Detection of Fake and Fraudulent Faces via Neural Memory Networks

Tharindu Fernando\textsuperscript{a}, Member, IEEE, Clinton Fookes\textsuperscript{b}, Senior Member, IEEE, Simon Denman\textsuperscript{c}, Member, IEEE, and Sridha Sridharan\textsuperscript{d}, Life Senior Member, IEEE

Abstract—Advances in computer vision have brought us to the point where we have the ability to synthesise realistic fake content. Such approaches are seen as a source of disinformation and mistrust, and pose serious concerns to governments around the world. Convolutional Neural Networks (CNNs) demonstrate encouraging results when detecting fake images that arise from the specific type of manipulation they are trained on. However, this success has not transitioned to unseen manipulation types, resulting in a significant gap in the line-of-defense. We propose a Hierarchical Attention Memory Network (HAMN), motivated by the social cognition processes of the human brain, for the detection of fake faces. Through visual cues and by utilising knowledge stored in neural memories, we allow the network to reason about the perceived face and anticipate its future semantic embeddings. This renders a generalisable face tampering detection framework. Experimental results demonstrate the proposed approach achieves superior performance for fake and fraudulent face detection.

Index Terms—Neural memory networks, fake and fraudulent face detection, image and video forensics.

I. INTRODUCTION

W ith social media becoming a primary source of information for many people, and with more than 100 million hours of video content watched daily \cite{1}, fake news has quickly risen as a significant threat to society and democracy. As demonstrated by the DeepFake app \cite{2}, the effortless and seamless nature of synthesising realistic fake content could have a crucial impact on people’s lives and severely undermine the trust in digital media \cite{3}, \cite{4}. In addition, for different kinds of applications such as biometric-based authentication \cite{5}, \cite{6} filtering out forged images and videos is a prominent issue \cite{3}.

While deep neural networks \cite{1}, \cite{3}, \cite{7} have attained striking levels of accuracy when detecting certain types of image tampering attacks, their performance is often disappointing when presented with images or videos with unseen manipulation methods \cite{8}, \cite{9}. It is observed that the underlying neural network can quickly overfit to a specific artefact left by the tampering method observed during training \cite{9}, and thus methods lack transferability. Furthermore, methods for image tampering detection are not directly transferable to videos due to artefacts and quality loss resulting from the video compression. As such, there exists a critical need for automated systems which capture the fundamentals of video and image content tampering, and have the ability to generalise to different manipulation attacks.

In this paper, we propose a deep learning framework to detect fake and fraudulent faces in images and video which is motivated by the social perception \cite{10} and social cognition \cite{11} processes in the human brain. Recent neuroscientific research \cite{12}, \cite{13} has reported that humans try to predict the observed person’s mental state through visual cues such as facial expression and by utilising specific knowledge stored in the observer’s memories. Furthermore, the authors of \cite{13} conclude that during this prediction process, the lack of social presence and social attributes cause the uncanny valley effect \cite{12} in the brain when humans view computer generated and tampered media. Inspired by the uncanny valley effect of the human brain, we propose a deep learning framework for detecting fake and fraudulent faces where we jointly predict the semantic embeddings of the future face state. This enforces the model to learn social context of the perceived face. However, instead of directly replicating the process of humans, we leverage the success of multi-task learning to jointly predict if a face is real or fake, and predict the future face state. Even though the human brain fails to detect the cleverly tampered or synthetic faces generated by machines, we propose to utilise the notion of social context in the design of the proposed method. Our approach which is detailed in Sec. 3, though inspired by the human cognition process, enables us to design a deep machine learning network that far exceeds the performance of humans in discriminating real faces from fake and fraudulent ones.

We exploit Neural Memory Networks (NMNs) to facilitate the above two tasks by mapping long-term dependencies in the data domain. Current state-of-the-art NMNs such as the Neural Turing Machine (NTM) \cite{14} and Tree Memory Networks (TMNs) \cite{15} have demonstrated encouraging results when the models are required to maintain additional capacity to recover long-term dependencies. However, through proper comparative illustrations and evaluations we clearly demonstrate that the current state-of-the-arts do not optimally map these long-term dependencies due to architectural deficiencies and propose a
Proposed Hierarchical Attention Memory Network (HAMN) framework for tampered face detection and future face semantic embedding prediction: In the input module visual facial features are extracted through a pre-trained ResNet [16]. The extracted embeddings are rearranged in a sequence and passed through a bi-directional GRU [17] to map their relationships. Input attention is then employed to extract informative embeddings, deriving a vector to query the memory. The memory module outputs information regarding the authenticity of the face region and its future behaviour. In the learning process we utilise a discriminator network which either receives ground truth CNN semantic embeddings \( \Delta \) frames ahead, or synthesised semantic embeddings from the memory output. Using these inputs it learns to classify the embeddings are real or fake. The future semantic predictor tries to fool the discriminator in the process, and this adversarial loss is coupled with a fake face classification loss to form the complete objective of the HAMN.

novel Hierarchical Attention Memory Network (HAMN) to overcome these limitations.

In addition, as there doesn’t exist an optimal, off the shelf loss function for joint learning of the fake face detection and future semantic embedding anticipation tasks, we propose an adversarial learning platform to automatically learn a loss function that considers both tasks.

An overview of the proposed framework is presented in Fig. 1. Given an input image of a face, we first extract an embedding that captures facial semantics using a ResNet [16] model pre-trained on ImageNet [18]. This semantic embedding is used to query the memory. The memory output is used for two tasks: (i) for classification of the authenticity of the face; and (ii) the prediction of a future face semantic embedding. Facilitating the adversarial learning framework, the predicted semantic embeddings are passed through a discriminator. During offline training, the discriminator receives either a predicted or a ground truth future embedding, alongside the same input frame as the generator, and learns to classify real embeddings from fake ones. This training process requires pairs of face images and their representation at future time-steps. For testing, it can be applied to authenticate either images or videos. Furthermore, we would like to point out that the memory, \( M^t \), is composed of input frame embeddings that are in an arbitrary order (depending on the order of the input), not consecutive frames from a video. The contents of the memory itself changes with every time step, \( t \), and thus evolves through out the training process. Therefore, the applications of the proposed method are not limited to videos. Please refer to Sec. III-A for further detail regarding the motivation for using images instead of videos.

To the best of our knowledge, this is the first work to employ NMNs in the context of multimedia forensics as well as the first work attempting multi-task learning of NMNs through adversarial training. To this end we make the following contributions:

- We propose a method to detect tampered face images that is inspired by the human social cognition process, and use joint learning to predict the future face state, supporting the primary tampered media detection task.
- We propose a novel neural memory architecture for hierarchical semantic embedding propagation and prediction.
- We introduce an adversarial learning paradigm to train the memory architecture where the model learns a task specific loss for both fake face classification and future semantic generation tasks.
- We perform extensive evaluations on the proposed network using multiple public benchmarks where we outperform baseline methods by a substantial margin, especially for unseen tampering categories.
- We analyse what is being activated in the proposed hierarchical attention memory models during learning and interpret the learning process.

Considering the critical need for a fully automated system for the detection of fake, fraudulent, and tampered faces, we demonstrate the proposed system on its primary application of face tampering detection. However, its applications are not constrained to this domain and the framework is directly applicable to any machine learning setting which requires the hierarchical capture of long-term dependencies.

II. RELATED WORK

Related work within the scope of this paper can be categorized into human social perception and cognition (Sec. II-A), face manipulation techniques (Sec. II-B), face manipulation detection techniques (Sec. II-C), and memory architectures (Sec. II-D).

A. Human Social Perception and Cognition

Recent neuroscientific research [12], [13] has investigated human brain activities related to the uncanny valley [12] effect, where humans report an “eerie feeling” when observing faces which lack natural attributes like emotions. Their observations suggest a two-stage process in human face cognition. In the first phase, the visual input is matched to a face template which is denoted by strong activities in dedicated areas of the visual cortex (i.e. occipital face area [19] and fusiform face area [20]). In the second phase, they detected a subsequent activation as the observed face is evaluated for social perception. Most interestingly, they conclude that the observed face region is evaluated by inferring the state of the observed person’s mind [12]. Through visual cues such as facial expression and by utilising specific knowledge stored in observer’s memories, humans try to understand the observed person’s mental state (e.g. he or she is is angry). Subsequently, through a human’s natural social cognition processes, they build a theory about
that particular person and their emotions (e.g. why is he or she angry?). Most importantly these processes are applied even if the observed face is not familiar to the perceiver. In such cases, the inferences are indexed based on visual cues in their semantic memory for later use [13]. In addition, [13] illustrates that computer generated and tampered media generates the uncanny valley effect due to the lack of social presence and social attributes. In the proposed work we exploit this aspects of human social perception and cognition process. However, instead of directly replicating the two phase process of humans, we utilise the recent advances in machine learning to automatically detect fake and fraudulent faces, which are designed to fool the humans.

B. Face Manipulation Techniques

From one of the first attempts reported in Dale et al. [21], video face replacement has rapidly advanced. In [22] the authors investigated methods to preserve expressions while replacing faces; while [23] uses 3D face capture to match the mouth movements between the actors and the rubber.

One of the most notable methods among video face replacement techniques is DeepFake [2] which aims to replace the face in an input image with a target face. The core idea underpinning DeepFake is to train two auto-encoders, which share a common encoder, in parallel [1]. The auto-encoders seek to represent the input image with a reduced number of parameters and output an approximation of the original input.

At test time, by using the shared encoder and the decoder of $B$ one can generate a mapping from input $A$ to a synthesised image $B$. During encoding, the encoder is expected to segregate fundamental facial attributes from other information such as the background. Hence during decoding, the decoder should map these facts to the target face considering it’s contextual information.

The Face2Face [24] method transfers facial expressions from an input to a target frame. The technique first reconstructs and tracks the input source and target faces through a dense tracking system. Then using the expression translation mechanism proposed in [24] the authors translate the expressions of the source face to the target face.

In a different line of work the authors in [25] propose an image-to-image translation where they learned a mapping between the faces from computer graphics and real face images. Suwaganakorn et al. [26] propose a mapping between audio and lip movements and Averbuch-Elor et al. [27] propose a system to bring portraits to life.

Considering the recent success of GANs for synthesising realistic content, numerous such methods have been proposed to alter face images. For instance, in [28], [29] authors propose to alter face attributes including age, moustaches and smiling. A method to improve the image quality of the synthesised content is proposed in [30]. Most recently, Karras et al. [31] propose a GAN based face generator which learns to alter high level facial attributes such as pose and identity, as well as the stochastic variation of features such as freckles and hair in the synthesised content. In addition, the authors of [32] propose a dictionary learning based method to reduce the artifacts introduced during the fake face synthesis process. Specifically, they first train a linear dictionary to capture the patterns of real face images. Then they try to represent the fake images using the learned dictionary. The authors illustrate that this process significantly reduces the artifact patterns that the fake face generation techniques introduce.

These techniques have demonstrated their superior ability to generate realistic-looking faces. However, the authors in [1] showed that the synthesised faces from DeepFake lack the expressiveness of real examples. Furthermore, considering the results presented in [3], [24], we observe that the process introduces unnatural attributes to different face regions. Hence, the proposed future face semantic prediction based framework facilitates an ideal platform to learn a conditional distribution to segregate examples that contain those manipulations from those that do not.

C. Face Manipulation Detection Techniques

There exist several CNN based methods to detect image manipulations. For instance, in [7] the authors introduce a novel predictive kernel to detect pixels that deviate from the logical local structural relationships in the neighbourhood. The kernel first predicts the pixel value of the centre pixel in the given window and subtracts the actual pixel value to get the prediction error. These prediction errors are propagated to subsequent layers. Quan et al. [33] proposed the use of 3D convolutional kernels to learn discriminative features and separate natural images from computer generated graphics. In a different line of work Li et al. [34] proposed a method to detect fake video content by solely utilising eye blink patterns. They show that real video content contains a unique eye blinking pattern which fake content lacks. In [8] the authors investigate the efficiency of capsule networks for fake content detection. Furthermore, in [35] the authors propose to extract neuron activations from multiple levels of the CNN feature extractor and concatenate these features together. They illustrate that by extracting features from multiple levels of the CNN they can capture more subtle features.

However, as shown in [9], CNN based methods tend to overfit to a particular type of artefact, limiting the applicability of these methods for general biometric and security applications. Furthermore, reenactment techniques do not carry the artefacts of blinking patterns as they generally tamper with the nose and mouth regions. Hence, the general applicability of [34] is questionable. The authors in [36] exploit both spatial and temporal information through utilising a recurrent convolutional model.

A method that considers transferability in face forgery detection is presented in [9]. They consider an encoder-decoder framework and explicitly force the encoder to learn two separate encoding spaces for real and fake images. As a result, the total loss of their system is the combination of this classification loss and the reconstruction loss of the decoder.

In a different line of work researchers and have investigated the possibility of utilising inconsistency between the claimed camera parameters and the characteristics of perspective
distortion [37] and illumination conditions [38] to determine the integrity of input images. A multi-task learning paradigm is proposed in [39] where the authors combine the fake face classification task together with the segmentation of manipulated regions. The authors designed the network to share information between the tasks, however they did not observe a significant performance improvement from the joint learning of both tasks. In addition, a method for detection and a localisation of GAN based face manipulations is proposed in [40]. The authors are motivated by the fact that GAN based face manipulations introduce unnatural textures to the image when it upsamples the low-resolution images to high-resolution. The authors utilise a GAN to generate grayscale maps to denote the manipulated regions in the input image. However, the application of this method is limited to GAN based face manipulation techniques and cannot be applied to detect face identity or expression swaps.

Furthermore, we would like to acknowledge the work done by Qi et al. [41] where they illustrate that fake and fraudulent faces can be detected by observing the remote heart-rate representation. In particular, they show that normal heart rhythms are disrupted during the fake face generation process and to detect those miniscule periodic changes they propose generating a motion-magnified spatial-temporal representation. These features are passed through a temporal attention network which generates a real-fake classification. The main limitation of this work is that it requires videos for both training and testing of the model, which limits its applications for tasks such as fake and fraudulent face detection in face authentication where (in general) a single image is utilised as the input. In contrast, the proposed method utilises videos only for model training and the model can be tested with both images and videos.

D. Memory Architectures

With the tremendous success achieved by recurrent neural networks such as Long Short-term Memory (LSTM) [42] networks and Gated Recurrent Units (GRUs) [17], numerous works have employed what are termed “memory modules”. Memory modules are expected to store important facts and map long-term dependencies when questioned with a particular input. However, experiments conducted in [15], [43] demonstrated that the memory modules in LSTMs and GRUs only map relationships within an input sequence, and disregard dependencies between different inputs.

This motivates the need for external memory components. Fig. 2 depicts the basic structure of a memory module. First, the input, \( f_t \), is passed through an input controller to encode the input and generate a query, \( q_t \), to question the memory. The memory receives this query and using the facts stored in the current memory state, \( M^{t-1} \), it synthesises the memory output, \( z_t^' \), which is passed through an output controller to generate the memory output, \( r_t^' \). The process finishes by updating the memory and propagating it to the next time step using an update controller. The update controller uses the current timestep’s input \( f_t \) and the memory output \( z_t^' \) to generate the next memory state, \( M^t \).

![Fig. 2. Overview of an external memory which is composed of input, output and update controllers. The input controller determines what facts within the input are used to query the memory. The output controller determines what portion of the memory is passed as output. The update controller updates the memory and advances it to the next time step.](image)

The use of external memory in gaining traction in numerous domains, including language modelling [44], visual question answering [45], [46], trajectory prediction [15], [47] and reinforcement learning [48]. In the seminal work of [14], Graves et. al proposed a Neural Turing Machine (NTM) which uses this concept of an external memory to achieve superior results when recalling long-term dependencies and improved the generalisability of the learned representation, compared to conventional deep learned models like LSTMs.

However, even with the augmented capacity offered by a memory network, the flat memory structure of [14], [44]–[46], [49]–[51] hinders the ability of the network to determine and exploit the long-term dependencies stored in memory. This structure ignores important historical behaviour which occurs only over long time periods, and is regularly observed in applications such as aircraft and pedestrian trajectory modelling [15], [47].

Authors in [15], [47] demonstrated the value of a hierarchical attention structure over a flat structure when generating inferences. However, we observe multiple deficiencies in those architectures which restrict their utility for the face manipulation detection task. The hierarchical architectures of [15], [47] compress the information, but pay no attention to the current query when doing so. Hierarchical information compression is useful when extracting informative content from a myriad of facts stored in memory, however, we speculate that different facts are informative under different contexts; and the hierarchical operation can act as a bottleneck for information flow. Secondly, the tree structure mixes semantics from different input embeddings in the hierarchy. Hence, it is ineffective for faces as attributes from different faces should remain separate. In contrast, we propose a hierarchical attention memory architecture which preserves the identity of individual facts stored in the memory and effectively attends to those facts through a hierarchical attention mechanism to retrieve query specific information.

III. HIERARCHICAL ATTENTION MEMORY NETWORK MODULE

A. Motivation

As discussed earlier, prior works in neuroscience have shown that humans try to predict an observed person’s mental
state and when this anticipation process fails due to a lack of social presence and social attributes in the observed face, it can cause the uncanny valley effect [13]. Moreover, we observe that synthesised faces also lack the expressiveness of real examples and can have unnatural facial expressions [1]. These observations motivate the proposed approach and we propose a system that detects fake and fraudulent faces by discriminating between real and fake facial expressions and facial attributes. Specifically, we borrow inspiration from the human social perception and cognition processes when designing the auxiliary task of our multi-task learning framework. We propose to adopt NMNs such that, similar to humans, the proposed framework can use historical observations to assist the prediction task, rather than just employ the information from the current input. However, to directly replicate the human social perception and cognition processes we require annotated data regarding social attributes, emotions or theories about the perceived face, which are not available in publicly released fake face detection datasets. Therefore, we design our auxiliary task to predict the semantic features of the perceived face \( \Delta \) frames ahead. Hence, we do not directly mimic the human social perception and cognition processes, but rather we draw inspiration from these processes in the design of our auxiliary task.

Given an input image the proposed method anticipates how an abstract representation of the face would appear in \( \Delta \) time steps. Therefore, instead of predicting the observed person’s mental state, the proposed method attempts to predict future facial attributes corresponding to the change in the facial expression. Even though the proposed method is not directly mimicking this human cognition process, by anticipating future facial attributes the method learns to discriminate between how the facial attributes change in real and fake examples.

A naive way to design such a system would be to use the entire temporal structure of the given video from the modeled dataset and sequentially map the temporal evolution of the facial expressions/attributes. However, this limits applications of the system to videos only. In contrast, a system that operates at a frame-level is more flexible. Furthermore, we observe that the majority of existing state-of-the-art frameworks for recognising fake faces operate at the frame-level. Hence, to allow direct comparisons we design a framework that only utilises a single frame as the input to the system. It should be noted that the proposed system requires image pairs (i.e. current and future frames) only for the training process. At test-time in requires only a single frame. For more details, see supplementary materials.

The proposed Hierarchical Attention Memory Network (HAMN) follows a completely different approach to the current state-of-the-art face manipulation detection techniques such as [3], [9], [36], [38], [39]. We utilise NMNs to produce long-term dependency mappings regarding the basic attributes of the observed face. Furthermore, in contrast to [14], [15], [44]–[47] the proposed memory addressing mechanism preserves the identity of facts stored in the memory, and effectively attends to those facts through a hierarchical attention mechanism to retrieve query specific information. The overall architecture of the proposed HAMN consists of an input module, an input attention layer, and a memory module. These components are described in the following subsections.

### B. Input Module

Before illustrating the structure of the proposed model we would like to clarify that the notion of time refers to time with respect to the state of memory, rather than within the input video sequence. Hence time, \( t \), evolves throughout the training and testing process as opposed to being reset at the end of each input sequence. We extract features from input images using a pre-trained ResNet model trained on ImageNet. During pre-processing the input image is resized to \( 224 \times 224 \) and we extract features from the “Activation-85” layer which has an output dimensionality, \( d = 14 \times 14 \times 256 \). Hence, as shown in Fig. 3, the output has \( 14 \times 14 = 196 \) local patches, each containing 256 features. The feature extraction layer is evaluated experimentally (please see supplementary materials for details).

Formally, let the input face image \( f^t \) at time \( t \) contain \( k \in [1, K] \) patches. We summarise the information from neighbouring patches using a bidirectional GRU [46]. In the forward pass, \( f_k^t \), of the GRU it reads \( f^t \) from patches 1 to \( K \); and in the backward pass, \( f_k^t \), it reads from \( K \) to 1 using,

\[
\begin{align*}
     f_k^t &= GRU_{forward}(f_k^t; f_{k-1}^t), \\
     f_k^t &= GRU_{backward}(f_k^t; f_{k+1}^t),
\end{align*}
\]

and concatenates the forward and backward vectors to generate a summary of the patch \( k \),

\[
    f_k^t = [f_k^t; f_k^t].
\]
The generation of a sequential representation from images is common within the machine learning community. For instance, in [46] the authors utilise a similar GRU layer to summarise the input features from the input image. Furthermore, in [52] a similar unstacking process is utilised in order to represent relationships among image patches.

C. Input Attention

Crucially, not all patches contribute equally when representing the salient attributes of a face. Hence, we utilise an attention mechanism that learns to pay varying levels of attention to patches when aggregating local patches into a single query vector, $q_t$. Specifically, we pass the encoded patches of the input image through a single layer MLP [53] and obtain a representation, $v_t$, using,

$$v_t^i = \tanh(W_f f_t^i + b_f), \quad (3)$$

where $W_f$ and $b_f$ are the weight and bias of the MLP. Then we measure the importance of the current patch, $f_t^i$, using the similarity of $v_t^i$ with a patch level context vector, $v_f$. Following [54], the functionality of $v_f$ is implemented using a dense layer, which has the same dimension as $v_t^i$, and the weights of this layer are randomly initialised. The parameters of this layer are updated by jointly back propagating its errors with rest of the components in the proposed framework.

The similarity score values are then normalised using a softmax function,

$$\beta_t^i = \frac{\exp([v_t^i]^T v_f)}{\sum_k \exp([v_t^k]^T v_f)}, \quad (4)$$

generating a query vector to summarise the input image,

$$q_t = \sum_k \beta_t^i f_t^i. \quad (5)$$

D. Memory Module

The proposed hierarchical attention memory module is illustrated in Fig. 4. When generating an inference we encode the information using two levels of attention, namely, patch level and memory level. Let the memory module, $M^{t-1}$, at time $t-1$ hold $L$ face embeddings, $I_i$, where $i \in [1, L]$ and each image contains $K$ patches, where $p_{i,k}$ is the $k^{th}$ patch of image $i$.

1) Patch Level Attention: Following Sec. III-C we summarise each patch of an image embedding that resides in the memory using,

$$\overrightarrow{p_{i,k}} = GRU_{fwd}(p_{i,k}, \overrightarrow{p_{i,k}}),$$

$$\overleftarrow{p_{i,k}} = GRU_{bwd}(p_{i,k}, \overleftarrow{p_{i,k}}),$$

$$\overrightarrow{p_{i,k}} = \{\overrightarrow{p_{i,k}}, \overleftarrow{p_{i,k}}\}. \quad (6)$$

Motivated by [46], in order to measure the similarity between $p_{i,k}$ that resides in memory and the input image...
1: for each $i \in L$ do 
2: for each $k \in K$ do 
3: $x_{i,k} = \frac{[\hat{p}_{i,k} \cdot f'_k; |\hat{p}_{i,k} - f'_k|]}{\exp((u_{i,k})^\top u_x)}$, 
4: end for 
5: end for 

patch $f'_k$, we first generate an augmented vector by multiplying the patches of $I_t$ with the equivalent patch in $f'$, and concatenating this with the absolute difference between the patches.

Then we employ attention to determine the informative patches,

$$
\begin{align*}
  u_{i,k} &= \tanh(W_x x_{i,k} + b_x), \\
  \alpha_{i,k} &= \frac{\exp((u_{i,k})^\top u_x)}{\sum_k \exp((u_{i,k})^\top u_x)}, \\
  \rho_i &= \sum_k \alpha_{i,k} x_{i,k},
\end{align*}
$$

(7)

where $W_x$ and $b_x$ are the weights and bias of a separate single layer MLP [55]. $u_x$ is a context vector which is randomly initialised and jointly learned during training, and $\alpha_{i,k}$ are the normalised score values quantifying the similarity between patch $f'_k$ of the current input image and patch $p_{i,k}$ of the image, $I_t$, in memory. Drawing similarities to the way the human brain operates, this process measures how the attributes are similar with respect to past experiences.

2) Memory Level Attention and Memory Output: We apply another level of encoding with attention to summarise the similarity at the image level. Specifically,

$$
\begin{align*}
  \overrightarrow{\rho}_t &= GRU_{fwd}(\rho_t, \overrightarrow{\rho}_{t-1}), \\
  \overleftarrow{\rho}_t &= GRU_{bwd}(\rho_t, \overleftarrow{\rho}_{t+1}), \\
  \overrightarrow{\rho}'_t &= [\overrightarrow{\rho}_t; \overleftarrow{\rho}_t].
\end{align*}
$$

(8)

Then we generate an augmented context vector using multiple interactions between the memory content and the query vector, $q'$, such that,

$$
\begin{align*}
  z_i &= [\overrightarrow{\rho}'_t \cdot q'; \overleftarrow{\rho}'_t \cdot r'^{-1}; \overrightarrow{\rho}'_t - q' ]; \forall i \in L,
\end{align*}
$$

(9)

where $r'^{-1}$ is the memory output at time step $t - 1$. Now the output of the memory read operation $r'$ at time step $t$ can be generated by,

$$
\begin{align*}
  o_i &= \tanh(W_z z_i + b_z), \\
  \gamma_i &= \frac{\exp(o_i^\top o_z)}{\sum_k \exp(o_k^\top o_z)}, \\
  r' &= \sum_k \gamma_i z_i,
\end{align*}
$$

(10)

3) Memory Update: In the update procedure we directly append the current input, $f'$, to the previous memory state, $M^{t-1}$, and propagate the memory to the next time step using,

$$
M^t = [M^{t-1}_2; f'].
$$

(11)

We remove the oldest entry when appending a new embedding, maintaining a constant size.

We emphasise the fact that this update operation is quite distinct to the standard memory update operations used by [14], [46], [47]. We specifically utilise this operation to preserve the integrity of the stored embedding. If we use the operations of [14], [46], [47] this would partially update the memory content, resulting in face embeddings that are a combination of two or more faces. Hence, it would make even a real face appear fraudulent. Therefore, we simply append the new observation to the memory stack as the update procedure.

To extract relevant information from this ‘raw’ face embedding hierarchical attention is employed, which searches through the individual memory slots as well as through the patches of the content within those individual slots. This allows us to locate relevant content from the memory stack to answer the current query, without altering individual face embeddings.

4) Fake Face Classification and Future Face Semantic Embedding Regression: When performing the input face classification, $\hat{y}'$, (i.e real or manipulated face) we directly pass the memory output, $r'$, through a single layer MLP,

$$
\hat{y}' = \text{softmax}(W_r r' + b_y).
$$

(12)

For the generation of future face patches we need to synthesise a sequence of $K$ patches. When decoding this information from the memory output, $r'$, we apply the same strategy that we applied when encoding the patch information into a vector representation. Specifically,

$$
\begin{align*}
  \overrightarrow{h}'_k &= GRU_{fwd}(h'_k, \overrightarrow{h}'_{k-1}), \\
  \overleftarrow{h}'_k &= GRU_{bwd}(h'_k, \overleftarrow{h}'_{k+1}), \\
  \overrightarrow{h}'_k &= [\overrightarrow{h}_k; \overleftarrow{h}_k],
\end{align*}
$$

(13)

where $h'_0 = r'$. Then the relevant future face embeddings are predicted by,

$$
\overrightarrow{h}'_k = \text{Relu}(W_h \overrightarrow{h}'_k + b_h).
$$

(14)

IV. MODEL LEARNING

It is tedious to hand engineer a loss for future face embedding prediction. Hence, motivated by the ability of GANs to seamlessly learn a task-specific loss in numerous domains [30], [56]–[60], we employ a generative adversarial framework.

GANs are comprised of two components, a Generator, $G$, and a Discriminator, $D$, where they partake in a two player game. Using a random noise vector, $z'$; $G$ tries to synthesises realistic future face embeddings, $\overrightarrow{h}'$, and tries to fool $D$. $D$ tries to separate the synthesised embeddings, $\overrightarrow{h}'$, from real examples, $\eta'$. Hence the loss of this process is not hand engineered and the framework learns a custom loss for the task at hand. This objective can be written as,

$$
V = \min_G \max_D \mathbb{E}[\log D(\eta') + \mathbb{E}[\log(1 - D(G(z')))].
$$

(15)

However, in Eq. 15 $G$ synthesises embeddings without considering the current input. Hence, we draw inspiration from the conditional GAN [61], where $G$ learns a conditional mapping from a random noise vector, $z'$, and the current
memory output, \( r' \), to an output, \( \hat{y}' : G(z', r') \rightarrow \hat{y}' \). This augmented objective, \( \hat{V} \), can be written as,

\[
\hat{V} = \min_{G} \max_{D} E(\log D(r', \hat{y}')) + E(\log(1 - D(r', G(z', r'))) \tag{16}
\]

We couple the objective in Eq. 16 with the fake face classification loss, allowing the model to jointly learn an objective that considers both tasks. In addition and following GAN literature \([56, 57, 61]\), we add \( L_2 \) regularisation of the synthesised embeddings to encourage \( G \) to generate realistic embeddings. Our final objective, \( V^* \), can be defined as,

\[
V^* = \hat{V} + \log(\hat{y}') + \sum_{k} ||\hat{\eta}^k_t - \eta^k_t||^2. \tag{17}
\]

Therefore, throughout the learning process the memory stack, \( L \), evolves as we stack the incoming frame embeddings. At the test-time we initialise \( L \) to the state it was at the end of the learning process and it evolves through the testing process as well. However, it should be noted that the test videos were completely separate to those used during training.

V. EXPERIMENTS

In this section we provide the implementation details of the proposed architecture and apply it to detect digital forgeries in three widely used public benchmarks.

A. Datasets

1) FaceForensics Dataset: The FaceForensics dataset \([3]\) was collected from YouTube. Videos are at least 480p in resolution, and are tagged with “face”, “newscaster” or “newsprogram”. To generate tampered faces, the authors use the Face2Face method \([24]\) between two random videos. The dataset contains 704 training videos (364,256 images), 150 validation videos (76,309 images), and 150 test videos (78,562 images).

2) FaceForensics++ Dataset: This dataset \([62]\) is an extended version of FaceForensics and contains face manipulations from FaceSwap \([63]\) and DeepFakes \([2]\). FaceSwap \([63]\) is a light weight editing tool which copies the face region from one image to another using face maker positions. The original videos are taken from youtube and manually screened to remove occlusions. The dataset consists of 1,000 original videos and 3,000 manipulated videos (1,000 for each category) from Face2Face, FaceSwap and DeepFake methods. Similar to \([62]\) we select 720 videos for training, 140 for validation and 140 for testing.

3) FakeFace in the Wild (FFW) Dataset: The FFW dataset \([64]\) is constructed using a set of public videos from YouTube, and contains a wide array of fake content generated through computer graphics, GANs, manual and automatic tampering techniques, and their combinations. Therefore, it provides an ideal setting to evaluate the generalisability of the proposed methodology under a diverse set of manipulations. Videos have a variable duration from 2-74 seconds and have at least 480p resolution. In addition to these 150 forged videos, the FFW dataset contains 150 real face videos from FaceForensics \([3]\).

B. Implementation Details

Following \([1]\), we applied the Viola-Jones face detector \([65]\) to detect faces. We extract every 20\(^{th}\) frame and the respective frame 15 frames ahead as the input-output pair for the future frame prediction task. To balance the dataset, frames are selected for extraction such that an equal number of samples are extracted from each video. For all the GRUs we set the hidden state dimension to 300. Values for hyper-parameters memory length, \( L = 200 \); and the number of patches, \( K = 196 \); are evaluated experimentally (please see supplementary materials for details).

For comparisons we utilise three baseline memory modules in our evaluations, Neural Turing Machine (NTM) \([14]\), the Dynamic Memory Network (DMN) \([46]\) and Tree Memory Network (TMN) \([15]\). For fair comparison we train these methods using the same ResNet features utilised by the proposed method and we set the LSTM hidden state dimension of the NTM, DMN and TMN modules to 300 and memory lengths to 200. The extraction depth of the TMN is evaluated experimentally and is set to 3. We train these memories to directly classify the input image using supervised learning and binary cross entropy loss.

C. Face Reenactment Detection Using FaceForensics Dataset

The ability of the proposed method to detect facial reenactments is measured using the FaceForensics dataset \([3]\). We strictly adhere to the author’s guidelines when pre-processing the data and used the same training, testing and validation splits as \([3]\). Following \([8]\), we report the classification results in terms of video and frame level accuracy. Video level classifications are obtained by aggregating frame level predictions over the entire video and obtaining the most voted class label.

Facial reenactment detection evaluations on FaceForensics \([3]\) at the video and frame level are presented in Tab. I.
When analysing the results it is clear that baseline system performance degrades significantly when the video compression level increases. This ratifies the observations presented in [9], [64] where the authors speculate that when the compression level increases the specific artefact that the system is focusing on degrades in clarity and significance. Hence, when the visual clarity degrades the CNN based tampered face detection systems such as Meso-4 [1], CapsuleForensics [8], Rössler et al. [3], and Nguyen et al. [66] fail. For instance, in Tab. I, for Rössler et al. [3] we observe a 12.12% degradation of performance between no-compression and strong-compression classes. In contrast, the performance degradation between no-compression and strong-compression classes for the memory based models is not that significant. Even though the baseline DMN and TMN architectures fail to attain satisfactory performance due to there inherent architectural deficiencies in memory structure, the performance difference between the compression classes is approximately 5%. We speculate from these results that our memory based models, which are comparing and contrasting the similarities between the observed facts stored in the memory, are not merely focusing on a specific artefact in the observed images (e.g. compression artefacts) but mapping the overall structure of the face and the long-term dependencies between examples.

Comparing the NTM, DMN and TMN memory architectures with the proposed memory architecture, we observe that they fail to achieve satisfactory performance due to inherent deficiencies. We speculate that the flat memory structure of NTM and DMN fails to propagate useful information to the output, while the TMN mixes the embeddings from different historical observations, making it difficult to discriminate the faces. In the proposed HAMN model we rectify these deficiencies and further augment the performance through the joint learning of the future embeddings. Through evaluations in Sec. VI-A.2 and face embedding prediction in supplementary materials we demonstrate that the two tasks complement each other.

D. Evaluations Against Different Manipulation Types Using FaceForensics++

We evaluate the robustness of the proposed method to different state-of-the-art face manipulation techniques using FaceForensics++. In Tab. II we compare the accuracies of the systems when trained on all manipulation types together.

Similar to the evaluations on the FaceForensics dataset (Sec. V-C) we observe that the performance of the baselines degrades rapidly as compression increases; while the proposed method achieves consistent accuracy across varied compression levels, regardless of the additional DeepFake and FaceSwap manipulation categories.

To further illustrate this we manually added Gaussian blur and noise to the test set of the FaceForensics++ dataset. We evaluate the model performance at different blur and noise levels by controlling the standard deviation parameter of the Gaussian kernel. For comparison, we evaluate the performance of the Meso-4 and MesoInception-4 models using the implementations provided by the authors. Furthermore, we evaluate the baseline memory model, Tree Memory.

Tab. III illustrates these results. Similar to the results presented in Tab. II we observe that the performance of the CNN based baseline models, Meso-4 and MesoInception-4, rapidly degrades when the quality of the input degrades. However, both memory based approaches have comparatively less sensitivity to the input image quality.

Tab. IV reports the accuracies for models trained on FaceForensics from Sec. V-C (i.e trained using only Face2Face manipulations), tested with the unseen DeepFake and FaceSwap manipulation types. Following [39] we use light compression videos (quantisation = 23) for this experiment and reported accuracies are at the image level.

From the results in Tab. IV, it is clear all baseline methods struggle to detect the unseen attacks. The Xception [69] model, which achieves commendable accuracy with the known attacks in Tab. II, struggles to generalise to unknown attacks as the method is focusing on specific artefacts in the training data left by the process that creates the fake face, rather than learning a generalisable representation to segregate the two classes. In contrast, the performance gap for the memory based systems for seen and unseen attack types is considerably lower, demonstrating the importance of long-term dependency modelling using memory architectures to discriminate between fake and real faces.

E. Detecting Unknown Attacks Using FakeFace in the Wild (FFW) Dataset

Following [64], we measure the classification accuracy in terms of Equal Error Rates (EER), Attack Presentation Classification Error Rate (APCER) [1] and Bonafide Presentation Classification Error Rate (BPCER) [1] under three evaluation settings: (i) TestSet-I where there are 1,500 real face and 1,500 fake face images tampered with known attacks; (ii) TestSet-II with 1,500 real and 1,500 fake face samples with unknown attacks; and (iii) TestSet-III which is comprised of 1,776 real and 1,576 fake faces generated using the FaceSwap and SwapMe applications proposed by [67].

1 available at https://github.com/DariusAf/MesoNet

| Method            | Accuracy for different compression levels |
|-------------------|-----------------------------------------|
|                  | 0 (None) | 23 (Light) | 40 (Strong) |
| MesoInception-4   | 96.51    | 85.51      | 75.65       |
| ResNet [16]       | 88.24    | 81.10      | 62.15       |
| Cozzolino et al.  | 98.56    | 79.56      | 56.38       |
| Xception [69]     | 99.41    | 97.53      | 85.49       |
| NTM [14]          | 78.32    | 74.24      | 72.40       |
| DMN [46]          | 80.25    | 75.23      | 73.04       |
| Tree Memory [15]  | 82.13    | 78.33      | 73.14       |
| HAMN              | 99.43    | 96.65      | 97.02       |
TABLE III
EVALUATION OF MODEL PERFORMANCE UNDER DIFFERENT BLUR AND NOISE LEVELS AT THE FRAME LEVEL ON THE FACE FORENSICS++ DATASET [62] (HIGHER IS BETTER). CURRENT STATE-OF-THE-ART METHODS FOR FAKE FACE DETECTION ARE SHOWN WITH A PINK BACKGROUND, BASELINE MEMORY MODELS ARE SHOWN WITH A BLUE BACKGROUND AND THE PROPOSED HAMN METHOD IS SHOWN WITH A WHITE BACKGROUND

| Method          | Gaussian Blur       | Gaussian Noise      |
|-----------------|----------------------|---------------------|
|                 | sigma = 2            | sigma = 4           | sigma = 6          | sigma = 0.05 | sigma = 0.1 | sigma = 0.25 |
| Meso-4 [1]      | 56.74                | 45.98               | 40.65              | 56.12        | 50.68       | 43.95        |
| Mesolception-4 [1] | 72.58              | 61.05               | 50.54              | 70.45        | 64.57       | 58.70        |
| Tree Memory [5] | 68.76                | 62.53               | 53.47              | 67.59        | 61.11       | 58.56        |
| HAMN            | 78.98                | 72.68               | 68.50              | 77.68        | 74.35       | 71.09        |

TABLE IV
EVALUATION AGAINST UNSEEN MANIPULATION TYPES AT THE FRAME LEVEL ON THE FACE FORENSICS++ DATASET [62] USING THE MODEL TRAINED IN SEC. V-C (HIGHER IS BETTER).
CURRENT STATE-OF-THE-ART METHODS FOR FAKE FACE DETECTION ARE SHOWN WITH A PINK BACKGROUND, BASELINE MEMORY MODELS ARE SHOWN WITH A BLUE BACKGROUND AND THE PROPOSED HAMN METHOD IS SHOWN WITH A WHITE BACKGROUND

| Method     | Accuracy | DeepFake | FaceSwap |
|------------|----------|----------|----------|
| ResNet16 [16] | 43.10    | 38.19    |          |
| Xception [69] | 55.12    | 50.38    |          |
| Cozzolino et al. [68] | 62.61    | 52.29    |          |
| Nguyen et al. [39] | 52.32    | 54.07    |          |
| NTM [14]   | 50.45    | 46.35    |          |
| DMN [46]   | 50.92    | 47.13    |          |
| TMN [15]   | 51.34    | 48.19    |          |
| HAMN       | 84.12    | 86.53    |          |

TABLE V
PERFORMANCE ON KNOWN FAKE FACES FROM TEST SET-I OF FFW [64]. WE REPORT APCER, BPCER AND EER [1] AS PERFORMANCE METRICS, (LOWER VALUES ARE BETTER).
CURRENT STATE-OF-THE-ART METHODS FOR FAKE FACE DETECTION ARE SHOWN WITH A PINK BACKGROUND, BASELINE MEMORY MODELS ARE SHOWN WITH A BLUE BACKGROUND AND THE PROPOSED HAMN METHOD IS SHOWN WITH A WHITE BACKGROUND

| Method        | APCER | BPCER | EER |
|---------------|-------|-------|-----|
| LBP [70]      | 3.80  | 2.87  | 3.33|
| AlexNet [71]  | 7.80  | 1.73  | 3.73|
| VGG19 [72]    | 2.47  | 0.47  | 1.40|
| ResNet50 [16] | 2.27  | 0.47  | 1.40|
| Xception [69] | 2.47  | 0.13  | 1.07|
| Inception [73] | 0.67  | 0.47  | 0.53|
| Meso-4 [1]    | 0.61  | 0.59  | 0.56|
| Mesolception-4 [1] | 0.55  | 0.56  | 0.53|
| NTM [14]      | 1.55  | 0.48  | 1.98|
| DMN [46]      | 1.44  | 0.42  | 1.98|
| Tree Memory [15] | 1.53  | 0.23  | 1.51|
| HAMN          | 0.12  | 0.09  | 0.10|

TABLE VI
PERFORMANCE ON UNKNOWN FAKE FACES FROM TEST SET-II OF FFW [64]. WE REPORT APCER, BPCER AND EER [1] AS PERFORMANCE METRICS, (LOWER VALUES ARE BETTER).
CURRENT STATE-OF-THE-ART METHODS FOR FAKE FACE DETECTION ARE SHOWN WITH A PINK BACKGROUND, BASELINE MEMORY MODELS ARE SHOWN WITH A BLUE BACKGROUND AND THE PROPOSED HAMN METHOD IS SHOWN WITH A WHITE BACKGROUND

| Method        | APCER | BPCER | EER |
|---------------|-------|-------|-----|
| LBP [70]      | 89.00 | 2.87  | 48.73|
| AlexNet [71]  | 91.47 | 1.73  | 32.13|
| VGG19 [72]    | 90.73 | 0.47  | 29.40|
| ResNet50 [16] | 89.53 | 0.47  | 30.33|
| Xception [69] | 93.20 | 0.13  | 25.87|
| Inception [73] | 91.93 | 0.47  | 27.47|
| Meso-4 [1]    | 93.90 | 1.05  | 31.13|
| Mesolception-4 [1] | 93.71 | 0.89  | 29.10|
| NTM [14]      | 93.39 | 2.94  | 43.53|
| DMN [46]      | 93.12 | 2.92  | 42.1 |
| Tree Memory [15] | 88.62 | 1.31  | 34.10|
| HAMN          | 49.95 | 0.12  | 12.51|

TABLE VII
PERFORMANCE ON UNKNOWN FAKE FACES FROM TEST SET-III OF FFW [64]. WE REPORT APCER, BPCER AND EER [1] AS PERFORMANCE METRICS, (LOWER VALUES ARE BETTER).
CURRENT STATE-OF-THE-ART METHODS FOR FAKE FACE DETECTION ARE SHOWN WITH A PINK BACKGROUND, BASELINE MEMORY MODELS ARE SHOWN WITH A BLUE BACKGROUND AND THE PROPOSED HAMN METHOD IS SHOWN WITH A WHITE BACKGROUND

| Method        | APCER | BPCER | EER |
|---------------|-------|-------|-----|
| LBP [70]      | 90.16 | 3.43  | 46.05|
| AlexNet [71]  | 94.04 | 5.01  | 43.03|
| VGG19 [72]    | 97.27 | 2.31  | 44.93|
| ResNet50 [16] | 89.40 | 8.22  | 43.79|
| Xception [69] | 93.15 | 3.43  | 40.99|
| Inception [73] | 71.64 | 22.58 | 46.39|
| Meso-4 [1]    | 75.34 | 12.13 | 45.12|
| Mesolception-4 [1] | 73.31 | 10.12 | 43.10|
| NTM [14]      | 83.45 | 14.93 | 48.80|
| DMN [46]      | 83.42 | 14.95 | 46.12|
| Tree Memory [15] | 79.09 | 10.15 | 43.72|
| HAMN          | 44.89 | 1.51  | 14.12|

As per [64], we perform comparative evaluations with the texture based LBP [70] method, and state-of-the-art CNN architectures, AlexNet [71], VGG19 [72], ResNet50 [16], Xception [69], and GoogleLetNet/InceptionV3 [73]; and the popular MesoNet-4 and MesoInception-4 [1] due to the public availability of their implementations.

Comparing Tab. VI and VII with Tab. V, the performance of all the baseline systems in terms of APCER and EER drops significantly showing the baseline system’s lack of transferability. For instance, we observe lower APCER for Meso-4 [1] and MesoInception-4 [1] architectures, compared to Inception [73] in Tab. V, but comparatively higher APCER in Tab. VI and VII. This clearly demonstrate that the state-of-the-art methods for fake face classification are focusing on
specific artefacts in the training data left by the fake face creation process, limiting their applicability.

Considering the performance of the memory architectures, the worst performance observed is for the NTM model highlighting the inadequacy of a single level attention in the memory retrieval mechanism. A slight performance boost is observed through the introduction of a tree-structured memory, however, due to the mixing of semantics between different embeddings that reside in the memory, this method fails to effectively learn to discriminate fake images from real. However, the performance degradation between seen and unseen attacks (in terms of APCER) is comparatively low compared to CNN based systems such as GoogleLetNet/InceptionV3 [73], and the popular MesoNet-4 and MesoInception-4 [1], supporting the notion that the long term dependency modelling of memory networks improves performance. Focusing on these properties, the proposed HAMN method, by building theories about the observed face in terms of their appearance and their anticipated behaviour, has successfully gained an insight into the fundamentals of real human faces, allowing it to successfully discriminate real examples from fake.

In order to better clarify this point we extract memory outputs $r^t$ from Eq. 10 of the proposed HAMN model for 300 randomly selected examples from the TestSet-III of FFW [64]. We then applied PCA [74] to plot the embeddings in 2D. For the same examples we extract embeddings from the flatten layer of MesoInception-4 which uses CNNs to encode its inputs. We note that the models haven’t seen the manipulation types in TestSet-III during model training. The resultant plots are shown in Fig. 5. Confirming the aforementioned observations, we see a better segregation between real and fake examples with the memory based features compared to the CNN based features of MesoInception-4. We believe this is because the memory network allows the proposed framework to not only utilise features from the input face in the learning process, but it also allows those features to be compared and contrasted with the features of the faces that the model has seen previously. This could be viewed as augmenting the ‘raw’ CNN feature input with historically observed features such that the real/fake classifier receives better features for the task at hand. Therefore, even if the model hasn’t seen the manipulation types within TestSet-III of FFW, it has the capacity to understand the general representation of a real and fake face

(i.e. by querying across hundreds of semantic representations stored in its memory bank). Hence, the retrieved feature vector is much richer (and discriminative) compared to using the raw CNN features (as in MesoInception-4). This allows the proposed model to learn a better feature representation, and a better segregation between the real and fake classes. Hence we observe the value of further processing the input CNN derived feature inputs through the proposed memory network. In contrast, the raw CNN features produced by MesoInception-4 are less discriminate and we believe this is largely because they tend to focus on fragile features such as artefacts left by the fake face generator [9]. Hence the characteristics that the CNN model is looking for are specific to particular manipulation methods, and do not generalise to different methods. This is clearly evident in Fig. 5 (b) where we see the overlaps between real and fake classes.

VI. DISCUSSION

A. Ablation Study

1) Importance of Hierarchical Attention: In order to evaluate the contribution of hierarchical attention within the proposed framework we conducted a series of ablation experiments where we remove attention at the input ($\beta$) (see Sec. III-C), patch ($\alpha$) (see Sec. III-D.1) and memory output levels ($\gamma$) (see Sec III-D.2). In these variants we directly aggregate the vectors into a single output vector. Evaluation results on unseen using FaceForensics++ dataset are presented in Tab. VIII.

Firstly, the most significant contribution from attention is observed when encoding the memory output, as denoted by the highest degradation in performance. Secondly, we observe a significant impact when generating the image representation using patch level attention. This verifies our hypothesis that hierarchical attention is important when acquiring knowledge from stored memories. This is why even the models with individual attention levels, HAMN / $\alpha$ and HAMN / $\gamma$, haven’t been able to obtain good performance. Thirdly, we observe a substantial contribution from $\beta$, which helps generate a query effectively from the current input.

2) Importance of the GAN Learning Objective: We evaluate the contribution of the GAN learning framework and

| Ablation Model | Acc. Real | Acc. Fake |
|---------------|----------|-----------|
| HAMN / $\beta$ | 80.15 | 82.68 |
| HAMN / $\alpha$ | 77.19 | 79.41 |
| HAMN / $\gamma$ | 75.33 | 76.10 |
| HAMN / ($\beta + \alpha$) | 75.98 | 77.15 |
| HAMN / ($\beta + \gamma$) | 72.15 | 74.63 |
| HAMN / ($\alpha + \gamma$) | 74.01 | 75.69 |
| HAMN / ($\alpha + \beta + \gamma$) | 70.25 | 72.78 |
| HAMN | 84.12 | 86.53 |

Fig. 5. 2D visualisation of the embedding space for the proposed HAMN (a), and MesoInception-4 [1] (b) for 300 randomly selected images from the real and fake examples from the TestSet-III of FFW [64]. In the visualisations blue stars denote the fake examples while green circles denote the real examples.
the contribution from the future face prediction task when detecting fake faces. In the ablation model, HAMN/GAN, we removed the GAN learning objective and trained it using supervised learning with a combination of classification loss and Mean Squared Error (MSE) between the predicted and ground truth future face semantics. The model, HAMN/(GAN + η), removes the future face prediction task and directly optimises the fake face classification objective. Neither of these methods uses GAN learning.

From Tab. IX, we speculate that the future face classification task and GAN learning objective are equally important. Although we expect better performance for HAMN/(GAN + η) without the future face embedding prediction overhead, this model suffers the highest performance drop, suggesting that understanding and anticipating future emotions and interactions regarding the perceived face helps detect unrealistic faces.

3) Importance of Multi-Task Learning: In this section we utilise the FaceForensics++ dataset [62] and evaluate the contribution from the auxiliary task. Using the model trained on FaceForensics dataset (Sec. V-C) we test the model using the test set of FaceForensics++ dataset’s three fake face classes, Face2Face which the model has already seen in the training data, and the unseen DeepFake and FaceSwap classes. The results are presented in Tabs. X and XI, respectively. For comparisons we report the results observed by Nguyen et al. in [39] as well as different variants of the state-of-the-art NTM framework [14].

We observe that our observations contradict the results of [39], which found no significant contribution from the auxiliary task for unseen attack detection and a reduction in performance from the auxiliary task for the detection of seen attacks. In contrast we observe a substantial contribution from the proposed multi-task learning paradigm for both seen and unseen attacks. In addition, the comparisons with the state-of-the-art NTM framework [14] reveal that even when the method is augmented with the future semantic embedding anticipation auxiliary task and the GAN learning of the memory embeddings, the NTM fails to outperform the proposed HAMN model due to the deficiencies of the memory structure. However, we observe an increase in the fake face detection accuracies in the NTM + η and NTM + (GAN + η) ablation models when the baseline NTM is augmented, highlighting the contributions of those individual components to our primary task.

4) Discussion: Through the results presented in Tabs. VIII - XI we clearly illustrate that the performance gain of our approach is due to our three novel contributions: the introduction of a joint learning framework which is inspired by recent neuroscience findings and predicts future face embeddings as an auxiliary task; the novel memory structure which utilises hierarchical attention for memory output generation and preserves the integrity of the stored embeddings; and automatic learning of a loss function for the multiple tasks at hand through the GAN learning framework.

To further demonstrate these merits, we visualise the embedding spaces for the HAMN and HAMN/η ablation model presented in Tab. XI. We randomly selected 300 images from the real and DeepFake classes and extracted memory outputs rτ from Eq. 10 for these inputs. We applied PCA [74] to plot the embeddings in 2D, which are shown in Fig. 6. Analysing these plots it is clear that through the joint learning of both tasks the proposed HAMN model has been able better separate the real and fake faces compared to solely learning the single fake face classification task.

Despite these encouraging results, it is still not completely clear why the future semantic embedding prediction task aids the fake face detection task. To explore this, we investigate whether there is a clear distinction between the way that a real face evolves over time and the way that a fake face evolves.
behaves (i.e. any unnatural attributes). Furthermore, the evaluations should illustrate how the identity factor (uniqueness of the perceived face) impacts the future semantic embedding prediction (and the segregation of the real and fake classes).

To address these concerns in the following evaluation we generate a 2D visualisation of the predicted and ground truth embedding from 20 real (indicated as circles) and 20 fake (indicated as stars) randomly selected images from the FaceForensics++ dataset [62]. Each colour in the plot represents a particular example. We observe in the illustration provided in Fig. 7 that the predicted embeddings for individual identities are segregated and most importantly the real and fake examples have been clearly separated in the 2D space. This validates our hypothesis that there exists a significant distinction between the ways that real and synthesised faces behave. We believe the lack of expressiveness and emotions have introduced unnatural behavioural attributes to the synthesised faces, which the proposed method has leveraged to attain its superior results. We would like to re-iterate the notion of the uncanny valley effect, which also shows that synthesised content has different characteristics compared to real content. The proposed future semantic embedding prediction task has effectively leveraged this feature of the synthesised faces for the fake face detection task. Furthermore, factors related to identity haven’t impacted this segregation, demonstrating that the proposed method learns a representation that is unique to a particular face (i.e. depends on the input face representation), which can also distinguish between the real and fake classes.

B. What is Actually Being Activated?

In Fig. 8 we visualise the $\gamma$ activation values, resulting from Eq. 10, for the content of the memory for a sample input. We have overlayed the input attention $\beta$, in yellow, on input (a brighter intensity corresponds to higher activations). As $L = 200$ there exist 200 memory slots which we denote $l_1$ to $l_{200}$. For different peaks and valleys in the memory activation, we also show what input image embeddings are stored at that particular memory index.

We observe that the HAMN provides higher responses for similar face attribute patterns that the model has seen in the long-term history. It should be noted that even though the model hasn’t seen the input image before, it is compares and contrast between faces with similar facial attributes by capturing the underlying semantics in regions such as the eye brows, nose and head pose (see peaks between $l_{50}$ to $l_{150}$). The HAMN model is measuring how much the current input is similar to the individual observations stored in memory by comparing and contrasting the fundamental facial attributes, allowing it to detect fake faces. This also highlights the importance of preserving the attributes of individual images separately in the proposed patch encoding mechanism. If this is not done (i.e. such as in the TMN) performance drops as the attributes of different faces are mixed.

In Fig. 9 we analyse the patch level attention $\alpha$ (i.e Eq. 6) given to individual face regions of the face for sample images. We populate a 2D heat map using the activations. As $K = 14 \times 14$, we upscale the heat map to fit the original image dimensions. As a result, there is only a rough correspondence. The activation values are shown in yellow and brighter intensity values correspond to higher activations.

In Fig. 9 we observe that eye, lip and cheek regions have a high level of attention to measure the image authenticity. These activation plots verify the importance of capturing the facial attributes hierarchically, while preserving their identity and mapping long-term dependencies for accurate detection of fraudulent and tampered faces; and enabling transferability among different types of attacks. We would like to refer the reader to supplementary materials for additional discussion.
on hyper-parameter evaluations, the quality of the predicted future face embeddings and theoretical comparisons between the proposed memory architecture and the current state-of-the-art methods.

VII. CONCLUSION

Motivated by the social perception and social cognition processes of the human brain, we have presented in this paper a Hierarchical Attention Memory Network (HAMN) architecture for the detection of fake and fraudulent faces. The main advantage of this method is the transferability of its learned representation across different, and most importantly, unseen face manipulation methods. By capturing both patch level and image level semantics, and effectively propagating the learned knowledge hierarchically through a memory architecture, the proposed method attains its ability to accurately anticipate the temporal evolution of the observed face, allowing it to discriminate fake faces from real ones. Through extensive evaluations, we have demonstrated the utility of the hierarchical modelling of the stored knowledge while preserving the identity of those facts, and provide visual evidence of how the memory retrieves stored facts while considering how they relate to the current input. This provides solid evidence about the underlying ability of the HAMN model to synthesise theories about faces and understand their temporal evolution; abilities which are fundamental across a number of tasks.

REFERENCES

[1] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, “MesoNet: A compact facial video forgery detection network,” in Proc. IEEE Int. Workshop Inf. Forensics Secur. (WIFS), Dec. 2018, pp. 1–7.

[2] Deepfake. Accessed: Feb. 20, 2019. [Online]. Available: https://www.deepfakes.club/openfaceswap-deepfakes-software/

[3] 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,” 2018, arXiv:1803.09179. [Online]. Available: http://arxiv.org/abs/1803.09179

[4] R. Chesney and D. K. Citron, “Deep fakes: A looming challenge for privacy, democracy, and national security,” California Law Rev., vol. 107, p. 1753, Jul. 2018.

[5] A. K. Jain, A. Ross, and S. Prabhakar, “An introduction to biometric recognition,” IEEE Trans. Circuits Syst. Video Technol., vol. 14, no. 1, pp. 4–20, Jan. 2004.

[6] M. Ferrara, A. Franco, and D. Maltoni, “The magic passport,” in Proc. IEEE Int. Joint Conf. Biometrics, Sep. 2014, pp. 1–7.

[7] B. Bayar and M. C. Stamm, “A deep learning approach to universal image manipulation detection using a novel convolutional layer,” in Proc. 4th ACM Workshop Inf. Hiding Multimedia Secur., Jun. 2016, pp. 5–10.

[8] H. H. Nguyen, J. Yamagishi, and I. Echizen, “Capsule-forensics: Using capsule networks to detect forged images and videos,” 2018, arXiv:1810.11215. [Online]. Available: http://arxiv.org/abs/1810.11215

[9] D. Cozzolino, J. Thies, A. Rössler, C. Riess, M. Nießner, and L. Verdoliva, “Forensictransfer: Weakly-supervised domain adaptation for forgery detection,” 2018, arXiv:1812.02510. [Online]. Available: http://arxiv.org/abs/1812.02510

[10] T. Allison, A. Puce, and G. McCarthy, “Social perception from visual cues: Role of the STS region,” Trends Cognit. Sci., vol. 4, no. 7, pp. 267–278, Jul. 2000.

[11] T. Fiske and S. E. Taylor, Social Cognition: From Brains to Culture. Newbury Park, CA, USA: Sage, 2013.

[12] S. Schindler, E. Zell, M. Botsch, and J. Kissler, “Differential effects of face-realism and emotion on event-related brain potentials and their implications for the uncanny valley theory,” Sci. Rep., vol. 7, no. 1, p. 45003, Jul. 2017.

[13] T. Wheatley, A. Weinberg, C. Looser, T. Moran, and G. Hajcak, “Mind perception: Real but not artificial faces sustain activity beyond the N170/VPP,” PLoS ONE, vol. 6, no. 3, Mar. 2011, Art. no. e17960.

[14] A. Graves, G. Wayne, and I. Danihelka, “Neural turing machines,” 2014, arXiv:1410.5401. [Online]. Available: http://arxiv.org/abs/1410.5401

[15] T. Fernando, S. Denman, A. McFadyen, S. Sridharan, and C. Fookes, “Tree memory networks for modelling long-term temporal dependencies,” Neurocomputing, vol. 304, pp. 64–81, Aug. 2018.

[16] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770–778.

[17] K. Cho et al., “Learning phrase representations using RNN encoder-decoder for statistical machine translation,” 2014, arXiv:1406.1078. [Online]. Available: http://arxiv.org/abs/1406.1078

[18] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2009, pp. 248–255.

[19] D. Pitcher, V. Walsh, G. Yovel, and B. Duchaine, “TMS evidence for the involvement of the right occipital face area in early face processing,” Curr. Biol., vol. 17, no. 18, pp. 1568–1573, Sep. 2007.

[20] N. Kanwisher, J. McDermott, and M. M. Chun, “The fusiform face area: A module in human extrastriate cortex specialized for face perception,” J. Neurosci., vol. 17, no. 11, pp. 4302–4311, Jun. 1997.

[21] K. De, K. Sunkavalli, M. K. Johnson, D. Vlasic, W. Matusik, and H. Pfister, “Video face replacement,” ACM Trans. Graph., vol. 30, no. 6, p. 130, 2011.

[22] P. Garrido, L. Valgards, O. Rehmsen, T. Thormaehlen, P. Perez, and C. Theobalt, “Automatic face reenactment,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2014, pp. 4217–4224.

[23] P. Garrido et al., “VDub: Modifying face video of actors for plausible visual alignment to a dubbed audio track,” Comput. Graph. Forum, vol. 34, no. 2, pp. 193–204, May 2015.

[24] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Niessner, “Face2Face: Real-time face capture and reenactment of RGB videos,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 2387–2395.

[25] H. Kim et al., “Deep video portraits,” ACM Trans. Graph., vol. 37, no. 4, p. 163, 2018.

[26] S. Suvajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, “Synthesizing obama: Learning lip sync from audio,” ACM Trans. Graph., vol. 36, no. 4, p. 95, Jul. 2017.

[27] H.-A. Elor, D. Cohen-Or, J. Kopf, and M. F. Cohen, “Bringing portraits to life,” ACM Trans. Graph., vol. 36, no. 6, p. 196, 2017.

[28] G. Lample et al., “Fader networks: Manipulating images by sliding attributes,” in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 5967–5976.

[29] P. Upchurch et al., “Deep feature interpolation for image content changes,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 7064–7073.

[30] T. Karras, T. Aila, S. Laine, and J. Lehtinen, “Progressive growing of GANs for improved quality, stability, and variation,” 2017, arXiv:1710.10196. [Online]. Available: http://arxiv.org/abs/1710.10196

[31] T. Karras, S. Laine, and T. Aila, “A style-based generator architecture for generative adversarial networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 4401–4410.

[32] Y. Huang et al., “FakePolisher: Making DeepFakes more detection-evasive by shallow reconstruction,” in Proc. 28th ACM Int. Conf. Multimedia, Oct. 2020, pp. 1217–1226.

[33] W. Quan, K. Wang, D.-M. Yan, and X. Zhang, “Distinguishing between real and computer-generated images using convolutional neural networks,” IEEE Trans. Inf. Forensics Security, vol. 13, no. 11, pp. 2772–2787, Nov. 2018.
E. Parisotto and R. Salakhutdinov, “Neural map: Structured memory et al.
M. Gemici M. Malinowski and M. Fritz, “A multi-world approach to question T. Munkhdalai and H. Yu, “Neural semantic encoders,” in T. Fernando, S. Denman, S. Sridharan, and C. Fookes, “Soft FERNANDO et al. Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, “Star- A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, et al. M. Qi H. Qi et al. Y. Huang et al. M. Ng H. Qi et al. M. Malinowski and M. Fritz, “A multi-world approach to question T. Munkhdalai and H. Yu, “Neural semantic encoders,” in T. Fernando, S. Denman, S. Sridharan, and C. Fookes, “Soft+ hard-wired attention: An LSTM framework for human trajectory prediction and abnormal event detection,” Neural Netw., vol. 108, pp. 466–478, Dec. 2018. T. Munkhdalai and H. Yu, “Neural semantic encoders,” in Proc. 15th Conf. Eur. Chap. Assoc. Comput. Linguistics, vol. 1, 2017, pp. 387–407. M. Malinowski and M. Fritz, “A multi-world approach to question answering about real-world scenes based on uncertain input,” in Proc. Adv. Neural Inf. Process. Syst., 2014, pp. 1682–1690. C. Xiong, S. Merity, and R. Socher, “Dynamic memory networks for visual and textual question answering,” in Proc. Int. Conf. Mach. Learn., 2016, pp. 2397–2406. T. Fernando, S. Denman, S. Sridharan, and C. Fookes, “Pedestrian trajectory prediction with structured memory hierarchies,” in Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases. Cham, Switzerland: Springer, 2018, pp. 241–256. E. Parratito and R. Salakhutdinov, “Neural map: Structured memory for deep reinforcement learning,” 2017, arXiv:1702.08360. [Online]. Available: http://arxiv.org/abs/1702.08360 S. Sukhbaatar et al., “End-to-end memory networks,” in Proc. Adv. Neural Inf. Process. Syst., vol. 2015, pp. 2440–2448. M. Denman et al., “Generative temporal models with memory,” 2017, arXiv:1702.04649. [Online]. Available: http://arxiv.org/abs/1702.04649 A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, “Meta-learning with memory-augmented neural networks,” in Proc. Int. Conf. Mach. Learn., 2016, pp. 1842–1850. A. Santoro et al., “A simple neural network module for relational reasoning,” in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 4967–4976. D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning internal representations by error propagation,” Inst. Cogn. Sci., Univ. California San Diego, La Jolla, CA, USA, ICS Rep. 8506, 1985. Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, “Hierarchical attention networks for document classification,” in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol., 2016, pp. 1480–1489. S. K. Pal and S. Mitra, “Multilayer perceptron, fuzzy sets, and classification,” IEEE Trans. Neural Netw., vol. 3, no. 5, pp. 683–697, Sep. 1992. Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, “StarGAN: Unified generative adversarial networks for multi-domain Image-to-Image translation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 8797–8797. M. Zhang, K. T. Ma, J. H. Lim, Q. Zhao, and J. Feng, “Deep future gaze: Gaze anticipation on egocentric videos using adversarial networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 4372–4381. T. Fernando, S. Denman, S. Sridharan, and C. Fookes, “Task specific visual saliency prediction with memory augmented conditional generative adversarial networks,” in Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), Mar. 2018, pp. 1539–1548. A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, “Social GAN: Socially acceptable trajectories with generative adversarial networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 2255–2264. A. Sadeghian, V. Kosaraju, A. Sadeghian, N. Hirose, S. Hamid Rezatofighi, and S. Savarese, “SoPhie: An attentive GAN for predicting paths of pedestrians under physical constraints,” 2018, arXiv:1806.01482. P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” 2017, arXiv:1611.07004. [Online]. Available: https://arxiv.org/abs/1611.07004 A. Rössler, D. Cuzzolinio, L. Verdoliva, C. Riess, J. Thies, and Nießner, “FaceForensics++: Learning to detect manipulated facial images,” 2019, arXiv:1901.08971. [Online]. Available: http://arxiv.org/abs/1901.08971 Faceswap. Accessed: May 29, 2019. [Online]. Available: https://github.com/MarcKowalski/FaceSwap A. Khodabakhsh, R. Ramachandra, K. Raja, P. Wasnik, and C. Busch, “Fake face detection methods: Can they be generalized?” in Proc. Int. Conf. Biometrics Special Interest Group (BIOSIG), Sep. 2018, pp. 1–6. P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR), vol. 1, Dec. 2001, pp. I–I. H. H. Nguyen, Y. N. D. Tieu, H. Q. Nguyen-Son, V. Nozick, J. Yamagishi, and I. Echizen, “Modular convolutional neural network for discriminating between computer-generated images and photographic images,” in Proc. 13th Int. Conf. Availability, Rel. Secur., Aug. 2018, pp. 1–10. P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, “Two-stream neural networks for tampered face detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jul. 2017, pp. 1831–1839. D. Cuzzolinio, G. Poggi, and L. Verdoliva, “Recasting residual-based local descriptors as convolutional neural networks: An application to image forgery detection,” in Proc. 5th ACM Workshop Inf. Hiding Multimedia Secur., 2017, pp. 159–164. F. Chollet, “Xception: Deep learning with depthwise separable convolutions,” 2016, arXiv:1610.02357. [Online]. Available: http://arxiv.org/abs/1610.02357 T. Ahonen, A. Hadid, and M. Pietikainen, “Face description with local binary patterns: Application to face recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 28, no. 12, pp. 2037–2041, Dec. 2006. A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagennet classification with deep convolutional neural networks,” in Proc. Adv. Neural Inf. Process. Syst., 2012, pp. 1097–1105. K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, arXiv:1409.1556. [Online]. Available: http://arxiv.org/abs/1409.1556 C. Szegedy et al., “Going deeper with convolutions,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 1–9. I. Jolliffe, Principal Component Analysis. New York, NY, USA: Springer, 2011. Tharindu Fernando (Member, IEEE) received the B.Sc. degree in computer science from the University of Peradeniya, Sri Lanka, and the Ph.D. degree from the Queensland University of Technology (QUT), Brisbane, QLD, Australia. He is currently a Post-Doctoral Research Fellow with the SAIVT Research Program, School Electrical Engineering and Computer Science, QUT. His research interests include human behavior analysis and prediction. Clinton Fookes (Senior Member, IEEE) received the B.Eng. degree in aerospace/aviations and the M.B.A. and Ph.D. degrees from the Queensland University of Technology (QUT), Brisbane, QLD, Australia. He is currently a Professor with the Science and Engineering Faculty, QUT. His research interests include computer vision, machine learning, and pattern recognition areas. He is a Senior Member of an Australian Institute of Policy and Science Young Tall Poppy, an Australian Museum Eureka Prize winner, and a Senior Fulbright Scholar. He serves on the Editorial Board for the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY.
Simon Denman (Member, IEEE) received the B.Eng. degree in electrical and the B.I.T. and Ph.D. degrees in object tracking from the Queensland University of Technology (QUT), Brisbane, QLD, Australia. He is currently a Senior Lecturer with the School of Electrical Engineering and Computer Science, QUT. His research interests include intelligent surveillance, video analytics, and video-based recognition.

Sridha Sridharan (Life Senior Member, IEEE) received the B.Sc. degree in electrical engineering and the M.Sc. degree in communication engineering from the University of Manchester, U.K., and the Ph.D. degree from the University of New South Wales, Sydney, NSW, Australia. He is currently with the Queensland University of Technology (QUT), where he is a Professor with the School Electrical Engineering and Computer Science. He is also the Leader of the Research Program in Speech, Audio, Image, and Video Technologies (SAIVT), QUT, with strong focus in the areas of computer vision, pattern recognition, and machine learning. He has published more than 600 articles consisting of publications in journals and in refereed international conferences in the areas of Image and Speech technologies during the period 1990–2019. During this period, he has also graduated 75 Ph.D. students in the areas of Image and Speech technologies. He has also received a number of research grants from various funding bodies including Commonwealth competitive funding schemes, such as the Australian Research Council (ARC) and the National Security Science and Technology (NSST) unit. Several of his research outcomes have been commercialized.