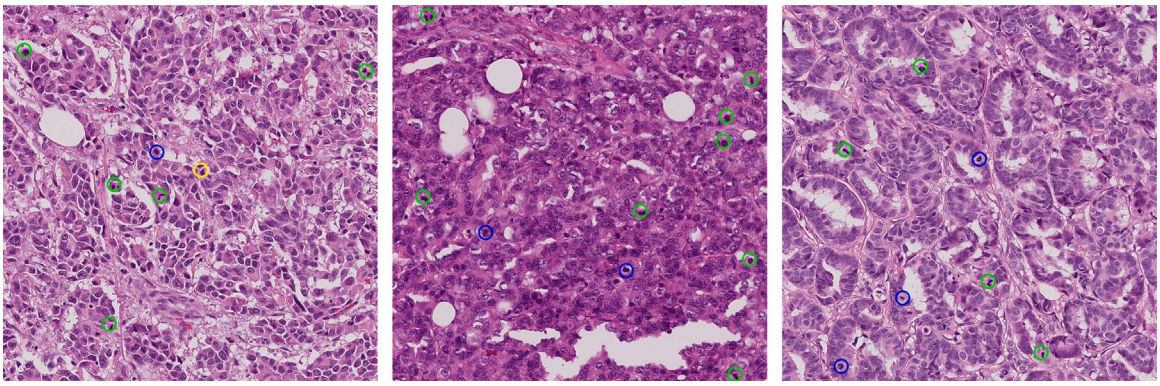


Fig. 3. Qualitative evaluation of the proposed UPS-Student model for mitosis localization on the test subset of TUPAC16 dataset. The three high-magnification fields are extracted from the test subset. Green: true positive; Blue: false negative; Yellow: false positive.

Table 2
Performance comparison of the proposed model with existing methods on test subset of MITOS14 dataset. The presented figures of merit are F1-score, precision, and recall, obtained for mitosis detection. The best result for each metric is highlighted in bold, and the proposed methods are emphasized in gray. FT: Fine-tuning of the model on MITOS14 training split.

Method	F1 score	Recall	Precision	Multiple phases	Location supervision	External data	External validation
MiotsisDetection - Lei et al. (2021)	0.400	–	–	×	×	×	×
DeepMitosis - Li et al. (2018)	0.437	0.443	0.431	×	×	×	
MaskMitosis - Sebai et al. (2020)	0.475	0.453	0.500	×	×	×	
UTS - Student (ours) w/o FT	0.476	0.345	0.767		×		×
CasNN - Chen et al. (2016)	0.482	0.478		×			
FMDet - Wang et al. (2022)	0.490	0.556	0.438		×	×	×
UTS - Teacher (ours) w/o FT (Fernandez-Martín et al., 2022)	0.505	0.375	0.774				×
SegMitos - Li et al. (2019)	0.562	0.502	0.637		×		
Akram et al. (2018)	0.620	0.496	**0.828**	×	×	×	
UTS - Teacher (ours) w/ FT (Fernandez-Martín et al., 2022)	0.660	0.672	0.649				
UTS - Student (ours) w/ FT	0.689	**0.728**	0.654		×		
Rehman et al. (2022)[a]	**0.863**	–	–		×	×	

[a] Ensemble of hand-crafted features models.

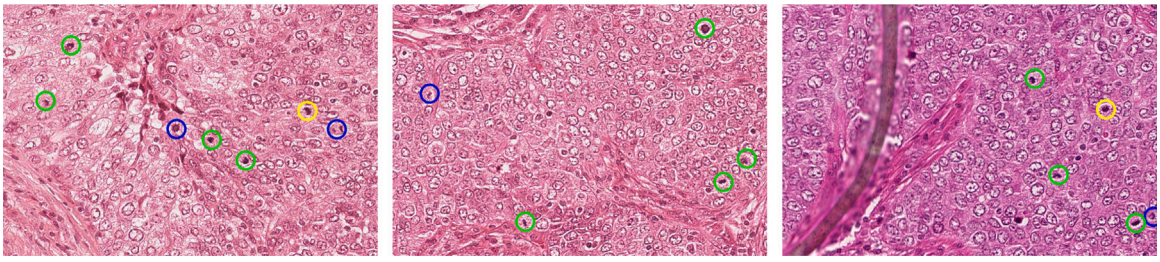


Fig. 4. Qualitative evaluation of the proposed UTS-Student model for mitosis localization on three high-magnification fields from the test subset of MITOS14 dataset. The first and second cases correspond to samples digitized using the Aperio scan, and the third case is a sample from a Hamamatsu scanning device. Green: true positive; Blue: false negative; Yellow: false positive.

of the Student-UTS model performance are introduced in Fig. 4. The obtained results using the models trained on the TUPAC16 dataset suffer a considerable performance drop when externally tested on MITOS14. Nevertheless, it can be noticed that this phenomenon is common in other works (see Table 2 FMDet (Wang et al., 2022) or MitosisDetection (Lei et al., 2021)). It is worth mentioning that still, the weakly supervised strategy achieves remarkable precision when tested under domain shift, in contrast to these deep learning methods that use supervision at the pixel level for training. In addition, the UTS method also shows promising performance with respect to other works that are directly trained on the target dataset (see Table 2 DeepMitosis (Li et al., 2018) or MaskMitosis (Sebai et al., 2020)). It is noticeable that the Teacher model of the UTS pipeline reaches the best results compared to the Student counterpart. Although the Student model is trained with a clean set of samples, the hard labeled samples are distilled with respect to the source domain annotators, which may not concur with the bias of the pathologists in the target domain, thus hindering the performance of the later. Once the UTS pipeline is trained on MITOS14, results are again competitive compared to the main core of neural-networks-based previous literature, and the Student model reaches an F1-score of 0.689. This corresponds to an improvement of nearly a 3% over the Teacher model, similar to the behavior observed on TUPAC16.

Generalization to other datasets: MIDOG21. In addition, experiments using the recently released MIDOG21 dataset are introduced. First, the direct inference of the model trained on TUPAC16 dataset without adaptation on this out-of-distribution (OOD) dataset is evaluated. Second, the UTS Teacher-Student setting is trained on the proposed MIDOG21 (in-distribution, ID) train split (see Section 4.2.1) under two scenarios: using the original clean dataset and using the same samples, but including the hard negative figures (MIDOG21_{HN}). In addition, although trained and tested on different, unaccessible data

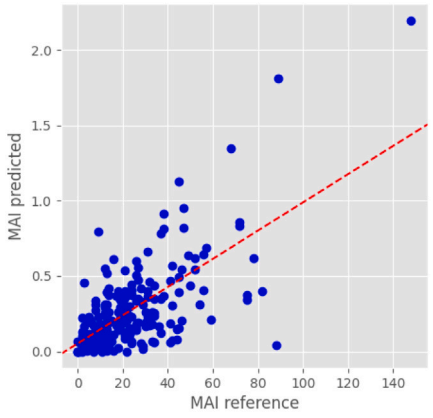
Table 3
Performance comparison of the proposed model with existing methods on MIDOG21 dataset. The presented figures of merit are F1-score. The proposed methods are emphasized in gray. ID: testing in-distribution -i.e. datasets used during training; OOD: testing on external datasets. HN: Training dataset containing annotated hard negatives.

Method	Training	MIDOG21 testing	
		ID	OOD
Challenge leaderboard			
Lb1 (Yang et al., 2021)	MIDOG21	0.793	0.714
Lb2 (Jahanifar et al., 2021)		0.837	0.717
Lb3 (Fick et al., 2021)		0.848	0.661
Other partitions			
RetinaNet (Aubreville, 2021)	TUPAC16	–	0.523
UTS - Teacher (Fernandez-Martín et al., 2022)		–	0.676
UTS - Student		–	0.664
UTS - Teacher (Fernandez-Martín et al., 2022)	MIDOG21	0.708	–
UTS - Student		0.713	–
UTS - Teacher (Fernandez-Martín et al., 2022)	$MIDOG21_{HN}$	0.604	–
UTS - Student		0.642	–

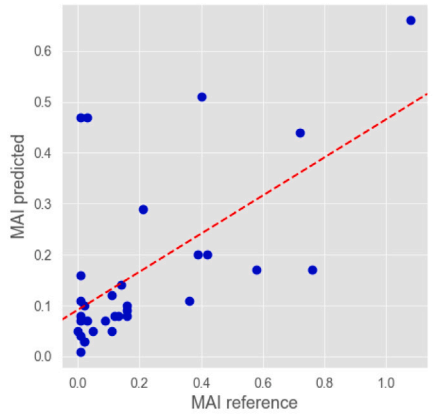
splits, different reference methods are included. First, the RetinaNet model trained on TUPAC16 and evaluated on MIDOG21 (Aubreville, 2021) is used. Second, the three leaderboard methods of the original challenge are included in the comparison. It is worth mentioning that these works include complementary domain-generalization strategies to tackle the competition objectives, such as Fourier-space-based stain augmentation (Yang et al., 2021), Cycle-GAN augmentation (Fick et al., 2021), or pixel-level annotations and multiple refinement stages (Jahanifar et al., 2021). For these works, the in-distribution score is obtained using the metrics found in Scanner A (which was available during the development stage) and the OOD performance using the average of Scanner D to F. The aforementioned results for mitosis localization in terms of F1-score are introduced in Table 3. As previously observed in MITOS14 experiments, direct generalization using the pretrained Student model on TUPAC16 is slightly worse compared to the Teacher model (*see Table 3, second block)*. Nevertheless, training the UTS pipeline in-domain shows again positive trends (*see Table 3, third block)*, which are exacerbated when training with noisy labeled mitosis, in which Student model brings substantial improvements of ∼4% (*see Table 3, last block)*. These observations suggest that the proposed hard negative distillation is especially effective when dealing with non-curated datasets such as TUPAC16 or $MIDOG21_{HN}$, but it might be less effective when different pathologists have carried out an in-depth inter-agreement annotation process, such as the clean MIDOG21. Finally, it is worth mentioning that the obtained OOD results approach the ones observed in the leaderboard challenge, in which range localization scores of $\sim[0.661, 0.717]$.

5.1.2. MAI counting at the WSI-level

The mitosis detection model trained on the TUPAC16 dataset was applied to the SUH dataset and the CCMCT-MEL to determine the proliferation score at the whole-slide image level. In the SUH dataset, the correlation between the automatically predicted proliferation score and the pathologist-annotated mitosis score was measured using Spearman's and Pearson's correlation coefficients, which were found to be 0.556 and 0.665, with a 95% confidence interval of $[0.454, 0.643]$ and $[0.564, 0.767]$, respectively. On the other hand, regarding the CCMCT-MEL dataset, the obtained Spearman's and Pearson's correlation coefficients between the annotated MAI and the predicted MAI in the hotspot ROI, were found to be 0.580 and 0.602, with a 95% confidence interval of $[0.277, 0.779]$ and $[0.305, 0.900]$, respectively. It is important to highlight that these results were obtained using the best threshold for binarizing the output probabilities determined for TUPAC16 dataset. However as there has not been any fine-tuning of the model parameters to these two external datasets, if this threshold is adapted to the new domains, evaluation metrics could increase (i.e. with by using few samples to adapt the threshold, Spearman's and Pearson's correlation coefficients could reach 0.666 and 0.738, respectively). The



(a) SUH



(b) CCMCT-MEL

Fig. 5. Scattered representation of the reference and predicted MAI on the SUH (*top*), and CCMCT-MEL (*bottom*) datsaets. The red line represents the linear adjustment between both variables.

corresponding scattered representations of reference and predicted MAI for the SUH and the CCMCT-MEL datasets are illustrated in Fig. 5. Note that the reference MAI follows the methodology performed by pathologists in the clinical practice, and the proposed predicted proliferation score is equivalent to the number of mitoses in one patch. These results underscore the remarkable generalization capability of the model across diverse datasets, spanning various organs and species, including both breast and cutaneous tumors in both human and canine

Table 4
Performance comparison of the different configurations of the WSS proposed model, in terms of aggregation strategies. Results are presented for mitosis localization and image-level classification.

Configuration	F1 score image-level	F1 score localization
Embedding - mean	0.762	0.134
Embedding - max	**0.772**	0.234
AttentionMIL (Ilse et al., 2018)	0.768	0.014
Instance - mean	0.753	0.004
Instance - max	0.761	**0.729**

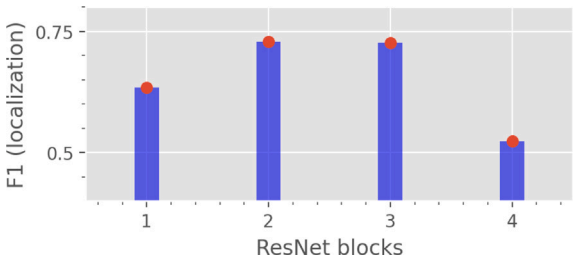


Fig. 6. Ablation study on the number of residual blocks used for feature extraction. Metric presented for mitosis localization.

beings. This adaptability highlights its potential as a robust and objective evaluation tool for tumor proliferation assessment, effectively mitigating the inherent observer variability associated with manual mitosis counting.

In Fig. 5(a), there is an outlier where the proposed model predicts a close-to-zero mitosis score per patch, whereas the pathologist annotated a considerably high MAI in that WSI. Upon a thorough review of this case with the pathologist, it was discovered that the reference MAI score was mainly calculated from an area rich in adipocytes that was discarded by the tissue segmentation model (TSM) for failing to meet the minimum tumor tissue threshold. Thus, this constitutes a limitation of the proposed method in evaluating the MAI score at the WSI level, as it relies on the TSM to identify tumor-containing patches. Additionally, in Fig. 5(b), two outliers are observed where the model predicts a high number of mitoses, whereas the number of mitoses in the hotspot ROI was close to zero. After a comprehensive examination of these two cases, it was determined that the two outliers could be mitigated by increasing the prediction probability threshold for considering a prediction as mitoses. Therefore, this constitutes a limitation of the different domain adaptability if no fine-tuning of the network parameters is performed to obtain the best performance in the new domains.

5.2. Ablation experiments

In the following, ablation experiments are depicted to motivate the choice of the different components of the proposed method. Unless stated the opposite, the ablation experiments were carried out on the TUPAC16-auxiliary dataset.

Weakly supervised setting. The study of the configuration of the Weakly Supervised Setting (WSS) model architecture begins with the exploration of various prominent configurations. This exploration encompasses embedding-based approaches that aggregate spatial features before reaching the classification layers and instance-based approaches that apply the classification layer spatially. Additionally, different aggregation methods, including mean and max operations, as well as the trainable attentionMIL mechanism (Ilse et al., 2018), are employed. The results are presented in Table 4. The figures of merit demonstrate that while all methods achieve similar results at the image level, only the instance-based approach with maximum aggregation exhibits satisfactory performance in mitosis localization. This is attributed to its unique ability to penalize false positive localization during training.

Table 5
Effect of directly integrating the location information into the weakly supervised formulation, depicted in Eq. (3), in the training subset. The metric presented is the F1-score for mitotic figure localization.

Method	Subset	
	Train	Test
Teacher	0.716	**0.729**
Teacher w/ L_c	**0.771**	0.673

On the importance of the feature complexity. Convolutional neural networks combine stacked convolutional and pooling operations, which merge spatial information. Thus, later layers in CNNs extract high-level features with complex shapes and low spatial resolution. Although CNNs for classification tasks usually benefit from deep structures, it is observed that spatial resolution and low-level features are vital for mitosis localization, as shown in Fig. 6. For that reason, only 3 residual blocks of ResNet-18 architecture were used for the proposed method.

Location constrained WSS. In the following, the effect of including the information regarding the mitosis position in the weakly supervised formulation during training is studied. The integration trough a centroid-based constraint formulation is described in Eq. (3). The Teacher model is trained, and the relative weight of the location constraint, β_{loc}, is empirically optimized to 0.01. Relevant performance metrics are recorded during training for both the constrained and unconstrained Teacher formulation, and these are depicted in Fig. 7. Concretely, the classification performance at the image level (*top*) and the constraint satisfaction (*bottom*), in the form of the euclidean distance, are presented. Also, Table 5 indicates the performance of the selected model on training and testing subsets for mitosis localization. Finally, qualitative visualization of the observed effect is depicted in Fig. 8. Results show that the location constraint successfully improves the location performance regarding the reference annotations (see Fig. 7 *bottom* and Table 5). However, the classification results at the image level in the validation subset worsen the more the constraint is met (see Fig. 7 *top*). These results suggest that forcing the model to focus on certain cell structures as mitotic figures hinders the model's generalization. This conclusion aligns with previous literature on noise distillation, highlighting that memorizing the training subset of uncertain cases might undermine its generalization (Arpit et al., 2017). Thus, the weakly supervised formulation (see Eq. (1)) guides the optimization more flexibly, as it leaves it up to the model to choose which cells in the image it considers anomalous based on their visual characteristics. Furthermore, this formulation, thanks to maximum aggregation, allows the classification as a mitosis of multiple cells in positive images, which may have been overlooked by the annotator.

Pseudolabeling vs. Discarding. The proposed UTS-Student model is trained on a clean subset of low-uncertainty samples. This subset is obtained discarding samples on the training subset, thanks to the flexibility of the weakly supervised setting to choose the mitotic figure on positive-labeled images during training, which may differ from those annotated - and in this case, they are discarded (see Section 3.3 for further details). Nevertheless, this process costs decreasing the number of samples used for training the Student model. Relevant prior literature regarding knowledge distillation on noisy datasets for image-level classification relies on pseudolabeling strategies (Zhang et al., 2020b; Shi and Jain, 2019), which present a softer transition between distilled subsets, and allow maintaining all available samples during training. Nevertheless, as introduced in Table 6, in this case, pseudolabeling offers fewer improvements compared to the proposed pipeline. This may be due to the incapacity of this approach to leverage noisy predictions that are actually correct at the global level, but they differ from the local figure used to establish this classification (see Fig. 8 for qualitative examples of these cases on the training subset).

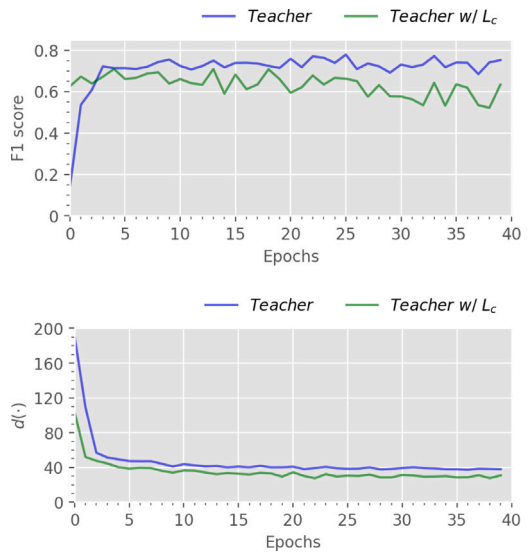


Fig. 7. Experiment on the effect of introducing a location constraint to the weakly supervised mitosis detection Teacher model. The model convergence is monitored in terms of the F1-score (*top*) of image-level predictions on the validation subset and the average euclidean distance of predicted and reference mitosis on the training dataset (*bottom*).

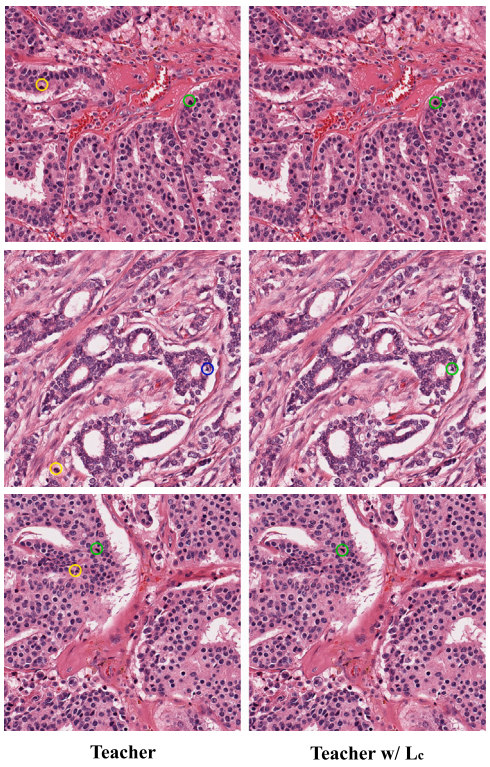


Fig. 8. Qualitative evaluation of the effect of introducing on the weakly supervised formulation (*left*) the location constraint depicted in Eq. (1) on the training subsets (*right*). Green: true positive; Blue: false negative; Yellow: false positive.

On the role of noise for uncertainty distillation. The investigation delving into the properties of the uncertainty distillation strategy within the Uninformed Teacher-Student (UTS) pipeline involves a closer examination. First, the impact of noise integration on the performance of both the Teacher and Student models in the context of strong augmentation is explored. This is accomplished by training the Teacher model separately using weak and strong augmentation. Subsequently,

Table 6
Ablation experiment on distilling the UTS-Teacher prediction on different strategies - i.e. pseudolabeling, or the proposed location-driven distillation.

	UTS - Student	
	Pseudolabels	Location distill. (Ours)
Localization F1-score	0.741	**0.767**

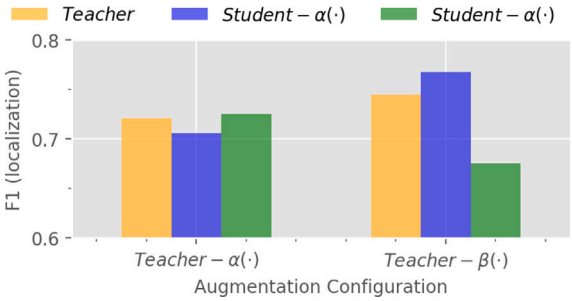


Fig. 9. Ablation study of the effect of noise integration, in the form of strong augmentations, to the distillation of clean samples for UTS-Student training. α: weak augmentations (e.g. random mirroring or rotations). β: strong augmentations (e.g. optical and grid distortion).

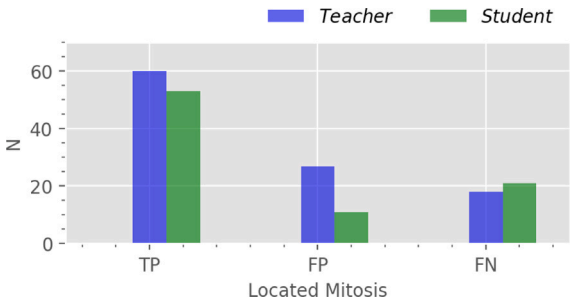


Fig. 10. Detailed results in terms of true positives (TP), false positives (FP), and false negatives (FN) mitosis detections of the proposed UTS Teacher and Student models on the test subset of TUPAC16 dataset.

Student models are trained for both options, employing weak and strong augmentation once more. Results for each pipeline are presented in Fig. 9. It can be observed that the pipeline using weak augmentation for Teacher model training reaches worse performance compared to the pipeline using strong augmentation, and any improvement is obtained by distilling hard samples for the Student training. When noise is incorporated into Teacher training in the form of strong augmentation, it contributes to a gain in generalization. Interestingly, after distilling uncertain samples, only weak augmentations reinforce the previously observed improvements. These results show the importance of noise to retrieve uncertain cases, but they also suggest that strong augmentation can hinder the model performance when using only distilled clean datasets.

In the following, the improvements observed in the F1-score for mitosis localization observed in the UTS pipeline between Teacher and Student modules are disentangled. The true positives (TP), false positives (FP), and false negatives (FN) detected in the test subset are presented in Fig. 10. The main contribution of training the Student model on clean, distilled samples is that it only produces classification on certain mitotic figures, which considerably reduces the number of false positives predictions while maintaining consistent performance in terms of true positives and false negatives.

Scaling performance trough model capacity. The results presented in this work are carried out using the popular lightweight ResNet-18 (RN18) architecture. Nevertheless, model capacity is a well-known

Table 7
Ablation experiment on the effect of model capacity on the UTS localization performance under hard-negative (HN) labeled samples during training, using the MIDOG21 dataset. Models are trained using ResNet-18, and a larger-capacity model, ResNet-50. The metric presented is the F1 score for mitosis localization.

Method	Architecture	MIDOG21 testing	
		HN	Clean
Training with clean MIDOG21			
Teacher	RN18	–	0.708
Student		–	0.713
Teacher	RN50	–	0.726
Student		–	**0.728**
Training with hard negatives - MIDOG21$_{HN}$			
Teacher	RN18	0.618	0.604
Student		0.646	0.642
Teacher	RN50	0.599	0.567
Student		0.637	**0.668**

critical factor for deep learning models' performance. Thus, Table 7 depicts the localization performance of UTS models using a larger-complexity backbone, ResNet-50 (RN50). The ablation experiment is performed on the MIDOG21 dataset, to evaluate the effect of this choice when dealing with hard negative samples. Results are presented in Table 7. Concretely, the ablation experiment consists of training UTS model using both architectures and training subsets (clean and with hard negatives labeled), and it is tested on both testing configurations.

Training the UTS pipeline with the RN50 configuration (*see* Table 7, *first block*) with the curated MIDOG21 subset, the model performance slightly increases in ∼1% for both Teacher and Student models. On the other hand, when using a noisy dataset (*see* Table 7, *second block*), the effect of model capacity is exacerbated. First, RN50 might be able to capture the hard negative patterns during Teacher training, thus performing remarkably worse on the clean testing subset. Nevertheless, this allows a better distillation process, and the Student model reaches an outstanding performance, showing improvements of ∼12% over the Teacher model. These results are closer to using the same configuration, but training on the manually cleaned dataset (only $\sim -6\%$ worse), compared to using RN18. It is worth mentioning that this model has been trained with ∼2714 hard negative mitotic figures, compared to a clean dataset with ∼1721 mitotic figures. These results demonstrate the promising performance of the proposed Uninformed Teacher-Student and the importance of model capacity for scaling its performance.

6. Discussion

Despite the recent advances in the literature for mitosis detection, existing literature still presents certain limitations in this difficult task. First, the state-of-the-art work is based on publicly available datasets from challenges, with the resulting ranking refinement heuristics common to these events, which emphasizes the difficulty of direct comparisons between solutions. This is exacerbated by the reporting of results in hidden test sets, which are no longer maintained over time by the organizers.

Because of all these limitations, in this work, the authors do not aim to present the proposed method as the solution to the problem of mitosis detection. Rather, they address the problem from a fresh perspective based on weakly supervised learning. This strategy allow dealing with the inherent noise in mitosis annotation, which presents large inter-annotator inter-variability in hard samples (Veta et al., 2016). The promising properties of this approach are reflected in the relative improvements observed when eliminating hard samples with the proposed uninformed distillation method. However, this work has certain limitations. First, the distillation stage results in the elimination of samples during training, which might reduce the generalization capabilities of the model by decreasing the data used. Additionally, the use of a two-stage Teacher-Student pipeline might be inefficient, and current practices on model distillation point out parallel and parameter-reused teacher networks as an interesting venue (Chen et al., 2022). By sharing parameters, distilling knowledge, and optimizing towards the same objectives, the parallel teacher-student approach may hold promise for further improving the performance and convergence properties. Second, the reported figures of merit are not directly comparable with the ones presented in some previous works due to the unavailability of the hidden test sets from the used datasets, the use of validation subsets, or the unavailability of official implementations. However, it should be noted that the authors have made every effort to ensure the proper validation and reproducibility of the proposed framework, including the evaluation of such methods under domain shift.

Finally, authors would like to highlight the scope shift from mitosis detection to WSI-level MAI scoring. While the main core of the previous literature focuses on individual mitosis detection, it is well-known that while two pathologists might differ between concrete mitotic figures, they present a large correlation for WSI-level counting (Veta et al., 2016), which is the ultimate goal of the diagnostic aid system. Thus, authors believe this task should be given greater attention in the future. Therefore, the proposed models have been evaluated for this task in two external databases, where new challenges have been opened. The way pathologists use only one subjectively selected high-power field, the representativeness of this area, and the robustness of tissue segmentation models as a pre-processing step for computer-aided solutions are appealing future research directions. Other promising avenues for future research could involve extending the proposed model to handle multiple cell types (i.e. typical and atypical mitoses). In this line, recent datasets, such as CCMCT (Bertram et al., 2019), provide valuable annotations for multiple cell types and their classification. Although our method is not primarily oriented to this objective, but to the overall assessment of mitotic activity within a tissue sample, currently our UTS framework is unable to tackle these particular inter-class uncertainties, which is an interesting venue for the future.

7. Conclusions

This work presents a novel deep learning model for weakly supervised mitosis localization on H&E breast cancer histology images. In addition, a Uninformed Teacher-Student pipeline, which takes advantage of the uninformed nature of weak supervision, is introduced for hard negative mining. Concretelly, this strategy leverages strong augmentations to distill hard samples and measure dissimilarities between the predicted and annotated mitosis. Comprehensive experiments have shown that this approach is competitive with state-of-the-art methods on three popular open-access datasets. These results have demonstrated the feasibility of the weakly supervised approach, which challenges the efficiency of previous methods that often require multiple stages and strong mitosis location information. In addition, training a Student model with a clean dataset, distilled trough the UTS pipeline, has shown consistent improvements of ∼4% on TUPAC16, MITOS14, and a version of MIDOG21 that includes unfiltered hard negative cases. Thus, the proposed UTS pipeline offers a promising strategy to deal with noisy annotations for mitosis detection. Additionally, the model trained on TUPAC16 was validated for the estimation of the mitotic activity index (MAI) at the whole slide image (WSI) level on two external datasets. The experiments have demonstrated a moderate correlation with regard pathologists annotations, and robust generalization capabilities to new datasets. These findings suggest the potential of this methodology as an objective evaluation tool for tumor proliferation in breast cancer, mitigating the inherent inter and intra-observer variability in manual mitosis counting.

CRediT authorship contribution statement

Claudio Fernandez-Martín: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Julio Silva-Rodriguez:** Conceptualization, Methodology, Software, Investigation, Writing – review & editing, Visualization. **Umay Kiraz:** Resources. **Sandra Morales:** Conceptualization, Supervision, Project administration, Funding acquisition. **Emiel A.M. Janssen:** Resources, Supervision, Project administration, Funding acquisition. **Valery Naranjo:** Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Claudio Fernandez-Martin reports financial support was provided by European Union. Sandra Morales reports financial support was provided by Polytechnic University of Valencia.

Funding

This work was funded by the Horizon 2020 European Union research and innovation programme under the Marie Sklodowska Curie grant agreement No 860627 (CLARIFY Project). The work of Sandra Morales has been co-funded by the Universitat Politècnica de València, Spain through the program PAID-10-20. The work of J. Silva-Rodríguez was carried out during his previous position at Universitat Politècnica de València. This work was partially funded by Generalitat Valenciana through project CIPROM/2022/20 and with Ayuda a Primeros Proyectos de Investigación (PAID-06-23), Vicerrectorado de Investigación of the Universitat Politècnica de València. Funding for open access charge: Universitat Politècnica de València (PAID-12-23).

Code availability

The code to train the proposed method is available through our GitHub repository (https://github.com/cvblab/Mitosis-UTS).

Data availability

The experiments carried out in this work rely largely on open-acces datasets. Concretelly, these can be accessed on the following links: TUPAC16 (https://tupac.grand-challenge.org/), MITOS14 (https://mitos-atypia-14.grand-challenge.org/), MIDOG21 (https://imig.science/midog2021/) and CCMCT (https://github.com/DeepMicroscopy/). The private dataset SUH is accessible upon request to the Stavanger University Hospital research group.

References

Akram, S.U., Qaiser, T., Graham, S., Kannala, J., Heikkilä, J., Rajpoot, N., 2018. Leveraging unlabeled whole-slide-images for mitosis detection. In: Medical Image Computing and Computer Assisted Intervention (MICCAI) - Workshop on Computational Pathology (COMPAY). 11039 LNCS, pp. 69–77.

Arpit, D., Jastrzębski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M.S., Maharaj, T., Fischer, A., Courville, A., Bengio, Y., Lacoste-Julien, S., 2017. A closer look at memorization in deep networks. In: Proceedings of the International Conference on Machine Learning (ICML). pp. 1–10.

Arvaniti, E., Fricker, K.S., Moret, M., Rupp, N., Hermanns, T., Fankhauser, C., Wey, N., Wild, P.J., Ruschoff, J.H., Claassen, M., 2018. Automated Gleason grading of prostate cancer tissue microarrays via deep learning. Sci. Rep. 8 (1).

Aubreville, M., 2021. Quantifying the scanner-induced domain gap in mitosis detection. In: Medical Image with Deep Learning (MIDL).

Aubreville, M., Bertram, C., Veta, M., Klopfleisch, R., Stathonikos, N., Breininger, K., ter Hoeve, N., Ciompi, F., Maier, A., 2021. Quantifying the scanner-induced domain gap in mitosis detection. In: Medical Image with Deep Learning (MIDL) - Short Paper. pp. 1–4.

Aubreville, M., Stathonikos, N., Bertram, C.A., Klopfleisch, R., ter Hoeve, N., Ciompi, F., Wilm, F., Marzahl, C., Donovan, T.A., Maier, A., Breen, J., Ravikumar, N., Chung, Y., Park, J., Nateghi, R., Pourakpour, F., Fick, R.H., Ben Hadj, S., Jahanifar, M., Shephard, A., Dexl, J., Wittenberg, T., Kondo, S., Lafarge, M.W., Koelzer, V.H., Liang, J., Wang, Y., Long, X., Liu, J., Razavi, S., Khademi, A., Yang, S., Wang, X., Erber, R., Klang, A., Lipnik, K., Bolfa, P., Dark, M.J., Wasinger, G., Veta, M., Breininger, K., 2023. Mitosis domain generalization in histopathology images — The MIDOG challenge. Med. Image Anal. 84, 102699.

Baak, J.P., van Diest, P.J., Voorhorst, F.J., van der Wall, E., Beex, L.V., Vermorken, J.B., Janssen, E.A., 2005. Prospective multicenter validation of the independent prognostic value of the mitotic activity index in lymph node-negative breast cancer patients younger than 55 years. J. Clin. Oncol. 23 (25), 1–9.

Belharbi, S., Rony, J., Dolz, J., Ayed, I.B., Mccaffrey, L., Granger, E., 2021. Deep interpretable classification and weakly-supervised segmentation of histology images via max-min uncertainty. IEEE Trans. Med. Imaging 41 (3), 1.

Bernhardt, M., Castro, D.C., Tanno, R., Schwaighofer, A., Tezcan, K.C., Monteiro, M., Bannur, S., Lungren, M.P., Nori, A., Glocker, B., Alvarez-Valle, J., Oktay, O., 2022. Active label cleaning for improved dataset quality under resource constraints. Nature Commun. 13 (1).

Bertram, C.A., Aubreville, M., Marzahl, C., et al., 2019. A large-scale dataset for mitotic figure assessment on whole slide images of canine cutaneous mast cell tumor. Sci. Data 6, 274.

Chen, H., Dou, Q., Wang, X., Qin, J., Heng, P.-A., 2016. Mitosis detection in breast cancer histology images via deep cascaded networks. In: AAAI Conference on Artificial Intelligence. pp. 1160–1166.

Chen, D., Mei, J.-P., Zhang, H., Wang, C., Feng, Y., Chen, C., 2022. Knowledge distillation with the reused teacher classifier. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11933–11942.

Ciresan, D.C., Giusti, A., Gambardella, L.M., Schmidhuber, J., 2013a. Mitosis detection in breast cancer histology images with deep neural networks. In: Medical Image Computing and Computer Assisted Intervention (MICCAI). pp. 411–418.

Ciresan, D.C., Meier, U., Masci, J., Gambardella, L.M., Urgen Schmidhuber, J., 2013b. Flexible, high performance convolutional neural networks for image classification. In: Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence Flexible. pp. 1237–1242.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. ImageNet: A large-scale hierarchical image database. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1–8.

Elmore, J.G., Longton, G.M., Carney, P.A., Geller, B.M., Onega, T., Tosteson, A.N., Nelson, H.D., Pepe, M.S., Allison, K.H., Schnitt, S.J., O'Malley, F.P., Weaver, D.L., 2015. Diagnostic concordance among pathologists interpreting breast biopsy specimens. JAMA - J. Am. Med. Assoc. 313 (11), 1112–1132.

Fernandez-Martín, C., Kiraz, U., Silva-Rodríguez, J., Morales, S., Janssen, E.A.M., Naranjo, V., 2022. Challenging mitosis detection algorithms: Global labels allow centroid localization. In: International Conference on Intelligent Data Engineering and Automated Learning (IDEAL).

Fick, R.H.J., Moshayedi, A., Roy, G., Dedieu, J., Petit, S., Hadj, S.B., 2021. Domainspecific cycle-GAN augmentation improves domain generalizability for mitosis detection. In: Biomedical Image Registration, Domain Generalisation and Out-of-Distribution Analysis. MICCAI MIDOG Challenge.

Ghesu, F.C., Georgescu, B., Mansoor, A., Yoo, Y., Gibson, E., Vishwanath, R.S., Balachandran, A., Balter, J.M., Cao, Y., Singh, R., Digumarthy, S.R., Kalra, M.K., Grbic, S., Comaniciu, D., 2021. Quantifying and leveraging predictive uncertainty for medical image assessment. Med. Image Anal. 68.

Guo, S., Huang, W., Zhang, H., Zhuang, C., Dong, D., Scott, M.R., Huang, D., 2018. CurriculumNet: Weakly supervised learning from large-scale web images. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 1–16.

Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., Sugiyama, M., 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In: Advances in Neural Information Processing Systems (NeurIPS). pp. 1–13.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1–12.

Hwang, M., Wang, D., Wu, C., Jiang, W.C., Kong, X.X., Hwang, K.S., Ding, K., 2020. A fuzzy segmentation method to learn classification of mitosis. Int. J. Fuzzy Syst. 22 (5).

Ilse, M., Tomczak, J.M., Welling, M., 2018. Attention-based deep multiple instance learning. In: International Conference on Machine Learning (ICML). pp. 1–16.

Jahanifar, M., Shepard, A., Zamanitajeddin, N., Bashir, R.M.S., Bilal, M., Khurram, S.A., Minhas, F., Rajpoot, N., 2021. Stain-robust mitotic figure detection for the mitosis domain generalization challenge. In: Biomedical Image Registration, Domain Generalisation and Out-of-Distribution Analysis. MICCAI MIDOG Challenge.

Jia, Z., Huang, X., Chang, E.I.-C., Xu, Y., 2017. Constrained deep weak supervision for histopathology image segmentation. IEEE Trans. Med. Imaging 36.

Ju, L., Wang, X., Wang, L., Mahapatra, D., Zhao, X., Zhou, Q., Liu, T., Ge, Z., 2022. Improving medical images classification with label noise using dual-uncertainty estimation. IEEE Trans. Med. Imaging 41 (6).

Kervadec, H., Dolz, J., Yuan, J., Desrosiers, C., Granger, E., Ayed, I.B., 2022. Constrained deep networks: Lagrangian optimization via log-barrier extensions. In: European Signal Processing Conference.

Lafarge, M.W., Pluim, J.P.W., Eppenhof, K.A.J., Moeskops, P., Veta, M., 2017. Domain-adversarial neural networks to address the appearance variability of histopathology images. In: Medical Image Computing and Computer Assisted Intervention (MICCAI) - Workshop on Deep Learning in Medical Image Analysis (DLMIA). pp. 1–8.

Lei, H., Liu, S., Elazab, A., Gong, X., Lei, B., 2021. Attention-guided multi-branch convolutional neural network for Mitosis detection from Histopathological images. IEEE J. Biomed. Health Inf. 25 (2), 358–370.

Leibig, C., Allken, V., Ayhan, M.S., Berens, P., Wahl, S., 2017. Leveraging uncertainty information from deep neural networks for disease detection. Sci. Rep. 7 (1).

Li, C., Wang, X., Liu, W., Latecki, L.J., 2018. DeepMitosis: Mitosis detection via deep detection, verification and segmentation networks. Med. Image Anal. 45 (1), 1–13.

Li, C., Wang, X., Liu, W., Latecki, L.J., Wang, B., Huang, J., 2019. Weakly supervised mitosis detection in breast histopathology images using concentric loss. Med. Image Anal. 53, 165–178.

López-Pérez, M., Morales-Álvarez, P., Cooper, L.A.D., Molina, R., Katsaggelos, A.K., 2023. Deep Gaussian processes for classification with multiple noisy annotators. Application to breast cancer tissue classification. IEEE Access 11.

Lu, W., 2021. A two-phase mitosis detection approach based on U-shaped network. BioMed Res. Int. 2021.

Ludovic, R., Daniel, R., Nicolas, L., Maria, K., Humayun, I., Jacques, K., Frédérique, C., Catherine, G., Gilles, L.N., Metin N, G., 2013. Mitosis detection in breast cancer histological images An ICPR 2012 contest. J. Pathol. Inform. 4 (8), 1–7.

Macenko, M., et al., 2009. A method for normalizing histology slides for quantitative analysis. In: International Symposium on Biomedical Imaging (ISBI). pp. 1107–1110.

Mahmood, T., Arsalan, M., Owais, M., Lee, M.B., Park, K.R., 2020. Artificial intelligence-based mitosis detection in breast cancer histopathology images using faster R-CNN and deep CNNs. J. Clin. Med. 9 (3), 1–25.

Maroof, N., Khan, A., Qureshi, S.A., Rehman, A.u., Khalil, R.K., Shim, S.O., 2020. Mitosis detection in breast cancer histopathology images using hybrid feature space. Photodiagn. Photodyn. Ther. 31.

Nateghi, R., Danyali, H., Helfroush, M.S., 2021. A deep learning approach for mitosis detection: Application in tumor proliferation prediction from whole slide images. Artif. Intell. Med. 114 (1), 1–10.

Oquab, M., Laptev, I., Sivic, J., 2015. Is object localization for free? Weakly-supervised learning with convolutional neural networks. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). pp. 685–694.

Paeng, K., Hwang, S., Park, S., Kim, M., 2017. A unified framework for tumor proliferation score prediction in breast histopathology. In: Medical Image Computing and Computer Assisted Intervention (MICCAI) - Workshop on Deep Learning in Medical Image Analysis (DLMIA). pp. 1–8.

Pathak, D., Krähenbühl, P., Darrell, T., 2015. Constrained convolutional neural networks for weakly supervised segmentation. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 1–12.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Rehman, M.U., Akhtar, S., Zakwan, M., Mahmood, M.H., 2022. Novel architecture with selected feature vector for effective classification of mitotic and non-mitotic cells in breast cancer histology images. Biomed. Signal Process. Control 71, 103212.

Roux, L., Racoceanu, D., Capron, F., Calvo, J., Attieh, E., Naour, G.L., Gloaguen, A., 2014. MITOS & ATYPIA detection of mitosis and evaluation of nuclear atypia score in breast cancer histological images. In: International Conference on Pattern Recognition (ICPR). pp. 1–8.

Sabeena Beevi, K., Nair, M.S., Bindu, G.R., 2019. Automatic mitosis detection in breast histopathology images using convolutional neural network based deep transfer learning. Biocybern. Biomed. Eng. 39.

Saha, M., Chakraborty, C., Racoceanu, D., 2018. Efficient deep learning model for mitosis detection using breast histopathology images. Comput. Med. Imaging Graph. 64.

Sebai, M., Wang, X., Wang, T., 2020. MaskMitosis: a deep learning framework for fully supervised, weakly supervised, and unsupervised mitosis detection in histopathology images. Med. Biol. Eng. Comput. 58 (7).

Shi, Y., Jain, A.K., 2019. Probabilistic face embeddings. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV).

Sigirci, I.O., Albayrak, A., Bilgin, G., 2022. Detection of mitotic cells in breast cancer histopathological images using deep versus handcrafted features. Multimedia Tools Appl. 81.

Silva-Rodrguez, J., Schmidt, A., Sales, M.A., Molina, R., Naranjo, V., 2022. Proportion constrained weakly supervised histopathology image classification. Comput. Biol. Med. 147.

Silva-Rodríguez, J., Colomer, A., Naranjo, V., 2021. WeGleNet: A weakly-supervised convolutional neural network for the semantic segmentation of gleason grades in prostate histology images. Comput. Med. Imaging Graph. 88 (1), 1–10.

Silva-Rodríguez, J., Colomer, A., Sales, M.a.A., Molina, R., Naranjo, V., 2020. Going deeper through the gleason scoring scale: An automatic end-to-end system for histology prostate grading and cribriform pattern detection. Comput. Methods Programs Biomed. 195 (1), 1–18.

Sohail, A., Khan, A., Wahab, N., Zameer, A., Khan, S., 2021. A multi-phase deep CNN based mitosis detection framework for breast cancer histopathological images. Sci. Rep. 11 (1), 1–18.

Srinidhi, C.L., Ciga, O., Martel, A.L., 2021. Deep neural network models for computational histopathology: A survey. Med. Image Anal. 67, 101813.

Tellez, D., Balkenhol, M., Otte-Höller, I., Van De Loo, R., Vogels, R., Bult, P., Wauters, C., Vreuls, W., Mol, S., Karssemeijer, N., Litjens, G., Van Der Laak, J., Ciompi, F., 2018. Whole-slide mitosis detection in H&E breast histology using PHH3 as a reference to train distilled stain-invariant convolutional networks. IEEE Trans. Med. Imaging 37 (9).

Veta, M., Heng, Y.J., Stathonikos, N., Bejnordi, B.E., Beca, F., Wollmann, T., Rohr, K., Shah, M.A., Wang, D., Rousson, M., Hedlund, M., Tellez, D., Ciompi, F., Zerhouni, E., Lanyi, D., Viana, M., Kovalev, V., Liauchuk, V., Phoulady, H.A., Qaiser, T., Graham, S., Rajpoot, N., Sjöblom, E., Molin, J., Paeng, K., Hwang, S., Park, S., Jia, Z., Chang, E.I., Xu, Y., Beck, A.H., van Diest, P.J., Pluim, J.P., 2019. Predicting breast tumor proliferation from whole-slide images: The TUPAC16 challenge. Med. Image Anal. 54, 111–121.

Veta, M., Van Diest, P.J., Jiwa, M., Al-Janabi, S., Pluim, J.P., 2016. Mitosis counting in breast cancer: Object-level interobserver agreement and comparison to an automatic method. PLoS One 11 (8), 1–13.

Wahab, N., Khan, A., Lee, Y.S., 2019. Transfer learning based deep CNN for segmentation and detection of mitoses in breast cancer histopathological images. Microscopy 68 (3), 216–233.

Wang, Y., Zhang, J., Kan, M., Shan, S., Chen, X., 2020. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12275–12284.

Wang, X., Zhang, J., Yang, S., Xiang, J., Luo, F., Wang, M., Zhang, J., Yang, W., Huang, J., Han, X., 2022. A generalizable and robust deep learning algorithm for mitosis detection in multicenter breast histopathological images. Med. Image Anal. 1 (1), 1–20.

Wollmann, T., Rohr, K., 2021. Deep consensus network: Aggregating predictions to improve object detection in microscopy images. Med. Image Anal. 70 (1), 1–14.

Yang, S., Luo, F., Zhang, J., Wang, X., 2021. Sk-unet model with Fourier domain for mitosis detection. In: Biomedical Image Registration, Domain Generalisation and Out-of-Distribution Analysis. MICCAI MIDOG Challenge.

Zerhouni, E., Lanyi, D., Viana, M., Gabrani, M., 2017. Wide residual networks for mitosis detection. In: International Symposium on Biomedical Imaging (ISBI). pp. 924–928.

Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O., 2016. Understanding deep learning requires rethinking generalization. In: Proceedings of the Internaitonal Conference on Learning Representations (ICLR). pp. 1–15.

Zhang, Z., Zhang, H., Arik, S.O., Lee, H., Pfister, T., 2020a. Distilling effective supervision from severe label noise. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1–11.

Zhang, Z., Zhang, H., Arik, S.O., Lee, H., Pfister, T., 2020b. Distilling effective supervision from severe label noise. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR).