Supervised detection of Alternaria solani on ultra-high-resolution modified RGB UAV images

Jana Wieme, Sam Leroux, Simon Cool, Jan Pieters, Wouter Maes

Jana Wieme, Sam Leroux, Simon Cool, Jan Pieters, Wouter Maes, "Supervised detection of Alternaria solani on ultra-high-resolution modified RGB UAV images," Proc. SPIE 12262, Remote Sensing for Agriculture, Ecosystems, and Hydrology XXIV, 122620U (28 October 2022); doi: 10.1117/12.2638911

Event: SPIE Remote Sensing, 2022, Berlin, Germany

Supervised detection of Alternaria solani on ultra-high-resolution modified RGB UAV images

Jana Wieme[a, b, c], Sam Leroux[b], Simon Cool[c], Jan Pieters[a], and Wouter Maes[a]

[a]Department of Plants and Crops - Ghent University, Coupure Links 653, 9000 Ghent, Belgium
[b]IDLab, Department of Information Technology at Ghent University - imec,
Technologiepark-Zwijnaarde 15, 9052 Ghent, Belgium
[c]Technology and Food Unit - ILVO, Burg. van Gansberghelaan 115, 9820 Merelbeke, Belgium

ABSTRACT

Potato cultivation is regularly affected by *Alternaria solani*, a destructive foliar pathogen causing early blight, a premature defoliation of potato plants resulting in yield losses. Currently, Alternaria is treated through preventive application of chemical crop protection productions, following warnings based on weather predictions and visual observations. Automatic detection could make the mapping of early blight more accurate, reducing production losses and application of crop protection products. Current research explores the potential of deep learning of high resolution imagery within precision agriculture, mainly using supervised learning. However, available datasets are often limited in size and variation, which reduces the robustness of the developed models. Here, we present a convolutional network to detect Alternaria and evaluate the influence of sampling size, sampling balance and sampling accuracy on the model performance. These analyses are based on ultra-high-resolution datasets of modified RGB cameras obtained with unmanned aerial vehicles (UAV) and collected over experimental in-field Alternaria trials. By using this varied dataset instead of a single-time dataset, higher accuracies are achieved. The method is relatively robust for imbalances of the training dataset. Further, we show that labeling quality plays a role, but that an error of up of to 20% of labeling is acceptable for good results. In conclusion, extra variability leads to more robust disease detection, desirable for in-field application.

Keywords: Alternaria solani, potato crops, supervised deep learning, UAV, ultra-high resolution, modified RGB, labeling quality

1. INTRODUCTION

Although potato is the third-most important food crop in terms of global consumption - with an annual production of 360 million tons,[1] potato yield is relaively unstable and fluctuating, due to diseases, pests and abiotic stress. One of these stressors is *Alternaria solani*, worldwide the second-most destructive foliar pathogen in potato cultivation causing premature defoliation and substantial yield losses. Nowadays, a cocktail of chemical crop protection products is used on regular intervals to prevent the outbreak of the disease, following early warnings based on weather forecasts and visual inspections. These chemical products have large economic and environmental impacts, which could be reduced if more automated symptom detection were possible.[2]

In the recent past, applications of automatic detection of deviating patterns using deep neural networks (DNNs) emerged in a multitude of sectors.[3] The strength of DNNs for agricultural application is the ability to learn features directly from data, enabling them to automatically capture the variability in cultivars, growth stages, growth conditions, and others. Consequently, a large number of studies within the precision agriculture

Remote Sensing for Agriculture, Ecosystems, and Hydrology XXIV, edited by
Christopher M. U. Neale, Antonino Maltese, Proc. of SPIE Vol. 12262,
122620U · © 2022 SPIE · 0277-786X · doi: 10.1117/12.2638911

context used this deep-learning technique in recent years.[4,5] The use of DNNs also emerges in research within potato cultivation in applications of automated disease detection on tubers,[6,7] leaves,[8,9] and the entire plants.[10]

However, deep learning also has it shortcomings, especially concerning the high need of data. This is reflected in the high presence of deep-learning techniques in research, but their rather limited adoption in practice.[5] The importance of a large and diverse dataset should not be underestimated. The successful use of deep learning in various fields is driven partially by the availability of large-scale, labeled and public datasets, such as ImageNet,[11] which are not readily available within the agricultural sector. Besides, to exploit the strength of DNNs and obtain a robust model, the dataset must be sufficiently large to effectively cover the intrinsic variability of agricultural data.[12] On top of that, current state-of-the-art models are mostly based on supervised learning, which, in addition to a rich dataset, also requires correct labeling. As labeling is a cost- and time-consuming process, it is a bottleneck in supervised deep learning models.[5]

The contributions of this paper are twofold. First, we use state-of-the-art techniques to develop a binary classification model, classifying small parts of images as Alternaria (1) or not (0). We use data collected on field infection trials during two growing seasons in two different fields and with two different cultivars. Second, we aim at investigating how some characteristics of a labeled dataset are influencing the results of the binary classification model. To this end, three research questions are addressed:

i What is the minimal and optimal size of the labeled dataset?

ii How important is the balance of classes in a dataset for binary classification?

iii How do corrupt labels affect model quality?

The remainder of this paper is organized as follows. Section 2 describes the data acquisition, pre-processing workflow, labeling strategy and model. Section 3 presents the result of the binary classification model, as well as gives insights into the characteristics of the labeled dataset and their impact on the model. Finally, Section 4 concludes this paper and provides directions for further research.

2. MATERIALS AND METHODS

2.1 Field trial and image acquisition

During two sequential growing seasons (2019 and 2020), a field trial of the method based on Van De Vijver et al.[13] was established. The field trial was located at ILVO (Lemberge, Merelbeke, Belgium; 50.986535, 3.773013). In 2019, four plots of 3 m × 3 m were inoculated with a concentrations of *Alternaria solani* spores of 3×10^3 per ml on July 30. In 2020, four plots of 4 m × 3 m were inoculated with a concentration of 2×10^3 per ml and four plots of the same size were inoculated with a lower concentration of 0.5×10^3 per ml on July 6. The cultivar was Spunta in 2019 and Bintje in 2020. The plots were visually monitored on a daily basis, starting from the day after inoculation. Data acquisition took place on a regular basis at 2-3 day intervals according to weather conditions. The data used in this work were collected on August 5, 8 and 12, 2019 and July 11, 13 and 15, 2020, respectively 6, 9 and 13 and 5, 7 and 9 after inoculation. Figure 1 presents an overview image of the field trial of 2020, showing the eight infected plots.

Images were captured with a Sony Alpha 7R III modified RGB camera equipped with a 135 mm lens (Zeiss), installed on a DJI M600 PRO unmanned aerial vehicle (UAV) and at 10m flight height. A gimbal (DJI Ronin-MX)

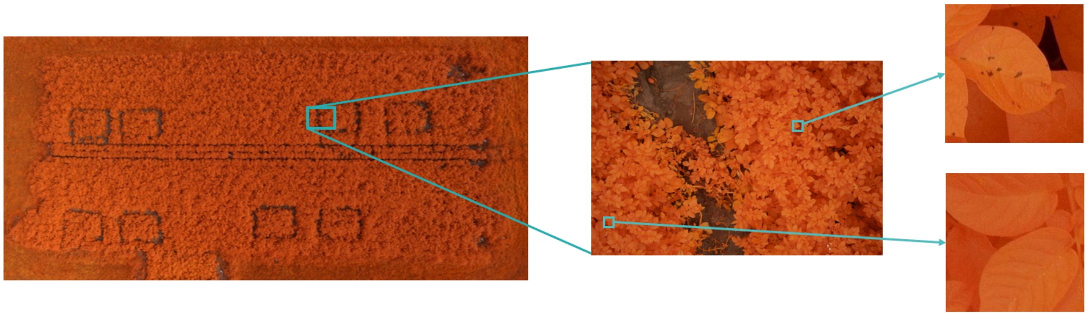

Figure 1: *Left*: overview image of the field trial in July 2020 with 8 plots of 4 m × 3 m; *Middle*: example of modified RGB image; *Right*: Alternaria (top) and other patch (bottom).

was used to stabilize the camera and keep it in nadir-looking position. A modified RGB camera (red band replaced by a NIR band) was used instead of a regular RGB camera. The advantage of this type of camera is that the lesions caused by structural damage from *Alternaria solani* are more visible in the near-infrared spectrum.[13]

Each image covers an area of about 2.5 m × 1.5 m and has a resolution of 0.3 mm. In Figure 1, an example image is given of the area covered by an individual image. Based on the UAV logs, coordinates were assigned to all images.

2.2 Pre-processing and labeling workflow

Labeling is performed on image patches of 256 × 256 pixels, as shown on the right of Figure 1. For this purpose, the full images (7952 × 5304 pixels) were cut into grid pattern. From the available data of both years, a total of 12000 patches were selected for the labeling campaign, with equal distribution over the six measurement days and plots.

To obtain sufficient variation, as well as to avoid selecting irrelevant patches, a pre-processing selection pipeline was set up. In this pipeline, a predefined number of random images were selected. Proportionally, more images were selected situated in Alternaria plots (72%), than in the control parts of the field, as images from Alternaria plots also include sections without Alternaria. From these selected images, four subimages (512 × 512 pixels) are selected which meet the following criteria: (i) sufficient sharpness, calculated using the perceptual-based no-reference objective image sharpness metric (CPBD);[14] (ii) lightness threshold to avoid shadow areas, based on the lightness value in hue, saturation, and lightness (HSL); and (iii) sufficient vegetation to avoid background images, based on a modified vegetation index (green leaf index (GLI)[15] adapted to modified RGB by swapping the green and red band in the formula). Four selected zones of each image are then further divided in four patches of 256 × 256 pixels.

Labelbox[16] was used to label these 12000 patches, with a class (Alternaria, Healthy, Dubious or Background) assigned to each patch by three different labelers. Afterwards, only patches equally labeled by all labelers, so having 100% consensus, were taken into account for further processing. For the binary classification problem, this was reduced to two classes: Alternaria (1) or not (0). Patches with label "Healthy" or "Background" were both placed in category 0, patches with label "Dubious" were disregarded. Afterwards, two small extra subsets were labeled in a separate Labelbox project to include even more variation. The first subset contains 466 patches picked from the data that did not meet the desired criteria of the pre-processing pipeline. The second subset contains 1221 patches from images taken from the rest of the field (not in the Alternaria plots). After processing the various labels, a dataset of 3640 patches with and 2486 patches without Alternaria was extracted with 100% consensus among the labelers.

2.3 Model and experimental datasets

The labeled dataset is split into train/test data with ratio of 85/15. For training, the whole training set (0: 2112 patches, 1: 3093 patches) is used, but the test set is limited to a balanced version (374 patches of each category) to be able to compare the different experiments in Section 3. We trained a convolutional neural network (CNN) with five convolutional layers (including max pooling) and five fully connected layers, each followed by a leaky rectified linear unit (ReLU) activation, except the last one. The last fully connected layer is followed by a Sigmoid activation function to map the resulting output values in the range of $[0, 1]$. The PyTorch framework[17] is used to train the model on an NVIDIA GeForce RTX 3070 GPU with CUDA 11.[18] The initial learning rate was 0.001 and Adam[19] is used to optimize the gradient descent with L2 regularization, during 100 epochs.

3. RESULTS AND DISCUSSION

Table 1 summarizes the precision, recall, F1 score and accuracy of the described binary classification model averaged over 20 runs. In Table 2, the corresponding normalized confusion matrix is given. These results are based on the balanced test set of 748 labeled patches of different days in 2019 and 2020, averaged over 20 runs.

In the following subsections, the experiments and results concerning the three characteristics of the dataset associated with our three research questions, are described.

Table 1: Resulting metrics for the binary classification of patches averaged over 20 runs. TP: true positives, FP: false positives, TN: true negatives, FN: false negatives

Metric	Definition	Value
Precision	$\frac{TP}{TP+FP}$	0.876
Recall	$\frac{TP}{TP+FN}$	0.921
F1 score	$\frac{TP}{TP+0.5\times(FP+FN)}$	0.896
Accuracy	$\frac{TP+TN}{TP+FP+TN+FN}$	0.892

Table 2: Resulting normalized confusion matrix on the balanced test set averaged over 20 runs.

		Predicted	
		Other	Alternaria
True	Other	0.87	0.13
	Alternaria	0.08	0.92

3.1 Impact of the amount of data

As mentioned in Section 1, deep learning is known for its need of large labeled datasets. In this experiment, we focus on the required size of the labeled datasets. Therefore, the size of the training data is ranged from 200 to

4200 patches, always balanced between both categories. It always starts with a random selection of 200 patches, and adds 200 extra random patches in each step. The experiment is averaged over 12 runs.

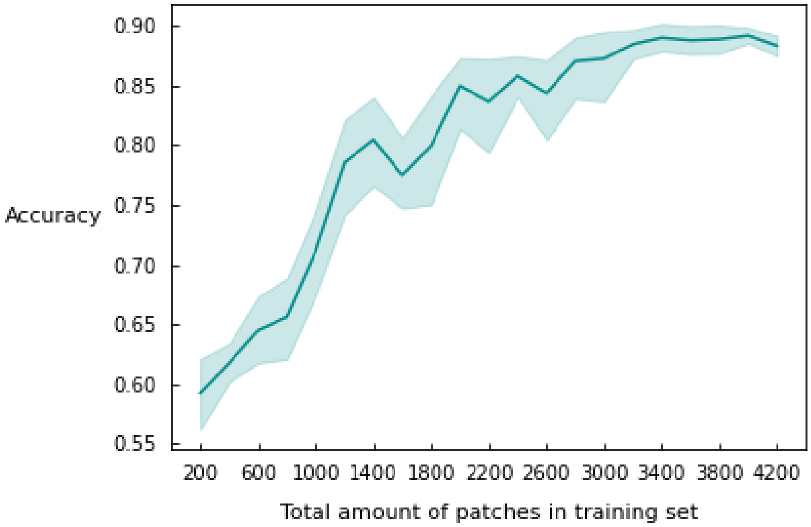

Figure 2: Graphical representation of the effect of the amount of labeled patches on the accuracy with 95% confidence intervals over 12 runs.

Figure 2 displays the result of the experiment. As expected, the accuracy increases when a larger amount of patches is used. The increase in accuracy is most noticeable for smaller labeled datasets and converges slightly around 3200 patches. This shows that adding additional data does indeed help to improve the model, although it will not help endlessly if no other things are changed. Further optimization of the variation in the dataset might well improve this. Consequently, depending on the application and model, it is important to use a sufficiently large dataset.

The confidence intervals also show that the variation in training data also influences the results. As the selection is made randomly out of 3093 Alternaria and 2112 other patches, every run is using a different training set. Of course, the higher the amount of patches, the less pronounced this random selection effect is, since by the end most of the data has been used and is therefore quite similar in all runs.

3.2 Impact of the balance in the dataset

The balance between the classes in the dataset is commonly regarded as an important characteristic of a good dataset. As this is an application of binary classification, the training set is varied from 0 to 100% patches labeled as Alternaria, in 5% increments. The total amount of patches remains constant and is limited to 2100. The test set is perfectly balanced. The experiment is averaged over 12 runs.

Figure 3 shows that at the extreme percentages (0 and 100), the accuracy is much lower as the model is only trained on one class. The outcome however indicates that a perfectly balanced dataset is not required to give good results, as ratios between 25 and 75% Alternaria patches generate similar results in overall accuracy.

3.3 Sensitivity to labeling quality

A third important characteristic of a good dataset for a supervised deep learning model concerns the quality of the labels. For these experiments, all patches were independently labeled three times by different labelers. Only

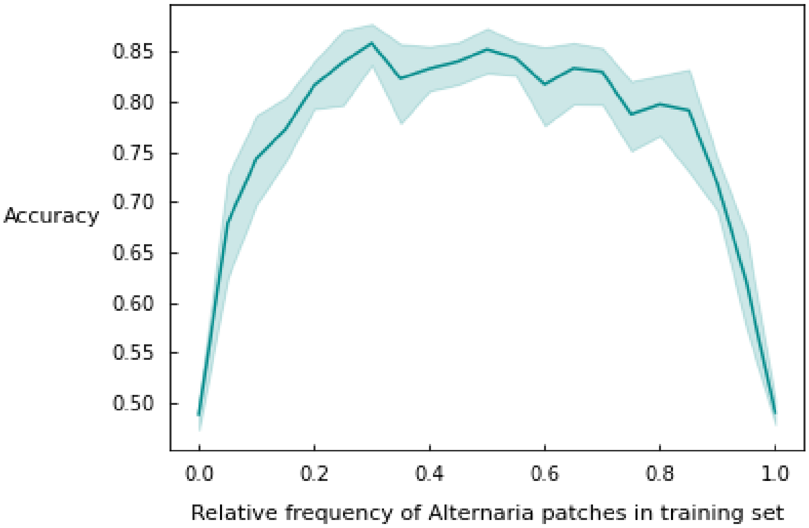

Figure 3: Graphical representation of the effect of the percentage Alternaria patches in the training set on the accuracy with 95% confidence intervals over 12 runs.

patches that were assigned the same class by all labelers, and thus have 100% consensus, were taken into account. Here, we want to assess the robustness of the method for dealing with wrongfully assigned labels.

The experiment tests the sensitivity of the model to correct labels through a varying amount of corrupted labels. The total training set continuously consists of 4200 patches, with a 50/50 balance. However, each step, an additional 5% of labels per category are inverted and thus incorrectly provided to the model during the training process. The experiment is averaged over 10 runs.

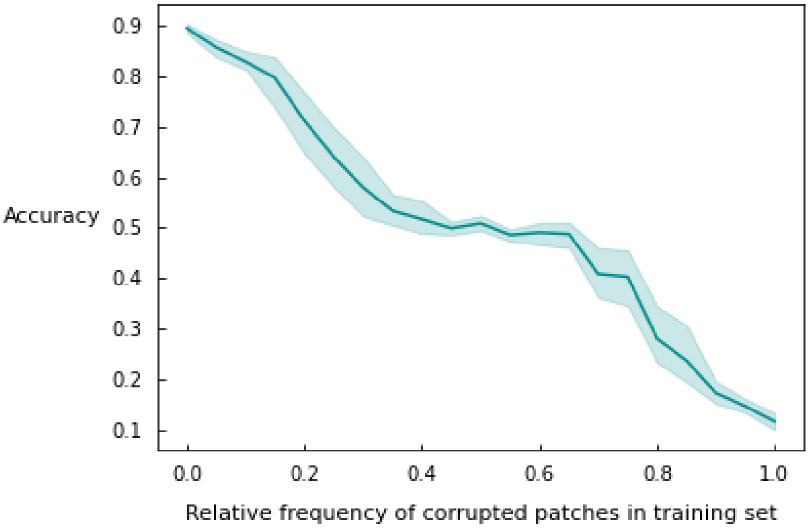

Figure 4: Graphical representation of the effect of the percentage corrupted patches per category in the training set on the accuracy with 95% confidence intervals over 10 runs.

The decreasing curve in 4 indicates that the accuracy is strongly affected by the quality of the labels and, as expected, that the model appears to be very sensitive to corrupted labels. When 15% of the labels per category are found to be wrong, there is a 10% drop in accuracy as compared to a fully correctly labeled set. From 20% onward, there is a steeper drop in accuracy. When about half of the labels are wrong, the model is no longer capable of achieving accuracies above 50%; once past half corrupt labels, the model actually learns the opposite.

4. CONCLUSIONS AND FUTURE WORK

The focus in this paper is on the potential of DNNs for supervised detection of *Alternaria solani* on ultra-high-resolution modified RGB images. To this end, a convolutional neural network was proposed to conduct binary classification on image patches of two consecutive growing seasons. The test results shows that the precision and recall are 87.6% and 92.1% respectively. These high scores are presumably due to the fact that the training dataset consisted of very diverse conditions by using data from different days in the disease progress, different growing seasons, different cultivars and data captured in different weather conditions.

Furthermore, we assessed the importance of three dataset characteristics. Quantitative experiments demonstrated that:

(1) the amount of labeled patches in the training set has a direct impact on the accuracy. In general, a large number of training data is required (in our case, 3200 points) to achieve good results, but a further increase does not lead to a significant improvement of the model performance.

(2) it is not essential to have a perfect 50/50 balance of labeled data. In fact, imbalances of the datasets between 25 and 75% still showed good results in our study.

(3) it is important to be aware of the quality of the labels. A trade-off can be made between the effort and time required for accurate labeling, relative to robustness. However, there is some room for error: an error margin of up to 15% causes a 10% drop in accuracy.

Future work includes a study of the overall analysis of limiting the training data to certain measurement days in order to check how well the model can adapt to the variation in data in terms of disease progression, growing season or weather conditions. Another interesting future research path includes analyzing quality aspects of the labeling campaign, such as the percentage of mislabeled patches per labeler as this work showed that quality is significant. In addition, it may be of interest to develop techniques that require less extensive labeling, such as active labeling or weakly supervision; this will remain the focus of future work.

ACKNOWLEDGMENTS

J.W. is funded by grant 1SE3921N of Research Foundation Flanders (FWO). Experimental data was gathered within the Proeftuin Smart Farming 4.0 project (180503) within the Industry 4.0 Living Labs with funding from Flanders innovation & entrepreneurship (VLAIO, Belgium) and in the Horizon 2020 project SmartAgriHubs - Connecting the dots to unleash the innovation potential for digital transformation of the European agrifood sector with funding from the European Union under grant agreement No. 818182.

REFERENCES

[1] FAO, "Crops and livestock products," (2022).

[2] Afzaal, H., Farooque, A. A., Schumann, A. W., Hussain, N., McKenzie-Gopsill, A., Esau, T., Abbas, F., and Acharya, B., "Detection of a potato disease (early blight) using artificial intelligence," *Remote Sensing* **13**(3) (2021).

[3] LeCun, Y., Bengio, Y., and Hinton, G., "Deep learning," *nature* **521**(7553), 436–444 (2015).

[4] Barbedo, J., "A review on the main challenges in automatic plant disease identification based on visible range images," *Biosystems Engineering* **144**, 52–60 (04 2016).

[5] Kamilaris, A. and Prenafeta-Boldú, F. X., "Deep learning in agriculture: A survey," *Computers and Electronics in Agriculture* **147**, 70 – 90 (2018).

[6] Oppenheim, D., Shani, G., Erlich, O., and Tsror, L., "Using deep learning for image-based potato tuber disease detection," *Phytopathology* **109** (12 2018).

[7] Arshaghi, A., Ashourin, M., and Ghabeli, L., "Detection and classification of potato diseases using a new convolution neural network architecture," *Traitement du Signal* **38** (12 2021).

[8] Mohanty, S. P., Hughes, D. P., and Salathé, M., "Using deep learning for image-based plant disease detection," *Frontiers in Plant Science* **7**, 1419 (2016).

[9] Rashid, J., Khan, I., Ali, G., Almotiri, S. H., AlGhamdi, M. A., and Masood, K., "Multi-level deep learning model for potato leaf disease recognition," *Electronics* **10**(17) (2021).

[10] Afonso, M., Blok, P. M., Polder, G., van der Wolf, J. M., and Kamp, J., "Blackleg detection in potato plants using convolutional neural networks," *IFAC papers online* **52**(30) (2019).

[11] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L., "Imagenet: A large-scale hierarchical image database," in [*2009 IEEE Conference on Computer Vision and Pattern Recognition*], 248–255 (2009).

[12] Liu, J. and Wang, X., "Plant diseases and pests detection based on deep learning: a review," *Plant Methods* **17** (02 2021).

[13] Van de Vijver, R., Mertens, K., Heungens, K., Somers, B., Nuyttens, D., Borra-Serrano, I., Lootens, P., Roldan-Ruiz, I., Vangeyte, J., and Saeys, W., "In-field detection of altemaria solani in potato crops using hyperspectral imaging," *Computers and electronics in agriculture* **168** (JAN 2020).

[14] Bhor, P., Gargote, R., Vhorkate, R., Yawle, P. R. U., and Bairagi, V. K., "A no reference image blur detection using cumulative probability blur detection (cpbd) metric," *International Journal of Science and Modern Engineering* **1**(5) (2013).

[15] Louhaichi, M., Borman, M., and Johnson, D., "Spatially located platform and aerial photography for documentation of grazing impacts on wheat," *Geocarto International* **16** (04 2001).

[16] "Labelbox." https://labelbox.com (2022).

[17] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S., "Pytorch: An imperative style, high-performance deep learning library," in [*Advances in Neural Information Processing Systems 32*], Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., eds., 8024–8035, Curran Associates, Inc. (2019).

[18] Nickolls, J. R., Buck, I., Garland, M., and Skadron, K., "Scalable parallel programming with cuda," *2008 IEEE Hot Chips 20 Symposium (HCS)* , 1–2 (2008).

[19] Kingma, D. P. and Ba, J., "Adam: A method for stochastic optimization.," in [*ICLR (Poster)*], Bengio, Y. and LeCun, Y., eds. (2015).