AggNet: Deep Learning From Crowds for Mitosis Detection in Breast Cancer Histology Images

6 authors, including:

Shadi Albarqouni
Helmholtz Zentrum München
121 PUBLICATIONS   **2,574** CITATIONS

SEE PROFILE

Felix Achilles
Technische Universität München
15 PUBLICATIONS   **567** CITATIONS

SEE PROFILE

Vasileios Belagiannis
Ulm University
67 PUBLICATIONS   **2,556** CITATIONS

SEE PROFILE

Stefanie Demirci
Technische Universität München
45 PUBLICATIONS   **616** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Project X-ray Image Quality Improvement View project

Project Single-View X-ray Interventional Imaging View project

AggNet: Deep Learning From Crowds for Mitosis Detection in Breast Cancer Histology Images

Shadi Albarqouni*, *Student Member, IEEE*, Christoph Baur, Felix Achilles, *Student Member, IEEE*,
Vasileios Belagiannis, *Student Member, IEEE*, Stefanie Demirci, and Nassir Navab, *Member, IEEE*

Abstract—The lack of publicly available ground-truth data has been identified as the major challenge for transferring recent developments in deep learning to the biomedical imaging domain. Though crowdsourcing has enabled annotation of large scale databases for real world images, its application for biomedical purposes requires a deeper understanding and hence, more precise definition of the actual annotation task. The fact that expert tasks are being outsourced to non-expert users may lead to noisy annotations introducing disagreement between users. Despite being a valuable resource for learning annotation models from crowdsourcing, conventional machine-learning methods may have difficulties dealing with noisy annotations during training. In this manuscript, we present a new concept for learning from crowds that handle data aggregation directly as part of the learning process of the convolutional neural network (CNN) via additional crowdsourcing layer (*AggNet*). Besides, we present an experimental study on learning from crowds designed to answer the following questions. 1) Can deep CNN be trained with data collected from crowdsourcing? 2) How to adapt the CNN to train on multiple types of annotation datasets (ground truth and crowd-based)? 3) How does the choice of annotation and aggregation affect the accuracy? Our experimental setup involved Annot8, a self-implemented web-platform based on Crowdflower API realizing image annotation tasks for a publicly available biomedical image database. Our results give valuable insights into the functionality of deep CNN learning from crowd annotations and prove the necessity of data aggregation integration.

Index Terms—Aggregation, crowdsourcing, deep learning, gamification, online learning.

I. INTRODUCTION

CROWDSOURCING is a type of participative online activity in which an individual, an institution, a non-profit organization or a company proposes to a group of individuals of varying knowledge, heterogeneity, and number, via flexible open call, the voluntary undertaking of a task [1]. It was first introduced by Jeff Howe and Mark Robinson in 2005 using the internet for outsourcing work to a crowd of people [2]. Being initially considered as market research strategy [3], it is nowadays widely seen as an economical way to recruit crowds for tedious and time-consuming tasks such as annotations for character recognition [4], image classification [5], and natural language processing [6]. As a result of this trend, many crowdsourcing plattforms such as Amazon Mechanical Turk (AMT)[1], Games with a Purpose[2] [7], Crowdflower[3], and LabelMe[4] have emerged within the past decade. Here, users are not only confronted with simple, every-day tasks, but are also engaged in highly complex processes involving innovation creation.

A good example for this, is the medical domain where, very recently, crowdsourcing has been presented as a solution to the immense lack in publicly available ground-truth data. Various applications such as medical pictogram [8], correspondence finding for stereo endoscopic imaging [9], device detection in angiographic sequences [10], telepathology [11], and medical image segmentation [12] and classification [13] have already shown that crowdsourcing can provide efficient and inexpensive data annotation. With the very recent launch of the CrowdTruth framework[5], IBM, Google and Amsterdam University have paved the way towards machine-human computing for collecting ground-truth annotation data on text, images and videos in the medical domain. Similarly, Celi *et al.* [14] organised several events and data marathons, where engineers, data scientists, and clinicians were invited to address specific challenges during the clinical routines and procedures. As a result, many innovative ideas and prototypes have been developed, clinicians as well as medical students become part of a data-driven learning system. The most astonishing fact about crowdsourcing studies in the medical domain, however, is the conclusion that a crowd of non-professional, inexperienced users do not underperform medical experts [9], [15].

[1]Amazon Mechanical Turk, https://www.mturk.com

[2]Game with a Purpose, https://www.gwap.com

[3]CrowdFlower, http://www.crowdflower.com/

[4]LabelMe, http://labelme.csail.mit.edu/

[5]CrowdTruth framework—http://crowdtruth.org

Improving the crowd's quality is very essential for being able to generate a reliable ground-truth and creating an interest within the research community. Redundancy and Aggregation (R&A) (i.e., majority voting) is the baseline approach that has been proposed in this context [16], [17]. However, there is no control on the sensitivity and specificity of single participants. All aforementioned crowdsourcing platforms integrate qualification tests in order to restrict "noisy" annotations. This information can then be incorporated into the ground-truth generation process via aggregation. Recently, Raykar *et al.* [18] have proposed a probabilistic model for supervised learning to evaluate different users and estimate the ground-truth labels. Having such ground-truth is very important for both training many machine learning algorithms as well as for evaluation.

Indeed, deep learning has advanced the field of computer vision the last few years [19] leading to powerful methods for various applications such as object classification [5], detection [20], segmentation [21], robust regression [22] and depth prediction [23]. The most established realization of deep learning are Convolutional Neural Networks (ConvNets or CNN) that have also been successfully applied for biomedical imaging purposes [24]–[27]. The bottleneck, however, for deep CNN to yield decent accuracy is the availability of a large number of annotated training samples. In particular in the biomedical domain, sufficient ressources are not available.

We believe that crowdsourcing platforms will engage various crowds to collaborate with clinicians and frontline healthcare workers in translating questions into methodologies and innovative solutions of which ground truth data is an essential part. However, it is not clear how state-of-the-art machine learning methods behave when fed with training data consisting of reliable (expert) and unreliable (crowd) annotations [15]. As suggested by Aroyo *et al.* [15], it is our goal to evaluate the trustworthiness of participants and integrate this knowledge into the analysis and further processing of annotations.

In this manuscript, we present a first attempt to apply the concept of learning from crowds within a biomedical environment. Being inspired by prominent previous work in this field [18], [26], we define the specific contribution of our own work as: i) Learning of a multi-scale CNN model for mitosis detection, ii) Incorporation of aggregation schemes into CNN layers, and iii) Augmentation and retraining of the CNN model with crowd's annotation labels.

In our analysis comparing performance of the CNN model when incorporating different types of aggregations schemes, we aim at answering the following questions: i) Can deep CNN be trained with data collected from crowdsourcing and is it robust against "noisy" labels?, ii) How to adapt the CNN when we have both ground-truth label and multiple annotations that could be "noisy"?, and iii) How is the accuracy compared to that obtained by ground-truth or majority voting?

In this manuscript, after recapitulating previous work in this field, we introduce *AggNet*, a novel aggregation layer that is integrated into our multi-scale CNN. We further present an analysis of the behavior of CNN with and without aggregation on a publicly available large-scale pathological dataset (including ground truth annotations). However, to the best of our knowledge, there has not yet been any effort to incorporate this information into machine learning algorithms analyzing the quality of models learned from non-expert annotations.
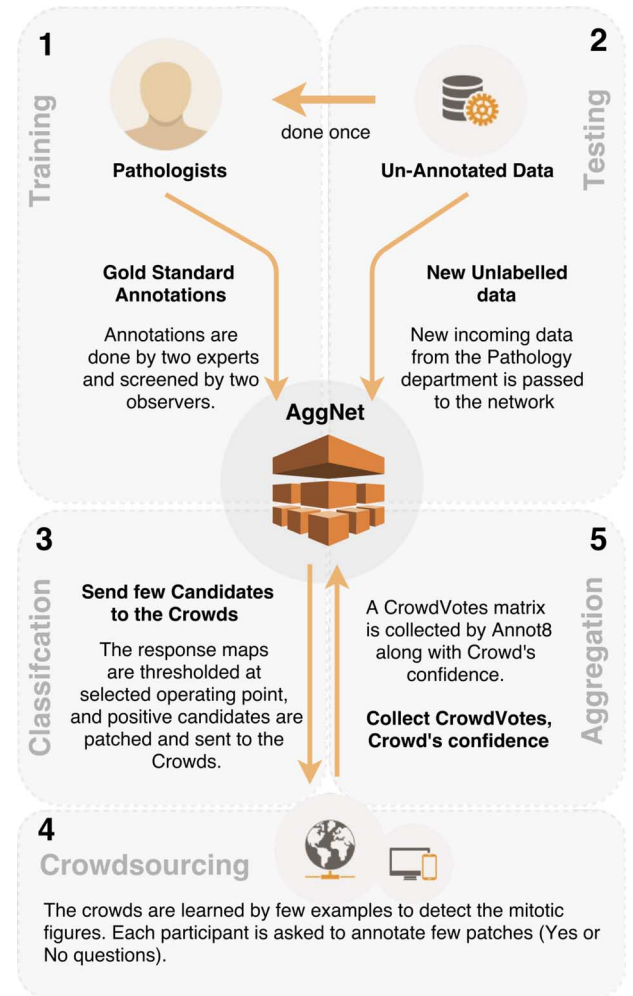

Fig. 1. *AggNet* Framework: (1) The multi-scale CNN model is trained from gold-standard annotations. (2) Then for any incoming unlabelled image, (3) the *AggNet* will produce a response map which is thresholded at selected optimal operating point. (4) These few resulting positive candidates are outsourced to crowds. (5) *AggNet* collects back the crowd votes and jointly aggregates the ground truth and refine the CNN model.

II. METHODOLOGY

In this section, we introduce the proposed CNN for aggregating annotations from crowds in conjunction with learning a model for a challenging classification task. Unlike typical supervised methods, which learn a model from ground truth labeled data, learning from crowd annotations is different in the sense that there may be (possibly noisy) multiple labels for the same sample. Our idea is to learn multiple CNN models with the same basic architecture on different image scales (cf. step 1 in Fig. 1), perform mitosis detection using these models (cf. step 2 in Fig. 1) and provide the crowds with detected mitosis candidates for annotation (cf. step 3 in Fig. 1). The collected annotations are then passed to the existing CNN (cf. step 5 in Fig. 1) with our aggregation layer attached in order to refine the models and simultaneously generate a ground-truth. This multi-scale approach ensures that we have redundant responses of the same data instances at different scales, with the goal to increase robustness of both aggregation and classification.
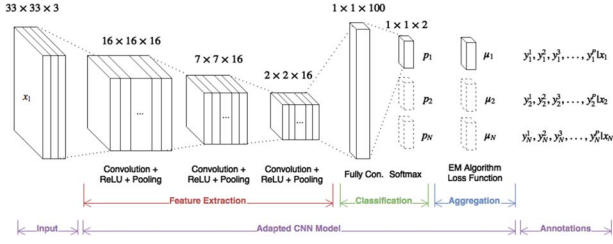
Fig. 2. *AggNet* architecture: The same CNN architecture is used for different scales, where p_i, μ_i, y_i^j represents the classifier output, the aggregated label, and the crowdvotes respectively.

Notation

The input to our network is an observation set $D = \{x_i, y_i^j; i = 1, \ldots, N, j = 1, \ldots, P\}$ containing N instances of $x_i \in \mathbb{R}^d$ (RGB image as d-dimensional vector) with corresponding labels $y_i \in C$ (i.e., $C := \{0, 1\}$ for binary classification) annotated by P independent participants. The goal is to learn a robust CNN model, represented by $f : \mathcal{X} \to \mathcal{Y}$, from aggregated labels which generalizes well on unseen data:

$$\hat{p} = f(\mathbf{x}, \mathbf{y}; \theta), \tag{1}$$

where $\hat{p}$ is the predicted label for an unseen image $\mathbf{x}$, and θ is the learned model parameter.

A. Multi-Scale CNN Model

Our network architecture consists of three convolutional blocks followed by two fully connected (FC) layers as shown in Fig. 2. Each convolutional block consists of a convolutional layer followed by a rectified linear unit (ReLU) [28] and max-pooling layer. The output of the softmax layer is the probabilistic score of the mitotic figures.

In our proposed multi-scale CNN model the input image is first down-sampled to different scales (i.e., 0.33, 0.66 and 1). Then, 33×33 patches are collected and passed to the model (scale-wise). On new unlabelled data, we apply the learned model to mirrored and rotated versions (0, 90, 180, and 270 deg) of each image and compute a final detection map (FDM) as the mean of all those detection results.

FDM of different scales are then geometrically averaged to filter out weak responses. By doing this, we aim at obtaining more accurate detections.

During learning from crowd annotations phase, we augment the CNN architecture with our novel aggregation layer (AG) (Section II-C) in order to i) aggregate the ground-truth from crowdvotes matrix, ii) compute the sensitivity and specificity of each annotator, and iii) jointly learn the classifier by back propagating the derivative of the loss function. We refer to this augmented architecture as *AggNet*.

B. Aggregation Layer (AG)

The straightforward method to aggregate labels annotated by users, is to employ majority voting (MV) [29]:

$$\mu = \begin{cases} 1 & \bar{y} \geq 0.5 \\ 0 & \bar{y} < 0.5 \end{cases} \tag{2}$$

where $\bar{y} = \frac{1}{|P|} \sum_{j=1}^{P} y^j$ is the average label of P users. However, this strategy assumes all users to be on an equal level of trustworthiness.

In our framework, we integrate the method initially proposed by Raykar *et al.* [18], showing a good performance in many applications [17]. On top of our CNN architecture, we aggregate labels μ, estimate the sensitivity α^j and the specificity β^j for each annotator $j \in P$, and jointly learn the classifier. The method is solved using the well-known expectation-maximization (EM) algorithm and adapted to learn the softmax classifier as follows:

- **Initialization:** Using the crowdvotes matrix $\mathbf{Y}$, the aggregated labels μ_i initialized with majority voting, α^j and β^j are initially computed from μ_i.
- **E-Step:** Given the observation set $\mathbf{D}$ and a current estimate of parameters $\psi := \{\alpha, \beta, \mu\}$, the conditional expectation is computed as

$$\mathbb{E}\{\ln Pr[\mathbf{D}, \boldsymbol{g}|\psi]\} = \sum_{i=1}^{N} \mu_i \ln p_i a_i + (1 - \mu_i) \ln(1 - p_i) b_i, \tag{3}$$

where

$$a_i = \prod_{j=1}^{P} [\alpha^j]^{y_i^j} [1 - \alpha^j]^{1 - y_i^j},$$
$$b_i = \prod_{j=1}^{P} [\beta^j]^{1 - y_i^j} [1 - \beta^j]^{y_i^j},$$
$$p_i = \sigma(z_i) = e^{z_{ic}} / \sum_{c=1}^{C} e^{z_{ic}}, \text{ the output of softmax layer,}$$
$$z_i = \mathbf{w}^T \mathbf{x_i}, \text{ the output of FC layer,}$$
$$\boldsymbol{g} \quad \text{is the hidden variable (ground-truth),}$$

and the expectation is with respect to $Pr[\boldsymbol{g}|\mathbf{D}, \psi]$.
Using Bayes' theorem, the aggregated labels μ_i can be computed as follows:

$$\mu_i = \frac{a_i p_i}{a_i p_i + b_i(1 - p_i)}. \tag{4}$$

The loss function in our aggregation layer (AG) is defined as the Negative log-likelihood:

$$\mathcal{L}(\mu_i, p_i) = -\mathbb{E}\{\ln Pr[\mathbf{D}, \boldsymbol{g}|\psi]\}. \tag{5}$$

- **M-Step:** Based on the observation set $\mathbf{D}$ and the current estimate μ_i, the model parameters ψ can be computed by taking the derivative of $\mathcal{L}$ with respect to each parameter and equate it to zero. The updates for α^j and β^j can be obtained as follows:

$$\alpha^j = \frac{\sum_{i=1}^{N} \mu_i y_i^j}{\sum_{i=1}^{N} \mu_i}, \quad \beta^j = \frac{\sum_{i=1}^{N} (1 - \mu_i)(1 - y_i^j)}{\sum_{i=1}^{N} (1 - \mu_i)}. \tag{6}$$

The softmax function is non-linear and the gradient with respect to parameter $\mathbf{w}$ should be back propagated to the CNN layers [30]. For this purpose, we can employ the chain rule:

$$\frac{\partial \mathcal{L}}{\partial w} = \frac{\partial \mathcal{L}}{\partial p_i} \frac{\partial p_i}{\partial z_i} \frac{\partial z_i}{\partial w}, \tag{7}$$

where

$$\frac{\partial \mathcal{L}}{\partial p_i} = \frac{\mu_i - p_i}{p_i(1-p_i)}, \text{ the output of AG Layer,}$$

$$\frac{\partial p_i}{\partial z_i} = p_i(\delta_{ij} - p_j), \text{ the output of softmax layer[6],}$$

$$\frac{\partial z_i}{\partial w} = \mathbf{x}_i, \text{ the output of FC layer.}$$

Then, weights are updated using Stochastic Gradient Descent (SGD) [31].

It is notable that **E-Step** and **M-Step** are computed in forward and backward propagation respectively, which means one EM iteration per epoch. The refining process is stopped when the loss function output barely changes to avoid overfitting.

To this end, the aggregation method takes into account the sensitivity and specificity of each annotator to aggregate the labels. Furthermore, the algorithm is adapted to handle:

- **Trustworthiness:** Some crowdsourcing platforms provide the customer with a single accuracy score γ that each user achieved on a qualitative test for a specific task. It has been suggested by Raykar *et al.* [18] to model a prior distribution on sensitivity and specificity to trust some participants more than others. With only a single accuracy score as provided by our scenario, this is however not possible.
- **Missing Labels:** A common failure in crowdsourcing is that some users annotate a few samples only. If all these samples happen to fall within one single class, sensitivity or specificity remains *unknown*. Furthermore, the user who annotates few samples only might be equally or even more trusted than a user who annotates more samples.

Therefore, we reformulate α^j and β^j, without loss of generality, in such a way to augment the number of True Positives (TP) and True Negatives (TN) when the user has high confidence (i.e., accuracy score) as follows:

$$\alpha^j = \frac{\tau\gamma^j|N_p| + \sum_{i=1}^{N_p}\mu_i y_i^j}{\tau\gamma^j|N_p| + \sum_{i=1}^{N}\mu_i} = \frac{(1+\tau\gamma^j)TP}{(1+\tau\gamma^j)TP + FN},$$

$$\beta^j = \frac{\tau\gamma^j|N_n| + \sum_{i=1}^{N_n}(1-\mu_i)(1-y_i^j)}{\tau\gamma^j|N_n| + \sum_{i=1}^{N}(1-\mu_i)} = \frac{(1+\tau\gamma^j)TN}{(1+\tau\gamma^j)TN + FP}, \quad (8)$$

where γ^j is the accuracy score for a particular user and τ is the hyper-parameter that leverage the user's confidence. To avoid numerical issues, we set the sensitivity and specificity to 0.5 for *unknown* cases.

III. EXPERIMENTS AND RESULTS

We have designed our experimental setup such that first, the proposed multi-scale CNN architecture is validated before evaluating the aggregated labels from the crowdvotes and validating the proposed augmented CNN (*AggNet*).

1) Dataset: We have validated our proposed network on the publicly available MICCAI-AMIDA13 challenge dataset[7]. It contains annotated histology images of a total of 23 patients,

[6]The Kronecker delta, $\delta_{ij} = \begin{cases} 1 & i=j \\ 0 & i \neq j \end{cases}$

[7]AMIDA13: http://amida13.isi.uu.nl/

who underwent invasive breast biopsy. During this medical examination, sections of suspicious breast tissue are collected and stained using hematoxylin and eosin (H&E). A histology RGB image of $2k \times 2k$ is then acquired with the Aperio ScanScope XT scanner at $40X$ magnification and with a spatial resolution of 0.25 μm/pixel. Then, a region of interest is identified and digitized to several high power field (HPF) images. The standard procedure in pathology is to count the mitotic figures in this area for the purpose of cancer grading (mitotic count score criteria). The annotation in the AMIDA13 dataset was done by two expert pathologists. Concordant annotations of both experts were taken as ground truth objects directly, whereas discordant cases were presented to two additional observers, such that the ground truth have been agreed upon by at least two experts. The reader is referred to [32] for more information about the dataset and its clinical/pathological background. In our experiment, we learn the proposed initial multi-scale model from 12 patients (311 HPF images), validate on 20% of the training set (60 HPF images) and test it on the whole testing data of AMIDA Challenge, including 11 patients (295 HPF images).

2) Implementation Details: Each input RGB image is first pre-processed by staining appearance normalization [33]. Then, small patches of 33×33 are collected. Furthermore, to handle highly imbalanced data, patches showing positive classes are augmented with rotation and mirroring in such a way to leverages the ratio of positive to negative classes about (3:7). The muti-scale CNN is implemented using MATLAB and MatConvNet [34] and conducted on an Intel i7 machine with a GeForce GT 750M graphics card. Concerning the network parameters, the learning rate is set to 1×10^{-3}, momentum to 0.9, weight decay to 5×10^{-4}, and the batch size fixed to 200 samples. Note that some of these parameters are changed in the refining process, i.e., learning rate is set to 5×10^{-5} and the batch size is set to the whole crowdsourcing set. For the sake of reusability and to overcome the limitations of the crowdsourcing platform Crowdflower, we have designed and implemented Annot8[8], a Ruby-on-Rails based web-platform, allowing registered users to create datasets, upload images and labels, and categorize the labels with the help of a powerful tagging system. Collections of existing labels can be sent to and crowdsourced labels can be imported from Crowdflower easily. Our web-platform also offers an online image processing frontend for on-demand patch extraction and computation of biomedical image filtering. On the participant side, each user was introduced briefly about the disease and the instructions of the actual task showing some good and bad examples as shown in Fig. 3. Then, participants had to conduct a few test questions for quality control purposes. Without being made aware of the quiz mode, each annotator was presented with patches with known labels. Only then, he/she started to annotate five patches presented along with the filtered images. In order to ensure continuous quality control, a few randomly seeded test patches were still shown during the actual annotation job.

3) Evaluation Metrics: We calculate different validation measures for comparison purpose, such as Recall $= \frac{TP}{TP+FN}$

[8]Annot8: http://vmnavab14.informatik.tu-muenchen.de/
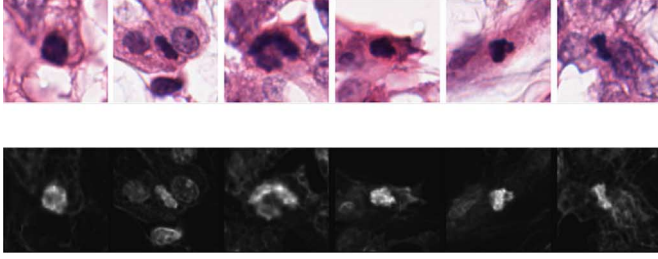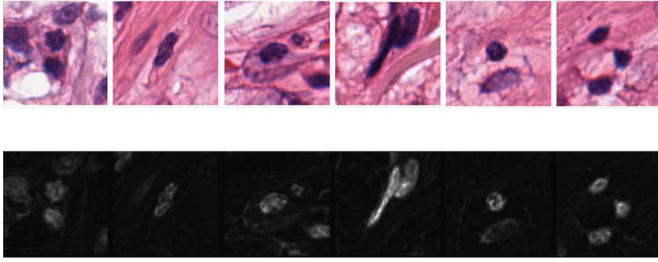
Tips and Examples
Mitosis:

The second row shows the corresponding so called "blueRatio" representation of the mitotic figures. Note how they have very bright spots!

Non-Mitosis

The second row shows the corresponding so called "blueRatio" representation of the non-mitotic figures. Note how they do not have such bright spots as the mitotic blue ratio representations!

Tips:

Mitotic figures usually look much more irregular than non-mitotic ones and appear to be very dark. In their blue ratio representation, they tend to have very bright spots. Watch out for very dark blobs that have a highly irregular (not really circle-like) shape and that have very bright spots in their blue ratio representation!

Fig. 3. Instructions and guidelines.

and Precision $= \frac{TP}{TP+FP}$, where TP, FP and FN represent the true positives, false positives and false negatives respectively.

We further employ widely used statistical measures, such as F_1-score $= \frac{2 \times TP}{FP+FN+2 \times TP}$, Receiver Operating Characteristics (ROC) and its Area Under Curve (AUC).

For measuring the improvement of multi-scale CNN over single scales, we compute the mean and standard variation of Relative Changes (RC) of different scales, i.e., $\mu_{RC} = \frac{1}{|Scales|} \sum_{Scales} \frac{MultiScale-xScale}{xScale}$.

Proof-of-Concept Evaluation

The objective of this experimental setup is to analyze the functionality of the entire *AggNet* framework (cf. Fig. 1), specifically i) the accuracy of multi-scale CNN, ii) the performance of our novel aggregation layer when fed with noisy annotations, and iii) the influence of the augmented model on the detection quality of the multi-scale CNN. In order to perform quantitative analysis with respect to real ground-truth data, we decided to employ the AMIDA13 Challenge training dataset only as it comes with ground truth annotations (cf. Table I). The setup is designed according to our overall framework pipeline depicted in Fig. 1. First, a model is learned on 8 random patients of the AMIDA13 Challenge training dataset and tested on different 3 patients. Training and testing is performed employing

TABLE I
DATASETS SPECIFICATIONS

	Proof of Concept	Use-Case
Model	8 Patients of Training set (318 HPF images)	The Entire Training set (371 HPF images)
Testing	3 Patients of Training set (22 HPF images)	The Entire Testing set (295 HPF images)
Crowdsourcing	Positive Candidates (550 Patches)	Positive Candidates (750 Patches)

TABLE II
F_1-SCORES

	Patient 9	Patient 11	Patient 12	Overall
0.33-Scale	**0.8000**	0.5833	**0.7778**	0.6479
0.66-Scale	0.5000	0.5556	0.6957	0.5882
Orig.-Scale	0.5000	0.5490	0.6957	0.5854
Multi-Scale	**0.8000**	**0.7368**	0.7368	**0.7419**
Improvement	40%±36	39%±11	2.2%±6.4	22.5%±6.8

our novel multi-scale CNN *AggNet*. Then, response maps of *AggNet* are thresholded at a lower operating point ensuring a large number of positive candidates, repatched and sent to CrowdFlower using our Annot8 web-platform.

In this experiment, we have crowdsourced around 550 patches, where each patch was annotated by 10 participants at least, resulting in more than 5500 labels stored in the crowdvotes matrix $\mathbf{Y}$. We have then evaluated results according to the different aspects related to the objectives defined for this specific experimental setup:

4) Multi-Scale CNN: In order to measure the performance of the multi-scale CNN, we train the network on three different scales (0.33, 0.66 and 1) individually. For inference, we geometrically average the resulting FDMs from all scales to obtain the final positive responses. As alternative method to extract positive responses, we also threshold the FDM for each scale. For each remaining response, we check whether it is a TP, FP or FN detection. A positive response is considered a TP if its Euclidean distance to a ground-truth mitotic figure is less or equal than 30 px. Multiple responses within the same radius around a mitotic figure are counted as a single TP. If there is no response for a mitotic figure within this radius, we count a FN detection. Any responses that are not inside any 30 px region around a mitotic figure are counted as FP. It should be noted that a 30 px radius (7.5μm) is used on the original scale, however, this is adjusted for different scales. Using these numbers, we calculate the Precision, Recall, and F_1-score over all HPF slides at once and also per patient. Table II shows the F_1-score of 22 HPF images, the corresponding testing dataset, while the bar plot in Fig. 5 displays the other metrics. It is obvious that the multi-scale CNN approach pushed the overall F_1-score about $22.5\% \pm 6.8$, which validates our initial hypothesis of the proposed multi-scale CNN approach yielding a more robust classification due to detection consensus at various scales.

5) Aggregated Labels (AL): To investigate the aggregated labels of the crowdvotes matrix $\mathbf{Y}$, we first run majority voting (MV) [29] and GLAD [35] methods on $\mathbf{Y}$ without any quality control, referred to as MV-NoQ and GLAD-NoQ respectively.
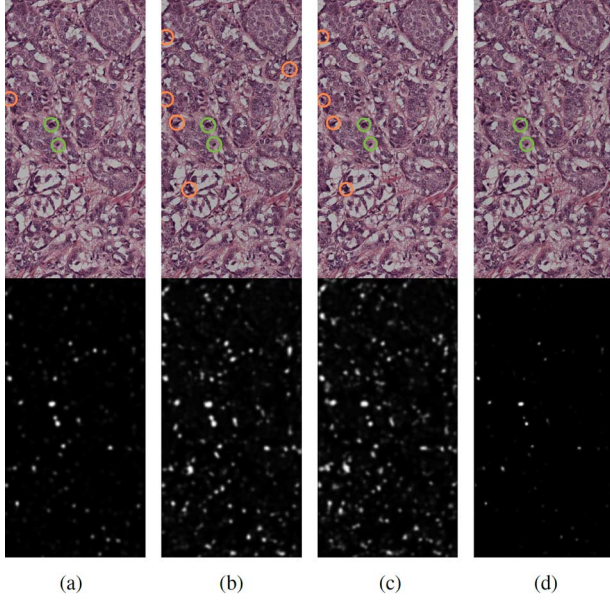
(a)   (b)   (c)   (d)

Fig. 4. First row shows results of one single image using multi-scale CNN, Green: the true positives, Orange: the false positives. Second row shows the corresponding final detection map (FDM) before thresholding. Best viewed in color. (a) 0.33-Scale (b) 0.66-Scale (c) Orig.-Scale (d) Mulit-Scale.
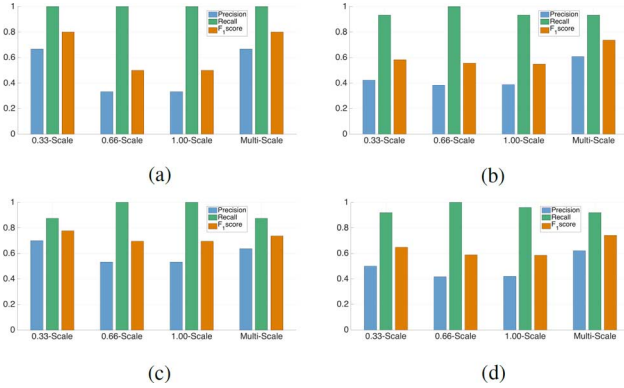

Fig. 5. Evaluation Metrics: Precision, Recall, and F_1-score of Patients 9, 11 and 12. (a) Patient 9 (b) Patient 11 (c) Patient 12 (d) Overall.

TABLE III
AGGREGATED LABELS

	Aggregation Method	Quality Control	Prior
MV-NoQ	MV	No	-
GLAD-NoQ	GLAD	No	-
AG-NoQ	Proposed	No	-
AG-NoQ-τ	Proposed	No	τ
HP-70Q	HP	70%	-
GLAD-70Q	GLAD	70%	-
AG-70Q	Proposed	70%	-
AG-70Q-τ	Proposed	70%	τ

Then, we use the existing 0.33-scale model from the previous multi-scale CNN experiment and augment it with our AG layer. Subsequently, we retrain it further for 100 epochs. We refer to this architecture as *AggNet*, and the aggregation results as AG-NoQ.

Further, to test the quality control, we filtered first the crowd-votes matrix $\mathbf{Y}$ in order to keep only the annotations from the
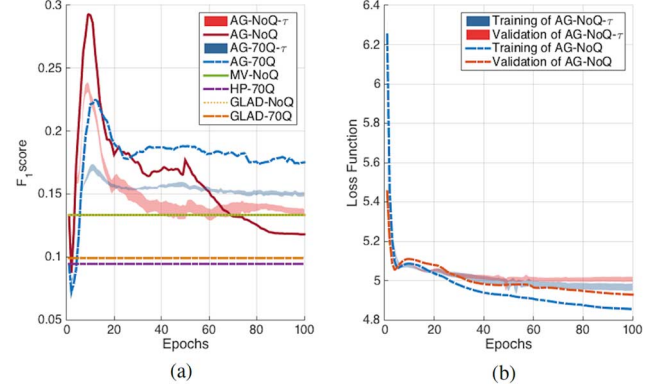

Fig. 6. (a) The aggregated labels of the crowdsourcing set are evaluated using the F_1-score metric. (b) The loss function barely changes at 3–8 epochs before starting to overfit and the gap between the validation and training curves becomes significant. The shaded area depicts the change of τ. (a) F_1-score (b) Loss function.
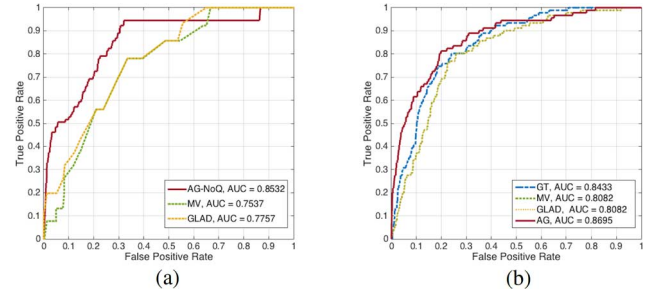

Fig. 7. ROC curves of the (a) aggregated labels using MV, GLAD and the proposed AG-NoQ, (b) the augmented models AM-GT, AM-MV, AM-GLAD and *AggNet* as well. (a)Aggregated Labels (b) Augmented Models.

TABLE IV
AUGMENTED MODELS

	AM-GT	AM-MV	AM-GLAD	*AggNet*
F_1-score	**0.6250**	0.6097	0.6097	0.6133
AUC	0.8433	0.8082	0.8082	**0.8695**
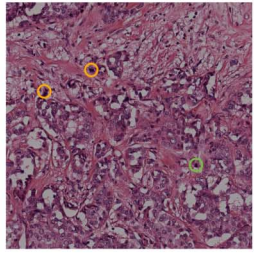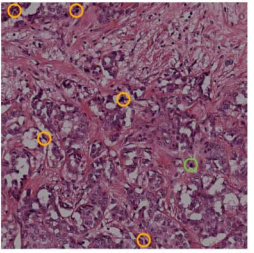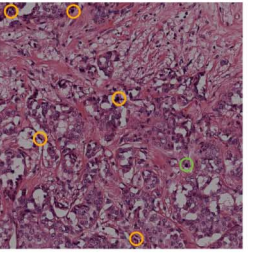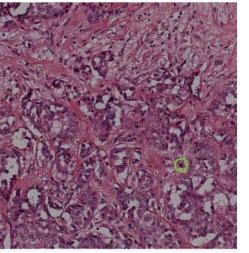
users who achieved more than 70% accuracy score in their qualitative test (in a quiz stage, each user had to annotate a few samples extracted from training data with known gr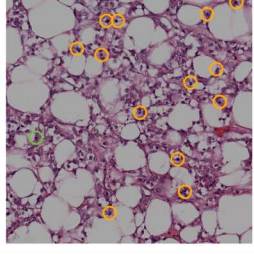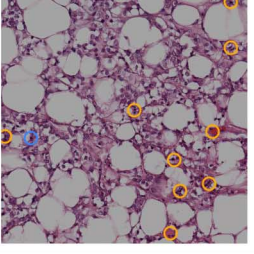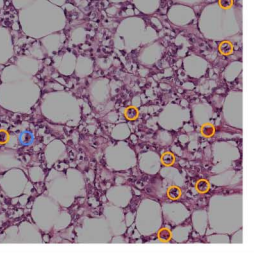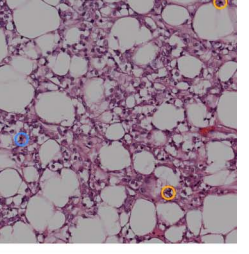ound-truth). Then, we run the same aggregation methods, however, MV is replaced with Honeypot (HP) [36]. We refer to the aggregated labels using 70% quality control as HP-70Q, GLAD-70Q, and AG-70Q.

In addition, to figure out how the accuracy scores γ of the participant can influence the proposed aggregation method, we validate the hyper-parameter $\tau = [0.1, 0.2, \ldots, 1]$ (ref. Section II-C) and run similar experiments, referred to as AG-NoQ-τ and AG-70Q-τ respectively.

To evaluate the aforementioned aggregated labels of crowd-sourced patches (cf. Table III), we compute the F_1-score between the ground-truth and the aggregated labels of different methods as well as the proposed method (cf. Fig. 6(a)). Further, to give the reader more insight on the probabilistic scores of different aggregation methods, we plot the ROC curves of the significant methods as shown in Fig. 7(a).

TABLE V
AGGREGATION AND DETECTION RESULTS

Augmented Models	AM-GT	AM-MV	AM-GLAD	The proposed *AggNet*
Perfect				
Off				

Unlike the weak agreement of the aggregated labels of both MV-NoQ and GLAD-NoQ with the ground-truth, both AG-NoQ and AG-NoQ-τ (-70Q as well) achieve an out-performing agreement at the first few epochs before getting decayed and saturated at still good agreement compared with the other aggregation methods as shown in Fig. 6(a). MV-NoQ and GLAD-NoQ coincide in this setup due to the choice of our label threshold $= 0.5$. The ROC curve of the respective methods shows that GLAD is slightly superior to MV.

Interestingly, the aggregation methods without any quality control perform better than the ones with 70% quality control, which shows that the quality control strategy should be revised for such challenging data. Notably, AG-NoQ-τ as well as AG-70Q-τ, which incorporate the gamer accuracies as a prior, underperform the other AG-models. Once again, this shows that the quality control strategy and thus the accuracy is a potential bottleneck.

6) Augmented Models (AM): We further investigate how aggregated labels obtained from the previous experiment, influence the detection quality of our CNN compared to the ground-truth model. For this purpose, we compute several augmented models based on 0.33-scale of the initially trained ground truth model (GT). Besides *AggNet*, which we obtain by attaching our AG layer to GT, we compute three additional distinct models by retraining with aggregated labels from MV and GLAD as well as the real ground truth labels. We refer to the augmented models as AM-MV, AM-GLAD and AM-GT respectively. In fact, we retrain for 100 epochs, but pick the best performing model for each. Note that AM-GT is the 0.33-scale CNN model, however, its operating point is set to the same threshold that was used to select positive candidates for crowdsourcing. Once retrained, we utilize the models to perform mitosis detection on the corresponding testing set. Fig. 7(b) and Table IV nicely outline how the proposed *AggNet* model almost performs as good as the augmented ground truth model AM-GT and easily outperforms both AM-MV and AM-GLAD (around 7.6% of AUC). Further-

more, some HPF images are visualized in Table V to show how different augmented models perform on their optimal operating points. It is worth noticing also here how the proposed *AggNet* hits only the ground-truth in the perfect scenario outperforming even the AM-GT, or miss some ground-truth while having very few false positive in off scenario.

A. Use Case Evaluation

This experiment aims at proving the overall impact of *AggNet* on a large standardized dataset (entire AMIDA13 Challenge training and testing datasets).

To first evaluate the performance of our proposed multi-scale CNN, we have participated in the AMIDA13 Challenge with our novel approach achieving 0.433 as overall F_1-score. As reported by the challenge organizers, our method yields rank three of 15 participating methodologies.

Response maps resulting from the AMIDA13 Challenge testing dataset on 0.33-Scale have then been thresholded at an operating point of 0.99 (calculation based on the dataset) and repatched samples have been forwarded to CrowdFlower. Similarly to the previous experimental setup, crowdvotes $\mathbf{Y}$ and confidences have been fed back to *AggNet* in order to augment the previously trained model. For performance evaluation, different augmented models have also participated in the AMIDA13 Challenge. Table VII shows the evaluation metrics of different models augmented with MV, GLAD and our robust aggregation layer (*AggNet*). The overall F_1-score of *AggNet* easily outperforms the other augmented models of MV and GLAD, however, it falls slightly behind the previously trained model (0.33-Scale).

IV. DISCUSSION

Our results confirm that aggregation and deep learning from crowd annotations using the proposed *AggNet* is robust to "noisy" labels and positively influences the performance of our CNN in the refining phase.

TABLE VI
USE CASE RESULTS

	Precision	Recall	Overall F_1-score
0.33-Scale	0.211	0.538	0.303
0.66-Scale	0.296	**0.583**	0.393
Orig.-Scale	0.172	0.400	0.241
Multi-Scale	**0.441**	0.424	**0.433**
Improvement	105%±54	-14%±18	44%±35

TABLE VII
USE CASE AUGMENTED MODELS

	0.33-Scale	AM-MV	AM-GLAD	*AggNet*
Precision	0.211	0.006	0.006	**0.374**
Recall	**0.538**	0.004	0.004	0.208
F_1-score	**0.303**	0.004	0.005	0.267

Proper selection of operating point for crowdsourcing is quite challenging, for instance, it can be chosen based on the validation set, however, it might not be optimal for crowdsourcing where you need more positives. Therefore, it is recommended to plot the AUC, which provides the clinicians with more flexibility to run the model at the optimal operating point from their perspective. The augmented model *AggNet* has a gain of 7.6% in AUC at lower than the optimal operating point, however, this can not be computed for the use case experiment due to the limited number of submissions to the AMIDA Challenge.

However, the aggregation results from quality-filtered crowd annotations shed the light on the applied quality control. Surprisingly, they turned out to be worse than the aggregation from the complete, "noisy" set of crowd votes. This is clearly related to the high ambiguity and the high level of difficulty in detecting mitotic figures in general. Indeed, such quality tests need to be carefully planned and well designed in order to make sure they do not carry more "noise" than the actual crowdsourcing task, which we believe, the latter was the case in our experiments. However, this problem can also be related to the community of the crowd itself [37].

The low agreement among the crowd together with the small number of patches might be the reasons of the noticed decay in the aggregation F_1-score curve, which leads to an overfitting after only few epochs (see Fig. 6). Nevertheless, it is obvious that our robust aggregation layer can detect the novices (spammers) and weight their votes less. Therefore, our *AggNet* outperforms easily the other augmented models, where the spammers hurt the aggregations (cf. Table VII). Fig. 8 shows the accuracy scores γ (based on the qualitative test) and the spammer scores S_p (based on the participant's sensitivity α and specificity β, where $S_p = (\alpha+\beta-1)^2$ [38]) of 100 participants in the crowdsourcing task.

During our research, the very natural question arose whether it may be possible to learn a model from crowdsourcing labels alone. For this purpose, we ran two additional crowdsourcing experiments. First, we utilized the crowdsourcing set (i.e., 5500 patches) and the binary classification crowdvotes to learn a model from scratch. Very soon, we realized that this model quickly overfited due to the small number of training instances.
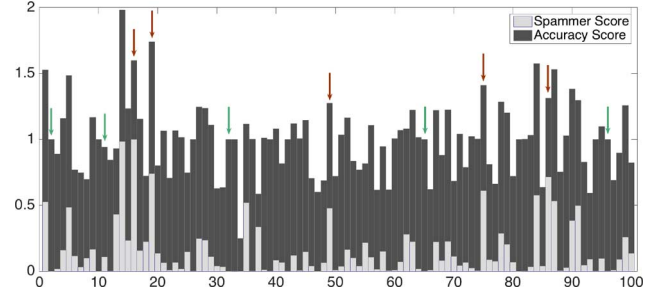

Fig. 8. Participants Analysis: accuracy and spammer scores of 100 participants. Arrows in green show some participants achieve high accuracy scores in the qualitative test, however, they are spammer. Arrows in red show very few participants who have good accuracy score as well as spammer score. Note that spammer score "0" means the participant is spammer.

Second, instead of publishing patches only, we asked the crowd to label full HPF images. In this case, however, due to insufficient settings within CrowdFlower platform, each participant labeled only few potential mitotic figures per image and left most of the challenging cases besides. This led to a large number of missing annotations and poor overall agreement among participants, which renders aggregation and training impossible. Still, these initial experiments gave us evidence that it is very difficult and maybe even impossible to entirely outsource the task of labeling mitotic figures in histology images to crowds. Instead, we decided to rather augment a small and narrow model learned from expert labels with wide but noisy crowd annotations to enhance variations encoded within the model.

Validating our methodology based on the smallest scale of our initial model is theoretically feasible. However, in future work, we want to conduct even more involved experiments including also the models of the other scales. This includes independent crowd sourcing rounds for each scale separately since retraining all the scales of the multi-scale model from the same crowd-sourced and aggregated labels hurts the concept of "redundancy & aggregation". Additionally, we want to consider multi-class classification which is, due to the binary nature of sensitivity and specificity, not directly possible, but can be performed in an one-vs-all fashion.

V. CONCLUSION

In this paper, we have introduced a novel concept for learning from crowds. Our new multi-scale CNN *AggNet* is designed to handle data aggregation directly as part of the learning process via an additional crowdsourcing layer. In our experimental study, we have further presented valuable insights into the functionality of deep CNN learning from crowd annotations and proven the impact of our novel aggregation scheme. To the best of knowledge, this is the first time that deep learning has been applied to generate a ground-truth labeling from non-expert crowd annotation in a biomedical context.

Although data aggregation is certainly necessary to learn from crowds, computational aggregation models have a limited impact, in particular if noisy crowd annotations are not significant, i.e., do not arise from ambiguous contexts. Besides clear guidelines, non-expert users need to be motivated to perform the task until the very end. Gamification is the ultimate solution

here and we will focus future work on novel solutions on how to transform complex expert tasks in the biomedical domain into a game for non-expert users.

ACKNOWLEDGMENT

REFERENCES

[1] E. Estellés-Arolas and F. González-Ladrón-De-Guevara, "Towards an integrated crowdsourcing definition," *J. Inf. Sci.*, vol. 38, no. 2, pp. 189–200, 2012.

[2] J. Howe, "The rise of crowdsourcing," *Wired Mag.*, vol. 14, no. 6, pp. 1–4, 2006.

[3] F. Kleemann, G. Voß, and K. Rieder, "Un(der)paid innovators: The commercial utiliza-tion of consumer work through crowdsourcing," *STI Studies*, vol. 4, no. 1, 2008.

[4] L. Von Ahn, B. Maurer, C. McMillen, D. Abraham, and M. Blum, "Recaptcha: Human-based character recognition via web security measures," *Science*, vol. 321, no. 5895, pp. 1465–1468, 2008.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Adv. Neural Inf. Proc. Sys.*, 2012, pp. 1097–1105.

[6] O. Inel *et al.*, "Crowdtruth: Machine-human computation framework for harnessing disagreement in gathering annotated data," in *Proc. Int. Semantic Web Conf., Part II*, 2014, pp. 486–504.

[7] L. Von Ahn, "Games with a purpose," *Comput.*, vol. 39, no. 6, pp. 92–94, 2006.

[8] B. Yu, M. Willis, P. Sun, and J. Wang, "Crowdsourcing participatory evaluation of medical pictograms using amazon mechanical turk," *J. Med. Internet Res.*, vol. 15, no. 6, p. E108, 2013.

[9] L. Maier-Hein *et al.*, "Crowdsourcing for reference correspondence generation in endoscopic images," in *Medical Image Computing and Comput.-Assisted Intervention—MICCAI 2014*, 2014, pp. 349–356.

[10] D. Volpi *et al.*, "Online tracking of interventional devices for endovascular aortic repair," *Int. J. Comp. Assist. Radiol. Surg.*, vol. 10, no. 6, pp. 773–781, 2015.

[11] S. Mavandadi *et al.*, "Biogames: A platform for crowd-sourced biomedical image analysis and telediagnosis," *Games Health J.*, vol. 1, no. 5, pp. 373–376, 2012.

[12] D. Gurari *et al.*, "How to collect segmentations for biomedical images? A benchmark evaluating the performance of experts, crowdsourced non-experts, and algorithms," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2015, pp. 1169–1176.

[13] A. Foncubierta Rodríguez and H. Müller, "Ground truth generation in medical imaging: A crowdsourcing-based iterative approach," *Proc. ACM Multim. Work. Crowdsourcing Multimedia*, pp. 9–14, 2012.

[14] L. A. Celi, A. Ippolito, R. A. Montgomery, C. Moses, and D. J. Stone, "Crowdsourcing knowledge discovery and innovations in medicine," *J. Med. Internet Res.*, vol. 16, no. 9, 2014.

[15] L. Aroyo and C. Welty, "Truth is a lie: Crowd truth and the seven myths of human annotation," *AI Mag.*, vol. 36, no. 1, 2015.

[16] V. S. Sheng, F. Provost, and P. G. Ipeirotis, "Get another label? Improving data quality and data mining using multiple, noisy labelers," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2008, pp. 614–622.

[17] N. Q. V. Hung, N. Tam, L. Tran, and K. Aberer, "An evaluation of aggregation techniques in crowdsourcing," *Web Inf. Syst. Eng.*, pp. 1–15, 2013.

[18] V. C. Raykar *et al.*, "Learning from crowds," *J. Mach. Learn. Res.*, vol. 11, pp. 1297–1322, 2010.

[19] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, 2015.

[20] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[21] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440.

[22] V. Belagiannis, C. Rupprecht, G. Carneiro, and N. Navab, "Robust optimization for deep regression," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 2830–2838.

[23] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Adv. Neural Inf. Proc. Syst.*, 2014.

[24] Y. Xie, F. Xing, X. Kong, H. Su, and L. Yang, "Beyond classification: Structured regression for robust cell detection using convolutional neural network," in *Medical Image Computing and Comput.-Assisted Intervention—MICCAI 2015*, 2015, pp. 358–365.

[25] F. Liu and L. Yang, "A novel cell detection method using deep convolutional neural network and maximum-weight independent set," in *Medical Image Computing and Comput.-Assisted Intervention—MICCAI 2015*, 2015, pp. 349–357.

[26] D. C. Ciresan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Mitosis detection in breast cancer histology images with deep neural networks," in *Medical Image Computing and Comput.-Assisted Intervention—MICCAI 2013*, 2013, pp. 411–418.

[27] G. Carneiro, J. C. Nascimento, and A. Freitas, "The segmentation of the left ventricle of the heart from ultrasound data using deep learning architectures and derivative-based search methods," *IEEE Trans. Image Process.*, vol. 21, no. 3, pp. 968–982, Mar. 2012.

[28] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proc. Int. Conf. Mach. Learn.*, 2010, pp. 807–814.

[29] L. I. Kuncheva, C. J. Whitaker, C. A. Shipp, and R. P. Duin, "Limits on the majority vote accuracy in classifier fusion," *Pattern Anal. Appl.*, vol. 6, no. 1, pp. 22–31, 2003.

[30] Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, "Efficient backprop," in *Neural Netw., Tricks of the Trade*, 2012, pp. 9–48.

[31] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proc. Computational Statist.*, 2010, pp. 177–186.

[32] M. Veta *et al.*, "Assessment of algorithms for mitosis detection in breast cancer histopathology images," *Med. Image Anal.*, vol. 20, no. 1, pp. 237–248, 2015.

[33] M. Macenko *et al.*, "A method for normalizing histology slides for quantitative analysis," in *Proc. ISBI*, 2009, vol. 9, pp. 1107–1110.

[34] A. Vedaldi and K. Lenc, "Matconvnet-convolutional neural networks for MATLAB," in *Proc. ACM Int. Conf. Multimedia*, 2015.

[35] J. Whitehill, T.-f. Wu, J. Bergsma, J. R. Movellan, and P. L. Ruvolo, "Whose vote should count more: Optimal integration of labels from labelers of unknown expertise," *Adv. Neural Inf. Process. Syst.*, pp. 2035–2043, 2009.

[36] K. Lee, J. Caverlee, and S. Webb, "The social honeypot project: protecting online communities from spammers," in *Proc. Int. Conf. World Wide Web*, 2010, pp. 1139–1140.

[37] M. Venanzi, J. Guiver, G. Kazai, P. Kohli, and M. Shokouhi, "Community-based bayesian aggregation models for crowdsourcing," in *Proc. Int. Conf. World Wide Web.*, 2014, pp. 155–164.

[38] V. C. Raykar and S. Yu, "Ranking annotators for crowdsourced labeling tasks," in *Adv. Neural Inf. Proc. Sys.*, 2011, pp. 1809–1817.