Title

Saliency-driven system models for cell analysis with deep learning

Permalink

https://escholarship.org/uc/item/3vv6b76k

Authors

Ferreira, Daniel S
Ramalho, Geraldo LB
Torres, Débora
et al.

Publication Date

2019-12-01

DOI

10.1016/j.cmpb.2019.105053

Peer reviewed

Saliency-driven system models for cell analysis with deep learning[*]

Geraldo L. B. Ramalho, Débora Torres, Alessandra H. G. Tobias, Mariana T. Rezende, Fátima N. S. Medeiros, Andrea G. C. Bianchi, Cláudia M. Carneiro, Daniela M. Ushizima

Abstract

Background and objectives: Saliency refers to the visual perception quality that makes objects in a scene to stand out from others and attract attention. While computational saliency models can simulate the expert's visual attention, there is little evidence about how these models perform when used to predict the cytopathologist's eye fixations. Saliency models may be the key to instrumenting fast object detection on large Pap smear slides under real noisy conditions, artifacts, and cell occlusions. This paper describes how our computational schemes retrieve regions of interest (ROI) of clinical relevance using visual attention models. We also compare the performance of different computed saliency models as part of cell screening tasks, aiming to design a computer-aided diagnosis systems that supports cytopathologists.

Method: We record eye fixation maps from cytopathologists at work, and compare with 13 different saliency prediction algorithms, including deep learning. We develop cell-specific convolutional neural networks (CNN) to investigate the impact of bottom-up and top-down factors on saliency prediction from real routine exams. By combining the eye tracking data from pathologists with computed saliency models, we assess algorithms reliability in identifying clinically relevant cells.

Results: The proposed cell-specific CNN model outperforms all other saliency prediction methods, particularly regarding the number of false positives. Our algorithm also detects the most clinically relevant cells, which are among the three top salient regions, with accuracy above 98% for all diseases, except carcinoma (87%). Bottom-up methods performed satisfactorily, with saliency maps that enabled ROI detection above 75% for carcinoma and 86% for other pathologies.

Conclusions: ROIs extraction using our saliency prediction methods enabled ranking the most relevant clinical areas within the image, a viable data reduction strategy to guide automatic analyses of Pap smear slides. Top-down factors for saliency prediction on cell images increases the accuracy of the estimated maps while bottom-up algorithms proved to be useful for predicting the cytopathologist's eye fixations depending on parameters, such as the num-

[*] Accepted to Journal: Computer Methods and Programs in Biomedicine, https://doi.org/10.1016/j.cmpb.2019.105053

ber of false positive and negative. Our contributions are: comparison among 13 state-of-the-art saliency models to cytopathologists' visual attention and deliver a method that the associate the most conspicuous regions to clinically relevant cells.

Keywords: saliency prediction, convolutional neural network, cell analysis, eye tracking experiment.

1 Introduction

Visual attention consists of a set of cognitive processes that enables focusing on a region or object while ignoring irrelevant stimuli from the environment, which allows humans and other animals to extract relevant information from complex input scenes [Carrasco, 2011]. The attention arises from both stimuli-driven factors (bottom-up attention) and task-driven factors (top-down attention) [Zhang and Lin, 2013]. Bottom-up attention is fast, involuntary and guided by visual distinctness or rarity using low-level image information such as orientation, color, intensity and texture. Top-down attention is slower, voluntary and based on a task or an intention, being strongly influenced by the prior knowledge and experience of the observer [Yarbus, 1967, Polatsek et al., 2018].

Computational systems can model the visual attention as a saliency map, which is a topographical map representing the conspicuity of each pixel in an image. Researchers have proposed several algorithms and applied them to fields such as computer vision, robotics, and medical image analysis [Murabito et al., 2018, Loukas et al., 2018, Nguyen et al., 2018]. Bottom-up methods usually model the image low-level features, for example, color, contrast, orientation, and others, and may use different approaches, e.g., cognitive concepts, probabilistic frameworks, spectral analysis, etc., to generate saliency maps [Borji and Itti, 2013]. Typically, bottom-up models refer to biological processes. They are designed to reveal certain image regions that are different from the surroundings, without dealing with cognitive phenomena that make these regions relevant.

In contrast, top-down methods are drawn to react to the image semantics, such as task demands and expectations Judd et al. [2009]. Early top-down attention models were driven by handcrafted features learned by training on human visual attention data sets [Borji et al., 2015]. Recently, the advances in deep learning [Zhang et al., 2018] and the increasing availability of large annotated databases [Winkler and Subramanian, 2013, Jiang et al., 2015] have enabled saliency models to perform end-to-end learning and consequently achieve strong improvements Borji [2018].

The eDN (ensembles of Deep Networks) model Vig et al. [2014] was the first effort to apply CNN for saliency prediction. It identifies saliency predictive instances of a richly-parameterized biology-inspired hierarchical model and then combines them using a linear SVM. Liu et al. [2018] introduced a multiresolution CNN (Mr-CNN) to learn both the bottom-up and top-down factors simultaneously. In this model, three different CNNs were trained at different scales and, then, combined by two fully connected layers for final saliency prediction. SALICON (Saliency in Context) Huang et al. [2015] explores the contextual information to reduce the semantic gap between the model prediction and the eye fixations. It concatenates two pre-trained CNNs, each on a different image scale (fine and coarse), to create its saliency map. A relevant contribution of Huang et al. [2015] is to fine-tune CNNs with saliency metric as an objective function. DeepFix Kruthiventi et al. [2017] innovated being the first algorithm to apply Fully Convolutional Neural Networks (FCNN) for saliency prediction. The authors also presented the novel Location Based Convolutional (LBC) to capture object-level semantics at multiple scales. Liu and Han [2018] proposed the Deep Spatial Contextual Long-term Recurrent Convolutional Neural Network (DSCLRCN model). DSCLRCN learns saliency-related local features of image regions using CNN and incorporates global and scene contexts to reveal the final saliency map. SAM (Saliency Attentive Models) Nets Cornia et al. [2018] predicts saliency by combining a FCNN with an attentive recurrent mechanism. An exhaustive study about the deep saliency models can be found in Borji [2018].

Most of the aforementioned predictions models were designed to estimate human eye fixations on natural images, which typically represent photographs of everyday scenes [Borji and Itti, 2013, Borji et al., 2015]. Moreover, several databases [Winkler and Subramanian, 2013], comprehensive and up-to-date benchmarks [Borji and Itti, 2013, Borji et al., 2015] and public challenges [Jiang et al., 2015, Bylinskii et al., 2015] have focused on the investigation of visual attention models for natural images. Recently, there is an increasing interest in understanding the expert's patterned eye movement for medical image analysis [Matsumoto et al., 2011, Li et al., 2016, Lévêque et al., 2018], and how this ability can support the development of automatic diagnosis systems [Guan et al., 2018]. However, the number of studies that links visual attention models and cell analysis is limited. Particularly, algorithms that simulate the human selective attention can enable fast object detection from large Pap smear slides by ranking the most relevant clinical areas within the images Zhang et al. [2013]. It can represent a promising strategy to drive the focus of classification, image compression, and other routines. In fact, the identification of image parts that are relevant to a cytologist may solve a current challenge: to design real-time applications to analyze Pap smear images under real noisy conditions, artifacts, and cell occlusions [William et al., 2018].

Previous work proposed by Coombes and Culverhouse [2003] employed visual attention theory to analyze cells. These authors used an eye tracking device for collecting the cytopathologist's visual data and identify manually marked salient features that are valuable for the development of quality assurance models on smear slide screening. Coombes and Culverhouse [2003] concluded that the use of saliency maps for providing feedback to the cytopathologist may reduce the diagnostic divergence on regular screening. Another finding was the high correlation between the expert's eye fixations and cell staining in the image. Zhang et al. [2013] proposed a method for abnormal cervical cell detection based on the bottom-up attention mechanism and the top-down information, such as size and color of abnormal nuclei. However, Zhang et al. [2013] used only liquid-based cytological images [Zhu et al., 2007], and concluded that visual attention mechanisms support finding diagnostic-relevant cells without massive processing of the whole image.

Different from previous approaches, such as in [Coombes and Culverhouse, 2003] and [Zhang et al., 2013], our paper investigates the feasibility of using saliency prediction methods to support screening of cervical cells from real routine exams using the conventional Pap smears, a harder task, but highly necessary as most of the countries still rely on this exam modality. The main contribution of this work consists in applying CNNs and state-of-the-art algorithms to predict cytopathologist's eye fixation and extract ROIs from conventional Pap smear images. Particularly, we are interested in confirming two major hypotheses:

- **Hypothesis 1 (H1):** Bottom-up methods present comparably accurate results to the predictions obtained by expert's visual attention.
- **Hypothesis 2 (H2):** State-of-the-art saliency models can detect clinically relevant ROI from Pap smear images.

In order to address these hypotheses, this paper describes four contributions: 1) quantitative evaluation of state-of-the-art saliency models in comparison to the cytopathologists' visual attention. Furthermore, we report the bottom-up methods that can detect relevant areas for domain experts; 2) verification of Coombes and Culverhouse [2003] findings for conventional Pap smear images that there is a high correlation between low-level features of the abnormal cells, such as brightness and color appearance, and the cytopathologist's attention on task-driven cervical cell image analysis. In addition, we show that the cytopathologist's gaze is strongly guided by top-down factors; 3) training of saliency prediction models using a CNN-based framework with two different neural networks (VGG-16 [Simonyan and Zisserman, 2014] and ResNet-50 [He et al., 2016]) based on Pap smear images using our human attention maps as ground truth; and 4) detecting which cell lesions are best identified by the saliency prediction algorithms if applied as a ROI extractor for diagnosis purposes.

In addition, this paper makes our saliency data set containing 5654 cervical cells from 232 images from real exams available at `http://dx.doi.org/10.17632/bk45c9yxb9.1`. To the best of our knowledge, this work is a pioneer in publishing cytopathologist's visual attention data recorded by an eye tracking device, contributing both to selective visual attention and to cell analysis reproducibility.

The remainder of this paper is organized as follows. Section 2 presents our cervical cell image database, the proposed methodology to collecting the cytopathologists's attention data, the surveyed saliency methods, and our performance evaluation methodology. The experimental results are discussed in Section 3, and our conclusions and future directions are drawn in Section 4.

2 Materials and Methods

2.1 CRIC database

The Center for Recognition and Inspection of Cells (CRIC) database contains digitized Pap smear images which were acquired with a Carl Zeiss microscope equipped with a Zeiss AxioCam MRc camera at 40× magnification. The acquired images have 0.255 μm/pixel and a resolution of 1392×1040 pixels (8-bit). The specimens were prepared via conventional Pap smears and contain cervical cells as well as other artifacts often collected as part of the exams. The cervical cells in the CRIC dataset are labeled into normal and, where abnormal uses the following classification: Atypical Squamous Cells of Undermined Significance (ASC-US); Atypical Squamous Cells of High Significance (ASC-H); Low-grade Squamous Intraepithelial Lesion (LSIL); High-grade Squamous Intraepithelial Lesion (HSIL); and Carcinoma (CA); patterns [Nayar and Wilbur, 2015] which were manually classified by three cytopathologists. Figure 1a displays image samples and illustrates the organization of the CRIC dataset.

We introduce a new dataset, named CRICVA[1] (CRIC Visual Attention) Ferreira et al. [2019], using 232 images of the CRIC. We arranged this database according to the following requirements: 1) images with cells distributed uniformly on the whole visual area; 2) images with different artifacts, such as blood cells, inflammatory cells, distorted cells, overlapping objects, mucus, etc; 3) images selected by the two-party vote; and 4) 80 images randomly selected from the remaining CRIC images (Fig. 1b). The random selection aimed to reduce the selection bias[2] in the voting process. Figure 1d displays the contents of the CRICVA dataset.

2.2 Collecting visual attention data

2.2.1 Subjects

We recorded the visual attention data from three cytopathologists, all of them with normal or corrected-to-normal vision via lens glasses. All of the subjects are experts in cervical care and reading conventional Pap smear slides. The mean age of the participants is 34.3 years (46, 30, and 27) and the length of their career in cytopathology laboratory is 20, 8 and 3 years, respectively. The two most experienced participants have Ph.D. degrees and the other has a Master degree in cytopathology. The screeners signed the informed written consent to participate in this project. The study has been approved by the Ethics Committee of Universidade Federal do Ceará under protocol number 2.439.252. We conducted all activities in accordance with the ethical guidelines defined by the Declaration of Helsinki and Brazilian laws.

2.2.2 Eye tracking task procedure

We used an eye tracking device to collect attention data from cytopathologists during their analyses of cervical cell images. Here, we focused on eye fixation, which is one of the eye movement parameters obtained by the eye tracker. Fixation is a period wherein the eye remains still, reflecting the conspicuity of a particular area of an image [Carrasco, 2011]. We represented the locations of eye fixations as a binary matrix, as shown in Figure 1f. From this matrix, we extracted attention heat maps (Fig. 1g) by convolving an isotropic 2D Gaussian of 1° (one degree) of visual angle centered on each fixation, where one degree of visual angle stands for an estimate of the size of the fovea [Torralba et al., 2006, Le Meur and Baccino, 2013]. We identified salient image regions for experts by overlapping the attention map with the input image as color-coded in Figure 1h, where dark red indicates the most conspicuous areas, and dark blue otherwise.

We performed our experiments using an EyeLink 1000 system designed by SR Research Ltd., Mississauga, Canada, with a sampling rate of 1000 Hz on the right eye recording (Fig. 1c). We created the tasks using SR Research Experiment Builder V4 and presented a sequence of eight trials for each participant on a Dell E178FPC at 60 Hz. Each trial was composed by the cervical cell images according to the following quantities: [Trial: Number of images]: 01: 26; 02: 26; 03–06: 25, and 07–08: 40. We exposed the cy-

[1] Cervical cell images and eye fixation data are available at http://dx.doi.org/10.17632/bk45c9yxb9.1

[2] Selection bias is related to the preference for a particular kind of image during the composition of the database [Torralba and Efros, 2011].

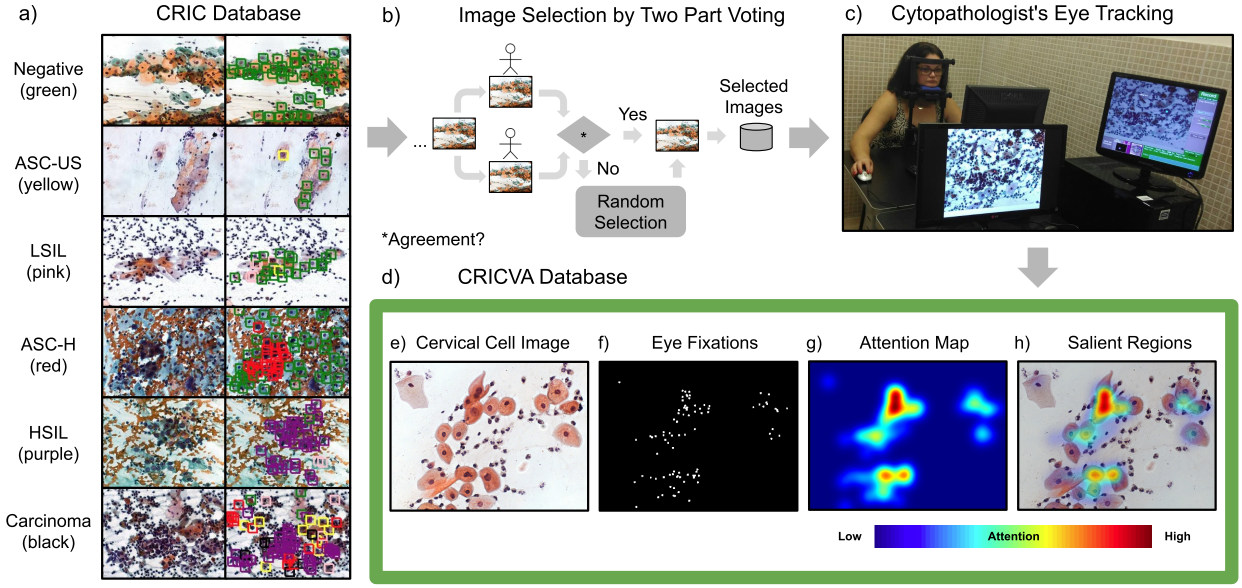


Figure 1: Overview of the CRICVA database. a) Images from the CRIC dataset with cell labels. b) Our methodology to select CRICVA images in order to reduce the selection bias. c) Collecting cytopathologist's eye movements by an eye tracking device. d) Visual attention data contained in the proposed CRICVA dataset. e) RGB cervical cell image. f) Cytopathologist's eye fixations as binary map. g) The attention map created by placing a Gaussian at each fixation position. h) Overlap between the input image and the attention map to highlight the salient regions on the image. Dark red regions correspond to the most conspicuous regions whereas dark blue the others.

topathologists to the same amount of images in each trial. Our motivation for adopting different image quantities across sessions was to increase the ability of our database to understand cytopathologist screening performance over time. Before each trial, the cytopathologists carried out a nine-point calibration procedure to map the eye-fixation to the screen coordinates. Furthermore, the participant had the opportunity to relax and report any discomfort with the experiment in the trial intervals.

We conducted a task-driven experiment in which the cytopathologist interpreted each cervical cell image and marked the abnormal cells with mouse clicks. The participants had free time to analyze the images, and the mouse clicks were not visible on the monitor. The cervical cell images were resized to 1280 x 1024 pixels, keeping the original aspect ratio by adding white pixel lines at the image bottom. Every time the cytopathologist pushed the space bar of the keyboard connected to the eye tracker, a new cell image appeared on the monitor. The images came into view in a randomized order to reduce any potential bias in the presentation of a sequence of images, and the participants could not return to the previous image. At the end of each trial, the participant data were recorded and a new round started.

2.2.3 Consistency across the participants

Inspired by Volokitin et al. [2016] and Bylinskii et al. [2018], we measured the gaze agreement among the cytopathologists for each image. We created a fixation map for each cytopathologist and then used it to predict fixations of the excluded ones. The Area Under the Receiver Operating Characteristic Curve (AUC-Judd) [Judd et al., 2009] metric was used for performance evaluation of our approach. The average AUC-Judd between all cytopathologists was 0.823 (SD=±0.007) and the p-value was 0.117 (Kruskal-Wallis ANOVA test with $\alpha = 0.05$) [Kruskal and Wallis, 1952], considering the whole image dataset. We also observed the intra-participant performance across sessions and we found the average values for AUC-Judd equal to 0.818 (SD=±0.016), 0.813 (SD=±0.018), and 0.829 (SD=±0.016) for each cytopathologist, respectively. These results suggest a high consistency among the participants' eye fixation patterns on the CRICVA database, evidencing the existence of well-defined clinical ROIs on Pap smear images and supporting the investigation of our H2 hypothesis.

2.2.4 Center bias analysis

The human tendency to look for objects near the central image region is a frequent bias in computer vision databases [Tatler, 2007, Borji et al., 2014]. Moreover, these datasets tend to be subjected to the photographers bias in framing relevant objects in the central region of the image [Borji et al., 2015]. To understand how the center bias occurs in the proposed dataset, we analyzed the Average Annotation Map (AAM) of the CRICVA database, as shown in Figure 2a. The AAM conveys the average of the visual attention ground-truth annotations of the whole image data set [Borji et al., 2015]. To analyze the dispersion of the AAM, we contrasted the horizontal and vertical midlines of the AAM with those obtained from a Gaussian blob at the center of the image, as illustrated in Figures 2b and 2c, respectively. The Gaussian kernel was set at 1° of visual angle to reflect estimates of fovea size [Tatler, 2007]. Although the AAM has a larger activation near the image center, there exists a significant variation in the spatial distribution of the activations, indicating that the cytopathologists spent a considerable amount of time analyzing objects far from the image center.

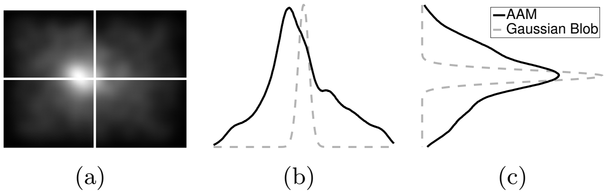


Figure 2: Center bias analysis of the CRICVA dataset. a) Average Annotation Map (AAM). b) Horizontal and c) vertical mid-lines of AAM and Gaussian blob at the image center.

2.3 Saliency methods

We investigated 10 bottom-up methods and two variants of a CNN-based model for top-down saliency prediction, covering different categories of algorithms [Borji and Itti, 2013]. We also considered the model proposed by Zhang et al. [2013] since it is designed to explore top-down factors in cervical cell images. We summarized the studied methods and introduced the model abbreviations adopted in the rest of the paper in Table 1. All algorithms, except the one proposed by Zhang et al. [2013], were validated on public data sets and chosen for this work according to the following criteria: 1) the input is a single image; 2) the source code is publicly available; 3) they present high performance in well-established saliency ranking list; 4) the runtime is less than three seconds per image, and 5) they are state-of-the-art algorithms or stand for a benchmark in the literature. Based on these requirements, we visited the MIT Saliency Benchmark [Bylinskii et al., 2015], sorted the ranking results by NSS (Normalized Scanpath Saliency) and picked out the top-ranking CNN-based algorithm: SAMv and SAMr [Cornia et al., 2018], in which the feature maps are extracted by the VGG-16 and ResNet-50, respectively. We also selected five non-deep methods: BMS [Zhang and Sclaroff, 2013], LDS [Fang et al., 2017], FES [Tavakoli et al., 2011], SWD [Duan et al., 2011], and UHM [Tavakoli and Laaksonen, 2016]. We complemented our study with the IT [Itti et al., 1998], GB [Harel et al., 2006], SR [Hou and Zhang, 2007], SS [Hou et al., 2012], and SIM [Murray et al., 2011] bottom-up models from the extensive benchmark introduced by Borji et al. [2015].

2.4 SAM model setup

We employed two different CNN architectures as the backbone for the SAM model. As proposed by Cornia et al. [2018], we combined a dilated version of the VGG-16 network into the SAM pipeline and trained it on the eye-fixation data from natural images using the SALICON [Jiang et al., 2015] dataset (SVS: SAM VGG-16 on SALICON) and cervical cell images (SVC: SAM VGG-16 on CRICVA). We also used the SAM approach based on the dilated ResNet-50 network, obtaining two other versions: SRS (SAM ResNet-50 on SALICON) and SRC (SAM ResNet-50 on CRICVA).

The SALICON database comprises 10,000 training images, 5,000 validation images, and 5,000 testing images divided into 80 categories, being the largest available data set for saliency prediction in natural images. Regarding cell images, we carried out the SAM model training and validation procedures on random sets of the CRICVA database, containing 130 and 29 images, respectively. We drew the CRICVA testing database from the 73 remaining images. Table 2 outlines the key information about the SAM model usage.

We adopted the default configuration of the SAM model as described by Cornia et al. [2018]. The initial weights of dilated CNN were defined with those of the VGG-16 and ResNet-50 models trained on ImageNet [Krizhevsky et al., 2012] and the recurrent weights matrices of the Attentive ConvLSTM were initialized as random orthogonal matrices. We resized the images to 240 x 320 and employed the same overall loss function proposed by Cornia et al. [2018].

2.5 Saliency prediction analysis

Figure 3 presents our methodology to investigate H1 within the saliency prediction analysis module. We focused our experiments on evaluating how well the surveyed saliency models predict where cytopathologists look at when analyzing cervical cell images. To this end, we processed the color (RGB) cervical cell image by the surveyed saliency models and performed evaluation of estimated maps against the visual data collected by the eye tracking device. We applied the following state-of-the-art metrics to assess the results and measure the influence of top-down factors on the cytopathologist's analysis.

2.5.1 Evaluation metrics

Bylinskii et al. [2018] categorized the metrics for saliency model evaluation in location-based and distribution-based according to the data representation. The location-based metrics use the discrete fixation locations (Fig. 1f) as saliency maps. The distribution-based metrics consider both attention maps and predicted saliency maps as continuous distributions. According to Bylinskii et al. [2018], the distribution-based metrics allow incorporating uncertainty in the measurements, such as the errors in eye tracking and imprecision of human eye position on the screen. Additionally, the distribution-based metrics are more robust to few observers than location-based ones since they extrapolate the data to model the behavior of more observers.

We employed five well-established measures[3] for performance evaluation of the surveyed models and we appraised our results in terms of false positives and false negatives. The location-based metrics that we adopted are the AUC-Judd [Judd et al., 2009] and the Normalized Scanpath Saliency (NSS). The AUC-Judd metric gives a high score for high-valued predictions placed at fixed locations, but it ignores low-valued false positives. The NSS metric is equally affected by false positives and negatives.

For the distribution-based category, we choose the Linear Correlation Coefficient (CC), the Similarity between Distributions (SIM) and the Kullback-Leibler divergence (KL) metrics. CC is symmetric and penalizes false positives and false negatives equally. SIM computes the intersection between two distributions, being more sensitive to false negatives than false positives. KL corresponds to an asymmetric dissimilarity metric highly sensitive to false negatives. Lower values of KL indicate better results. A broader study about evaluation metrics designed for saliency models is available in [Bylinskii et al., 2018].

2.6 ROI selection analysis

The second module of the proposed methodology investigates the H2 hypothesis. We measured

[3] The eye fixation evaluation metrics are available at https://github.com/cvzoya/saliency/tree/master/code_forMetrics.

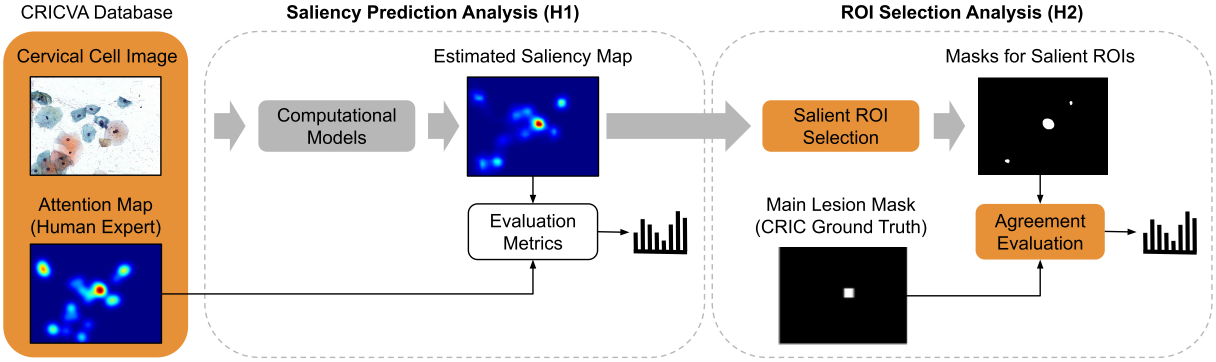


Figure 3: Our investigation consists of two steps: 1) evaluation of the surveyed saliency models according to state-of-the-art metrics (H1 test), and 2) ROI selection analysis to determine whether the most relevant lesion in the image is highlighted on the estimated saliency map (H2 test). The orange boxes represent additional contributions of this work.

the probability of the most relevant cell lesions to be signed as a salient object by the surveyed models. Inspired by Bylinskii et al. [2016], we mark as salient objects those whose positions are presented on the predicted saliency map as highlighted regions. We introduce the Algorithm 1 to draw the $nr = 3$ separated regions with the highest values for each estimated saliency map. If this map has $nr < 3$ salient ROIs, our algorithm thresholds it at the 95th percentile.

We define the agreement rate $\varphi \in [0, 1]$ between the salient ROIs and the location of the most significant cell lesion (le) in the image as:

$$\varphi = \frac{1}{N_{le}} \sum_{i=1}^{N_{le}} \# \{\forall r \in Z_i \mid (r \cap G_i) \neq \emptyset\}, \qquad (1)$$

where $\# \{\cdot\}$ stands for the cardinality, N_{le} is the number of testing images with le lesion, Z is the binary mask for high-density saliency regions, r is the salient ROI and G represents the ground-truth binary mask signaling the 100 x 100 pixels of le area. Then, we considered *identified* if any part of the main cell lesion matched the ROI. Otherwise, we assumed that the model could not identify the location of the main lesion, appropriately.

3 Experimental Results

In this section, we describe the analyses and experiments to test each hypothesis and validate the contributions of different algorithms. We also report the bottom-up models with the best performance on cervical cell images in terms of false positive and false negative. The parameters used for each algorithm are those published in the original papers listed in Table 1.

3.1 Saliency prediction analysis

We investigated H1 on 73 images (CRICVA testing set). We first conducted the analysis of classic (non-deep) bottom-up methods. Afterward, we analyzed the performance of the top-down and CNN-based methods. We also performed experiments with a center prior baseline model, which consists of a Gaussian blob (kernel width at 3 degrees) at the center of the image. We employed this baseline to reveal the existence of capture bias[4] in the CRIC database and to analyze the influence of center bias on the performance of the surveyed methods.

Algorithm 1: ROI selection algorithm for finding the highest density regions on the estimated saliency maps.

```
function ROIselect ($M, nr$);
Input  : (*float matrix* $\in$ [0,1]) $M$: Estimated saliency map; (*int*) $nr$: number of required regions
Output: (*logical matrix*) $Z$: Binary mask signaling the ROI locations
$th = 0.95$;
$step = 0.05$;
$countr = 0$;
while ($countr < nr$) and ($th > 0$) do
    $Z$ =logical($M. * (M >= th)$);
    $countr$ =getNumberOfSeparatedRegions($Z$);
    $th = th - step$;
end
if ($th <= 0$) then
    $Z$ =logical($M >= 0.95$);
end
  ▷ Ranking the regions in descending order based on the energy of the respective pixels and returning up to $nr$ most salient separated regions.
$Z$ = getNRMostSalientRegions($Z$,$M$,$nr$);
return $Z$;
```

3.1.1 Bottom-up methods

Figure 4 shows the estimated saliency maps from all algorithms for an input cervical cell image. The quantitative comparison of these maps is presented in Figure 5.

Overall, the bottom-up algorithms tended to highlight false positive attention regions on cervical cell images. Since these methods were mainly designed to simulate the human low-level visual attention, features such as the contrast between the cells and background, and the staining of image structures, had a significant impact on saliency prediction.

The SIC method was strongly affected by the brightness and color appearance, as a result of the color perception methodology developed by Otazu et al. [2010]. Thus, the SIC algorithm tended to highlight image regions of high contrast in relation to its surroundings, such as cell-background transition, artifact presence, and background noise. Therefore this model is seldom adequate for saliency prediction on cervical cell image applications that require low scores of false positive. In special circumstances, the SIC approach can be useful to detect image areas with high contrast objects, such as normal nuclei and neutrophils.

[4] Capture bias conveys some tendency of the photographers to position the targets on the picture during the scene imaging [Torralba and Efros, 2011].

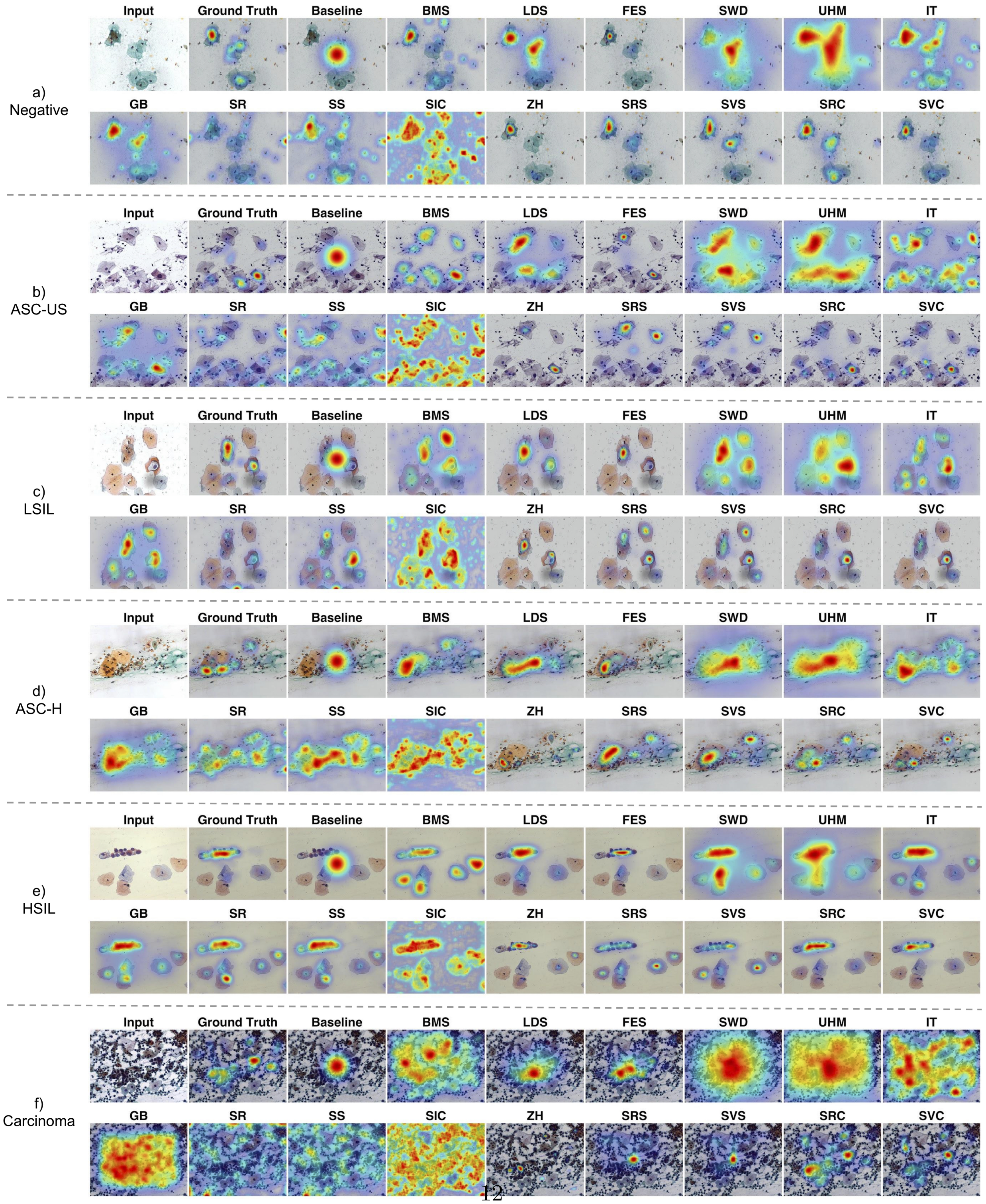


Figure 4: Computed saliency maps by the surveyed methods overlapped upon the image. a) to f) display our results for each case of the CRICVA dataset. Some models are able to indicate conspicuous image regions on cervical cell analysis, whereas others may be ineffective for prediction of cytopathologist's gaze when on noisy images and clustered cells.

An algorithm that produced saliency maps visually close to the cytopathologist's attention maps was FES, which revealed that the use of local features in a Bayesian framework enables detection of abnormal cells. The FES method's nature is implicitly biased by the approach used for estimating the prior probability distributions. The original FES implementation uses the AAM over a training set for approximating the priors required by the Bayesian approach. We processed the CRICVA AAM (Fig. 2a) and excluded test data in accordance with Tavakoli et al. [2011] to estimate the FES prior distributions and run our experiments. Although the FES algorithm emphasized salient objects close to the image center due to the AAM information, this algorithm found abnormal cell regions far from the central region. The BMS, LDS, and GB techniques also presented high saliency values at eye fixation locations, but presented higher sensitivity to false positives than the FES algorithm.

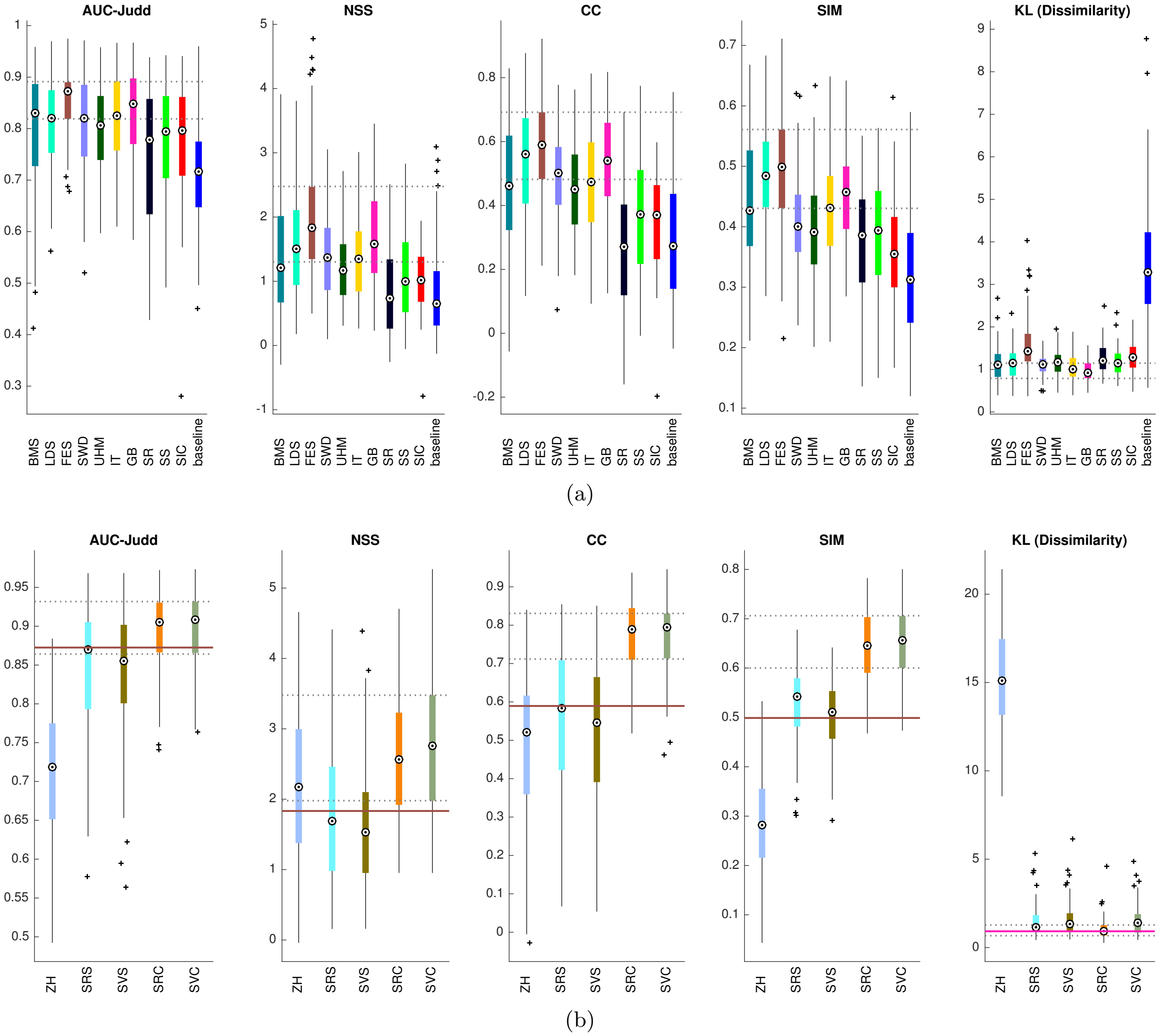
We also performed the Kruskal-Wallis statistical test [Kruskal and Wallis, 1952] with post-hoc Nemenyi test ($\alpha = 0.05$) [Hollander et al., 2013] to find the model results that differ significantly from each other. Figure 6 reports all pairwise comparisons for each studied evaluation metric.

In terms of the AUC-Judd measure, the FES method outperformed all the other bottom-up models. However, it did not differ statistically from BMS, LDS, SWD, UHM, and IT as well as GB. Based on the AUC-Judd features discussed by Bylinskii et al. [2018], we noticed that these methods may be suitable for cervical cell image analysis slightly affected by false positives. Afterwards, we extended our investigation by considering equally the effects of false positives and false negatives. The NSS measure pointed out LDS, FES, and GB as the best-performing classic method, and the CC and SIM metrics confirmed this finding. Therefore, our tests showed that the LDS, FES, and GB algorithms are feasible for saliency prediction on cervical cell images that require accuracy in identifying visual attention regions. We found a few exceptions, which corresponds to the cases where there are large number of image artifacts (Fig. 4f). For these cases, the use of classic bottom-up methods may be ineffective. Although the FES algorithm scored well the location-task of the salient regions in most cases, the size of its blobs was usually smaller to those of the ground truth. This fact undermined the performance evaluation of the FES method by the KL metric, which is highly sensitive to pixel false negatives. These are some indications that the FES additional settings may boost its performance, but further investigations might be necessary to confirm such biases.

The results of the center prior baseline confirmed that part of the expert's visual attention was devoted to targets in the center of the image. However, the best bottom-up methods, according to the evaluation metrics, reached salient regions far from the image center, showing that they captured satisfactorily the low-level features of the targets. Based on this analysis, we validated the H1 hypothesis by arguing that the cytopathologist's visual attention is highly correlated with the low-level features of the abnormal cells. In addition, the appropriate modeling of the cell attributes may allow the use of fast bottom-up algorithms as part of a saliency prediction framework for cell image analysis.

3.1.2 Top-down models

Figure 4 shows that the target-driven approach implemented by the ZH model represents a



(a)

(b)

Figure 5: Quantitative evaluation for a) bottom-up and b) top-down and CNN-based methods. The dotted horizontal line stands for the interquartile range of the best model. For each metric, the colored horizontal solid line marks the performance of the best bottom-up algorithm. The axis ranges are defined according to the evaluation metric limits [Bylinskii et al., 2018].

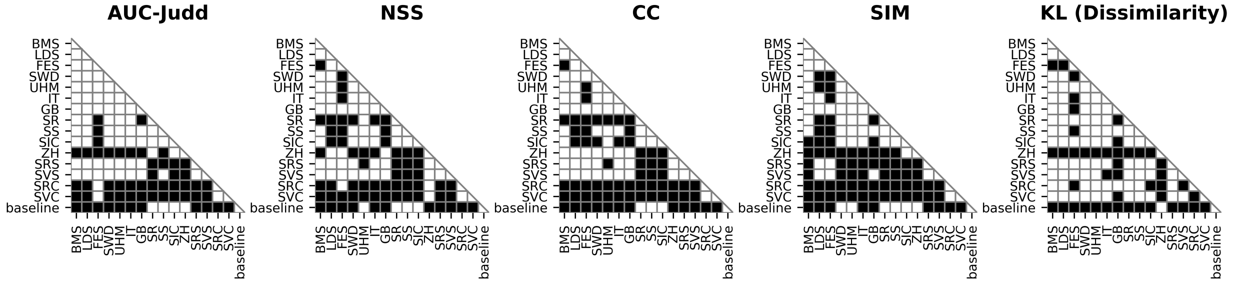


Figure 6: Pairwise comparisons for all surveyed methods using the Kruskal-Wallis statistical test with the post-hoc Nemenyi test. The black boxes represent the pairs with significant difference at $\alpha = 0.05$.

promising strategy to localize conspicuous areas for cytopathologists on cervical cell images. However, the saliency maps may be sparse for cytopathologist's eye fixation prediction. Since the ZH model is driven by an abnormal cell template matching, some salient areas can be neglected by ineffective correspondence between the template and the shape of the cells (Fig. 4b). In conventional Pap smears, some abnormal nuclei can present distorted shapes due to the clinical acquisition process, high overlap and intensity variations. These factors reduce the performance of the annular template adopted by the ZH model, leading it to produce saliency maps confined to specific image regions. These observations are supported by our quantitative analysis presented in Figure 5. AUC-Judd revealed that the ZH model underweighted relevant attention areas (false negative), although NSS confirmed assertiveness for some regions. Furthermore, the distribution-based metrics (CC, SIM, KL) suggested a significant distance between the attention areas demanded by the cytopathologists and those estimated by the ZH model.

Our experiments showed that SVC and SRC indicated improved saliency maps in comparison with all other surveyed models. Since SVS and SRS performed similarly to the best bottom-up methods (Fig. 6), we argue that SVC and SRC learned top-down features that are relevant to driving the cytopathologist's visual attention. Particularly, we demonstrated that the transfer learning from a CNN trained on a large-scale data set from a different domain is also suitable for saliency prediction on cervical cell images, mainly for the low-level feature modeling on earlier layers. Although our experiments have revealed differences between the top-down factors on natural and cervical cell image analysis, we found evidence that the CNN-based frameworks, trained on a small database with cytopathologist's attention maps, can predict valuable saliency maps on Pap smear images through transfer learning. An interpretation of the CNN results here was that there are features relevant to both domains, possibly through low-level vision primitives. Furthermore, our results pointed out that the top-down factors guide the cytopathologist's attention during the diagnosis task, mainly reducing the sensitivity to brightness and high contrast areas.

By exploring CNN architectures for both

natural and cervical cell images, we analyzed the correlation between the activations of the CNN layers when trained in the SALICON and CRICVA databases and we presented our results in Figure 7. We observed that approximately the first half of the layers have correlation coefficient above 95% in their activations, which confirmed that the CNN lower layers represent low-level features and build upwards toward higher-level representation on upper layers. For CNN training on cell images, our analysis may be used to adjust some fine-tuning parameters, such as the number of lower layers to be frozen and the learning rate. Additionally, our analysis showed that the activations of the lower layers of both CNN architectures were similar for visual attention regardless of the database purpose.

The Kruskal-Wallis and post-hoc Nemenyi statistical tests did not identify statistical differences between SVC and SRC for all evaluation metrics, except KL (Fig. 6). This led us to conclude that both variants may be equally chosen for cervical cell applications according to the availability of the computational resources.

3.1.3 Runtime performance

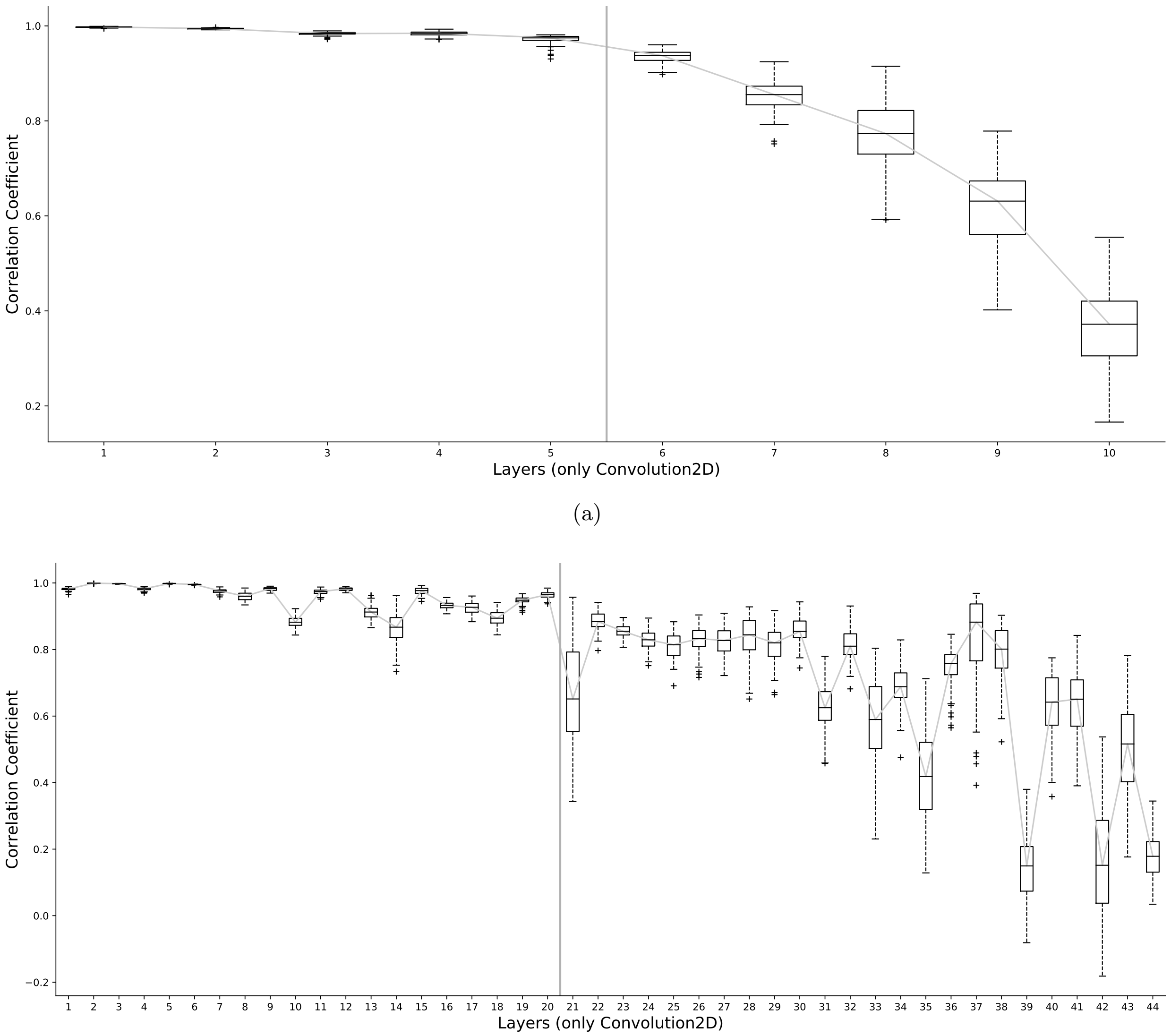
The average time of the surveyed methods on the CRICVA database was ranked in Table 3, based on algorithms whose source code were publicly available and considering the experimental configurations suggested in their respective scientific articles. The bottom-up approaches, apart from the SIC method, performed faster than the deep learning algorithms. The results in Figure 5 show that the bottom-up methods may be a viable solution for cervical cell image analysis and applications in which computational resources are limited – this motivated us to report results using a computer with an Intel (I7-4770HQ) CPU (2.2 GHz) and 16 GB RAM.

We summarized the contributions of the H1 hypothesis by reporting fast bottom-up techniques that can identify the most conspicuous regions within a cell image for analysis by cytopathologists. In addition, we confirmed a finding described by Coombes and Culverhouse [2003] that there is a high correlation between some low-level features of the abnormal cells and the expert's eye-fixation pattern. We also trained a state-of-the-art CNN-based framework using VGG-16 and ResNet-50 for saliency prediction using our expert's attention maps as ground-truth, achieving accurate saliency maps on conventional Pap smear images.

3.2 ROI selection analysis

The investigation of the H2 hypothesis required no eye fixations data; instead, it considers only the lesion annotations available on the CRIC database. Thus, we extended our experiment to the remaining abnormal images on this database, resulting in a total of 207 images. The distribution of image labels according to the main cell lesion is #{ASC-US: 52, LSIL: 100, ASC-H: 26, HSIL: 21 and Carcinoma: 8}.

Table 4 presents the φ results for all surveyed methods, considering three separated regions. The SRC model identified the location of the most relevant cell lesion for all considered pathologies, except for one LSIL case where four regions were required. This demonstrates the potentiality of this algorithm as a ROI extractor for automated cell pre-screening systems for cervical cells. In fact, it localized the lesion region within a few candidate ones. SVC also showed valuable results for ROI-based systems, however it was less robust in detecting regions with ASC-US and carcinoma lesions than the SRC method.



(a)

(b)

Figure 7: Correlation coefficient between the activations of a) VGG-16 and b) ResNet-50 layers trained on the SALICON and CRICVA databases. Layers up to 95% of correlation are shown on the left of the vertical line.

The ZH model achieved competitive results. As the authors of ZH only reported the results for liquid-based cell images Zhang et al. [2013], our findings extend the application of this method to conventional Pap smears. Some bottom-up strategies, such as those adopted by FES, and

GB, may be suitable for ROI selection in cervical cell images, especially in images with sparse cell clumps and low presence of dark artifacts. Although SWD and UHM performed well in our experiments, Figure 4 suggests the existence of large false positives, which may restrict the use of these methods for some applications. These results, in addition to those described in Figure 5, point out that the FES method is our best bottom-up strategy for ROI selection on cervical cell images from accurate saliency maps.

Nevertheless, we argue that the φ analysis is robust only for the ASC-US, LSIL, and ASC-H lesions. For these cases, our center prior baseline analysis revealed the absence of significant capture bias, confirming the performance of the surveyed methods. For HSIL and carcinoma, further works should consider: 1) a larger and more diverse CRIC database of cell samples, especially cases for carcinoma and 2) avoidance of the center prior baseline, which scored high (above 90%) for the current HSIL images, suggesting that HSIL lesions were mostly located on the image center, which may have influenced the performance of the methods.

According to H2, our experiments confirmed the reliability of several saliency prediction models in identifying critical cells for the diagnosis. Other contributions included the proposed algorithm for salient ROI selection (Algorithm 1) and the agreement rate (Equation 1) that quantified the performance of the surveyed models applied to ROI extraction. Finally, we emphasized the importance of organizing visual attention databases for specific domains, such as medical and cell imaging to benchmark algorithmic advances.

4 Conclusion and Future Directions

This paper evaluated the performance of state-of-the-art saliency models applied to conventional Pap smear image analysis. We investigated 10 bottom-up algorithms, one target-driven model that highlights conspicuous abnormal cell regions, and two variants of a CNN-based framework trained on natural images and cervical cell microscopy, using VGG-16 and ResNet-50 networks as backbone. Our results revealed that top-down factors could guide the cytopathologist's attention on task-driven analysis. In addition, bottom-up methods could also recover relevant cells for accurate diagnosis, although at the expense of false positives.

We also observed high correlation (above 95%) between the first half of CNN layers trained on natural, and cervical cell image databases. Figure 7 illustrated a strategy to identify a specific CNN layer to fine-tune databases of cervical cells for saliency prediction purposes. Furthermore, we showed that a transfer learning approach from a different domain allows CNN methodologies to achieve promising saliency prediction on cervical cell images, even using a small cervical cell image database. Future work might explore similarities across domains as part of schemes to address CNN interpretability.

The CNN-based models trained on cervical cell images outperformed the surveyed algorithms mainly because they achieved lower false positives on the estimated saliency maps and remained sensitive to relevant cell regions. These algorithms identified the most important region on the image among the three most salient regions. Thus, it confirmed the applicability of these algorithms to extract ROIs from cervical cell images. For LSIL, ASC-H and HSIL cases, our results also revealed fast bottom-up algorithms with similar ROI extractor performance to the CNN-based models with aggreement above 86%. This result indicated that there are feasible algorithms for applications with low availability of memory and computational power.

The CRIC database represented a step forward in benchmarking algorithms, but it showed bias toward the HSIL images. For these cases, the proposed center prior baseline, which consisted of a Gaussian blob at the center of the image, indicated that about 90% of HSIL lesion (when it is the most significant in the image) is located near to image center. Since the HSIL cells represent high-risk lesions, it was natural for photographers to position them close to the image center. Thus, we recommend further researches to investigate the performance of saliency methods on HSIL images free of capture bias. On the other hand, we did not find significant evidence that the capture bias influences the saliency prediction on images with other pathologies.

We identified two main limitations of our approach: 1) the CRIC database contained few HSIL #{21} and carcinoma #{8} images. This statistically restricted the ROI extractor analysis for these diseases. 2) The participation of only three cytopathologists reduced the accuracy of the location-based evaluation metrics. Additional research with more cell image databases, more lesion cases and more cytopathologists are needed to better assess the application of saliency prediction techniques as region ranking for diagnosis systems and for optimizing parameters of supervised models.

In the future, other saliency models that may improve the eye fixation prediction of cytopathologists on cervical imaging need to be investigated for several applications. Furthermore, eye tracking studies on image perception within cell analysis would be beneficial for the whole medical imaging community, especially to understand scan patterns and the reasons for diagnostic error.

Acknowledgment

This work was supported by CAPES/CNPq-PVE (401442/2014-4) and CNPq (306600/2016-1), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, PPSUS/FAPEMIG (APQ-03740-17), Propp/UFOP, the Center for Recognition and Inspection of Cells (CRIC), the Moore-Sloan Foundation, and a research fellowship from CNPq (Carneiro, CM). Partial work on the development of machine learning algorithms was supported by the Office of Science, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. The authors would like to thank Dr. José Soares de Andrade Júnior and Dr. Humberto de Andrade Carmona for authorizing the use of the eye tracking device at the Complex Systems Laboratory (Department of Physics/UFC). Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Department of Energy or the University of California. Declarations of interest: none.

References

Marisa Carrasco. Visual attention: The past 25 years. *Vision Research*, 51(13):1484–1525, 2011.

Liming Zhang and Weisi Lin. *Selective visual attention: Computational models and applications*. Wiley-IEEE Press, 1 edition, 2013. ISBN 9780470828120.

Alfred L Yarbus. *Eye movement and vision (translated from the russian edition by Basil Haigh.)*. Plenum Press, New York, 1967.

Patrik Polatsek, Manuela Waldner, Ivan Viola, Peter Kapec, and Wanda Benesova. Exploring visual attention and saliency modeling for task-based visual analysis. *Computers & Graphics*, 72(2):26–38, 2018.

Francesca Murabito, Concetto Spampinato, Simone Palazzo, Daniela Giordano, Konstantin Pogorelov, and Michael Riegler. Top-down saliency detection driven by visual classification. *Computer Vision and Image Understanding*, 172:67–76, 2018.

Constantinos Loukas, Christos Varytimidis, Konstantinos Rapantzikos, and Meletios A Kanakis. Keyframe extraction from laparoscopic videos based on visual saliency detection. *Computer Methods and Programs in Biomedicine*, 165:13–23, 2018.

Tam V Nguyen, Qi Zhao, and Shuicheng Yan. Attentive systems: A survey. *International Journal of Computer Vision*, 126(1):86–110, 2018.

Ali Borji and Laurent Itti. State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):185–207, 2013.

Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *IEEE 12th International Conference on Computer Vision (ICCV 2009)*, pages 2106–2113, Kyoto, Japan, September - October 2009. IEEE.

Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. Salient object detection: A benchmark. *IEEE Transactions on Image Processing*, 24 (12):5706–5722, 2015.

Qingchen Zhang, Laurence T Yang, Zhikui Chen, and Peng Li. A survey on deep learning for big data. *Information Fusion*, 42:146–157, 2018.

Stefan Winkler and Ramanathan Subramanian. Overview of eye tracking datasets. In *Fifth International Workshop on Quality of Multimedia Experience (QoMEX 2013)*, pages 212–217, Klagenfurt am Wörthersee, Austria, July 2013. IEEE.

Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. SALICON: Saliency in context. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015)*, pages 1072–1080, Boston, USA, June 2015. IEEE.

Ali Borji. Saliency prediction in the deep learning era: An empirical investigation. *arXiv preprint arXiv:1810.03716*, 2018.

Eleonora Vig, Michael Dorr, and David Cox. Large-scale optimization of hierarchical features for saliency prediction in natural images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2798–2805, 2014.

Nian Liu, Junwei Han, Tianming Liu, and Xuelong Li. Learning to predict eye fixations via multiresolution convolutional neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 29(2):392–404, 2018.

Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao. SALICON: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *IEEE International Conference on Computer Vision*, pages 262–270, 2015.

Srinivas SS Kruthiventi, Kumar Ayush, and R Venkatesh Babu. Deepfix: A fully convolutional neural network for predicting human eye fixations. *IEEE Transactions on Image Processing*, 26(9):4446–4456, 2017.

Nian Liu and Junwei Han. A deep spatial contextual long-term recurrent convolutional network for saliency detection. *IEEE Transactions on Image Processing*, 27(7):3264–3274, 2018.

Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. Predicting human eye fixations via an LSTM-based saliency attentive model. *IEEE Transactions on Image Processing*, 27(10):5142–5154, 2018.

Zoya Bylinskii, Tilke Judd, Ali Borji, Laurent Itti, Frédo Durand, Aude Oliva, and Antonio Torralba. MIT saliency benchmark. `http://saliency.mit.edu/`, 2015.

Hideyuki Matsumoto, Yasuo Terao, Akihiro Yugeta, Hideki Fukuda, Masaki Emoto, Toshiaki Furubayashi, Tomoko Okano, Ritsuko Hanajima, and Yoshikazu Ugawa. Where do neurologists look when viewing brain CT images? An eye-tracking study involving stroke cases. *PloS One*, 6(12):e28928, 2011.

Rui Li, Pengcheng Shi, Jeff Pelz, Cecilia O Alm, and Anne R Haake. Modeling eye movement patterns to characterize perceptual skill in image-based diagnostic reasoning processes. *Computer Vision and Image Understanding*, 151:138–152, 2016.

Lucie Lévêque, Hilde Bosmans, Lesley Cockmartin, and Hantao Liu. State of the art: Eye-tracking studies in medical imaging. *IEEE Access*, 6:37023–37034, 2018.

Qingji Guan, Yaping Huang, Zhun Zhong, Zhedong Zheng, Liang Zheng, and Yi Yang. Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification. *arXiv preprint arXiv:1801.09927*, 2018.

Jian-Wei Zhang, Min-Chao Lian, Wan-Peng Wang, and Lin Zhu. Detection of abnormal nuclei in cervical smear images based on visual attention model. In *IEEE International Conference on Machine Learning and Cybernetics (ICMLC 2013)*, volume 2, pages 920–924, Tianjin, China, July 2013. IEEE.

Wasswa William, Andrew Ware, Annabella Habinka Basaza-Ejiri, and Johnes Obungoloch. A review of image analysis and machine learning techniques for automated cervical cancer screening from pap-smear images. *Computer Methods and Programs in Biomedicine*, 164:15–22, 2018.

LR Coombes and PF Culverhouse. Pattern recognition in cervical cytological slide images. In *Fifth International Conference on Advances in Pattern Recognition (ICAPR 2003)*, Calcutta, India, December 2003. ICAPR.

Jie Zhu, Ingrid Norman, Kristina Elfgren, Vera Gaberi, Bjorn Hagmar, Anders Hjerpe, and Sonia Andersson. A comparison of liquid-based cytology and pap smear as a screening method for cervical cancer. *Oncology Reports*, 18(1):157–160, 2007.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)*, pages 770–778, Las Vegas, USA, June 2016. IEEE.

Ritu Nayar and David C Wilbur. The Pap test and Bethesda 2014. *Acta Cytologica*, 59(2): 121–132, 2015.

Daniel S. Ferreira, Geraldo L. B. Ramalho, Débora Torres, Alessandra H. G. Tobias, Mariana T. Rezende, Fátima N. S. Medeiros, Andrea G. C. Bianchi, Cláudia M. Carneiro, and Daniela M. Ushizima. CRICVA Database, 2019. Mendeley Data, v1, `http://dx.doi.org/10.17632/bk45c9yxb9.1`.

Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2011)*, pages 1521–1528, Colorado Springs, USA, June 2011. IEEE.

Antonio Torralba, Aude Oliva, Monica S Castelhano, and John M Henderson. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological Review*, 113 (4):766–786, 2006.

Olivier Le Meur and Thierry Baccino. Methods for comparing scanpaths and saliency maps: strengths and weaknesses. *Behavior Research Methods*, 45(1):251–266, 2013.

Anna Volokitin, Michael Gygli, and Xavier Boix. Predicting when saliency maps are accurate and eye fixations consistent. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 544–552, 2016.

Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(3):740–757, 2018.

William H Kruskal and W Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47 (260):583–621, 1952.

Benjamin W Tatler. The central fixation bias in scene viewing: selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, 7(14):1–17, 2007.

Ali Borji, Ming-Ming Cheng, Qibin Hou, Huaizu Jiang, and Jia Li. Salient object detection: a survey. *arXiv preprint arXiv:1411.5878*, 2014.

Jianming Zhang and Stan Sclaroff. Saliency detection: A boolean map approach. In *IEEE International Conference on Computer Vision (ICCV 2013)*, pages 153–160, Sydney, Australia, December 2013. IEEE.

Shu Fang, Jia Li, Yonghong Tian, Tiejun Huang, and Xiaowu Chen. Learning discriminative subspaces on random contrasts for image saliency analysis. *IEEE Transactions on Neural Networks and Learning Systems*, 28(5): 1095–1108, 2017.

Hamed Rezazadegan Tavakoli, Esa Rahtu, and Janne Heikkilä. Fast and efficient saliency detection using sparse sampling and kernel density estimation. In *Scandinavian Conference on Image Analysis (SCIA 2011)*, pages 666–675, Ystad, Sweden, May 2011. Springer.

Lijuan Duan, Chunpeng Wu, Jun Miao, Laiyun Qing, and Yu Fu. Visual saliency detection by spatially weighted dissimilarity. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2011)*, pages 473–480, Colorado Springs, USA, June 2011. IEEE.

Hamed R Tavakoli and Jorma Laaksonen. Bottom-up fixation prediction using unsupervised hierarchical models. In *Asian Conference on Computer Vision (ACCV 2016)*, pages 287–302, Taipei, Taiwan, November 2016. Springer.

Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20 (11):1254–1259, 1998.

Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. In *Advances in Neural Information Processing Systems (NIPS 2006)*, pages 545–552, Vancouver, Canada, December 2006. NIPS.

Xiaodi Hou and Liqing Zhang. Saliency detection: A spectral residual approach. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2007)*, pages 1–8, Minneapolis, USA, June 2007. IEEE.

Xiaodi Hou, Jonathan Harel, and Christof Koch. Image signature: Highlighting sparse salient regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(1):194–201, 2012.

Naila Murray, Maria Vanrell, Xavier Otazu, and C. Alejandro Parraga. Saliency estimation using a non-parametric low-level vision model. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2011)*, pages 433–440, Colorado Springs, USA, June 2011. IEEE.

Xavier Otazu, C. Alejandro Parraga, and Maria Vanrell. Toward a unified chromatic induction model. *Journal of Vision*, 10(12):1–24, 2010.

Marta Wesoła, Artur Lipiński, and Michał Jeleń. Morphometry in the cytological diagnosis of cervical smears. *Advances in Clinical and Experimental Medicine: Official Organ Wroclaw Medical University*, 23(2):289–293, 2014.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS 2012)*, pages 1097–1105, Lake Tahoe, USA, December 2012.

Zoya Bylinskii, Adrià Recasens, Ali Borji, Aude Oliva, Antonio Torralba, and Frédo Durand. Where should saliency models look next? In *European Conference on Computer Vision (ECCV 2016)*, pages 809–824, Amsterdam, Netherlands, October 2016. Springer.

Myles Hollander, Douglas A Wolfe, and Eric Chicken. *Nonparametric statistical methods*, volume 751. John Wiley & Sons, New York, USA, 2013.

Table 1: Surveyed saliency prediction models. {T: Top-down model, B: Bottom-up model}

#	Model Abbreviation		Description	Year	Cat.	Ben.
	Literature	Here				
1	SAMv (Saliency Attentive Model - VGG-16) [Cornia et al., 2018]	SVS, SVC	This algorithm uses the VGG-16 CNN as backbone and incorporates an Attentive Convolutional Long Short-Term Memory network (Attentive ConvLSTM) to predict eye fixations in the images, without handling a temporal sequence. This algorithm has a center prior component able to learn the center bias of the database.	2018	T	MIT
2	SAMr (Saliency Attentive Model - ResNet-50) [Cornia et al., 2018]	SRS, SRC	The same approach of the SAMv model, but using the ResNet-50 network as backbone.	2018	T	
3	BMS (Boolean Map based Saliency) [Zhang and Sclaroff, 2013]	BMS	This strategy characterizes an image by a set of binary images, which are created by randomly thresholding the image features maps in a whitened feature space. Given an input image, this algorithm uses the topological analysis of the Boolean maps to discover surrounding regions and estimate the saliency map.	2013	B	
4	LDS (Learning Discriminative Subspaces) [Fang et al., 2017]	LDS	This saliency map estimation is based on the learning of a set of discriminative subspaces. These subspaces have to perform the best in popping out targets and suppressing artifacts. LDS creates the candidate subspaces based on the principal component analysis.	2017	B	
5	FES (Fast and Efficient Saliency) [Tavakoli et al., 2011]	FES	This algorithm uses a center-surround approach to estimate saliency of local feature contrast in a Bayesian framework. It estimates the needed probability distributions using the sparse sampling and the kernel density estimation.	2011	B	
6	SWD (Spatially Weighted Dissimilarity Saliency) [Duan et al., 2011]	SWD	This algorithm is based on the integration of dissimilarities and spatial distance between image patches and the center bias. The spatial distance weighs the corresponding dissimilarities and the principal component analysis is adopted for dimension reduction. The center bias is addressed by a weighting mechanism.	2011	B	
7	UHM (Unsupervised Hierarchical Models) [Tavakoli and Laaksonen, 2016]	UHM	This unsupervised multi-scale hierarchical saliency model explores both local and global saliency concepts. This approach adopts independent subspace analysis (ISA), which is equivalent to a two-layer neural architecture. The algorithm obtains a hierarchical representation of the input, stacking the ISA networks together, as done in deep models.	2016	B	
8	IT (Itti's Saliency Model) [Itti et al., 1998] - implementation by Harel et al. [2006]	IT	This work is the pioneer of saliency prediction and it is considered the purely bottom-up model. It extracts low-level features using the local center-surround differences of intensity, color and orientation features at multiple spatial scales. Then, fusion of across-scale and normalization of these maps produces three conspicuity maps, which are combined to yield the saliency map.	1998	B	[Borji et al., 2015]
9	GB (Graph-Based Visual Saliency) [Harel et al., 2006]	GB	This model extracts the low-level features similar to IT. Then, it uses a Markov chain to construct a fully connected graph which joins all grid locations (nodes) for each feature map. The weight between two nodes is defined as the dissimilarity of the feature values and their spatial distance. The saliency map is estimated based on the equilibrium distribution.	2016	B	
10	SR (Spectral Residual Approach) [Hou and Zhang, 2007]	SR	This approach is independent of features, categories, or any prior knowledge about the objects. It conducts the saliency estimation by exploring the properties of the backgrounds. It evaluates the log-spectrum of an input image and extracts the spectral residual. Then, the spectral residual is transformed into the spatial domain to obtain the saliency map.	2007	B	
11	SS (Sparse Salient Regions) [Hou et al., 2012]	SS	The authors used the sign function of the Discrete Cosine Transform (DCT) of an image to generate a signature, containing mainly information about the image foreground. The algorithm explores this information to detect regions and generate saliency maps.	2012	B	
12	SIM (Saliency by Induction Mechanisms) [Murray et al., 2011]	SIC	This methodology consists in processing the visual stimuli according to the early human visual pathway (e.g. color-opponent, luminance channels and multi-scale decomposition). Afterward, the algorithm simulates the inhibition mechanisms of the visual cortex cells and integrates information at multiples scales by an inverse wavelet transform. It is based on the unified color induction model developed by Otazu et al. [2010]	2012	B	
13	ZH (Zhang et al. [2013] Detection of Abnormal Nuclei in Cervical Smear Images) Zhang et al. [2013]	ZH	This work explores the bottom-up attention mechanism and a target-driven strategy for abnormal cell detection in liquid-based cervical smear images. This model consists in extracting conspicuous image regions according to both direction and brightness features and then modulating this information by a high response area obtained by an annular template matching model. The authors designed the annular template from abnormal nuclei statistics.	2013	T	*

Table 2: Variants of the SAM model used in this work

Model	**CNN**	**Database**	**Sets**
SVS	VGG-16	SALICON	#{Train}: 10K, #{Validation}: 5K, #{Test}: 5K
SRS	ResNet-50		
SVC	VGG-16	CRICVA	#{Train}: 130, #{Validation}: 29, #{Test}: 73
SRC	ResNet-50		

Table 3: Ranking of the average execution time (seconds per image). CNN-based models are in bold. {Mat: Matlab, Py: Python}.

Method	SR	SS	FES	IT	ZH	GB	BMS	LDS	UHM	SWD	**SVC**	**SVS**	**SRC**	**SRS**	SIC
Code	Mat	Mat	Mat	Mat	Mat	Mat	Py	Mat	Mat	Mat	**Py**	**Py**	**Py**	**Py**	Mat
Time	0.01	0.02	0.22	0.23	0.50	0.51	0.63	0.96	1.36	2.38	**3.23**	**3.27**	**3.85**	**3.86**	10.03

Table 4: Agreement rate φ (Equation 1) between the $nr = 3$ salient ROIs and the most significant lesion in the image. The best results for each pathology are in bold. { M: Models, L: Lesions}

L **M**	ASC-US	LSIL	ASC-H	HSIL	Carcinoma
BMS	0.731	0.820	0.808	0.857	0.500
LDS	0.692	0.860	0.885	0.952	0.625
FES	0.865	0.860	0.962	**1.000**	0.750
SWD	0.808	0.900	0.923	**1.000**	0.625
UHM	0.712	0.810	0.846	**1.000**	0.625
IT	0.462	0.680	0.923	**1.000**	0.375
GB	0.730	0.870	0.923	0.904	0.500
SR	0.173	0.280	0.384	0.571	0.000
SS	0.385	0.450	0.731	0.762	0.250
SIC	0.153	0.390	0.730	0.571	0.375
ZH	0.846	0.930	0.884	0.761	0.375
SRS	0.865	0.860	0.846	0.952	0.625
SVS	0.731	0.840	0.885	0.952	0.500
SRC	**1.000**	0.990	**1.000**	**1.000**	**1.000**
SVC	0.981	**1.000**	**1.000**	**1.000**	0.875
Experts*	1.000	0.989	1.000	1.000	0.786
Baseline	0.365	0.570	0.653	0.904	0.250

* computed from all images of CRICVA. The experts reached $\varphi = 1$ for LSIL and Carcinoma with $nr = 4$ and $nr = 5$ salient ROIs, respectively.