Attention Aware Wavelet-based Detection of Morphed Face Images

Poorya Aghdaie, Baaria Chaudhary, Sobhan Soleymani, Jeremy Dawson, Nasser M. Nasrabadi
West Virginia University

Abstract

Morphed images have exploited loopholes in the face recognition checkpoints, e.g., Credential Authentication Technology (CAT), used by Transportation Security Administration (TSA), which is a non-trivial security concern. To overcome the risks incurred due to morphed presentations, we propose a wavelet-based morph detection methodology which adopts an end-to-end trainable soft attention mechanism . Our attention-based deep neural network (DNN) focuses on the salient Regions of Interest (ROI) which have the most spatial support for morph detector decision function, i.e, morph class binary softmax output. A retrospective of morph synthesizing procedure aids us to speculate the ROI as regions around facial landmarks , particularly for the case of landmark-based morphing techniques. Moreover, our attention-based DNN is adapted to the wavelet space, where inputs of the network are coarse-to-fine spectral representations, 48 stacked wavelet sub-bands to be exact. We evaluate performance of the proposed framework using three datasets, VISAPP17, LMA, and MorGAN. In addition, as attention maps can be a robust indicator whether a probe image under investigation is genuine or counterfeit, we analyze the estimated attention maps for both a bona fide image and its corresponding morphed image. Finally, we present an ablation study on the efficacy of utilizing attention mechanism for the sake of morph detection.

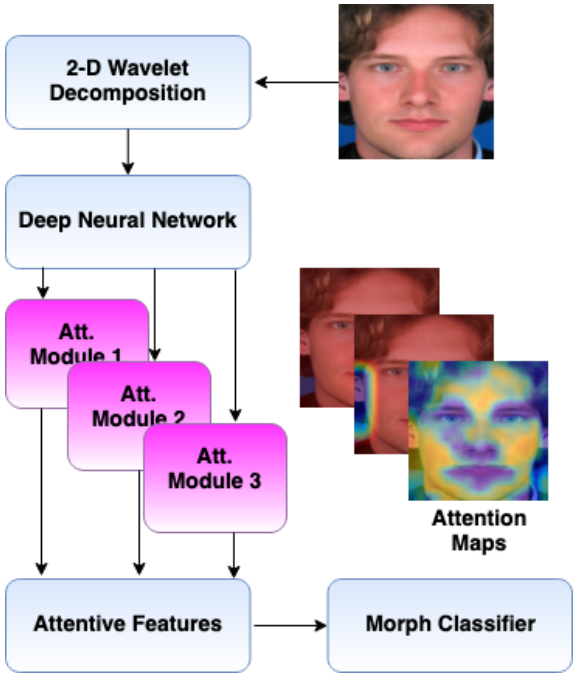

Figure 1. Our Proposed Deep Attention-based Morph Detector. Wavelet sub-bands of the input image is fed into our DNN during training phase of the network. Three Attention modules, i.e., Att. modules, generate the new attention weighted features, i.e., attentive features, as well as the attention maps. Attentive features are used to detect morphed images. Please note that the attention maps are generated after training the DNN.

1. Introduction

Robust, reliable verification systems are the crucial backbones of biometric document authentication protocols, that are to operate flawlessly. Although image morphing is not a new paradigm, it was first identified as a security concern by Ferrara *et al*. [8], who explained how a criminal can dodge a border control checkpoint using a travel document that was issued with a morphed image. The goal of the face image morphing attack is to synthesize a forged imaged from two composing original images such that the artificially crafted morphed image can be verified against the two original images not only visually, but also in the feature space by a classifier [28]. Moreover, morphed samples can be labeled as hard positive samples in comparison to negative genuine samples because morphed samples are synthesized to intentionally lie on the negative samples' manifold. Similar to adversarially perturbed data samples that fool classification networks into a wrong predicted class [11,16], morphed images are crafted to lead a verifier into a false acceptance.

Detecting morphed images has garnered a great deal of attention from the biometrics research community because of its crucial impact on the security protocols [21], especially those used for authenticating travel documents. The vast majority of research efforts has dealt with morphing attacks through either using hand-crafted texture features

to find a discriminative hyperplane between the positive (morphed), and negative (genuine) samples [7, 27, 30, 42], or harvesting those features for learning a deep classifier [6, 29]. Recently, the visual attention mechanism has taken computer vision community with storm. First introduced in [20], the visual attention mechanism has emerged as a powerful by-product of DNNs, which can boost visual recognition performance on a variety of datasets considerably [10, 13, 17, 38, 40].

In this paper, we present an attention-based DNN in the wavelet domain for detecting morphed samples. To the best of our knowledge, this is the first work which incorporates attention mechanism into a deep morph detector. Our proposed network employs attention to focus on Regions of Interest (ROI) in terms of morph detection, that are specifically landmarks around the eyes and hairline in the landmark-based facial image morphing attacks.

Wavelet sub-bands of an image represent information with different time-frequency granularity that are adapted to our DNN as input. The soft attention mechanism used in a given layer of our DNN retains spatial regions in the layer's resulting feature maps that represent the discriminative regions, and discard those pixels that are outside the discriminative regions. Fig. 1 shows an overview of our proposed deep attention-based morph detector. We utilize wavelet sub-bands instead of the raw images since we can easily discard frequency contents, sub-bands, which are not discriminative for morph detection such as the low-low (LL) sub-bands. Most importantly, we validate performance of our method through extensive experiments on the three morph datastes: VISAPP17 [19], LMA [5], and MorGAN [5]. Moreover, estimated attention maps are obtained for both real and morphed images. The contribution of this work are as follows:

- Incorporating an end-to-end trainable soft attention mechanism into deep morph detector network.
- Tailoring wavelet sub-bands for our deep attention-based morph detector.
- Training our deep attention-based network using the three datasets, as well as a combination of all the three datasets, which is coined "universal" dataset.

2. Related Works

2.1. Morph Generation

Facial morph generation techniques are categorized into two types, i.e., landmark-based morphing [5, 8, 19, 31], and GAN-based morphing [5, 37]. In the landmark-based morphing attack, appearance of a resulting morphed image is associated with that of two underlying subject's bona fide face images, while geometric locations of its landmarks are the average of the corresponding landmarks in the two bona fide images [32]. By applying Delaunay triangulation on the two bona fide images, corresponding regions on the two facial images are further warped and mixed through alpha blending to synthesize the morphed image. Generative Adversarial Networks (GANs) are also employed for synthesizing morphed images. In [5], morphed images are generated using a GAN which incorporates an encoder in its generator to model latent space. In addition, morphed images can be generated using StyleGAN [1, 15].

2.2. Morph Detection

Different texture descriptors such as Local Binary Patterns (LBP), Histogram of Gradients (HOG), Speeded Up Robust Features (SURF), and Scale-Invariant Feature Transform (SIFT) are utilized for detecting morphed images [6, 23, 24, 29]. Discrepancies in locations of facial landmarks in a morphed image, and a live capture of the corresponding bona fide can be exploited for morph detection [26]. Convolutional Neural Networks (CNN), as well as deep embedding features have represented promising performance in morph detection [2,4,9,25]. In [7], spectral behaviour of Photo Response Non-Uniformity (PRNU) is studied to detect morphed images. The resulting noise artifact in the face morphing pipeline can be adopted for morph detection. In particular, residual noise is an established indicator for morph detection [35, 36].

2.3. Attention Mechanism

Attention mechanism has been widely used for the visual recognition tasks such as image caption generation and visual question answering (VQA) [13,40]. Two dominant categories of the attention mechanism are the soft deterministic attention and the hard stochastic attention [40]. The soft attention can either be adopted in a post-hoc manner, or it can be trained along with a DNN using back-propagation [13]. The hard attention mechanism is trained using a method called REINFORCE [39]. Attentive recurrent neural networks (RNNs) [20] are another variant of networks where expolits attention mechanism to amplify ROI and suppress background clutter.

3. Our Framework

Our attention-based morph detector is displayed in Fig. 2. Based on Fig. 2, the input images are initially decomposed into 48 uniform wavelet sub-bands that are further stacked channel-wise and then passed to our morph detector. Our morph detector leverages three attention modules at three different convolutional layers, denoted by $\mathcal{L}_1$, $\mathcal{L}_2$, and $\mathcal{L}_3$. The local feature vectors resulting from the three convolutional layers $\mathcal{L}_1$, $\mathcal{L}_2$, and $\mathcal{L}_3$ are denoted by $L_{feat.1}$, $L_{feat.2}$, and $L_{feat.3}$, respectively. The attention weighted local features for a given convolutional layer are obtained

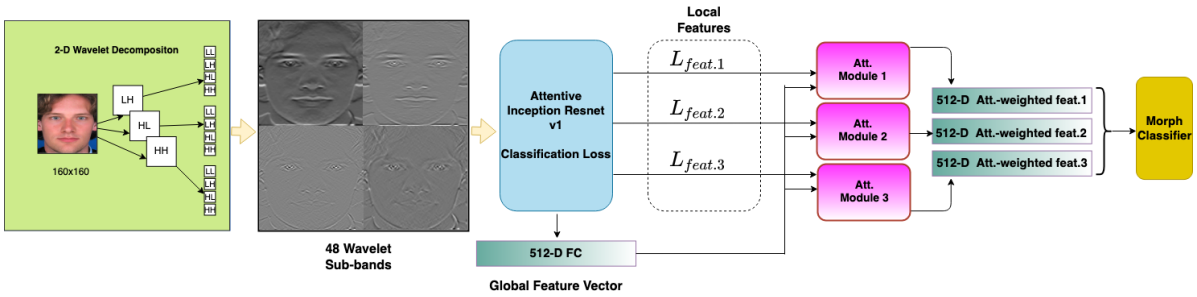

Figure 2. Our deep attention-based morph detector. The input images are initially decomposed into 48 uniform wavelet sub-bands, which are fed into our morph detector. Attention modules are placed at three convolutional layers, namely $\mathcal{L}_1$, $\mathcal{L}_2$, and $\mathcal{L}_3$. The $L_{feat.1}$, $L_{feat.2}$, and $L_{feat.3}$ represent the local features vectors of layers $\mathcal{L}_1$, $\mathcal{L}_2$, and $\mathcal{L}_3$, respectively. The 512-D attention weighted local features in a layer, shown by 512-D Att. weighted feat., are obtained using the local features of the layer, and the 512-D FC global feature vector. The three resulting attention weighted features are concatenated to form our new attended features.

using the layer's local features, and the global feature vector resulting from the 512-D fully connected (FC) layer in our network, e.g., the first 512-D attention weighted local features in the $\mathcal{L}_1$ layer, shown by 512-D Att. weighted feat.1, are obtained using the local features of $\mathcal{L}_1$, that is to say $L_{feat.1}$, and the 512-D FC global feature vector. The three resulting attention weighted features are concatenated and passed into a new FC layer with 512×3 neurons.

3.1. Uniform Wavelet Decomposition

Most artifacts due to facial image morphing techniques lie within the high frequency spectrum, and using wavelet decomposition allows us to cherry-pick the desired wavelet sub-bands by discarding the low-frequency sub-bands. Therefore, using specific wavelet sub-bands instead of the original image is highly justified in our study. We apply three-level undecimated 2-D wavelet decomposition on both bona fide and morphed images. Analyzing the wavelet sub-bands of a bona fide and its corresponding morphed image justifies considering the high frequency spectra for the task of morph detection. In other words, we discard the low-low (LL) wavelet sub-band after first level of decomposition, and we keep the low-high (LH), high-low (HL), high-high (HH) for the second and third levels of decomposition. In total, 48 wavelet sub-bands are stacked channel-wise, which are utilized as the input to our attention-based morph detector. Decomposing an RGB image into 48 wavelet sub-bands leads to decoupled spectra, focusing on the frequency contents that are discriminative in terms of distinguishing between bona fide and morphed images.

3.2. Integrating Attention-Weighted Features

To distinguish between bona fide and morphed images, we adopt the end-to-end trainable soft attention mechanism introduced in [13]. This soft attention mechanism is differentiable with respect to the network parameters. We show that our attention-based network can meticulously focus on the regions that contribute the most to detecting morphed images. We insert three attention modules at three different convolutional layers $\mathcal{L}_1$, $\mathcal{L}_2$, and $\mathcal{L}_3$ in our DNN. Therefore, as presented in Fig. 2, instead of a single global feature vector, that is the 512-D fully connected layer (FC) output, we concatenate the three attention-weighted local feature vectors at three different convolutional layers to accomplish the classification task. These attention maps at each convolutional layer reveals the importance of each spatial location in the layers' feature maps.

Suppose that a spatial local feature vector in the location $i \in \{1, 2, .., n\}$ in the convolutional layer $\mathcal{L}_k$, $1 \leq k \leq 3$, is shown by $\boldsymbol{\ell}_i^{\mathcal{L}_k}$. As presented in Fig. 2, $L_{feat.k} = \{\boldsymbol{\ell}_1^{\mathcal{L}_k}, \boldsymbol{\ell}_2^{\mathcal{L}_k}, ..., \boldsymbol{\ell}_n^{\mathcal{L}_k}\}$. The compatibility score for each spatial location, i, represents the importance of that pixel for detecting morphed images. The compatibility score for local feature vector $\boldsymbol{\ell}_i^{\mathcal{L}_k}$ is given as:

$$c_i^{\mathcal{L}_k} = \langle \boldsymbol{\ell}_i^{\mathcal{L}_k}, \boldsymbol{g} \rangle, i \in \{1, 2, .., n\}, \tag{1}$$

where $\boldsymbol{g}$ designates the global feature vector, that is the 512-D output of the fully connected layer and $\langle ., . \rangle$ represents the inner product. We further normalize the computed compatibility scores in a given convolutional layer $\mathcal{L}_k$ using the softmax normalization, which is given as:

$$a_i^{\mathcal{L}_k} = \frac{\exp(c_i^{\mathcal{L}_k})}{\Sigma_{i=1}^{i=n} \exp(c_i^{\mathcal{L}_k})}, i \in \{1, 2, .., n\}. \tag{2}$$

A linear combination of the local feature vectors $\boldsymbol{\ell}_i^{\mathcal{L}_k}$ and the attention weights $a_i^{\mathcal{L}_k}$ yields the attentive local descriptor for the given convolutional layer $\mathcal{L}_k$. The global feature vector, i.e., attention-weighted feature vector, can be written as:

$$\boldsymbol{g}_a^{\mathcal{L}_k} = \Sigma_{i=1}^{i=n} a_i^{\mathcal{L}_k} \boldsymbol{\ell}_i^{\mathcal{L}_k}. \tag{3}$$

We concatenate the estimated attention weighted local features at three different convolutional layers which are fed

Table 1. Performance of single morph detection: D-EER%, BPCER@APCER=5%, and BPCER@APCER=10%.

Dataset	Algorithm	D-EER	5%	10%
VISAPP17	BSIF+SVM [14]	16.51	35.61	26.79
	SIFT+SVM [18]	38.59	82.40	75.60
	LBP+SVM [22]	38.00	77.10	67.90
	SURF+SVM [3]	30.45	84.70	69.40
	RGB+DNN [33]	1.76	0.588	0.58
	Ours	**0.00**	**0.00**	**0.00**
LMA	BSIF+SVM [14]	33.05	78.34	62.86
	SIFT+SVM [18]	33.30	83.40	72.00
	LBP+SVM [22]	28.00	58.60	51.40
	SURF+SVM [3]	37.40	79.50	70.00
	RGB+DNN [33]	9.10	15.18	7.49
	Ours	**8.71**	**17.86**	**6.52**
MorGAN	BSIF+SVM [14]	1.57	1.42	1.30
	SIFT+SVM [18]	43.50	93.20	84.20
	LBP+SVM [22]	20.10	52.70	32.30
	SURF+SVM [3]	39.95	80.00	72.60
	RGB+DNN [33]	2.44	1.88	1.50
	Ours	**0.00**	**0.00**	**0.00**

Table 2. Performance of single morph detection: D-EER%, BPCER@APCER=5%, and BPCER@APCER=10%.

Train	Test	Algorithm	D-EER	5%	10%
Universal(VISAPP17+LMA+MorGAN)	VISAPP17	BSIF+SVM [14]	35.00	67.20	59.00
		SIFT+SVM [18]	27.00	83.20	70.90
		LBP+SVM [22]	37.67	72.50	59.50
		SURF+SVM [3]	31.00	79.40	70.10
		RGB+DNN [33]	0.00	0.00	0.00
		Ours	**0.00**	**0.00**	**0.00**
	LMA	BSIF+SVM [14]	30.00	70.42	57.60
		SIFT+SVM [18]	28.31	67.70	50.00
		LBP+SVM [22]	29.00	61.50	51.20
		SURF+SVM [3]	33.40	74.50	62.70
		RGB+DNN [33]	7.80	13.00	6.10
		Ours	**8.11**	**14.21**	**6.83**
	MorGAN	BSIF+SVM [14]	28.80	62.42	45.70
		SIFT+SVM [18]	47.60	92.30	88.60
		LBP+SVM [22]	31.20	62.00	55.60
		SURF+SVM [3]	38.67	76.00	70.00
		RGB+DNN [33]	4.69	4.70	2.74
		Ours	**2.59**	**1.50**	**0.89**
	Universal	BSIF+SVM [14]	23.74	51.42	38.67
		SIFT+SVM [18]	37.21	87.45	76.71
		LBP+SVM [22]	38.80	91.36	83.40
		SURF+SVM [3]	36.00	75.50	65.76
		RGB+DNN [33]	5.57	6.08	3.00
		Ours	**6.42**	**7.58**	**3.46**

into a FC layer having size of 512×3 followed by a 2-neuron FC layer, which generates the binary logits for detecting morphed images.

4. Experimental Setup

4.1. Datasets

In this study, three different morphed image datasets are used that are, the VISAPP17 [19], LMA [5], and MorGAN [5]. The VISAPP17 dataset is generated using landmark-based face morphing attack, followed by splicing. In the landmark-based morphing pipeline locations of the corresponding landmarks in two bona fide subjects are averaged, and facial regions are divided using Delaunay triangulation before their alpha blending. LMA is a landmark-based morphed image dataset, and MorGAN dataset is generated using a generative model, GAN to be exact. Contrary to the landmark-based morphing attack, which captures geometry of underlying bona fide images, GAN-based morphing attacks synthesize morphed images after capturing the underlying distributions of bona fide facial images.

MTCNN [41] is utilized for face detection and alignment. Face images are resized to 160×160 pixels. For each dataset, 50% of the subjects are considered for training while the other 50% are used for the test set. In addition, 15% of the training set is selected during model optimization as the validation set. The train-test split is disjoint, with no overlapping bona fides, morphs, or bona fides contributing to morphs. In addition to the individual datasets, we combine the three datasets into a *universal dataset*. Regarding the universal dataset, the training set includes 1631 bona fide, and 1183 morphed samples. The validation set contains 462 bona fide, and 167 morphed subjects. In addition, the test set is composed of 1631 bona fide, and 1183 morphed images.

4.2. Training Setup

For the backbone of our attention-based morph detector, we employ Inception-ResNet-v1 [33] , which harnesses the residual skips [12], as well as the revised version of the Inception network [34]. We add three attention modules to the network at $\mathcal{L}_1$ = *"conv2d_4b"*, $\mathcal{L}_2$ = *"mixed_6a"*, and $\mathcal{L}_3$ = *"mixed_7a"*. Since the number of channels in the resulting feature vectors related to the three convolutional layers are not 512-D, we project the feature vectors to new vectors where number of channels are 512. The projection in each convolutional layer is achieved using the 1×1 convolutional filters, where 512 kernels with the size of 1×1 are employed. The Adam optimizer updates the weights of our DNN accelerated using two 12 GB TITAN X (Pascal) GPUs. Batch size of 8 is considered for training.

4.3. Performance of the Attention-based Morph Detector

Standard quantitative measures are used to evaluate the effectiveness of our proposed method. The first measure

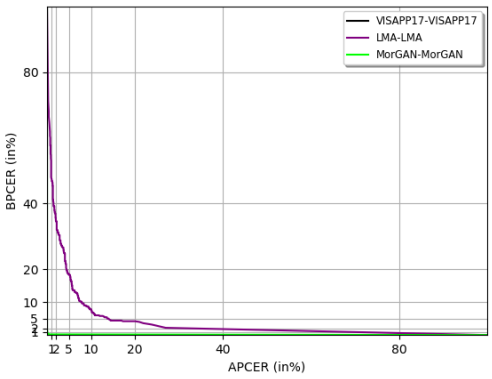

Figure 3. DET curves when our attention-based morph detector is trained using the individual datasets.

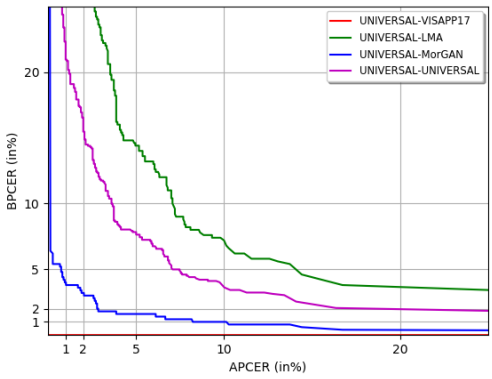

Figure 4. DET curves when our attention-based morph detector is trained using the training portion of the universal dataset.

is Attack Presentation Classification Error Rate (APCER), which is the percentage of morphed images that are classified as bona fide. The second measure is Bona Fide Presentation Classification Error Rate (BPCER), which represents the percentage of bona fide samples that are classified as morphed. If we label the morphed class as positive and the bona fide class as negative , APCER, and BPCER are equivalent to false negative rate and false positive rate, respectively. Detection error trade-off (DET) curves represent performance of our attention-based DNN. D-EER stands for the Detection Equal Error Rate, where APCER equals BPCER. BPCER5 represents BPCER rate for APCER=5%, and BPCER10 represents BPCER rate for APCER=10%.

We train our attention-based DNN using the three datasets, that are the VISAPP17, LMA, and MorGAN. Table 1 delineates the performance of the baseline methods, as well as our attention-based morph detector for the three datasets. In addition, Fig. 3 depicts the detection error trade-off (DET) curves for the three datasets.

Moreover, we scrutinize the scenario where all three datasets are combined, which was coined the universal dataset. Therefore, we train our network using the training portion of the universal dataset, and test set comes from all individual datasets, as well as the testing portion of the universal dataset. The performance of our attention-based morph detector when trained on the universal dataset is summarised in Table 2, and Fig. 4 depicts the DET curves when the attention-based DNN is trained using the universal dataset. Our attention-based morph detector can detect morphed samples in the VISAPP17 and MorGAN datasets accurately when the network is trained on each dataset.

4.4. Estimated Attention Maps

The estimated attention maps, resulting from the three attentions modules are shown in Fig. 5. It is worth mentioning that the heatmaps of the resulting attention maps are applied to the image for visualization purpose. The first row, shows the bona fide image, and its corresponding attention maps from the three different convolutional layers. The second row is related to the attention maps of the morphed image. Comparing the attention map of the $\mathcal{L}_1$ for the bona fide image with that of the morphed image reveals that the morphed images has more attended areas that is caused by the morphing attack pipeline, which comprises landmark manipulation for this image, coming from the VISAPP17 dataset. Given the attention maps of the $\mathcal{L}_2$ and $\mathcal{L}_3$, there are salient impacted regions in the feature maps of the morphed images, while there is no obvious attentive regions in the bona fide image.

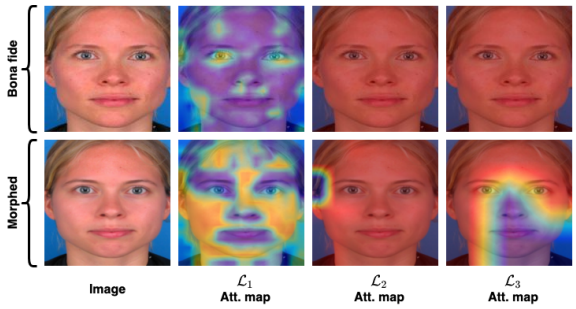

Figure 5. Estimated attention maps for a bona fide and the corresponding morphed image obtained from the three attention modules.

4.5. Ablation Study

In this section, we delve into the effect of the attention modules on the performance of our attention-based morph detector. To this end, we compare the performance of the morph detector when the number of attention modules are one, two, or three. We have already studied the case where

Table 3. Performance of single morph detection for the different number of attention-modules: D-EER%, BPCER@APCER=5%, and BPCER@APCER=10%.

Dataset	Att. Layers	D-EER	5%	10%
VISAPP17	$\mathcal{L}_3$	00.00	00.00	00.00
	$\mathcal{L}_2$+$\mathcal{L}_3$	00.00	00.00	00.00
	$\mathcal{L}_1$+$\mathcal{L}_2$+$\mathcal{L}_3$	00.00	00.00	00.00
LMA	$\mathcal{L}_3$	12.45	21.23	15.18
	$\mathcal{L}_2$+$\mathcal{L}_3$	12.12	23.58	17.21
	$\mathcal{L}_1$+$\mathcal{L}_2$+$\mathcal{L}_3$	8.71	17.86	6.52
MorGAN	$\mathcal{L}_3$	00.00	00.00	00.00
	$\mathcal{L}_2$+$\mathcal{L}_3$	00.00	00.00	00.00
	$\mathcal{L}_1$+$\mathcal{L}_2$+$\mathcal{L}_3$	00.00	00.00	00.00

number of attention modules are three in the section 4.3. We plot the DET curves for the individual datasets for every number of attention modules. Table 3 delineates the performance of our morph detector when trained on the individual datasets for the following cases: 1- One attention module placed at $\mathcal{L}_3$ = *"mixed_7a"*, 2- two attention modules placed at $\mathcal{L}_3$ = *"mixed_7a"* and $\mathcal{L}_2$ = *"mixed_6a"*, 3- three attention modules at $\mathcal{L}_3$ = *"mixed_7a"*, $\mathcal{L}_2$ = *"mixed_6a"*, and $\mathcal{L}_1$ = *"conv2d_4b"*. Also, Table 4 summarizes the performance of our morph detector when trained on the universal dataset for the above-mentioned cases.

Fig. 6 depicts the performance of our morph detector using the two attention modules when our morph detector is trained using the individual datasets. Fig. 7 shows the performance of our morph detector using the two attention modules when our DNN is trained using the universal dataset. Moreover, Fig. 8 displays the performance of our attention-based morph detector with the one attention module that is trained on the individual datasets, and Fig. 9 displays the performance of our attention-based morph detector using the one attention module when the network is trained on the universal dataset. It is evident from Table 3 that the most accurate morph detection for the LMA dataset is achieved when there are three attention modules in our proposed network. Concerning Table 4, the attention-based DNN with three attention modules outperforms the network which has either one or two attention modules.

5. Conclusion

In this paper, we study the application of attention mechanism for detecting morphed images. More importantly, our attention-based model is adapted to a wavelet-based Inception-ResNet-v1, where all input images are decomposed into 48 wavelet sub-bands. The three integrated attention modules can emphasize the artifacts stem from the morphing attack, leading to detecting morphed images accurately. Most importantly, our attention-based morph detector can detect morphed images in the VISAPP17 and

Table 4. Performance of the universal training set single morph detection for different number of attention-modules: D-EER%, BPCER@APCER=5%, and BPCER@APCER=10%.

Train	Test	Att. Layers	D-EER	5%	10%
Universal	VISAPP17	$\mathcal{L}_3$	00.00	00.00	00.00
		$\mathcal{L}_2$+$\mathcal{L}_3$	00.00	00.00	00.00
		$\mathcal{L}_1$+$\mathcal{L}_2$+$\mathcal{L}_3$	00.00	00.00	00.00
	LMA	$\mathcal{L}_3$	14.37	27.23	16.54
		$\mathcal{L}_2$+$\mathcal{L}_3$	13.24	35.36	18.61
		$\mathcal{L}_1$+$\mathcal{L}_2$+$\mathcal{L}_3$	8.11	14.21	6.83
	MorGAN	$\mathcal{L}_3$	7.21	6.31	5.02
		$\mathcal{L}_2$+$\mathcal{L}_3$	7.14	7.86	4.91
		$\mathcal{L}_1$+$\mathcal{L}_2$+$\mathcal{L}_3$	2.59	1.50	0.89
	Universal	$\mathcal{L}_3$	8.91	12.21	8.27
		$\mathcal{L}_2$+$\mathcal{L}_3$	9.95	12.23	8.93
		$\mathcal{L}_1$+$\mathcal{L}_2$+$\mathcal{L}_3$	6.42	7.58	3.46

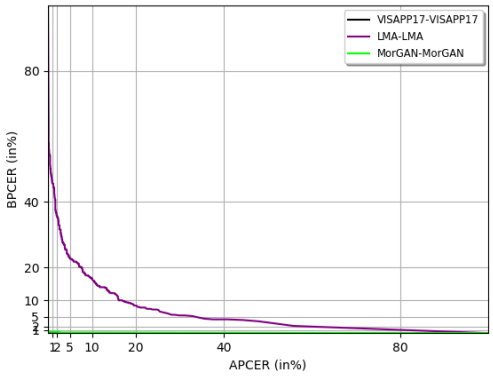

Figure 6. DET curves for the individual datasets for two attention modules.

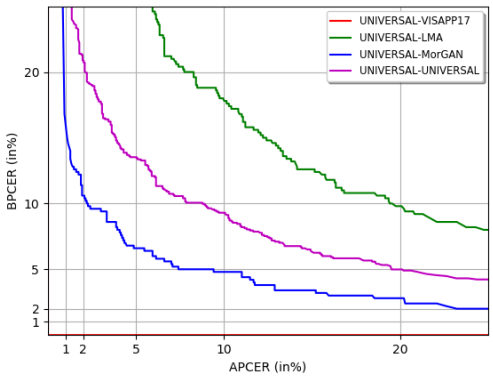

Figure 7. DET curves for the universal datasets for two attention modules.

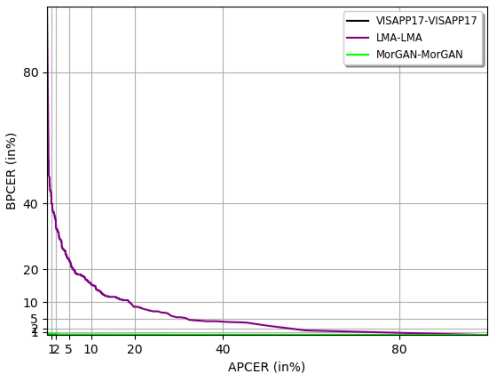

Figure 8. DET curves for the individual datasets for one attention module.

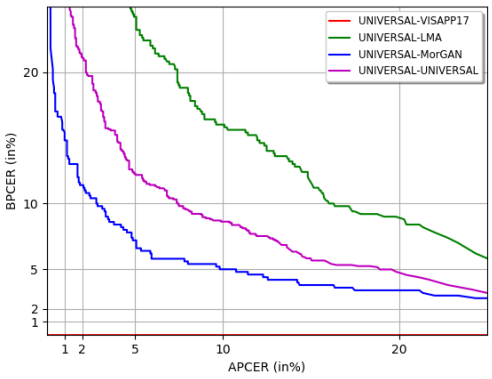

Figure 9. DET curves for the universal datasets for one attention module.

MorGAN datasets accurately. Displayed attention maps substantiates the effectiveness our algorithm in detecting morphed images, because morphed images have substantial attentive pixels compared to bona fide images. Finally, our ablation study proves the superior performance of our attention-based morph detector that uses three attention modules in comparison to a network that has either one or two attention modules.

References

[1] R. Abdal, Y. Qin, and P. Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4432–4441, 2019.

[2] P. Aghdaie, B. Chaudhary, S. Soleymani, J. Dawson, and N. M. Nasrabadi. Detection of morphed face images using discriminative wavelet sub-bands. *arXiv preprint arXiv:2106.08565*, 2021.

[3] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. In *European conference on computer vision*, pages 404–417. Springer, 2006.

[4] B. Chaudhary, P. Aghdaie, S. Soleymani, J. Dawson, and N. M. Nasrabadi. Differential morph face detection using discriminative wavelet sub-bands. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1425–1434, 2021.

[5] N. Damer, A. M. Saladie, A. Braun, and A. Kuijper. Morgan: Recognition vulnerability and attack detectability of face morphing attacks created by generative adversarial network. In *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–10. IEEE, 2018.

[6] L. Debiasi, N. Damer, A. M. Saladié, C. Rathgeb, U. Scherhag, C. Busch, F. Kirchbuchner, and A. Uhl. On the detection of gan-based face morphs using established morph detectors. In *International Conference on Image Analysis and Processing*, pages 345–356. Springer, 2019.

[7] L. Debiasi, C. Rathgeb, U. Scherhag, A. Uhl, and C. Busch. Prnu variance analysis for morphed face image detection. In *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–9. IEEE, 2018.

[8] M. Ferrara, A. Franco, and D. Maltoni. The magic passport. In *IEEE International Joint Conference on Biometrics*, pages 1–7. IEEE, 2014.

[9] M. Ferrara, A. Franco, and D. Maltoni. Face morphing detection in the presence of printing/scanning and heterogeneous image sources. *arXiv preprint arXiv:1901.08811*, 2019.

[10] H. Fukui, T. Hirakawa, T. Yamashita, and H. Fujiyoshi. Attention branch network: Learning of attention mechanism for visual explanation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10705–10714, 2019.

[11] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[13] S. Jetley, N. A. Lord, N. Lee, and P. H. Torr. Learn to pay attention. *arXiv preprint arXiv:1804.02391*, 2018.

[14] J. Kannala and E. Rahtu. Bsif: Binarized statistical image features. In *Proceedings of the 21st international conference on pattern recognition (ICPR2012)*, pages 1363–1366. IEEE, 2012.

[15] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[16] A. Kurakin, I. Goodfellow, S. Bengio, et al. Adversarial examples in the physical world, 2016.

[17] Y. Li, J. Zeng, S. Shan, and X. Chen. Occlusion aware facial expression recognition using cnn with attention mechanism. *IEEE Transactions on Image Processing*, 28(5):2439–2450, 2018.

[18] D. G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1150–1157. Ieee, 1999.

[19] A. Makrushin, T. Neubert, and J. Dittmann. Automatic generation and detection of visually faultless facial morphs. In *International Conference on Computer Vision Theory and Applications*, volume 7, pages 39–50. SCITEPRESS, 2017.

[20] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu. Recurrent models of visual attention. *arXiv preprint arXiv:1406.6247*, 2014.

[21] M. Ngan, M. Ngan, P. Grother, K. Hanaoka, and J. Kuo. *Face recognition vendor test (frvt) part 4: Morph-performance of automated face morph detection*. US Department of Commerce, National Institute of Standards and Technology, 2020.

[22] T. Ojala, M. Pietikainen, and D. Harwood. Performance evaluation of texture measures with classification based on kullback discrimination of distributions. In *Proceedings of 12th International Conference on Pattern Recognition*, volume 1, pages 582–585. IEEE, 1994.

[23] R. Raghavendra, K. Raja, S. Venkatesh, and C. Busch. Face morphing versus face averaging: Vulnerability and detection. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 555–563, 2017.

[24] R. Raghavendra, K. B. Raja, and C. Busch. Detecting morphed face images. In *2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–7, 2016.

[25] K. Raja, S. Venkatesh, R. Christoph Busch, et al. Transferable deep-cnn features for detecting digital and print-scanned morphed face images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 10–18, 2017.

[26] U. Scherhag, D. Budhrani, M. Gomez-Barrero, and C. Busch. Detecting morphed face images using facial landmarks. In *International Conference on Image and Signal Processing*, pages 444–452. Springer, 2018.

[27] U. Scherhag, L. Debiasi, C. Rathgeb, C. Busch, and A. Uhl. Detection of face morphing attacks based on prnu analysis. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 1(4):302–317, 2019.

[28] U. Scherhag, R. Raghavendra, K. B. Raja, M. Gomez-Barrero, C. Rathgeb, and C. Busch. On the vulnerability of face recognition systems towards morphed face attacks. In *2017 5th International Workshop on Biometrics and Forensics (IWBF)*, pages 1–6. IEEE, 2017.

[29] U. Scherhag, C. Rathgeb, and C. Busch. Morph deterction from single face image: A multi-algorithm fusion approach. In *Proceedings of the 2018 2nd International Conference on Biometric Engineering and Applications*, pages 6–12, 2018.

[30] C. Seibold, A. Hilsmann, and P. Eisert. Reflection analysis for face morphing attack detection. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 1022–1026. IEEE, 2018.

[31] C. Seibold, W. Samek, A. Hilsmann, and P. Eisert. Accurate and robust neural networks for security related applications exampled by face morphing attacks. *arXiv preprint arXiv:1806.04265*, 2018.

[32] S. Soleymani, A. Dabouei, F. Taherkhani, J. Dawson, and N. M. Nasrabadi. Mutual information maximization on disentangled representations for differential morph detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1731–1741, 2021.

[33] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, inception resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*, 2017.

[34] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[35] S. Venkatesh, R. Ramachandra, K. Raja, L. Spreeuwers, R. Veldhuis, and C. Busch. Morphed face detection based on deep color residual noise. In *2019 Ninth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6. IEEE, 2019.

[36] S. Venkatesh, R. Ramachandra, K. Raja, L. Spreeuwers, R. Veldhuis, and C. Busch. Detecting morphed face attacks using residual noise from deep multi-scale context aggregation network. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 280–289, 2020.

[37] S. Venkatesh, H. Zhang, R. Ramachandra, K. Raja, N. Damer, and C. Busch. Can gan generated morphs threaten face recognition systems equally as landmark based morphs?-vulnerability and detection. In *2020 8th International Workshop on Biometrics and Forensics (IWBF)*, pages 1–6. IEEE, 2020.

[38] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang. Residual attention network for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2017.

[39] R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

[40] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015.

[41] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[42] L.-B. Zhang, F. Peng, and M. Long. Face morphing detection using fourier spectrum of sensor pattern noise. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2018.