PERFORMANCE EVALUATION: TRIPLET LOSS IMPLEMENTATION ON ALEXNET WITH MTCNN IN DETECTING DEEPFAKE

Rich Tristan Lim
Department of Computer,
Information Sciences, and
Mathematics
University of San Carlos
Cebu City, Philippines
richlim16@gmail.com

Angie Ceniza-Canillo,
PHD
Department of Computer,
Information Sciences, and
Mathematics
University of San Carlos
Cebu City, Philippines
amceniza@usc.edu.ph

Raymond Anthony Aya-ay
Department of Computer,
Information Sciences, and
Mathematics
University of San Carlos
Cebu City, Philippines
raymond.p.ayaay@gmail.com

ABSTRACT

Deep fakes are media, usually in the form of a video, that have been altered to change certain features from the original video, such as swapping the face of the subjects to make them appear as someone else, this form of deep fake is also known as FaceSwap. This is done by using an existing video and replacing the actor with someone else's appearance or likeness. They are created using artificial intelligence, where real images are fed to a system and are trained by utilizing two parts - one which creates fake images and the other which spots the fakes until it cannot distinguish between the real and fake. This paper analyzes an already existing deepfake detection system known as AlexNet, aided by a Multi-Task Cascaded Convolutional Neural Network, and seeks to find the effects of implementing a Triplet Loss Network. The system is given a video, which will be preprocessed and analyzed to be classified as either deep fake or authentic. Training and testing of the machine is to be done over pre-existing datasets, namely: the Celeb-DF dataset, the DeepFake Detection Challenge dataset, and the FaceForensics++ dataset. After several iterations of testing, the authors of this paper recorded the following results from applying different loss functions to the proposed model; Binary Cross Entropy yielded an AUC score of 75.58%, Semi Hard Triplets yielded a score of 72.04%, Contrastive Loss yielded a score of 62.37%, and lastly, Sparse CategoricalCross Entropy yielded a score of 55.58%.

The authors of this paper will then compare the results to measure the effects of implementing the Triplet Loss Network to the base system of AlexNet with MTCNN. This will be done by comparing the true positive rate over the false positive rate of the models.

CCS CONCEPTS

• Software and its engineering • Software creation and management • Software verification and validation
• Empirical software validation

KEYWORDS

Deep fake detection, deep fakes, Deep Neural Network, MTCNN, Triplet Loss

ACM Reference format:

Rich Tristan Lim, Angie Ceniza-Canillo and Raymond Anthony Aya-ay. 2024. A PERFORMANCE EVALUATION: TRIPLET LOSS IMPLEMENTATION ON ALEXNET WITH MTCNN IN DETECTING DEEPFAKES. In *Proceedings of Philippine Computing Science Congress (PCSC2024). Laguna, Philippines, 6 pages.*

1 Introduction

Deepfakes generally refer to AI-generated videos where the identities of the subjects have been swapped for those of another person. These videos were originally created with conventional computer graphics methods, but due to the recent advances in deep learning networks and the considerable increase in the computing power of personal computers, newer deep fake videos are easier to make and are much harder to detect due to their continued evolution and improvement. In particular, Generative Adversarial Networks (GANs) have been used to generate newer generations of deep fake videos. While deepfakes can be used for beneficial purposes, they can also be weaponized. Such deep fakes can be used to spread misinformation, propaganda, and fake news in general. One of the earliest examples of widespread deep fakes is the collection of pornographic videos of celebrities posted on Reddit (Harris, 2018), where the videos had the faces of celebrities superimposed into the pornographic videos to make it look like the celebrities were the main subjects of the videos.

Such fraudulent videos have the potential for grave consequences, such as discrediting institutions or even political candidates, thereby tipping electoral outcomes (Chesney & Citron, 2019) and damaging international relations.

Having been inspired by previous studies on this topic, the authors of this paper sought to combine two existing studies, namely, a study in 2020 by Xie, Chatterjee, Liu, Roy, & Kossi that used a modified version of a convolutional neural network (CNN) called AlexNet, and another study in 2020 by Bhavsar, Kumar, & Verma that used the Xception architecture with MTCNN and applied Triplet Loss to the embeddings. This research aims to find the effects of applying the Triplet Loss Network to AlexNet with MTCNN, as well as understanding why the results are such.

2 Related Literature

2.1 Generating Deep Fakes

Major improvements and advances in computer graphics, computer vision, and machine learning have led to the development of deep fake image, video, and audio as well as their continuous improvement. Due to the fact that AI-synthesized content is relatively new (Agarwal et al. 2019), there have been continuous improvements in the development. For example, a study by Yuezun, Ming-Ching, and Siwei (2018) made the observation that there was an irregularity in the blinking done by individuals in face-swap deep fakes due to the fact that the training data used to synthesize faces usually did not depict the person with their eyes closed. It did not take long after this study was made public for synthesis techniques to make the necessary changes to render this method of detecting deep fakes less effective.

Generative Adversarial Networks (GANs) are a form of deep neural networks for both supervised and semi-supervised learning (Creswell et al., 2018) that has been used in generating deep fakes. These models require a large set of training data to create deep fake media, and the larger the training set is the more realistic and indistinguishable the result will be. GANs have two neural network components: a generator and a discriminator. These two can be thought of as an art forger and an art expert. The model uses the generator to train on the training set provided to generate the deep fake. This is then given to the discriminator which has been trained to differentiate the real data from the fake data (Almars, 2021). However, the drawback of requiring large amounts of training data is offset by the large amount of publicly accessible data online, especially for public figures and celebrities. Although deepfakes usually require a large number of images to create a realistic forgery, techniques have already been developed where one can generate a fake video by feeding it only one photo such as a selfie (Westurlund, 2019).

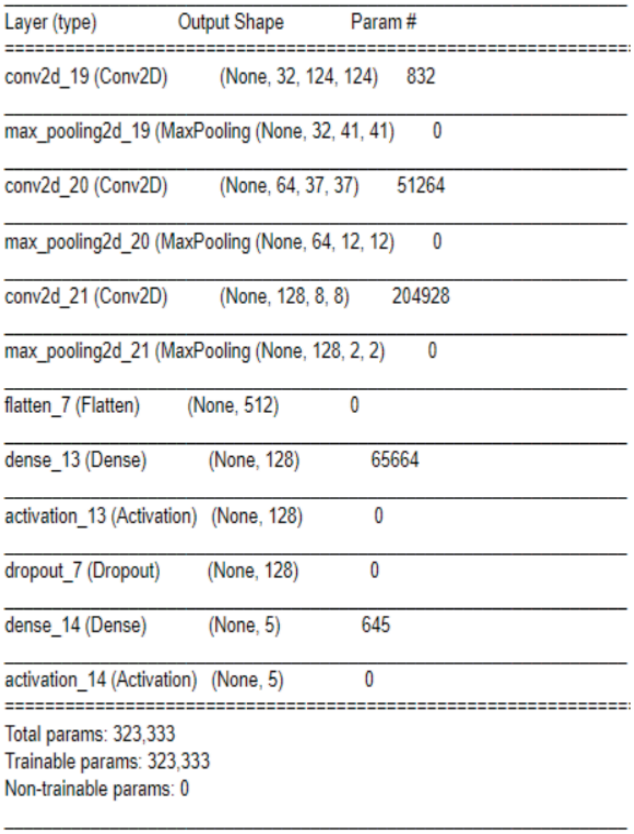
2.2 Detecting Deep Fake Videos

While there are potentially beneficial applications for deep fake technology, because of the fact that videos of an individual have a significant impact on their image and reputation, they can also be weaponized in ways that far outweigh their potential benefits (Citron & Chesney, 2019). The potential threats can range from revenge porn, to a politician saying outrageous or controversial things, causing political or religious tensions between countries, or even of a company official making statements of the company to affect the outcome of the stock market (Ngyuen et al., 2019). Due to the potential damage, as well as how rapid such videos can spread in the current digital environment.

There are multiple methods of how a DeepFake can be detected. General Network-based Methods regard detection as a frame-level classification task which is finished by CNNs. Temporal Consistency-based Methods identify DeepFakes by detecting inconsistencies between adjacent frames due to the defects of the algorithm. Visual Artifacts-based Methods use the intrinsic image discrepancies found in blending boundaries, called artifacts. Camera Fingerprints-based Methods use the different traces that are left by devices in captured images. Which helps acknowledge that faces and background images are from different devices. In DeepFake generation, it is difficult to synthesize humans with believable behavior, thus these biological signals are extracted to detect DeepFake videos, this is an example of a Biological Signals-based Method (Yu et al., 2021). These methods can be categorized to Low-level Approaches and High-Level Approaches.

2.3 Alexnet

AlexNet is a Convolutional Neural Network that contains many layers. For this study, a modified lighter version of AlexNet based on a study by Daniel, Chatterjee, Liu, & Roy (2020) will be used. Instead of the more in depth model, three convolutional layers, three max-pooling layers, one flatten later, one dense layer, one activation layer, and an optional dropout layer will be used.



```
Layer (type)              Output Shape           Param #
=================================================================
conv2d_19 (Conv2D)        (None, 32, 124, 124)   832
_________________________________________________________________
max_pooling2d_19 (MaxPooling (None, 32, 41, 41)  0
_________________________________________________________________
conv2d_20 (Conv2D)        (None, 64, 37, 37)     51264
_________________________________________________________________
max_pooling2d_20 (MaxPooling (None, 64, 12, 12)  0
_________________________________________________________________
conv2d_21 (Conv2D)        (None, 128, 8, 8)      204928
_________________________________________________________________
max_pooling2d_21 (MaxPooling (None, 128, 2, 2)   0
_________________________________________________________________
flatten_7 (Flatten)       (None, 512)            0
_________________________________________________________________
dense_13 (Dense)          (None, 128)            65664
_________________________________________________________________
activation_13 (Activation) (None, 128)           0
_________________________________________________________________
dropout_7 (Dropout)       (None, 128)            0
_________________________________________________________________
dense_14 (Dense)          (None, 5)              645
_________________________________________________________________
activation_14 (Activation) (None, 5)             0
=================================================================
Total params: 323,333
Trainable params: 323,333
Non-trainable params: 0
_________________________________________________________________
```

Figure 1. **AlexNet Structure**

2.4 **Multi-task Cascaded Networks**

Multi-task Cascaded CNNs (MTCNN) extracts faces from frames using face detection and alignment to boost performance. It mainly consists of three parts: Proposal networks, which detects faces across multiple resolutions by generating a list of candidate windows and using it for classifying which are the faces and estimating bounding box regression vectors and non-maximum suppression. Refine net rejects the false candidates. Finally, the output network will output five facial landmarks (Xiang & Zhu, 2017).

2.5 **Triplet Network**

Triplet network is a type of metric learning that requires three sample input images, which are the anchor sample (A), the positive sample (P) which is of the same class as the anchor, and the negative sample (N) which is of a different class. Triplet loss is used to calculate the loss of estimation results of the three samples. The network minimizes the distance between P and A, and maximizes the distance between the N and A, this formula is shown in Figure 2.

$$\|f(x)\|_2 = 1$$

$$\sum_i^N \left[\overset{1}{\boxed{\|f(x_i^a) - f(x_i^p)\|_2^2}} - \overset{2}{\boxed{\|f(x_i^a) - f(x_i^n)\|_2^2}} + \alpha \right]_+$$

$$\text{where } [z]_+ = max(z, 0)$$

Figure 2. **Triplet Loss Formula**

The first item is the distance between the anchor and positive and the second item is the distance between the anchor and negative. The value of the first item is learned to be smaller while the second item is bigger. If the difference is smaller than minus alpha, the loss would become zero and the network parameters would not be updated at all (Hoffer & Ailon, 2014).

In a study by Bhavsar, Kumar, & Verma (2020) the triplet loss network was implemented in which they used semi-hard triplets, where the negative is farther from the anchor than the positive, but still produces a positive loss. In this study, the negative triplet is a deepfake while the positive and anchor triplets are genuine. The researchers used FaceNet to generate face embeddings of a 512 dimension vector and applied triplet loss to those embeddings. From there the network learns the discriminative features to the embeddings of the original and the manipulated faces separately.

3 **Technical Background**

3.1 **Deepfake**

It is a generated media in which a person in an existing image or video is replaced with someone else's likeness. While the act of faking content is not new, deep fakes leverage powerful techniques from machine learning and artificial intelligence to manipulate or generate visual and audio content to appear more authentic to the viewer. (Kietzmann, 2020). The most significant problem with deep fakes is the scope, scale, and sophistication of the technology, and how easy it is to generate, since almost anyone with a computer can create convincing fake videos with free applications and online services (Fletcher, 2018).

3.2 Deep Learning

It is a category of machine learning algorithms that uses numerous layers to gradually extract higher-level features from the raw input. These neural networks attempt to simulate the behavior of the human brain. While a neural network with a single layer can still make approximate predictions, additional hidden layers can help to optimize and refine for accuracy (IBM Cloud Education, 2020). For example, in image processing, lower layers may identify edges, while higher layers may identify the concepts relevant to a human such as digits or letters or faces.

4 Methodology

4.1 Research Environment and Respondents

The data used in this study is based on a previous research by Li et al. (2020) entitled Celeb-DF where the authors found that several datasets being used by studies on deep fake detection had a considerable amount of data that was too easy to detect and, therefore, were not representative of current deep fake videos that one is more likely to find. As a result of this, they created a dataset which was more sophisticated and proved to result in lower average detections when compared to several other popular datasets. Similar results were also shown in a later study by Tolosona et al. (2020), in which they compared several deep fake detectors against different datasets and found that Celeb-DF consistently had lower detection scores along with the DFDC database (Dolhansky et al., 2019), which is a public database released by Meta in collaboration with other companies such as Microsoft, Amazon, and MIT to be used in their Deep Fake Detection Challenge which was created to boost development in this field. FaceForensics++ is another modern dataset that was commonly used among recent studies in deep fake detection and will serve as another benchmark for comparison of results.

4.2 Research Instrument or Sources of Data

This study uses two recent databases of the 2nd generation of deep fakes, namely Celeb-DF, DFDC, and one database of the 1st generation, FaceForensics++. These three were chosen because they are the latest databases for deep fake videos and have also proven to be more sophisticated and challenging for deep fake detectors (Li et al., 2020; Deepfake Detection Challenge Dataset, 2020). DFDC will consist of two versions, one of which contains around 5,000 videos that feature two facial modification algorithms and the other version contains around 124,000 videos that feature eight facial modification algorithms. Celeb-DF will contain around 5,600 high-quality videos of celebrities and FaceForensics++ will contain 1,000 real videos taken from Youtube and will have 4 types of deep fake videos of varying qualities.

4.3 Research Procedure

4.3.1 Gathering of Data. The general objective of this study is to measure the effects of implementing triplet loss to the CNN, particularly in classifying deep fake videos from authentic videos that have not been tampered in a way that changes the identity of the subjects in the video. The datasets used were developed for the purpose of providing a more challenging and sophisticated training set for deep fake detectors to improve the effectiveness of following studies in this field of study (Li et al., 2020). Prior studies have proven these datasets to be the most challenging to test against when compared to other popular datasets.

4.3.2 Treatment of Data. All the datasets utilized in this study are composed of videos which will be extracted into frames so the models will focus more on the facial imagery. The OpenCV package will be used to read each video file and extract every 5th frame of the video, as well as reducing the resolution of each frame. An MTCNN to extract faces out of the frames and place them in the center of alignment .

4.4 Conceptual Framework

The deep learning model used in this research is based on a modified, lighter version of AlexNet by Xie, Chatterjee, Liu, Roy, & Kossi (2020) to detect deep fake videos from real videos. Initially, the faces are extracted from the frames of the videos via MTCNN to place five landmarks for each face. Afterwards the authors of this paper will take the processed frames and generate a 512 dimension vector using the modified AlexNet. One of the models will apply semi-hard triplets to the embeddings generated by AlexNet and the other model will not. The results will be measured and compared to determine the effects of applying the Triplet Loss Network. The main purpose of this study is to find the benefits and drawbacks of applying the Triplet Loss Network to the deep fake detection system.

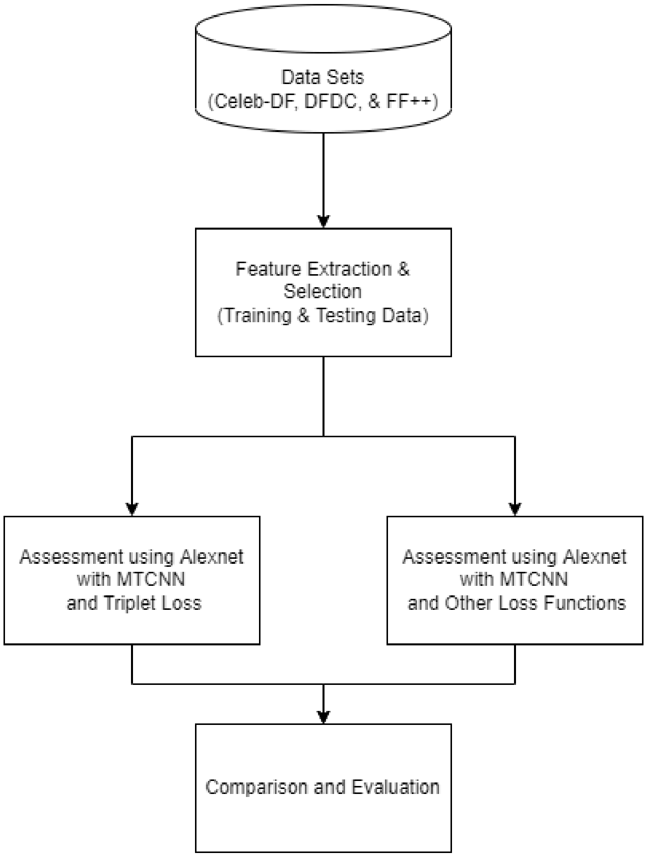

Data Sets
(Celeb-DF, DFDC, & FF++)
Feature Extraction & Selection
(Training & Testing Data)
Assessment using Alexnet with MTCNN and Triplet Loss
Assessment using Alexnet with MTCNN and Other Loss Functions
Comparison and Evaluation

Figure 3. **Conceptual Framework**

5. Results and Analysis

The authors of this paper started testing each dataset individually to see how they affected the results of the model. The results in Figure 4 show that Celeb-DF V2 resulted in higher performance scores in accuracy, whereas FF++ and DFDC yielded closer scores in relation to each other. However, all the average scores were within 20% of each other, therefore, the researchers concluded that this difference would not significantly skew the results of the model.

With a combination of all three datasets at 5 frames per second for a total of 450,000 frames, however no significant difference in performance was found when compared to a smaller subset of 75,000 frames, thus, succeeding tests were performed with the smaller number of frames (75,000 frames) with 25,000 frames coming from each dataset.

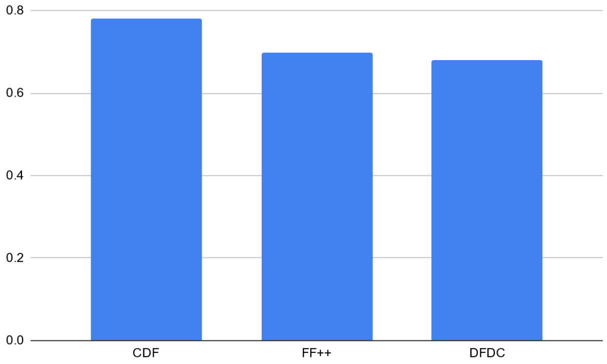


Figure 4. **Performance Scores in Accuracy of Each Dataset**

In the training of the model all frames were extracted with MTCNN applied and the model used was AlexNet in all tests. The different Loss functions were tested and the results for their average accuracy are recorded in the bar chart illustrated below in Figure 5. The results show that the one outlier in the results is contrastive loss, with a significantly lower accuracy score compared to the other loss functions, with an accuracy of 69.54%, whereas the other three are within a 5% difference. The highest accuracy was achieved by using Binary Cross Entropy with an accuracy of 89.08%, followed by Sparse Categorical Entropy with an accuracy of 87.44%, and Semi Hard Triplets achieved an accuracy of 85.71%.

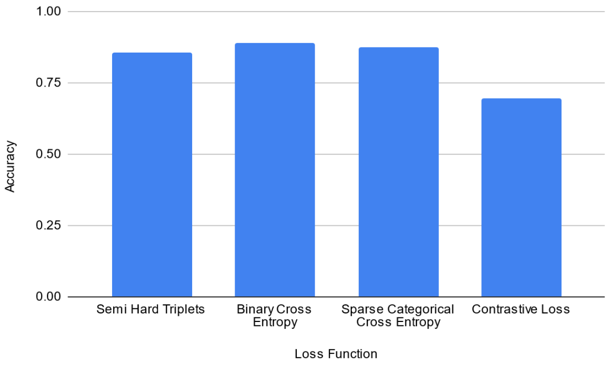


Figure 5. **Average Accuracy Scores of Each Loss Function**

When the results were finalized with AUC scores as shown in Figure 6 below, it is seen that Binary Cross Entropy still maintains the highest performance. However Sparse Categorical Cross Entropy is now the lowest performing, with Semi Hard Triplets and Contrastive Loss having better AUC scores, meaning that although using Sparse Categorical Cross Entropy for loss yielded in higher accuracy, Semi Hard Triplets and Contrastive Loss actually, this may hint at poorer positive class classification in comparison to negative classification for Sparse Categorical Cross Entropy, as ROC scores are biased to positive classes.

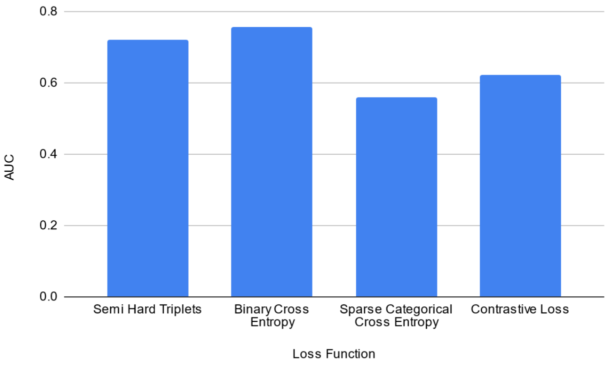


Figure 6. **Average AUC Scores of Each Loss Function**

6. Conclusion and Recommendation

In this work, the authors of this paper presented a deep study for classification of deep fake videos using a benchmark testing procedure with MTCNN and Alexnet Deep Learning Model, to collect and analyze standardized results that give a better understanding of possible improvements in the field of deep fake detection with differing loss functions. The effects of different datasets were studied and how they affect the performance results as well as testing the results of having a combined dataset. For future work, our aim is to use a more modern dataset and test it on the newer generation of deep fake videos to evaluate their effectiveness on newer methods.

In a previous study by Hoffer & Ailon (2014), they stated that triplet loss, despite not specializing in classification, does perform well when compared to other models specific for classification tasks. Our findings prove that triplet loss does work well, but is outperformed by loss functions specific to binary classification tasks such as deepfake detection. Therefore, the authors of this paper conclude that triplet loss works well for classification tasks, but its method of learning differences does not benefit in the task of binary classification. Instead it is better to use loss functions specific to binary classification tasks.

ACKNOWLEDGMENTS

Firstly, the authors of this paper would like to thank God, their families, friends, and everyone who has provided support and guidance to them in one way or another.

REFERENCES

[1] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, H. Li: Protecting world leaders against deep fakes. In: IEEE Conference on Computer Vision and Pattern Recognition, Workshop on Media Forensics. pp. 38–45 (2019)

[2] A.M. Almars (2021). Deepfakes Detection Techniques Using Deep Learning: A Survey. Journal of Computer and Communications, 9(5), 20-35. scirp.org. doi: 10.4236/jcc.2021.95003

[3] R. Chesney, & D. Citron. (2019). Deepfakes and the new disinformation war: The coming age of post-truth geopolitics. Foreign Aff., 98, 147.

[4] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, & A.A. Bharath (2018). Generative Adversarial Networks: An Overview. IEEE Signal Processing Magazine, 35(1), 53-65. doi: 10.1109/MSP.2017.2765202

[5] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, & C.C. Ferrer (2019). The Deepfake Detection Challenge (DFDC) Preview Dataset. arXiv Preprint. arXiv:1910.08854

[6] J. Fletcher (2018). Deepfakes, Artificial Intelligence, and Some Kind of Dystopia: The New Faces of Online Post-Fact Performance. Theatre Journal, 70(4), 455-471. Project MUSE. doi:10.1353/tj.2018.0097

[7] D. Harris (2018). Deepfakes: False pornography is here and the law cannot protect you. Duke L. & Tech. Rev., 17, 99.

[8] E. Hoffer, & N. Ailon (2014). Deep metric learning using Triplet network. arXiv. https://arxiv.org/abs/1412.6622

[9] A. Kumar, A. Bhavsar, & R. Verma (2020). Detecting deepfakes with metric learning. In 2020 8th international workshop on biometrics and forensics (IWBF)(pp.1-6).IEEE.

[10] D. Xie, P. Chatterjee, Z. Liu, K. Roy and E. Kossi, "DeepFake Detection on Publicly Available Datasets using Modified AlexNet," 2020 IEEE Symposium Series on Computational Intelligence (SSCI), Canberra, ACT, Australia, 2020, pp. 1866-1871, doi: 10.1109/SSCI47803.2020.9308428.

[11] R. Tolosona, S. Romero-Tapiador, J. Fierrez, & R. Vera-Rodriguez (2020). DeepFakes Evolution: Analysis of Facial Regions and Fake Detection Performance. Biometrics and Data Pattern Analytics - BiDA Lab, Universidad Autonoma de Madrid.
[11] M. Westurlund (2019). The Emergence of Deepfake Technology: A Review. Technology Innovation Management Review, 9(11), 39-52.http://doi.org/10.22215/timreview/1282

[12] J. Xiang, & G. Zhu (2017). Joint Face detection and Facial Expression Recognition with MTCNN. 2017 4th International Conference on Information Science and Control Engineering. doi:10.1109/ICISCE.2017.95

[13] Y. Li, M. -C. Chang and S. Lyu, "In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking," *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, Hong Kong, China, 2018, pp. 1-7, doi: 10.1109/WIFS.2018.8630787.

[14] P. Yu, Z. Xia, J. Fei, & Y. Lu (2021). A Survey on Deepfake Video Detection. IET Biometrics, 10(6), 607-624. https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/bme2.12031