İSTANBUL TECHNICAL UNIVERSITY
FACULTY OF COMPUTER AND INFORMATICS

PERFORMANCE ANALYSIS OF DEEPFAKE DETECTION USING DIFFERENT FACIAL REGIONS

Graduation Project Interim Report

Egehan Orta
150160124

Department: Computer Engineering
Division: Computer Engineering

Advisor: Prof. Dr. Hazım Kemal EKENEL

February 2021

Statement of Authenticity

I/we hereby declare that in this study

1. all the content influenced from external references are cited clearly and in detail,
2. and all the remaining sections, especially the theoretical studies and implemented software/hardware that constitute the fundamental essence of this study is originated by my/our individual authenticity

İstanbul, February 2021

Egehan Orta

PERFORMANCE ANALYSIS OF DEEPFAKE DETECTION USING DIFFERENT FACIAL REGIONS

(SUMMARY)

Computer Vision is one of today's the most popular fields of study of Computer Science. Thanks to the opportunities by developing technology and widespread use of technology, having more processing power and visual data have led to positive impact on development speed of Computer Vision in recent years. Among the subjects of Computer Vision, there are subjects such as face recognition systems, object tracking systems and visual manipulations, whose usage has increased in recent years. Deepfake, one of the subheadings of visual manipulation, is undoubtedly in the top ranks of the most popular topics thanks to social media.

Deepfake is visual data that created using the deep learning to replace human faces inside the video or image with another face within the video or image. Deepfake, whose name comes from the combination of "deep learning" and "fake", is still an active field of study of Computer Vision. Successful models developed by people working in the field of Computer Vision, have an ability to create good deepfakes such that people cannot understood whether the created visual data is real or fake.

Deepfake process, which is easily performed even by using mobile phone applications within the scope of the possibilities offered by developing technology, has been recognized by the whole world by being included in the camera filters of social media applications. Deepfake images, which have a common usage as entertainment, pose a great threat in terms of spreading false information. Since it is difficult to detect whether data is deepfake or not, and there exist a threat of using deepfake to harm people or communities, people working in field of Computer Vision have started to building models to detect deepfake. Although great success has been achieved in deepfake detection, the increase in detection difficulty is inevitable with the developments in the deepfake field. For this reason, studies are continuing to increase performance of deepfake detectors by using the weaknesses of deepfake technology. This situation has also been the focus of attention in this study, and considering that it will contribute to the development of deepfake detection models, it has been tried to determine which facial region has got more importance to detect deepfake.

FARKLI YÜZ BÖLGELERİ KULLANILARAK YAPILAN DEEPFAKE TESPİTİNİN PERFORMANS ANALİZİ

(ÖZET)

Bilgisayarla Görü günümüzün popüler Bilgisayar Bilimi çalışma alanlarından bir tanesidir. Gelişen teknolojinin sunduğu fırsatlar ve teknolojinin yaygın kullanımı sayesinde daha fazla işlem gücüne ve daha fazla görsel veriye sahip olmak, son yıllarda Bilgisayarla Görü'nün gelişim hızına olumlu katkılar sağlamıştır. Son yıllarda kullanımında artış yaşanan Bilgisayarla Görü konularının başında yüz tanıma sistemleri, obje takip sistemleri ve görsel manipülasyonlar gibi konular yer almaktadır. Görsel manipülasyonun alt başlıklarından olan deepfake konusu da şüphesiz sosyal medya sayesinde en popüler konuların baş sıralarında yer almaktadır.

Derin öğrenme kullanılarak bir videodaki veya fotoğraftaki insan yüzlerinin farklı video veya fotoğraflardaki insan yüzleri ile değiştirilmesiyle oluşan görsel veriye deepfake adı verilmektedir. İsmi İngilizce "deep learning" ve "fake" kelimelerinin birleşiminden gelen deepfake, Bilgisayarla Görü alanının halen aktif bir çalışma konusudur. Bilgisayarla görü alanında çalışan kişilerin geliştirdikleri başarılı modeller, yaratılan görsel verilerin gerçek olup olmadığı insanlar tarafından anlaşılamayacak derecede iyi deepfake verisi oluşturma kabiliyetine sahiptir.

Gelişen teknolojinin sunduğu fırsatlar çerçevesinde cep telefonlarındaki uygulamaların kullanımıyla bile kolaylıkla gerçekleştirilen deepfake işlemi sosyal medya uygulamalarının kamera filtrelerinde de yer alarak tüm dünya tarafından tanınmıştır. Eğlence alanında kullanımı yaygın olan deepfake görüntüleri yalan bilgi yayılımı konusunda çok büyük bir tehdit oluşturmaktadır. Yaratılan deepfake görüntülerinin gerçekliğinin kolaylıkla tespit edilememesi ve sahte görüntüler oluşturularak kişilere veya topluluklara zarar verilebilme potansiyelinin yüksek oluşu bilgisayarla görü alanında çalışma yapan kişilerin dikkatini çekmiş ve deepfake tespiti yapabilen modellerin geliştirilmesine sebep olmuştur. Her ne kadar deepfake tespitinde büyük başarılar sağlansa da deepfake alanındaki gelişmelerle tespit zorluğundaki artış kaçınılmazdır. Bu sebeple deepfake teknolojisinin zayıflıkları kullanılarak deepfake tespitindeki performansın yükseltilmesi için çalışmalar sürdürülmektedir. Bu durum bu çalışmanın da ilgi odağı olmuş ve deepfake tespit modellerinin gelişmesine katkı sağlayacağı düşünülerek, yüzdeki farklı bölgeler kullanılarak deepfake tespitinde hangi bölgenin ne derecede önem taşıdığı tespit edilmeye çalışılmıştır.

Contents

1 Introduction and Problem Definition

Image processing techniques have been the focus of attention of people dealing with computer science for years. After discovering how to process visual data thanks to the technology we have, various studies conducted to increase the use of this data have contributed greatly to the development of computer vision. After we realized we could process and manipulate visual data by using computers, with the help of mathematics, the way to find elements such as edges, corners, objects and even movement directions in images was made possible. The world of science was not limited to these inventions, it continued the development of the field of computer vision by using the opportunities provided by the developing and renewed technology. Developments in the field of machine learning and artificial intelligence have taken the field of computer vision to a different dimension, allowing visual data to be learned, processed and manipulated by machines. Today, with the technological power we have, the widespread use of artificial intelligence and computer vision has increased, and thanks to mobile phones and computers, these technologies have turned into technologies that are easily accessible by everyone.

The development and combination of artificial intelligence and computer vision undoubtedly led to very important discoveries. Today, using these technologies, many applications such as face recognition, motion analysis and tumor detection have been developed, as well as visual syntheses such as the development of 3D objects from 2D objects, creation of virtual cities and image transformations called deepfake. Deepfake, one of these issues, is at the heart of this project and studies on deepfake detection have been carried out in this project.

The deepfake process, which is performed by synthesizing fake images using deep learning, takes as the main goal the replacement of a person's face in a video or photo with a face video or photograph of another person. Scientists conducting studies in this field have developed successful models that create deepfakes so well that people cannot distinguish between real and fake. Deepfake is still an active field of computer vision.

Although Deepfake is an academic research area, it has become a technology that has attracted the attention of the whole world with the effect of the widespread use of mobile phones and social media. Nowadays, it is possible to easily create deepfakes even with popular social media applications. Although deepfake seems like a successful and amusing technology, there exist serious problems it may create. For example, today, thanks to the use of deepfake, while people who are not alive can be included in movies, by manipulating the videos of important statesmen, the way has been enabled for the creation of dangerous discourses. For this reason, the subject of deepfake detection has entered the radar of the Computer Vision field and models that detect deepfake have started to be developed by people working in this field. Although among the developed detection models, there are models that have achieved great success, deepfake is still a developing field of study that tries to cover its weaknesses. As deepfakes develop, deepfake detection models will also be developed. This development of detection models will only be possible if weaknesses of deepfakes can be found. This study aims to contribute to the detection of weak spots in deepfakes by measuring the importance of different facial regions in deepfake detection.

2 Literature Survey

Sharing images or videos is one of the most popular activity in today's world. People like to share visual data on the internet, and recently, fake data called deepfake has also been included in this visual data. It is predicted that deepfake elements can cause many problems as well as having entertainment features. For this reason, people working in the Computer Vision field have tried to create deepfake detection models and have achieved successful results in this.

Rösslar et al. have made a great and important contribution to deepfake detection systems with their work, and has become the performance metric of many studies in this field. They created a dataset that has 1000 original youtube videos, 363 original videos of paid actors, 3068 manipulated videos called Deepfake Detection Dataset and manipulated video datasets each containing 1000 videos created by Deepfakes, Face2Face, FaceSwap and NeuralTextures methods. They served this dataset as in low quality, high quality and raw. They also created and tested detection models on this dataset with for every quality group. They prepared a report with 5 different models. As a result of the models they made, they were able to obtain 99.26 percent success as a binary detection. Furthermore, they created a benchmark dataset to make able to other researchers compare their models' performance. [1]

A research done by Li and Lyu showed that deepfakes have got serious weaknesses. In their research they trained the models without any deepfake data. They collected images as positive examples and applied basic image operations like resize and blur to create negative examples. Since images got degraded after deepfake operations, they tried to simulate these degradations with image operations. They trained 4 different methods: VGG16 [2], ResNet50, ResNet101 and ResNet152 [3]. They evaluated these methods with 2 different datasets called UDFV [4] and DeepfakeTIMIT [5]. Top results according to these datasets are 97.4 percent for UADFV [4], 99.3 percent and 93.2 percent for DeepfakeTIMIT's low-quality and high-quality dataset [5] respectively. Their methods showed that Some features of deepfakes offer great advantages in terms of detection. [6]

According to the research of Ciftci, Demir and Yin, deepfakes can be detected by using information gained from not whole face but some regions. They extracted photoplethysmography (PPG) signals from different ROIs and created their models by using these signals. They used FaceForensics++ [1] dataset to train and test models. They made their researches with 7 different models. Their results showed that their approach can detect deepfakes with 97.29% accuracy and the source of deepfake method with 93.39% accuracy. This research is showed that some regions in face has got more information for deepfake detection. [7]

The topic of analyzing the importance of different facial regions, is engaged attention of Tolosano et al. and they made a research on this topic. They masked eye, nose, mouth and rest face area and trained models separately. They used UADFV [4], FaceForensics++ [1], Celeb-DF [8] and DFDC [9]. They trained each facial region and each dataset separately. They used Xception [10] and Capsule Network [11]. Based on their results some facial parts give more accuracy than whole face, also their approach gave more accurate results than some of previous researches. [12]

3 Novel Aspects and Technological Contributions

This project is aimed to find important facial regions for deepfake detection. Unlike similar studies, in this project, it was decided to use all of the datasets together instead of using them separately. Also, in the study that Tolosano et al. made [12], Xception [10] and Capsule Network [11] are used as model, however this project will focus on different models. To avoid from long training times, 20% of whole data will be used as model selector. Vggface2 [13], Vgg16, Vgg19 [2] , Xception [10], ResNet50 and ResNet152 [3] will be tested during the model selection phase. While measuring the performance of models and different facial regions during the model selection phase, top 2 or 3 model will be selected as main model based on models' average scores. Thus, powerful deepfake detection models will be created at the end of this study.

3.1. Datasets

Three different datasets are selected for this project. These datasets will be used combined during the training phase of model. The selected models are: FaceForensics++ high-quality (c23 compression) dataset [1], Celeb-DF [8] dataset and DeepfakeTIMIT high-quality dataset [5]. Information about datasets can be seen in Table 3.1.

Table 3.1: A Sample table

	Original Video Count	***Manipulated Video Count***
FaceForensics++ [1]	1363	8068
Celeb-DF [8]	890	5639
DeepfakeTIMIT [5]	0	320
Total	**2253**	**14027**

3.1. General Structure

Videos of these datasets have got average duration of 15 second. 10 frames at equal distance from each other will be extracted from videos by using OpenCV [14]. Then, 68 facial spots will be found by using DLIB [15] and ROIs will be extracted. Five facial regions will be extracted from the 20% split of whole dataset for model selection phase. According to the results of model selection phase some ROIs can be discarded in the final models. The selected facial regions will be eye area, mouth area, midface area, nose area and chin area. While creating these areas below boundaries are used for regions as in order top, bottom, left and right edges.

- **Eye Area:** Top 18-27 / Mid-point of bottom 37-48 and spot 29 / Leftmost of 18 and 37 / Rightmost 27 and 46.
- **Mouth Area:** Bottom 32-36 / Bottom 55-6 / Leftmost of 49,50,60 and 61 / Rightmost 54,55,56 and 65

- **Midface Area:** Bottom 37-48 / Top 49-55 / Leftmost 1-17 / Rightmost 1-17
- **Nose Area:** Top 49-55 / Top 18-27 / Spot 32 / Spot 36
- **Chin Area:** Top 49-55 / Bottom 1-17 / Leftmost 1-17 / Rightmost 1-17

Example image can be seen in Figure 3.1.

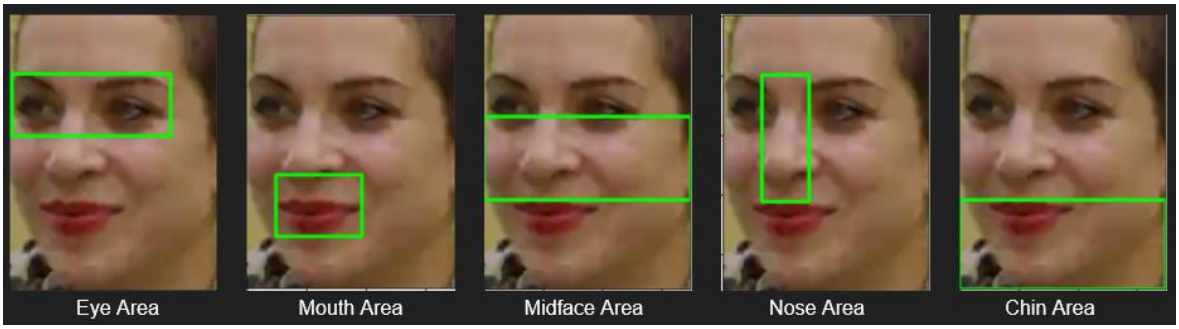

Figure 3.1: A sample figure from FaceForensics++ Dataset [1]

4 System Requirements

4.1 Functional Requirements

- Researchers must be able to get information about used model and corresponding scores.
- People who want to detect deepfake videos should be use the trained models of this project.
- It should assist other researchers in deepfake detection by presenting the importance of different facial regions.

4.2 Non-Functional Requirements

- Models must be dataset independent and give similar results on different datasets.
- Models should be able to successfully distinguish whether deepfake or not.
- Models should be work with images that are not extracted from videos.

4.3 Use Cases / User Stories

A researcher is working on deepfake detection systems but ger model gives lower performance than her expectations. She makes a literature search about deepfakes and finds this project. After examine the results of this project she decides to give different weights on different facial regions and improve her model.

A researcher is trying to find out if a video of people wearing masks is deepfake. However, he does not have any data to train a model. He searches for a model to test videos that he has and finds this project's models. He downloads final eye model and tests videos that he has. Finally, he gets the results of test and learns whether the videos are deepfake or not.

5 Project Plan

5.1 Project Resources

Hardware:

- A decent computer to extract frames from videos
- GPU powered system to train detection models
- 500 GB free disk space

Software:

- Python (3.7.7+) [16]
- DLIB Python library [15]
- OpenCV Python library [14]
- TensorFlow Python library [17]
- NumPy Python library [18]
- Jupyter Notebook for Python [19]
- Nvidia CUDA [20]

5.2 Work Breakdown

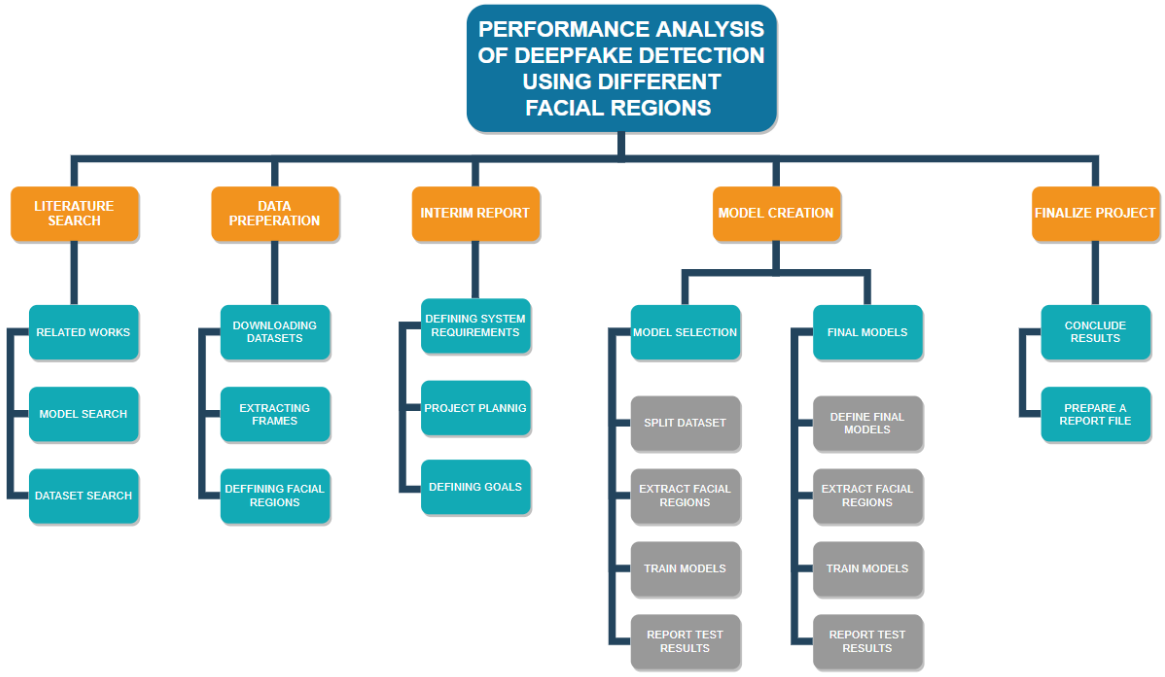

Figure 5.1: The Work Breakdown Structure of the project

5.3 Time Plan

#	TASK	START DATE	DUE DATE	DURATION
1.1	Previous Researches on Deepfake Detection	09/09/2020	04/10/2020	25
1.2	Knowledge Revision Based on Researches	04/10/2020	11/10/2020	7
2.1	Data Collection	11/10/2020	01/11/2020	21
2.2	Data Cleansing and Exploratory Data Analysis	01/11/2020	06/01/2021	66
3	Interim Report	06/01/2021	08/02/2021	33
4.1	Data Creation for Model Selection	08/02/2021	21/02/2021	13
4.2	Model Evolution	21/02/2021	07/03/2021	14
5.1	Data Creation for Final Models	21/02/2021	14/03/2021	21
5.2	Define and Implement Final Models	14/03/2021	30/03/2021	16
5.3	Apply Trained Models on Test Dataset	30/03/2021	05/04/2021	6
6	Report of Results	05/04/2021	30/05/2021	55

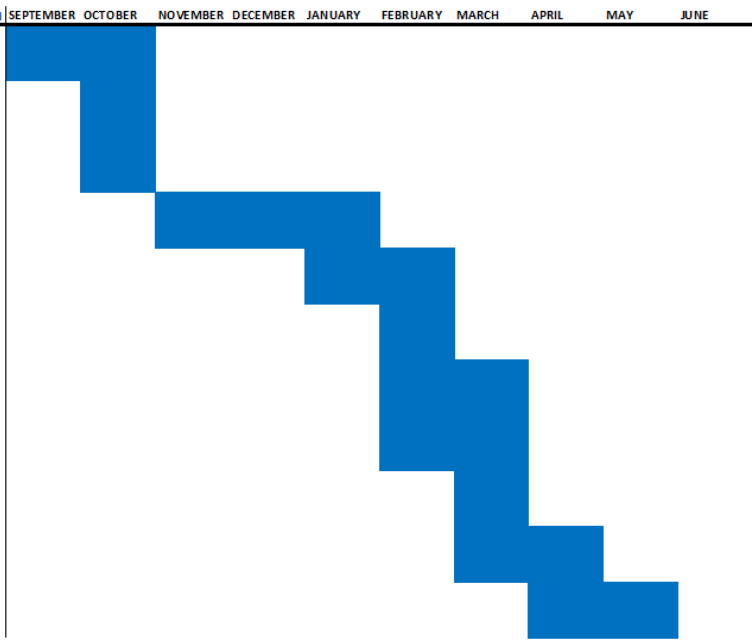

Figure 5.2: The Gantt Diagram of the project

6 Goals and Evaluation Criteria

At the end of this project at least 6 models (eye area, midface area, nose area, mouth area, chin area and whole face) must be obtained. Since similar projects' accuracies are above 80, this project's test accuracy in distinguishing deepfake must be above 80. While training models, output of models must be calculated by using Logistic Sigmoid Function and class of input must be decided based on Logistic Sigmoid Function's result. Total accuracy must be reported.

6.1 Logistic Sigmoid Function

$$f(x) = \frac{1}{1 + e^{-x}}$$

Logistic Sigmoid Function is a mathematical function that gives output in range bounded with 0 and 1. Logistic Sigmoid Function is real function that takes all real values as input. It is differentiable and has got non-negative derivative on all point and has got only 1 inflection point. This function generally used in binary classification models. If output closes to zero class will be labeled as 0 and if output closes to one class will be labeled as 1.

6.2 Accuracy

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

True Positive (TP): The count of correct predictions of positive class.
True Negative (TN): The count of correct predictions of negative class.
False Positive (FP): The count of false predictions of positive class.
False Negative (FN): The count of false predictions of negative class.

Accuracy defines the quality of correctness. Generally defined as number in percent. High rates of accuracy in test data generally mean that the model has been well trained.

7 References

[1] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies and M. Nießner, "FaceForensics++: Learning to Detect Manipulated Facial Images," in *ICCV*, 2019.

[2] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," 2015.

[3] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," in *CVPR*, 2016.

[4] X. Yang, Y. Li and S. Lyu, "Exposing Deep Fakes Using Inconsistent Head Poses," in *ICASSP*, 2019.

[5] P. Korshunov and S. Marcel, "DeepFakes: a New Threat to Face Recognition? Assessment and Detection," 2018.

[6] Y. Li and S. Lyu, "Exposing DeepFake Videos By Detecting Face Warping Artifacts," 2019.

[7] U. A. Ciftci, İ. Demir and L. Yin, "How Do the Hearts of Deep Fakes Beat? Deep Fake Source Detection via Interpreting Residuals with Biological Signals," 2020.

[8] Y. Li, P. Sun, H. Qi and S. Lyu, "Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics," in *IEEE Conference on Computer Vision and Patten Recognition (CVPR)*, Seattle, WA, United States, 2020.

[9] B. Dolhansky, R. Howes, B. Pflaum, N. Baram and C. C. Ferrer, "The Deepfake Detection Challenge (DFDC) Preview Dataset," 2019.

[10] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," in *IEEE/CVF Conference on Computer Vision and Pattern*, 2017.

[11] H. H. Nguyen, J. Yamagishi and I. Echizen, "Use of a Capsule Network to Detect Fake Images and Videos," 2019.

[12] R. Tolosana, S. Romero-Tapiador, J. Fierrez and R. Vera-Rodriguez, "DeepFakes Evolution: Analysis of Facial Regions and Fake Detection Performance," 2020.

[13] Q. Cao, L. Shen, W. Xie, O. M. Parkhi and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," 2018.

[14] OpenCV, "Open Source Computer Vision Library," 2015.

[15] D. E. King, "Dlib-ml: A Machine Learning Toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755-1758, 2009.

[16] G. Van Rossum and F. L. Drake, Python 3 Reference Manual, Scotts Valley, CA: CreateSpace, 2009.

[17] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard and R. Jozefowicz, "TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems," 2015.

[18] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk and Br, "Array programming with NumPy," *Nature*, vol. 585, pp. 357-362, September 2020.

[19] T. Kluyver, B. Ragan-Kelley, F. Perez, B. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. Hamrick, J. Grout, S. Corlay, P. Ivanov, D. Avila, S. Abdalla and C. Willing, "Jupyter Notebooks -- a publishing format for reproducible computational workflows,"

in *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, F. Loizides and B. Schmidt, Eds., IOS Press, 2016, pp. 87 - 90.

[20] NVIDIA, P. Vingelmann and F. H. Fitzek, "CUDA," 2020. [Online]. Available: https://developer.nvidia.com/cuda-toolkit. [Accessed 8 February 2021].