IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 5, May 2024

DOI: 10.17148/IJARCCE.2024.13588

3D FACE RECONSTRUCTION AND DEEP FAKE DETECTION

Prof. Nitisha Rajgure[1], Deep Gandhi[2], Mayur Bagade[3], Manisha Badhe[4]

Professor, Department of Computer Engineering, Zeal College of Engineering and Research, Pune, Maharashtra[1]

Student, Department of Computer Engineering, Zeal College of Engineering and Research, Pune, Maharashtra[2-4]

Abstract: In today's rapidly evolving digital landscape, the security and integrity of software applications are paramount. As technology progresses, so do the intricacies of cyber threats, highlighting the critical importance of identifying and resolving vulnerabilities. Addressing this need, we present "3D Face Reconstruction and Deepfake Detection," a project marking a significant advancement in the fields of computer vision and deep learning. We employ Volumetric Convolutional Neural Networks (CNNs) to reconstruct 3D facial models with precision and accuracy, leveraging the feed-forward properties of CNNs to ensure stability and efficiency. This innovative approach enhances the quality of 3D reconstructions, showcasing the potential of deep learning in solving complex real-world problems. Equally important, our project integrates an advanced deepfake detection system using MesoNet, which efficiently identifies synthetic facial images and ensures the authenticity of the reconstructed 3D models. By leveraging a custom dataset that combines various standard datasets, our deepfake detection model achieves high accuracy and robustness, minimizing false positives and negatives. The dual focus on 3D face reconstruction and deepfake detection exemplifies the power of machine learning in capturing intricate facial features and structures while simultaneously safeguarding against digital threats. "3D Face Reconstruction and Deepfake Detection" represents a pivotal step at the intersection of technology and innovation, redefining the processes of 3D face reconstruction and deepfake detection, and making a significant contribution to the fields of computer vision, digital security, and 3D modeling.

I. INTRODUCTION

In the face of an escalating threat posed by facial Deepfakes, this research introduces a pioneering approach centered around the application of Volumetric Regression Networks (VRN) for robust and efficient deepfake detection. As the digital era progresses, the sophistication of facial manipulation techniques like Deepfakes poses significant challenges to the security and integrity of digital identities. The paramount objective of this project is to elevate the accuracy and dependability of authentication processes by harnessing the transformative capabilities of 3D facial reconstruction. By converting 2D facial images into comprehensive volumetric models, this project transcends the limitations inherent in conventional methods, providing a more accurate and detailed analysis of facial features.

A critical component of our approach is the use of MesoNet for deepfake detection, chosen over other technologies due to its effectiveness in capturing mesoscopic properties of images and its relatively low computational requirements. While techniques like GANs and 3DMM fitting offer robust solutions, MesoNet stands out for its balance of efficiency and performance in detecting subtle manipulations in facial imagery. This research aims to establish a cutting-edge solution that mitigates contemporary challenges in facial manipulation and lays the groundwork for a secure, adaptive, and forward-looking paradigm in digital identity verification. By addressing the intricacies of facial Deepfakes, this project seeks to redefine authentication standards, counteracting the dynamic landscape of emerging threats and setting a new benchmark for security in the digital age.

II. 3DMM

Advanced statistical methods called 3D Morphable Models (3DMM) are used to represent 3D facial forms and textures. They allow for point-to-point correspondence between facial reconstructions and make it easier to convert between different types of faces. In order to align the generated face with a photograph, the conventional method uses inverse rendering and optimizes factors like as shape, texture, posture, and lighting. This approach uses 2D and 3D picture datasets along with dimensionality reduction techniques like Principal Component Analysis (PCA). Nevertheless, tiny datasets of about 200 people constrained the diversity and detail that early models, such as the Basel Face Model (BFM), could represent. By building a large-scale 3DMM with scans from 10,000 people, Booth et al. solved this drawback and improved the model's capacity to represent a variety of facial traits.

DOI: 10.17148/IJARCCE.2024.13588

Advanced statistical methods called 3D Morphable Models (3DMM) are used to represent 3D facial forms and textures. They allow for point-to-point correspondence between facial reconstructions and make it easier to convert between different types of faces. In order to align the generated face with a photograph, the conventional method uses inverse rendering and optimizes factors like as shape, texture, posture, and lighting. This approach uses 2D and 3D picture datasets along with dimensionality reduction techniques like Principal Component Analysis (PCA). Nevertheless, tiny datasets of about 200 people constrained the diversity and detail that early models, such as the Basel Face Model (BFM), could represent. By building a large-scale 3DMM with scans from 10,000 people, Booth et al. solved this drawback and improved the model's capacity to represent a variety of facial traits.

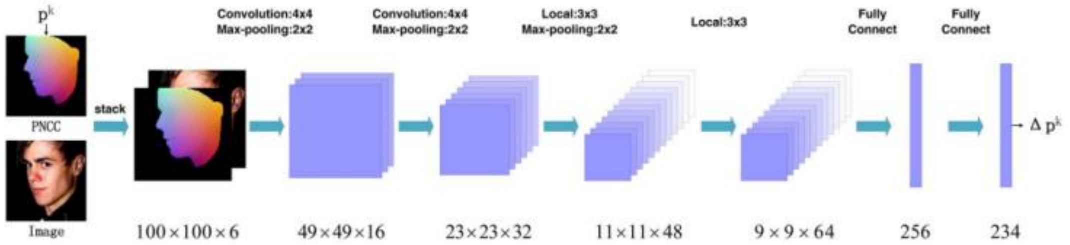


Figure 1: An overview of 3DDFA

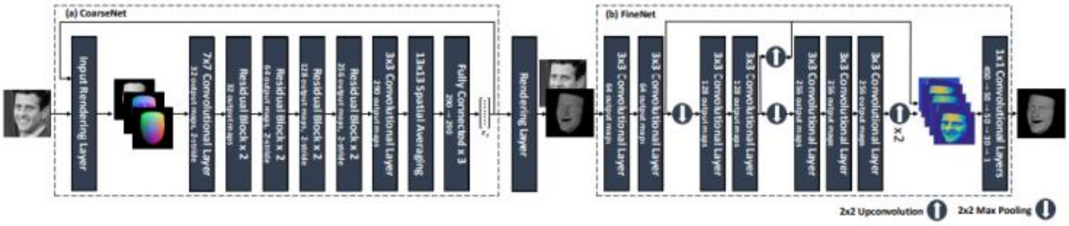


Figure 2: The end-to-end network

Related Work in 3D Face Reconstruction and Depth Estimation Using CNNs:
This section reviews closely related work in 3D face reconstruction, depth estimation with CNNs, and 3D representation modeling with CNNs.

3D Face Reconstruction: While a comprehensive review of the literature on 3D face reconstruction is beyond the scope of this paper, it is highlighted that this method makes minimal assumptions, requiring only a single 2D image to reconstruct a full 3D facial structure, and it works under various poses and expressions. Among single-image methods, the most relevant works are those based on 3D Morphable Model (3DMM) fitting and joint face reconstruction and alignment. For instance, some methods utilize a multi-feature approach to 3DMM fitting with non-linear least-squares optimization, achieving good accuracy with proper initialization. More recent methods have shifted towards estimating 3DMM parameters using CNN regression rather than non-linear optimization. Some approaches estimate 3DMM parameters in multiple steps, each employing a different CNN, focusing on 3D face alignment through a sparse set of landmarks rather than full face reconstruction. Other state-of-the-art methods use a single CNN applied iteratively to estimate model parameters from a 2D image and a 3D representation from the previous iteration. Notable landmark-based 3DMM fitting methods also exist which emphasize incremental refinement of the fit.

Differences in Our Method:

Direct 3D Reconstruction: Unlike the aforementioned methods, our approach bypasses the fitting of a 3DMM altogether, directly producing a 3D volumetric representation of facial geometry.

CNN Architecture: Our method employs a CNN architecture designed to make spatial predictions at the voxel level, in contrast to other networks that predict 3DMM parameters holistically.

DOI: 10.17148/IJARCCE.2024.13588

Versatility: Our method can reconstruct faces from completely unconstrained images, covering a wide range of poses, expressions, and occlusions. Compared to other methods, our approach shows significant performance improvements. Compared to shape-from-shading methods, which capture finer details, our current method does not capture such fine details. However, we believe this limitation is primarily due to the dataset used. With appropriate training data, our method has the potential to learn and replicate finer facial details as well.

CNN-Based Depth Estimation for Facial Structures

Our research is inspired by previous works demonstrating that a Convolutional Neural Network (CNN) can be trained to directly predict depth values from pixels using a single image as input. However, this approach differs from earlier studies in three significant ways. Firstly, while the earlier works focused on general scenes primarily featuring rigid objects, this study concentrates on faces, which are deformable objects. Secondly, instead of learning a mapping from 2D images to 2D depth maps, it is shown that it is possible to map 2D images to complete 3D facial structures, including the non-visible parts of the face. Thirdly, the previous studies utilized a multi-scale approach, processing images from low to high resolution. In contrast, facial images are processed at a fixed scale, assuming this is provided by a face detector. The CNN architecture is based on a state-of-the-art bottom-up top-down module that enables analysis and combination of CNN features at different resolutions, ultimately making predictions at the voxel level. This innovative approach allows for more accurate and detailed 3D facial structure reconstruction from 2D images, leveraging the capabilities of advanced CNN architectures to capture complex facial features.

Recent Work on 3D Face Reconstruction

Recent advancements in 3D face reconstruction have leveraged CNNs to directly regress 3D volumes from images. One notable effort employs LSTM to regress the 3D structure of various object classes from multiple images. Our approach differs in that it treats the reconstruction process as a semantic segmentation problem, regressing a 3D volume spatially aligned with the input image from a single image in one step. This method produces a much larger volume (192 × 192 × 200) compared to the smaller volumes used in previous methods.

Additionally, another method decomposes an input 3D shape into shape primitives to reassemble the given shape using a CNN. Unlike approaches focusing on sparse 3D landmarks for human pose estimation, our method uses a 3D volumetric representation to effectively learn dense 3D facial geometry. Recent methods for 3D Morphable Model (3DMM) fitting have introduced significant improvements, such as using CNNs to produce coarse facial geometry followed by a secondary network for refinement, and employing very deep CNNs for enhanced detail and accuracy. Our approach bypasses traditional 3DMM fitting entirely, directly producing a 3D volumetric representation of facial geometry, which allows for more detailed and accurate spatial predictions at the voxel level, resulting in superior performance and the ability to handle a wide range of facial poses, expressions, and occlusions.

3DFaceGAN

In our project, the technological aspects and theoretical foundations of 3DFaceGAN for 3D facial reconstruction and authenticity detection have been extensively studied. This approach involves unique training processes, network architectures, and loss functions that are highly relevant to the specific application. Although GANs were not implemented directly, the methodologies used in 3DFaceGAN were referenced to inform model development. For example, traditional GANs typically employ discriminator architectures with logit outputs; however, in this analysis, the use of an autoencoder as the discriminator was explored, as suggested in the original 3DFaceGAN framework. Additionally, methods for handling data annotated with multiple labels were considered, further enhancing understanding and capabilities. By studying these advanced techniques, valuable insights were leveraged while employing alternative, more stable architectures like Volumetric CNNs for the project.

Deepfake Detection Technologies:

Meso-4

The Meso-4 network, used in the experiments, features a simplified architecture that maintains effectiveness while enhancing efficiency. It starts with four layers of successive convolutions and pooling, followed by a dense network with one hidden layer.

To enhance generalization, ReLU activation functions are employed in the convolutional layers to introduce non-linearities, along with Batch Normalization to regularize their output and prevent the vanishing gradient effect. Additionally, the fully-connected layers use Dropout to further regularize and improve their robustness. With a total of 27,977 trainable parameters, this network achieves the desired results efficiently.

DOI: 10.17148/IJARCCE.2024.13588

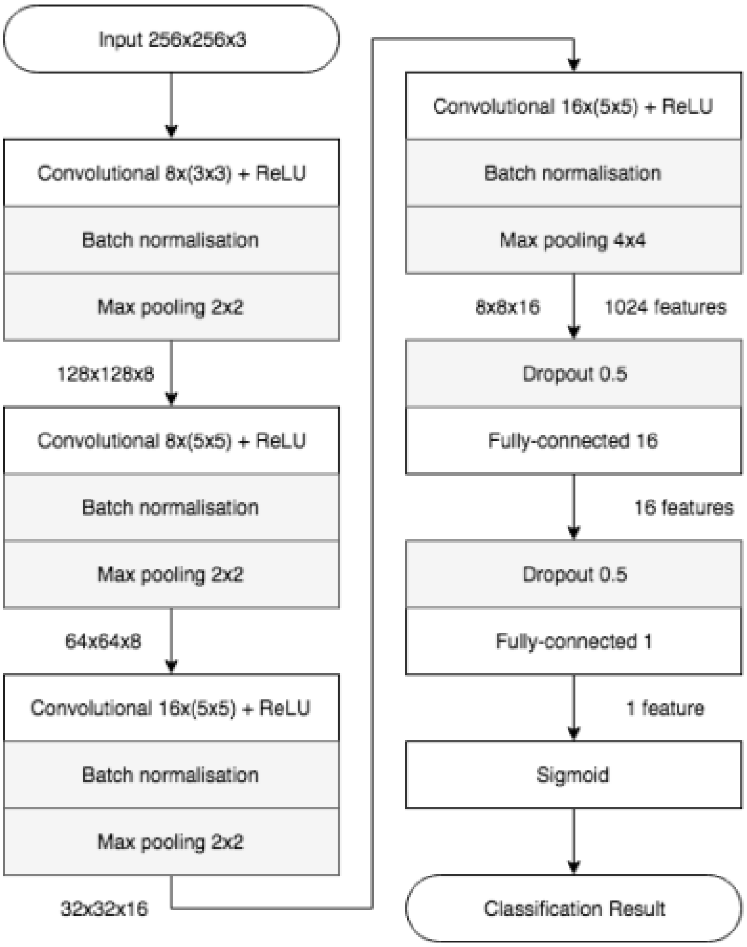


Figure 4. The network architecture of Meso-4. Layers and parameters are displayed in the boxes, output sizes next to the arrows.

XceptionNet:
XceptionNet is a sophisticated deep learning model used for image-based face detection and deepfake detection. It employs depthwise separable convolutions to reduce the number of parameters and computational cost compared to traditional convolutional networks. Trained on extensive datasets, XceptionNet has demonstrated high accuracy in detecting manipulated images, making it a powerful tool in face detection tasks. However, this high performance comes at the cost of substantial computational resources, often necessitating high-end GPUs for both training and inference. This requirement can be a significant limitation in environments with restricted hardware resources.

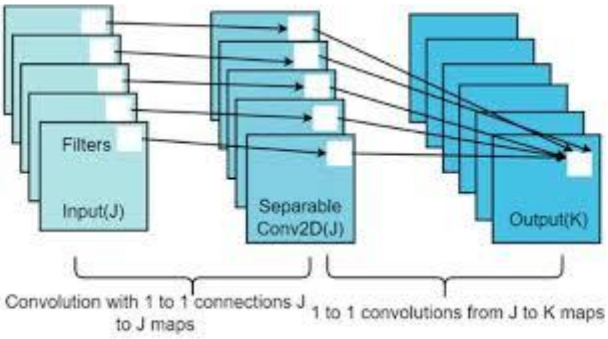

Filters
Input(J)
Separable Conv2D(J)
Output(K)
Convolution with 1 to 1 connections J to J maps
1 to 1 convolutions from J to K maps

Fig6. VRN architecture is similar to above shown Architecture

DOI: 10.17148/IJARCCE.2024.13588

Choice of MesoNet over XceptionNet:

XceptionNet is a sophisticated deep learning model used for image-based face detection and deepfake detection. It employs depth wise separable convolutions to reduce the number of parameters and computational cost compared to traditional convolutional networks. Trained on extensive datasets, XceptionNet has demonstrated high accuracy in detecting manipulated images, making it a powerful tool in face detection tasks. However, this high performance comes at the cost of substantial computational resources, often necessitating high-end GPUs for both training and inference. This requirement can be a significant limitation in environments with restricted hardware resources.

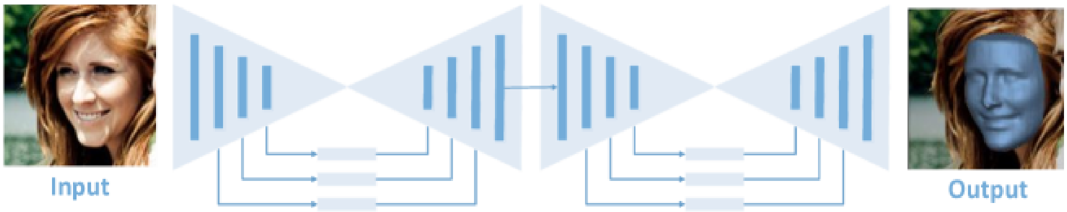
In contrast, MesoNet offers a more computationally efficient solution for deepfake detection. Designed with fewer parameters and a simpler network structure, MesoNet is better suited for real-time applications and devices with limited hardware capabilities. While MesoNet may not achieve the same level of accuracy as XceptionNet, it still provides robust performance, with accuracy rates around 80-85%, and can be deployed more easily in diverse and resource-constrained environments.

Choosing MesoNet over XceptionNet for our project allows us to balance performance and resource efficiency, enabling the deployment of deepfake detection systems without the need for high-end hardware. This decision enhances the accessibility and practicality of our solution, ensuring that it can be widely used in various real-world scenarios where computational resources may be limited.

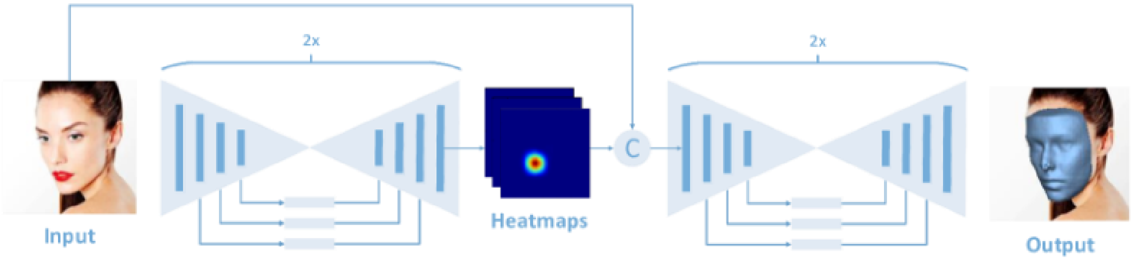
Deepfake Datasets:

The dataset for Deepfake detection is custom-made, comprising a combination of several standard datasets. It includes images used to train autoencoders for the forgery task, which required several days of training with conventional processors to achieve realistic results. Unlike traditional approaches limited to two specific faces, this dataset provides a more diverse range of faces by downloading numerous publicly available videos from the internet. The dataset consists of forged faces with a minimum resolution of 854 × 480 pixels, compressed using the H.264 codec at varying levels.

Additionally, real face images from various internet sources were included, maintaining the same resolutions. To ensure dataset quality, all faces were extracted using the Viola-Jones detector and aligned using a trained neural network for facial landmark detection. The dataset was further refined through manual review to remove misalignment and incorrect face detections. For unbiased classification, both classes (real and forged images) were balanced in terms of resolution quality.

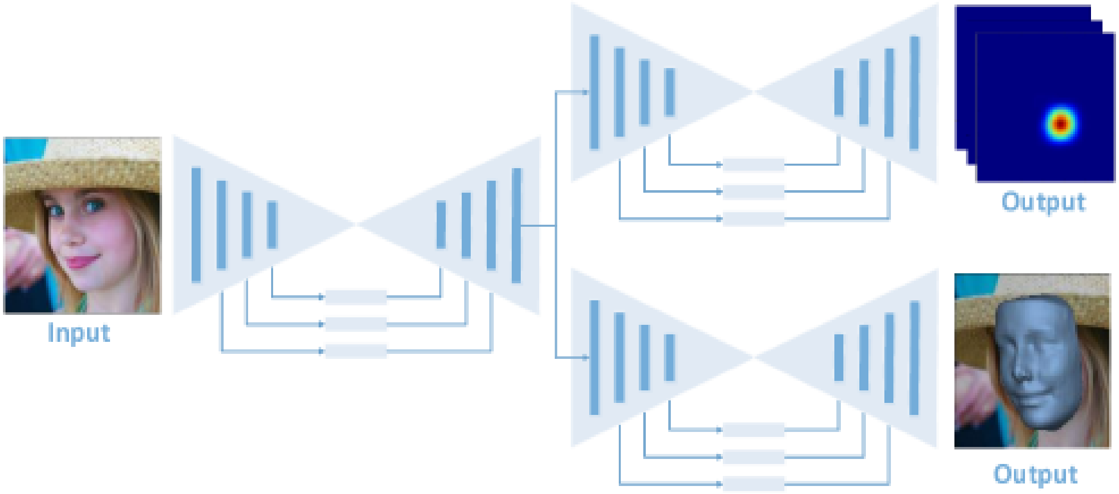


(a)The proposed *Volumetric Regression Network (VRN)* accepts as input an RGB input and directly regresses a 3D volume completely bypassing the fitting of a 3DMM. Each rectangle is a residual module of 256 features.



(b)The proposed *VRN - Guided* architecture firsts detect the 2D projection of the 3D landmarks, and stacks these with the original image. This stack is fed into the reconstruction network, which directly regresses the volume.

DOI: 10.17148/IJARCCE.2024.13588



(c)The proposed *VRN - Multitask* architecture regresses both the 3D facial volume and a set of sparse facial landmarks.

Figure 5: An overview of the proposed three architectures for Volumetric Regression: *Volumetric Regression Network (VRN)*, *VRN - Guided* and *VRN - Multitask*.

Volumetric Regression Network (VRN).

The Volumetric Regression Network (VRN) is designed to map 2D facial images to corresponding 3D volumes. It employs a CNN architecture based on the "hourglass network," consisting of two stacked hourglass modules without intermediate supervision. The input is an RGB image, and the output is a volume of 192 × 192 × 200 real values. The network uses an encoding/decoding structure with convolutional layers to compute a feature representation, which is then processed back to the spatial domain to establish spatial correspondence.

The second hourglass refines the output from the first one. The VRN is trained using the sigmoid cross-entropy loss function and produces a 3D volume at test time, from which the outer 3D facial mesh is recovered. Additionally, the Multitask VRN regresses both 68 facial landmarks and the 3D face structure, and a guided volumetric regression method uses facial landmarks for reconstruction guidance during training and inference.

To enhance the efficiency and robustness of this process, a fast parameter regression strategy based on a lightweight network is proposed, combined with an attention mechanism and Graph Convolutional Networks (GCNs). MobileNet is employed for efficient and quick feature extraction from images, utilizing depthwise separable convolutions to reduce computational cost while maintaining accuracy.

This allows for fast and stable feature extraction, critical for real-time applications. Furthermore, an attention mechanism is incorporated to focus on the context-aware representation of facial features. This mechanism generates attention masks that weight different channels of the feature map, enhancing relevant features and suppressing irrelevant background information. The resulting content-aware matrix ensures that the extracted features are highly relevant to the 3D reconstruction task.

Additionally, both static and dynamic GCNs are introduced to improve the robustness of parameter regression. Static GCNs use a fixed adjacency matrix to perform convolution operations on non-Euclidean structured data, leveraging neighboring node information for state updates.

Dynamic GCNs adapt the adjacency matrix based on input features, allowing for adaptive information spreading and faster learning of local semantic information. By combining these advanced techniques with VRNs, this approach ensures high-quality, realistic 3D model generation from 2D images. This comprehensive methodology leverages the strengths of lightweight networks, attention mechanisms, and GCNs to achieve efficient and robust 3D face reconstruction.

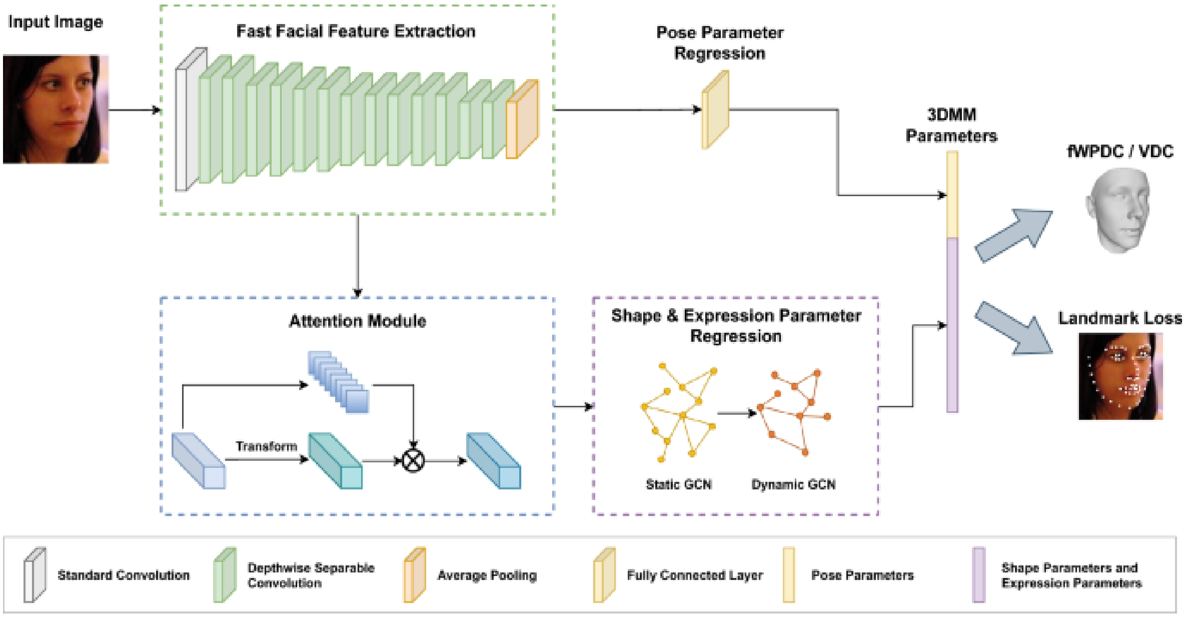

Input Image
Fast Facial Feature Extraction
Pose Parameter Regression
3DMM Parameters
fWPDC / VDC
Landmark Loss
Attention Module
Transform
Shape & Expression Parameter Regression
Static GCN
Dynamic GCN
Standard Convolution
Depthwise Separable Convolution
Average Pooling
Fully Connected Layer
Pose Parameters
Shape Parameters and Expression Parameters

Fig6. VRN architecture is similar to above shown Architecture

Technology Stack and Implementation

It has been proposed to use Django as the primary web framework due to its strong compatibility with Python, which is crucial for integrating deep learning models into the application. Django's robust architecture and ease of use allow for seamless handling of the backend processes involved in deepfake detection and 3D facial reconstruction. The use of Python ensures leveraging powerful deep learning libraries such as TensorFlow and Keras, facilitating the development and deployment of the MesoNet model.

Additionally, the project employs ThreeJS, a JavaScript library, for rendering 3D models. This allows for the visualization of the reconstructed 3D facial models in an interactive and user-friendly manner. By using ThreeJS, the user experience is enhanced, making it possible to view and interact with the 3D models directly in the web browser. The integration of ThreeJS with Django, using Vite for optimized development and build processes, ensures that the application remains responsive and efficient.

Transfer Learning

Using these efficient pre-trained models of VRN and MesoNet ,we optimize the last few layers of their architectures for our own custom dataset and use backpropagation to learn optimal weights ,over multiple training iterations.
In summary, the combination of Django, MesoNet, and ThreeJS provides a robust framework for addressing the challenges of deepfake detection and 3D facial reconstruction. By leveraging the strengths of each technology, the project offers a comprehensive solution that is both effective and accessible, setting a new standard in the field of digital identity verification.

Summarization

The project '3D Face Reconstruction and Deepfake Detection' integrates Django for backend operations, facilitating seamless integration of deep learning models. Users upload images through Django, which are processed using MesoNet for deepfake detection. Further refinement is done with 3D Morphable Model (3DMM) fitting and Volumetric Regression Networks (VRN). CNNs are preferred over GANs for their feed-forward properties and stable hardware requirements. A custom dataset enhances deepfake detection accuracy. Rendered 3D models using ThreeJS and Vite provide an interactive experience, with real-time outputs like evolution scores and confusion matrices for deepfake detection, ensuring high-quality results and robustness.

III. APPLICATIONS AND FUTURE SCOPE

Facial Animation: This project can be used in the entertainment industry for creating realistic facial animations in video games, movies, and virtual reality experiences.

DOI: 10.17148/IJARCCE.2024.13588

Virtual Try-On: E-commerce platforms can use it to offer virtual try-on solutions, allowing customers to see how products like eyeglasses, makeup, or accessories look on their own 3D-rendered face.

Personalized Healthcare: It can be applied in personalized healthcare, where it helps in designing custom-fitted medical devices, such as dental implants, prosthetics, and customized orthodontic treatments.

Virtual Cosmetic Surgery Simulation: Plastic surgeons can use it to simulate the potential results of cosmetic procedures on a patient's face, helping them make informed decisions.

Gaming and Character Customization: In gaming, it can enable players to create highly detailed and realistic character avatars based on their facial features.

Emotion and Expression Analysis: It can be used in emotion recognition applications to analyze facial expressions for user experience improvements or market research.

Security and Surveillance: In security and surveillance systems, this project can enhance facial recognition by providing 3D facial models, making it more robust against deepfake manipulation like variations in lighting, pose, and facial expressions.

IV. LIMITATIONS

Resource-Intensive Processing: The project demands substantial resources, especially during training and real-time processing, requiring high-performance GPUs and significant computational power, which can be costly and resource-intensive.

Complex Development Requirements: Building and maintaining a system based on Volumetric Regression Networks (VRN) and deepfake detection is intricate and resource-intensive, requiring specialized expertise in deep learning and computer vision, making development and maintenance challenging.

Data Privacy Concerns: Handling biometric data, even in a highly secure system, raises data privacy concerns, necessitating strict adherence to data protection measures and compliance with relevant regulations to ensure user privacy.

False Positives and Negatives: While effective, deepfake detection models may still generate false positives or false negatives, inconveniencing legitimate users or leading to false alarms when confronting advanced deepfake techniques.

Evolving Deepfake Threats: With deepfake technology continually advancing, the project may require frequent updates and enhancements to stay ahead of new threats, necessitating keeping the system up-to-date to address the evolving landscape of deepfake techniques.

V. RESULTS

In the Django app ,the binary probabilities of fake or real class of face images ,as calculated by the models, are displayed. In the vite server ,the reconstructed 3D model obj file from the user uploaded 2d face image, is rendered using three.js.

VI. CONCLUSION

Our project effectively combines advanced technologies to tackle the challenging tasks of 3D facial reconstruction and Deepfake detection. Utilizing Django as the backend framework, we provide an intuitive interface for image uploads, which are processed through a sophisticated pipeline. MesoNet is employed for initial image processing, focusing on the mesoscopic properties of images to detect tampering. This ensures robust handling of various input qualities. For 3D reconstruction, we use 3D Morphable Model (3DMM) fitting techniques to create detailed 3D representations from 2D images, and further refine these using Volumetric Regression Networks (VRN). The VRN's architecture, featuring encoding/decoding and hourglass modules, maintains high spatial correspondence and detail in the generated 3D models.

Our custom dataset, combining multiple standard datasets, provides a diverse training set, improving the robustness and accuracy of our models. This dataset ensures our models are well-equipped to handle various scenarios and data variations, which is crucial for reliable Deepfake detection.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 5, May 2024

DOI: 10.17148/IJARCCE.2024.13588

The processed data is converted into JSON format for rendering. Using ThreeJS and Vite, we display the 3D models in a new terminal window, offering dynamic and interactive visualization. Meanwhile, the Django server continues to run, providing real-time outputs such as evolution scores and confusion matrices, essential for evaluating the performance of our detection algorithms. This integrated approach demonstrates the effectiveness of combining cutting-edge technologies in computer vision and machine learning, advancing the fields of facial recognition and reconstruction while contributing significantly to combating Deepfake technologies.

REFERENCES

[1] Xiaoyu Chen, Hongliang Li, Qingbo Wu, King Ngi Ngan, and Linfeng Xu. High-quality r-cnn object detection using multi-path detection calibration network. IEEE Transactions on Circuits and Systems for Video Technology, 31(2):715–727, 2020.

[2] Zhu, X., Lei, Z., Liu, X., Shi, H., & Li, S. Z. (2021). 3D Face Reconstruction from a Single Image with Hierarchical Deep Learning. IEEE Transactions on Pattern Analysis and Machine Intelligence https://ieeexplore.ieee.org/document/9382950. (2021).

[3] Zhao, G., & Pietikäinen, M. (2021). A Comprehensive Study of Deep Learning for 3D Facial Action Unit Detection. Pattern Recognition https://www.sciencedirect.com/science/article/pii/S0031320320304843, (2021).

[4] Ching Y., & Yuxing Mao (2020). A survey of techniques for Face Reconstruction, IEEE Transactions on some approaches in face reconstruction and the methodology. (2021)

[5] Tzionas, D., & Zafeiriou, S. (2021). A Survey on 3D Face Reconstruction Methods. Journal of Imaging. [https://link.springer.com/article/10.1007/s00138-2021-01193-8] (2021).

[6] Stylianos Moschoglou, Stylianos Ploumpis. 3DFaceGAN: Adversarial Nets for 3D Face Representation, Generation, and Translation. arXiv:1905.00307 (2019)

[7] Sam P. Tarassoli & Matthew Shield. Facial Reconstruction: A Systematic Review of Current Image Acquisition and Processing Techniques DOI:10.3389/fsurg.2020.537616 (2020)

[8] Arian Sabaghi Marzieh Oghbaie. Deep Learning meets Liveness Detection: Recent Advancements and Challenges. https://doi.org/10.48550/arXiv.2112.14796 (2022)

[9] Sandra Mau, Farhad Dadgostar A Face Biometric Benchmarking Review and Characterization. (2020).

[10] Zeyu Ruan, Changqing Zou. SADRNet: Self-Aligned Dual Face Regression Networks for Robust 3D Dense Face Alignment and Reconstruction. ResearchGate, (2021)