Robust Ensemble Morph Detection with Domain Generalization

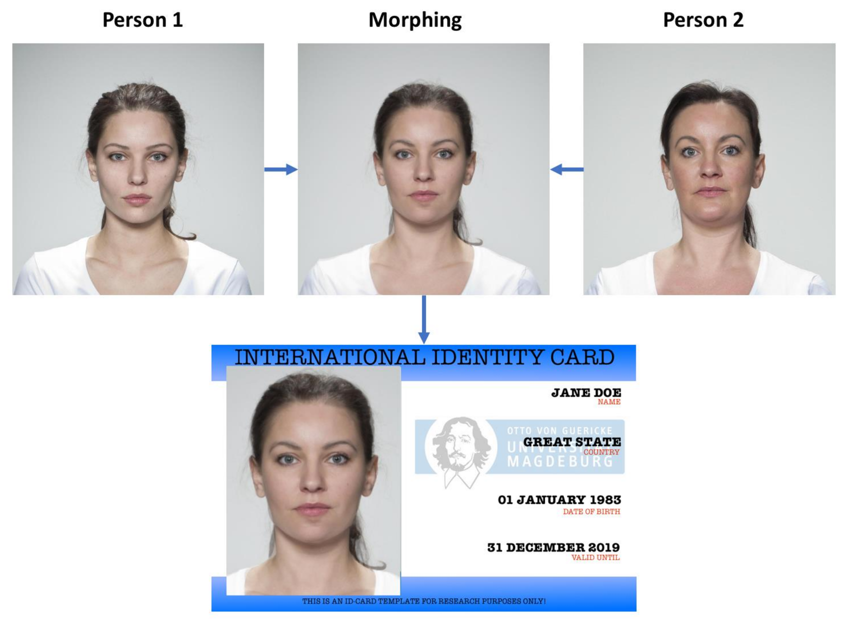
WestVirginiaUniversity
BENJAMIN M. STATLER COLLEGE OF ENGINEERING AND MINERAL RESOURCES

Hossein Kashiani, Shoaib Meraj Sami, Sobhan Soleymani, Nasser M. Nasrabadi
Lane Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, WV

CITeR NSF

Problem Statement

Face morphing is an image manipulation where two faces are blended together. At the time of passport enrollment, the passport photo can be easily manipulated with a morphing attack without the requirement of advanced forgery.



Motivation

Reliable detection of morphed face images can reduce vulnerability especially in highly secured applications including border control.

Contribution

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.

Challenges

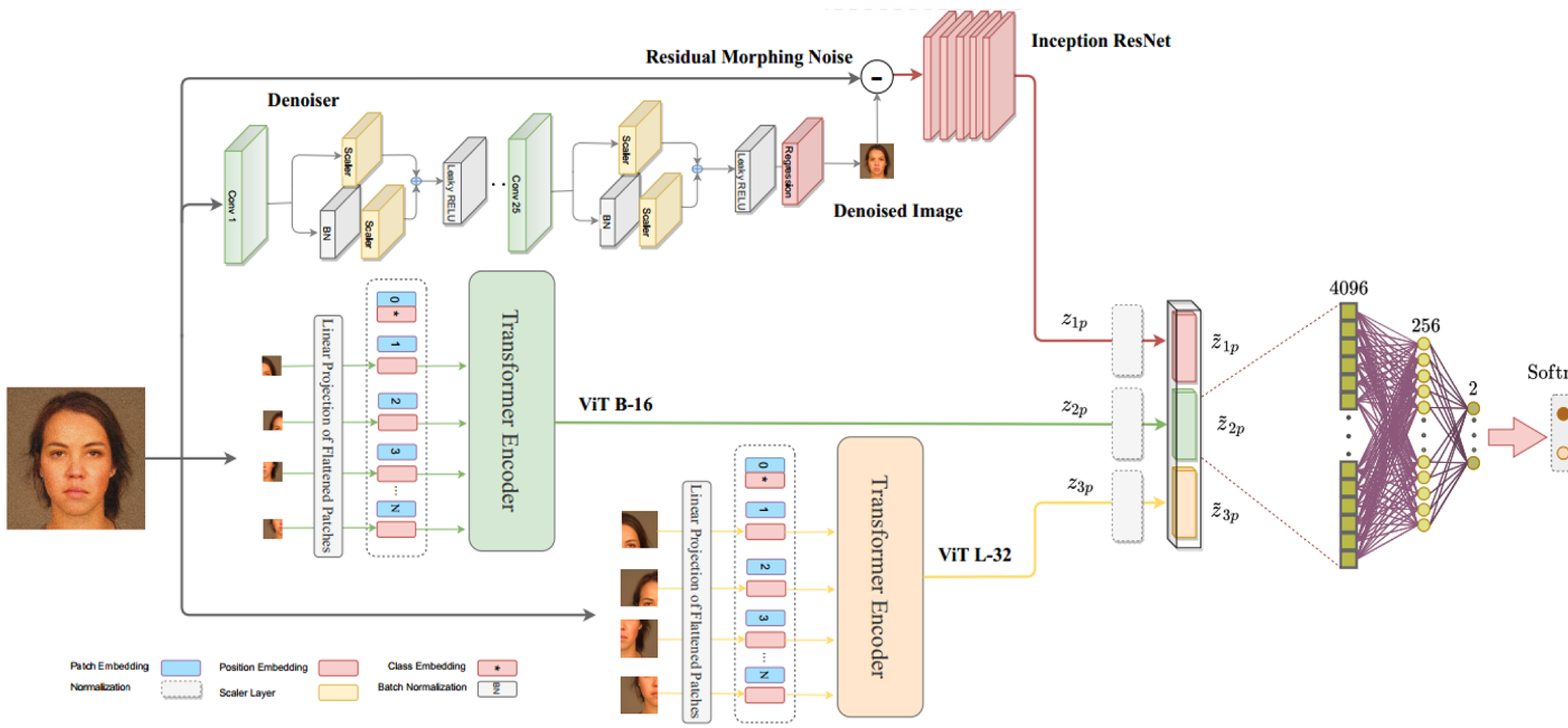
- There exist large domain shifts between different morph attacks.

- Face recognition systems are also known to be susceptible to adversarial examples.

Proposed Method

- We propose an ensemble morph attack detection model which highly generalizes to a wide range of morphing attacks. It incorporates the inductive bias of CNNs and long-range dependencies in Transformers to include the strengths of both CNN and Transformer architectures at the same time.
- Since the RGB residual morphing noise is effective for morph detection, we learn a CNN denoiser to calculate the residual artifacts for morph attack detection task.
- To construct our ensemble model, we train a feed-forward network to compute the matching scores of all single models in the fusion phase.
- To improve the effectiveness of our adversarial training, we craft the adversarial examples with high transferability using the model-based ensembling attack as follows:

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



Generalization Results

- We try different fusion strategies for our ensemble model that include soft voting, feature-based super learner, and score-based super learner strategies.
- From this experiment, we can deduce that the ensemble model with ViT B-16 [1], ViT L-32 [1], N-ResNet [2], and the feature-based super learner components outperforms its competitor models on different unseen test sets.

Table 1, Morph detection results for different fusion strategies (AUC)

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)
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
ResNet	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	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	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	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	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

Evaluation

- To explore the generalization capability of our ensemble model, we benchmark it on a wide range of unseen target domains. It includes FERET [3], FRLL [4], FRGC [5], and AMSL [6] datasets with different landmark-based and GAN-based morphing attacks. The landmark-based attacks include Facemorpher [4], OpenCV [4], and WebMorph [4] and GAN-based attacks include MIPGAN [7], StyleGAN2, and Print and Scan attack.
- In the robustness evaluation, we utilize new adversarial attacks in a white-box and black-box settings.

Robustness and SOTA Results

- The comparison results demonstrate that the proposed robust ensemble model maintains its superior performance on clean accuracy and also significantly surpasses the state-of-the-art studies.
- It is observed that the robust ensemble model gains substantial improvements over the baseline ensemble model against different adversarial attack in white-box and black-box settings.

Table 3, Comparison results with different studies on FRLL test set

Target	D-EER	BPCER (1%)	BPCER (10%)
MixFacenet - SMDD	3.87	23.53	0.49
PW-MAD - SMDD	2.20	26.47	0.49
Inception - SMDD	3.17	30.39	0.49
Denoising based method	1.96	5.39	00.0
Ensemble Model	0.98	0.98	00.0
Robust Ensemble Model	0.98	0.98	00.0

Table 2, Morph detection results against different adversarial attacks (AUC)

	Target	DIFGSM	MIFGSM	TIFGSM	TPGD	Square	C&W
White-Box	Ensemble Model	84.60	80.67	83.26	71.39	74.17	74.80
	Robust Ensemble Model	88.87	91.82	89.43	96.22	94.04	91.82
Black-Box	Ensemble Model	32.0	49.9	15.2	80.4	91.8	86.8
	Robust Ensemble Model	98.0	97.6	97.4	98.6	92.9	91.5

References

1. A. Dosovitskiy, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In International Conference on Learning Representations, 2020
2. C. Szegedy, et al. Inception-v4, inception-resnet and the impact of residual connections on learning. In Thirty-first AAAI Conference on Artificial Intelligence, 2017.
3. P. Phillips, et al. The feret database and evaluation procedure for face-recognition algorithms. Image and Vision Computing, 16:295–306, 1998.
4. E. Sarkar, et al. Vulnerability analysis of face morphing attacks from landmarks and generative adversarial networks. arXiv preprint, Oct. 2020.
5. P. J. Phillips, et al. 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.
6. T. Neubert, et al. Extended stirtrace benchmarking of biometric and forensic qualities of morphed face images. IET Biometrics, 7(4):325–332, 2018.
7. H. Zhang, et al. 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.