

Robust Ensemble Morph Detection with Domain Generalization

Hossein Kashiani, Shoaib Meraj Sami, Sobhan Soleymani, Nasser M. Nasrabadi
West Virginia University

{hk00014, sms00052, ssoleyma}@mix.wvu.edu, nasser.nasrabadi@mail.wvu.edu

Abstract

Although a substantial amount of studies is dedicated to morph detection, most of them fail to generalize for morph faces outside of their training paradigm. Moreover, recent morph detection methods are highly vulnerable to adversarial attacks. In this paper, we intend to learn a morph detection model with high generalization to a wide range of morphing attacks and high robustness against different adversarial attacks. To this aim, we develop an ensemble of convolutional neural networks (CNNs) and Transformer models to benefit from their capabilities simultaneously. To improve the robust accuracy of the ensemble model, we employ multi-perturbation adversarial training and generate adversarial examples with high transferability for several single models. Our exhaustive evaluations demonstrate that the proposed robust ensemble model generalizes to several morphing attacks and face datasets. In addition, we validate that our robust ensemble model gain better robustness against several adversarial attacks while outperforming the state-of-the-art studies.

1. Introduction

Face recognition systems are built upon the hypothesis that the face is uniquely linked to the identity. When it comes to face verification scenarios, morph attacks are potential threats due to their capacity to break this unique connection. Face morph attacks are particularly dangerous for border security since a morphed passport photo allows unauthorized entry to be unnoticed. Through this loophole, a criminal can apply for a passport using a morphed face. Face morphing is an image manipulation technique in which two faces are blended together to form a morphed image of high quality. At the time of passport enrollment, the passport photo can be easily manipulated with a morphing attack without the requirement of advanced passport forgery technologies. Therefore, face recognition systems rely highly on the morph attack detection.

A typical assumption in morph detection studies [43, 9, 2, 1] is that the source and target data are independent and

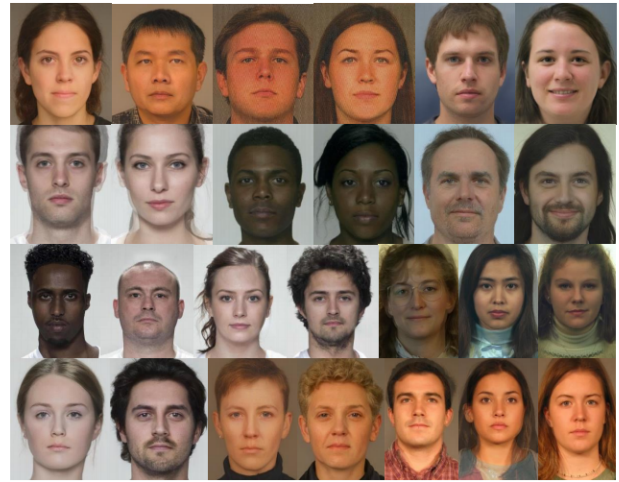


Figure 1: Domain shift between different morphing attacks in different datasets, including FERET [39], FRLL [18, 42], FRGC [40], and AMSL [36]. Morphs correspond to landmark-based morphing methods such as Facemorpher [41] and WebMorph [17] and GAN-based morphing methods such as MIPGAN [53] and StyleGAN2 [42].

identically distributed (*i.i.d.*) [44, 6, 15]. This assumption neglects the domain shift challenge in the real world (Figure 1). As a result, when applied to out-of-distribution samples, a learning agent that is only trained on source data often performs much worse. The domain shift issue may be mitigated to some extent by including some data from the target domain, but this is a non-trivial task since preparing image annotations is both costly and time-consuming. Moreover, despite the remarkable success of deep CNNs in face recognition systems, they are known to be susceptible to another type of attack known as adversarial attack. An adversarial attack attempts to generate an adversarial example with indistinguishable and purposeful perturbations, forcing CNNs to an incorrect prediction. In this regard, the vulnerability of morph detection systems to adversarial examples have serious consequences.

In this study, we develop an ensemble of deep mod-

els trained on a morph source domain to improve domain generalization. It should be noted that neural networks mostly do not attain perfect performance owing to the overwhelming number of local minima. In other words, neural networks might fall into a variety of local minima. As a result, individual neural networks do not operate optimally in various areas of the feature space, and the errors do not have a large positive correlation between them [23]. To address this issue, we adopt an ensemble of well-trained neural networks. Our evaluations demonstrate that the ensemble learning can partly mitigate the domain shift challenge. In this respect, we study an ensemble model, which comprises of a ViT B-16 Transformer [22], ViT L-32 Transformer [22], and a noise-aware Inception ResNet network [26, 49]. Noise-aware Inception ResNet network is built upon Venkatesh *et al.*, [49] with an Inception ResNet backbone. Domain generalization needs translation equivariance, local receptive fields, weight sharing, and long-range dependencies. Integration of CNN and Transformer models can incorporate these objectives at the same time. Therefore, we propose an ensemble model to include the strengths of both CNN and Transformer architectures [49, 26, 22, 46, 28] simultaneously.

In addition, to robustify our ensemble model against adversarial perturbations, we first generate adversarial examples with high transferability in a black-box setting. To this end, we generate a wide range of adversarial perturbations with different methods such that our adversarial examples fool different morph detection methods in a white-box setting. Then, we robustify our ensemble model using multi-perturbation adversarial training [25]. With the multi-perturbation adversarial training, we enforce the output of the proposed ensemble model to remain nearly within an l_p ball of every training sample to enhance robust accuracy while keeping favorable clean accuracy. Our major contributions in this paper are:

- We integrate CNN and Transformer models and propose an ensemble model for morph detection that highly generalizes to a wide range of morphing attacks.
- We craft highly transferable adversarial examples for multi-perturbation adversarial training to improve the adversarial robustness of our ensemble models.
- We carry out extensive evaluations on different datasets to prove the generalization capability and adversarial robustness of our ensemble model.

2. Related Works

2.1. Morph Detection

Morph attack detection methods are divided into single and differential morph detection. In the single morph detection, the detector only aims to detect the potential morphed

image for final classification. This corresponds to morph detection during the passport application, when the applicant uploads his passport picture. On the other hand, differential morph detection utilizes an additional trusted image of the real subject for its detection. To be more specific, it makes a comparison between the potential morphed image and the trusted reference image.

Recently, deep learning models have been widely used for morph detection. Authors in [6] employ a conditional generative network (cGAN) for differential face morph attack detection. The cGAN learns to implicitly extract identities from the morphed image conditioned on the trusted reference image. Aghdaie *et al.* [3] exploit wavelet domain analysis to explore spatial frequency information for morph detection. They train a CNN-based morph classifier on the decomposed wavelet sub-bands of the morphed and bona fide images. Authors in [49] propose a deep multi-scale context aggregation model to capture residual morphing noise. The pre-trained AlexNet [30] is then utilized to compute the textural features of the residual morphing noise for morph attack detection. Recently, Damer *et al.* [14] introduce a synthetic-based development dataset for morph attack detection to address the legal issues of utilizing biometric information.

2.2. Domain Generalization and Ensemble models

Machine learning paradigms [37, 35, 4, 38, 34] are built upon the assumption that training and test data have the same probability distributions. When this hypothesis is even marginally broken, as most real-life settings, a significant drop in performance can be observed. Domain generalization aims to address this issue and mainly covers the events in which the target source is unavailable during the training phase. In this respect, ensemble learning is investigated for domain generalization. Domain-specific neural networks [57], and weight averaging [8] are a number of domain generalization-based ensemble learning studies. In ensemble learning, we seek to learn several instances of the same model with different initializations, or we learn different explicit models. In this respect, Xu *et al.* [51] demonstrate that naively integrating diverse training distributions from different source domains can contribute to out-of-distribution generalization. In our study, domain generalization from a single source domain is targeted using different state-of-the-art explicit deep learning models. We construct our ensemble model from CNN and Transformer models since the integration of these models enables us to capture long-term and short-term connections in the visual data.

2.3. Adversarial Robustness

While face recognition is improved significantly over recent years thanks to the breakthroughs in deep learning, the

tion [10] to finally calculate the residual morphing artifacts. Given the calculated residual morphing artifacts, we utilize the Inception-ResNet-V1 model [45] for morph detection. In this paper, we call this model the noise-aware ResNet model.

3.2. Transformer Classifier

Due to their capability in capturing long-term dependencies between different regions of the data, Transformers are considered in many tasks such as object recognition, image classification, and face recognition. Morph detection problem can also benefit from this capability since long-term dependencies between different regions in facial images can provide important information. We leverage from the ViT B-16 and ViT L-32 [22] Transformers for morph detection task in our ensemble model.

Tokenization A ViT Transformer applies attention to small patches of input images instead of pixels. To be more specific, ViT divides a 2D image $x \in \mathbb{R}^{H \times W \times 3}$ into $R \times R$ flattened patches $x_f \in \mathbb{R}^{N \times (R^2 \times 3)}$ uniformly, where $(H \times W)$, R^2 , 3, and N stand for spatial dimensions of the input image and patches, three RGB channels, and the number of total patches, respectively. It is worth noting that ViT underperforms for both extremely small and large patch sizes since the former induce too much computational cost, while the latter cannot encode the low-level context in the input images. The first layer of ViT projects the flattened patches (also called patch tokens) into latent D-dimensional embeddings [22]. To learn about the structure of the input data, a positional embedding is added to the flattened patches which contains information about the location of patches. An extra classification token x_{class} is also appended to the embedded patches. The content of this classification token interacts with all the remaining patch tokens in the self-attention module, accumulating data for the eventual classification [22]. With this structure, nearby patches contain more comparable position embeddings, and the embeddings of patches in the same row or column are identical. Thus, the ViT utilizes the latent position embedding to consider distances within the input image patches. The resulting vector corresponding to different patch tokens is then fed into the transformer encoders. Note that the x_{class} in the Transformer encoder finally results in y output [22].

Encoder Block: Each block of a Transformer encoder is composed of a feed-forward network (FFN) and a multi-headed self-attention module (MSA). The hidden layer of the FFN is constructed from a two-layer multi-layer perceptron as well as a Gaussian Error Linear Unit (GELU) non-linearity layer. Around each sub-layer, residual shortcuts are adopted, which are preceded by a layer normalization

(LN). The encoder of ViT works as follows:

$$y = \hat{x} + \text{FFN}(\text{LN}(\hat{x})), \quad (1)$$

$$\hat{x} = x + \text{MSA}(\text{LN}(x)), \quad (2)$$

where x and y are the out of tokenization phase and the final output class, respectively. A series of encoders component constitutes a Transformer encoder.

Multi-Head Self-Attention: Multi-head attention extends single-head self-attention to encompass various complicated connections among distinct image patches. The self-attention mechanism is a fundamental building block of Transformers that systematically simulates the relationship between all pixels in an image by giving a pairwise attention score between every two patch tokens in the terms of global contextual information. To this end, the embeddings of $R \times R$ pixel patches, x_f , are mapped to three learnable matrices: query ($\mathbf{Q} \in \mathbb{R}^{N+1 \times C}$), key ($\mathbf{K} \in \mathbb{R}^{N+1 \times C}$), and value ($\mathbf{V} \in \mathbb{R}^{N+1 \times C}$). Then, the self-attention mechanism is defined as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right) \mathbf{V}, \quad (3)$$

where d_k is the dimension of $\mathbf{K}$ matrix. This single self-attention process is performed for each head in parallel in the MSA module. Then, the output of various head units is concatenated along the channel dimension.

3.3. Fusion

To construct our ensemble model, we train a FFN to compute the matching scores of all single models in the fusion phase. For this goal, we capture score z_i from the output of each model $i \in 1, 2, 3$. Since the large scores which belong to a certain model can dominate the others and force them to be small, the scores of single models are first normalized as follows :

$$\tilde{z}_{ip} = \frac{e^{z_{ip}}}{\sum_j e^{z_{ij}}}, \quad i = 1, 2, 3, \quad (4)$$

where $\tilde{z}_{ip}$ is the normalized score of a single model. Finally, the normalized scores of all single models are ensembled for final classification as follows:

$$y = \text{FFN}[\tilde{z}_{1p}; \dots; \tilde{z}_{ip}], \quad (5)$$

where FFN denotes a two-layer feed forward network with hidden layers of 256 and 2 units. The FC layers are also activated with a RELU function. Note that, during the training, we randomly set a limited number of these scores to zero to cope with the overfitting.

3.4. Multi-Perturbation Adversarial Training

Adversarial training is one of the main types of defense against adversarial attacks that aims to train moderately robust DNNs with adversarial examples [47]. In adversarial training, we require to delicately set the perturbation level ϵ for the generation of adversarial examples in the training phase. Using small values of ϵ in adversarial training fails to generate hard adversarial examples, and consequently, the final trained model would lack adversarial robustness. On the other side, large values of ϵ would improve the adversarial robustness at the cost of a significant clean accuracy drop. As such, we need to tune the perturbation level ϵ to strike the right balance between robust and clean accuracy. Moreover, the adversarially trained models with a certain type of adversarial perturbation are still vulnerable to other types of adversarial attacks [47]. To cope with this issue, we adopt ensemble-based multi-perturbation adversarial training. To this end, we augment training data with a wide range of adversarial perturbations. In a white-box setting, we fool each single models and generate adversarial examples with different perturbation level ϵ . It is worth highlighting that several types of adversarial attacks are included in order to improve the network robustness.

To improve the effectiveness of our adversarial training, we require to craft the adversarial examples with high transferability. Conventionally, the adversarial examples can fool the white-box models more easily compared to the black-box models. The adversarial perturbations generally do not transfer well to the black-box models with different topologies [27]. Transferability, which allows adversarial examples to transfer to networks of unknown structures, makes adversarial examples even more harmful. To address this challenge and generate highly transferable adversarial examples, we utilize the model-based ensembling Attack [32]. In this approach, we assume that if an adversarial example succeeds in fooling several models, it can also fool black-box models with high probability [32]. With that in mind, we produce adversarial examples for an ensemble of single models in a white-box setting by optimizing the following equation:

$$\operatorname{argmax}_{x^*} -\log \left(\left(\sum_{i=1}^n \alpha_i J_i(x^*) \right) \cdot 1_y \right) + \lambda d(x, x^*), \quad (6)$$

where x , y , and n denote the input image, the ground-truth class, and the number of single models. In addition, J_i , α_i and d indicate the output of a single model, ensemble weights, and the distance between input and perturbed images. This optimization results in the alignment of the decision boundaries of the individual models. Therefore, the crafted adversarial example would have higher transferability in the black-box setting. Equipped with

these highly transferable adversarial examples, we employ multi-perturbation adversarial training over single models. To be more specific, we apply this optimization (Equation 6) to several untargeted adversarial attacks to craft highly transferable adversarial images and utilize them in the adversarial training of the ensemble model. These attacks include fast gradient sign method (FGSM) [24], basic interactive method (BIM) [31], Random initialization FGSM (RFGSM) [48], projected gradient method (PGD), [33], PGDL [33], Trade-off PGD (TPGD) [54], Auto-PGD (APGD) [11], APGD-Targeted (APGDT) [11], SmoothFool [13], and AutoAttack [11].

4. Evaluations

We benchmark the generalization performance of the proposed ensemble model on a wide range of unseen morph attacks. In a real-world scenario, the morphing technology used by a criminal attacker may be novel and unknown. As such, it is reasonable to evaluate morph detection methods on several "unseen" datasets to obtain an accurate representation of how the network performs. To this end, a wide range of target domains are employed in our investigations. To assess network performance, we rely on the Area Under the Receiver Operating Characteristic Curve (AUC), Attack Presentation Classification Error Rate (APCER), and Bona fide Presentation Classification Error Rate (BPCER), which are the standard measures for morph attack detection. APCER stands for Attack Presentation Classification Error Rate or the rate at which morphs are incorrectly categorized as bona fides. BPCER, on the other hand, is the rate at which bona fide images are wrongly identified as morphs [44].

4.1. Experimental setup

To train the single models, the input images are first pre-processed with the MTCNN framework [55]. Faces are detected, aligned, and resized to 512×512 using the MTCNN framework. All training images are further augmented with horizontal flips. Additionally, the batch generator is weighted to mitigate the class imbalance at every iteration. We utilize our private dataset as the source domain, which includes about 9,052 bona fide and 1,2991 morphed images. The target domains consist of five separate datasets, including FERET [39], FRLL [18, 42], FRGC [40], and AMSL [36] datasets. To train the base single models on the source domain, we use the mini-batch stochastic gradient descent optimization with batch size 256, and the Adam solver. The initial learning rate is set to 5×10^{-5} , and it is decreased by a factor of 2 at 20^{th} and 30^{th} epochs. In the adversarial training, to strike a right robustness/accuracy balance, the adversarial examples are crafted with $\epsilon = 2/255$ and $\epsilon = 4/255$ perturbation levels. In the test phase, the adversarial attacks are produced in a white-box and black-box

Table 1: Test datasets utilized in our experiments.

Dataset Number	Name	Morph Images	Bona fide Images
1	AMSL [36]	2175	204
2	FRLL AMSL [42]	6525	204
3	FRLL Webmorpher [42]	1221	204
4	FRLL OpenCV [42]	1221	204
5	FRLL StyleGAN [42]	1221	204
6	FRLL Facemorpher [42]	1221	204
7	FERET OpenCV [42]	529	1413
8	FERET StyleGAN [42]	529	1413
9	FERET Facemorpher [42]	529	1413
10	FRGC OpenCV [42]	964	3038
11	FRGC StyleGAN [42]	964	3038
12	FRGC Facemorpher [42]	964	3038
13	FRGC MIPGAN [53]	747	373
14	FRGC MIPGAN + PRINT AND SCAN [53]	747	376

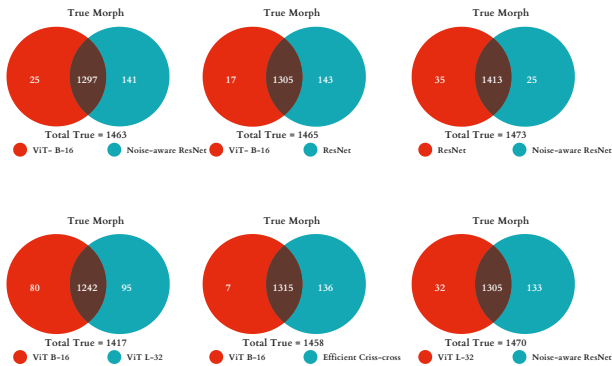


Figure 3: The number of morphs images that are correctly classified as the morph class.

settings with translation-invariant FGSM (TIFGSM) [20], Carlini and Wagner (C&W) [7], momentum iterative FGSM (MIFGSM) [19], Square [5], Diverse Inputs Iterative FGSM (DIFGSM) [50]. In the black-box settings, the adversarial images are crafted using a black-box model and transfer to the ensemble model. Our experiments in the training and tests phases are conducted on three 12 GB TITAN X (Pascal) GPUs.

4.2. Results and Analysis

To explore the generalization capability of our ensemble model on a wide range of domains, we employ FERET [39], FRLL [18, 42], FRGC [40], and AMSL [36] datasets in our evaluations. Further, different landmark-based and GAN-based morphing attacks are used to generate morph images on the datasets. The landmark-based attacks include Facemorpher [42], OpenCV [42], and WebMorph [42] and GAN-based attacks include MIPGAN [53], StyleGAN2 [42], and Print and Scan attacks [53]. Table 1 presents the details of different test sets that are used in our experiments.

Single Models: To create our ensemble model, several single models based on the CNN and Transformer architectures are trained individually using the source domain. To demonstrate that the error distribution in morph detection is not the same between different models and the proposed ensemble model contributes to the overall performance of morph detection, we apply each single model to the morph images created by the StyleGAN attack [42]. Figure 3 represents the Venn diagram of morph detection between the ViT B-16, ViT L-32, Inception-ResNet-V1 (abbreviated as ResNet), Noise-aware ResNet, and EffitientNet Criss-cross models. In the EffitientNet Criss-cross model, the Criss-cross attention [28] is integrated into the EffitientNet architecture [46]. The number of true morphs detected by each single model and both of them is provided. The results reveal that the errors are not totally correlated between the single models and they do not perform similarly in different parts of feature space. This demonstrates that the ensemble model, which is composed of different single models, can promote morph detection performance and generalize better to different domains. To select the best models in the ensemble model, we gauge the performance of the single models on 14 different test sets (Table 1). The comparison results in terms of AUC and APCER (@BPCER=0.1) are reported in Table 2. The results show that in some target domains such as #1, #2, #4, and #6, single models generally perform satisfactorily. However, in most cases such as target datasets #3, #5, #7, #8, #9, #10, #11, #12, and #13, the single models perform differently. For instance, ViT B-16, ResNet, and Noise-aware ResNet models significantly outperform the other single models in target datasets #3, #5, #9, respectively. This analysis indicates that an ensemble model constructed from these single models could potentially gain noticeable improvements in comparison with the single models, mainly for the challenging target domains.

4.3. Ensemble Model

To find the best configuration of the ensemble model, we conduct two ablation studies as in Table 3. In the first study, the number of single models, as well as their combinations, are ablated in the ensemble model. The soft voting strategy is adopted to fuse different single models. More specifically, the probability scores of the single models are averaged in the ensemble model. For the sake of brevity, only the best combinations are reported in Table 3. It is observed that naively averaging the probability scores of the single models could generally boost the generalization of the ensemble model to different target domains. This is largely ascribed to the lower variance in the ensemble models compared to the single models. In addition, by drawing a comparison between different configurations of the ensemble model, we observe that the ViT B-16 + ViT L-32 + N-ResNet [22, 49] combination generally yields the best per-

Table 2: Morph detection results for different single models in terms of AUC, APCER (@BPCER=1%), and BPCER (@APCER=1%) metrics. N-ResNet denotes the noise-aware ResNet method [49].

	Methods	AMSL	FRLL AMSL	FRLL Webomapper	FRLL OpenCV	FRLL StyleGAN	FRLL Faceomapper	FERET OpenCV	FERET StyleGAN	FERET Faceomapper	FRGC OpenCV	FRGC StyleGAN	FRGC Faceomapper	FRGC MPCGAN	FRGC MPCGAN (PRNT)
AUC	ResNet [26]	99.32	98.77	79.09	99.99	99.16	99.94	94.14	92.7	94.31	97.24	92.28	96.36	81.24	68.45
	N-ResNet [26, 49]	99.63	99.46	87.4	99.95	97.71	99.83	94.51	95.02	95.88	99.65	97.07	99.61	91.68	73.91
	EfficientNet [46, 28]	99.95	99.73	83.52	99.85	90.71	99.80	89.01	90.88	92.97	92.64	68.75	92.07	95.66	75.87
	ViT B-16 [22]	99.62	99.41	90.70	99.84	86.37	99.61	93.35	90.77	93.38	99.01	87.77	97.54	88.75	83.56
	ViT L-32[22]	99.20	99.09	85.81	99.69	93.38	99.52	92.76	91.97	92.69	98.75	89.46	97.31	80.00	82.16
APCER	ResNet [26]	6.86	8.82	72.54	00.0	7.35	00.0	43.94	50.74	43.38	17.70	39.4	21.95	84.182	91.48
	N-ResNet [26, 49]	6.37	8.33	76.96	00.0	24.01	1.96	54.91	36.51	43.94	2.40	23.86	2.50	63.80	62.23
	EfficientNet [46, 28]	0.98	3.43	58.33	1.47	53.92	1.47	67.72	55.41	53.07	37.12	61.68	39.10	30.29	82.71
	ViT B-16 [22]	5.74	8.85	75.51	1.47	76.92	4.09	39.50	66.91	38.75	15.04	74.68	29.77	70.95	74.43
	ViT L-32 [22]	8.78	8.70	80.75	2.04	49.09	4.50	63.13	64.83	65.97	21.16	76.55	36.41	76.17	89.29
BPCER	ResNet [26]	24.55	36.78	98.44	0.32	27.00	0.40	44.23	68.62	46.88	81.63	93.98	89.73	98.39	98.25
	N-ResNet [26, 49]	5.97	7.75	70.35	0.0	30.76	1.71	39.69	61.62	35.91	14.00	70.33	16.39	94.64	99.59
	EfficientNet [46, 28]	1.79	8.93	95.08	2.62	72.42	2.86	65.97	71.64	57.08	79.56	99.37	80.80	93.17	90.62
	ViT B-16 [22]	6.86	6.86	55.39	2.94	77.94	6.37	61.07	66.31	56.40	17.281	69.32	31.764	71.04	89.89
	ViT L-32 [22]	13.72	20.09	69.11	6.37	57.84	9.80	44.16	62.42	40.90	20.17	70.53	41.70	89.81	86.43

Table 3: Morph detection results for different combinations of ensemble models in terms of APCER (@BPCER=1%) and AUC metrics. N-ResNet denotes the noise-aware ResNet method [49].

	Methods	AMSL	FRLL AMSL	FRLL Webomapper	FRLL OpenCV	FRLL StyleGAN	FRLL Faceomapper	FERET OpenCV	FERET StyleGAN	FERET Faceomapper	FRGC OpenCV	FRGC StyleGAN	FRGC Faceomapper	FRGC MPCGAN	FRGC MPCGAN (PRNT)
AUC	ViT B-16 + ResNet	99.70	99.45	90.97	99.97	96.76	99.90	94.42	93.96	94.48	99.01	94.36	98.35	89.09	84.49
	N-ResNet + ResNet	99.74	99.54	87.08	99.97	99.02	99.91	94.85	94.92	95.64	98.93	96.57	98.53	87.83	80.01
	ViT L-32 + ResNet	99.69	99.39	87.39	100	98.19	99.94	95.22	94.75	95.21	98.70	93.89	97.90	84.02	80.94
	ViT B-16 + ViT L-32	99.92	99.95	91.57	99.94	92.59	99.88	94.51	92.75	94.69	99.32	89.85	98.02	85.22	84.54
	ViT B-16 + ViT L-32 + N-ResNet	99.91	99.87	94.10	99.97	97.47	99.89	95.45	94.90	96.08	99.84	95.58	99.47	90.23	85.61
	ViT B-16 + ViT L-32 + EfficientNet	99.97	99.92	89.60	99.96	93.53	99.90	93.89	93.35	94.85	99.40	89.21	98.53	92.30	85.89
	ResNet+ N-ResNet + EfficientNet	99.85	99.75	87.78	99.95	97.91	99.88	93.89	94.43	95.27	99.36	94.91	99.12	92.87	86.57
	ViT B-16 + ViT L-32 + EfficientNet +ResNet	99.73	99.52	89.74	99.98	96.47	99.91	94.43	94.32	95.23	99.43	92.42	98.90	91.98	85.95
	ViT B-16 + ViT L-32 + EfficientNet + ResNet +N-ResNet	99.87	99.78	92.22	99.97	97.71	99.90	94.85	94.99	95.85	99.65	94.90	99.35	92.69	86.16
	ViT B-16 + ResNet	2.34	5.48	76.41	0.08	32.14	0.24	31.75	50.28	35.16	17.84	59.23	29.46	76.84	80.05
APCER	N-ResNet + ResNet	3.03	5.14	68.38	00.0	14.89	00.0	28.73	37.05	28.73	17.53	48.13	26.86	61.84	92.77
	ViT L-32 + ResNet	5.24	11.55	86.89	0.16	8.26	0.24	33.64	43.85	37.42	23.65	58.50	35.16	78.44	89.55
	ViT B-16 + ViT L-32	1.19	1.57	61.75	0.40	55.31	0.73	39.69	60.86	39.50	11.72	73.96	26.76	73.49	74.69
	ViT B-16 + ViT L-32 + N-ResNet	1.74	3.50	54.62	0.32	19.80	0.65	31.75	46.69	30.62	2.17	51.14	8.92	63.58	74.29
	ViT B-16 + ViT L-32 + EfficientNet	0.22	0.91	75.83	0.24	52.37	0.32	38.75	55.95	37.80	9.647	75.93	21.16	70.14	74.02
	ResNet+ N-ResNet + EfficientNet	1.97	4.16	77.72	0.16	24.87	0.32	32.51	39.13	29.86	10.06	49.37	17.21	61.17	86.57
	ViT B-16 + ViT L-32 + EfficientNet +ResNet	0.09	0.41	73.30	0.16	29.78	0.24	34.21	46.31	34.59	7.36	55.60	16.49	73.62	72.28
	ViT B-16 + ViT L-32 + EfficientNet + ResNet +N-ResNet	1.97	4.30	80.01	0.16	29.21	0.32	33.08	39.5	29.86	3.73	48.44	7.98	58.63	72.28

formance among the ensemble models with two and three components. This evaluation also points out that the larger number of single models would not necessarily boost the generalization of the final ensemble model, and may only incurs more computational costs.

In the second study, different fusion strategies are ablated in Table 4. The fusion strategies include soft voting, feature-based super learner, and score-based super learner strategies. In our score-based super learner, we freeze the single models and train another learner, which is a single-layer FFN, to weigh the output scores of the single models. Also, in our feature-based super learner, the output features from the last FC layers of different single models are concatenated and fed to a two-layer FFN for final morph classification (Figure 2)). Since the ViT B-16 + ViT L-32 +

N-ResNet model relatively outperforms the other configurations in Table 4, we set them as the components of the final selected ensemble model. The comparisons indicate that the feature-based super learner outperforms other fusion strategies to some extent. From this experiment, we can deduce that the ensemble model with ViT B-16, ViT L-32, N-ResNet, and the feature-based super learner components outperforms its competitor ensemble models.

4.4. Robustness

In this section, the robust accuracy of the selected ensemble model is targeted. For this objective, multi-perturbation adversarial training is employed. To benchmark the effectiveness of the model-based ensembling attack [32] in the adversarial training, we gauge the robustness of the ensem-

Table 4: Morph detection results for the selected ensemble model (ViT B-16 + ViT L-32 + N-ResNet) for different fusion strategies. The evaluations are in terms of APCER (@BPCER=1%) and AUC metrics.

Methods		AMSL	FRLL AMSL	FRLL Webmorpher	FRLL OpenCV	FRLL StyleGAN	FRLL Facemorpher	FERET OpenCV	FERET StyleGAN	FERET Facemorpher	FRGC OpenCV	FRGC StyleGAN	FRGC Facemorpher	FRGC MIPGAN	FRGC MIPGAN (PRINT)
AUC	Soft Voting	99.91	99.87	94.10	99.97	97.47	99.89	95.45	94.90	96.08	99.84	95.58	99.47	90.23	85.61
	Max Voting	99.74	99.63	88.92	99.96	97.77	99.86	95.71	95.32	96.29	99.83	95.08	99.37	88.30	85.11
	Score-based Super Learner	99.81	99.73	93.22	99.96	97.82	99.87	95.36	94.99	96.40	99.86	96.73	99.63	91.29	86.75
	Feature-based Super Learner	99.91	99.83	92.08	99.98	98.08	99.89	95.83	95.78	96.70	99.78	96.48	99.69	91.86	85.81
APCER	Soft Voting	1.74	3.50	54.62	0.32	19.80	0.65	31.75	46.69	30.62	2.17	51.14	8.92	63.58	74.29
	Max Voting	1.93	3.69	53.89	0.32	20.45	0.65	31.37	46.69	30.62	2.17	51.65	9.85	63.58	74.29
	Score-based Super Learner	3.63	5.08	62.98	0.32	29.78	1.22	31.19	46.88	30.43	1.45	44.19	5.39	56.35	69.61
	Feature-based Super Learner	1.70	3.21	60.19	0.16	26.43	0.40	28.16	40.26	24.76	1.97	47.09	3.42	55.55	72.04

Table 5: Comparison of the selected ensemble model and the robust ensemble model with adversarial training. The robust accuracy against different adversarial attacks is in terms of AUC/APCER@BPCER=1% metrics on a subset of FRGC MIPGAN. In black-box attacks, stronger perturbation levels are utilized to generate adversarial perturbations in comparison to white-box attacks.

Target		DIFGSM [50]	MIFGSM [19]	TIFGSM [20]	TPGD [54]	Square [5]	C&W [7]
White-Box	Ensemble Model	84.60 / 93.75	80.67 / 96.875	83.26 / 95.08	71.39 / 97.32	74.17 / 100	74.80 / 100
	Robust Ensemble Model	88.87 / 87.05	91.82 / 59.37	89.43 / 84.87	96.22 / 30.35	94.04 / 59.37	91.82 / 59.37
Black-Box	Ensemble Model	32.0 / 100	49.9 / 99.5	15.2 / 100	80.4 / 44.6	91.8 / 78.1	86.8 / 93.7
	Robust Ensemble Model	98.0 / 28.1	97.6 / 30.8	97.4 / 47.3	98.6 / 19.2	92.9 / 54.9	91.5 / 59.8

Table 6: Comparison of the selected robust ensemble model with state-of-the-art morph detection models in the FRLL and the LMA-DRD [16] datasets. The results are in terms of EER, and BPCER (@APCER=1% and 10%) metrics.

Target		D-EER	BPCER (1%)	BPCER (10%)
FRLL	MixFacenet - SMDD [14]	3.87	23.53	0.49
	PW-MAD - SMDD [14]	2.20	26.47	0.49
	Inception - SMDD [14]	3.17	30.39	0.49
	Denosing based method [49]	1.96	5.39	00.0
	Ensemble Model	0.98	0.98	00.0
	Robust Ensemble Model	0.98	0.98	00.0
LMA-DRD	MixFacenet - SMDD [14]	19.42	79.67	31.71
	PW-MAD - SMDD [14]	17.06	80.49	25.20
	Inception - SMDD [14]	15.11	69.92	30.89
	Ensemble Model	13.64	72.35	17.07

ble model that has been adversarially trained. In the robustness evaluations, we utilize new adversarial attacks of which the ensemble model is not aware. Based on Table 5, the performance of the ensemble model against unseen adversarial attacks in the white-box and black-box settings drops significantly. It is observed that the robust ensemble model gains substantial improvements over the baseline ensemble model against DIFGSM [50], MIFGSM [19], TIFGSM [20], TPGD [54], Square [5] C&W [7] at-

tacks in both white-box and black-box settings. In short, taking these results into account, we can substantiate that the multi-perturbation adversarial training with the model-based ensembling optimization [32] improves the robust accuracy of our final model against several adversarial attacks in white-box and black-box settings. In the last experiment, we make a comparison between the proposed robust ensemble model and various state-of-the-art studies. This experiment is also carried out to find out whether the multi-perturbation adversarial training detrimentally hurts the clean accuracy of the ensemble model or not. The results in Table 6 demonstrate that the proposed robust ensemble model maintains its superior performance on clean accuracy and also significantly surpasses the state-of-the-art studies.

5. Conclusions

In this paper, we present a morph attack detection with strong generalization ability to different morph attacks and high robustness against adversarial attack. By combining CNN and Transformer models in our ensemble model, we capture both long-term and short-term relationships in morph images to improve domain generalization. To strengthen the robustness of our model against adversarial attacks, we employ multi-perturbation adversarial training with highly transferable adversarial examples. Experimental results on several datasets demonstrate the generalization ability and adversarial robustness of our proposed model.

Acknowledgements. This material is based upon a work supported by the Center for Identification Technology Research and the National Science Foundation under Grant 1650474.

References

- [1] P. Aghdaie, B. Chaudhary, S. Soleymani, J. Dawson, and N. M. Nasrabadi. Attention aware wavelet-based detection of morphed face images. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–8, 2021.
- [2] P. Aghdaie, B. Chaudhary, S. Soleymani, J. Dawson, and N. M. Nasrabadi. Detection of morphed face images using discriminative wavelet sub-bands. In *2021 IEEE International Workshop on Biometrics and Forensics (IWBf)*, pages 1–6, 2021.
- [3] P. Aghdaie, B. Chaudhary, S. Soleymani, J. Dawson, and N. M. Nasrabadi. Morph detection enhanced by structured group sparsity. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 311–320, 2022.
- [4] N. Alipour and R. P. Hasanzadeh. Superpixel-based brain tumor segmentation in mr images using an extended local fuzzy active contour model. *Multimedia Tools and Applications*, 80(6):8835–8859, 2021.
- [5] M. Andriushchenko, F. Croce, N. Flammarion, and M. Hein. Square attack: a query-efficient black-box adversarial attack via random search. In *Proceedings of the European Conference on Computer Vision*, pages 484–501, 2020.
- [6] S. Banerjee and A. Ross. Conditional identity disentanglement for differential face morph detection. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–8, 2021.
- [7] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57, 2017.
- [8] J. Cha, H. Cho, K. Lee, S. Park, Y. Lee, and S. Park. Domain generalization needs stochastic weight averaging for robustness on domain shifts. *arXiv e-prints*, pages arXiv:2102, 2021.
- [9] 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.
- [10] Q. Chen, J. Xu, and V. Koltun. Fast image processing with fully-convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2497–2506, 2017.
- [11] F. Croce and M. Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on Machine Learning*, pages 2206–2216, 2020.
- [12] J. Cui, S. Liu, L. Wang, and J. Jia. Learnable boundary guided adversarial training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15721–15730, 2021.
- [13] A. Dabouei, S. Soleymani, F. Taherkhani, J. Dawson, and N. Nasrabadi. Smoothfool: An efficient framework for computing smooth adversarial perturbations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2665–2674, 2020.
- [14] N. Damer, C. A. F. López, M. Fang, N. Spiller, M. V. Pham, and F. Boutros. Privacy-friendly synthetic data for the development of face morphing attack detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2022.
- [15] N. Damer, A. M. Saladie, S. Zienert, Y. Wainakh, P. Terhöst, F. Kirchbuchner, and A. Kuijper. To detect or not to detect: The right faces to morph. In *2019 International Conference on Biometrics (ICB)*, pages 1–8, 2019.
- [16] N. Damer, N. Spiller, M. Fang, F. Boutros, F. Kirchbuchner, and A. Kuijper. Pw-mad: Pixel-wise supervision for generalized face morphing attack detection. In *International Symposium on Visual Computing*, pages 291–304, 2021.
- [17] L. DeBruine. *debruine/webmorpher*: Beta release 2, January 2018.
- [18] L. DeBruine and B. Jones. Face research lab london set, May 2017.
- [19] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9185–9193, 2018.
- [20] Y. Dong, T. Pang, H. Su, and J. Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4312–4321, 2019.
- [21] Y. Dong, H. Su, B. Wu, Z. Li, W. Liu, T. Zhang, and J. Zhu. Efficient decision-based black-box adversarial attacks on face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7714–7722, 2019.
- [22] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.
- [23] M. Ganaie, M. Hu, et al. Ensemble deep learning: A review. *Neural Networks*, 2021.
- [24] I. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.
- [25] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.
- [26] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.
- [27] Q. Huang, I. Katsman, H. He, Z. Gu, S. Belongie, and S.-N. Lim. Enhancing adversarial example transferability with an intermediate level attack. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4733–4742, 2019.
- [28] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu. CCNet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 603–612, 2019.

- [29] S. A. Kilany, A. Mahfouz, A. M. Zaki, and A. Sayed. Analysis of adversarial attacks on face verification systems. In *The International Conference on Artificial Intelligence and Computer Vision*, pages 463–472, 2021.
- [30] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.
- [31] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial examples in the physical world. *ICLR Workshop*, 2017.
- [32] Y. Liu, X. Chen, C. Liu, and D. Song. Delving into transferable adversarial examples and black-box attacks. In *International Conference on Learning Representations*, 2017.
- [33] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.
- [34] S. Mosharafian, M. Razzaghpour, Y. P. Fallah, and J. M. Velni. Gaussian process based stochastic model predictive control for cooperative adaptive cruise control. In *2021 IEEE Vehicular Networking Conference (VNC)*, pages 17–23, 2021.
- [35] M. Mostofa, S. Mohamadi, J. Dawson, and N. M. Nasrabadi. Deep gan-based cross-spectral cross-resolution iris recognition. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(4):443–463, 2021.
- [36] T. Neubert, A. Makrushin, M. Hildebrandt, C. Kraetzer, and J. Dittmann. Extended stirtrace benchmarking of biometric and forensic qualities of morphed face images. *IET Biometrics*, 7(4):325–332, 2018.
- [37] H. Nodehi and A. Shahbahrami. Multi-metric re-identification for online multi-person tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(1):147–159, 2021.
- [38] M. Nourelahi, F. Dadboud, H. Khalili, A. Niakan, and H. Parsaei. A machine learning model for predicting favorable outcome in severe traumatic brain injury patients after 6 months. *Acute and critical care*, 37(1):45–52, 2022.
- [39] P. Phillips, H. Wechsler, J. R. Huang, and P. J. Rauss. The feret database and evaluation procedure for face-recognition algorithms. *Image and Vision Computing*, 16:295–306, 1998.
- [40] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the face recognition grand challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, volume 1, pages 947–954, 2005.
- [41] A. Quek. Facemorpher, Jan 2019.
- [42] E. Sarkar, P. Korshunov, L. Colbois, and S. Marcel. Vulnerability analysis of face morphing attacks from landmarks and generative adversarial networks. *arXiv preprint*, Oct. 2020.
- [43] 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, 2017.
- [44] 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.
- [45] 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.
- [46] M. Tan and Q. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114, 2019.
- [47] F. Tramèr and D. Boneh. Adversarial training and robustness for multiple perturbations. *Advances in Neural Information Processing Systems*, 32, 2019.
- [48] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations*, 2018.
- [49] 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.
- [50] C. Xie, Z. Zhang, Y. Zhou, S. Bai, J. Wang, Z. Ren, and A. L. Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2730–2739, 2019.
- [51] K. Xu, M. Zhang, J. Li, S. S. Du, K.-I. Kawarabayashi, and S. Jegelka. How neural networks extrapolate: From feedforward to graph neural networks. In *International Conference on Learning Representations*, 2021.
- [52] Y. Xu, Q. Zhang, J. Zhang, and D. Tao. Vitae: Vision transformer advanced by exploring intrinsic inductive bias. *Advances in Neural Information Processing Systems*, 34, 2021.
- [53] H. Zhang, S. Venkatesh, R. Ramachandra, K. Raja, N. Damer, and C. 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.
- [54] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. El Ghaoui, and M. Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on Machine Learning*, pages 7472–7482, 2019.
- [55] 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.
- [56] Y. Zhong and W. Deng. Towards transferable adversarial attack against deep face recognition. *IEEE Transactions on Information Forensics and Security*, 16:1452–1466, 2020.
- [57] K. Zhou, Y. Yang, Y. Qiao, and T. Xiang. Domain adaptive ensemble learning. *IEEE Transactions on Image Processing*, 30:8008–8018, 2021.