2023 3rd International Conference on Computing and Information Technology (ICCIT) | 979-8-3503-2148-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICCIT58132.2023.10273909

A Novel Framework for Detection of Digital Face Video Manipulation using Deep Learning

Sanskriti Chandra[1], Shivam Saxena[2], Santosh Kumar[3], Mithilesh Kumar Chaube[4], Srinivas KG [5],
Saeed Hamood Alsamhi[6], Edward Curry[6], Abdu Saif[7]
{sanskriti20102, shivam20102, santosh, mithilesh, srinivasa}@iiitnr.edu.in
{saeed.alsamhi, edward.curry}@insight-centre.org
abdu.saif@taiz.edu.ye
[1,2,5]Department of Data Science and Artificial Intelligence
[3]Department of Computer Science and Engineering
[4]Department of Mathematical Sciences
[1,2,3,4,5] IIIT-Naya Raipur, Chhattisgarh, India
[6]Insight SFI Research Centre for Data Analytics,
University of Galway, Galway, Ireland
[7] Faculty of Engineering & Information Technology, Taiz, Yemen

***Abstract*—Digital face manipulation and classification have recently attracted the attention of academia and industry worldwide. Researchers have developed deep learning and computer vision techniques for detecting face manipulations, and it has become a challenging task to differentiate between authentic and manipulated face images manually. The challenge results in the decline of authenticity in digital media content. In this paper, we propose a framework for the classification of manipulating face images using the EfficientNet learning model. The proposed framework takes four digital facial forgeries: Face-Swap, Face-2-Face, DeepFakes, and neural textures. Multiple manipulation techniques are used to process manipulated faces, such as the Blaze-face tracking method, to determine the locations of the face images and pixel coordinates. The proposed framework is used first to identify the type of face manipulation and then to perform detection of the tampered regions in the face images. The proposed framework provided an automated benchmark that considers all four modification techniques in a realistic situation. The results show that the proposed framework outperforms existing approaches regarding accuracy and efficiency. Furthermore, the proposed framework is suitable for detecting digital face video manipulation in various applications, including forensics and security.**

***Index Terms*—Machine learning, deep learning, Face-Swap, neural textures, Face-2-Face**

I. Introduction

Rapid advancement in the computer vision and deep learning paradigms has recently proliferated due to massive video face verification and manipulation applications. Generative models have evolved immensely in the last few years. Generative Adversarial Network (GAN)-based video and image generation has become very accessible due to the availability of implementable tools and software to any user for manipulating face images. Hence, manipulating faces may pose severe challenges for threat and security to society. Several GAN-based video and image methods have become available for different users in the public domain. Generating digital faces in such fake videos/images involves high computation, which increases anxiety and suspicion of the digital content over the Internet. It is a big challenge for humans to differentiate between real and fake faces in a video, significantly when the video is compressed or has low resolution.

In 2017, the DeepFake phenomenon got attention and needed face forensics to identify manipulated facial images and video regions. DeepFakes effects' popularity is highly used to deliberately create misleading or deceptive audiovisual content by manipulating face images, audio, and video. To classify the manipulated faces, digital forensic methods are used for DeepFakes analysis by finding facts associated with the contents, which have been de-contextualized in terms of time, form, and location. Based on the available literature, accurate face forgery detection would have an immediate and far-reaching impact in alleviating the malicious intents of DeepFakes, such as face manipulation, face morph, face recognition attacks, and fake news [10], [14].

Both false news and DeepFakes are generated with the similar objective of deceiving the user. However, because visual data is much easier for the user to accept than other types of information, modified images can be much more influential [11]. A de-contextualized and manipulated image or video will likely have a far more significant effect than a text. They could, for instance, make a reporter cover a news item incorrectly. Reporters must thus always be conscious of this threat and meticulously check the reliability of their information sources. The main limitation of other state-of-the-art methods is their unsatisfactory results in detecting realistically manipulated videos. Results are frequently reported for synthetically altered facial videos (i.e., those created by automatically copying and pasting sections in various places), whose validity can be instantly measured by basic visual inspection, while making manually falsified sequences is time-consuming.

A. Motivation and Contributions

A large number of facial manipulation approaches have been developed in recent years by researchers across the world. Earlier, there were limited resources and less domain expertise for facial forgery, and digital content was more reliable. However, with the advancement in deep learning that removes manual editing steps like auto-encoders and GAN and an increased number of datasets available, generating even non-existing faces or altering real faces is easy. It is nearly impossible to distinguish between a natural face and a manipulated face by human observation. It poses a severe problem for threats and has serious repercussions on our society as such advanced manipulation approaches can be used for illegal activities or to spread fake news in the public domain.

In the last few decades, much research has been done on face manipulation techniques in both images and videos. In the context of digital face manipulation, DeepFakes is the most popular term and is becoming a significant concern for the public concerning authentication. DeepFake refers to all the false digital content generated using deep learning techniques [3], [4]. The worst DeepFakes applications include untruths, false websites, and financial fraud [5]. We proposed a framework for the classification of facial manipulation in video sequences using the EfficentNet classification technique to solve the problem of face manipulation. Four digital facial forgeries are used as input to model training for classifying [13], [15] any unknown/query (test image) is manipulated face and normal face in video sequence: Face-Swap, Face-2-Face (F2F), DeepFakes, and neural textures. Multiple manipulation techniques are used to process manipulated faces, such as the Blaze-face tracking method, to determine the locations of the face images and pixel coordinates to ensure the manipulation of the face or not in a given video sequence.

B. Literature Review

Recently, numerous video forensics methods have been presented for various jobs [7]. Research on forensic decision-making utilizing edge energy information from stochastic pictures is being done by Rhee [8]. From JPEG-compressed pictures of an actual picture with a varied Q-Factor and a query image, edge information is retrieved using SA (steaking artifacts) and SPAM (subtractive-pixel-adjacency matrix). This data is compared to the TCJCR (threshold by the conjunction of Jpeg ratios) to identify picture modification. TP (True Positive) and FN (False Negative) are, respectively, 87.20% and 13.80%. Using the mean score of the Wavelet-transform coefficient, Jeon et al. In [9] introduced the identification of copy-moving operation images [9]. Additionally, the authors provided a post-processing technique that can improve detection performance even under post-processing settings, like adding noise or compression to mask the manipulation. Research is being done by Bayram et al [10]. To identify picture alteration using binary matching in images. It is considered that the bit planes' binary texture properties will vary between a source and a modified picture. The authors of [10] focused on concentrate on scaling up images, rotating attacks, adjusting brightness values, blurring attacks, and sharpening attacks.

Binary texture statistics are used to gauge how similar binary pictures are. In many cases, identity or emotion transfer is accomplished using graphics-based methods by reconstructing the 3-D models of the source and destination faces, then using the matching 3-D geometry to shift between them. Thies et al. [11] proposed a presentation of an expression switch for face recreation with an RGB-D camera that is particularly noteworthy. A technology called F2F [12] recreates faces in real-time with just an RGB camera. The widely supported [13] transmits the whole 3D-head posture, rotation, emotion, and eye movements from an origin actor to a human face film of a target actor, rather than only changing expression. The visage of "Synthesizing Obama" [15] animates in response to an audio signal input. Face-Swap changes a 3D model's identity while keeping the expressions intact. Convolutional Neural Networks (CNN) are frequently employed to carry out classification tasks with excellent outcomes, and their more recent employment in face verification and classification techniques has shown their significant potential in security monitoring [16]–[18]. The enormous capability of CNNs comes from convolutional layers' capacity to identify numerous characteristics (vital to the task at hand) and derive models that capture ever-more complex ideas as a network's depth grows.

II. Proposed Framework

In this part, we provide a technique for identifying video face modification, determining whether a video frame has a real or manipulated face. In the proposed work, we use advancements in deep learning to capture facial characteristics using CNN technique for forgery detection. To avoid unnecessary computation, we use blaze-face to extract only facial regions. The extracted face is given input to EfficientNet model, which performs compound scaling using a constant ratio to scale up all depth, width, and resolution and classifies the individual face into five classes: original, F2F, Face-Swap, Neural texture, and Deep Fake. The working of the proposed framework is shown in Fig. 1.

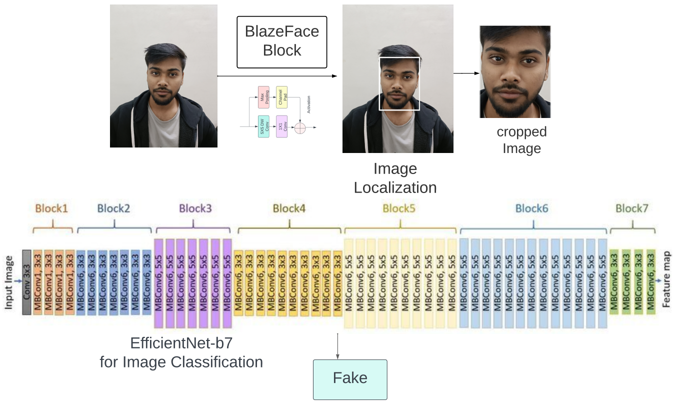

Fig. 1. Flow diagram of the proposed framework.

A. Dataset description

For facial forgery detection, we have used the FaceForencies++ dataset, which consists of 1000 original video sequences manipulated facial videos. The dataset is obtained by applying four facial manipulation methods: Face-Swap, F2F, DeepFake, and neural textures. F2F and Face-Swap are computer graphics-based approaches, and DeepFake and neural texture are two learning-based approaches for facial manipulation in video sequences, each having a source and a target video. DeepFake alters each frame of the target sequence. On the other hand, Face-Swaps and neural textures alter only the minimum number of frames across the source and target sequences.

1) Face-2-Face: F2F is a facial reconstruction system that keeps the target person's identity while transferring the expressions from a source to a target video. The initial version used two video input streams, and keyframe selection was done manually. In order to re-create the face under varying illumination and emotions, these frames are combined to create a complex restructuring of the face. We modify the F2F technique to produce reenactment alterations to analyze our video database automatically. We do a pre-processing run on each video before tracking the facial emotions throughout the remaining frames. Here, we use the first few frames to create a temporary face identity (3D model). Based on such identity rebuilding, we follow the movie and compute each frame's rigid posture, lighting, and expression parameters, much as was done in the basic F2F implementation. We create the reenactment video results by transferring the 76 blend shape coefficients of each frame's source expression parameters to the target video. The original paper has further information on the reenactment method.

2) Face-Swap: Face-Swap is one of the graphics-based approaches for facial manipulation. This method moves a facial region from a source video to a target video by extracting weak feature points of the source face. This approach uses a blend shape for fitting a 3-D specific model using these feature points. By leveraging the texture from the input picture, this model is back-projected onto the target image while minimizing the disparity between the projected form and the local features. Finally, the created model is combined with the photograph, and any necessary color adjustments are made. We repeat these procedures for all source and destination frame pairings until a video is finished. The implementation is efficient on the CPU and has a small computational footprint.

3) DeepFake: A facial image frame from a source video replaces a face in a target frame sequence. The technique is built on two auto-encoders that share an encoder and are trained to reconstruct training pictures of the source and target faces, respectively. In order to trim and align the photos, a face detector is used. The learned encoder and decoder of the original face are applied to the targeted face to produce a false picture. Finally, the auto-encoder output is combined with the remainder of the image using Poisson image editing.

4) Neural-Textures: The Neural-texture of the target individual, comprising a rendering network, is learned using the original video data. This was discovered using an adversarial loss and a photometric reconstruction loss. We employ patch-based GAN-loss in our approach. Tracked geometry is utilized throughout the train and test phases and is the foundation of the Neural Textures technique. To produce these results, we make use of F2F's tracking module. The eye area is left unaltered; we alter the facial emotions related to the mouth region.

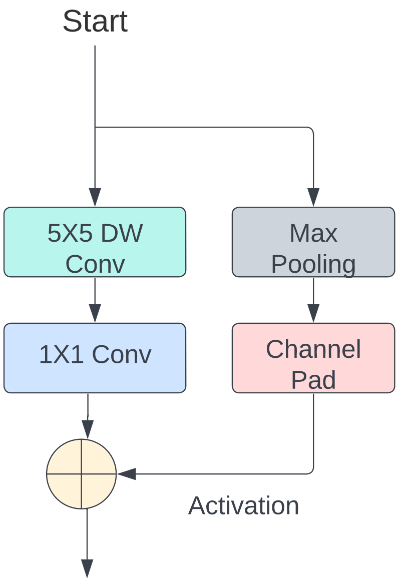

Fig. 2. Working of Blaze-face model for determine the location of the face and its critical points.

A robust face-tracking method blaze-face has been used to process the input image frames. Using blaze-face it is possible to concurrently determine the location of the face and its critical points. The other four crucial parts are the eyes, nose, ears, and mouth. Also conceivable is the simultaneous detection of many We utilize the data to identify the area of the image that the face occupies, and we feed this information into a trained classification network to get the prediction results. Fig 2 shows the working of the Blaze-face model to determine the location of the face and its critical points.

B. Classification using EfficientNet

EfficientNet is a convolutional neural network used for image classification purposes. The EfficientNet is a compound scaling-based neural network that uses a compound coefficient to scale all three dimensions: depth, width, and resolution. In traditional CNNs, dimension scaling is done arbitrarily in terms of network depth. As the network depth increases, similar to increasing the number of layers, the model complexity and its capability to extract complex features increase. This leads to more accurate classification results. The model architecture of the EffecientNet classification is shown in Fig. 3.

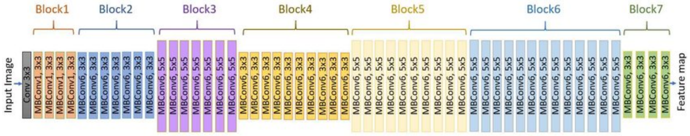

Fig. 3. EffecientNet-Architecture

In EfficientNet, a compound scaling of the network is performed using a constant ratio in all dimensions. It applies a grid search approach to obtain a relationship among all three scaling dimensions of the base network. In in-depth scaling, we increase the number of layers in the network to extract more complex features. In resolution scaling, the resolution of the input image is scaled up. This automatically leads to increases in network depth. In ”width scaling,” we widen the network, and that is to add more channels as you proceed through classification. The variety of feature mappings is increased to extract more data from the image tensors. More extensive networks can represent fine-grained characteristics. The only issue is determining how much the network must be expanded to improve performance. This way, we can obtain appropriate scaling coefficients to scale the base network by the desired dimensions. Based on floating-point-operations-per-second (FLOPS), it improves both precision and efficiency. This newly created architecture uses movable inverted bottleneck convolution (MBConv).

III. Result and Discussion

In this section, the experimental results are discussed based on different benchmark settings. Given an input face, we classify the face into five different classes namely original, F2F, Face-Swap, Neural texture, and DeepFake. As the EfficientNet technique is used for video manipulation of face images for scaling up the dimensions of depth, width, and resolution, the proposed framework is achieved 95.99% accuracy for classifying real faces and manipulated faces using the proposed multiclass classification model.

Although deep learning techniques have demonstrated exceptional effectiveness in detecting face manipulations in video sequences, the quality of manipulation has been growing as they not only manipulate the faces with advanced techniques but also apply post-processing techniques after tampering to remove any traces of face manipulation in video sequences. For performance evaluation of the proposed framework, the confusion matrices-based measures are used for measuring precision, recall, and f1-score. The measured performance of the proposed system is shown in Fig. 4, and Fig. 5, Eq. 1-Eq. 4). The test validation accuracy of the proposed framework is shown in Fig. 6. Based on overall observations, the accuracy of the proposed framework increases when the number of data samples (real faces and manipulated faces) increases per epoch during 5-cross validation results.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4}$$

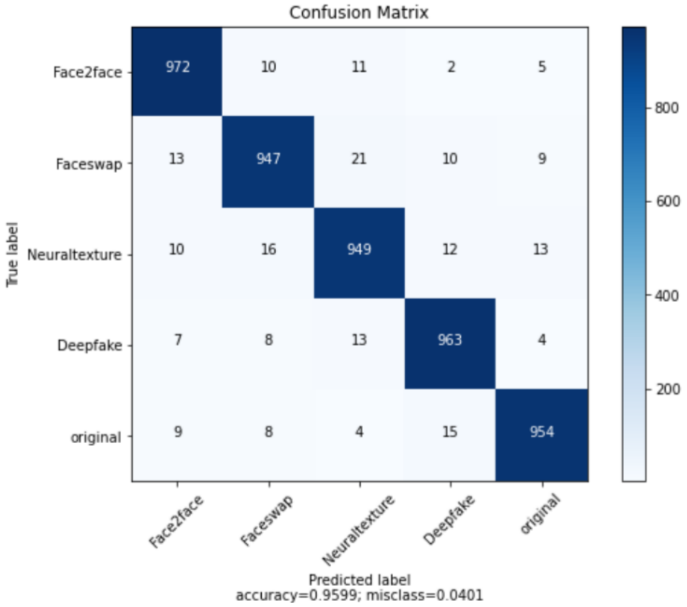

Fig. 4. Fusion Confusion-matrix

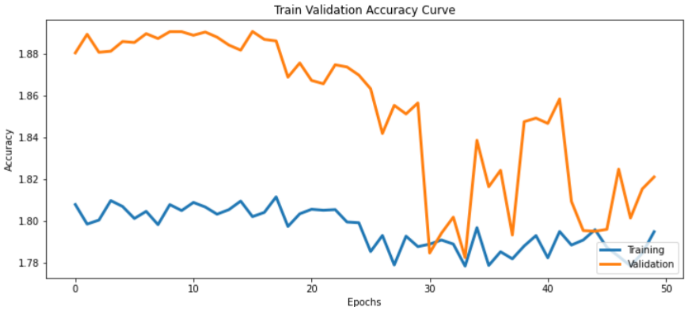

Fig. 5. Train validation Accuracy

Where, TP= True positive, TN = True negative, TP = True positive, FP=False positive.

IV. Conclusion and Future Directions

In this work, a novel framework is proposed for the classification of facial manipulation in video sequences using the

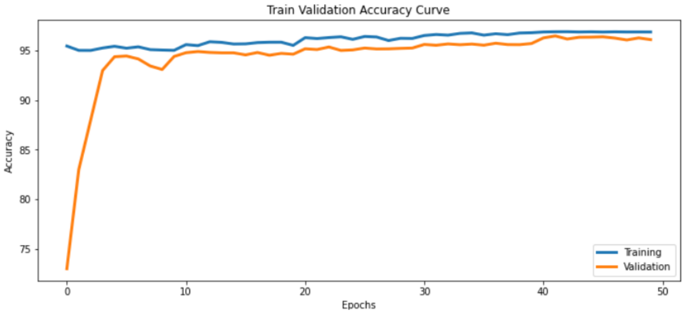

Fig. 6. Test validation Accuracy

blaze-face technique. It is used as a facial region extractor to extract distinct facial regions from video to train models for the classification of real and manipulated faces using the efficient net technique. The proposed framework uses a binary cross-entropy method to compare the predicted probabilities to actual class output for the classification of real face or manipulated faces in video sequences. Based on observations, our proposed framework showed that the detection of manipulated video achieved with high-level accuracy and extraction of the Region of I interest (RoI) from given images took less time compared to the existing method. Future work can be focused on leveraging a multimodal framework for video manipulation in real-time. Another area of research is the introduction of forgery detection technologies into social media platforms to improve their effectiveness in dealing with the pervasive effects of forgery and reducing their effects.

Acknowledgment

This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number SFI/12/RC/2289_P2.

References

[1] H. K. Sahu, S. Kumar, S. H. Alsamhi, M. K. Chaube and E. Curry, ”Novel Framework for Alzheimer Early Diagnosis using Inductive Transfer Learning Techniques,” 2022 2nd International Conference on Emerging Smart Technologies and Applications (eSmarTA), Ibb, Yemen, 2022, pp. 1-7, doi: 10.1109/eSmarTA56775.2022.9935379.

[2] Kumar, Santosh, Mithilesh Kumar Chaube, Saeed Hamood Alsamhi, Sachin Kumar Gupta, Mohsen Guizani, Raffaele Gravina, and Giancarlo Fortino. ”A novel multimodal fusion framework for early diagnosis and accurate classification of COVID-19 patients using X-ray images and speech signal processing techniques.” Computer Methods and programs in biomedicine 226 (2022): 107109.

[3] G. Amato et al., ”Face Verification and Recognition for Digital Forensics and Information Security,” 2019 7th International Symposium on Digital Forensics and Security (ISDFS), 2019, pp. 1-6.

[4] L. Verdoliva, ”Media Forensics and DeepFakes: An Overview,” in IEEE Journal of Selected Topics in Signal Processing, vol. 14, no. 5, pp. 910-932, Aug. 2020.

[5] Kietzmann J, Lee LW, McCarthy IP, Kietzmann TC (2020) Deepfakes: trick or treat? Business Horizons 63(2):135–146

[6] G. Amato et al., ”Face Verification and Recognition for Digital Forensics and Information Security,” 2019 7th International Symposium on Digital Forensics and Security (ISDFS), 2019, pp. 1-6, doi: 10.1109/ISDFS.2019.8757511.

[7] Anderson Rocha, Walter Scheirer, Terrance Boult, and Siome Goldenstein. 2011. Vision of the unseen: Current trends and challenges in digital image and video forensics. ACM Comput. Surv. 43, 4, Article 26 (October 2011), 42 pages.

[8] K. H. Rhee, “Image Forensic Decision Algorithm using Edge Energy Information of Forgery Image,” Journal of the Institute of Electronics and Information Engineers, vol. 51, no. 3. The Institute of Electronics Engineers of Korea, pp. 75–81, 25-Mar-2014.

[9] Jeon, J.-J., Park, S.-H., Kim, Y.-I., and Eom, I.-K., 2014, “Copy-Move Forged Image Detection Using Average of Singular Values of Wavelet Coefficients”, Journal of Korean Institute of Information Technology, 12(11), pp. 119-126.

[10] Korshunov, Pavel, and Sebastien Marcel. ”Deepfakes: a new threat to face recognition.” Assessment and detection (2018).

[11] Thies, J., Zollhöfer, M., Nießner, M., Valgaerts, L., Stamminger, M., & Theobalt, C. (2015). Real-time expression transfer for facial reenactment. ACM Transactions on Graphics (TOG), 34, 1 - 14.

[12] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt and M. Nießner, ”Face2Face: Real-Time Face Capture and Reenactment of RGB Videos,” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2387-2395.

[13] Kim, H., Garrido, P., Tewari, A., Xu, W., Thies, J., Niessner, M., Pérez, P., Richardt, C., Zollhöfer, M. and Theobalt, C., 2018. Deep video portraits. ACM Transactions on Graphics (TOG), 37(4), pp.1-14.

[14] Saleh, H., Alharbi, A., & Alsamhi, S. H. (2021). OPCNN-FAKE: Optimized convolutional neural network for fake news detection. IEEE Access, 9, 129471-129489.

[15] Supasorn Suwajanakorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. 2017. Synthesizing Obama: learning lip sync from audio. ACM Trans. Graph. 36, 4, Article 95 (August 2017), 13 pages.

[16] Granger, E., Kiran, M. and Blais-Morin, L.A., 2017, November. A comparison of CNN-based face and head detectors for real-time video surveillance applications. In 2017 7th IEEE International Conference on Image Processing Theory, Tools and Applications (IPTA) (pp. 1-7).

[17] H. Kavalionak, C. Gennaro, G. Amato, C. Vairo, C. Perciante, C. Meghini, et al., ”Distributed video surveillance using smart cameras”, Journal of Grid Computing, pp. 1-19, 2018.

[18] Mokbal, F. M. M., Wang, D., Osman, M., Yang, P., & Alsamhi, S. H. (2022). An efficient intrusion detection framework based on embedding feature selection and ensemble learning technique. Int. Arab J. Inf. Technol., 19(2), 237-248.

[19] Putra, Tryan & Rufaida, Syahidah & Leu, Jenq-Shiou. (2020). Enhanced Skin Condition Prediction Through Machine Learning Using Dynamic Training and Testing Augmentation. IEEE Access. PP. 1-1.

[20] Belhassen Bayar and Matthew C. Stamm. 2016. A Deep Learning Approach to Universal Image Manipulation Detection Using a New Convolutional Layer. In Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security (IH&amp;MMSec ’16). Association for Computing Machinery, New York, NY, USA, 5–10.

[21] P. Nemani, G. S. Krishna, N. Ramisetty, B. D. S. Sai and S. Kumar, ”Deep Learning based Holistic Speaker Independent Visual Speech Recognition,” in IEEE Transactions on Artificial Intelligence, 2022, doi: 10.1109/TAI.2022.3220190.

[22] Mishra, P., Kumar, S. and Chaube, M.K., 2021. Dissimilarity-based regularized learning of charts. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 17(4), pp.1-23.