Cleaning Adversarial Perturbations Via Residual Generative Network For Face Verification

Yuying Su, Guangling Sun, Weiqi Fan, Xiaofeng Lu and Zhi Liu
School of Communication and Information Engineering,
Shanghai University, Shanghai 200444, China

Introduction

Recent researches show that Deep Neural Networks (DNNs) are vulnerable to adversarial perturbations injected into input samples. We investigated a defense method for **face verification**: a deep **Residual Generative Network** (ResGN) is learned to clean adversarial perturbations. The proposed novel training framework is composed of ResGN, pre-trained VGG-Face network and FaceNet network.

The parameters of ResGN are optimized by minimizing a **joint loss** consisting of a **pixel loss**, a **texture loss** and a **verification loss**, in which they measure content errors, subjective visual perception errors and verification task errors between cleaned image and legitimate image respectively. Specially, the latter two are provided by VGG-Face and FaceNet respectively and have essential contributions for improving verification performance of cleaned image.

Attack

Given pairs of facial images labeled 'same person' or 'not same person', a face verification system will determine whether the pair of facial images is 'same person' or 'not same person'.

Obviously, adversarial facial images can confuse the verification system to output an incorrect verification result.

Dodging and **impersonation** are two types of attacks: given an input facial image, dodging intends to make it identified as any one different from its genuine identity; impersonation intends to make it identified as a specified identity.

Results

Dataset: **Labeled Faces in the Wild (LFW)**

Table 1: The verification accuracy of original input examples and cleaned examples using the ResGN optimized with seven losses and Randomization.

Type / Loss	Dodging	Impersonation	Legitimate
pixel	99.16	99.68	90.52
texture	99.06	99.67	97.91
verification	88.36	95.31	98.31
pixel+ texture	99.25	99.82	97.86
pixel+ verification	99.26	99.83	98.70
texture+ verification	95.99	97.57	98.83
joint	**99.45**	**99.92**	**99.18**
Randomization	65.28	61.81	97.65
Original input Example	0.84	43.21	100

Methods

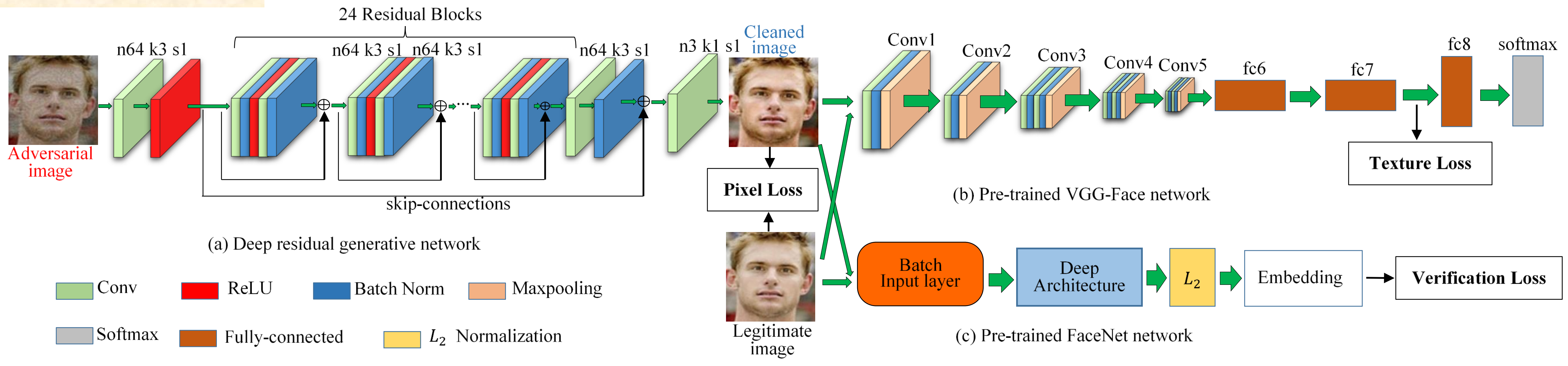

Fig.1: The proposed training framework for ResGN. Corresponding number of feature maps (n), kernel size (k) and stride (s) are marked for each convolutional layer in (a).

- The parameters of ResGN are optimized by minimizing a joint loss.

- The joint loss $\mathcal{L}_{joint}$ consists of a pixel loss, a texture loss and a verification loss and is the weighted summation of them.

$$\mathcal{L}_{joint} = \alpha * \mathcal{L}_{pixel} + \beta * \mathcal{L}_{texture} + \gamma * \mathcal{L}_{verification}. \qquad (1)$$

- Given an adversarial image, ResGN will output the cleaned image, the pixel loss $\mathcal{L}_{pixel}$ is calculated using the cleaned image I_L and corresponding legitimate image I_C.

$$\mathcal{L}_{pixel} = \frac{1}{N_p}\sum_{j=1}^{N_p}(\boldsymbol{C}_j^{I_C} - \boldsymbol{C}_j^{I_L})^2 , \qquad (2)$$

where N_p is the total number of pixels in each image.$\boldsymbol{C}_j^{I_C}$ *and* $\boldsymbol{C}_j^{I_L}$ denote the color of jth pixel in RGB space corresponding to I_C and I_L respectively.

- Similarly, the two images are both fed into VGG-Face and FaceNet network to calculate the texture loss $\mathcal{L}_{texture}$ and the verification loss $\mathcal{L}_{verification}$ respectively.

$$\mathcal{L}_{texture} = \frac{1}{N_t}\sum_{j=1}^{N_t}(\mathcal{F}_j^{I_C} - \mathcal{F}_j^{I_L})^2 , \qquad (3)$$

where $\boldsymbol{\mathcal{F}}^{I_C} \in \mathcal{R}^{N_t}$ and $\boldsymbol{\mathcal{F}}^{I_L} \in \mathcal{R}^{N_t}$ denote the feature maps of $I_C$and I_L,which are available from the second fully connected layer of VGG-Face. N_t is the dimension of the feature map and equals to the node number of the second fully connected layer.

$$\mathcal{L}_{verification} = -\sum_{i=0}^{K-1} p_i \cdot log(q_i) \qquad (4)$$

The K is the number of category. p denotes the actual label of pair of faces and q represents the output probability between legitimate image and cleaned image.

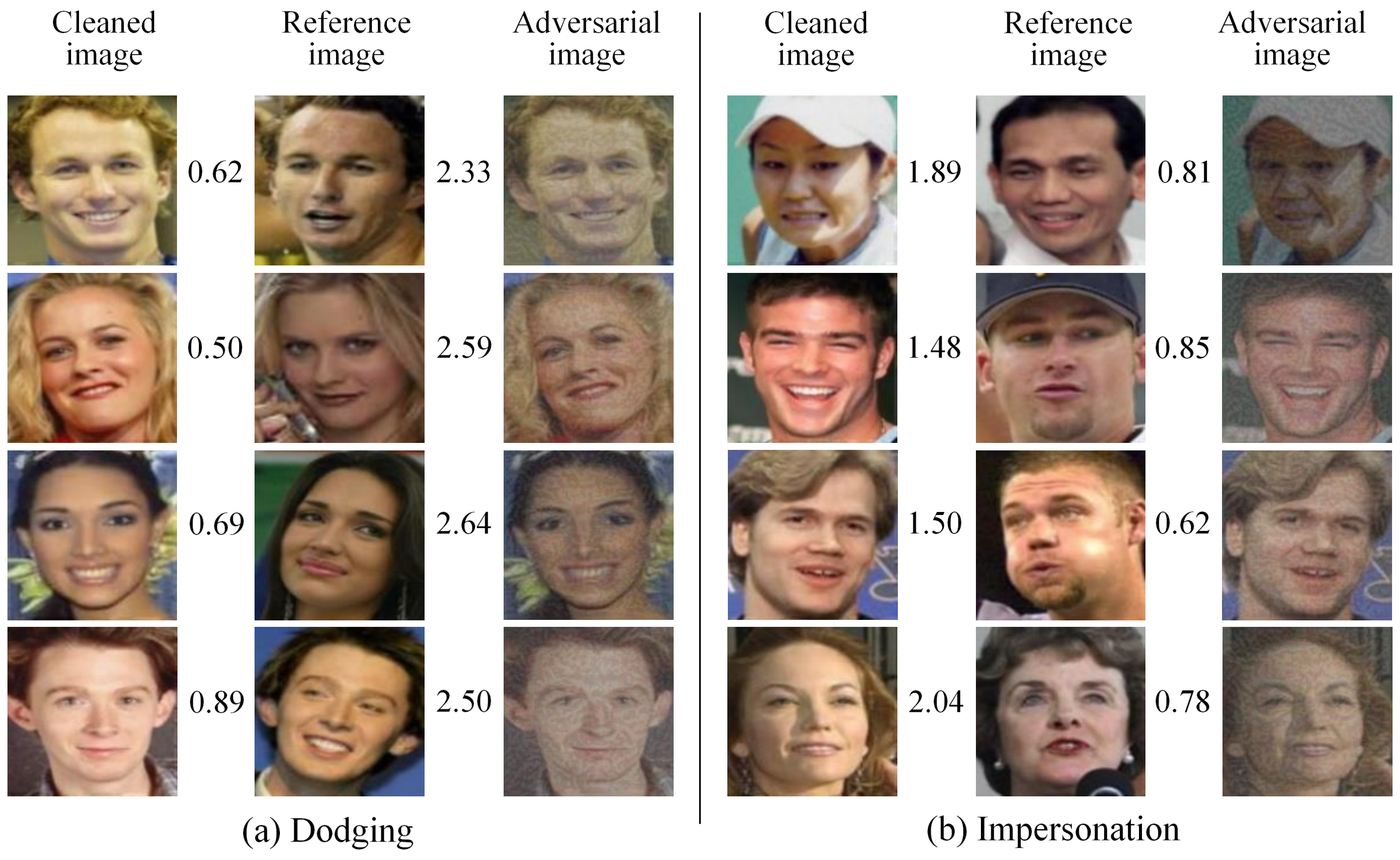

Fig.2: Adversarial examples of perturbation successfully cleaned. The value between two images is a Euclidean distance computed with FaceNet network.

Conclusion

- The proposed ResGN has a **flexible adaptivity** in that ResGN can incorporate with any pre-trained network applied to other face analysis task, such as face identification and facial attribute classification.

- In future, we will learn ResGN from more advanced adversarial examples and combine ResGN with other defensive technique to enhance security of image recognition system.

The link where you could search for this paper on IEEE Xplore is https://ieeexplore.ieee.org/document/8683327.
If you want to know more information, you could contact us. Email: sunguangling@shu.edu.cn, suyuyingshu@shu.edu.cn.