ISSN 2464-4617 (print)
ISSN 2464-4625 (DVD)
Computer Science Research Notes
CSRN 3201
WSCG 2022 Proceedings

Parallel YOLO-based Model for Real-time Mitosis Counting

Robin Yancey
Department of Computer Science
University of California, Davis
USA (95616) Davis, CA
reyancey@ucdavis.edu

ABSTRACT

It is estimated that breast cancer incidences will increase by more than 50% by 2030 from 2011. Mitosis counting is one of the most commonly used methods of assessing the level of progression, and is a routine task for every patient diagnosed with invasive cancer. Although mitotic count is the strongest prognostic value, it is a tedious and subjective task with poor reproducibility, especially for non-experts. Object detection networks such as Faster RCNN have recently been adapted to medical applications to automatically localize regions of interest better than a CNN alone. However, the speed and accuracy of newer state-of-the-art models such as YOLO are now leaders in object detection, which had yet be applied to mitosis counting. Moreover, combining results of multiple YOLO versions run in parallel and increasing the size of the data in a way that is appropriate for the specific task are some of the other methods can be used to further improve the score overall. Using these techniques the highest F-scores of 0.95 and 0.96 on the MITOS-ATYPIA 2014 challenge and MITOS-ATYPIA 2012 challenge mitosis counting datasets are achieved, respectively.

Keywords

YOLO, deep learning, mitosis counting, breast cancer, histopathology, machine learning, real-time detection

1 Introduction

1.1 Mitotic Count & Issues

The Nottingham Grading System (NGS) is recommended by various professional bodies internationally (World Health Organization [WHO], American Joint Committee on Cancer [AJCC], European Union [EU], and the Royal College of Pathologists (UK RCPath) [17]. It says that tubule formation, nuclear pleomorphism, and mitotic index should each be rated from 1 to 3, with the final score ranging between 3 and 9. This is divided into three grades: Grade 1, score 3-5, well differentiated; Grade 2, score 6-7, moderately differentiated; and Grade 3, score 8-9, poorly differentiated [1].

When pathologists need to make this assessment of the tumor for mitotic count, they start by finding the region with the highest proliferative activity. The mitotic count is used to predict the aggressiveness of a tumor and is defined in a region from ten consecutive high-power fields (HPF) within a space of $2mm^2$. Variation in phase and slide preparation techniques make it possible to misdiagnose. They also often have a low density and can look different depending on whether the mitosis is in one of the four main phases: prophase, metaphase, anaphase, and telophase.

The shape of the cell itself differs significantly for each phase. For example, when in telophase it is split into to separate regions even though they are still one connected mitotic cell. Apoptotic cells (or cells going through preprogrammed cell death) and other scattered pieces of waste on the slides can also easily be confused with mitoses, having a similar dark spotty appearance. Further, mitotic nuclei often resemble many other hyperchromatic cellular bodies such as necrotic and non-dividing dense nuclei, making detection of mitosis more difficult on tissue [27]. The variation in the process of obtaining the slides using different scanners and different preparation techniques may also make distinguishing cells more exhausting. Worse yet, pathologists can get tired and it can make it harder to make proper judgement on slides when trained pathologists need to examine hundreds of high power fields (HPF) of histology images, in a short amount of time. Biopsies can take up to ten days before the patient receives results [18].

The increasing numbers of breast cancer incidences calls for a more time-and cost-efficient method of prognosis, which could later even help to provide care to impoverished regions. Automatic image analysis

has recently proven to be a possible solution, with inter-observer agreement when tested against the human judgement [28].

2 Related Work

2.1 Automatic and Machine Learning Methods

The use and development of automatic detection methods of mitosis counting have gradually been increasing since the end of the 20th century in order to make doctors' jobs easier and more efficient [16]. Due to the recent progress in digital medication, a large amount of of data has became available for use in the medical studies. Machine learning has helped to discover new characteristics of cancer mutations by sorting through more image data than humanly possible and simultaneously analyzing all of the millions of image pixels undetectable to the human eye. For example, in the field of histopathology, machine learning algorithms have been used for analysis of scanned slides to assist in tasks including diagnosis [9]. The use of computing in image analysis may reduce variability in interpretation, improve classification accuracy, and provide clinicians (or those in training) with a second opinion [9]. Existing methods use either handcrafted features captured by specific morphological, statistical, or textural attributes determined by a pathologist or features are automatically learned through the use of convolutional neural networks (CNN).

2.2 Deep Learning Methods

With the help of their strong self-learning qualities, deep learning networks, especially neural networks have also been heavily investigated in medical image processing [26]. CNN's have made a significant impact in machine learning for image classification, segmentation, object detection, and computer vision tasks [5]. Medical applications in particular, such as mitosis detection, cell nucleus segmentation and tissue classification tasks have also been popular tasks for CNN's. This is because pathological images are texture-like in nature, making them ideal task to learn with their shift invariance and pooling operations. Deep learning methods often outperform traditional methods such as use the of handcrafted features alone since feature extractors and can be classifiers simultaneously optimized [23] [33].

CNNs are well-suited to learn high level features such as mitotic figures, which is likely what made these methods winners of the ICPR2012 [22], ICPR 2014 [21], and AMIDA 2013 challenges [29] [18]. These well-known mitosis counting competitions were held at conferences and now are publicly available datasets commonly used for research, further discussed in 3.3.

Ciresan et al., won ICPR 2012 using deep max-pooling convolutional neural networks to classify each image pixel using a patch centered on the pixel as context. The simple CNN consisted of five convolutional layers with max pooling layer, and two fully connected layers [3]. A similar model has also been successfully used in detecting mitoses from the AMIDA13 challenge [12], where a multi-column neural network is used to classify image patches and generate the precise image descriptors.

2.3 Object Detection for Histopathological Image Analysis

On the other hand, it has now become well-known that a basic CNN alone lacks cell level supervision and often requires limiting the size of the input image. This is done so that sub-image features can be learned from localized regions of the image rather than the full image context consisting of multiple objects as well as non-objects (regions of interest). Moreover, object detection or precise localization is actually a more common task than full-image classification in medical applications. Consequently, deep learning methods designed originally for object detection such as R-CNN, Faster R-CNN, and Mask R-CNN have been applied to this to target specific frames from within the image which have been deduced from a ROI (Region of Interest).

For example, Lu et al. cascade detection algorithm based on segmentation and classification and reached 0.83 on the ICPR 2012 data set and 0.58 on the ICPR 2014 data set. It used a cascaded convolutional neural network based on UNet, which consisted of three parts: semantic segmentation and classification to detect mitosis. First UNet is used for segmentation to locate the candidate set of mitotic targets. Second, the cell nucleus is located by means of semantic segmentation to obtain accurate image blocks of mitotic and non-mitotic cells via a Vnet. Third, the cell image output block is used to train a CNN to do binary classification and this area is checked for mitosis [14]. Sebai et al. developed a multi-task deep learning framework for both object detection and instance segmentation tasks using Mask RCNN. First, it is used for segmentation to estimate the mitosis mask labels for the weakly annotated mitosis dataset. This produces the mitosis mask and bounding box labels for training another mitosis detection and instance segmentation model for mitosis detection on the other dataset [25]. They obtained an F-score of 0.86 on the 2012 ICPR dataset and an F-score of 0.48 on the 2014 ICPR dataset. Rao used Faster-RCNN to achieve the highest F-score of 0.96 when their model was trained and tested on all three challenge datasets above combined [18]. This is 6.22% more accurate than

the previous high score of 0.90 achieved by the model proposed by [24].

3 Materials & Methods

The first goal was to test the newer and more advanced object detection networks such as YOLOv3, YOLOv4-scaled, YOLOv5, and YOLOR for the mitosis counting task. The second goal was to try a number of different methods of increasing the size of the training data to further improve prediction accuracy. This included adding images from multiple scanners, combining the two different contest datasets, and multiple forms of image augmentation. Image augmentation helps to reduce overfitting while artificially enlarging the dataset [10]. The third goal was to try running the best YOLO models in parallel for improved accuracy since the training and inference times were exceptionally short compared to former methods.

3.1 Finding the Optimal Model & Configuration

Once the best augmentation combination was found, different numbers of epochs and versions were tested for each YOLO version model to find the optimal setup for speed and accuracy for this specific application. Once this was found, it was used for the further testing.

3.1.1 Combining YOLOv5m-p5 with YOLOR

Both YOLOv5m-p5 and YOLOR consistently produced the highest F-scores, but with different predictions. Also, YOLOR predicted its highest scores at quicker runtimes. Therefore, when the bounding box and confidence scores of each of the predictions made by YOLOv5m-p5 and YOLOR were averaged, the overall results and runtimes could be optimized. This helps to refine the results without loss in efficiency because each of the models can be trained and tested in parallel on a separate cloud GPU.

3.2 Increasing the Size of Training Data

3.2.1 Alternate Data Augmentation

Data augmentation can help add more samples, while increasing variability and diversity in the appearance of each mitotic region. This makes the model more robust towards new examples that show up in the test set with similar characteristics.

The types of augmentation tested included none, blur, noise, rotation, mosaic, brightness, and exposure, and each was compared to when no augmentation was applied. In each case a new training image was added to the dataset for each image augmentation and the unfiltered image was still included in the training set.

3.2.2 Multiple Scanners

To test for the potential change in accuracy by adding data of multiple scanners, training was done with the model on the images from each scanner alone and testing on images from the same scanner on which it was trained. It was then compared that to the results of training on both scanner data combined to see if adding data from the other scanner helps predict. Next, testing was done by alternating the training and testing data to test on data from another scanner besides the one of which it was trained on. These tests are also interesting or useful for realistic situations in which similar training data from the same scanner for the image is missing.

3.2.3 Multiple Databases

Then, to assess the effect of combining data from multiple databases to the training set, the ICPR 2012 training set was combined with the ICPR 2014 training set. If the predictions on the test images from one database alone are better when the model is trained with data from both then this helps us determine how overall useful this could be in real life cancer detection, as well. For example, we could continue to add data to the training set and keep updating the weights to get better predictions on any test set from a new patient.

3.3 Datasets & Preparation

3.3.1 Contest Datasets

The models were trained with two different open datasets from the International Conference on Pattern Recognition (ICPR) of breast cancer histopathology in 2012 [22] and 2014 [21] developed to address this challenging issue. The data is for mitosis counting in images stained with standard hematoxylin and eosin (H&E) dyes obtained from breast biopsies. The hematoxylin stains cell nuclei a purplish blue, while eosin stains the extracellular matrix and cytoplasm pink (and blood cells in red). The Aperio Scanscope XT and the Hamamatsu Nanozoomer 2.0-HT slide scanners have different resolution and are used to produce RGB high-power fields (HPFs). Annotations for the image coordinates of each mitosis are made by two senior pathologists, where if one disagrees a third will give the final say. The ICPR 2012 X40 resolution training dataset consists of 35 images with 226 mitotic cells. The original HPFs are of size 2084 $\times$ 2084 pixels. The ICPR 2014 X40 resolution training set provided consists of 1,136 frames containing a total of 749 labeled mitotic cells. Aperio images are sized 1539 $\times$

1376 pixels, and Hamamatsu images are 1663 × 1485 pixels.

Since the time of the contests both have been very commonly re-used among the research in this area thus far, so testing with these datasets will help compare the results to other published works.

Preprocessing The annotations of *official* test set for the ICPR contests is unavailable to the public so part of this training set was used for the test set. This is similar to what most other research groups (such as those referenced) have done, in order to be able to check the correctness of predictions by their developed framework. Here, the test set was selected randomly by extracting a set of images containing approximately the average number of mitosis in one slide from the provided training set. Also, similar to other groups referenced, the training set was artificially augmented to increase the density of mitosis and avoid class imbalance. Images needed to be cropped due to the small size and number of the mitosis compared to the very large size of the original HPF images.They were then expanded in order for the mitosis to be large enough for the smallest detectable size of the network aspect scales. These patches in mitotic regions were cropped into approximately 64 equal sized subsections from each HPF after being converted to JPEG. The bounding box coordinates were then created by adding 25 pixels in the upper left and lower right directions from the derived provided centroid coordinates.

Finally, each image is expanded to 416×416 and its coordinates are scaled upward accordingly. This is the appropriate input image size and scale for the YOLO network setup, which does internal data augmentation to rotate and resize the images internally. In order to provide consistency and better prediction accuracy (of both large and small objects), it is best for each image to have the same height and width. For the annotations, the first two coordinates are the centroid, while the second two coordinates are the width and height. So the new second coordinates were modified from the cropped images from the original training set provided and calculated by subtracting $x_2 - x_1$ and $y_2 - y_1$ and the first by adding half of that to the original, then they are divided by the height and width of the image.

The following calculation were used to generate the new coordinates for x and y:

- $x = (x_1 + (x_2 - x_1) \cdot 1/2) \cdot 1/w$
- $y = (y_1 + (y_2 - y_1) \cdot 1/2) \cdot 1/h$

3.3.2 Accuracy Calculation

The score for the tests here was calculated using the F-score, in the same way as the contestants. According to the contest evaluation criteria, a correct detection (true positive) is the one that lies within 32 pixels from this centroid of the ground truth mitosis. This is a harmonic mean of precision and recall (sensitivity), as described below.

$$F - score = \frac{2 \cdot (precision \cdot sensitivity)}{(precision + sensitivity)} \quad (1)$$

The precision measures how accurate the predictions are using the percentage of the correct predictions out of the total. It is calculated using the FP which represents the number of false positive predictions, and TP which is the number of true positive predictions, as shown below:

$$Precision = \frac{TP}{FP + TP} \quad (2)$$

The recall measures how well all the positives are found in the test set, where FN is the number of false negatives (those ground truths which were not detected), as shown below

$$Recall = \frac{TP}{FN + TP} \quad (3)$$

3.4 Software & Hardware

Google Colab cloud was used for the GPU access. The architecture is limited to NVIDIA P100 or T4, with RAM to 25 GB.

3.4.1 YOLOv5 and YOLOv3

The YOLOv5 [7] implementation used is written in the Ultralytics framework [6]. The repository also contains the model parameters and layers for the YOLOv3 network.

3.4.2 YOLOv4-Scaled

The official implementation of YOLOv4-Scaled [30] makes use of the Pytorch framework. Yolov4-csp from the yolov4-large branch for cloud GPU was used.

3.4.3 YOLOR

The official implementation of YOLOR [31] is on Github. The *yolor-p6.cfg* was used.

4 Results

4.0.1 Results of Data Augmentation

The Table 1 below shows some of the different augmentation techniques which were applied to the dataset and compared to the test without augmentation. The *Augmentation Type* column is the type of augmentation applied described above.

Augmentation Type	ICPR 2014 F-score	ICPR 2012 F-score
None	0.77	0.67
Exposure	0.94	0.96
Brightness	0.94	0.94
Blur	0.92	0.93

Table 1: Tests with YOLOR with Different Data Augmentation Techniques Applied to each Dataset

For each of the tests, the YOLOR model was trained for 120 epochs and a batch size of 8. There is a significant increase in the score for the dataset with any image augmentation that was tested. Over 3 trials on ICPR 2014, each type, mosaic blur, rotation, noise produced an F-score of 0.92, while a combination of techniques around 0.94. Exposure (and brightness) (changes of +/- 25 %) were consistently the highest on both datasets. For further testing only one augmentation technique was applied to the datasets since the combination of multiple augmentations did not significantly effect the results, besides increasing the training time.

Further, when no augmentation was applied and the training time was increased to the same amount as all of the other tests (the number of epochs were doubled), the F-score was still not as high as it was with augmentation; it only increased to 0.87 and 0.86 for ICPR 2012 and ICPR 2014, respectively. The training time for ICPR 2012 was around 0.32 hours for each test with augmented data, while the training time was around 1.4 hours for ICPR 2014.

4.0.2 Models & Versions

The YOLO version/model, and combined training and testing runtimes (in hours), are shown in the Model and Time columns, respectively, of the tables below. 120 epochs were evaluated in order to maximize the F-score.

YOLOv5 Each preset model scale and size of YOLOv5 was tested with a batch size of 16 for both the p5 and p6 versions, but the p5 version performed better. The resulting scores with each of the different model scales are shown in Tables 2 and 3 for ICPR 2014 and 2012 datasets, respectively.

Overall the *m* model consistently had a slightly higher F-score. For ICPR 2014 and 2012, F-scores of up to 0.95 and 0.96, respectively, were achieved using YOLOv5m with the augmented training data when trained for 120 epochs. The version *l* and *x* required far longer runtime without an increase in F-score for both datasets.

YOLOv4-Scaled Table 4 shows the tests with the scaled YOLOv4-csp with a batch size of 16 and a range of numbers of epochs. For the ICPR 2014 dataset it takes around 60 epochs for the F-score to reach its highest F-score of around 0.93. The runtime was similar than YOLOv5m for a lower F-score. However, it produced faster speed and higher accuracy than YOLOv3.

Model	Time (hrs)	Precision	Recall	F-score
YOLOv5s	1.52	0.93	0.95	0.94
YOLOv5m	1.53	0.95	0.95	0.95
YOLOv5l	2.29	0.90	0.94	0.92
YOLOv5x	4.09	0.95	0.94	0.94

Table 2: Tests with YOLOv5-p5 for 120 Epochs with ICPR 2014

Model	Time (hrs)	Precision	Recall	F-score
YOLOv5s	0.18	0.9	1.0	0.94
YOLOv5m	0.26	0.96	0.96	0.96
YOLOv5l	0.50	0.96	0.96	0.96
YOLOv5x	0.79	0.92	0.96	0.94

Table 3: Tests with YOLOv5-p5 for 120 Epochs with ICPR 2012

Epochs	Time (hrs)	Precision	Recall	F-score
20	0.49	0.90	0.84	0.87
40	0.95	0.89	0.93	0.91
60	1.44	0.90	0.95	0.93
80	1.91	0.89	0.95	0.92

Table 4: Tests with YOLOv4-Scaled with ICPR 2014

Epochs	Time (hrs)	Precision	Recall	F-score
20	1.35	0.91	0.77	0.84
60	1.63	0.90	0.92	0.91
120	2.69	0.90	0.92	0.91

Table 5: Tests with YOLOv3 with ICPR 2014

Epochs	Time (hrs)	Precision	Recall	F-score
20	0.09	0.82	1.0	0.90
60	0.25	0.82	1.0	0.92
120	0.50	0.81	1.0	0.91

Table 6: Tests with YOLOv3 with ICPR 2012

YOLOv3 Table 5 and Table 6 shows the tests with the scaled YOLOv3 model with a batch size of 16 and a range of numbers of epochs for each dateset. It takes around 60 epochs for the F-scores to reach their highest of around 0.91 and 0.92 for ICPR 2014 and 2012, respectively. Not only is the F-score much lower, but the runtime is much longer than the YOLOv5 and YOLOv4-scaled models for both datasets.

Epochs	Time (hrs)	Precision	Recall	F-score
30	0.84	0.94	0.91	0.93
60	1.34	0.91	0.89	0.90
120	2.64	0.92	1.0	0.92

Table 7: Tests with YOLOR on ICPR 2014

Epochs	Time (hrs)	Precision	Recall	F-score
30	0.16	0.92	0.92	0.92
60	0.32	0.97	0.93	0.95
120	0.65	0.93	0.94	0.94

Table 8: Tests with YOLOR on ICPR 2012

Time (hrs)	Precision	Recall	F-score
0.18	0.97	0.94	0.96

Table 9: Tests with YOLOv5-p5 combined with YOLOR for 30 Epochs with ICPR 2012

YOLOR Table 7 shows the tests with YOLOR when trained with increasing numbers of epochs, using a batch size of 8. This model has lower runtime and a higher F-score for any number of epochs. Only 30 epochs are required to reach the highest F-score for the model of 0.93.

Table 8 shows the tests with YOLOR when trained with increasing numbers of epochs, using a batch size of 8 for ICPR 2012. On this dataset this model provides a similar runtime and F-score to YOLOv5.

4.0.3 Combining YOLOv5m-p5 with YOLOR

The YOLOv5m-p5 model with YOLOR model run in parallel on separate GPUs at the same time, for only 30 epochs. Since the predictions made with YOLOv5m-p5 and with YOLOR were both very high yet both had different predictions, the combination of predictions was used to produce consistently the highest final scores and lowest runtimes (over multiple tests), as shown in Table 9. For example, YOLOv5m-p5 helped to eliminate false positives predicted by YOLOR, resulting in a higher precision than YOLOR alone and a lower runtime.

4.0.4 Combining Both Datasets

The Tables 10 and 11 below show the results of combining the ICPR 2012 and 2014 training datasets. For each test, YOLOR was trained for both 60 and 120 epochs with a batch size of 8.

By combining the training sets, some of the highest F-scores were obtained on both the ICPR 2012 test set and the 2014 test set. The runtime for training the ICPR 2014 dataset with YOLOR was also the lowest with the highest F-score. Although the training time was much higher, the highest F-score was obtained on the ICPR 2012 test dataset.

Epochs	Time (hrs)	Precision	Recall	F-score
60	0.44	0.93	0.96	0.94
120	0.87	0.90	0.98	0.94

Table 10: Tests with Combined Training Sets and YOLOR on ICPR 2014 Test Set

Epochs	Time (hrs)	Precision	Recall	F-score
60	1.5	0.93	0.96	0.92
120	3.2	0.89	1.0	0.96

Table 11: Tests with Combined Training Sets and YOLOR on ICPR 2012 Test Set

Train Dataset	Test Dataset	F-score
Aperio & Hamamatsu	Aperio	0.94
Aperio	Aperio	0.94
Aperio	Hamamatsu	0.81
Aperio & Hamamatsu	Hamamatsu	0.91
Hamamatsu	Hamamatsu	0.95
Hamamatsu	Aperio	0.95

Table 12: YOLOR with Different Combinations of Scanners for Train and Test Datasets ICPR 2014

4.0.5 Adding the Data from Another Scanner

As shown in the results in the Table 12 below, adding the Hamamatsu Nanozoomer 2.0-HT slide scanner data to the training set did not help in prediction in the tests on the Aperio Scanscope XT scanner data. However, when the training set consisted of the Hamamatsu combined with the Aperio or just consisted of the Hamamatsu the network predicted Hamamatsu scanner dataset alone, better. Interestingly, the network predicted the Hamamatsu slide scanner test set best when the Aperio data was removed from the training set. Therefore, when the network was trained on both scanner data it was not able to better predict the images from the test set consisting of one scanner alone. Adding the data from another scanner to the training set also significantly increases the runtime for training.

Interestingly, the Hamamatsu *only* training set helped to predict both the Aperio test set alone and the Hamamatsu alone test set the best, but only when it was trained with the Aperio images excluded. The Hamamatsu *only* training set actually produced a very slight increase in the F-score by 0.01.

5 Discussion

5.0.1 YOLOv3

It took the YOLOv3 model over an hour and a half to train to reach a high F-score of 0.91 on ICPR 2014, while other models reached up to 0.95. It also took longer and only reached 0.92 with ICPR 2012, which had a high of 0.96.

5.0.2 YOLOv3 vs. Newer Models

The neck of YOLOv4 and YOLOv5 is a PANet (Path Aggregation Network) which uses a more advanced technique called *path aggregation* to help preserve more of the spatial information in instance segmentation [13]. Since the complexity of the features in a CNN increases as the image passes through the network as spatial resolution of the image decreases, the pixel-level feature masks are extracted in layers far from the the deeper layers of the network.

On the other hand, an FPN is used in YOLOv3. This uses a top-down path through the CNN layers to extract and combine the semantically rich features with the precise localization information. This can be time-consuming for large objects or large networks because the information must be passed on through hundreds of layers. Wherefore, PANet takes a short-cut connection from both a bottom-up path as well as the top-down path originally taken by FPN. This makes for clean short cut paths from upper to lower layers, which are only around ten layers.

5.0.3 YOLOR

YOLOR [32] and YOLOv5 came extremely close in runtime and accuracy but YOLOv5 was not quite as fast. YOLOR improves upon the former models with an unified network architecture which combines the implicit and explicit knowledge in order to optimize the Kernel Space Alignment, multi-task learning, and prediction refinement for learning implicit features [32].

5.0.4 YOLOR & YOLOv5m-p5 in Parallel

Since both of these models achieved the top scores in the shortest runtimes, running them in parallel was superior. Both consistently produced the highest scores since each predicted slightly different. The YOLOv5m-p5 model helped to increase precision in YOLOR by eliminating false-positives.

5.0.5 Combining Datasets

The highest F-scores on the test sets were obtained by combining the two datasets, which proves the real-life potential for the use of deep learning frameworks for mitosis counting. If continuously adding data from other databases helps improve the prediction accuracy on any given test dataset, we would be able to keep updating the model weights by training or fine-tuning with new datasets for better results. The more variety in the training examples there are, the better the network learns the features of mitosis and is able to adapt to the slightly different features contained in the test set.

5.0.6 Combining Scanner Data

More tests need to be run with combinations of scanner data, since it was not necessarily the case that the combination produced better results than one alone. For example, better predictions were made with the Hamamatsu scanner data for each test.

5.0.7 Data Augmentation

For ICPR 2014 and 2012, the augmented data helped to increase the F-score by 0.17 and 0.28, respectively. Note, that when the original dataset (without augmentation) training time was increased to the same training time as adding the augmented data (eg. by doubling the number of epochs) the F-score only increased by about 0.1 (for each dataset). Therefore, the data augmentation up-front was critical to obtaining the very high F-score.

Differences in Augmentation Types The difference between the results of augmentation types is likely due to the fact that, YOLOv4 and on introduced new data augmentation techniques. Possibly, the combination of two mosaic augmentation applications compounded upon one another reduced the networks ability to learn from those examples because the region of interest became to small for the network parameters. For example, four images would have became eight different images, so the part of the bounding box would be cut off from most of the images.

The exposure (and brightness) likely had a slightly *better* result because HPF slides often have these types of variations in real-life. Parameters that can contribute to discrepancy in the representation of the HPF include scanner optics, camera sensors, and digital resolution, scan resolution, image viewer, monitor size, aspect ratio, and display resolution [8].

5.1 Comparison to Other Models

Accuracy The F-score is achieved is significantly higher than contest winners score of 0.356. Additionally, Table 13 shows a comparison to some of other top-performing groups. The YOLO-based model also achieved at least as high of an F-score as others who have trained and tested their models on test sets extracted from the ICPR public training dataset. [18]

Model	F-score	Inference Time (s)
CasNN [2]	0.482	4.62
Lightweight RCNN [12]	0.427	0.83
DeepMitosis (DeepDet+Seg+Ver) [11]	0.437	0.72
FRCNN [18]	0.503	0.58
MS-FRCNN [34]	0.507	0.55
Cascaded w/ U-net [14]	0.576	-
Faster R-CNN and deep CNNs [15]	0.691	-
Deep Cascaded + HC [24]	0.900	0.300
YOLOv5/R	0.950	0.110
MITOS-RCNN [18]	0.955	0.500

Table 13: Performance Comparison on ICPR 2014 Test Set

and [24] both trained their model on all 3 datasets, so the resulting F-scores are not necessarily comparable.

Further, most of the top performers in literature (such as those listed below) either used a version of a regional CNN (eg. RCNN) or a deep cascaded network (as the contest winners did).

Time Analysis The RCNN-based models above require multiple hours to train [34]. For example, the Lightweight RCNN approach still requires 11.4 hours to train. Without multiple GPUs some frameworks would even be infeasible, such as, [18] which requires 5 Tesla NVIDIA K80 GPUs and 3 parameter servers. This is better than the fully CNN-based approaches initially proposed in the ICPR contests which required days to weeks to train even with a GPU [4]. However, neither type of framework is appropriate for clinical use. On the other hand, the YOLO-based real-time model only takes 15-30 minutes to train on ICPR 2012 and around and hour on ICPR 2014. Not only is the training time far shorter than other models, but the inference time per full HPF is only about 0.11 which is significantly shorter than other models as shown in Table 13.

YOLO Models vs. RCNN Models YOLO [19], which stands for *You Only Look Once*), combines the CNN used to predict the bounding boxes and the class probabilities for each box, rather than separating the two (as in RCNN). Further, the later versions of YOLO used here improve further by including innovations such as Cross-Scaled-Partial (CSP) connections [30], a Path Aggregation Network (PANet) [13], and optimized Data Augmentation. Hence, this model can be many times faster than Faster RCNN, while still maintaining accuracy.

6 Conclusion

In this paper, YOLO-based [20] models were tested as a tool for mitosis counting. Multiple models and versions of YOLO were compared with different types of augmentation, and then top models were run in parallel for superior results. The model out-performed all earlier methods on the ICPR contest datasets when YOLOR [32] was run in parallel with YOLOv5m-p5 on two separate cloud GPUs with exposure augmentations added and the results were averaged. Additionally, it took a fraction of the time for both training and testing, making it clinically applicable.

REFERENCES

[1] Akinfenwa. Atanda, Mohammed. Imam, Ali. Umar, Ibrahim. Yusuf, and Shamsu. Bello. Audit of nottingham system grades assigned to breast cancer cases in a Teaching Hospital. *Annals of Tropical Pathology*, 8(2):104–107, 2017.

[2] Hao Chen, Qi Dou, Xi Wang, Jing Qin, and Pheng-Ann Heng. Mitosis detection in breast cancer histology images via deep cascaded networks. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 1160–1166. AAAI Press, 2016.

[3] Dan C. Ciresan, Alessandro Giusti, Luca Maria Gambardella, and Jürgen Schmidhuber. Mitosis detection in breast cancer histology images with deep neural networks. *Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention*, 16 Pt 2:411–8, 2013.

[4] Dan CireÅan, Alessandro Giusti, Luca Maria Gambardella, and JÃ$\frac{1}{4}$rgen Schmidhuber. Mitosis detection in breast cancer histology images with deep neural networks. volume 16, pages 411–8, 09 2013.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[6] Glenn Jocher, Alex Stoken, Jirka Borovec, NanoCode012, Ayush Chaurasia, TaoXie, Liu Changyu, Abhiram V, Laughing, tkianai, yxNONG, Adam Hogan, lorenzomammana, AlexWang1900, Jan Hajek, Laurentiu Diaconu, Marc, Yonghye Kwon, oleg, wanghaoyang0106, Yann Defretin, Aditya Lohia, ml5ah, Ben Milanko, Benjamin Fineran, Daniel Khromov, Ding Yiwei, Doug, Durgesh, and Francisco Ingham. ultralytics/yolov5: v5.0 - YOLOv5-P6 1280 models, AWS, Supervise.ly and YouTube integrations, April 2021.

[7] Glenn Jocher, Alex Stoken, Ayush Chaurasia, Jirka Borovec, NanoCode012, TaoXie, Yonghye Kwon, Kalen Michael, Liu Changyu, Jiacong Fang, Abhiram V, Laughing, tkianai, yxNONG, Piotr Skalski, Adam Hogan, Jebastin Nadar, imyhxy, Lorenzo Mammana, AlexWang1900, Cristi Fati, Diego Montes, Jan Hajek, Laurentiu Diaconu, Mai Thanh Minh, Marc, albinxavi, fatih, oleg, and wanghaoyang0106. ultralytics/yolov5: v6.0 - YOLOv5n 'Nano' models, Roboflow integration, TensorFlow export, OpenCV DNN support, October 2021.

[8] David Kim, Liron Pantanowitz, Peter SchÃ$\frac{1}{4}$ffler, DigVijay Yarlagadda, Orly Ardon, VictorE Reuter, Meera Hameed, DavidS Klimstra, and MatthewG Hanna. (re) defining the high-power field for digital pathology. *Journal of Pathology Informatics*, 11:33, 10 2020.

[9] Daisuke Komura and Shumpei Ishikawa. Machine learning methods for histopathological image analysis. *Computational and Structural Biotechnology Journal*, 16:34 – 42, 2018.

[10] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, page 1097â1105, Red Hook, NY, USA, 2012. Curran Associates Inc.

[11] Chao Li, Xinggang Wang, Wenyu Liu, and Longin Jan Latecki. Deepmitosis: Mitosis detection via deep detection, verification and segmentation networks. *Medical Image Analysis*, 45:121 – 133, 2018.

[12] Yuguang Li, Ezgi Mercan, Stevan Knezevitch, Joann G. Elmore, and Linda G. Shapiro. Efficient and accurate mitosis detection - a lightweight rcnn approach. In *ICPRAM*, 2018.

[13] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. *CoRR*, abs/1803.01534, 2018.

[14] Xi Lu, Zejun You, Miaomiao Sun, Jing Wu, and Zhihong Zhang. Breast cancer mitotic cell detection using cascade convolutional neural network with u-net. *Mathematical Biosciences and Engineering*, 18:673–695, 04 2021.

[15] Tahir Mahmood, Muhammad Arsalan, Muhammad Owais, Min Beom Lee, and Kang Ryoung Park. Artificial intelligence-based mitosis detection in breast cancer histopathology images using faster r-cnn and deep cnns. *Journal of Clinical Medicine*, 9(3), 2020.

[16] Xipeng Pan, Yinghua Lu, Rushi Lan, Zhenbing Liu, Zujun Qin, Huadeng Wang, and Zaiyi Liu. Mitosis detection techniques in h&e stained breast cancer pathological images: A comprehensive review. *Computers & Electrical Engineering*, 91:107038, 2021.

[17] Emad Rakha, Jorge Reis-Filho, Frederick Baehner, David Dabbs, Thomas Decker, Vincenzo Eusebi, Stephen Fox, Shu Ichihara, Jocelyne Jacquemier, Sr Lakhani, JosÃ© Palacios, Andrea Richardson, Stuart Schnitt, Fernando Schmitt, Puay-Hoon Tan, Gary Tse, Sunil Badve, and Ian Ellis. Breast cancer prognostic classification in the molecular era: the role of histological grade. *Breast cancer research : BCR*, 12:207, 07 2010.

[18] Siddhant Rao. MITOS-RCNN: A novel approach to mitotic figure detection in breast cancer histopathology images using region based convolutional neural networks. *CoRR*, abs/1807.01788, 2018.

[19] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. *CoRR*, abs/1506.02640, 2015.

[20] Joseph Redmon and Ali Farhadi. YOLO9000: better, faster, stronger. *CoRR*, abs/1612.08242, 2016.

[21] L. Roux. Mitos-atypia-14 - grand challenge, 2014.

[22] Ludovic. Roux, Daniel. Racoceanu, Nicolas. LomÃ©nie, Maria. Kulikova, Humayun. Irshad, Jacques. Klossa, FrÃ©dÃ©rique. Capron, Catherine. Genestie, Gilles. Naour, and Metin. Gurcan. Mitosis detection in breast cancer histological images An ICPR 2012 contest. *Journal of Pathology Informatics*, 4(1):8, 2013.

[23] Monjoy Saha, Chandan Chakraborty, and Daniel Racoceanu. Efficient deep learning model for mitosis detection using breast histopathology images. *Computerized Medical Imaging and Graphics*, 64, 12 2017.

[24] Monjoy Saha, Chandan Chakraborty, and Daniel Racoceanu. Efficient deep learning model for mitosis detection using breast histopathology images. *Computerized Medical Imaging and Graphics*, 64, 12 2017.

[25] Meriem Sebai, Xinggang Wang, and Tianjiang Wang. Maskmitosis: a deep learning framework for fully supervised, weakly supervised, and unsupervised mitosis detection in histopathology images. *Medical & Biological Engineering & Computing*, 58, 05 2020.

[26] Nida Shahid, Tim Rappon, and Whitney Berta. Applications of artificial neural networks in health care organizational decision-making: A scoping review. *PLOS ONE*, 14:1–22, 02 2019.

[27] Mitko Veta, Yujing J. Heng, Nikolas Stathonikos, Babak Ehteshami Bejnordi, Francisco Beca, Thomas Wollmann, Karl Rohr, Manan A. Shah, Dayong Wang, Mikaël Rousson, Martin Hedlund, David Tellez, Francesco Ciompi, Erwan Zerhouni, David Lanyi, Matheus Palhares Viana, Vassili A. Kovalev, Vitali Liauchuk, Hady Ahmady Phoulady, Talha Qaiser, Simon Graham, Nasir M. Rajpoot, Erik Sjöblom, Jesper Molin, Kyunghyun Paeng, Sangheum Hwang, Sunggyun Park, Zhipeng Jia, Eric I-Chao Chang, Yan Xu, Andrew H. Beck, Paul J. van Diest, and Josien P. W. Pluim. Predicting breast tumor proliferation from wholeâslide images: The tupac16 challenge. *Medical Image Analysis*, 54:111â121, 2019.

[28] Mitko Veta, Paul J. van Diest, Mehdi Jiwa, Shaimaa Al-Janabi, and Josien P. W. Pluim. Mitosis counting in breast cancer: Object-level interobserver agreement and comparison to an automatic method. *PLoS ONE*, 11(8), 8 2016.

[29] Mitko Veta, Paul J. van Diest, Stefan M. Willems, Haibo Wang, Anant Madabhushi, Angel Cruz-Roa, Fabio A. González, Anders Boesen Lindbo Larsen, Jacob S. Vestergaard, Anders B. Dahl, Dan C. Ciresan, Jürgen Schmidhuber, Alessandro Giusti, Luca Maria Gambardella, F. Boray Tek, Thomas Walter, Ching-Wei Wang, Satoshi Kondo, Bogdan J. Matuszewski, Frédéric Precioso, Violet Snell, Josef Kittler, Teófilo Emídio de Campos, Adnan Mujahid Khan, Nasir M. Rajpoot, Evdokia Arkoumani, Miangela M. Lacle, Max A. Viergever, and Josien P. W. Pluim. Assessment of algorithms for mitosis detection in breast cancer histopathology images. *CoRR*, abs/1411.5825, 2014.

[30] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Scaled-YOLOv4: Scaling cross stage partial network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13029–13038, June 2021.

[31] Chien-Yao Wang, I-Hau Yeh, and Hong-Yuan Mark Liao. You only learn one representation: Unified network for multiple tasks. *arXiv preprint arXiv:2105.04206*, 2021.

[32] Chien-Yao Wang, I-Hau Yeh, and Hong-Yuan Mark Liao. You only learn one representation: Unified network for multiple tasks. *CoRR*, abs/2105.04206, 2021.

[33] Haibo Wang, Angel Cruz-Roa, Ajay Basavanhally, Hannah Gilmore, Natalie Shih, Mike Feldman, John Tomaszewski, Fabio GonzÃ¡lez, and Anant Madabhushi. Mitosis detection in breast cancer pathology images by combining handcrafted and convolutional neural network features. *Journal of Medical Imaging*, 1:1–8, 12 2014.

[34] Robin Elizabeth Yancey. Multi-stream faster rcnn for mitosis counting in breast cancer images, 2020.