CROWD SOURCING IMAGE SEGMENTATION WITH iaSTAPLE

Dmitrij Schlesinger[⋆†] *Florian Jug*[⋆‡] *Gene Myers*[‡] *Carsten Rother*[†] *Dagmar Kainmüller*[‡]

[†] Computer Vision Lab Dresden, Dresden University of Technology
[‡] Max Planck Institute of Molecular Cell Biology and Genetics
[⋆] authors contributed equally

ABSTRACT

We propose a novel label fusion technique as well as a crowdsourcing protocol to efficiently obtain accurate epithelial cell segmentations from non-expert crowd workers. Our label fusion technique simultaneously estimates the true segmentation, the performance levels of individual crowd workers, and an image segmentation model in the form of a pairwise Markov random field. We term our approach image-aware STAPLE (iaSTAPLE) since our image segmentation model seamlessly integrates into the well-known and widely used STAPLE approach. In an evaluation on a light microscopy dataset containing more than 5000 membrane labeled epithelial cells of a fly wing, we show that iaSTAPLE outperforms STAPLE in terms of segmentation accuracy as well as in terms of the accuracy of estimated crowd worker performance levels, and is able to correctly segment 99% of all cells when compared to expert segmentations. These results show that iaSTAPLE is a highly useful tool for crowd sourcing image segmentation.

***Index Terms*—** Epithelial cell segmentation, Crowdsourcing, Markovian Random Fields, iaSTAPLE

I. INTRODUCTION

Cell segmentation is an important and ubiquitous step in the process of scientific discovery in biology and in clinical applications. Ground truth (GT) segmentations of cells in microscopic images are not only needed for biomedical analyses, but also for training and evaluating automated segmentation methods and anatomy models. Due to rapidly evolving microscopes, fluorescent dyes, and markers, there is no such thing as a “default” appearance of cells in microscopic images. It is therefore common practice to manually generate GT segmentations for each new task to be solved. Since this is a tedious and time consuming task, the amount and quality of available GT is often rather low.

We propose an alternative solution for GT generation: Instead of asking a few (expensive) experts to generate GT, we aim at crowd-sourcing this task to many non-expert workers.

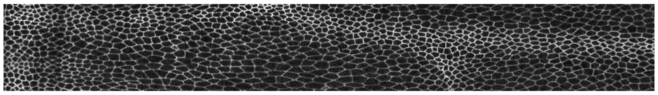


Fig. 1. Microscopic image of a developing fly wing (crop).

Our technique extends the popular STAPLE approach [1] by simultaneously estimating not only the true segmentation and user performance levels, but jointly also learning an image segmentation model in the form of a pairwise Markovian random field (MRF, cf. Section III). We call our new method “iaSTAPLE”, for **i**mage-**a**ware STAPLE. Our method yields “more informed” estimates than STAPLE by means of the additional image model, and it enables the segmentation of image regions that were not processed by crowd workers at all. This leads to significantly improved overall performance, as we demonstrate in a quantitative evaluation.

Landman et al. [2] showed that STAPLE offers a straightforward way to deal with partial as well as multiple segmentations by the same user. Their approach, termed STAPLER, thus reveals an important feature of STAPLE not shown in [1]. In the remainder of this document we will, for simplicity, use the abbreviation STAPLE to refer to both, the original STAPLE [1] and STAPLER [2].

The LOP STAPLE approach [3] employs an MRF-based segmentation model. However, as opposed to our proposed method, LOP STAPLE uses a hand-crafted, un-trained MRF model and relies on approximate mean field based inference. Non-local STAPLE [4] shares our idea of making use of image intensities. However, intensities are not exploited in the form of an appearance model for segmentation as we propose, but in a different context, namely to improve the registration of atlases. The approach closest to our own work is probably iSTAPLE [5], where image intensities are used as an appearance model for segmentation. However, iSTAPLE employs a pixel-independent segmentation model as well as a simple Gaussian probability distribution of gray-values as appearance model. Our method, in contrast, uses a learned pairwise MRF for segmentations and a Gaussian Mixture model with shading to model appearances.

Our main contributions are: (i) A novel crowd-sourcing workflow for efficiently segmenting all cells in membrane labeled epithelia (cf. Section II). (ii) A generic label fusion technique that can be used in the context of many image segmentation problems. We demonstrate that this is a viable

We are grateful to the Pallas Ludens GmbH and especially to D. Kondermann for providing the crowd sourcing platform. Financial support by the BMBF for the competence center for Big Data ScaDS and the 031A099 project is gratefully acknowledged. We also thank the European Research Council (ERC, grant agreement No 647769) for their support. Computations were performed at the ZIH at TU Dresden.

approach for the exemplary application of segmenting membrane labeled epithelial cells in fluorescence microscopy images (cf. Figure 1). Furthermore, we show quantitatively that iaSTAPLE's user performance estimates are more accurate than those of STAPLE/STAPLER. This finding suggests that our user performance estimates may serve as valuable input for automated quality control in crowd-sourced image segmentation, complementing supplementary information like click behavior and performance on test tasks (see e.g. [6]).

II. CROWDSOURCING EPITHELIAL CELL SEGMENTATION

Microscopy images of cell epithelia often contain thousands of cells. Epithelial cells are densely packed with no space in-between. Hence, cell membranes appear as a honeycomb-like polygonal mesh that covers the whole image. Cell size varies considerably, with a scale factor of 10 not being unusual. Figure 1 shows part of a fly wing. The cell membranes appear bright due to a fluorescent membrane marker that is genetically introduced into the organism.

Algorithm 1 Crowdsourcing Epithelial Cell Segmentation.

```
Input: Set $T$ of image tiles $t$, outer loop iterations $K$.
Output: $P$, a set of closed polygons $p$ describing segmented cell outlines.
$P = \emptyset$;
for $k = 1 \dots K$ do
    $T_k = T$                          ▷ The set of image tiles not fully covered
                                         at the current iteration of the outer loop.
    for $t \in T_k$ do $P_t = \emptyset$   ▷ The current set of polygons for each tile.
    repeat
        for $t \in T_k$ do $P_t = \text{CrowdAnnotate}(t, P_t)$
        $T_k = \text{GetNotCoveredTiles}(T, \bigcup_{t \in T} P_t)$
    until $T_k = \emptyset$ or $T_k$ did not change.
    $P = P \cup \bigcup_{t \in T} P_t$

function CrowdAnnotate($t, P_t$)
    if more than 20 visible cells in $t$ are not yet in $P_t$ then
        Outline 20 fully visible cells that are not yet in $P_t$.
    else Outline all fully visible cells that are not yet in $P_t$.
    return $P_t \cup$ newly drawn polygons.

function GetNotCoveredTiles($T, P$)
    Mark areas covered by polygons $P$ as foreground in binary image $c$.
    Dilate foreground of $c$ by known membrane width.
    $T_{nc} = \emptyset$          ▷ The set of not fully covered tiles to be returned.
    for $t \in T$ do
        if there are more background pixels in $t$ than a threshold then
            $T_{nc} = T_{nc} \cup t$
    return $T_{nc}$
```

An individual epithelial cell can be segmented easily by drawing its outline as a closed polygon. We use Amazon Mechanical Turk as crowd sourcing platform, where workers are paid by completed task and not by the hour. Hence all tasks given to the crowd should, by design, be similarly time consuming. To satisfy this criterion in the face of highly varying cell sizes, we define a task to be to outline a fixed number of cells in a given image tile of comfortably visualizable size. The respective instruction as given to crowd workers can be found at `bioimagecomputing.com/crowd-instructions`.

In Algorithm 1 we describe a workflow for yielding (almost) complete coverage of the whole image by means of these tasks performed on overlapping image tiles of fixed size. The inner **repeat**-loop aims at covering the whole image once, while the outer loop can be used to obtain multiple coverages.

III. iaSTAPLE

Model. Let x be an observation (image) and y be a hidden variable, i.e. the binary segmentation in our case. Formally, the labeling is a mapping $y : R \to L$, $L = \{0, 1\}$ (called label set), that assigns 0 (background) or 1 (foreground) to each pixel $i \in R$. By $y_i \in L$, we will denote the label chosen in the pixel i. Analogously, the image $x : R \to C$ is a mapping assigning a gray value $c \in C$ to each pixel. The gray value of pixel i is denoted by x_i. Let us denote user inputs by $z^u : R \to L$, where u indicates a particular user, so if we have m users, $u = 1 \dots m$. Our key idea is to consider user inputs as "additional observations" – the u-th user "observes" the true scene y and gives his/her opinion about the label $l \in L$ of each pixel $i \in R$. It is also reasonable to assume that users are conditionally independent from each other given a labeling y. To summarize, the joint probability distribution for all model constituents is[1]

$$p(x, y, z^1, \dots z^m; \theta_p, \theta_a, p_u, u{=}1 \dots m) = \\ = p(y; \theta_p) \cdot p(x|y; \theta_a) \cdot \prod_{u=1}^{m} p_u(z^u|y), \quad (1)$$

with θ_p the parameters of the prior probability distribution of labelings $p(y; \theta_p)$, and θ_a the parameter of the gray value model $p(x|y; \theta_a)$. Each user is characterized by his own, initially unknown probability distribution $p_u(z^u|y)$ reflecting his/her *reliability*. Let us consider all model parts in more detail.

Our prior model $p(y; \theta_p)$ is a Markovian Random Field (MRF) over the graph $G = (R, E)$, where the node set R corresponds do the pixel grid and the edge set E reflecting pixel neighborhoods. The associated energy is

$$E(y, \theta_p) = \sum_{i \in R} \psi_0(y_i) + \sum_{c} \sum_{ij \in E_c} \psi_c(y_i, y_j). \quad (2)$$

Unary potentials $\psi_0 : L \to \mathbb{R}$ assign values to each label $l \in L$ of each pixel $i \in R$. Pairwise potentials are defined in a densely connected neighborhood, where edges E are partitioned into *edge classes* E_c. Classes are characterized by the translational vector that connects the corresponding pixels (see e.g. [7], [8]). Pairwise potentials $\psi_c : L \times L \to \mathbb{R}$ are class specific and shared by all edges of the same class. Hence the free model parameters θ_p are the unary potentials ψ_0, and the pairwise potentials ψ_c for all c, and the prior probability distribution is

$$p(y; \theta_p) = \frac{1}{Z(\theta_p)} \exp\bigl[-E(y, \theta_p)\bigr], \\ \text{with } \; Z(\theta_p) = \sum_{y} \exp\bigl[-E(y, \theta_p)\bigr]. \quad (3)$$

[1] Parameters are separated from random variables by a semicolon.

Like STAPLE [1], we assume conditional independence of user inputs at each pixel, given a labeling y. So $p_u(z^u|y) = \prod_i p_u(z_i^u|y_i)$, with z_i^u denoting the label given by user u in the pixel i. Hence, the unknown parameters of user models are user specific conditional probability distributions $p_u : L \times L \rightarrow \mathbb{R}$, with $p_u(l'|l) \geq 0$ and $\sum_{l'} p_u(l'|l) = 1$ for all $l \in L$. The key difference between our approach and STAPLE is that our model is aware of the input image. At each pixel we use a Gaussian Mixture to model the gray value distribution, with label specific weights. In addition, we use a so-called shading field [9] that allows smoothly varying deviations of the appearance model. Formally, the shading is a mapping $s : R \rightarrow \mathbb{C}$, like the image itself, yet it is required to be spatially smooth. Given a shading field s and a labeling y, the appearance model is

$$p(x|y,s) = \prod_i \sum_j w_{y_i j} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{\big((x_i - s_i) - \mu_j\big)^2}{2\sigma^2} \right], \tag{4}$$

with the prior probability distribution for shading being a Gaussian MRF, μ_j being the centers of Gaussians, a common σ to all Gaussians, $w_{l,j}$ being label specific weights, and the initially unknown shading s.

Learning. We formulate the learning problem according to the Maximum Likelihood principle and describe here the learning of user models p_u. All other model components are learned as in [7], [9]. Given an image x and user annotations z^u, $u = 1 \ldots m$, the task is to maximize the logarithm of their joint probability, which is obtained by marginalization over hidden y:

$$\ln p(x, z^1, \ldots, z^m; \theta_p, \theta_a, p_u, u{=}1 \ldots m) =$$

$$= \ln \sum_y \left[\frac{1}{Z(\theta_p)} \exp\big[-E(x, y, \theta_p, \theta_a)\big] \cdot \right. \tag{5}$$

$$\left. \cdot \prod_u \prod_i p_u(z_i^u|y_i) \right] \rightarrow \max_{p_u, u=1\ldots m}, \tag{6}$$

where $E(x, y, \theta_p, \theta_a)$ summarizes all energy terms that are fixed for the moment, i.e. the prior energy (2) and logarithms of (4). We use Expectation Maximization as follows: (i) In the E-step the marginal posterior label probabilities

$$p(y_i{=}l|x, z_1, \ldots, z_m) = \sum_{y:y_i=l} p(y|x, z_1, \ldots, z_m) \tag{7}$$

for each pixel and each label are computed, followed by (ii), the M-step, where the following optimization problem is solved:

$$\sum_i \sum_l \sum_u p(y_i{=}l|x, z^1, \ldots, z^m) \cdot \ln p_u(z_i^u|l) \rightarrow \max_{p_u, u=1\ldots m}. \tag{8}$$

Let $I_{ul'}$ be the set of pixels that user u assigned label l'. Probabilities $p_u(l'|l)$ solving (8) are proportional to the *frequencies* $p_u(l'|l) \propto n_u(l', l)$, where

$$n_u(l', l) = \sum_{i \in I_{ul'}} p(y_i{=}l|x, z^1, \ldots, z^m). \tag{9}$$

Since the computation of marginal probabilities in (7) is intractable for general MRFs, we use Gibbs Sampling: We sample labelings $\hat{y}$ according to the posterior $p(y|x, z^1, \ldots, z^m)$, and count how many times which label was generated at which pixel. Doing so, we observe the following: Usually, it is necessary to have many independent samples $\hat{y}$ drawn from the target probability distribution $p(y|x, z^1, \ldots, z^m)$ in order to estimate the frequencies of (9). Our frequencies, however, are sufficient statistics of low order, i.e. for a particular user we need to estimate just two values $p_u(0|1)$ and $p_u(1|0)$. Hence, in order to accelerate the overall learning procedure we use *warm starts*, meaning that we use the sample obtained at the previous iteration of the EM-algorithm as the initialization and perform just one Gibbs Sampling iteration. Similar learning procedures are for example known as Persistent Contrastive Divergence [10]. The learning algorithm is summarized in Algorithm 2.

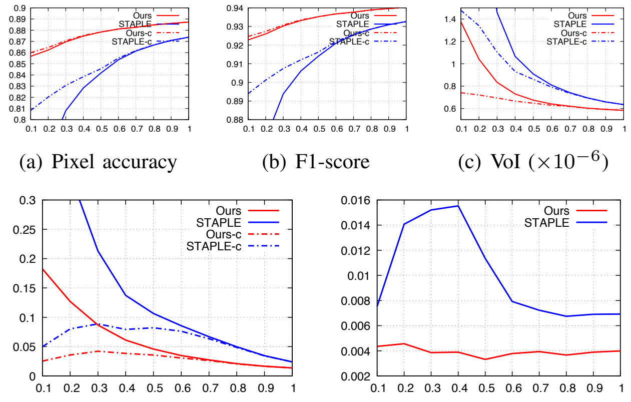
Algorithm 2 Learning the User Models.

```
Input: Image $x$, user inputs $z^u$, prior model parameters $\theta_p = (\psi_0, \psi_c)$,
       appearance model $\theta_a = (w, \mu, \sigma)$, initial user models $p_u^{(0)}$, initial
       random labeling $\hat{y}^{(0)}$.
Output: Learned user models $p_u$.
for $t := 0 \ldots T$ do
    Perform one iteration of Gibbs Sampling to draw a labeling
    from the posterior probability distribution $p(y|x, z^1, \ldots, z^m)$,
    starting from the labeling $\hat{y}^{(t)}$, using the current model parameters
    $(\theta_p, \theta_a, p_u^{(t)}, u = 1 \ldots m) \rightarrow$ obtain the new labeling $\hat{y}^{(t+1)}$.
    for all users $u$ do
        Compute frequencies $n_u(l', l) = \#(z_i^u{=}l', \hat{y}_i^{(t+1)}{=}l)$, $\forall l', l$.
        Perform the step toward the optimal value with a step size $\lambda$:
    $p_u^{(t+1)}(l'|l) = (1 - \lambda) \cdot p_u^{(t)}(l'|l) + \lambda \cdot \frac{n_u(l', l)}{\sum_{l''} n_u(l'', l)}, \quad \forall l', l$
```

Inference. We compute the Maximum Posterior Marginal decision (MPM) by computing the node-wise posterior marginal probabilities, followed by choosing the labelling which has the best marginal at each node. We do this, again, by using Gibbs Sampling.

IV. RESULTS AND DISCUSSION

Segmentation Evaluation. To evaluate iaSTAPLE, we randomly pick certain fractions of crowd annotations, ranging from 10% to 100% of 1584 annotations obtained for a 5000-cell image of a developing fly wing. Given such a fraction, we train our model and estimate the true segmentation. We compare results to a manually prepared ground truth (GT) according to three standard error measures, namely (i) pixel accuracy, (ii) F1-score, and (iii) variation of information (VoI). To compute the VoI, we partition the image by assessing connected components of the "interior" label and assigning all "membrane" pixels to the closest connected component. We repeat each experiment 10 times per given fraction and report respective average errors. We perform the same experiments for STAPLE/STAPLER. Fig. 2 (top row) shows resulting errors as a function of the respective fraction of annotations. Since STAPLE/STAPLER cannot handle regions that are not annotated by at least one user, we also evaluate all results exclusively on image areas that are covered by at least one annotation (see “STAPLE-c” and “Ours-c” in Fig. 2).



(a) Pixel accuracy (b) F1-score (c) VoI ($\times 10^{-6}$)

(d) under-segmentation (e) over-segmentation

Fig. 2. Comparison of the estimated segmentations. The suffix “-c” indicates that we evaluated the quality only on regions that are covered by at least one user annotation.

To complement the standard error measures by easily interpretable yet non-standard measures, we also assess *under-* and *over-segmentations*. An under-segmentation occurs if two neighboring GT cells are merged by mistake, an over-segmentation if a GT cell is split into two by mistake. We count under- and over-segmentations as follows: We first compute all connected background components (cells) from results and from GT. Then we perform a Hungarian Matching between the two sets of cells, where cells are allowed to be matched only if they are reasonably close. Unmatched cells in the segmentation results count as over-segmentations, while unmatched cells in the GT count as under-segmentations. In Figure. 2 (bottom row) we show percentages of under- and over-segmented GT cells as a function of the fraction of crowd annotations used in the experiment.

iaSTAPLE outperforms STAPLE/STAPLER in all experiments and all error measures, for partial image coverage as well as for full coverage. For most error measures, iaSTAPLE's segmentation accuracy surpasses STAPLE/STAPLER already when using less than 50% of the respective amount of crowd user inputs. In case of 100% crowd annotations, the total percentage of wrongly segmented GT cells for iaSTAPLE is only about 1%. We hypothesize that this accuracy is sufficient for iaSTAPLE's results to serve as ground truth for many practical applications. However, future work has to assess the respective inter-expert variability, or alternatively the impact of expert-GT vs crowd-GT on application-specific measures, like e.g. biological quantities of interest, or performance of GT-trained segmentation algorithms (cf. [11]).

User Rating. iaSTAPLE, like STAPLE/STAPLER, estimates performance measures for each user and label, $p_u(z_i^u{=}\text{membrane}|y_i{=}\text{cell})$ and $p_u(z_i^u{=}\text{cell}|y_i{=}\text{membrane})$.

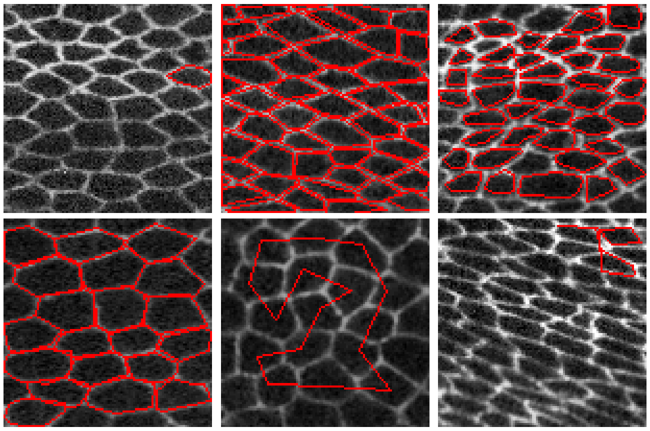


Fig. 3. Exemplary user annotations (red polygons overlaid with original images). Bottom row, left to right: Overall best, a very bad, and overall worst annotation.

To capture the quality of users with a single number we propose to use the user specific pixel accuracy, i.e. the number of correctly labeled pixels divided by the number of all pixels a user has annotated. To give an impression of the effectiveness of this measure, Fig. 3 shows the annotations by the respective best user (bottom left) as well as worst user (bottom right).

We evaluate estimated user ratings as follows: Let r_u be the rank of user u among all users in terms of the fraction of pixels where u agrees with the segmentation estimate yielded by STAPLE or iaSTAPLE. Let r_u^* be the rank of user u in terms of the fraction of pixels where u agrees with the ground truth segmentation. Thus, r_u represents an estimated user ranking yielded by STAPLE or iaSTAPLE, while r_u^* represents the respective true user ranking. The average absolute difference between estimated and true user ranking can be interpreted as the *quality* of the respective estimated user ranking, where lower (i.e. more similar to true ranking) is better; formally, $1/m \sum_{i=1}^{m} |r_u - r_u^*|$. Considering the whole set of user annotations (179 users), iaSTAPLE achieves a quality value of 8.63, while STAPLE achieves a value of 10.06. Hence in these terms iaSTAPLE estimates a more accurate user ranking than STAPLE. This result suggests that iaSTAPLE's user rating estimates are more useful than STAPLE's for improving automated quality control in crowd-sourced image segmentation.

V. CONCLUSION

We proposed iaSTAPLE, a novel image-aware method for label fusion, as well as an efficient crowd-sourcing workflow for epithelial cell segmentation. We applied this workflow on microscopy images of membrane labeled fly wing epithelia, and compared the results of iaSTAPLE against the well-known STAPLE. We showed that iaSTAPLE outperforms STAPLE/STAPLER both in terms of accuracy of the estimated segmentation, as well as in the quality of the estimated user ranking. These results make iaSTAPLE the method of our choice for crowd-sourced image segmentation.

VI. REFERENCES

[1] S Warfield, K Zou, and W Wells, "Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation," *Medical Imaging, IEEE Transactions on*, vol. 23, no. 7, pp. 903–921, 2004.

[2] B Landman, A Asman, A Scoggins, J Bogovic, F Xing, and J Prince, "Robust Statistical Fusion of Image Labels," *Medical Imaging, IEEE Transactions on*, vol. 31, no. 2, pp. 512–522, 2012.

[3] Alireza Akhondi Asl, Lennox Hoyte, Mark E. Lockhart, and Simon K. Warfield, "A logarithmic opinion pool based STAPLE algorithm for the fusion of segmentations with associated reliability weights," *IEEE Trans. Med. Imaging*, vol. 33, no. 10, pp. 1997–2009, 2014.

[4] Andrew J. Asman and Bennett A. Landman, "Non-local STAPLE: an intensity-driven multi-atlas rater model," in *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2012 - 15th International Conference, Nice, France, October 1-5, 2012, Proceedings, Part III*, 2012, pp. 426–434.

[5] Xiaofeng Liu, Albert Montillo, Ek T. Tan, and John F. Schenck, "istaple: improved label fusion for segmentation by combining STAPLE with image intensity," in *Medical Imaging 2013: Image Processing, Lake Buena Vista (Orlando Area), Florida, USA, February 10-12, 2013*, 2013, pp. 866–92O.

[6] Mehrnoosh Sameki, Danna Gurari, and Margrit Betke, "Predicting quality of crowdsourced image segmentations from crowd behavior," in *Third AAAI Conference on Human Computation and Crowdsourcing*, 2015.

[7] B. Flach and D. Schlesinger, "Modelling composite shapes by gibbs random fields," in *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2011.

[8] S. Nowozin, C. Rother, S. Bagon, T. Sharp, B. Yao, and P. Kohli, "Decision tree fields," in *Proc. of the International Conference of Computer Vision (ICCV)*, 2011.

[9] Dmitrij Schlesinger and Boris Flach, "A probabilistic segmentation scheme," in *Pattern Recognition*, Gerhard Rigoll, Ed. 2008, vol. 5096 of *LNCS*, pp. 183–192, Springer.

[10] T. Tieleman, "Training restricted boltzmann machines using approximations to the likelihood gradient," in *ICML*. ACM, 2008.

[11] Lena Maier-Hein, Sven Mersmann, Daniel Kondermann, Sebastian Bodenstedt, Alexandro Sanchez, Christian Stock, Hannes Kenngott, Mathias Eisenmann, and Stefanie Speidel, "Can masses of non-experts train highly accurate image classifiers?," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 438–445. Springer, 2014.