Eurographics Workshop on Visual Computing for Biology and Medicine (2017)
I. Hotz, D. Merhof, and C. Rieder (Editors)

A Guided Spatial Transformer Network for Histology Cell Differentiation

Marc Aubreville[1], Maximilian Krappmann[1], Christof Bertram[2], Robert Klopfleisch[2] and Andreas Maier[1]

[1]Pattern Recognition Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
[2]Institute of Veterinary Pathology, Free University Berlin, Germany

Abstract

Identification and counting of cells and mitotic figures is a standard task in diagnostic histopathology. Due to the large overall cell count on histological slides and the potential sparse prevalence of some relevant cell types or mitotic figures, retrieving annotation data for sufficient statistics is a tedious task and prone to a significant error in assessment. Automatic classification and segmentation is a classic task in digital pathology, yet it is not solved to a sufficient degree.
We present a novel approach for cell and mitotic figure classification, based on a deep convolutional network with an incorporated Spatial Transformer Network. The network was trained on a novel data set with ten thousand mitotic figures, about ten times more than previous data sets. The algorithm is able to derive the cell class (mitotic tumor cells, non-mitotic tumor cells and granulocytes) and their position within an image. The mean accuracy of the algorithm in a five-fold cross-validation is 91.45 %.
In our view, the approach is a promising step into the direction of a more objective and accurate, semi-automatized mitosis counting supporting the pathologist.

CCS Concepts
•***Computing methodologies*** → ***Object detection;*** *Neural networks;* •***Applied computing*** → *Bioinformatics;*

1. Introduction

The assessment of cell types in histology slides is a standard task in pathology. Especially in tumor diagnostics, determining the relative amount of mitotic figures, a marker for tumor proliferation and aggressiveness, is another important task for the diagnostic pathologist [RRL*13].

However, evaluation of the complete slide for mitotic figures (see figure 1) is usually too time consuming in routine diagnostics. Therefore it is suggested that only 10 high power fields (an area of assumed equal size used for statistic comparison), presumed to contain the highest density of mitoses, are subjectively chosen by the pathologist. The area of these fields is, however, not well-defined, as it depends on the optical properties of the microscope and which may vary significantly in their content of mitotic figures [MMG16]. The final count thus strongly depends on the randomly but not necessarily representatively selected high power fields thus the resulting mitotic count is usually observer-dependent [VvDW*14]. In addition, mitotic figures may be very variable in their histologic phenotype, which may also lead to inter-observer variability between pathologists.

The aim of this work is to develop a more objective and accurate, automatized approach to counting of mitotic figures by assisting pathologists in the selection of fields with the highest mitotic counts and with more constant parameters of mitotic figure identification. Detection and annotation of mitotic cells in histology slides is a well-known task in images processing, and subject of several challenges in recent years [VvDW*14, RRL*13].

Mitosis comprises a number of different phases in the cell cycle (prophase, metaphase, anaphase, and telophase). In each phase, the nucleus is shaped differently. This means that the variance in images showing a mitotic cell is high (see figure 1). On top of that, there is also atypical mitosis, adding yet another factor of variance to the picture. However, publicly available databases for mitosis detection feature a rather low number of mitotic figures (e.g. the 2014 ICPR MITOS-ATYPIA-14 dataset with 873 images, the 2012 ICPR dataset with 326 images [RRL*13], or the AMIDA13 dataset with 1083 images [VvDW*14]), especially for robust detection.

Automatic detection of mitotic figures has been widely performed using the classical machine learning workflow on textural, morphological and shape features (e.g. [SFHG12, Irs13]).Cireşan *et al.* were the first to employ deep learning-based approaches for mitosis detection [CGGS13], yielding significant improvements over traditional approaches [VvDW*14]. Yet, deep learning technologies suffer considerably from insufficient data amounts, as they

have a large number of trainable parameters and, because of this, are likely to overfit the data. Particularly in the field of mitosis detection, we assume that detection performance could be improved if the whole variance of mitotic processes can be captured in the networks, requesting for a substantial increase in training data for such networks.

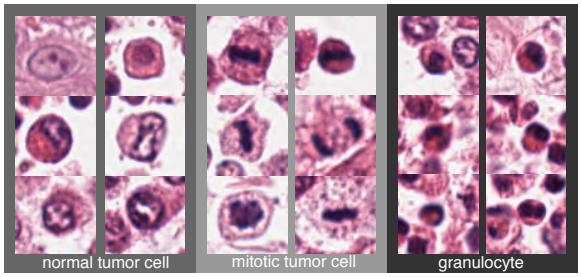

Figure 1: *Examples of cropped cells, slides stained with hematoxylin and eosin.*

2. Related Work

Typically, the process of object detection is parted into two subprocesses: Segmentation and classification. This setup is especially sensible for histology since the images represent a large amount of data and classification is usually the more complex process compared to segmentation. Sommer *et al.* used pixel-wise classification for candidate retrieval and then object shape and texture features for mitotic cell classification [SFHG12]. Irshad used active contour models for candidate selection and statistical and morphological features for classification [Irs13]. Those hand-crafted features have significant drawbacks, however: Given the often small data sets, automatic selection of features is prone to random correlation, while using higher-dimensional classification approaches on the complete set increases overfitting [Lea96]. Further, it is questionable, if those approaches can represent the variability in shape and texture of mitotic figures [CDW*16].

Triggered by the ground-breaking initial works of Lecun [LBBH98], Convolutional Neural Networks (CNN) have spread widely in the use for various image classification tasks. CNN-based recognition algorithms have won all major image recognition challenges in recent years because of their ability to capture complex shapes and still remain sensitive to minor variations in the image. In the field of mitosis detection, CNN-based approaches have been used for classification [CGGS13], feature extraction [WCRB*14] as well as candidate generation [CDW*16]. Yet, CNNs, through their inherent ability to capture complex structures, are also prone to overfitting, a problem which is usually targeted by data augmentation and regularization strategies like dropout and other mechanisms or by means of transfer learning. Another regularization strategy is to constrain the capacity of the approach [Goo16] by reducing effectively the free parameters of the model. We aim to attempt this by splitting the problem into an attention task and a classification task. The general issue however, that the training data might be a non-representative sample of the classification task and thus parts of real-world data are not recognized because the data set does not generalize well, can be best targeted with a bigger training data set, as it was the base for this work.

3. Material

For this study, digital histopathological images were acquired using Aperio ScanScope (Leica Biosystems Imaging, Inc., USA) slide scanner hardware at a magnification of 400x. Candidate patches for three different cell types (mitotic cells, eosinophilic granulocytes and normal tumor cells) were annotated by an expert with profound knowledge on cell differentiation and classified by a trained pathologist. The cells were selected from histologic images of 12 different paraffin-embedded canine mast cell tumors, stained with standard hematoxylin and eosin (H&E). In order to train a deep learning classifier with a sufficient amount of data, our emphasis was not on complete annotation of the slides but on finding enough candidates for the above-mentioned cell types within the image.

More specifically, the emphasis was on finding mitotic cells. Commonly, in all major related works, the number of mitotic cells in the data set was proportional to the actual occurrence in the respective slides, as whole slides where annotated, resulting in a relatively low number of mitotic cells. On the contrary, we purposely selected a similar number of cells from each category to not assume any priors in distribution.

We acknowledge that this procedure might add a certain bias in cell selection, and that our dataset might not be representative. However, this argument can also be made for the case where only a small number of mitotic figures is available. Further, because of the high inter-rater variability in mitosis expert classifications, we assume that an unbiased ground truth is hard to retrieve and a minor bias by image pre-selection can be tolerated. Finally, we do not target at finding all mitotic cells, but rather to guide the pathologist in finding a representative part of the slide and to thus reduce variability in expert grading.

In the data set, we have approx. 37,800 single annotations of cells of the three different types (about 10,400 granulocytes, 10,800 mitotic figures and 16,600 normal tumor cells). The majority of cells was rated by the pathologist to be normal tumor cells, however also a significant amount of mitotic cells and eosinophil granulocytes was annotated.

4. Methods

Spatial Transformer Networks (STN), first described by Jaderberg et. al, provide a learnable method to focus the attention of a classification network on a specific subpart of the original image [JSZ*15]. To achieve this, parameters of an affine transformation matrix θ are regressed by the network, alongside with the optimization of the actual classifier.

Spatial Transformer Networks were originally successfully employed on a distorted MNIST [LBBH98] data set, where translation, scale, rotation and clutter were used to increase the difficulty for the detection task. The approach has shown to be able to – without any prior knowledge about the actual transform that was applied beforehand – increase accuracy of the classification network by focusing its attention to the area where the number was present and by

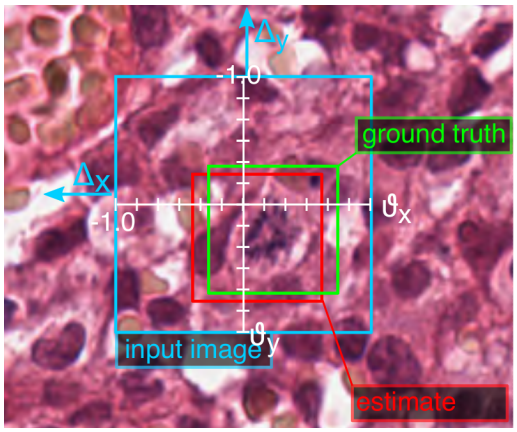

Figure 2: *Image preprocessing. The offset Δ_x, Δ_y is set randomly while keeping the cell within the image.*

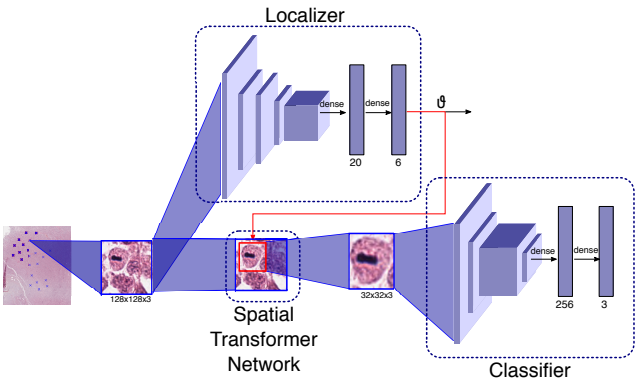

Figure 3: *Overview of the network. The cell's position is estimated by the localizer, which regresses an affine transform applied on the original image, to feed the classifier with cell images.*

compensating for the deformation [JSZ*15]. The approach can be used in a joined learning approach, where both the transform and the classification are learned end-to-end, something that could be described as a weakly supervised learning approach for the transformation. The optimization on the MNIST data set is, however, a much easier task than on real-world data. In a typical patch extracted from a histology slide, a lot of similar and valid objects may be contained in the image, and joined optimization suffers from local extrema in the gradient descend approach.

In this work, we aim to use STN as a method of not only directing the attention of a classification network to a sub-area of a larger image, and thus hopefully improving classification performance, but also as a segmentation approach to derive the information about where the respective cells are located.

We believe that Spatial Transformer Networks are an ideal candidate for this kind of task because they can be used to model two sources of natural variance into the machine learning process with a comparatively small overhead in complexity: Scaling and translation. Scaling is relevant in microscopy for two reasons: Firstly, the actual magnification of the microscope is dependent on the optical properties of the ocular, notably on the field number [MMG16]. Secondly, cells differ in size, dependent on their function and the species they originate from.

4.1. Image Preprocessing

All images were cropped around the cell center in a first processing step. In a second processing step, we introduce a random translation Δ_x, Δ_y to the origin area of the input image before cropping, so it is no longer centered around the cell, i.e. the cell can be anywhere on the image, with the restriction that the whole cell will be within the image (see figure 2). From the introduced translation, we can derive a new ground truth transformation vector

$$\theta = \begin{bmatrix} \vartheta_s & 0 & \vartheta_x \\ 0 & \vartheta_s & \vartheta_y \end{bmatrix} \tag{1}$$

where ϑ_s is the (in our case) fixed scaling vector. The scaling vector is dependent on the (manually chosen) expected cell size d_c. For our data, prior investigation has shown that all typical cells in our case are fully contained within an area of 64 px around the cell center, so with $d_i = 128\,px$ being the length of the input image, we can derive:

$$\vartheta_s = \frac{d_i}{d_c} = 0.5 \tag{2}$$

The (relative) coordinate grid for the STN is spaced from -1.0 to 1.0, with 0.0 being the center pixel. The translation elements of the ground truth transformation vector in eqn. 1 thus become:

$$\vartheta_{\{x,y\}} = -\frac{2\Delta_{\{x,y\}}}{d_i} \tag{3}$$

4.2. Network layout

Our network consists of three main blocks, as depicted in Figure 3: The localizer, the classifier and the Spatial Transformer Network. The localizer is a deep convolutional network with two stacked convolutional and max-pooling layers, one inception layer and two fully connected layers. It regresses an estimate $\widehat{\theta}$ of the transform matrix θ.

The classifier is a rather small convolutional neural network with 7 layers, using also convolutional, max-pooling and inception layers. It outputs a vector of dimension 3, which represents the class probabilities for the three cell types depicted in Figure 1.

Inception [SLJ*14] blocks were introduced by Szegedy *et al.* in 2014, and have been since then widely used in classification tasks. They are based on the idea that visual information should be processed at different scales, and described to be particularly useful for localization [SLJ*14]. We incorporated an inception layer, much like Szegedy, between the initial convolutional and max-pooling layers and the fully connected layers. In our case, the inception layer increased convergence and performance in both localizer and classifier.

4.3. Training

The network was trained with the TensorFlow framework using the Adam optimizer [KB14]. Each image was augmented with an arbitrarily rotated copy of itself to increase robustness of the system. To not assume priors for the cell types, the distributions for the training were made uniform by random deletion of non-minority classes within the training set. A five-fold cross-validation was used.

4.3.1. Classification network

In order to achieve good localization and classification performance, we propose a three stage process: In a first step, centered cell images are presented to the network, and the classification-part of the network is trained for 50 epochs using an initial learning rate of 10^{-3}. This serves as a good initialization of the network for later use. As loss function, denoting the (one hot coded) ground-truth cell class c and the estimated class probabilities $\widehat{c}$, standard cross-entropy is used:

$$l_{\text{cla}} = -\sum_{i=1}^{3} ln\left(\widehat{c_i}\right) \cdot c_i \tag{4}$$

4.3.2. Training of the localization network

In the next step, the localization network is trained. For this, the images were cropped with a random offset from the original image, as described in section 4.1. Knowledge of this random offset enables to define a ground truth transformation matrix θ for optimizing the network. This is used to regress the estimated transformation vector $\widehat{\theta}$ with its elements

$$\widehat{\theta} = \begin{bmatrix} \widehat{\vartheta_1} & \widehat{\vartheta_2} & \widehat{\vartheta_x} \\ \widehat{\vartheta_3} & \widehat{\vartheta_4} & \widehat{\vartheta_y} \end{bmatrix} \tag{5}$$

We want $\widehat{\theta}$ to be an affine transform with no skew and known scale ϑ_s. To achieve this, we first derive the scaling of the estimated transform as:

$$\widehat{\vartheta_{s_x}} = \sqrt{\widehat{\vartheta_1}^2 + \widehat{\vartheta_3}^2} \tag{6}$$

$$\widehat{\vartheta_{s_y}} = \sqrt{\widehat{\vartheta_2}^2 + \widehat{\vartheta_4}^2} \tag{7}$$

Further, we want the diagonal elements $\widehat{\vartheta_1}$ and $\widehat{\vartheta_4}$ to be equal and the off-diagonal elements $\widehat{\vartheta_2}$ and $\widehat{\vartheta_3}$ equal with opposite sign, resulting in a rotation matrix with scale. These constraints compile into the loss for the localization network:

$$\begin{aligned} l_{\text{loc}} = &\left|\widehat{\vartheta_x} - \vartheta_x\right|^2 + \left|\widehat{\vartheta_y} - \vartheta_y\right|^2 + \left|\widehat{\vartheta_{s_x}} - \vartheta_s\right|^2 + \\ &\left|\widehat{\vartheta_{s_y}} - \vartheta_s\right|^2 + \left|\widehat{\vartheta_1} - \widehat{\vartheta_4}\right|^2 + \left|\widehat{\vartheta_2} + \widehat{\vartheta_3}\right|^2 \end{aligned} \tag{8}$$

The rotation angle of the transform is a degree of freedom and thus not covered by the loss. The localization part of the network is trained for 200 epochs using an initial learning rate of 10^{-4}.

4.3.3. Final refinement of the classification network

Finally, the whole network is trained for 100 epochs, using an initial learning rate of 10^{-4}. This final step is calculated on the translated images that were estimated by the localization network and the STN, and it is using a combined loss:

$$l = l_{\text{loc}} + \kappa \cdot l_{\text{cla}} \tag{9}$$

This loss thus incorporates knowledge about the proper class of the image, about the position of the cell within the image and about the scaling of the patch representing the cell, yet the rotation angle is not known.

4.4. Baseline comparison

It is hard to compare our results to other authors' works, because unlike them, we consider different cell types within the image and our data set is sparse and not fully annotated. For a baseline comparison, we took a 12-layer CNN like the one described by Cireşan *et al.* [CGGS13] for Mitosis detection, but aimed at a three class problem and with an input size of 128x128 px. This classification network was trained for 200 epochs using an initial learning rate of 10^{-3}.

5. Results and Discussion

There were only minor differences in the results of the individual test sets in cross-validation, which is why we concatenated the respective test vectors and calculated the following metrics on the ensemble. We achieved an accuracy of 91.8 %, with precisions reaching from 90.4 % to 93.4 % and recall reaching from 90.1 % to 92.8 %, as described in table 1. Compared to the baseline CNN described in section 4.4, this is a significant increase, with the added benefit of retrieving also segmentation information.

Regarding misclassifications, it is noteworthy that for many false decisions the root cause of error seems to be within the scope of the localizer (see right column of figure 4). In the top and bottom examples depicted there, the localizer selected a different cell than the one originally annotated. Particularly for tumor cells, this is not always a definite fault, since we do not consider annotation information of the direct environment of the annotated cell. If, in a direct surrounding of a tumor cell, a granulocyte or mitotic cell is present, the localizer in fact behaves completely correct in presenting this cell to the classifier. Since we do not aim at finding or classifying all cells, this is no major drawback. In fact, we inherently prioritize classification this way: Since we crop around a known sparse event (mitotic cells or granulocyte), and give this label to our classifier, we incorporate the knowledge that sparse events are more important than others into the loss function.

We think that the acquired data set provides a good fundament for further approaches in mitosis detection, where in our opinion the lack of a sufficient amount of samples may limit the methodic progress.

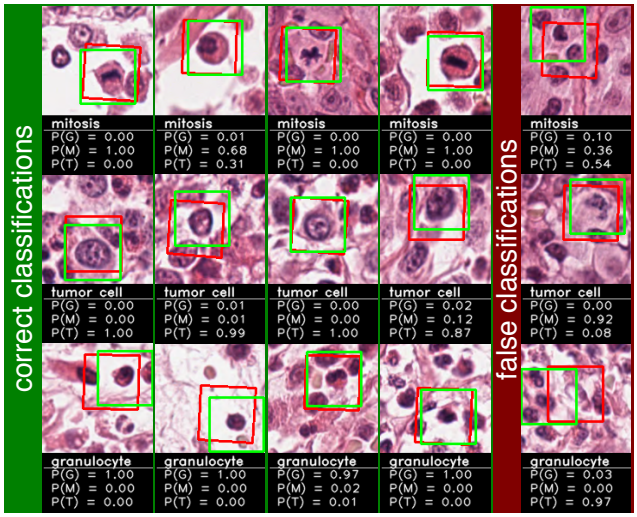

Figure 4: *Random choice of correct and false image classifications alongside with selected focus areas, as picked by the localizer. The probabilities denoted are those for the classes: P(G)=granulocytes, P(M)=mitosis, P(T)=normal tumor cells*

approach	name	precision	recall	f1-score
CNN baseline	granulocytes	0.847	0.898	0.872
	mitotic figures	0.822	0.853	0.837
	normal t. cells	0.916	0.859	0.887
	avg / total	**0.870**	**0.868**	**0.869**
CNN-STN	granulocytes	0.912	0.925	0.918
	mitotic figures	0.891	0.889	0.890
	normal t. cells	0.932	0.924	0.928
	avg / total	**0.915**	**0.915**	**0.915**

Table 1: *Overall classification results of the proposed network.*

The acquired data set is also a very interesting candidate for transfer learning. Assuming that many known CNN-approaches suffer from networks that partially do not have well defined filters due to lack of training data, in-domain transfer learning from our mitosis data to other, fully labeled data sets like the competition data sets should be beneficial.

6. Summary

In this work the potential of Spatial Transformer Networks within a convolutional neural network approach, applied to segmentation and classification tasks in digital histology images, has been shown.

The presented approach focuses the attention of a classification network to a part of the original image where most likely a sparsely distributed cell type (mitosis or granulocyte) can be found.

Further, we have acquired and introduced a data set of cell images from different classes of H&E stained histology images, with at least ten thousand pathologist-rated samples per class.

Modeling the localization and classification process independently but with a joint training cuts down on computational complexity of the overall system. We believe that this work is an important step towards a microscope-embeddable algorithm that can help the pathologist in counting of mitotic figures by finding a representative area within a histology slide, an algorithm which could reduce inter-rater-variability and thus improve overall quality of tumor grading systems.

References

[CDW*16] CHEN H., DOU Q., WANG X., QIN J., HENG P. A.: Mitosis Detection in Breast Cancer Histology Images via Deep Cascaded Networks. In *13th AAAI Conference on Artificial Intelligence* (2016). 2

[CGGS13] CIREŞAN D. C., GIUSTI A., GAMBARDELLA L. M., SCHMIDHUBER J.: Mitosis Detection in Breast Cancer Histology Images with Deep Neural Networks. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) 16*, Pt 2 (2013), 411–418. 1, 2, 4

[Goo16] GOODFELLOW I.: *Deep Learning*. The MIT Press, Cambridge, Massachusetts, 2016. 2

[Irs13] IRSHAD H.: Automated Mitosis Detection in Histopathology using Morphological and Multi-channel Statistics Features. *Journal of Pathology Informatics 4*, 1 (2013), 10. 1, 2

[JSZ*15] JADERBERG M., SIMONYAN K., ZISSERMAN A., ET AL.: Spatial transformer networks. In *Advances in Neural Information Processing Systems* (2015), pp. 2017–2025. 2, 3

[KB14] KINGMA, D, BA, J: Adam: A method for stochastic optimization. *arXiv.org* (2014). arXiv:1401.4983v4. 4

[LBBH98] LECUN Y., BOTTOU L., BENGIO Y., HAFFNER P.: Gradient-based Learning Applied to Document Recognition. In *Proceedings of the IEEE* (November 1998), vol. 86, pp. 2278–2324. 2

[Lea96] LEARDI R.: Genetic Algorithms in Feature Selection. In *Genetic Algorithms in Molecular Modeling*. Elsevier, 1996, pp. 67–86. 2

[MMG16] MEUTEN D. J., MOORE F. M., GEORGE J. W.: Mitotic Count and the Field of View Area. *Veterinary Pathology 53*, 1 (Jan. 2016), 7–9. 1, 3

[RRL*13] ROUX L., RACOCEANU D., LOMÉNIE N., KULIKOVA M., IRSHAD H., KLOSSA J., CAPRON F., GENESTIE C., LE NAOUR G., GURCAN M. N.: Mitosis Detection in Breast Cancer Histological Images - An ICPR 2012 Contest. *Journal of Pathology Informatics 4* (2013), 8. 1

[SFHG12] SOMMER C., FIASCHI L., HAMPRECHT F. A., GERLICH D. W.: Learning-based Mitotic Cell Detection in Histopathological Images. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)* (2012), IEEE, pp. 2306–2309. 1, 2

[SLJ*14] SZEGEDY C., LIU W., JIA Y., SERMANET P., REED S., ANGUELOV D., ERHAN D., VANHOUCKE V., RABINOVICH A.: Going Deeper with Convolutions. *arXiv.org* (Sept. 2014). arXiv:1409.4842v1. 3

[VvDW*14] VETA M., VAN DIEST P. J., WILLEMS S. M., WANG H., MADABHUSHI A., CRUZ-ROA A., GONZALEZ F., LARSEN A. B. L., VESTERGAARD J. S., DAHL A. B., SCHMIDHUBER J., GIUSTI A., GAMBARDELLA L. M., TEK F. B., WALTER T., WANG C.-W., KONDO S., MATUSZEWSKI B. J., PRECIOSO F., SNELL V., KITTLER J., DE CAMPOS T. E., KHAN A. M., RAJPOOT N. M., ARKOUMANI E., LACLE M. M., VIERGEVER M. A., PLUIM J. P. W.: Assessment of Algorithms for Mitosis Detection in Breast Cancer Histopathology Images. *arXiv.org*, 1 (Nov. 2014), 237–248. arXiv:1411.5825v1. 1

[WCRB*14] WANG H., CRUZ-ROA A., BASAVANHALLY A., GILMORE H., SHIH N., FELDMAN M., TOMASZEWSKI J., GONZALEZ F., MADABHUSHI A.: Mitosis Detection in Breast Cancer Pathology Images by Combining Handcrafted and Convolutional Neural Network Features. *Journal of Medical Imaging 1*, 3 (Oct. 2014), 034003. 2