Generalized Single-Image-Based Morphing Attack Detection Using Deep Representations from Vision Transformer

Haoyu Zhang[1*] Raghavendra Ramachandra[1] Kiran Raja[1]
Christoph Busch[1,2]
[1] Norwegian University of Science and Technology, Norway
[2] Darmstadt University of Applied Sciences, Germany
{haoyu.zhang, raghavendra.ramachandra, kiran.raja, christoph.busch}@ntnu.no
christoph.busch@h-da.de

Abstract

Face morphing attacks have posed severe threats to Face Recognition Systems (FRS), which are operated in border control and passport issuance use cases. Correspondingly, morphing attack detection algorithms (MAD) are needed to defend against such attacks. MAD approaches must be robust enough to handle unknown attacks in an open-set scenario where attacks can originate from various morphing generation algorithms, post-processing and the diversity of printers/scanners. The problem of generalization is further pronounced when the detection has to be made on a single suspected image. In this paper, we propose a generalized single-image-based MAD (S-MAD) algorithm by learning the encoding from Vision Transformer (ViT) architecture. Compared to CNN-based architectures, ViT model has the advantage on integrating local and global information and hence can be suitable to detect the morphing traces widely distributed among the face region. Extensive experiments are carried out on face morphing datasets generated using publicly available FRGC face datasets. Several state-of-the-art (SOTA) MAD algorithms, including representative ones that have been publicly evaluated, have been selected and benchmarked with our ViT-based approach. Obtained results demonstrate the improved detection performance of the proposed S-MAD method on inter-dataset testing (when different data is used for training and testing) and comparable performance on intra-dataset testing (when the same data is used for training and testing) experimental protocol.

*This work was supported by the European Union's Horizon 2020 Research and Innovation Program under Grant 883356.

1. Introduction

Face recognition systems (FRS) have been widely deployed in various security applications, such as passport issuance and automated border control (ABC)[7]. However, with the development of image manipulation techniques, FRS are becoming vulnerable to different kinds of attacks that may lead to security lapses [21] [29]. Morphing attack is one type of the attacks that targets to subvert FRS by combining biometric samples from 2 or more individuals into a single morphed image. Morphing attacks have been illustrated as an evolving threat to the FRS [2]. Morphing attack detection algorithms (MAD) have been therefore proposed to detect these attacks to improve the security of FRS.

Single-image-based morphing attack detection (S-MAD) aims to detect the face morphing attack based on a single image presented to the algorithm. The most common application scenario of S-MAD is validating the face photos submitted in passport or visa applications (physically/through online services) [29]. Another possible used case for S-MAD is the validation of an existing face image database, to validate that no morphed images are contained. Hence, the S-MAD algorithm should well generalize for different types of face images and anticipated image processing, such as digital, print-scanned and print-scanned-compression. In addition, there are various types of morphing algorithms that generate morphed face images with different characteristics, such as realistic texture and high face structure similarity. While many previous works have developed MAD approaches that can detect attacks efficiently for known kinds of morphing attacks, the performance tends to degrade when testing involves data stemming from different morphing methods and which were unseen during training. Fig. 1 illustrates an example of such a scenario when the S-MAD algorithm trained on the known attack (i.e., known morphing generation type) can easily miss detecting an attack from the unknown generation type [11].

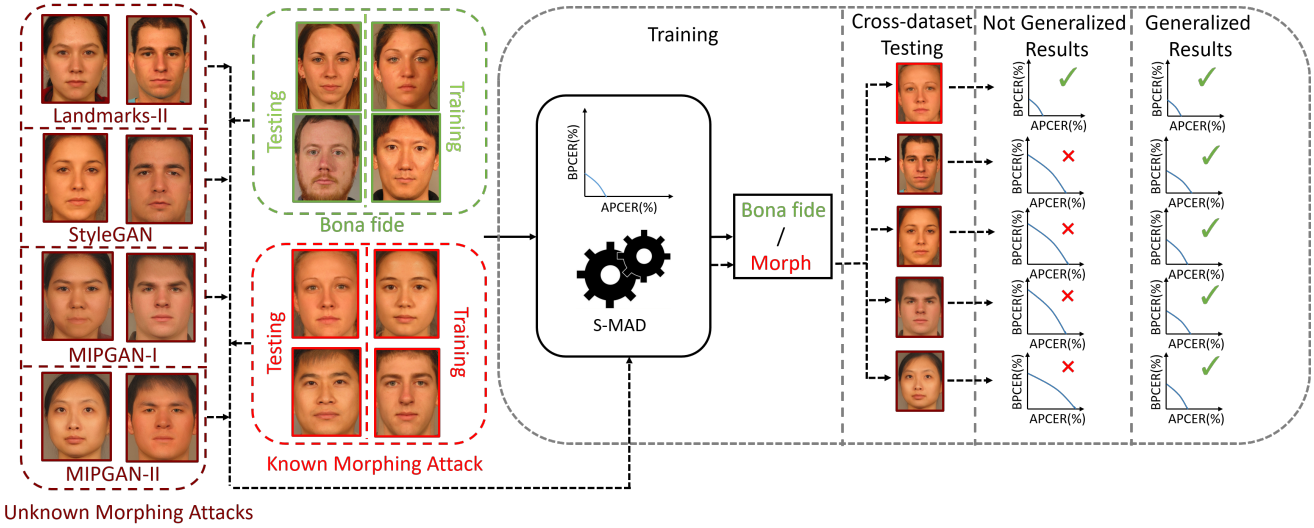

Figure 1. Hypothesised illustration of S-MAD as open-set problem: A model trained on known morphing attacks may fail at unknown morphing attacks.

Given the envisioned application scenario, it is crucial to improve the generalizability of S-MAD algorithm and to evaluate the detection performance in an open-set scenario by cross-dataset testing.

The existing S-MAD approaches are based on texture features [22], residual noise features, hybrid features, and deep learning features [29] [4] [18] [23]. With the achievement of deep convolutional neural networks (CNNs) in the field of image recognition, many researchers have applied pretrained CNNs and transfer learning to solve the S-MAD problems as binary classification problems [15] [4]. Although it has been shown that CNN-based methods may achieve better performance than S-MAD methods based on hand-crafted features, the generalizability of these approaches to print-scan images tends to be limited [11].

Recently, Vision Transformer (ViT) [1] has become popular in computer vision and has achieved impressive results on existing image recognition challenges. Transformer models [25] apply the concepts of natural language processing directly to images where an image is split into small patches and then projected as a sequence of linear embeddings, which further are treated as the input to a Transformer model. By applying the self-attention mechanism and without introducing strong image-specific inductive biases as CNNs, ViT has shown the capability to integrate information globally from low layers and has achieved state-of-the-art (SOTA) performance in different tasks with large-scale training data. Consequently, many works have been investigating the possibility of applying ViT to other tasks. In the case of MAD, the traces of morphing are widely distributed among the face region, and hence the algorithms should have a large receptive field and the capacity of integrating local and global information to be robust and generalized. Hence, We assert that the advantages of ViTs can improve S-MAD and investigate further if they improve the generalizability of the developed S-MAD algorithm.

Our Contributions: 1) We propose an S-MAD algorithm based on the deep representation from a pretrained vanilla ViT against other works using CNNs. 2) We investigate the applicability of the pure self-attention-based model in S-MAD tasks by conducting comprehensive cross-dataset testing with various morph generation types and different dataset types (digital/print-scan/print-scan compression). The generalizability and detection performance of the proposed approach is quantitatively evaluated and reported 3) We benchmark the proposed method together with other state-of-the-art S-MAD algorithms based on the ensemble of hand-crafted features [27], hybrid scale-space colour texture features [16] (reported in the testing report from National Institute of Standards and Technology [8]), deep CNN features [15], steerable features [17], Multi-modality approach (tested in Bologna Online Evaluation Platform [10][1], residual AutoEncoder [12], and Multi-level Deep Features [26] respectively. The analysis result indicates an improved generalizability on digital inputs.

2. Proposed Method

An overview of our proposed S-MAD method is described in Fig. 2. We first crop the face region using MTCNN [31] to detect face regions and then resize the cropped face image into 384 x 384 pixels to fit the input of ViT model. Then,

[1] https://biolab.csr.unibo.it/fvcongoing/UI/Form/BOEP.aspx

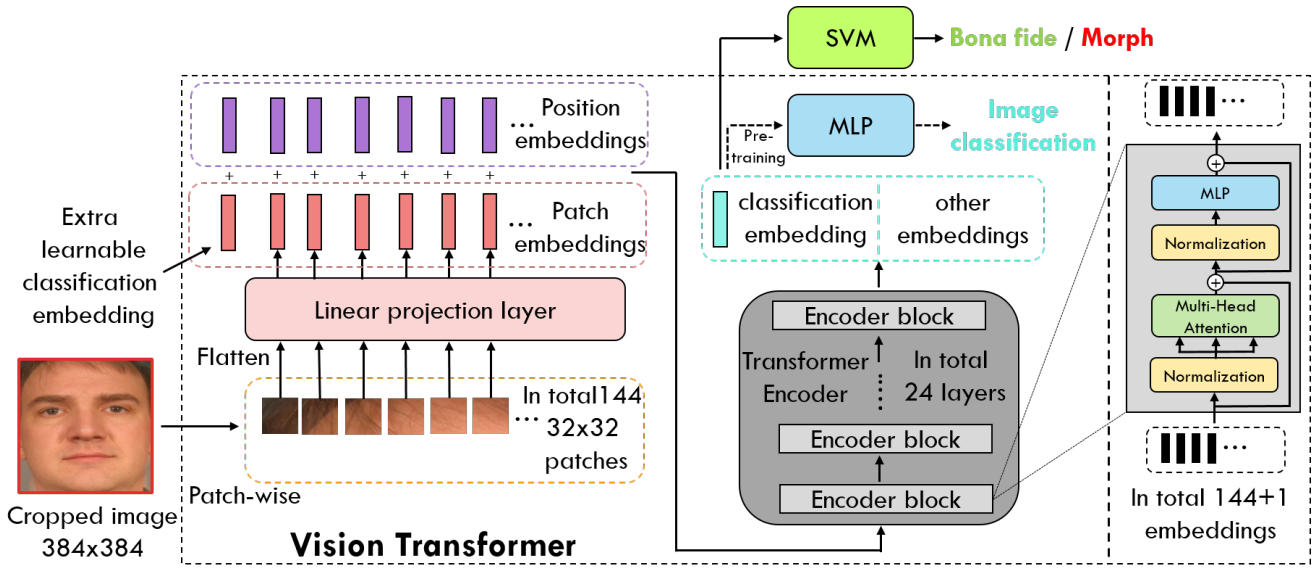

Figure 2. Overview of our proposed method using pretrained Vision Transformer model.

the input image is split into small patches x_p with the size of 32 x 32 pixels and then flattened and projected as patch embeddings through a learned linear projection layer with one layer of fully connected blocks for each embedding. Then an extra learnable classification embedding x_{class} is attached to the other patch embeddings as the learned image representation for further classification tasks. Similar to the design of the vanilla ViT, 1-D positional encoding is applied to generate position embeddings E_{pos} with the same length of the patch embeddings using sinusoidal functions. Each position embedding is added to the corresponding patch embedding hence the positional information can be encoded. Then the processed input z_0 can be noted as:

$$z_0 = [x_{class}; x_p^1 E; x_p^2 E; ...; x_p^N E] + E_{pos} \quad (1)$$

where $N = 144$ is the number of patches and E indicates the linear projection process. After processing the image into a sequence of embeddings, they are fed forward through the transformer encoder stacked with 24 layers of encoder blocks. Each encoder block includes a multi-head self-attention layer and a Multilayer Perceptron (MLP) block.

$$z'_l = MSA(LN(z_{l-1})) + z_{l-1},\ l = 1, ..., L. \quad (2)$$

$$z_l = MLP(LN(z'_l)) + z'_l,\ l = 1, ..., L. \quad (3)$$

The multi-head self-attention layer extends the key-query-value triplet into 16 sub-triplets and executes the computation of the self-attention mechanism in parallel, hence the model can learn to extract features from multiple different aspects.

During Pretraining of the ViT model, the classification token is linked to an extra MLP with a dimension of 4096 and then a softmax classifier for image classification. The model is pretrained on ImageNet21k [19] and ImageNet2012 dataset [20] with 1000 classes. To avoid duplicated training processes and achieve sustainability, we use the settings of hyper-parameters inspired by the original ViT paper [1]. As for model selection, we selected the ViT-L model with the large parameter size for higher capacity generalizability and large patch size to extract more local information. For the S-MAD task, we use the pretrained model to extract the classification tokens with the dimension of 1024 on our face morphing dataset. The extracted classification tokens will be considered as general deep representations and then we train a linear SVM classifier to solve the S-MAD problem as a binary classification task. The SVM classifier is chosen over training a deep-learning-based binary classifier due to its efficiency and robustness in preventing overfitting for small-to-medium size datasets.

3. Dataset

In order to conduct the cross-dataset testing comprehensively and simulate the operational use case, we use a database generated by various morphing algorithms and in different image processing methods (digital, print-scanned, print-scanned-compression). To simulate the passport use cases with face photos, our database is constructed based on selected morphed images from FRGC-V2 dataset [9] with high image quality and well-controlled capturing conditions

(e.g., pose variations). 140 unique subjects, including 93 male subjects and 47 female subjects are selected. For each subject, 7-24 mated samples with similar capturing conditions (image resolution, neutral expression, pose, illumination, etc.) are chosen. In total, 1270 bona fide samples are included.

As for the morphing algorithms, we selected the following five representative morphing algorithms including, two landmark-based algorithms Landmark-I [14] and Landmark-II [3], and 3 GAN-based algorithms StyleGAN-IWBF[28], MIPGAN-I and MIPGAN-II [30] to establish a diversity of unknown attacks. The samples are pre-processed to meet the ICAO 9303 requirements [5]. Pairs of parent images for the morphing process are selected following guidelines suggested in [14] [21] (e.g., isolating between different genders, pairing based on similarity score of an FRS model), as the attacker may spend as much as an effort to generate the morphing attacks in real cases. As our target is to train the model to learn patterns generated from morphing instead of general patterns from GANs, reconstructed bona fide images are applied to the datasets with GAN-based morphing algorithms. In this way, we can reduce the bias between bona fide and morph samples and can make the trained classifier generalize to other types of attacks that are not generated by the same GAN model.

To evaluate the generalizability of S-MAD algorithms on different types of images, 3 types are included in our database:

- **Digital**: Morph images are obtained from the morphing algorithms given digital parent images as input.
- **Print-scan**: Both generated morphs and bona fide images are printed using DNP-DS820 dye-sublimation photo printer and then re-digitized using the Canon office scanner with 300 dpi as suggested in ICAO 9303 requirements [5]. This is to simulate the process of a passport application.
- **Print-scan with compression**: Print-scanned images (morphs and bona fide) are compressed into less or equal to 15 KBs to simulate the images stored in the e-passport.

Overall, each dataset has 2500 morphed images and 1270 bona fide images. Given the 5 included morphing algorithms and 3 image processing types, in total 15 datasets are used in the database for further cross-dataset testing on S-MAD algorithms.

4. Experiments and Results

To evaluate the generalizability and robustness of our approach, we apply cross-dataset testing on different morphing algorithms within each dataset of different image processing types and benchmark it with the other selected SOTAs:

- Ensemble Features [27] uses ensembled features including LBP, HoG, and BSIF. The algorithm has been evaluated by public testing and included in NIST report [8]
- Hybrid Features [16] uses scaled colour space and trains independent classifiers based on the extracted LBP features.
- Deep Features [15] use pretrained VGG and AlexNet to extract transferable features and apply feature-level fusion for further classification.
- Steerable Features [17] extracts steerable pyramids from illuminance components and trains classifiers based on high-frequency components.
- Multi-Modality [13] crops the face image into different regions and extracts BSIF and LBP features. Independent classifiers are trained and score-level fusion is applied to output the final classification result. The algorithm has been evaluated in the Bologna Online Evaluation Platform [10] [2].
- Residual AutoEncoder [12] is a deep learning approach that consists of a skip-connected AutoEncoder and a ResNet18 Classifier. Guided by the designed loss functions, the model is trained to extract learnable residuals which can be used for further classification by the ResNet.
- Multi-level Deep Features [26] applies multi-level fusion on features extracted from AlexNet and ResNet50.

The selected baselines cover approaches based on handcrafted features, deep-learning-based transferable features, different fusion strategies, and trained deep-learning models.

More specifically, for each dataset generated with a specific morphing algorithm, we train the S-MAD algorithm on it and test with the datasets (generated by different morphing algorithms). This shows how the detection algorithms can generalize and to which extent they are robust with respect to unknown attacks. The performance of testing across different image processing types is not included as considering a model trained on print-scan data is often not used to detect attacks from digital data rather an ensemble is used. Instead, we report the performance of cross-dataset testing for the same image processing types (e.g., digital versus digital) to evaluate the generalizability of MAD algorithms.

To report the performance of each test, we employ standardized metrics such as Bona fide Presentation Classification Error Rate (BPCER) and Morphing Attack Classification Error Rate (MACER) following ISO/IEC CD 20059.2 [6] and measure the detection error trade-off by reporting BPCER@MACER=5% and BPCER@MACER=10%. To simplify and scalarize the results, Detection equal error rate (D-EER) is also reported. The lower D-EER numbers indicate better detection performances.

For the evaluation protocol, we evaluate both intra-dataset testing and inter-dataset testing but without crossing

[2] https://biolab.csr.unibo.it/FvcOnGoing/UI/Form/AlgResult.aspx?algId=8422

Table 1. Statistical analysis on the D-EER(%) computed for all cross-dataset testing results on FRGC morphing dataset.

S-MAD Algorithms	Digital		Print-scan		P.S. with Compression	
	μ	σ	μ	σ	μ	σ
Ensemble Features [27]	21.03	17.55	16.73	14.90	17.24	15.78
Hybrid Features [16]	26.11	20.26	19.07	16.04	16.28	14.23
Deep Features [15]	21.16	17.50	11.38	16.23	16.77	13.19
Steerable Features[17]	35.97	16.57	15.72	18.44	31.49	11.20
Multi-Modality [13]	18.05	15.51	**7.63**	**12.27**	13.57	13.72
Residual AutoEncoder [12]	15.95	17.38	16.01	15.25	14.92	13.97
Multi-level Deep Feature [26]	14.78	13.90	9.51	12.44	**13.38**	**12.36**
Proposed Method	**13.63**	**11.61**	18.33	14.32	19.09	13.61

image types (digital, print-scanned, and print-scanned and compressed). Detailed quantitative analysis is included in the supplementary material. To measure the overall generalizability of the MAD algorithms and establish the significance of the obtained results, we propose to conduct statistical analysis on the D-EER of the cross-dataset testing cases within each type of image, and also visualize this analysis as a boxplot. From the quantitative analysis in Tab. 1, it is shown that our approach has the lowest mean and standard deviation of D-EER for digital images. The mean value of D-EERs from the proposed method has decreased 1.15% compared to the best among the baselines. As visualized in Fig. 3, a similar observation can also be noticed by the similarly low median value as Residual AutoEncoder and Multi-level Deep Features. However, the range of error rate from our approach during testing is more narrowed, which indicates better robustness. In print-scan and print-scan with compression cases, a degradation of the detection performance of our algorithm can be noticed compared to the digital case. We reason this by 1) the Vision Transformer model is pretrained only with digital images and hence the extracted representation is less effective when transferred to another image processing type 2) compared to the digital images, print-scan and print-scan compression images are in a much lower resolution and can provide less information for the Vision Transformer model (which takes the input size of 384 x 384). For print-scan inputs, the Multi-modality approach and multi-level Deep Features approach achieved the best performances. As for further compressed print-scanned images, the multi-modality approach and Residual AutoEncoder approach are similar.

It is also shown that different algorithms perform inconsistently for the same testing cases. Ensemble Features, Hybrid Features, Deep Features, and Steerable Features are not generalizing well for most of the cross-testing cases, even for inter-testing among GAN-based morphed images. Our approach has shown considerable generalization when the test is crossing between landmark-based and GAN-based morphs in digital cases. The Residual AutoEncoder approach has also shown similar performance, while some extreme cases can be noticed for example in when the model is trained by MIPGAN-I dataset and tested on Landmark-II dataset. This might be caused by the randomness during the training process of the network. The multi-Modal approach in general has impressive performances on Print-scan and print-scan compression images. Compared with Deep Features approach with transferable CNN features, it is shown that the ViT-based features can achieve improvement in the generalizability of MAD for digital images. Meanwhile, the multi-level fusion of CNN features has shown an overall improvement in the Deep Features approach. It also achieves comparable results with ViT-based features in the digital case with considerable generalizability for print-scan and print-scan compression images.

Additionally, the interpretation of the proposed method and obtained results is studied. As shown in Fig. 4 - Fig. 6, T-SNE [24] plot is used to visualize the feature space of the proposed method with data using different processing processes.

Similarly as what we've observed in the cross-dataset testing results, features from morphs generated by StyleGAN-IWBF, MIPGAN-I and MIPGAN-II can be well-separated between the features from bona fide images, which indicates good generalization on cross-dataset test-

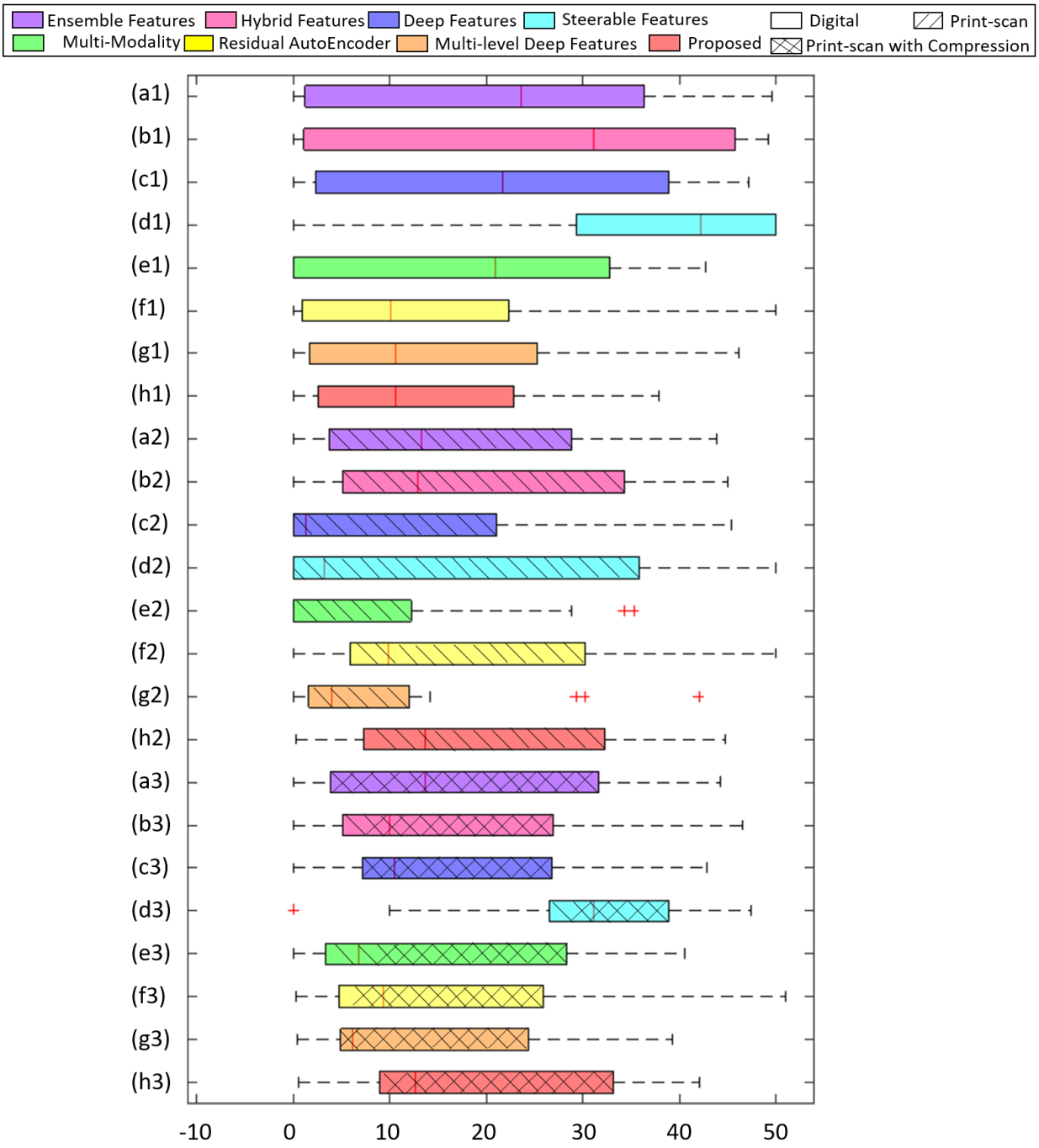

Figure 3. Boxplot of the statistical analysis on D-EER computed for all cross-dataset testing results on FRGC morph database. (a) Ensemble Features (b) Hybrid Features (c) Deep Features (d) Steerable Features(e) Multi-Modality (f) Residual AutoEncoder (g) Proposed Method. (1): Digital (2): Print-scan (3): Print-scan Compression

ing. Features of Landmark-I morphs are shown to be less seperable than GAN-based morphs. The overlap between features of morphed samples generated by Landmark-II method and features of bona fide samples also follows detection accuracy where this method is the most challenging one to classify and generalize. This shows that the post-processing on morphs can effectively make the generated attacks stronger. Meanwhile, when the processing type is changed from digital (Fig. 4) to print-scan (Fig. 5) and then to print-scan compression (Fig. 6), it becomes more difficult

to find sharable boundaries for classifying different types of morphs between bona fide samples.

5. Limitations

In this work we have applied a specific pretrained ViT model in order to be sustainable on computational powers, however, the influence of different hyper-parameters in the ViT model on the final performance of S-MAD tasks is still worth to be studied. Meanwhile, for the cross-dataset testing, we only conducted experiments with leave-one-out training. It is also interesting to evaluate S-MAD trained on a dataset mixed with multiple morphing algorithms and study on the learning capacity. By using the pretrained ViT model, our model has shown an improvement in generalizability for digital images, while it can be noticed that for intra-dataset testing the detection accuracies are overall less or equal for the other algorithms. Also as shown in our evaluation, the different algorithm performs inconsistently. Hence, it is reasonable to further explore fusion strategies or combine them with the multi-modality approach.

6. Conclusion

In this work, we proposed an S-MAD algorithm based on pretrained Vision Transformer model instead of existing deep-learning-based methods using CNNs. Motivated by the real application scenario of open-set testing, we use a morph dataset with three different image processing types and five different representative morphing algorithms, including both GAN-based and landmark-based algorithms for the cross-dataset testing. The proposed method is benchmarked against two selected SOTA algorithms. Based on the statistical analysis of the obtained results, it can be concluded that the proposed method based on the pure self-attention model can achieve notable improvement in the generalizability of the digital use cases. Despite the low performance for some cases in print-scan and print-scan compression images as noticed, one can note overall detection accuracy gain in cross-dataset testing while remaining comparable with the other SOTA algorithms.

To conduct a comprehensive evaluation of the detection performance of the proposed method, we have benchmarked several existing SOTA targeting generalized S-MAD tasks. Besides constructing a representative morph database we are trying to simulate the operational application scenario, further we will submit the algorithms to third-party tests such as NIST (National Institute of Standards and Technology) Face Analysis Technology Evaluation (FATE) [8] or Bologna Online Evaluation Platform (BOEP) [10]. In this work, the proposed method has not been submitted, but the selected reference algorithm based on hybrid features [16] has been tested in FRVT MORPH and the performance is reported in [8], and the Motimodality-based algorithm [13] has been evaluated in BOEP.

Meanwhile, it should also be noted that in this work we only applied the vanilla Vision Transformer model pretrained with digital images on the image classification task. Hence it remains future works to plug in improved Vision Transformer models or replace the pretraining strategy with MAD-related tasks on different types of images.

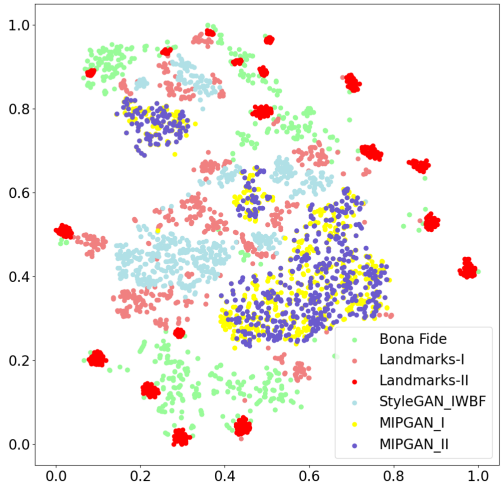

Figure 4. T-SNE plot of the feature space used in proposed method with digital images

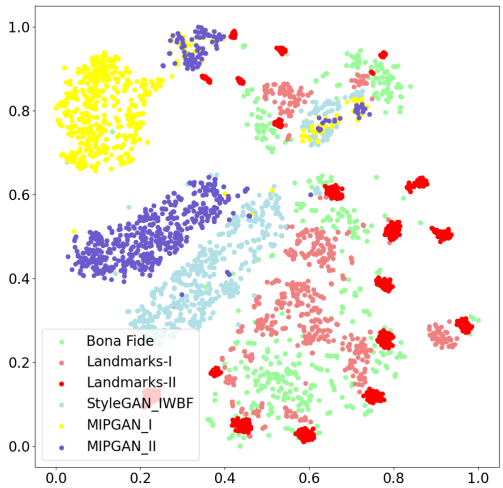

Figure 5. T-SNE plot of the feature space used in proposed method with print-scanned images

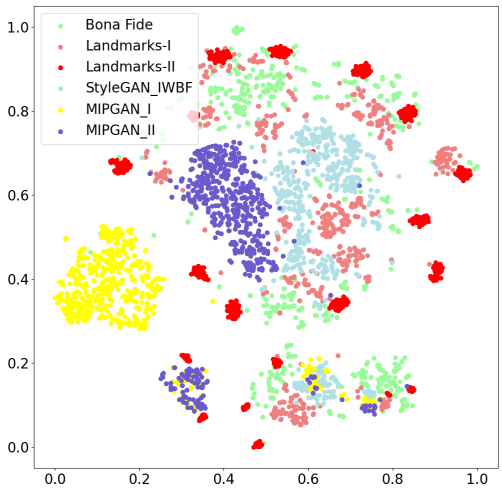

Figure 6. T-SNE plot of the feature space used in proposed method with print-scanned and compressed images

References

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 3

[2] Matteo Ferrara, Annalisa Franco, and Davide Maltoni. The magic passport. In *IEEE International Joint Conference on Biometrics*, pages 1–7. IEEE, 2014. 1

[3] Matteo Ferrara, Annalisa Franco, and Davide Maltoni. Decoupling texture blending and shape warping in face morphing. In *2019 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–5. IEEE, 2019. 4

[4] Matteo Ferrara, Annalisa Franco, and Davide Maltoni. Face morphing detection in the presence of printing/scanning and heterogeneous image sources. *arXiv preprint arXiv:1901.08811*, 2019. 2

[5] International Civil Aviation Organization. Machine readable passports – part 9 – deployment of biometric identification and electronic storage of data in eMRTDs, 2021. 4

[6] ISO/IEC JTC1 SC37 Biometrics. *ISO/IEC CD 20059.2 Methodologies to evaluate the resistance of biometric recognition systems to morphing attacks*. International Organization for Standardization, 2023. 4

[7] Anil K Jain and Stan Z Li. *Handbook of face recognition*. Springer, 2011. 1

[8] Mei Ngan, Patrick Grother, Kayee Hanaoka, and Jason Kuo. *Face Analysis Technology Evaluation (FATE) Part 4: MORPH - Performance of Automated Face Morph Detection: Morph-performance of automated face morph detection*. US Department of Commerce, National Institute of Standards and Technology, 2024. 2, 4, 7

[9] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, Jin Chang, K. Hoffman, J. Marques, Jaesik Min, and W. Worek. Overview of the face recognition grand challenge. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, pages 947–954 vol. 1, 2005. 3

[10] Kiran Raja, Matteo Ferrara, Annalisa Franco, Luuk Spreeuwers, Ilias Batskos, Florens de Wit, Marta Gomez-Barrero, Ulrich Scherhag, Daniel Fischer, Sushma Krupa Venkatesh, et al. Morphing attack detection-database, evaluation platform, and benchmarking. *IEEE transactions on information forensics and security*, 16:4336–4351, 2020. 2, 4, 7

[11] Kiran Raja, Matteo Ferrara, Annalisa Franco, Luuk Spreeuwers, Ilias Batskos, Florens de Wit, Marta Gomez-Barrero, Ulrich Scherhag, Daniel Fischer, Sushma Krupa Venkatesh, Jag Mohan Singh, Guoqiang Li, Loïc Bergeron, Sergey Isadskiy, Raghavendra Ramachandra, Christian Rathgeb, Dinusha Frings, Uwe Seidel, Fons Knopjes, Raymond Veldhuis, Davide Maltoni, and Christoph Busch. Morphing attack detection-database, evaluation platform, and benchmarking. *IEEE Transactions on Information Forensics and Security*, 16:4336–4351, 2021. 1, 2

[12] Kiran Raja, Gourav Gupta, Sushma Venkatesh, Raghavendra Ramachandra, and Christoph Busch. Towards generalized morphing attack detection by learning residuals. *Image and Vision Computing*, 126:104535, 2022. 2, 4, 5

[13] Raghavendra Ramachandra and Guoqiang Li. Multimodality for reliable single image based face morphing attack detection. *IEEE Access*, 10:82418–82433, 2022. 4, 5, 7

[14] Raghavendra Ramachandra, Kiran B Raja, Sushma Venkatesh, and Christoph Busch. Face morphing versus face averaging: Vulnerability and detection. In *IEEE International Joint Conference on Biometrics (IJCB)*, pages 555–563, 2017. 4

[15] Raghavendra Ramachandra, Kiran B. Raja, Sushma Venkatesh, and Christoph Busch. Transferable deep-cnn features for detecting digital and print-scanned morphed face images. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1822–1830, 2017. 2, 4, 5

[16] Raghavendra Ramachandra, Sushma Venkatesh, Kiran Raja, and Christoph Busch. Towards making morphing attack detection robust using hybrid scale-space colour texture features. In *2019 IEEE 5th International Conference on Identity, Security, and Behavior Analysis (ISBA)*, pages 1–8, 2019. 2, 4, 5, 7

[17] Raghavendra Ramachandra, Sushma Venkatesh, Kiran Raja, and Christoph Busch. Detecting face morphing attacks with collaborative representation of steerable features. In *Proceedings of 3rd International Conference on Computer Vision and Image Processing: CVIP 2018, Volume 1*, pages 255–265. Springer, 2019. 2, 4, 5

[18] Raghavendra Ramachandra, Sushma Venkatesh, Kiran Raja, and Christoph Busch. Detecting face morphing attacks with

collaborative representation of steerable features. In *Proceedings of 3rd International Conference on Computer Vision and Image Processing*, pages 255–265. Springer, 2020. 2

[19] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021. 3

[20] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 3

[21] Ulrich Scherhag, Andreas Nautsch, Christian Rathgeb, Marta Gomez-Barrero, Raymond NJ Veldhuis, Luuk Spreeuwers, Maikel Schils, Davide Maltoni, Patrick Grother, Sebastien Marcel, Breithaupt Ralph, Raghavendra Ramachandra, and Christoph Busch. Biometric systems under morphing attacks: Assessment of morphing techniques and vulnerability reporting. In *2017 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–7. IEEE, 2017. 1, 4

[22] Ulrich Scherhag, Christian Rathgeb, and Christoph 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. 2

[23] Clemens Seibold, Anna Hilsmann, and Peter Eisert. Style your face morph and improve your face morphing attack detector. In *2019 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–6. IEEE, 2019. 2

[24] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 5

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[26] Sushma Venkatesh. Multilevel fusion of deep features for reliable single image based face morphing attack detection. In *2022 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*. IEEE, 2022. 2, 4, 5

[27] Sushma Venkatesh, Raghavendra Ramachandra, Kiran Raja, and Christoph Busch. Single image face morphing attack detection using ensemble of features. In *23rd International Conference on Information Fusion*, pages 1–5, 2020. 2, 4, 5

[28] Sushma Venkatesh, Haoyu Zhang, Raghavendra Ramachandra, Kiran Raja, Naser Damer, and Christoph Busch. Can gan generated morphs threaten face recognition systems equally as landmark based morphs? - vulnerability and detection. In *2020 International Workshop on Biometrics and Forensics (IWBF)*, pages 1–6. IEEE, 2020. 4

[29] Sushma Venkatesh, Raghavendra Ramachandra, Kiran Raja, and Christoph Busch. Face morphing attack generation and detection: A comprehensive survey. *IEEE Transactions on Technology and Society*, 2(3):128–145, 2021. 1, 2

[30] Haoyu Zhang, Sushma Venkatesh, Raghavendra Ramachandra, Kiran Raja, Naser Damer, and Christoph Busch. Mipgan—generating strong and high quality morphing attacks using identity prior driven gan. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(3):365–383, 2021. 4

[31] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23 (10):1499–1503, 2016. 2