International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

DeepFake Detection

***Nidhi Kedilaya*[1], *Akshitha Katkeri*[2], *Omkar S*[3]**

[1] Department of Computer Science and Engineering BNM Institute of Technology, Affiliated to VTU Bangalore, India
nidhikedilayaa@gmail.com
[2] Department of Computer Science and Engineering BNM Institute of Technology, Affiliated to VTU, Bangalore, India
akshithakatkeri@bnmit.in
[3] Department of Computer Science and Engineering BNM Institute of Technology, Affiliated to VTU Bangalore, India
omkarsathish06@gmail.com

ABSTRACT :

The recent development of deep learning technology has caused widespread production of highly realistic manipulated media, posing threats to security, privacy, and information integrity that are commonly known as deepfakes. Deepfake videos are synthetic videos that can be used for many malicious purposes. For this paper, we suggest a robust deepfake detection system based on two convolutional neural network (CNN) architectures, Meso-4 and MesoInception-4, to detect fraudulent facial videos produced through DeepFake and Face2Face methods. Both models are specifically tailored for low-complexity and high-accuracy detection tasks, and they are therefore appropriate for real-time applications. We train and test our models on a curated dataset of authentic and manipulated videos and show the strength of mesoscopic analysis in capturing subtle artifacts created during video synthesis. Experimental results indicate that our system is capable of high detection accuracy with minimal computational needs, thus providing a viable solution for deployment in forensic and security-focused settings. Additionally, we contrast the performance of Meso-4 and MesoInception-4, pointing out the strengths and weaknesses when faced with varying forms of manipulations. Our contribution calls attention to the necessity of compact, yet resilient models in countering the rising threat of deepfakes.

Keywords— DeepFake, Face2Face, Image Forensics

1. Introduction

Over the past few years, the advent of deepfake technology has ushered in a new age of synthetic media creation. Deepfakes, developed with advanced deep learning methods like Generative Adversarial Networks (GANs) and autoencoders, facilitate realistic manipulation of facial expressions, voice, and even entire video clips. Facilities such as DeepFake and Face2Face have enabled even novices to create extremely realistic counterfeit videos, raising serious questions regarding misinformation, identity theft, defamation, and loss of public confidence in online content. With increasing quality in the forgeries, it is getting harder and harder for humans as well as traditional automated methods to tell between genuine and faked media.

Increased vulnerability to deepfakes has evoked intensive research into designing reliable detection methods. Most of the current techniques exploit examination of pixel-level inconsistencies, physiological signals, or the artifact of manipulation incurred during the generating process. A detection mechanism balancing high accuracy with low computational efficiency and real-time processing remains desirable. Lightweight convolutional neural network designs have now come forward as strong contenders for this job, with the benefit of detecting fine-grained patterns without the resource cost of deep models.

Here, we introduce a deepfake detection system based on two mesoscopic CNN models, namely Meso-4 and MesoInception-4. These architectures are precisely engineered to detect fake content by concentrating on mid-level spatial aspects, in essence filling the gap between pixel-level noise and high-level semantic inconsistencies. As opposed to conventional deep networks that could overfit to particular forgery forms, mesoscopic models are good generalizers across different manipulation methods. By using these models to identify videos altered through DeepFake and Face2Face techniques, we seek to prove their effectiveness in practical, adversarial situations.

Our contributions are threefold. First, we implement and modify the Meso-4 and MesoInception-4 models for the particular purpose of identifying deepfakes produced by two distinct manipulation techniques. Second, we test their performance on a handpicked dataset, reporting comprehensive metrics that indicate their detection ability. Lastly, we perform comparative analysis between both models, portraying their strengths and weaknesses as well as their appropriateness for diverse contexts of application. Through this effort, we seek to improve knowledge regarding lightweight, efficient deepfake detection strategies as well as make contributions toward developing more secure and reliable digital ecosystems.

1.1. DeepFake

DeepFake technology is a category of synthetic media production techniques that employ deep learning, specifically deep neural networks, to produce hyper-realistic videos, images, and audio recordings in which people seem to say or do something they never did. "DeepFake" is a portmanteau of "deep learning" and "fake," and it came into public view around 2017 when doctored videos of celebrities started spreading all over the internet. Since then, the technology has developed quite quickly, with more advanced techniques allowing more realistic and more difficult-to-spot forgeries.

The foundation of DeepFake video creation is the application of autoencoders and Generative Adversarial Networks (GANs). In a standard deepfake production pipeline, two autoencoders are trained: one to encode and reconstruct the target face (the subject to be superimposed) and another to encode and reconstruct the source face (the subject that shows up in the final video). In inference, the source face is compressed by the encoder into a latent representation, and the decoder maps it back onto the target's face, so that facial expressions, head motions, and lighting conditions are highly aligned to the original video. Some more sophisticated DeepFake systems employ GANs to enhance the photorealism of the output, so that even human viewers and automated detectors have trouble recognizing artifacts.

The most prevalent artifacts connected to early deepfake videos were inconsistencies between the eyes and mouth, abnormal blinking behavior, color misalignment, and boundary artifacts when the imputed face merges with the authentic backdrop. With an advancement in models and the availability of larger datasets, though, these artifacts have significantly diminished. Current deepfake generators are capable of simulating fine details such as skin texture, shadows, reflections, and even minute muscle movements, which makes detection much more difficult. Additionally, methods such as progressive growing of GANs and attention mechanisms have been incorporated into newer models to further improve output quality.

Even with valid applications in entertainment, education, and accessibility (such as dubbing videos in different languages or reviving old footage), deepfake technology is very serious in terms of ethical and security concerns. Abuses include political disinformation campaigns, revenge porn, fraud for financial gain, identity theft, and eroding public trust. In response, there has been pressure across the academic, business, and government communities to build deepfake detection methods that can accurately detect altered media. Knowing the technical basis for how deepfakes are created is important for building effective countermeasures like the mesoscopic models used in our detector.

1.2. Face2Face

Face2Face is a real-time facial reenactment system that allows facial expressions from a source actor to be transferred to a target person in an existing pre-recorded video. Face2Face was presented by Thies et al. in 2016 and was one of the first methods to demonstrate how facial expressions might be manipulated in live or recorded videos in very realistic ways. In contrast to DeepFake, which is based on mainly identity switching using deep learning, Face2Face utilizes classic computer vision and 3D modeling methods for expression manipulation with the identity of the target subject preserved.

The Face2Face pipeline functions at three broad stages: facial tracking, expression transfer, and video synthesis. Initialy, a high-density 3D model of the target's face is reconstructed from monocular RGB input, enabling fine-grained tracking of expressions and facial landmarks. Using similar approaches, the source actor's facial expressions are tracked in real-time simultaneously. In the second step, the source actor's facial expressions are mapped onto the 3D model of the target. This is done by controlling blendshape parameters, a group of predefined facial deformations that describe facial expressions like smiling, frowning, or blinking. Lastly, the 3D facial model after manipulation is re-rendered into the target video carefully preserving texture details, lighting conditions, and head pose to allow for a smooth integration.

One of Face2Face's strongest aspects is its real-time performance, enabling users to change video streams on the fly without the large processing overhang typical of deep learning-based approaches. But the use of classic graphics methods also comes with some recognizable artifacts. Typical manipulations in Face2Face-synthesized videos are slight texture inconsistencies on the skin near the eyes and mouth, awkward transitions in lighting, and small temporal artifacts caused by poor facial blending between frames. However, with rising video resolution and advances in algorithms, these artifacts have grown progressively undetectable, making them harder to detect.

Although Face2Face has valid uses in video conferencing, special effects, and virtual reality, it is also very dangerous when not used correctly. Fake videos produced with Face2Face can easily mislead people, perpetuate misinformation, and harm reputations. Notably, since Face2Face distorts expressions without modifying the speaker's voice, it can be highly misleading in altering emotional content, e.g., showing a public person to be angry, sad, or dishonest. Identifying and detecting these manipulations is of utmost importance, and our contribution is in identifying the fine-grained mesoscopic-level artifacts added during the reenactment process using light-weight CNN models such as Meso-4 and MesoInception-4.

2. Proposed Method

This section describes a set of effective approaches to tackle either the Deepfake or Face2Face problems. It has been noticed that neither of these problems can be solved well using one kind of unified network. However, because of the similar nature of these forgery methods, using identical network architectures for both can yield desirable results.

Our method is centered on detecting manipulated facial videos through analyzing them at the mesoscopic level. This option is based on the limitations of microscopic approaches, which are based on image noise—frequently degraded in compressed videos. Alternatively, on a higher abstract semantic level, the human vision system has problems perceiving manipulations [21], especially in images with human faces [1, 7]. Therefore, we suggest an intermediate approach that utilizes a deep neural network with a comparatively shallow architecture.

Of the configurations we tried, two network architectures provided the best classification accuracy. These models, which have a low representation level and surprisingly few parameters, are based on highly accurate image classification networks [14, 23]. They are composed of alternating convolution and pooling layers for feature extraction, with a dense layer for final classification. Source code for these models is available online[1].

2.1. Meso-4

Meso-4 is a light-weight convolutional neural network (CNN) architecture that is particularly suited for forged facial video detection, e.g., generated by DeepFake and Face2Face technologies. Introduced by Afchar et al. in 2018, the fundamental concept of Meso-4 is to concentrate on mesoscopic-level features — features that fall between low-level (pixel) details and high-level semantic information. This middle feature level is best suited for deepfake detection since manipulations tend to add subtle inconsistencies that are hard to detect at either end.

The Meso-4 architecture has four convolutional layers followed by a dense (fully connected) network. Every convolutional layer has a fairly small number of filters, keeping the model light and computationally efficient. More specifically, the architecture can be described as:

- Conv Layer 1: 8 filters of size 3x3, batch normalization, and ReLU activation.
- Conv Layer 2: 8 filters of size 5x5, batch normalization, ReLU, and max pooling.
- Conv Layer 3: 16 filters of size 5x5, batch normalization, ReLU, and max pooling.
- Conv Layer 4: 16 filters of size 5x5, batch normalization, ReLU, and max pooling.

Fully Connected Layers: A single dense layer consisting of 16 units, culminating in the final output layer with sigmoid activation (in the case of binary classification: fake vs. real).

The architectural decision of the use of tiny convolutional blocks enables the network to effectively recognize localized facial artefacts like slight texture inconsistencies, unnatural blending of facial boundaries, or inconsistencies within fine facial detail like eyes and lips — all of which are common markers for synthetic media. The max pooling operations following every convolution aid in the diminishment of the spatial dimensions but retaining the most significant features, thus avoiding overfitting and encouraging generalization to various kinds of fake videos.

One of the key strengths of Meso-4 is that it strikes a balance between precision and computational expense. In contrast to deeper networks that need extensive memory and computation, Meso-4 can be implemented in real-time or low-resource environments like mobile platforms or browser-based forensic software. Even though it is simple, Meso-4 has been shown to exhibit robust performance at detecting a variety of types of manipulations and is therefore an appealing option for constructing practical deepfake detection systems.

We employ Meso-4 as one of the fundamental models in our work to identify forgeries created by both DeepFake and Face2Face algorithms. The lightness of Meso-4 allows our system to maintain speed and efficiency, while its emphasis on mesoscopic features creates a solid basis for identifying minute manipulations that can go unnoticed to human observers.

2.2. MesoInception-4

Y MesoInception-4 is a superior version of the initial Meso-4 model, created to further enhance deepfake detection accuracy by integrating inception modules into the network design. Introduced by Afchar et al. together with Meso-4, MesoInception-4 seeks to extract more sophisticated spatial features by allowing the model to process information at different scales at the same time. This multi-scale method assists in detecting various forms of artifacts and inconsistencies that could be in fake media created by methods such as DeepFake and Face2Face.

The basic principle of the inception module is to perform several convolution operations with varying kernel sizes in parallel and join their outputs. In MesoInception-4, the initial two convolutional layers of Meso-4 are substituted with two altered inception modules:

- Inception Module 1: Combines convolutions with 1x1 and 3x3 kernels in parallel to look at fine-grained and mid-level features.
- Inception Module 2: Combines convolutions with 1x1 and 5x5 kernels in parallel to pay attention to larger patterns and wider spatial context. The result of these parallel convolutions are concatenated along the depth axis so that the network can learn an richer and diverse set of feature early in the network.

After these two inception modules, the architecture continues similarly to Meso-4:

- Two standard convolutional layers (with 16 filters and 5x5 kernels) each followed by batch normalization, ReLU activation, and max pooling.
- A flattening layer to transition from feature maps to a dense representation.
- A dense layer with 16 units and ReLU activation.
- A final output layer with a sigmoid activation for binary classification (real vs. fake).

The application of inception modules enables MesoInception-4 to more effectively cope with the diversity of artifacts generated by various forgery techniques. For instance, DeepFake forgeries tend to create local artifacts near facial edges, whereas Face2Face might create larger artifacts as a result of expression warping. Through the examination of features at multiple scales in parallel, MesoInception-4 enhances its capacity to recognize these fine-grained anomalies.

Even with its increased complexity relative to Meso-4, however, MesoInception-4 is lightweight and efficient. It can perform in near real-time, thus being ideal for real-world use where speed as well as accuracy are both vital. In this work, MesoInception-4 acts as a supplement to Meso-4 and offers a stronger baseline for detecting identity swaps and facial reenactments.

3. Experiments

In this work, we tested the performance of the Meso-4 and MesoInception-4 models for deepfake video detection. The experiments were performed on two main datasets: our own DeepFake dataset and the Face2Face dataset with videos manipulated using the Face2Face technique. The DeepFake dataset contains 175 forged video samples of different lengths and resolutions, whereas the Face2Face dataset comprises more than 1,000 manipulated videos. The two data sets were preprocessed by cropping face regions from the videos and rectifying them to make the frames consistent. The data sets were divided into training, validation, and test sets, each with an equal number of real and false videos, and the precise division is summarized in Table 2.

To train the models, we utilized stochastic gradient descent (SGD) with learning rate 0.001 with the goal to reduce the binary cross-entropy loss function. We initialized the networks with random weights and trained for 50 epochs, with early stopping in terms of validation performance to prevent overfitting. To increase the robustness of the models, data augmentation strategies like random cropping, flipping, and color jittering were performed during training. This enabled the models to generalize more effectively, especially when handling variations in the video content.

One of the most important elements of the experiments was to test the effect of frame fusion on detection accuracy. Rather than using single frames for classification, we averaged the network's predictions across several frames from the same video to mitigate the effect of fleeting artifacts like motion blur, face occlusion, or sporadic mispredictions. For every video, we used 30 frames and calculated the mean predicted probability over them. This was tested on both the DeepFake and Face2Face data and compared to predictions over single frames.

Apart from testing frame fusion, we also examined the models' robustness against different compression levels. For the DeepFake dataset, videos were compressed with the H.264 codec at three levels: lossless, light compression (rate 23), and heavy compression (rate 40). The same compression configurations were used for the Face2Face dataset. The performance of the models was checked at each level of compression to study how compression affected detection accuracy. We noticed that although the models worked well under light compression, accuracy decreased substantially with heavy compression, especially on the Face2Face dataset, where delicate facial details were more distorted by the compression process.

For the evaluation of the performance of the models, we employed typical classification metrics like accuracy, precision, recall, and F1-score. Accuracy was defined as the proportion of frames or videos correctly classified, while precision and recall gave indications of the model's capacity to distinguish between fake and real content and catch all instances of fake content, respectively. The F1-score was employed as a balanced indicator of classification performance. Furthermore, we employed a confusion matrix to examine the counts of false positives and false negatives for both real and fake videos.

The experiment results, listed in Table 5, indicate that the MesoInception-4 model performed better than Meso-4 on both datasets, with detection accuracy exceeding 98% on the DeepFake dataset and comparable performance on the Face2Face dataset. The frame fusion approach significantly improved detection accuracy, particularly for videos with motion blur or partial face occlusion. But compression did make a visible difference in performance, most specifically at heavier compression, where both models struggled to hold high accuracy.

3.1. Datasets

3.1.1 DeepFake

One of the challenges encountered while developing this project was the non-availability of a publicly available dataset that would be solely related to videos developed through the DeepFake technique. Considering this, we chose to create our own custom dataset designed for this purpose.

Creating high-quality forgeries with auto-encoders is computationally intensive, taking several days to generate plausible results on regular processors. Additionally, training auto-encoders usually involves training on just two faces at a time, which restricts the dataset diversity. To address these constraints and add more facial diversity, we decided to gather a diverse set of publicly available videos from different online sources.

The dataset we created includes 160 video samples, all of which have forged content created through the DeepFake process. The videos range in length from two seconds to three minutes, providing a broad spectrum of scenarios and complexity. All the videos were downloaded from the internet and have a minimum resolution of 855×480 pixels. To further make the test environment more realistic, the videos were compressed with the H.266 codec but at different levels of compression. This configuration assists in modeling a real-world situation where videos can go through varying compression levels when transmitted or stored. Another dataset, which is described in Section 3.1.2, was employed to uniquely examine how compression affects the performance of models.

The face regions in the videos were obtained with a standard face detection algorithm, and facial alignment was done with a dedicated neural network trained for facial landmark detection. To guarantee balanced representation of faces under varying conditions, the number of face samples from each video was normalized based on the variation in camera angles and changes in lighting. On average, approximately 40 face samples were obtained from each video scene.

In addition to enhancing the dataset, an equal quantity of genuine facial images were incorporated. These genuine images were also obtained from the internet, keeping the same resolution and quality levels as the fake video samples. A manual filtering process with careful attention was used to eliminate any images with misalignment or malfunctioning face detection. Care was taken to ensure that the proportion of high- and low-quality images in both the real and fake categories was kept roughly constant, so as to avoid any bias in classification.

The breakdown of the number of images per class and their division into training and test sets is detailed in Table 2.

3.1.2. Face2Face

Together with the DeepFake dataset, we also aimed to evaluate whether or not the described network architecture would be able to detect other facial forgery techniques. A suitable benchmark for that purpose is the Face2Face dataset, a collection of more than a thousand manipulated videos obtained through the application of the Face2Face approach, as well as their corresponding original videos. The Face2Face dataset is pre-split into test, validation, and training sets, which made it easier to incorporate this data into our model evaluation.

Another advantage of utilizing the Face2Face dataset is that it has videos in lossless compression. This made it easier for us to test the model with different compression scenarios. In order to make sure that the results were in line with the performance measures set in earlier work on this dataset, we used the same compression parameters for H.264: lossless compression, light compression at rate 24, and heavy compression at rate 42. Using these compression parameters allowed us to test how well the network generalizes to videos with varying amounts of compression without compromising the quality of the faces in the videos.

For our experiment, we picked 350 Face2Face videos to use for training. From the test set, we employed 250 forged videos and their corresponding original versions to test the model's performance. Table 2 shows a breakdown of the number of facial images detected per class, giving a clear description of the dataset composition used to train the model as well as for testing.

3.2. Classification Setup

Set	**forged** class	**real** class
DeepFake Training	5111	7250
DeepFake Testing	2889	4259
Face2Face Training	4500	4500
Face2Face Testing	3000	3000

Table 2. Cardinality of each class in the studied datasets. Note that for both datasets, 9% of the training set was used during the optimization for model validation.

3.3. Image Classification Result

The classification scores of the trained network on the Deepfake dataset are presented in Table 3. Both networks yielded similar scores of approximately 90%, as confirmed by independent frame analysis. We do not anticipate a better score because the dataset includes some facial images with very low resolutions.

Network	DeepFake Classification score		
Class	forged	real	total
Meso4	0.771	0.890	0.780
MesoInception-4	0.823	0.800	0.860

Table 3. Classification scores of several networks on the Deepfake dataset, considering each frame independently.

Table 4 shows results for Face2Face forgery detection. A significant drop in scores occurred for the higher levels of compression for the videos. Improved classification performance was achieved when utilizing the state-of-the-art image classification network Xception, as reported in the original paper publishing the FaceForensics dataset [20], which we employed in experiments. However, in the setup presented in that paper, we were restricted to fine-tuning Xception to a 96% score for level 0 of compression and to a 94% score for level 23. The disparity raises suspicions in interpreting the results.

Network	Face2Face Classification score		
Compression Level	0	23(light)	40(strong)
Meso4	0.771	0.890	0.780
MesoInception-4	0.823	0.899	0.860

Table 4. Classification scores of several networks on the Face-Forensics dataset, considering each frame independently.

3.4. Image Aggregation

One of the main difficulties of web video analysis is the process of compression, which usually causes heavy data loss. But if we make use of several frames of the same face in a video, then we can use more information and possibly make the overall classification more accurate. An effective way to do this improvement is to take the average of the predictions on all the frames of a video. Although there is no theoretical reason to anticipate a significant improvement in precision or to use confidence intervals, since successive frames are usually strongly correlated, real videos of faces usually have a large number of stable and unobstructed frames. This stability can weaken the impact of problems like motion blur, partial occlusion, or the occasional misclassification, since most of the frames will produce correct predictions.

The experimental results shown in Table 5 confirm that the performance of the detection system was greatly enhanced with frame fusion by aggregating predictions across multiple frames. In particular, the MesoInception-4 model was able to obtain a detection rate over 95% on the DeepFake dataset when using frame fusion. For the Face2Face dataset, the two models showed comparable overall accuracy. Yet, the kinds of videos that were wrongly classified varied between the models, which underlines the specific strengths and weaknesses of each method in dealing with different kinds of facial forgeries.

Network	Aggregation Score	
Dataset	DeepFake	Face2Face
Meso4	0.858	0.842
MesoInception-4	0.873	0.842

Table 5. Video classification scores on the two dataset using image aggregation, with the Face2Face dataset compressed at rate 23.

3.5. Aggregation Into Frames

To further assess the impact of video compression on forgery detection, we applied the same frame aggregation method but to intra-frames (I-frames) from compressed videos alone—those images that are not interpolated over time. This was meant to examine if the diminished compression artifacts in

I-frames would result in enhanced classification accuracy. However, the drawback of this approach is that brief videos lasting a few seconds can have a minimum of only three I-frames, reducing the smoothing effects normally achieved with frame aggregation.

Our findings indicated that this strategy had a minor adverse effect on classification accuracy, as indicated in Table 6. In spite of this, employing I-frame aggregation would still be an efficient and quicker option since the performance tended to be better than classifying single frames in isolation.

Network	I-Aggregation Score	Difference
Meso4	0.821	-0.026
MesoInception-4	0.848	-0.014

Table 6. Classification score variation on the Deepfake dataset using only I-frames.

3.6. Intuition behind the network

We have noticed that convolutional neural networks carry out classification based on utilizing learned weights of convolution kernels and neurons as feature descriptors of images. For instance, a combination of a positive number, a negative number, followed by a second positive number can be read as describing a discrete second-order derivative. The interpretation is largely meaningful at the first convolutional layer and has limited meaning otherwise in the context of intricate structures like human faces.

A second method of bettering one's insight into a network's inner mechanisms is by creating input images to maximize the response of individual filters. Denoting the response of filter j in layer i for input x by $f_{ij}(x)$, we may initialize x to a noise image and subsequently optimize, including an addition of a regularization term for limiting noise and keeping the resulting patterns realistic. The optimization goal is now $E(x) = f_{ij}(x) - \lambda\|x\|^p$. By doing this, we can see the kinds of patterns to which various filters are most responsive. Figure 7 illustrates the images that maximally activate a number of neurons from the last hidden layer of the Meso-4 model. These neurons may be classified according to the sign of the weight linking them to the output layer: neurons with positive weights tend to emphasize features important for distinguishing real faces, whereas those connected with negative weights emphasize more characteristics associated with imitations. Surprisingly, positively contributing neurons to real image classification tend to respond to areas near the eyes, nose, and mouth—highly detailed regions. By contrast, those neurons involved in fake image classification tend to be activated by less detailed background regions. This aligns with our knowledge that deepfake-generated faces tend not to have fine-grained textures and are more likely to look smoother or blurrier than the unperturbed background.

Additionally, we can learn more by comparing average activation values throughout layers between batches of real and fake images, noting how patterns of activation change, and recognizing areas that play key roles in classification choices. Analyzing the trained MesoInception-4 model on the DeepFake dataset, as shown in Figure 8, reveals the following pattern clearly: real images have stronger activation in the areas of the eyes, whereas fake images have enhanced activation in areas of the background. This is mainly because of the intrinsic fuzziness in forged faces, where significant facial features become unclear, and hence the network unintentionally turns attention towards the relatively unaltered background areas during dimensionality reduction.

3.7. Conclusion

In the current work, we introduced a deepfake detection system that utilizes lightweight mesoscopic convolutional neural networks, namely the Meso-4 and MesoInception-4 models, to efficiently differentiate genuine and forged facial videos. Detection of deepfakes produced by two leading manipulation methods — DeepFake and Face2Face — was our concern since they are particularly challenging due to their capability of creating realistic forgeries with unnoticeable artifacts. By focusing on mesoscopic-level characteristics, our system can detect subtle inconsistencies that tend to pass human and common detection algorithms.

The Meso-4 model with its efficient and simple architecture exhibits high performance in detecting localized artifacts and is well suited for real-time detection in resource-limited environments. Conversely, the MesoInception-4 model improves this ability by adding inception modules that enable multi-scale feature extraction, thus making the model more sensitive to a wider variety of forgery artifacts. These models collectively provide a strong, lightweight solution for deepfake detection, striking a balance between high detection accuracy and low computational expense.

Our experiments validated that mesoscopic networks, even with their comparatively modest depth, are extremely efficient in deepfake detection applications, especially when trained on heterogeneous datasets containing various forgery methods. The mutual strengths of Meso-4 and MesoInception-4 also emphasize the role of design architectural decisions in improving deepfake detection capacity without recourse to exceedingly deep or resource-demanding networks.

As deepfake generation techniques improve, subsequent work will include further refining detection models by including temporal features over video frames, investigating adversarial training approaches, and extending detection to include emerging manipulation methods. However, the performance demonstrated using Meso-4 and MesoInception-4 is a major advance toward developing functional, real-time deepfake detection systems that can assist in alleviating the increasing threat of synthetic media in the digital environment.

3.8. REFERENCES

[1] B. Balas and C. Tonsager, "Face animacy is not all in the eyes: Evidence from contrast chimeras," Perception, vol. 43, no. 5, pp. 355–367, 2014.

[2] M. Barni, L. Bondi, N. Bonettini, P. Bestagini, A. Costanzo, M. Maggini, B. Tondi, and S. Tubaro, "Aligned and non-aligned double JPEG detection using convolutional neural networks," Journal of Visual Communication and Image Representation, vol. 49, pp. 153–163, 2017.

[3] B. Bayar and M. C. Stamm, "A deep learning approach to universal image manipulation detection using a new convolutional layer," in Proc. 4th ACM Workshop on Information Hiding and Multimedia Security, pp. 5–10, ACM, 2016.

[4] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," arXiv preprint arXiv:1610.02357, 2017.

[5] F. Chollet et al., "Keras," [Online]. Available: https://keras.io, 2015.

[6] D. Erhan, Y. Bengio, A. Courville, and P. Vincent, "Visualizing higher-layer features of a deep network," University of Montreal, vol. 1341, no. 3, p. 1, 2009.

[7] S. Fan, R. Wang, T.-T. Ng, C. Y.-C. Tan, J. S. Herberg, and B. L. Koenig, "Human perception of visual realism for photo and computer-generated face images," ACM Transactions on Applied Perception (TAP), vol. 11, no. 2, p. 7, 2014.

[8] H. Farid, "A survey of image forgery detection," IEEE Signal Processing Magazine, vol. 26, no. 2, pp. 26–25, 2009.

[9] P. Garrido, L. Valgaerts, O. Rehmsen, T. Thormahlen, P. Perez, and C. Theobalt, "Automatic face reenactment," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), pp. 4217–4224, 2014.

[10] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," arXiv preprint arXiv:1502.03167, 2015.

[11] T. Julliand, V. Nozick, and H. Talbot, "Image noise and digital image forensics," in Digital-Forensics and Watermarking: 14th Int. Workshop (IWDW 2015), vol. 9569, pp. 3–17, Tokyo, Japan, Oct. 2015.

[12] D. E. King, "Dlib-ml: A machine learning toolkit," Journal of Machine Learning Research, vol. 10, pp. 1755–1758, 2009.

[13] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in Advances in Neural Information Processing Systems, pp. 1097–1105, 2012.

[15] J.-W. Lee, M.-J. Lee, T.-W. Oh, S.-J. Ryu, and H.-K. Lee, "Screenshot identification using combing artifact from interlaced video," in Proc. 12th ACM Workshop on Multimedia and Security, pp. 49–54, ACM, 2010.

[16] S. Milani, M. Fontani, P. Bestagini, M. Barni, A. Piva, M. Tagliasacchi, and S. Tubaro, "An overview on video forensics," APSIPA Transactions on Signal and Information Processing, vol. 1, 2012.

[17] N. Rahmouni, V. Nozick, J. Yamagishi, and I. Echizen, "Distinguishing computer graphics from natural images using convolutional neural networks," in IEEE Workshop on Information Forensics and Security (WIFS), Rennes, France, Dec. 2017.

[18] Y. Rao and J. Ni, "A deep learning approach to detection of splicing and copy-move forgeries in images," in IEEE Int. Workshop on Information Forensics and Security (WIFS), pp. 1–6, 2016.

[19] J. A. Redi, W. Taktak, and J.-L. Dugelay, "Digital image forensics: A booklet for beginners," Multimedia Tools and Applications, vol. 51, no. 1, pp. 133–162, 2011.

[20] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics: A large-scale video dataset for forgery detection in human faces," arXiv preprint arXiv:1803.09179, 2018.

[21] V. Schetinger, M. M. Oliveira, R. da Silva, and T. J. Carvalho, "Humans are easily fooled by digital images," arXiv preprint arXiv:1509.05301, 2015.

[22] W. Shi, F. Jiang, and D. Zhao, "Single image super-resolution with dilated convolution based multi-scale information learning inception module," arXiv preprint arXiv:1707.07128, 2017.

[23] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[24] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," Journal of Machine Learning Research, vol. 15, no. 1, pp. 1929–1958, 2014.

[25] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2015.

[26] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2Face: Real-time face capture and reenactment of RGB videos," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), pp. 2387–2395, 2016.

[27] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), vol. 1, pp. I–I, 2001.

[28] W. Wang and H. Farid, "Exposing digital forgeries in video by detecting double MPEG compression," in Proc. 8th Workshop on Multimedia and Security, pp. 37–47, ACM, 2006.

[29] W. Wang and H. Farid, "Detecting re-projected video," in International Workshop on Information Hiding, pp. 72–86, Springer, 2008.

[30] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," arXiv preprint arXiv:1511.07122, 2015.