http://dx.doi.org/10.14500/aro.12990

EfficientNet-Lite: Parameter-Efficient CNNs for Face Morph Attack Detection

Sheerko R. HmaSalah[1†], Abdulbasit K. Al-Talabani[1,2] and Aras T. Asaad[3,4]

[1]Department of Software Engineering, Faculty of Engineering, Koya University, Danielle Mitterrand Boulevard, Koya KOY45, Kurdistan Region – F.R. Iraq

[2]Department of Computer Science and Engineering, School of Science and Engineering, University of Kurdistan Hewlêr, Erbil, Iraq

[3]Department of Mathematics and Natural Sciences, American University of Iraq, Sulaimani, Kurdistan Region, Iraq

[4]School of Computing, University of Buckingham, Buckingham, UK

***Abstract*—Morphing attacks on facial images pose a serious threat to face recognition systems during identity verification. In this paper, we introduce a parameter-efficient family of convolutional neural networks, named EfficientNet-Lite (EN_Lite), for morph attack detection (MAD). The experimental framework uses eight datasets, four morph generation methods, and ten homogeneous and heterogeneous evaluation scenarios under a leave-one-dataset-out protocol. Performance was assessed using dataset-size-weighted average equal error rate (EER_{WA}), computed across the held-out datasets within each scenario. EN_Lite was evaluated using ImageNet-1K and VGGFace2-HQ pre-training, with the latter comprising 1,107,817 successfully processed facial images. The ImageNet-pretrained IEN_Lite-1 achieved a mean EER_{WA} of 2.99%, compared with 2.65% for EfficientNet-B3 while using approximately 76% fewer parameters. It also achieved a lower mean EER_{WA} than GhostNet-1.0x (4.01%), MobileNetV3-Large (5.03%), MobileNetV3-Small (6.03%), and ShuffleNet V2 (5.31%). The VGGFace2-HQ-pretrained variants produced higher error rates than their corresponding ImageNet-pretrained variants under the evaluated protocol. Performance remained strong in homogeneous scenarios but degraded under heterogeneous cross-morph evaluation. Overall, EN_Lite provides a favorable efficiency–performance trade-off for MAD, while high-security deployment should be assessed at application-specific operating points.**

***Index Terms*—Biometric security, Convolutional neural networks, EfficientNet, Morph attack detection, Transfer learning.**

I. Introduction

Facial recognition has become an important biometric technology for identification and authentication because of the unique and distinctive features of the human face. Yet the rapid development of deepfake algorithms and easy access to image editing tools pose serious security challenges. A morphing attack on facial biometrics was first identified in 2014 by Ferrara, Franco and Maltoni (2014) as a serious threat that can fool face recognition systems (FRS) deployed for the purpose of face verification and recognition in border controls and identity (ID) document issuance departments. Morphing is a process of combining facial features of two or more individuals to produce a visually realistic facial image that can bypass both automated FRSs as well as human inspectors (Ferrara, Franco and Maltoni, 2014; Scherhag, et al., 2019). In this work, we consider the single-image morph attack detection (S-MAD) scenario, where the detector classifies a single facial image as bona fide or morphed without a trusted reference image.

Despite significant progress by the forensics and multimedia research community, several fundamental challenges remain unsolved in the development of efficient and generalizable MAD algorithms. The field still lacks a comprehensive evaluation of deep convolutional neural network (CNN) architectures designed specifically for MAD, especially in terms of how model complexity affects detection performance. Existing studies in the literature mainly focus on individual architectures developed for other image analysis tasks or use a small number of datasets or rely on landmark-based morphing algorithms or deep learning (DL) based morphing strategies (Raja, et al., 2020). In the same vein, many studies provide conclusions based on a small number of datasets, which makes reproducing the same findings challenging (Jassim and Asaad, 2018; Asaad and Jassim, 2017). Furthermore, the computational requirements of state-of-the-art CNN models present significant challenges for real-world biometric systems where efficiency and speed are crucial or necessary when the deployment is aimed at mobile or embedded devices.

This paper addresses two key questions: (1) What efficiency–performance trade-off can compact CNNs

ARO-The Scientific Journal of Koya University
Vol. XIV, No.2 (2026), Article ID: ARO.12990. 9 pages
DOI: 10.14500/aro.12990
Received: 25 April 2026; Accepted: 24 September 2026
Regular research paper; Published: 04 November 2026
†Corresponding author's e-mail: sherko.rahman@koyauniversity.org
Copyright © 2026 Sheerko R. HmaSalah, Abdulbasit K. Al-Talabani and Aras T. Asaad. This is an open access article distributed under the Creative Commons Attribution License (CC BY-NC-SA 4.0).

provide for MAD compared with conventional and modern lightweight architectures? (2) How do ImageNet-1K and face-specific VGGFace2-HQ pre-training compare under the evaluated MAD protocol?

The main contributions of this research are: (1) A parameter-efficient EN_Lite family for MAD; (2) direct comparison with conventional and modern lightweight CNNs, including ShuffleNet V2, GhostNet-1.0x, and MobileNetV3; and (3) systematic cross-morph and cross-dataset evaluation under ImageNet-1K and large-scale VGGFace2-HQ pre-training.

II. Related Work

MAD has been studied using both landmark-based and DL-based morph generation approaches. Landmark-based methods generally rely on facial landmark alignment and blending operations to create morphed faces, as reflected in tools and methods, such as FaceMorpher, OpenCV-based, and AMSL pipelines (Quek, 2019; Mallick, 2016; Makrushin, Neubert and Dittmann, 2017). DL-based methods, by contrast, use generative models for face synthesis and morph generation. In particular, MIPGAN-I/II extended earlier StyleGAN-based morph generation by strengthening ID preservation through ID-prior optimization, while ReGenMorph used a fine-tuned StyleGAN encoder-generator pipeline to regenerate landmark-based morphs with improved visible realism (Karras, Laine and Aila, 2019; Zhang, et al., 2021; Damer, et al., 2021). These developments show that MAD systems should be evaluated under diverse morph generation conditions rather than relying on a single attack type.

For S-MAD, a wide range of CNN-based architectures has been explored. Previous studies have examined lightweight and conventional CNN backbones for related biometric security tasks, such as presentation attack detection and face anti-spoofing (Tapia and Busch, 2023; Aghdaie, et al., 2021), while MAD-specific research has considered different architectures and settings, including cross-dataset evaluation and single-image MAD protocols (Tapia and Busch, 2023; Colbois and Marcel, 2022; Huber, et al., 2022). Other studies have investigated ResNet-based methods, de-morphing approaches, hybrid frameworks, DenseNet, inception-based models, VGG variants, AlexNet, and more recently transformer-based or attention-based models (Venkatesh, et al., 2020; Caldeira, et al., 2023; Ibsen, et al., 2024; Seibold, et al., 2017; Kashiani, et al., 2022; Zhang, et al., 2024; Pimenta, Medvedev and Gonçalves, 2023). Survey papers also provide broader overviews of both DL and handcrafted MAD methods (Scherhag, et al., 2019; Venkatesh, et al., 2021; Hamza, et al., 2022). Zero-shot learning utilizing large language models, such as ChatGPT (GPT4-turbo), is explored as a new S-MAD algorithm together with an open-source vision model (Zhang, et al., 2026). Bona fide and morphed images are synthetically generated to comply with privacy restrictions, which is a major limitation of present LLM-based S-MAD approach (Zhang, et al., 2026). Inspired by differences in the variation of similarity between morph and bona fide faces, Hoang, Islam and Ahn (2025) proposed a DL model to de-morph live captured faces utilizing StyleGAN and the pixel-to-style-to-pixel encoder. Tapia, Russo and Busch (2025) proposed a new DL algorithm using transfer of print-scanned texture and then utilizing it for MAD detection to improve generalizability. Siamese networks and triplet semi-hard loss used a different MAD detection method to address the problem of cross-dataset and unknown morphing scenarios (Tapia, Schulz and Busch, 2025). Modern lightweight CNNs, such as MobileNetV3 (Howard, et al., 2019) and GhostNet (Han, et al., 2020) provide relevant efficiency-oriented baselines for assessing compact MAD architectures.

Despite these advances, the literature still provides limited systematic analysis of the trade-off between model complexity and detection performance in MAD. Many studies focus on individual architectures, specific morphing methods, or a small number of datasets, which makes general conclusions more difficult to establish (Tapia and Busch, 2023; Colbois and Marcel, 2022; Huber, et al., 2022; Venkatesh, et al., 2020). In addition, although transfer learning is central to deep MAD systems, it remains unclear whether pre-training on a general-purpose dataset, such as ImageNet or on a face-specific dataset is more effective for this task. These gaps form the motivation for the present study. Specifically, this work systematically evaluates parameter-efficient CNN models using eight datasets, four morph generation methods, several homogeneous and heterogeneous experimental scenarios, and two different pre-training strategies.

III. Dataset Construction and Morph Generation

A. Dataset Collection and Preparation

Eight publicly and privately available research datasets were used to provide demographic and acquisition diversity for the MAD evaluation. The selected datasets contain individuals from White, Black, Asian, Multiracial American, Latino, Hispanic, Kurdish, and East Asian backgrounds. Although KurdFace-1000 (Ibrahem and Abdelazeez, 2026) is an important Kurdish facial dataset designed for facial beauty prediction and multi-attribute learning. It was not included in the present study. The reason is that this work focuses specifically on morph attack detection, which requires source images suitable for morph generation and subsequent MAD evaluation rather than images annotated mainly for beauty-related attributes. To maintain consistency across the morph generation procedures, only images with neutral facial expression, frontal pose, clear facial visibility, no occlusion, no glasses, and minimal illumination variation were selected. Overall, the final dataset contained 2,625 bona fide facial images; for both bona fide and morphed images, a landmark-based facial convex-hull mask was applied, pixels outside the mask were replaced with a constant white background, and the resulting face crops were used for subsequent model-specific resizing. A detailed summary of the dataset characteristics is provided in Table I.

 http://dx.doi.org/10.14500/aro.12990

TABLE I
DATASET CHARACTERISTICS BY SOURCE

Source	Bona fide			Details		Age	Image size
	Male	Female	Total	Race	#		
CFD (Ma, Correll and Wittenbrink, 2015), CFD-INDEA (Lakshmi and Wittenbrink, 2021), CFD-MR (Ma, Kantner and Wittenbrink, 2021)	406	421	827	Asian American	109	17–65	1718×2444 JPEG
				Indian Asian	142		
				Multiracial American	88		
				Latino	108		
				Black	197		
				White	183		
FRLL (DeBruine and Jones, 2017)	53	49	102	White	69	18–54	1350×1350 JPEG
				East Asian	9		
				Black	13		
				East Asian White	1		
				West Asian	10		
FERET (Phillips, *et al.*, 1998)	350	180	530	White	282	-	512×768 BMP
				East Asian	58		
				Black	52		
				-	138		
Kurd (Hussein and Rashid, 2022)	110	60	170	West Asian	170	18–45	720×480 JPEG
Utrecht (Hancock, 2008)	47	20	67	-		-	900×1200 JPG
DEFACTO (Mahfoudi, *et al.*, 2019)	142	58	200	-		-	652×500 JPEG
FEI (Thomaz and Giraldi, 2010)	88	89	177	-		19–40	640×480 JPG
FRGC (Phillips, *et al.*, 2005)	318	234	552	White	378	18–70	1704×2272 1200×1600 JPEG
				Asian	112		
				Hispanic	15		
				Asian-Southern	13		
				Black-or-African-American	10		
				Asian-middle-Eastern	1		
				-	23		
Total	1514	1111	2625				

(M: Male, F: Female, #: Number of individuals)

To determine an appropriate morph-generation protocol, cosine similarity was computed using ArcFace embeddings (Deng, et al., 2019). The analysis showed that highly similar image pairs were more commonly observed within the same dataset than across different datasets, indicating greater visual consistency within each source. Accordingly, morphs were generated from intra-dataset image pairs, while cross-dataset evaluation was reserved for testing generalization to previously unseen conditions.

B. Morph Generation and Selection Strategy

Four morph generation methods were used in this study: Three landmark-based approaches, namely, OpenCV, FaceMorpher, and AMSL, and one GAN-based approach based on StyleGAN2-ADA, referred to throughout the paper as StyleGAN for simplicity and consistency with prior MAD literature. OpenCV and FaceMorpher were implemented following Mallick (2016) and Quek (2019), respectively, whereas AMSL was re-implemented following Makrushin, Neubert and Dittmann (2017). Additional methods, such as WebMorpher, MIPGAN-I, and MIPGAN-II were excluded because of implementation and compatibility constraints.

For landmark-based morphs, same-gender image pairs were generated within each dataset. For the GAN-based method, only the top 5% most similar genuine pairs within each dataset were used because exhaustive GAN-based generation was computationally impractical. To focus the evaluation on more challenging landmark-based cases, ArcFace was used for morph selection; a preliminary pair-selection analysis showed that retained same-gender landmark-morph candidates decreased from 355,309 at a 0.5 cutoff to 53,516 at 0.6 and 1,359 at 0.7, so 0.6 was used as a fixed dataset-construction cutoff rather than an optimized model threshold.

The final experimental setup included 2,625 bona fide images and 71,277 morphed images generated using OpenCV, FaceMorpher, StyleGAN, and AMSL. The detailed sample distribution across datasets and morphing methods is provided in Table II.

IV. METHODOLOGY

A. CNN Architecture Selection

For comparison, eight well-established CNN architectures were used as benchmark models: ShuffleNet V2 (Ma, et al., 2018), EfficientNet-B0 (Tan and Le, 2019), DenseNet-201 (Huang, et al., 2017), Inception-v3 (Szegedy, et al., 2016), ResNet-50 (He, et al., 2016), MobileNetV3-Small and MobileNetV3-Large (Howard, et al., 2019), and GhostNet-1.0x (Han, et al., 2020). These architectures span conventional and modern lightweight designs and provide direct reference points for assessing the EN_Lite efficiency–performance trade-off.

B. Training and Evaluation Protocol

To investigate the influence of morph generation conditions on detection performance, ten experimental scenarios were designed. These scenarios covered both homogeneous and heterogeneous settings, using OpenCV, FaceMorpher, AMSL, and StyleGAN morphs for training and held-out evaluation. The full scenario configuration is presented in Table III. Across the eight datasets, a leave-one-dataset-out protocol was adopted. For each scenario, one complete dataset was held out and used for epoch-level performance monitoring and evaluation, while the remaining seven datasets were used for model training. This procedure was repeated 8 times so that each dataset served once as the held-out evaluation set.

Training was performed using CrossEntropyLoss and the AdamW optimizer with a learning rate of 2×10^{-3}. During training, the held-out evaluation loss and performance metrics were monitored after each epoch. To reduce overfitting, early stopping was applied when the held-out loss showed no improvement for 10 consecutive epochs. Evaluation metrics were recorded at the epoch level, and the final reported results were obtained from the saved performance records. Unless required by model input constraints, the training setup was kept consistent across experiments, using a batch size of 32 on an NVIDIA GeForce RTX 4090 GPU with a fixed random seed for reproducibility.

Performance was evaluated using standard metrics from ISO/IEC (2023), namely, attack presentation classification error rate (APCER), bona fide presentation classification error rate (BPCER), and equal error rate (EER). APCER measures the proportion of attack presentations misclassified as bona fide, whereas BPCER measures the proportion of bona fide presentations misclassified as attacks. EER was used as the main numerical summary metric, while DET curves were used to visualize the BPCER behavior across APCER operating points.

Since the eight datasets and their corresponding morph types are imbalanced in size, the dataset-size-weighted EER was computed within each evaluation scenario. For model m and scenario s, the weighted average is defined as:

$$EER_{WA}^{m,s} = \left(\sum_{(d \in D_s)} EER_{(m,s,d)} w_{(d,s)} \right) / \left(\sum_{(d \in D_s)} w_{(d,s)} \right) \quad (1)$$

Where D_s is the set of datasets evaluated in scenario s, $EERm,s,d$ is the EER obtained by model m on dataset d under scenario s, and $w_{d,s}$ is the dataset-specific evaluation weight. The weight is defined as:

$$w_{(d,s)} = n_d^{(BF)} + \sum_{\left(t \in T_s^{(eval)}\right)} n_{(d,t)}^{(PA)} \quad (2)$$

Where n_d^{BF} is the number of bona fide samples in dataset d, $n_{d,t}^{PA}$ is the number of morphed samples of type t in dataset d, and T_s^{eval} is the set of morph types evaluated in scenario s, as specified in Table III. Therefore, each held-out dataset contributes in proportion to the number of bona fide and morphed samples actually evaluated in that scenario. The final mean EER_{WA} reported in Tables IV-VII is obtained by averaging $EER_{WA}^{(m,s)}$ across the ten evaluation scenarios.

To assess the reliability of the reported differences, EER_{WA} is reported as mean ± standard deviation (SD) across the ten scenarios. Exact two-sided paired Wilcoxon signed-rank tests were applied to scenario-wise EER_{WA} values when comparing selected models or pre-training strategies.

C. EN_Lite Design

To investigate whether compact models can maintain competitive MAD performance, a family of EfficientNet-Lite models, denoted EN_Lite, was developed by applying negative compound coefficients $\phi \in \{-1, -2, -3, -4, -5\}$ to

TABLE II
Distribution of Genuine and Morphed Images Across Datasets and Methods

#	Dataset	OpenCV	FaceMorpher	StyleGAN	AMSL	#Morphs	#Genuine
1	Chicago	18195	12880	8531	4398	44004	827
2	DEFACTO	317	188	541	29	1075	200
3	FRLL	261	152	127	55	595	102
4	KurdFace	1735	1456	388	1194	4773	170
5	Utrecht	59	55	63	23	200	67
6	FERET	811	543	3841	692	5887	530
7	FEI	1150	930	387	408	2875	177
8	FRGC	2632	3225	3883	2128	11868	552

TABLE III
Training and Held-out Evaluation Scenarios

#	Training	Held-out evaluation
S1	OpenCV	OpenCV
S2	FaceMorpher	FaceMorpher
S3	AMSL	AMSL
S4	StyleGAN	StyleGAN
S5	FaceMorpher+AMSL+StyleGAN+OpenCV	FaceMorpher+AMSL+StyleGAN+OpenCV
S6	StyleGAN	OpenCV+AMSL+FaceMorpher
S7	OpenCV+AMSL+FaceMorpher	StyleGAN
S8	OpenCV+AMSL+StyleGAN	FaceMorpher
S9	FaceMorpher+AMSL+StyleGAN	OpenCV
S10	FaceMorpher+OpenCV+StyleGAN	AMSL

(# Denotes Scenario ID)

TABLE IV
EfficientNet and EN_Lite Variant Specifications: EN_Lite (−5 to−1) and Standard (B0 to B4) Models with Compound Scaling Parameters for Binary Classification

Variant	α^{ϕ}	β^{ϕ}	γ^{ϕ}	Input	Dropout	Params (M)
EN_Lite-5	0.4	0.62	0.5	104×104	0	0.636
EN_Lite-4	0.48	0.68	0.57	128×128	0	0.753
EN_Lite-3	0.58	0.75	0.66	144×144	0.05	1.269
EN_Lite-2	0.69	0.83	0.76	168×168	0.1	1.741
EN_Lite-1	0.83	0.91	0.87	192×192	0.15	2.586
EfficientNet-B0	1	1	1	224×224	0.2	4.01
EfficientNet-B1	1.2	1.1	1.15	240×240	0.2	6.516
EfficientNet-B2	1.44	1.21	1.32	288×288	0.3	7.704
EfficientNet-B3	1.73	1.33	1.52	300×300	0.3	10.699
EfficientNet-B4	2.07	1.46	1.75	380×380	0.4	17.552

the EfficientNet scaling formulation with $\alpha = 1.2$, $\beta = 1.1$, and $\gamma = 1.15$.

EN_Lite retains the original EfficientNet architectural design, including the MBConv block structure, depthwise separable convolutions, squeeze-and-excitation components, activation functions, and stage-wise organization. No new convolutional block, attention module, or connection pattern was introduced; the modifications are limited to scaling down the network depth, width, and input resolution, together with the binary classification head and the variant-specific dropout values. The five resulting models, denoted EN_Lite-1 through EN_Lite-5, contain far fewer parameters than the standard EfficientNet family, as summarized in Table VIII.

D. Transfer Learning Strategies

Since no pre-trained weights exist for our EN_Lite architectures, all variants were implemented using two distinct pre-training strategies: (1) ImageNet-1K (Deng, et al., 2009) pre-training for 350 epochs using parameter settings nearly identical to those of the official EfficientNet backbone to ensure direct comparability; and (2) face-specific VGGFace2 (Cao, et al., 2018) pre-training using 1,107,817 successfully processed VGGFace2-HQ facial images with face-specific adaptations. The processed VGGFace2-HQ set substantially narrows the pre-training-scale difference between the two strategies.

V. Experimental Results

A. Standard CNN Architecture Performance

The performance of the eight baseline CNN architectures across the ten experimental scenarios is shown in Table IV. EfficientNet-B0 achieved the lowest mean EER_{WA} (3.47%), followed by GhostNet-1.0x (4.01%), MobileNetV3-Large

TABLE V
Performance of Adopted CNN Architectures

Model	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Mean±SD
DenseNet-201	5.24	3.58	0.15	0.40	4.75	26.55	6.98	7.42	7.14	3.74	6.60±7.47
EfficientNet-B0	1.22	1.30	0.08	0.17	1.68	20.67	5.67	1.31	1.93	0.68	**3.47**±6.24
Inception-v3	4.44	4.07	0.18	0.55	7.99	33.27	14.51	2.58	3.82	3.79	7.52±9.93
ResNet-50	3.17	3.51	0.24	1.06	3.92	37.85	9.19	2.54	3.57	2.21	6.73±11.19
ShuffleNet V2	2.56	2.42	0.18	0.20	3.73	19.20	13.98	3.01	3.84	4.01	5.31±6.22
GhostNet-1.0x	2.36	2.02	0.09	0.29	2.77	19.18	6.88	1.78	3.72	0.96	4.01±5.68
MobileNet V3-Small	2.96	2.72	0.48	0.83	5.07	28.05	8.88	3.26	5.71	2.37	6.03±8.12
MobileNet V3-Large	6.21	1.29	0.19	0.40	5.68	17.00	9.83	2.00	5.22	2.47	5.03±**5.19**

CNN: Convolutional neural networks, SD: Standard deviation. Bold values indicate the lowest mean EERWA and the lowest SD among the evaluated models.

TABLE VI
Performance of EfficientNet Variants

Model	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Mean±SD
EfficientNet-B1	0.72	0.79	0.09	0.04	1.37	17.20	4.08	1.05	1.40	1.10	2.78±5.19
EfficientNet-B2	1.65	1.12	0.07	0.20	2.21	22.19	8.87	1.21	2.26	0.73	4.05±6.86
EfficientNet-B3	1.31	1.55	0.00	0.04	1.42	12.84	6.32	1.09	1.52	0.37	**2.65±4.01**
EfficientNet-B4	0.73	0.74	0.02	0.16	1.88	12.69	8.48	0.61	2.13	1.82	2.93±4.22

SD: Standard deviation. Bold values indicate the lowest mean EERWA and the lowest SD among the evaluated models.

TABLE VII
Performance of EN_Lite Models Pre-Trained on ImageNet

Model	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Mean±SD
IEN_Lite-1	2.11	1.00	0.86	0.15	1.76	15.18	3.81	1.58	2.64	0.80	**2.99**±4.41
IEN_Lite-2	1.69	2.24	0.36	0.11	2.43	12.89	8.29	2.48	3.47	1.01	3.5±**4.02**
IEN_Lite-3	2.37	1.88	0.35	0.35	3.35	17.42	4.08	2.20	3.52	1.48	3.7±4.98
IEN_Lite-4	2.76	1.91	0.71	0.44	2.78	22.07	5.62	2.67	3.53	1.34	4.38±6.39
IEN_Lite-5	2.35	2.69	1.02	0.47	2.90	17.04	4.31	2.61	3.32	1.93	3.86±4.76

SD: Standard deviation. Bold values indicate the lowest mean EERWA and the lowest SD among the evaluated models.

TABLE VIII
Performance of EN_Lite Models Pre-Trained on VGGFace2

Model	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Mean±SD
VEN_Lite-1	2.87	2.95	0.23	0.40	3.50	27.88	16.21	3.54	4.38	1.73	6.37±8.81
VEN_Lite-2	3.98	2.96	0.35	0.28	3.40	24.71	8.12	3.29	5.16	2.44	**5.47**±7.13
VEN_Lite-3	6.28	3.41	0.35	0.35	3.91	21.53	11.05	2.91	3.83	1.85	5.55±6.42
VEN_Lite-4	5.39	3.81	0.70	0.45	4.31	24.65	10.70	4.69	5.66	4.15	6.45±6.99
VEN_Lite-5	4.43	3.29	1.28	0.51	5.84	21.65	7.02	4.34	4.71	4.69	5.78±**5.91**

SD: Standard deviation. Bold values indicate the lowest mean EERWA and the lowest SD among the evaluated models.

(5.03%), ShuffleNet V2 (5.31%), MobileNetV3-Small (6.03%), DenseNet-201 (6.60%), ResNet-50 (6.73%), and Inception-v3 (7.52%). Performance generally degraded under heterogeneous cross-morph evaluation, indicating that generalization to unseen morphing methods remains challenging.

B. EfficientNet Family Performance

The performance of EfficientNet variants B0–B4 across all evaluation scenarios is summarized in Table V. Among these models, EfficientNet-B3 achieved the lowest mean EER_{WA} (2.65%), followed by EfficientNet-B1 (2.78%), EfficientNet-B4 (2.93%), EfficientNet-B0 (3.47%), and EfficientNet-B2 (4.05%). Overall, the results suggest that scaling beyond B0 can slightly improve MAD performance, although the gains are not strictly monotonic across all variants.

C. ImageNet-Pretrained EN_Lite Performance

The mean performance of the ImageNet-pretrained EN_Lite (IEN_Lite) variants is summarized in Table VI. Among these models, IEN_Lite-1 achieved the best overall result, with the lowest mean EER_{WA} of 2.99%. The remaining variants showed slightly weaker performance, with mean EER_{WA} values of 3.5% for IEN_Lite-2, 3.7% for IEN_Lite-3, 4.38% for IEN_Lite-4, and 3.86% for IEN_Lite-5. These results indicate that moderate model reduction can preserve strong MAD performance, whereas further scaling down leads to less consistent gains.

D. VGGFace2-Pretrained EN_Lite Performance

The performance of the VGGFace2-pretrained EN_Lite (VEN_Lite) variants is shown in Table VII.

VEN_Lite-2 achieved the lowest mean EER_{WA} (5.47%), followed by VEN_Lite-3 (5.55%), VEN_Lite-5 (5.78%), VEN_Lite-1 (6.37%), and VEN_Lite-4 (6.45%). All VEN_Lite variants produced higher mean EER_{WA} than their corresponding IEN_Lite variants under the evaluated protocol.

The mean EER_{WA} values of the EfficientNet and EN_Lite variants are summarized in Fig. 1. EfficientNet-B3 achieved the lowest mean EER_{WA} (2.65%), while IEN_Lite-1 reached 2.99%. Among the VGGFace2-HQ-pretrained variants, VEN_Lite-2 achieved the lowest mean EER_{WA} (5.47%).

The operational error behavior of the evaluated models is further illustrated in Fig. 2 using scenario-balanced weighted-mean DET curves across Tables IV-VII. The curves show that EfficientNet-B3 and IEN_Lite-1 generally achieve lower BPCER at fixed APCER operating points, whereas the VEN_Lite variants exhibit weaker detection behavior.

Across S1–S10, IEN_Lite-1 increased mean EER_{WA} by 0.34% points relative to EfficientNet-B3 (2.99 ± 4.41% vs. 2.65 ± 4.01%) while reducing parameters by approximately 76%; the difference was not statistically significant (exact two-sided Wilcoxon signed-rank test, $p = 0.232$). Thus, IEN_Lite-1 indicates a parameter-efficient trade-off under this protocol, while deployment impact should be assessed at application-specific DET operating points rather than from EER alone. IEN_Lite variants also significantly outperformed their corresponding VEN_Lite variants (exact two-sided paired Wilcoxon signed-rank tests: $p = 0.0059$, 0.0117, 0.0078, 0.0059, and 0.0020 for variants 1–5, respectively). To check whether dataset-size weighting masked small-dataset behavior, the main comparison was also repeated with equal dataset weights; IEN_Lite-1 obtained lower unweighted mean EER than EfficientNet-B3 (4.39% vs. 5.01%) and lower SD across dataset-level mean EERs (1.88% vs. 2.92%).

VI. Discussion

The results show that increasing model size by itself is not enough to ensure better MAD performance. The small EfficientNet-B0 performed better than much larger models, such as ResNet-50 and Inception-v3, indicating that the architecture of the model matters more than the number of parameters. The most plausible reason behind the excellent performance of EfficientNet-B0 lies in its compound scaling mechanism that is far superior to the traditional scaling techniques in terms of handling the depth, width, and input resolution of the network. Performance dropped in scenarios where validation on unseen morphs. Specifically, in Scenario 6, the model was trained on StyleGAN but validated on landmark-based morphs. This can be interpreted as evidence that the model overfits to GAN-specific artifacts rather than learning generalizable morphing cues.

The results of EfficientNet variants suggest a non-linearity relationship between model capacity and the efficiency of MAD performance. The EfficientNet-B3 produced the lowest mean EER_{WA} value (2.65%), followed by B1, suggesting that mid-range models might have an advantage in terms of a good trade-off between representational capacity and generalization ability. In turn, B2 and B4 models exhibited weaker and inconsistent performances, respectively, implying that increasing the model capacity has limited benefit.

Among the ImageNet-pretrained EN_Lite variants, IEN_Lite-1 achieved the best average performance, followed by IEN_Lite-2 and IEN_Lite-3. However, performance slightly declined for the IEN_Lite-4 and IEN_Lite-5 variants. This suggests that moderate down-scaling can preserve much of the model's useful discriminative capacity, while excessive reduction may limit its ability to detect morphing artifacts reliably. The aim is not to claim EN_Lite as universally optimal among lightweight CNNs. Under the evaluated protocol, IEN_Lite-1 achieved a lower mean EER_{WA} (2.99%) than GhostNet-1.0x (4.01%), MobileNetV3-Large (5.03%), MobileNetV3-Small (6.03%), and ShuffleNet V2 (5.31%). These direct comparisons support a favorable efficiency–performance trade-off for IEN_Lite-1. Because EN_Lite scaling also reduces input resolution and channel capacity, the weaker performance of smaller variants may partly reflect reduced sensitivity to fine spatial or high-frequency morphing artifacts; frequency-domain sensitivity analysis is therefore left for future work.

The VGGFace2-HQ-pretrained EN_Lite variants remained weaker than their corresponding ImageNet-pretrained

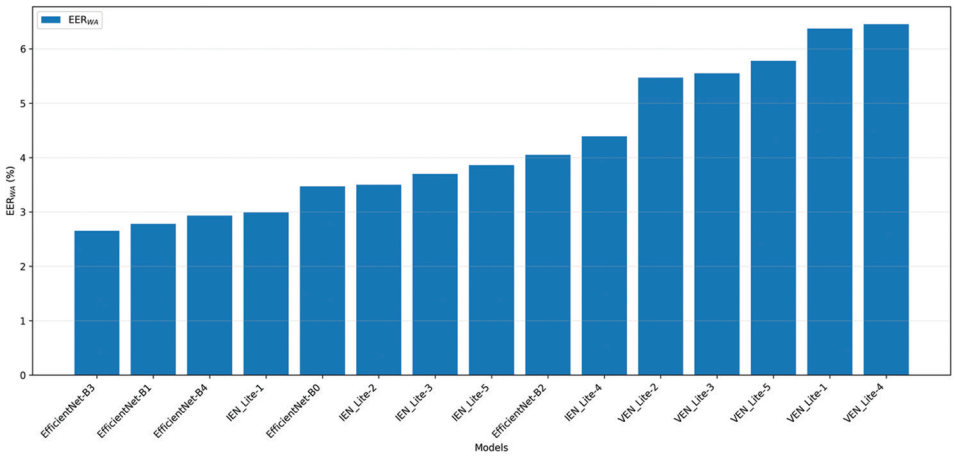


Fig. 1. Mean EERwa of EfficientNet and EN_Lite variants under ImageNet-1K and VGGFace2-HQ pre-training. Lower values indicate better performance.

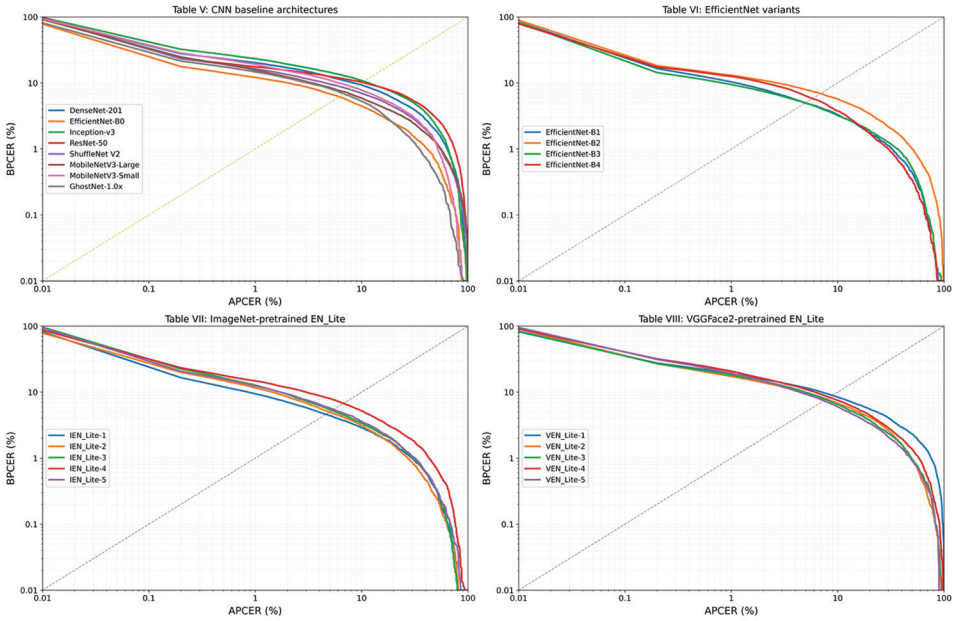


Fig. 2. Scenario-balanced weighted-mean DET curves for the models reported in Tables IV-VII. Lower curves indicate better performance, corresponding to a lower bona fide presentation classification error rate at a fixed attack presentation classification error rate operating point.

variants under the evaluated protocol, with VEN_Lite-2 achieving the lowest mean EER_{WA} of 5.47%. The paired differences between corresponding IEN_Lite and VEN_Lite variants were statistically significant. Nevertheless, because ImageNet-1K and VGGFace2-HQ differ in image content, class structure, and composition, these results should be interpreted as an empirical comparison under substantially more comparable pre-training scale rather than as evidence that generic pre-training is intrinsically superior to face-specific pre-training.

The leave-one-dataset-out design across the eight datasets provides a strong setting for evaluating MAD generalization, including performance under previously unseen morph types. Despite the low error rates observed in homogeneous scenarios, performance degraded substantially in heterogeneous cross-morph settings, especially in Scenario 6, where models were trained on StyleGAN morphs and evaluated on landmark-based morphs. This result indicates that present CNN-based MAD systems, including the proposed EN_Lite variants, still struggle to learn sufficiently robust morph-invariant representations under challenging cross-morph conditions.

StyleGAN morphs tend to introduce synthesis-related texture irregularities and high-frequency artifacts, whereas landmark-based morphs are more often characterized by local geometric distortions and blending inconsistencies. As a result, models trained on artifacts from a specific morphing method may become too specialized, limiting their ability to generalize to morphs produced by other techniques.

The relatively stronger results observed in mixed-training settings, where multiple morph types were included during training, suggest that exposure to a wider range of attack

methods can improve generalization. However, these settings still did not fully close the gap with homogeneous evaluation, indicating that training diversity alone is not sufficient.

Importantly, this cross-morph degradation was not limited to the lightweight variants. EfficientNet-B3 and IEN_Lite-1 both showed substantially higher heterogeneous than homogeneous EER_{WA}, with heterogeneous means of 4.43% and 4.80%, respectively. Thus, downsizing did not appear to uniquely cause the observed artifact-specific behavior for IEN_Lite-1, although excessive reduction may still limit robustness in smaller EN_Lite variants.

Several limitations remain. Dataset and morph-type imbalance may affect weighted averages. The GAN condition used only StyleGAN morphs from the top 5% most similar pairs because exhaustive generation was impractical. A fixed seed was used without multiple-seed repetitions, so statistics reflect scenario-wise rather than training-run variability. Although the two pre-training strategies are evaluated at substantially more comparable scales, ImageNet-1K and VGGFace2-HQ still differ in content and class structure. FLOPs, latency, memory use, calibration, and threshold stability remain future work.

VII. Conclusion

In this paper, we proposed EN_Lite as a parameter-efficient CNN family for MAD. IEN_Lite-1 achieved a mean EER_{WA} of 2.99%, compared with 2.65% for EfficientNet-B3 while reducing parameters by approximately 76%. Under the same evaluation protocol, it also achieved a lower mean EER_{WA} than GhostNet-1.0x (4.01%), MobileNetV3-Large (5.03%), MobileNetV3-Small (6.03%), and ShuffleNet V2 (5.31%), supporting a favorable efficiency-performance trade-off rather than automatic equivalence for high-security deployment. Using VGGFace2-HQ pre-training with 1,107,817 successfully processed facial images; the VEN_Lite variants produced higher error rates than their corresponding IEN_Lite variants. However, because ImageNet-1K and VGGFace2-HQ differ in content and class structure, this finding is interpreted as an empirical outcome under the evaluated protocol rather than evidence of an intrinsic advantage of generic pre-training.

Although all architectures achieved strong performance under homogeneous scenarios, performance noticeably degraded under heterogeneous scenarios, particularly when training only on StyleGAN morphs and evaluating on landmark-based morphs. This degradation may indicate that models trained exclusively on StyleGAN learn morph-specific features that do not generalize sufficiently to landmark-based morphs. In contrast, training on landmark-based morphing types generally resulted in better cross-morph performance, possibly because landmark-based methods share more similar artifact characteristics. These observations highlight the importance of including diverse morphing types in the training set.

In acknowledgment to the limitations of this work, future work should aim for a more balanced demographic composition, perform systematic hyperparameter tuning, and consider a broader set of landmark-based and GAN-based morphing techniques. Hence, the reported results should be interpreted as strong baseline findings rather than the full performance potential of the models.

References

Aghdaie, P., Chaudhary, B., Soleymani, S., Dawson, J., and Nasrabadi, N.M., 2021. Detection of morphed face images using discriminative wavelet sub-bands. In: *2021 IEEE International Workshop on Biometrics and Forensics (IWBF)*, IEEE, pp.1-6.

Asaad, A., and Jassim, S., 2017. Topological Data Analysis for Image Tampering Detection. Digital Forensics and Watermarking: 16th International Workshop, IWDW 2017, Magdeburg, Germany, August 23-25, 2017, Proceedings 16, Springer, pp.136-146.

Caldeira, E., Neto, P.C., Gonçalves, T., Damer, N., Sequeira, A.F., and Cardoso, J.S., 2023. Unveiling the two-faced truth: Disentangling morphed identities for face morphing detection. 2023 31st European Signal Processing Conference (EUSIPCO), IEEE, pp.955-959.

Cao, Q., Shen, L., Xie, W., Parkhi, O.M., and Zisserman, A., 2018. VGGFace2: A Dataset for Recognising Faces Across Pose and Age. In: *2018 13th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2018)*, IEEE, pp.67-74.

Colbois, L., and Marcel, S., 2022. On the Detection of Morphing Attacks Generated by GANs. In: *2022 International Conference of the Biometrics Special Interest Group (BIOSIG)*, IEEE, pp.1-5.

Damer, N., Raja, K., Süßmilch, M., Venkatesh, S., Boutros, F., Fang, M., Kirchbuchner, F., Ramachandra, R., and Kuijper, A., 2021. ReGenMorph: Visibly Realistic GAN Generated Face Morphing Attacks by Attack Re-Generation. In: Advances in Visual Computing: 16th International Symposium, ISVC 2021, Virtual Event, October 4-6, 2021, Proceedings, Part I, Springer, pp.251-264.

DeBruine, L., and Jones, B., 2017. Face Research Lab London Set. Figshare. Dataset. Available from: https://doi.org/10.6084/m9.figshare.5047666.v4 [Last accessed on 24 April 2024].

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., and Fei-Fei, L., 2009. ImageNet: A large-scale hierarchical image Database. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, pp.248-255.

Deng, J., Guo, J., Xue, N., and Zafeiriou, S., 2019. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.4690-4699.

Ferrara, M., Franco, A., and Maltoni, D., 2014. The Magic Passport. In: *IEEE International Joint Conference on Biometrics*, Clearwater, FL, USA, pp.1-7.

Hamza, M., Tehsin, S., Humayun, M., Almufareh, M.F., and Alfayad, M., 2022. A comprehensive review of face morph generation and detection of fraudulent identities. *Applied Sciences*, 12(24), p.12545.

Han, K., Wang, Y., Tian, Q., Guo, J., Xu, C., and Xu, C., 2020. GhostNet: More Features From Cheap Operations. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, pp.1580-1589.

Hancock, P., 2008. Psychological Image Collection at Stirling (PICS) – 2D Face Sets – Utrecht ECVP. Available from: https://pics.stir.ac.uk [Last accessed on 24 April 2024].

He, K., Zhang, X., Ren, S., and Sun, J., 2016. Deep Residual Learning for Image Recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.770-778.

Hoang, T.T., Islam, B.M.S., and Ahn, H., 2025. Face Morphing Attack Detection Using Similarity Score Patterns Between De-Morphed and Live Images. *Electronics*, 14(19), p.3851.

Howard, A., Sandler, M., Chu, G., Chen, L.C., Chen, B., Tan, M., Wang, W.,

Zhu, Y., Pang, R., Vasudevan, V., Le, Q.V., and Adam, H., 2019. Searching for MobileNetV3. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp.1314-1324.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K.Q., 2017. Densely Connected Convolutional Networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.4700-4708.

Huber, M., Boutros, F., Luu, A.T., Raja, K., Ramachandra, R., Damer, N., Neto, P.C., Gonçalves, T., Sequeira, A.F., and Cardoso, J.S., 2022. SYN-MAD 2022: Competition on Face Morphing Attack Detection Based on Privacy-Aware Synthetic Training Data. In: 2022 IEEE *International Joint Conference on Biometrics (IJCB)*, IEEE, pp.1-10.

Hussein, A.R., and Rashid, R.D., 2022. KurdFace morph dataset creation using OpenCV. *Science Journal of University of Zakho*, 10(4), pp.258-267.

Ibrahem, A.H., and Abdelazeez, A.M., 2026. KurdFace-1000: A new multi-attribute dataset for facial beauty prediction using multi-task learning. *ARO-The Scientific Journal of Koya University*, 14(1), pp.1-11.

Ibsen, M., Rathgeb, C., Marcel, S., and Busch, C., 2024. Multi-Channel Cross Modal Detection of Synthetic Face Images. In: *2024 12th International Workshop on Biometrics and Forensics (IWBF*), IEEE, pp.1-6.

ISO/IEC, 2023. ISO/IEC 30107-3 *Information Technology Biometric Presentation Attack Detection - Part 3: Testing and Reporting*. International Organization for Standardization, Switzerland.

Jassim, S., and Asaad, A., 2018. Automatic Detection of Image Morphing by Topology-Based Analysis. In: *2018 26th European Signal Processing Conference (EUSIPCO)*, IEEE, pp.1007-1011.

Karras, T., Laine, S., and Aila, T., 2019. A Style-Based Generator Architecture for Generative adversarial Networks. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.4401-4410.

Kashiani, H., Sami, S.M., Soleymani, S., and Nasrabadi, N.M., 2022. Robust Ensemble Morph Detection with Domain Generalization. In: *2022 IEEE International Joint Conference on Biometrics (IJCB)*, IEEE, pp.1-10.

Lakshmi, A., and Wittenbrink, B., 2021. The India face set: International and cultural boundaries impact face impressions and perceptions of category membership. *Frontiers in Psychology*, 12, p.627678.

Ma, D.S., Correll, J., and Wittenbrink, B., 2015. The Chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods*, 47, pp.1122-1135.

Ma, D.S., Kantner, J., and Wittenbrink, B., 2021. Chicago face database: Multiracial expansion. *Behavior Research Methods*, 53, pp.1289-1300.

Ma, N., Zhang, X., Zheng, H.T., and Sun, J., 2018. ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design. In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp.116-131.

Mahfoudi, G., Tajini, B., Retraint, F., Morain-Nicolier, F., Dugelay, J.L., and Marc, P., 2019. *DEFACTO: Image and Face Manipulation Dataset. In: 2019 27th European Signal Processing Conference (EUSIPCO)*, IEEE, pp.1-5.

Makrushin, A., Neubert, T., and Dittmann, J., 2017. Automatic Generation and Detection of Visually Faultless Facial Morphs. In: *International Conference on Computer Vision Theory and Applications, SciTePress*, pp.39-50.

Mallick, S., 2016. Face Morph Using OpenCV — C++/Python. Available from: https://learnopencv.com/face-morph-using-opencv-cpp-python [Last accessed on 24 Apr 2024].

Phillips, P.J., Flynn, P.J., Scruggs, T., Bowyer, K.W., Chang, J., Hoffman, K., Marques, J., Min, J., and Worek, W., 2005. Overview of the Face Recognition Grand Challenge. In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05*), IEEE, pp.947-954.

Phillips, P.J., Wechsler, H., Huang, J., and Rauss, P.J., 1998. The FERET database and evaluation procedure for face-recognition algorithms. *Image and Vision Computing*, 16, pp.295-306.

Pimenta, J.A., Medvedev, I., and Gonçalves, N., 2023. Impact of Image Context for Single Deep Learning Face Morphing Attack Detection. In: *2023 International Conference of the Biometrics Special Interest Group (BIOSIG)*, IEEE, pp.1-6.

Quek, A., 2019. FaceMorpher. Available from: https://github.com/alyssaq/face_morpher [Last accessed 24 Apr 2024].

Raja, K., Ferrara, M., Franco, A., Spreeuwers, L., Batskos, I., De Wit, F., Gomez-Barrero, M., Scherhag, U., Fischer, D., and Venkatesh, S.K., 2020. Morphing attack detection-database, evaluation platform, and benchmarking. *IEEE Transactions on Information Forensics and Security*, 16, pp.4336-4351.

Scherhag, U., Rathgeb, C., Merkle, J., Breithaupt, R., and Busch, C., 2019. Face recognition systems under morphing attacks: A survey. *IEEE Access*, 7, pp.23012-23026.

Seibold, C., Samek, W., Hilsmann, A., and Eisert, P., 2017. Detection of Face Morphing Attacks by Deep Learning. In: *DIGITAL Forensics and Watermarking: 16th International Workshop, IWDW 2017,* Magdeburg, Germany, August 23-25, 2017, Proceedings 16, Springer, pp.107-120.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z., 2016. Rethinking the Inception Architecture for Computer Vision. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.2818-2826.

Tan, M., and Le, Q., 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. In: *International Conference on Machine Learning*, PMLR, pp.6105-6114.

Tapia, J., and Busch, C., 2023. AlphaNet: Single morphing attack detection using multiple contributors. In: *2023 IEEE International Workshop on Information Forensics and Security (WIFS)*, IEEE, pp.1-6.

Tapia, J.E., Russo, M., and Busch, C., 2025. Generating automatically print/scan textures for morphing attack detection applications. *IEEE Access*, *13*, pp.55277-55289.

Tapia, J.E., Schulz, D., and Busch, C., 2025. Single-morphing attack detection using few-shot learning and triplet-loss. *Neurocomputing*, 636, pp.130033.

Thomaz, C.E., and Giraldi, G.A., 2010. A new ranking method for principal components analysis and its application to face image analysis. *Image and Vision Computing*, 28(6), pp.902-913.

Venkatesh, S., Ramachandra, R., Raja, K., and Busch, C., 2020. Single Image Face Morphing Attack Detection using Ensemble of Features. In: *2020 IEEE 23rd International Conference on Information Fusion* (FUSION), IEEE, pp.1-6.

Venkatesh, S., Ramachandra, R., Raja, K., and Busch, C., 2021. Face morphing attack generation and detection: A comprehensive survey. *IEEE Transactions on Technology and Society*, 2(3), pp.128-145.

Zhang, H., Ramachandra, R., Raja, K., and Busch, C., 2024. Generalized single-image-based morphing attack detection using deep representations from Vision Transformer. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.1510-1518.

Zhang, H., Ramachandra, R., Raja, K., and Busch, C., 2026. Chatgpt encounters morphing attack detection: Zero-shot mad with multi-modal large language models and general vision models. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 8(2), pp.154-168.

Zhang, H., Venkatesh, S., Ramachandra, R., Raja, K., Damer, N., and Busch, C., 2021. MIPGAN—generating strong and high quality morphing attacks using identity prior driven GAN. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(3), pp.365-383.