Computerized Medical Imaging and Graphics 112 (2024) 102327

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

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

Learning from crowds for automated histopathological image segmentation ☆

Miguel López-Pérez [a,*], Pablo Morales-Álvarez [b], Lee A.D. Cooper [c,d], Christopher Felicelli [c], Jeffery Goldstein [c], Brian Vadasz [c], Rafael Molina [a], Aggelos K. Katsaggelos [d,e]

[a] Department of Computer Science and Artificial Intelligence, University of Granada, Spain
[b] Department of Statistics and Operations Research, University of Granada, Spain
[c] Department of Pathology at Northwestern University, Chicago, USA
[d] Center for Computational Imaging and Signal Analytics, Northwestern University, Chicago, USA
[e] Department of Electrical and Computer Engineering at Northwestern University, Chicago, USA

ARTICLE INFO

Keywords:
Segmentation
Histopathology
Crowdsourcing
Cancer
Noisy labels

ABSTRACT

Automated semantic segmentation of histopathological images is an essential task in Computational Pathology (CPATH). The main limitation of Deep Learning (DL) to address this task is the scarcity of expert annotations. Crowdsourcing (CR) has emerged as a promising solution to reduce the individual (expert) annotation cost by distributing the labeling effort among a group of (non-expert) annotators. Extracting knowledge in this scenario is challenging, as it involves noisy annotations. Jointly learning the underlying (expert) segmentation and the annotators' expertise is currently a commonly used approach. Unfortunately, this approach is frequently carried out by learning a different neural network for each annotator, which scales poorly when the number of annotators grows. For this reason, this strategy cannot be easily applied to real-world CPATH segmentation. This paper proposes a new family of methods for CR segmentation of histopathological images. Our approach consists of two coupled networks: a segmentation network (for learning the expert segmentation) and an annotator network (for learning the annotators' expertise). We propose to estimate the annotators' behavior with only one network that receives the annotator ID as input, achieving scalability on the number of annotators. Our family is composed of three different models for the annotator network. Within this family, we propose a novel modeling of the annotator network in the CR segmentation literature, which considers the global features of the image. We validate our methods on a real-world dataset of Triple Negative Breast Cancer images labeled by several medical students. Our new CR modeling achieves a Dice coefficient of 0.7827, outperforming the well-known STAPLE (0.7039) and being competitive with the supervised method with expert labels (0.7723). The code is available at https://github.com/wizmik12/CRowd_Seg.

1. Introduction

Computational Pathology (CPATH) has made a breakthrough by incorporating Deep Learning (DL) to automatize relevant tasks in the analysis of histopathological Whole-Slide Images (WSIs) (Litjens et al., 2017; Aatresh et al., 2021; Van der Laak et al., 2021; Khaliliboroujeni et al., 2022). Among others, automated semantic segmentation has become a crucial task in CPATH (Foucart et al., 2022). Semantic segmentation decomposes the image's semantic content into multiple segments, e.g., which pixels belong to the tumor region. This estimation can quantify established biomarkers used in clinical practice for diagnosis and prognosis, thus assisting and supporting pathologists in decision-making (Salgado et al., 2015).

The main limitation of DL methods for segmentation in CPATH is the scarcity of expert annotations (Ben Hamida et al., 2022). The labeling process is time-consuming, challenging, and demanding for expert pathologists, who have limited availability. *Crowdsourcing* (CR) has emerged as a promising approach to labeling histopathological images. CR distributes the labeling effort by involving a large crowd with varying degrees of expertise. This strategy speeds up the labeling process and significantly reduces the expert's workload at the

☆ Supported by Spanish Ministry of Science and Innovation under Project PID2022-140189OB-C22, FEDER/Junta de Andalucia under Project P20_00286, FEDER/Junta de Andalucía, and Universidad de Granada under Project B-TIC-324-UGR20, and "Contrato puente" of the University of Granada. Funding for open access charge: Universidad de Granada / CBUA.

* Corresponding author.
E-mail addresses: mlopez@decsai.ugr.es (M. López-Pérez), pablomorales@ugr.es (P. Morales-Álvarez), lee.cooper@northwestern.edu (L.A.D. Cooper), rms@decsai.ugr.es (R. Molina), a-katsaggelos@northwestern.edu (A.K. Katsaggelos).

https://doi.org/10.1016/j.compmedimag.2024.102327
Received 16 June 2023; Received in revised form 20 October 2023; Accepted 12 December 2023
Available online 5 January 2024
0895-6111/© 2024 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

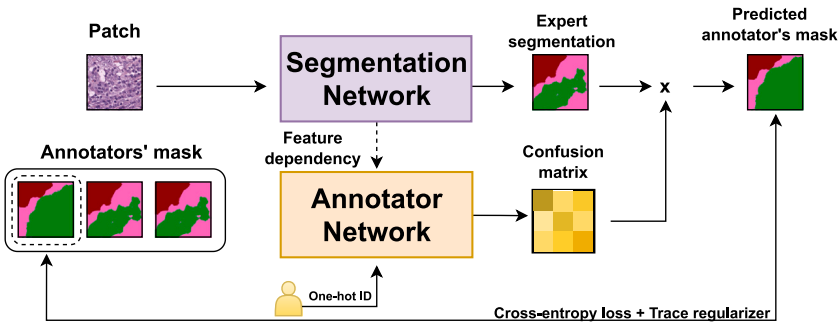

Fig. 1. General overview of the proposed framework for one-stage crowdsourcing segmentation. We propose two coupled networks: segmentation and annotator networks. The segmentation network aims to generate latent (expert) segmentation, and the annotator network provides the estimated confusion matrix for the annotator. Combining pixel-wise both outcomes, we generate the observed noisy masks of the multiple participants.

expense of introducing noisy labels by less experienced annotators. CR approaches have been utilized for labeling histological images and successfully applied to different tasks in CPATH (Grote et al., 2018; Amgad et al., 2019; Le et al., 2020; Amgad et al., 2022). However, expert pathologists reviewed and approved manual annotations in all cases, which is still tedious and challenging.

The need for experts can be avoided by directly learning the DL model using the (non-expert) CR annotations. In this case, the DL method has to be adapted to cope with the noisy labels provided by multiple annotators (Rodrigues and Pereira, 2018; Yang et al., 2022; Jiang et al., 2022). There are two big groups of methods to tackle this: two-stage and one-stage (end-to-end) methods. The two-stage methods aggregate the labels in a previous step, and then apply a supervised method. The most straightforward way to perform this approach is Majority Voting (MV). Notice that MV assumes that every annotator is equally reliable. When the crowd is not equally reliable and contains noisy annotators, methods achieve better results by estimating the expertise of the different annotators. For example, the well-known STAPLE iteratively estimates the annotators' expertise and the latent (expert) segmentation for each image independently (Warfield et al., 2004). STAPLE is widely applied to generate ground-truth annotations from several noisy annotators in segmentation tasks (Commowick et al., 2018; Nir et al., 2018).

In the two-stage CR methods, such as STAPLE and MV, label aggregation and model training are isolated processes. In contrast, one-stage CR methods jointly learn the model and the annotators' expertise, yielding better performance (Tanno et al., 2019; Karimi et al., 2020). This enhancement has been widely described in histopathological tissue classification with CR labels (Nir et al., 2018; Albarqouni et al., 2016; López-Pérez et al., 2021, 2023). Regarding CR segmentation, which is more complex than classification, one-stage approaches are challenging to implement. The first and only study proposing a one-stage CR method for segmentation was (Zhang et al., 2020), where the authors applied the method to toy examples and Computerized Tomography (CT) scan images. This method cannot be easily extended to real-world CPATH datasets because the proposed architecture scales poorly with the number of annotators. Specifically, they introduce a different neural network to model the behavior of each annotator, which hampers its application beyond a few annotators (up to five are used in their paper). Therefore, a scalable one-stage CR method has yet to be proposed and studied for real-world CPATH segmentation.

This paper proposes a new framework for one-stage CR semantic segmentation applied to real-world histopathological images. We propose a framework composed of two coupled networks: a segmentation network based on the U-Net architecture (Ronneberger et al., 2015) and an annotator network, see Fig. 1. The segmentation network aims to provide the expert (latent) segmentation. The annotator network predicts the confusion matrices (CMs), which model annotators' biases on the annotation. Since annotator modeling is a challenging and open problem, we propose three approaches with the following hierarchy:

1. CR Global: there is one CM for each annotator, but the CM does not depend on the patch or image that is being analyzed.
2. CR Image: the CM additionally depends on the patch or image being analyzed, but the same CM is applied to all the pixels of the patch/image.
3. CR Pixel: the CM depends on the annotator and the patch/image being analyzed, and one CM is obtained for each pixel.

Whereas the first and last ones had already been used in Zhang et al. (2020), the second one is a novel contribution that provides a trade-off between the other two. In the experiments, we will analyze the three approaches.

The contributions of this paper are summarized as follows:

- We propose a novel family of one-stage scalable CR methods for the segmentation of images labeled by multiple noisy annotators in CPATH. In contrast to widely used two-stage approaches (Warfield et al., 2004), our family is the first one in CPATH that learns the expert (latent) segmentation and the CM of the annotators jointly from noisy masks.
- Our novel CR family estimates the CM of every annotator with only one network. This modeling has two great advantages: (i) it enables involving a large crowd; (ii) it optimizes the annotator network with the information provided by every annotator. In contrast, other works resort to building several networks (one per annotator) (Zhang et al., 2020). This approach is not feasible in real-world CPATH tasks where a large crowd is hired (the method quickly goes out of memory), and each network would only be fed by the data of their respective annotator.
- We propose a novel modeling of the annotator network (CR Image). This modeling predicts a common CM for all the pixels within the same image. With this paradigm, we suggest that global features influence the annotators on the image. In contrast, other works estimate the CM for each pixel independently (Zhang et al., 2020).
- We apply our family of CR methods to a real-world Triple Negative Breast Cancer (TNBC) dataset labeled by 20 medical students. This is out of the scope for previous approach (Zhang et al., 2020), yielding memory overflow. A comprehensive evaluation of our methods on this real-world dataset demonstrates superior performance than the state-of-the-art STAPLE. Furthermore, our methods with noisy crowdsourcing annotations are competitive with the supervised model with expert annotations.

The remainder of the paper is structured as follows. Section 2 presents the proposed family of crowdsourcing methods. Section 3 details the experimental setup. Section 4 shows the experimental results. Section 5 provides the conclusions and outlines future lines of work. Appendix includes additional figures for better visualization of the results.

2. Methodology

2.1. Problem formulation

In this work, we address the use of crowdsourcing semantic segmentation techniques in CPATH. We have observed a training noisy dataset $\mathcal{D} = (\mathbf{X}, \tilde{\mathbf{Y}}) = \{(\mathbf{x}_n, \tilde{\mathbf{y}}_n^r) : n = 1, \dots, N; r \in R_n\}$, where $\mathbf{x}_n \in \mathbb{R}^{W \times H \times 3}$ is the histopathological patch and $\tilde{\mathbf{y}}_n^r \in \mathcal{Y}^{W \times H}$ is the segmentation mask provided by the rth annotator for the nth instance. We represent by $\mathcal{Y} = \{1, \dots, L\}$ the set of classes for our problem. Since each annotator may have annotated only some patches, we denote by $R_n \subseteq \{1, \dots, R\}$ the subset of annotators that provided the label for the nth patch. We assume that every patch has been annotated by at least one participant, i.e., $|R_n| \geq 1$. We also assume that there exists a true (expert) label for each patch $\mathbf{Y} = \{\mathbf{y}_n \in \mathcal{Y}^{W \times H}\}_{n=1,\dots,N}$ but this is not available during training, only in test to validate the models. The final goal is to jointly estimate each annotator's expertise and a segmentation model that predicts the expert (latent) segmentation $\mathbf{Y}$ from the observed noisy dataset $(\mathbf{X}, \tilde{\mathbf{Y}})$.

2.2. Probabilistic model

First, we make two assumptions of independence on the observed crowdsourcing labels (Zhang et al., 2020). Given the input patches: (1) the annotators independently provide crowdsourcing labels; (2) annotations over different pixels are independent. Hence, the observation model is given by

$$\mathrm{p}(\tilde{\mathbf{Y}}|\mathbf{X}) = \prod_{n=1}^{N} \prod_{r \in R_n} \prod_{w=1}^{W} \prod_{h=1}^{H} \mathrm{p}(\tilde{y}_{n,w,h}^{(r)}|\mathbf{x}_n). \tag{1}$$

Second, we assume that those observed noisy masks depend on the pixel's expert class and the annotator's expertise (Zhang et al., 2020; Tanno et al., 2019). Then, the likelihood of the observed noisy mask $\tilde{y}_n^{(r)}$ is modeled pixel-wise,

$$\mathrm{p}(\tilde{y}_{n,w,h}^{(r)}|\mathbf{x}_n) = \sum_{l=1}^{L} \mathrm{p}(\tilde{y}_{n,w,h}^{(r)}|y_{n,w,h} = l, \mathbf{x}_n)\mathrm{p}(y_{n,w,h} = l|\mathbf{x}_n), \tag{2}$$

where $\mathrm{p}(y_{n,h,w}|\mathbf{x}_n)$ represents the expert label distribution over the (w, h)th pixel of the nth patch and $\mathrm{p}(\tilde{y}_{n,w,h}^{(r)}|y_{n,w,h}, \mathbf{x}_n)$ how the rth annotator generates the noisy label given the expert (latent) label and the observed patch. To encode this distribution, we utilize a pixel-wise Confusion Matrix (CM) of size $L \times L$ per annotator, denoted by $\mathbf{A}^{(r)}(\mathbf{x}_n) \in [0, 1]^{W \times H \times L \times L}$. For each pixel, the ijth element of the CM represents $\mathrm{p}(\tilde{y}_{n,w,h}^{(r)} = i|y_{n,w,h} = j, \mathbf{x}_n)$. The probability that the rth annotator labels as class i a pixel whose expert (latent) class is j.

Our CR approach couples two networks: a segmentation network and an annotator network. For each patch $\mathbf{x}_n$, the segmentation network, parametrized by θ, provides an estimation, $\mathbf{p}_\theta(\mathbf{x}_n) \in \mathbb{R}^{W \times H \times L}$, of the expert (latent) segmentation while the annotator network, parametrized by ϕ, provides an estimation, $\{\mathbf{A}_\phi^{(r)}(\mathbf{x}_n)\}_{r=1}^{R}$, of the annotators' CMs. Subscripts θ and ϕ are employed to represent the estimated values derived from neural networks, while if they do not appear, they will refer to the true unknown values. Then, the product $\mathbf{p}_{\theta,\phi}^{(r)}(\mathbf{x}_n) := \mathbf{A}_\phi^{(r)}(\mathbf{x}_n) \cdot \mathbf{p}_\theta(\mathbf{x}_n)$ defines the estimated segmentation probability mask of each annotator. The "." refers to the element-wise matrix multiplications in the spatial W, H dimensions. At inference time, the output of the segmentation network $\mathbf{p}_\theta(\mathbf{x}_n)$ yields the estimation of the expert (latent) segmentation mask $\mathrm{p}(\mathbf{x}_n)$. In the following subsections, we describe the architecture of both networks and how the model is trained.

2.3. Segmentation network

The segmentation network aims to estimate the expert (latent) segmentation. Here, we rely on state-of-the-art segmentation architectures for supervised problems. We choose an encoder–decoder architecture based on the U-Net because of its success in biomedical domains (Ronneberger et al., 2015). Its combination of feature maps at multiple scales makes it a suitable option for CPATH segmentation tasks. We also consider ResNet34 as the backbone for the encoder part (He et al., 2016). ResNet is a prominent family of deep Neural Network architectures. They introduced the skip connection, also known as residual connections, to avoid information loss during the training of deep networks. Skip Connections enable to train very deep networks and boost the network's performance. Since segmentation problems in CPATH are challenging, we fine-tune the segmentation network with pre-trained weights on ImageNet.

2.4. Annotator network

The annotator network aims to estimate the annotators' CM, $\mathbf{A}^{(r)}(\mathbf{x}_n)$ (i.e., the expertise of each annotator). We utilize a deep neural network $\mathbf{A}_\phi^{(r)}(\mathbf{x}_n)$ to parametrize the CMs. In contrast to Zhang et al. (2020), we only build one network to model every annotator. This network receives as an input the one-hot encoding ID of the annotator and, depending on the modeling, also the feature map of the last decoder layer of the segmentation network. The one-hot encoding ID is just a vector with all zeros except for a one in the position corresponding to the annotator whose CM we want to predict. By providing this annotator-dependent input to the network, a different CM is obtained for each annotator, removing the need for having one branch for each annotator as in Zhang et al. (2020). This paradigm is useful in real-world CPATH tasks because of the large number of annotators that may be involved.

Our annotator network $\mathbf{A}_\phi^{(r)}(\mathbf{x}_n)$ is composed of three branches, one for the annotator, another for the features, and the final one that combines both outputs. The annotator branch, which is common to every annotator, outputs an embedding for the annotator,

$$f_{\phi_{\text{ann}}}(\mathbf{e}_r) = \mathbf{h}_r, \tag{3}$$

where $\mathbf{e}_r$ is the one-hot encoding of the rth annotator ID and $\phi_{\text{ann}} \subset \phi$ are the weights for the branch corresponding to the annotators. This embedding $\mathbf{h}_r$ only depends on the rth annotator and contains information about this annotator's 'general' expertise. For example, if the annotator is prone to confuse two specific classes.

Assuming that an annotator may behave differently for every sample, we utilize an image feature branch that outputs an embedding for the patch,

$$g_{\phi_{\text{feat}}}(\mathbf{x}_n) = \mathbf{o}_n, \tag{4}$$

where $\phi_{\text{feat}} \subset \phi$ are the weights for the image feature branch. This embedding $\mathbf{o}_n$ only depends on the nth patch and contains information about the difficulty to annotate of this patch. Notice that some samples may be more ambiguous and more difficult to categorize, while others may be trivial to annotate.

Finally, both embeddings are stacked together, and the final branch composed of fully connected layers is used to output the final CM for this annotator and patch,

$$z_{\phi_{\text{final}}}(\mathbf{h}_r, \mathbf{o}_n) = \mathbf{A}_\phi^{(r)}(\mathbf{x}_n), \tag{5}$$

where $\phi_{\text{final}} \subset \phi$ are the weights for the final branch. This final branch learns the CM from the patch and annotator embeddings. It identifies how the annotator will behave on that patch.

Since annotator behavior modeling is challenging, we define a family composed of three different models that follow the previous architecture (see Fig. 2):

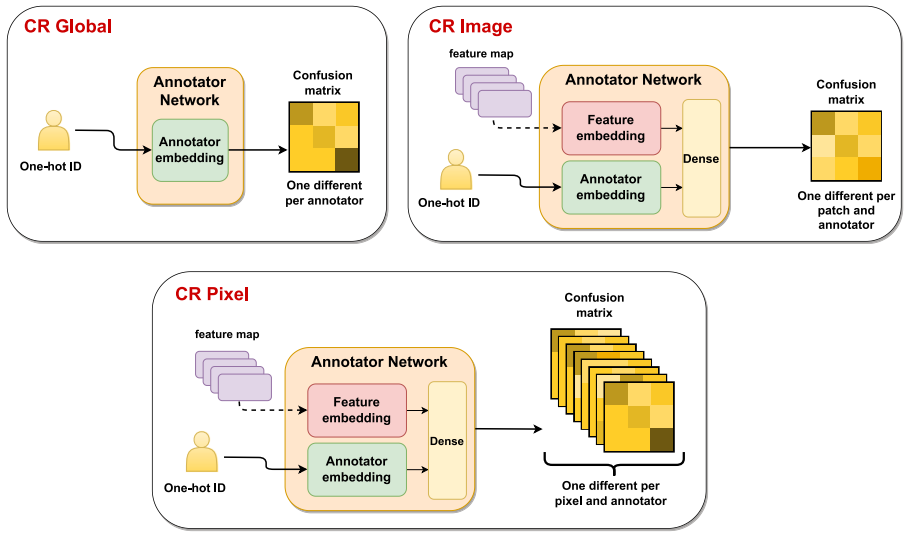

Fig. 2. The family of three proposed architectures for the annotator network. (i) CR Global: it does not consider the features to estimate the annotator expertise; (ii) CR Image: the annotator network has two modules. One is to compute an embedding for the annotator and another for the features of the patch from the segmentation network. Then, it combines both embeddings to estimate the CM for the patch; (iii) CR Pixel also has two modules, but in this case, it estimates a per-pixel CM.

CR Global. This model is the least complex. It assumes that CMs do not depend on the patch features but only on the annotators. In this model, we use only the annotator embedding to compute the CM of the annotator. The annotator will share the same CM for every patch,

$$\mathbf{A}_{\phi}^{(r)}(\mathbf{x}_n) \equiv \mathbf{A}_{\phi}^{(r)} := z_{\phi_{\text{final}}}(\mathbf{h}_r), \qquad \forall n \in N. \tag{6}$$

CR Image. This novel model in the literature has never been proposed for crowdsourcing segmentation. This model depends on the image features and assumes that every pixel within the patch shares the same CM (for each annotator). That is, the CM only depends on the global features of the patch and the annotator. Now, the CM may vary across different patches and annotators,

$$\mathbf{A}_{\phi}^{(r)}(\mathbf{x}_n)_{w,h} \equiv \mathbf{A}_{\phi}^{(r)}(\mathbf{x}_n) := z_{\phi_{\text{final}}}(\mathbf{h}_r, \mathbf{o}_n), \qquad 1 \le w \le W, 1 \le h \le H. \tag{7}$$

CR Pixel. This model is the most complex. It assumes that CMs depend on the annotator and the patch features as CR Image but estimates a CM per pixel. The CM may vary across different pixels and annotators:

$$\mathbf{A}_{\phi}^{(r)}(\mathbf{x}_n)_{w,h} := z_{\phi_{\text{final}}}(\mathbf{h}_r, \mathbf{o}_n)_{w,h}, \qquad 1 \le w \le W, 1 \le h \le H. \tag{8}$$

To predict the behavior of the rth annotator on the nth image, a forward pass is performed using the image and the one-hot encoding ID as the inputs of the segmentation and annotator networks, respectively. In the CR Image and CR Pixel models, the final map of the segmentation network will also be fed to the annotator network.

2.5. Loss function

This section presents how to learn the optimal parameters of both networks $\{\hat{\theta}, \hat{\phi}\}$. Every model of the proposed family is learned using the same procedure. We minimize the loss function, the sum of the negative log-likelihood (NLL) plus a regularizer. The optimization process is performed by using a method based on stochastic gradient descent.

First, the NLL is given by the sum of cross-entropy (CE) losses between the observed noisy segmentations and the estimated annotator label distributions,

$$-\log \mathrm{p}_{\theta,\phi}(\tilde{\mathbf{Y}}^{(1)}, \dots, \tilde{\mathbf{Y}}^{(R)}) = \sum_{n=1}^{N} \sum_{r=1}^{R} \mathbb{I}(\tilde{\mathbf{y}}_n^{(r)} \in R_n) \cdot \mathrm{CE}\left(\mathbf{A}_{\phi}^{(r)}(\mathbf{x}_n) \cdot \mathbf{p}_{\theta}(\mathbf{x}_n), \tilde{\mathbf{y}}_n^{(r)}\right), \tag{9}$$

where $\mathbb{I}$ is the indicator function. Remember that not all annotators may label the same patches. Minimizing the NLL encourages to make $\mathbf{p}_{\theta,\phi}^{(r)}(\mathbf{x}_n)$ as close as possible to the true distribution of the noisy annotators $\mathbf{p}^{(r)}(\mathbf{x}_n)$. However, this procedure may not lead to the expert (latent) segmentation since many combinations of the estimated CM $\mathbf{A}_{\phi}^{(r)}(\mathbf{x}_n)$ and the expert (latent) segmentation $\mathbf{p}_{\theta}(\mathbf{x}_n)$ can match the true annotators' distribution. To overcome this issue, some works (e.g., Zhang et al., 2020; Tanno et al., 2019) have added the trace of the estimated CMs to the loss function as a regularization term. Intuitively, the bigger the trace term the more reliable this annotator. Then, the combined loss is given by,

$$\mathcal{L}(\theta, \phi) := \sum_{n=1}^{N} \sum_{r=1}^{R} \mathbb{I}\left(\tilde{\mathbf{y}}_n^{(r)} \in R_n\right) \cdot \left[\mathrm{CE}\left(\mathbf{A}_{\phi}^{(r)}(\mathbf{x}_n) \cdot \mathbf{p}_{\theta}\left(\mathbf{x}_n\right), \tilde{\mathbf{y}}_n^{(r)}\right) + \lambda \cdot \mathrm{tr}\left(\mathbf{A}_{\phi}^{(r)}\left(\mathbf{x}_n\right)\right)\right], \tag{10}$$

where $\mathrm{tr}(\cdot)$ is the trace operator. We aim to find the optimal $\hat{\phi}$ and $\hat{\theta}$ as the solution to

$$\{\hat{\theta}, \hat{\phi}\} = \arg\min_{\{\theta,\phi\}} \mathcal{L}(\theta, \phi). \tag{11}$$

Then, the training process is the following, firstly, it is widely extended to set λ to a negative number for maximizing the CMs trace (i.e., the annotators' are reliable). Otherwise, there may be *identifiability* issues (Ruiz et al., 2023). Namely, the crowdsourcing method could consider the annotators unreliable and learn an incorrect concept. Secondly, after several steps during the training stage, some works have introduced setting λ to non-negative to encourage annotators to be as unreliable as possible (Tanno et al., 2019; Zhang et al., 2020). Intuitively, once the CE is optimized by relying on the annotators' labels, this loss finds the maximum confusion that adequately explains the noisy observations. We will study in the experimental section if minimizing the trace term is effective for our CR models.

3. Experimental setup

3.1. Data

We evaluate the proposed method for crowdsourcing segmentation of histopathological images with a public dataset of Triple Negative Breast Cancer images (Amgad et al., 2019). This dataset contains 151

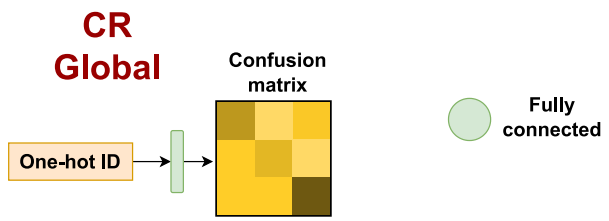

Fig. 3. Diagram of the architecture of the annotator network used in the CR Global model.

different Whole Slide Images stained with Hematoxylin and Eosin. Each WSI has a delineated Region of Interest (ROI) that contains representative tissue of predominant region classes and textures within each slide. In total, 161 ROIs are extracted and annotated by 20 medical students: ten of the ROIs are annotated by everyone, and the rest by only one participant. Furthermore, two senior pathologists provided curated labels that represent the ground-truth. We extract patches of 512×512 (training: 10,173; validation: 1264; test: 399) from the ROIs. We consider five classes: Other, tumor, stroma, inflammation, and necrosis. The dataset is imbalanced. Regions containing areas of tumor and stroma are over-represented. Necrosis and the Other class are rare. For this reason, class-wise metrics are essential to validate the methods.

3.2. Model hyperparameters

We set $\lambda = -1$ (maximize the trace regularizer) and a learning rate of 10^{-3} for five epochs to warm up the annotator network. Following the fifth epoch, we study the impact of the regularizer term (in the first experiment) and adjust the learning rate to 10^{-4}. We utilize a learning rate of 10^{-4} for the segmentation network during the whole training process. We utilize the Adam optimizer, a mini-batch size of 16, and run the methods three times with varying seeds. We train the models for 20 epochs.

3.3. Implementation details

Segmentation Network architecture. The encoder (i.e., ResNet34) downsamples the $512 \times 512 \times 3$ image 5 times to $16 \times 16 \times 512$. The number of channels scales up progressively, starting from 64, then to 64 again, 128, 256, and finally reaching 512. The decoder part is built symmetrically with upsampling modules and skip connections according to the U-Net architecture. We integrate this popular architecture in the literature through the utilization of the package of pre-trained U-Net backbones in Pytorch (Iakubovskii, 2019).

Annotator Network architecture. The practical implementation of the annotator network is as follows:

- CR Global. This method makes only use of the annotator branch. This branch applies a fully connected layer to the one-hot encoding ID associated to the annotator. This procedure makes the confusion matrices only annotator-aware. Fig. 3 depicts graphically this architecture.
- CR Image. This method uses the annotator, image, and final branches. The annotator and image branches operate in parallel. The annotator branch applies a fully connected layer to the one-hot encoding ID. The image branch applies a convolutional neural network to the last feature map of the segmentation network. Subsequently, embeddings from the annotator and image branches are stacked together. The confusion matrix is obtained by applying fully connected layers to this new embedding. This procedure makes the confusion matrices image- and annotator-aware. Fig. 4 depicts graphically this architecture.
- CR Pixel. This method comprises the annotator, image, and final branches. The annotator branch is defined as in the previous model. The image branch applies a convolutional neural network with no pooling layers. Consequently, the image branch yields an output shape of $W \times H \times C'$, being C' the dimension of the pixel-wise features. As before, the outputs of the image and annotator branches are stacked together, and the final branch applies fully connected layers to predict the confusion matrix for each pixel. This procedure makes the confusion matrices pixel- and annotator-aware. This method obtains the most fine-grained confusion matrices of the CR family. Fig. 5 depicts graphically this architecture.

Software. All our methods are implemented in Pytorch 1.10.1 and for the segmentation networks we use a specific library for loading architectures and pretrained models (Iakubovskii, 2019). All the models are run using an NVIDIA GeForce RTX 3090 GPU. The code will be available on GitHub upon acceptance.

3.4. Baseline methods

We compare the proposed CR methods against three well-known methods: (i) the Supervised Expert method, which is trained with expert labels provided by senior pathologists; (ii) Majority Voting (MV); and (iii) STAPLE. The latter two are well-known in the two-stage crowdsourcing literature and widely applied to histopathological images. They aggregate the noisy crowdsourcing masks in a previous step and then train the segmentation methods with these aggregated labels. The MV computes the mode of the annotations while the STAPLE performs a weighted average. STAPLE computes blindly the expertise of each annotator for every patch. For these three models, the hyperparameters and training process are the same as in the segmentation network of the proposed CR methods.

3.5. Evaluation metrics

Since the DICE coefficient is very extended in the semantic segmentation literature, we use it to evaluate the methods. The DICE measures the (pixel-wise) agreement between the predicted segmentation mask and the ground-truth. We assume the masks provided by senior pathologists to be the ground-truth in this problem. The formula of the DICE is given by

$$\mathrm{DICE}(\mathbf{X}, \mathbf{Y}) := \frac{2|\mathbf{X} \cap \mathbf{Y}|}{|\mathbf{X}| + |\mathbf{Y}|} \tag{12}$$

In our experiments, we calculate the DICE coefficient for each class. We also report two global metrics: the micro- and macro-average DICE. The micro-average DICE is computed pixel-wise across all the classes. Notice that missing a minority class will not hamper this metric. For the macro-average metric, the average of the DICE coefficients per class is computed. Each class contributes the same to the metric regardless of its representativity on the test set. The metrics are averaged through the three runs and the standard error $S = \sigma/\sqrt{3}$ is also reported.

4. Results

This section presents the empirical results of our family of CR methods. First, Section 4.1 analyzes the effect of the regularizer on the loss function. Then, we evaluate the studied methods with two complementary approaches: (i) quantitative (Section 4.2) provides the DICE coefficients for the studied methods; (ii) qualitative (Section 4.3) conducts a human evaluation by asking three expert pathologists to rate the predicted masks. Additionally, we assess the CR methods depending on the subjectivity across the different classes (Section 4.4) and through different sizes of the training set (Section 4.5).

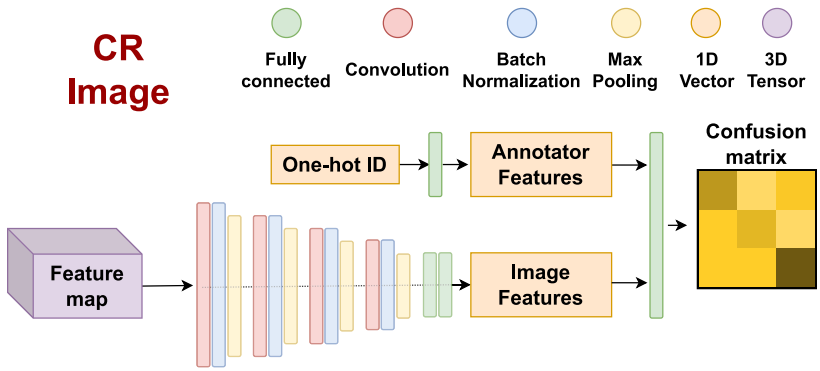

Fig. 4. Diagram of the architecture of the annotator network used in the CR Image model. The feature map comes from the last convolutional layer of the segmentation network.

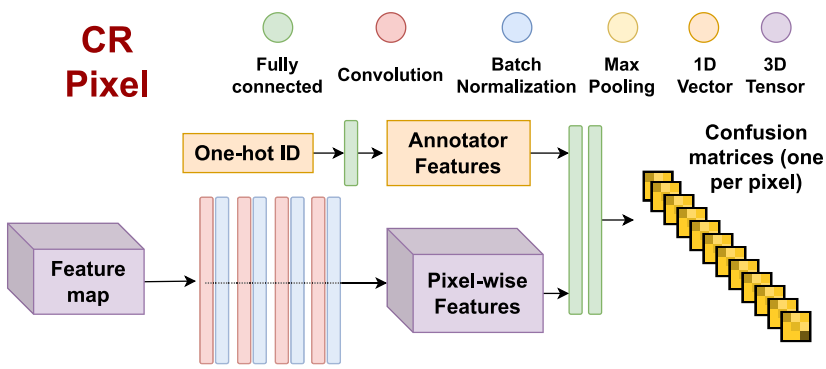

Fig. 5. Diagram of the architecture of the annotator network used in the CR Pixel model. The feature map comes from the last convolutional layer of the segmentation network.

4.1. Analyzing the weight λ for the regularizer

This section studies the impact of the regularizer on the loss function of our family of CR methods. As we discussed earlier, some works have used this regularizer to estimate the CM of the annotators. The idea behind this is that a higher weight for the regularizer will force the annotators to be more unreliable. We warm up every method with $\lambda = -1$ and then choose between several values for $\lambda = \{0, 0.01, 0.1, 0.5\}$. Fig. 6 shows the micro- and macro-average DICE. A high value of λ hampers the performance of more complex methods, CR Image and CR pixel. This fact makes sense since these methods have more flexibility to estimate the annotators' CM and are more influenced by the trace term. See that CR Global is robust to different values of λ with low variability ranges. However, the three CR methods achieve the best results when λ is low. These results show that the trace term does not improve our CR methods. By making the annotators more unreliable, we find identifiability issues, learning the wrong concept because of the flexibility of the more complex models, CR Image and CR Pixel.

Table 1 depicts the per-class DICE values, showing that less represented classes, such as necrosis or inflammation, are more challenging to detect. High λ values also hamper the performance on these classes. The three CR models perform well with low λ values. The best-performing model is CR Image, which generally achieves the best per-class results and outperforms the rest at necrosis, the most challenging class to detect.

In conclusion, in this first experiment, we find that our family of CR methods is competitive, especially our novel CR Image, which performs remarkably well. Furthermore, the trace regularizer term does not benefit our CR methods. For the rest of the experimental section, we remove this term, i.e., fixing $\lambda = 0$ after the warm-up period.

4.2. Comparison with other methods

Here, we assess the performance of the proposed CR family against other well-known methods. The objective of this comparison is twofold. First, we aim to determine how well our CR family performs compared to the supervised expert method trained with expert segmentation masks. These masks are assumed to be the ground-truth in histopathological images. Second, we would like to discern if our methods perform better than two-stage well-known CR methods (i.e., MV and STAPLE). In contrast to our framework, these methods combine the labels in a previous stage.

Table 2 and Fig. A.13 depict the DICE coefficients for this comparison. In this case, the supervised expert method achieves the best results in the micro-average (DICE: 0.7953 ± 0.0057), while the CR image stands out in terms of the macro-average metric (DICE: 0.7827 ± 0.0046). CR Image is also robust to different initializations, presenting minor variability. The rest of our CR models (CR Global and CR Pixel) outperform the crowdsourcing competitors, MV and STAPLE.

Table 3 and Fig. A.14 present the results per class for further analysis. Our family of CR methods performs well across all the classes and outperforms MV and STAPLE. CR Image is the best in more challenging classes, i.e., necrosis (DICE: 0.7364 ± 0.0162) and inflammation (DICE: 0.7556 ± 0.0180), and even superior to the supervised expert method (DICE in necrosis: 0.6819 ± 0.1216; DICE in inflammation: 0.7434 ± 0.0237).

We provide further insights about the results in Fig. 7. This figure shows the predicted segmentations obtained for different patches. The supervised expert method performs consistently well in the four examples (as it is expected). Then, our CR family exhibits a competitive behavior, specifically the CR Image. In the first row, our CR Image clearly detects the Other (cyan) class, in contrast to MV and STAPLE. In

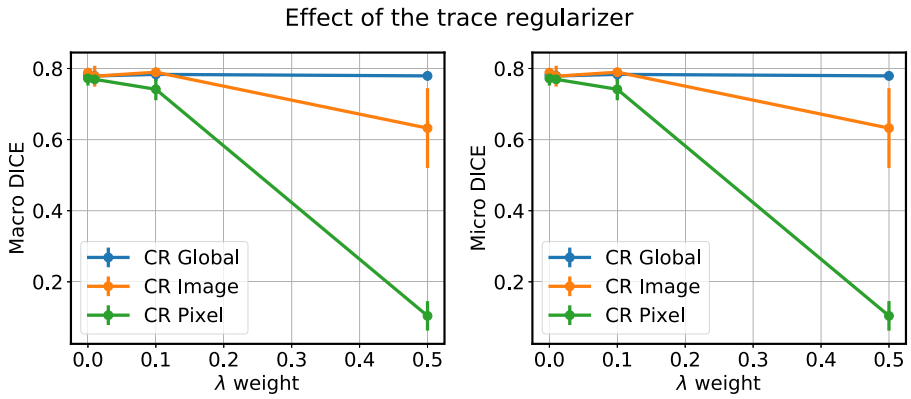

Fig. 6. DICE coefficients in the test set for the CR methods. The hyperparameter λ is varied in {0,0.01,0.1,0.5}. The results are averaged, and the bars represent the standard error, through three different runs. As λ grows, the performance of the CR methods does not improve.

Table 1
DICE coefficients per class in the test set for the CR methods. The hyperparameter λ is varied in {0,0.01,0.1,0.5}. The results are averaged, and the intervals represent the standard error, through three different runs. As the weight of the regularizer grows, the performance of the CR methods does not improve. CR Image performs better than the rest across almost all the classes. CR Image performs remarkably well at necrosis, the most challenging class to detect.

	λ	Other	Tumor	Stroma	Inflammation	Necrosis
CR global	0	0.8327 ± 0.0081	0.8364 ± 0.0030	0.7434 ± 0.0074	0.7146 ± 0.0405	0.6621 ± 0.0724
	0.01	0.8294 ± 0.0094	0.8311 ± 0.0077	0.7404 ± 0.0057	0.7228 ± 0.0235	0.662 ± 0.0616
	0.1	0.834 ± 0.0096	**0.8367 ± 0.0041**	0.7468 ± 0.0121	0.7061 ± 0.055	0.6955 ± 0.0361
	0.5	0.8069 ± 0.0069	0.8322 ± 0.0062	0.7461 ± 0.0082	0.7418 ± 0.0057	0.615 ± 0.0629
CR image	0	0.8423 ± 0.0048	0.8292 ± 0.0075	0.7501 ± 0.0062	0.7556 ± 0.0180	**0.7364 ± 0.0162**
	0.01	**0.8492 ± 0.0021**	0.8252 ± 0.0140	0.7478 ± 0.0059	0.7169 ± 0.0534	0.6113 ± 0.1020
	0.1	0.8343 ± 0.0150	0.8299 ± 0.0073	**0.7505 ± 0.0075**	**0.774 ± 0.0074**	0.7167 ± 0.0445
	0.5	0.6315 ± 0.1922	0.7738 ± 0.0202	0.6130 ± 0.0541	0.3021 ± 0.2096	0.2374 ± 0.2374
CR pixel	0	0.8145 ± 0.0198	0.8323 ± 0.0071	0.7290 ± 0.0235	0.7285 ± 0.0042	0.6176 ± 0.0987
	0.01	0.8298 ± 0.0049	0.8146 ± 0.0066	0.7300 ± 0.005	0.7336 ± 0.0024	0.6777 ± 0.0432
	0.1	0.7475 ± 0.0573	0.8253 ± 0.0039	0.6913 ± 0.0228	0.7184 ± 0.0136	0.3977 ± 0.1269
	0.5	0.0255 ± 0.0116	0.1519 ± 0.0657	0.1035 ± 0.0186	0.0212 ± 0.0062	0.03 ± 0.0180

the second row, STAPLE fails to identify "Inflammation" (green) when it is actually "Necrosis" (purple). Our CR methods detect the "Necrosis" and CR Image and CR pixel can even distinguish the small "Other" class in between "Necrosis" and "Tumor" (red). In the third row, CR Image and CR Pixel detect the "Inflammation" and MV does not. Finally, in the fourth row, most methods underestimate the tumoral area while CR Image provides an accurate prediction.

To summarize this section, our family of CR methods outperforms widely known methods in the literature, MV and STAPLE. Our methods benefit from one-stage learning in line with previous studies in the literature for classification (Karimi et al., 2020; López-Pérez et al., 2023; Tanno et al., 2019). Specifically, our CR family is an effective framework for dealing with multiple annotators in segmentation tasks in CPATH. Furthermore, our new modeling of the annotator network (i.e., CR Image) stands out in the problem of TNBC tissue segmentation. This novel modeling is a model between the previously formulated CR Pixel and CR Global in terms of complexity. These results show that CR Image satisfies a trade-off between complexity and performance. These results also highlight the potential of our crowdsourcing methods, with noisy masks being a feasible alternative to single expert labeling.

4.3. Expert evaluation

Semantic segmentation is a subjective task in histopathology, with high inter-observer variability. Even expert pathologists may disagree on the segmentations. To provide further insights, we conduct an additional human evaluation of the methods with expert pathologists. This evaluation is qualitative and differs from the previous one since it considers the global features of the patches instead of exhaustively the value of every pixel.

Table 2
Dice coefficients micro- and macro-average in the test set. The results represent the average through three different runs and the intervals are the standard error.

	Micro	Macro
MV	0.7687 ± 0.0056	0.7355 ± 0.0103
STAPLE	0.7549 ± 0.0053	0.7039 ± 0.0120
Supervised	**0.7953 ± 0.0057**	0.7723 ± 0.0120
CR Global	0.7812 ± 0.0067	0.7578 ± 0.0120
CR Image	0.7884 ± 0.0073	**0.7827 ± 0.0046**
CR Pixel	0.7721 ± 0.0116	0.7444 ± 0.0214

In this experiment, we randomly selected nine patches and presented them, along with predictions generated by six different methods (MV, STAPLE, Expert, CR Global, CR Image, and CR Pixel), to three expert pathologists. These patches mainly contained tumor patterns, which is the majority class in our setting. To ensure unbiased evaluation, the predictions were anonymized and shuffled. The experts were given two tasks: (i) determine whether each prediction was acceptable or not; (ii) rank the acceptable masks in order of quality, from the best to the worst. By following this approach, we aimed to assess the performance of the six methods and gather insights from the expert evaluations.

We measure the agreement level using Krippendorff's Alpha (Krippendorff, 2018). The value of this estimator ranges between −1 and 1, with −1 meaning total disagreement, 0 meaning randomness and 1 meaning total agreement. The three pathologists have a Krippendorff's Alpha value of 0.2729 across all the images for acceptable or not, which

Table 3
Dice coefficients per class in the test set. The results represent the average through three different runs and the intervals are the standard error.

	Other	Tumor	Stroma	Inflammation	Necrosis
MV	0.8113 ± 0.0109	0.8252 ± 0.0045	0.7462 ± 0.0127	0.6650 ± 0.0079	0.6297 ± 0.0460
STAPLE	0.8257 ± 0.0038	0.8089 ± 0.0050	0.7512 ± 0.0007	0.6224 ± 0.0356	0.5111 ± 0.0371
Supervised	0.8309 ± 0.0064	0.8322 ± 0.0038	**0.7731 ± 0.0035**	0.7434 ± 0.0237	0.6819 ± 0.01216
CR Global	0.8327 ± 0.0081	**0.8364 ± 0.0030**	0.7434 ± 0.0074	0.7146 ± 0.0405	0.6621 ± 0.0724
CR Image	**0.8423 ± 0.0048**	0.8292 ± 0.0075	0.7501 ± 0.0062	**0.7556 ± 0.0180**	**0.7364 ± 0.0162**
CR Pixel	0.8145 ± 0.0198	0.8323 ± 0.0071	0.7290 ± 0.0235	0.7285 ± 0.0042	0.6176 ± 0.0987

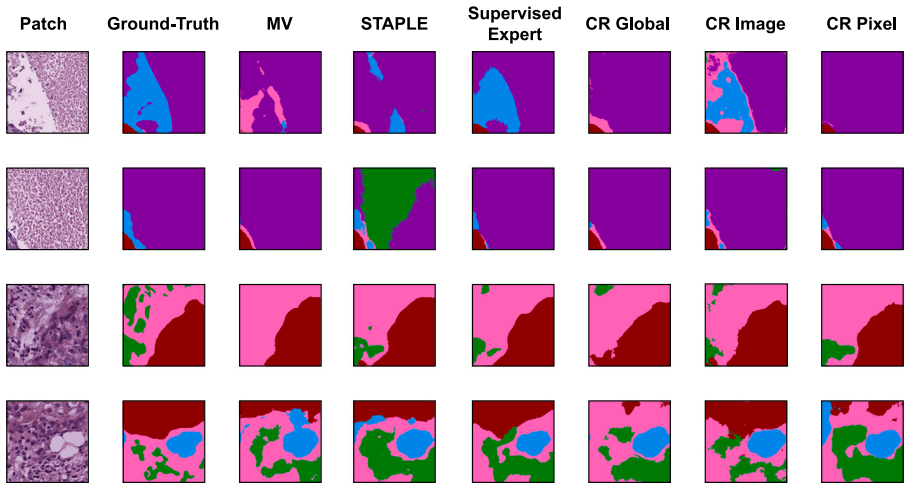

Fig. 7. Segmentation masks of the six studied methods in four different patches for visualization. The color legend is red: tumor, pink: stroma; green: inflammation; purple: necrosis; cyan: other. The ground-truth mask is the curated mask provided by Senior Pathologists. The supervised expert method is the only one which was trained with expert labels. The rest only had access to noisy masks provided by medical students. We see that the family of CR methods (Ours) performs fairly well and is competitive with the supervised expert method.

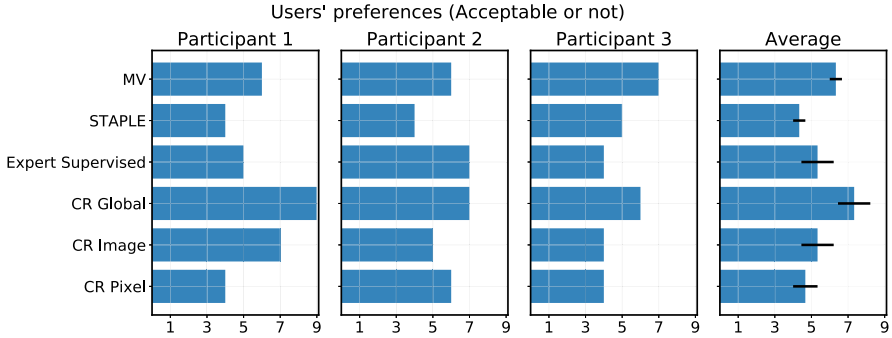

Fig. 8. Number of times the methods have been considered acceptable for each pathologist. The right figure represents the average and standard error of the three.

shows a low agreement. This result highlights the notable subjectivity and difficulty of this task.

We first analyze if the pathologists found the methods acceptable or not. Fig. 8 shows how often they said a method is acceptable. The last column summarizes this information with the average and the standard error across the three pathologists. We can see clearly that generally, they find CR Global satisfying most of the time. Also, CR Image is competitive, which agrees with the good results obtained in previous sections. Notice that CR Pixel does not achieve good performance on this evaluation. This may be due to the higher error through the three runs.

Since this task is very subjective, Fig. 9 shows the number of times that a method was found acceptable by the three of them, unanimously. CR Global outperforms the rest, while CR Image and MV are competitive. This result confirms the competitiveness of our CR family, and specifically of CR Global in this expert evaluation.

Finally, Fig. 10 shows the number of times each method was ranked first for each pathologist. On average, CR Global was systematically ranked first, while the rest were not too appealing to the pathologists. Indeed, the two first pathologists prefer CR Global most of the time.

From this experiment, we conclude that our CR family is competitive and obtains remarkable results according to three expert pathologists. Especially, CR Global performs remarkably well, which aligns with the previous experiment since this method was the best in tumor performance. Furthermore, CR Image is very competitive. We cannot strongly conclude which method is better because they were evaluated on only nine patches. These results complement those of the previous section and confirm that our CR Global and CR Image provide satisfying segmentation masks according to the visual analysis of three different pathologists. However, CR Pixel does not produce worthy segmentation masks for expert pathologists. The main problem of CR Pixel is the complex annotator network which outputs a Confusion Matrix per

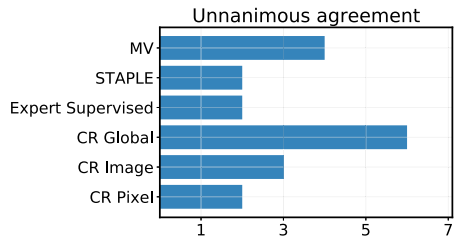

Fig. 9. Number of times that the methods have been considered acceptable by the three expert pathologists unanimously.

pixel. The number of classes, patches, and the difficulty/subjectivity inherent to this problem can harm this CR method.

4.4. Analysis of the subjectivity across different classes

This subsection further analyzes the performance of the CR methods depending on the subjectivity across the different classes. As indicated in the previous study by Amgad et al. (2019), the discordance between observers varies significantly by class. They computed the discordance between annotators across the classes utilizing the following formula

$$\nabla_{i,j,l} = 1 - 2 \times \frac{|I_l \cap J_l|}{|I_l| + |J_l|}, \tag{13}$$

where $\nabla_{i,j,l}$ is the discordance between the annotators i and j for the class l, with corresponding binary masks I_l and J_l for that class. This coefficient ranges in $[0, 1]$, with 1 indicating total discordance and 0 total agreement. They obtained that the median discordance for tumor, stroma, inflammation, and necrosis between non-experts and experts were 0.14, 0.27, 0.54, and 1.0, respectively.

Fig. 11 shows the performance of the CR methods against these values of discordance. While in the tumor class, the three CR methods are almost coincident, when referring to more subjective classes, i.e., Inflammation or Necrosis, CR Image outperforms the other two methods. As the discordance grows, the DICE of CR Global and CR Pixel drops and the gap between them and CR Image becomes wider.

We have shown through this analysis that CR Global performs fairly well across classes with a high agreement between annotators, especially in the tumor class, which is the dominant class (see DICE in tumor and the micro-average DICE in Tables 2 and 3). However, CR Image shows a remarkable performance across the different classes, outperforming CR Global in more subjective classes (see macro-average DICE and DICE in inflammation and necrosis in Tables 2 and 3). We hypothesize that the more flexible annotator network of the CR Image is important to learn the noisy annotators' behavior in these classes.

4.5. Robustness to the size of the training set

We assessed our crowdsourcing segmentation family's generalization capability and robustness against the lack of labeled data, which is a typical scenario in medical imaging. We used a random sample of the training data with 10%, 25%, 50%, and 75% of the total training data. We ran this experiment three times with different subsets.

Fig. 12 depicts the macro-average DICE coefficients for this experiment. We used the macro-average to underline the performance through the different classes. As expected, the supervised method trained with expert masks performs best when the training set is small. Conversely, the baseline methods that rely on mask aggregation (i.e., STAPLE and MV) perform worse across various training set sizes. Our CR family performs remarkably well under these varying training size conditions.

Upon closer examination of this experiment, we see that the more complex methods of our CR family (i.e., CR Image and CR Pixel) do not perform well when trained with a small number of samples. The limited number of labeled samples per annotator hinders obtaining reliable parameters in more complex models. In contrast, CR Global, which is the simplest method of our family, achieves satisfying performance in this case. CR Global also achieves its best performance with the 50% of the training set. By adding more data, the performance does not keep improving. Interestingly, CR Image's performance improves notably when trained with 75% of the dataset. The flexible annotator network of CR Image takes advantage of having access to more labeled data. Indeed, when trained with the entire dataset, CR Image stands out in terms of the macro-average DICE metric.

In conclusion, the simple annotator network makes CR Global the best option for scenarios involving a small training dataset because it can estimate a reliable model with a small number of labeled samples. The more flexible CR Image is the best when training with the entire dataset, achieving the best macro-average DICE and better modeling of more subjective classes.

5. Conclusion

This paper proposes a novel family of one-stage crowdsourcing segmentation methods for histopathological tissue. The proposed CR methods jointly estimate the annotators' expertise and the segmentation method, scaling on the number of annotators, and can be applied to real-world CPATH segmentation. We validate the methods in a real-world dataset composed of Triple Negative Breast Cancer histopathological images. Through extensive and exhaustive experimentation and evaluation, we show the effectiveness and potential of our CR methods from several perspectives. The results are remarkable in quantitative (reporting DICE metrics) and qualitative analysis (conducted with three expert pathologists), outperforming state-of-the-art STAPLE and showing comparable results against the supervised method trained with expert masks. Within our CR family, the new CR Image stands out for its remarkable performance in challenging and subjective classes, owing to the flexibility of its annotator network. Furthermore, CR Global takes advantage of its simplicity and achieves excellent results with small training sets.

In this work, we have assumed independence among the annotators, which is unreal. Future work will address new architectures for the annotator network considering correlations in their behavior. We also plan to consider how to introduce prior information on the annotators, such as years of experience or confidence in the annotation. Furthermore, the proposed methods can potentially be leveraged in other real-world medical imaging settings with multiple noisy annotators.

CRediT authorship contribution statement

Miguel López-Pérez: Software, Validation, Investigation, Formal analysis, Visualization, Writing – original draft. **Pablo Morales-Álvarez:** Software, Methodology, Writing – original draft. **Lee A.D. Cooper:** Conceptualization, Data curation, Visualization, Supervision, Writing – review & editing. **Christopher Felicelli:** Validation, Writing – review & editing. **Jeffery Goldstein:** Validation, Writing – review & editing. **Brian Vadasz:** Validation, Writing – review & editing. **Rafael Molina:** Methodology, Resources, Writing – review & editing, Supervision, Funding acquisition, Project administration. **Aggelos K. Katsaggelos:** Conceptualization, Supervision, Writing – review & editing, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

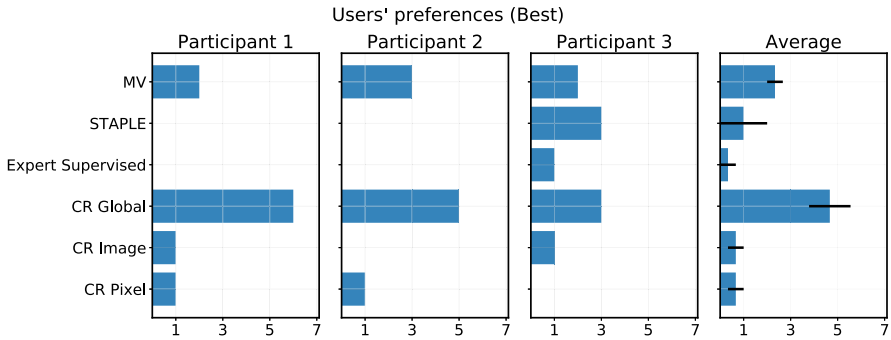

Fig. 10. Number of times the methods have been ranked as the best one for each pathologist. The right figure represents the average and standard error of the three.

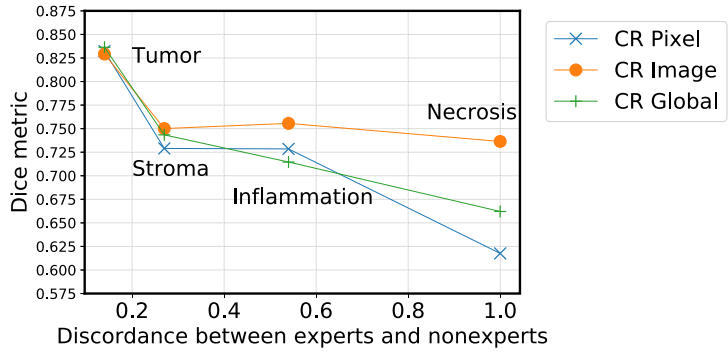

Fig. 11. Performance of the CR methods depending on the subjectivity of the classes. CR methods show a similar behavior in classes with less subjectivity. As the subjectivity grows, CR Image outperforms the rest.

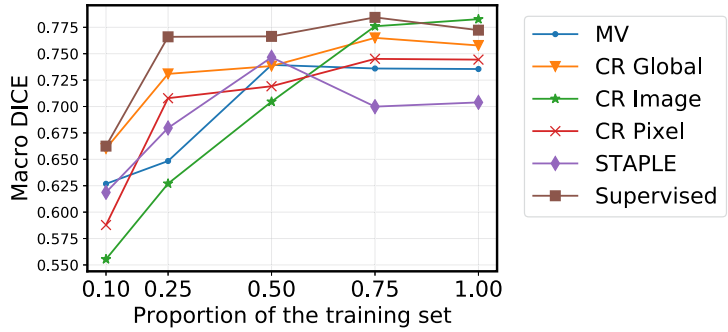

Fig. 12. Macro-average DICE coefficient in the test set of methods trained with a randomly selected subset of the training data. The results are averaged through three different runs.

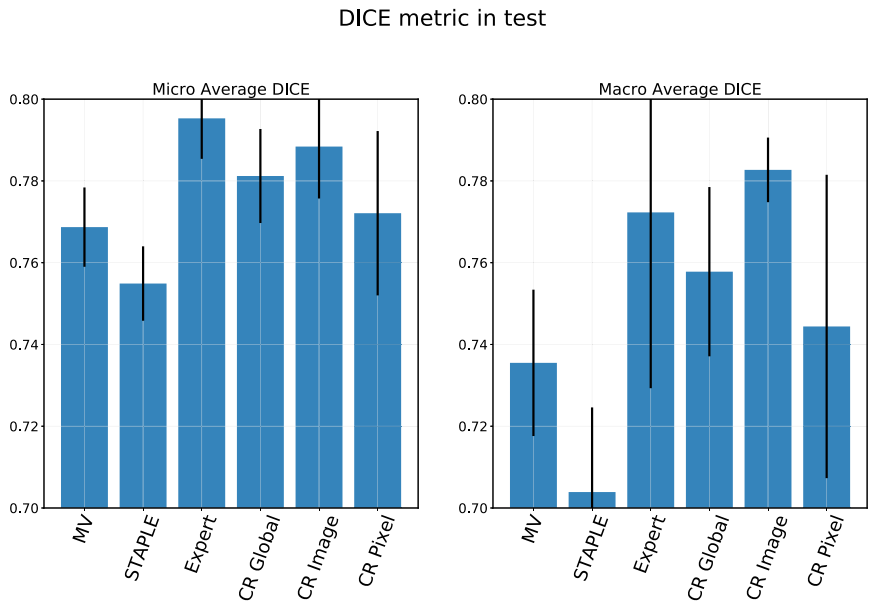

Fig. A.13. Left, micro-average DICE coefficients in the test set. Right, macro-average DICE coefficients in the test set. The bars represent the average through three different runs and the intervals are the standard error.

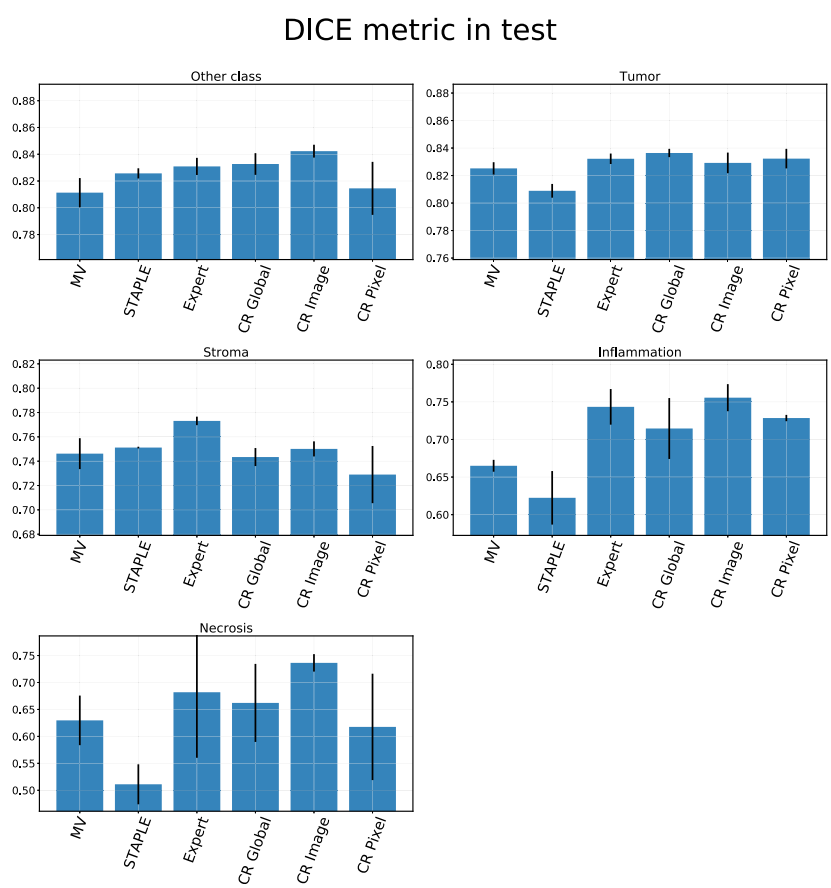

Fig. A.14. Per-class DICE coefficients in the test set. The results are reported for the six studied methods. The bars represent the average through three different runs and the intervals are the standard error.

Appendix. Additional figures

This appendix includes some additional figures to better visualize the results obtained in our work. Figs. A.13 and A.14 depict the DICE values and standard error through the three runs for the studied methods. These figures are commented on in the main text but we included them here for better readability.

References

Aatresh, A.A., Yatgiri, R.P., Chanchal, A.K., Kumar, A., Ravi, A., Das, D., Raghavendra, B., Lal, S., Kini, J., 2021. Efficient deep learning architecture with dimension-wise pyramid pooling for nuclei segmentation of histopathology images. Comput. Med. Imaging Graph. 93, 101975.

Albarqouni, S., Baur, C., Achilles, F., Belagiannis, V., Demirci, S., Navab, N., 2016. AggNet: Deep learning from crowds for mitosis detection in breast cancer histology images. IEEE Trans. Med. Imaging 35 (5), 1313–1321. http://dx.doi.org/10.1109/TMI.2016.2528120.

Amgad, M., Atteya, L.A., Hussein, H., Mohammed, K.H., Hafiz, E., Elsebaie, M.A., Alhusseiny, A.M., AlMoslemany, M.A., Elmatboly, A.M., Pappalardo, P.A., et al., 2022. NuCLS: A scalable crowdsourcing approach and dataset for nucleus classification and segmentation in breast cancer. GigaScience 11.

Amgad, M., Elfandy, H., Hussein, H., Atteya, L.A., Elsebaie, M.A., Abo Elnasr, L.S., Sakr, R.A., Salem, H.S., Ismail, A.F., Saad, A.M., et al., 2019. Structured crowdsourcing enables convolutional segmentation of histology images. Bioinformatics 35 (18), 3461–3467.

Ben Hamida, A., Devanne, M., Weber, J., Truntzer, C., Derangère, V., Ghiringhelli, F., Forestier, G., Wemmert, C., 2022. Weakly supervised learning using attention gates for colon cancer histopathological image segmentation. Artif. Intell. Med. (ISSN: 0933-3657) 133, 102407.

Commowick, O., Istace, A., Kain, M., et al., 2018. Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure. Sci. Rep. 8 (1), 13650, (2018).

Foucart, A., Debeir, O., Decaestecker, C., 2022. Shortcomings and areas for improvement in digital pathology image segmentation challenges. Comput. Med. Imaging Graph. 102155.

Grote, A., Schaadt, N.S., Forestier, G., Wemmert, C., Feuerhake, F., 2018. Crowdsourcing of histological image labeling and object delineation by medical students. IEEE Trans. Med. Imaging 38 (5), 1284–1294.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

Iakubovskii, P., 2019. Segmentation models pytorch. https://github.com/qubvel/segmentation_models.pytorch.

Jiang, L., Zhang, H., Tao, F., Li, C., 2022. Learning from crowds with multiple noisy label distribution propagation. IEEE Trans. Neural Netw. Learn. Syst. 33 (11), 6558–6568.

Karimi, D., Dou, H., Warfield, S.K., Gholipour, A., 2020. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. Med. Image Anal. 65, 101759.

Khaliliboroujeni, S., He, X., Jia, W., Amirgholipour, S., 2022. End-to-end metastasis detection of breast cancer from histopathology whole slide images. Comput. Med. Imaging Graph. 102, 102136.

Krippendorff, K., 2018. Content Analysis: An Introduction to its Methodology. Sage publications.

Van der Laak, J., Litjens, G., Ciompi, F., 2021. Deep learning in histopathology: the path to the clinic. Nat. Med. 27 (5), 775–784.

Le, H., Gupta, R., Hou, L., Abousamra, S., Fassler, D., Torre-Healy, L., Moffitt, R.A., Kurc, T., Samaras, D., Batiste, R., et al., 2020. Utilizing automated breast cancer detection to identify spatial distributions of tumor-infiltrating lymphocytes in invasive breast cancer. Am. J. Pathol. 190 (7), 1491–1504.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Med. Image Anal. 42, 60–88.

López-Pérez, M., Amgad, M., Morales-Álvarez, P., Ruiz, P., Cooper, L.A., Molina, R., Katsaggelos, A.K., 2021. Learning from crowds in digital pathology using scalable variational Gaussian processes. Sci. Rep. 11 (1), 1–9.

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, 6922–6934.

Nir, G., Hor, S., Karimi, D., Fazli, L., Skinnider, B.F., Tavassoli, P., Turbin, D., Villamil, C.F., Wang, G., Wilson, R.S., et al., 2018. Automatic grading of prostate cancer in digitized histopathology images: Learning from multiple experts. Med. Image Anal. 50, 167–180.

Rodrigues, F., Pereira, F., 2018. Deep learning from crowds. In: Proceedings of the AAAI Conference on Artificial Intelligence. 32.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer, pp. 234–241.

Ruiz, P., Morales-Álvarez, P., Coughlin, S., Molina, R., Katsaggelos, A.K., 2023. Probabilistic fusion of crowds and experts for the search of gravitational waves. Knowl.-Based Syst. 261, 110183.

Salgado, R., Denkert, C., Demaria, S., Sirtaine, N., Klauschen, F., Pruneri, G., Wienert, S., Van den Eynden, G., Baehner, F.L., Pénault-Llorca, F., et al., 2015. The evaluation of tumor-infiltrating lymphocytes (TILs) in breast cancer: recommendations by an international TILs Working Group 2014. Ann. Oncol. 26 (2), 259–271.

Tanno, R., Saeedi, A., Sankaranarayanan, S., Alexander, D.C., Silberman, N., 2019. Learning from noisy labels by regularized estimation of annotator confusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11244–11253.

Warfield, S.K., Zou, K.H., Wells, W.M., 2004. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Trans. Med. Imaging 23 (7), 903–921.

Yang, W., Li, C., Jiang, L., 2022. Learning from crowds with decision trees. Knowl. Inf. Syst. 64 (8), 2123–2140.

Zhang, L., Tanno, R., Xu, M.-C., Jin, C., Jacob, J., Cicarrelli, O., Barkhof, F., Alexander, D., 2020. Disentangling human error from ground truth in segmentation of medical images. Adv. Neural Inf. Process. Syst. 33, 15750–15762.