大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
National Institute of Informatics

National University
SOKENDAI
The Graduate University for Advanced Studies

Transformation on Computer–Generated Facial Image to Avoid Detection by Spoofing Detector

Huy H. Nguyen*, Ngoc-Dung T. Tieu, Hoang-Quoc Nguyen-Son, Junichi Yamagishi, and Isao Echizen

IEEE International Conference on Multimedia and Expo
July 23-27, 2018
San Diego, USA

Outline

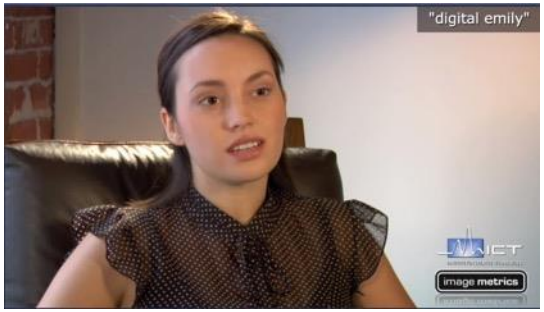
1. Motivation
2. Related Work
3. Proposed Method
4. Evaluation
5. Conclusion & Future Work

1. Motivation

1. Motivation

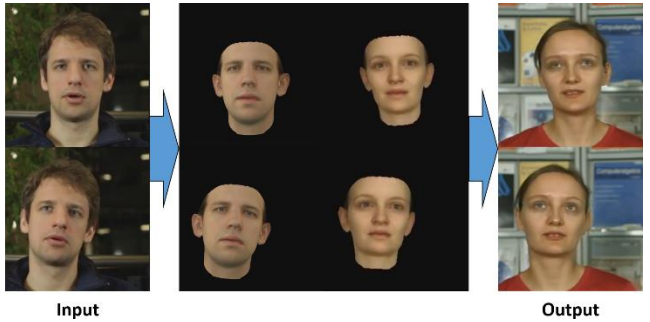
1.1. Presentation Attacks & Forgeries

Hard <Requirements for attackers to perform spoofing attacks> **Easy**



The Digital Emily Project [1] (SIGGRAPH 2008)

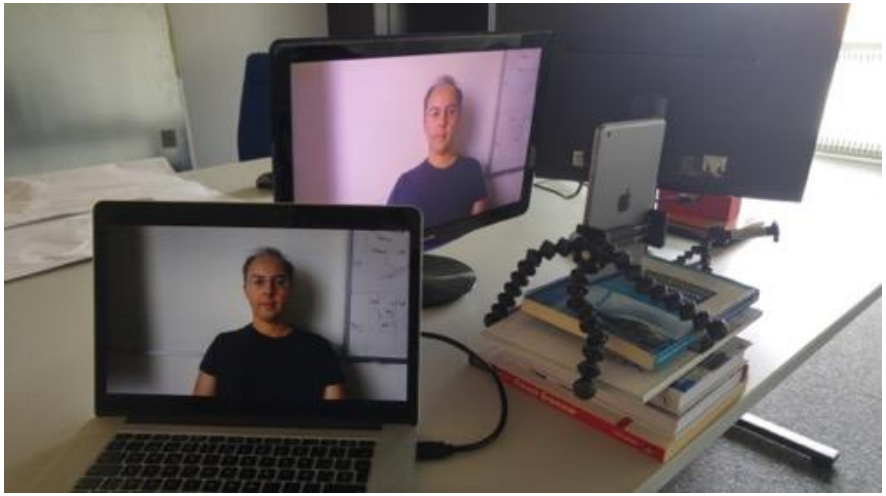
Face2Face: Real-time face capture & reenactment [2] (CVPR 2016)



DeepVideo Portraits = Face2Face + head poses [3] (SIGGRAPH 2018)

1. Motivation

1.1. Presentation Attacks & Forgeries



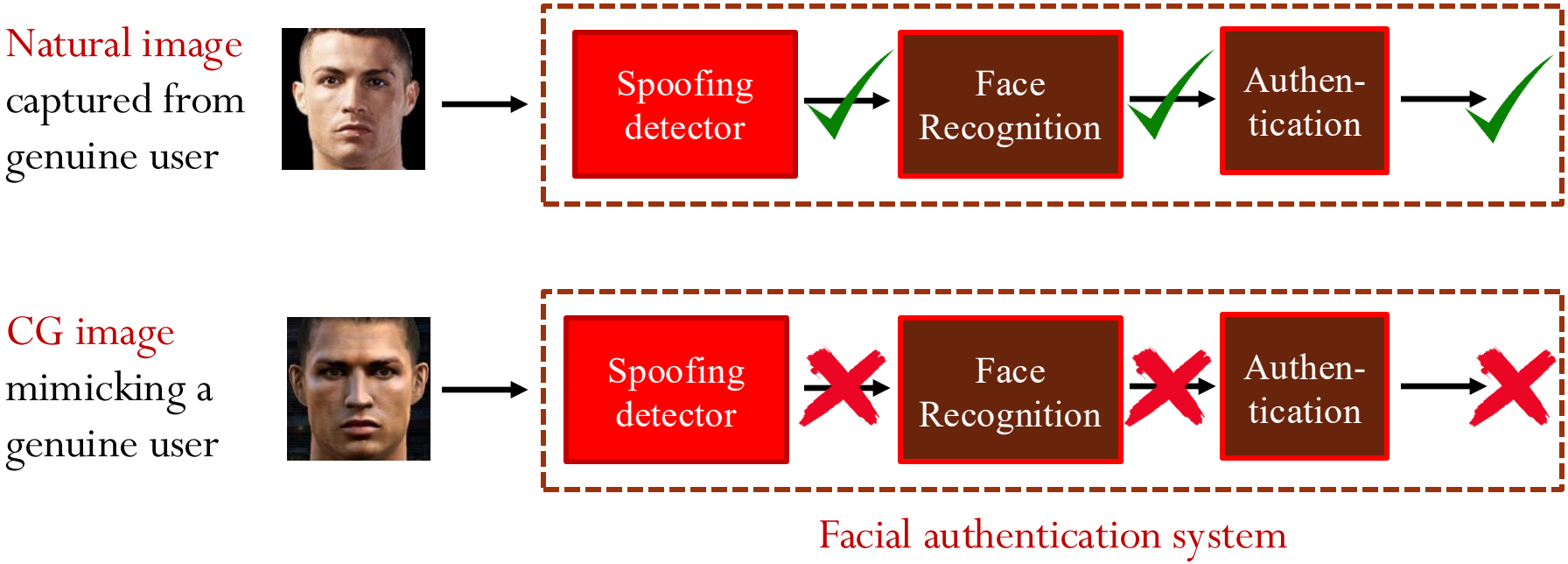
A presentation attack to break a facial authentication system
(Costa-Pazo et al., BioSig 2016)

Creating fake news/ Impersonation
(Thies et al., CVPR 2016)

1. Motivation

1.2. Spoofing Detectors

Spoofing detectors can be used to detect presentation attacks & forgeries.



1. Motivation

1.3. Proposed Method

Natural image captured from genuine user

Spoofing detector

CG image mimicking a genuine user

Spoofing detector

Transformation **(Proposed method)**

New image transformed from CG one

Spoofing detector

Existing spoofing detectors can be attacked !!!

2. Related Work

2. Related Work

2.1. Spoofing Detection

General images:

- Histogram of differential images (Wu et al.) [1]
- Multi-fractal and regression analysis (Peng et al.) [2]

Facial images:

- Smoothness property and local entropy (Nguyen et al.) [3]

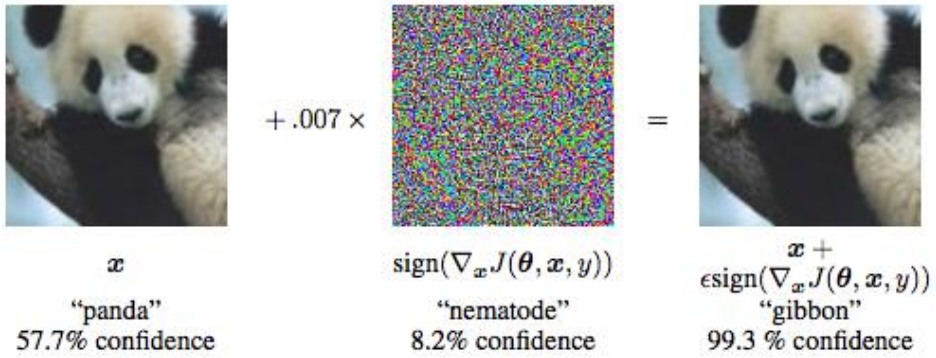
[1] Wu, Ruoyu, Xiaolong Li, and Bin Yang. "Identifying computer generated graphics via histogram features." ICIP 2011.

[2] Peng, Fei, et al. "Discrimination of natural images and computer generated graphics based on multi-fractal and regression analysis." AEU-International Journal of Electronics and Communications 71 (2017): 72-81.

[3] Echizen, Isao, et al. "Discriminating between computer-generated facial images and natural ones using smoothness property and local entropy." International Workshop on Digital Watermarking (IWDW) 2015.

2. Related Work

2.2. Adversarial Machine Learning



Goodfellow et al. "Explaining and harnessing adversarial examples." ICLR 2015.

Our work: Treating the spoofing detector as a **black-box**

→ No gradient information.

3. Proposed Method

3. Proposed method

3.1. Conditions

1. Spoofing detector is treated as a **black-box** (only knowing the classification output).
2. Transformed images must **preserve identity** of source images.
3. Natural images and CG counterparts are **not necessarily pairs**.

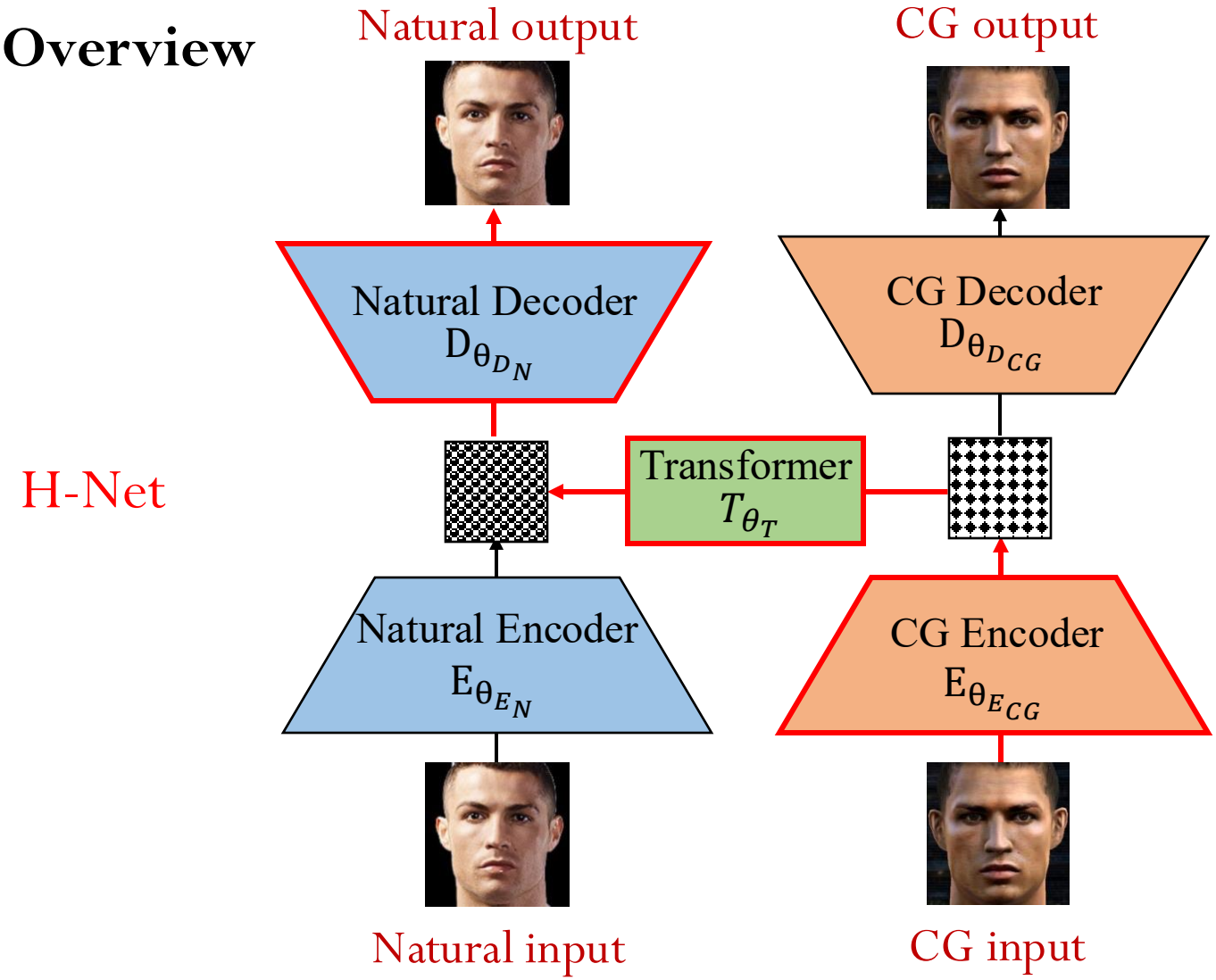
Ex:

Natural image

CG images

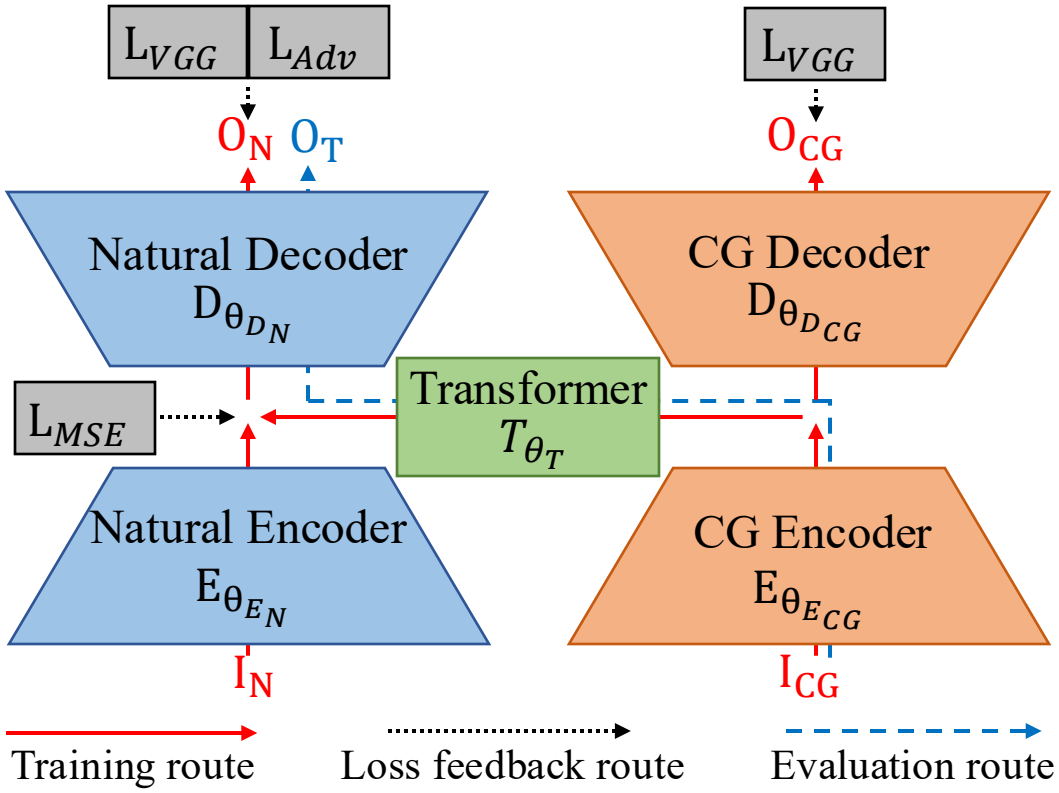
3. Proposed method

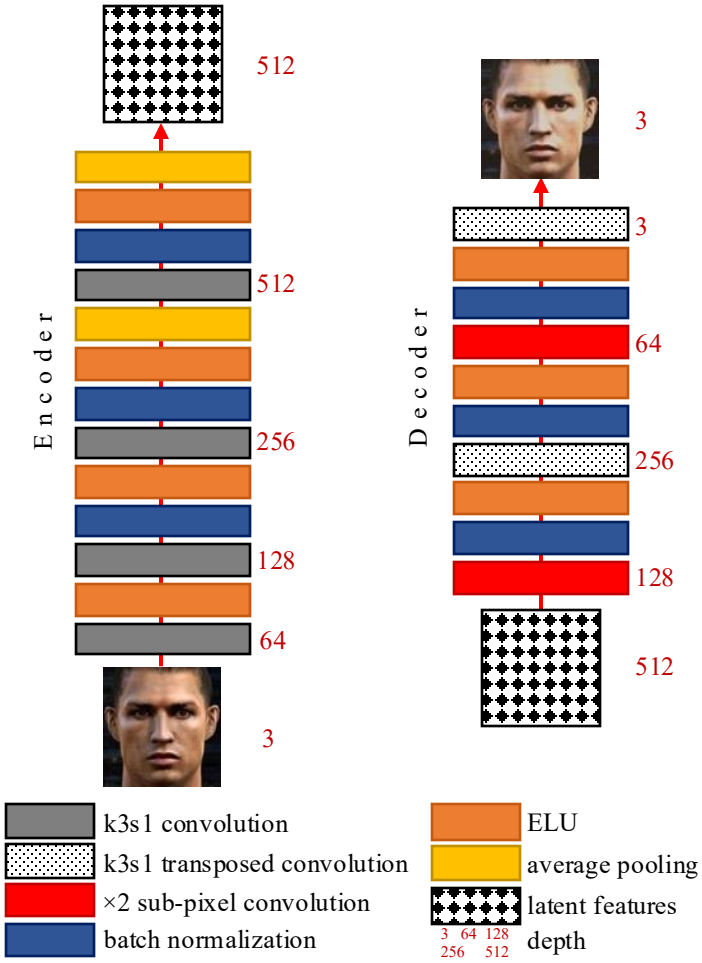
3.2. Overview



3. Proposed method

3.3. Network Architecture

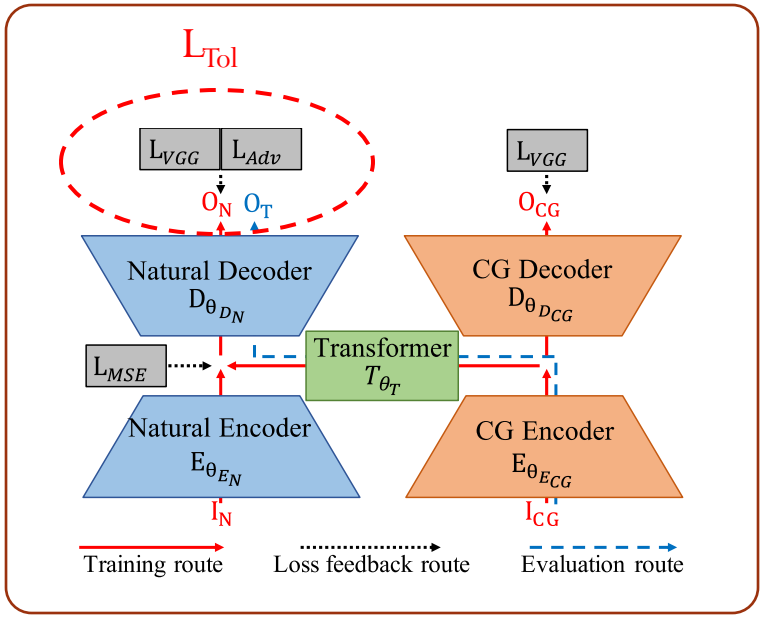




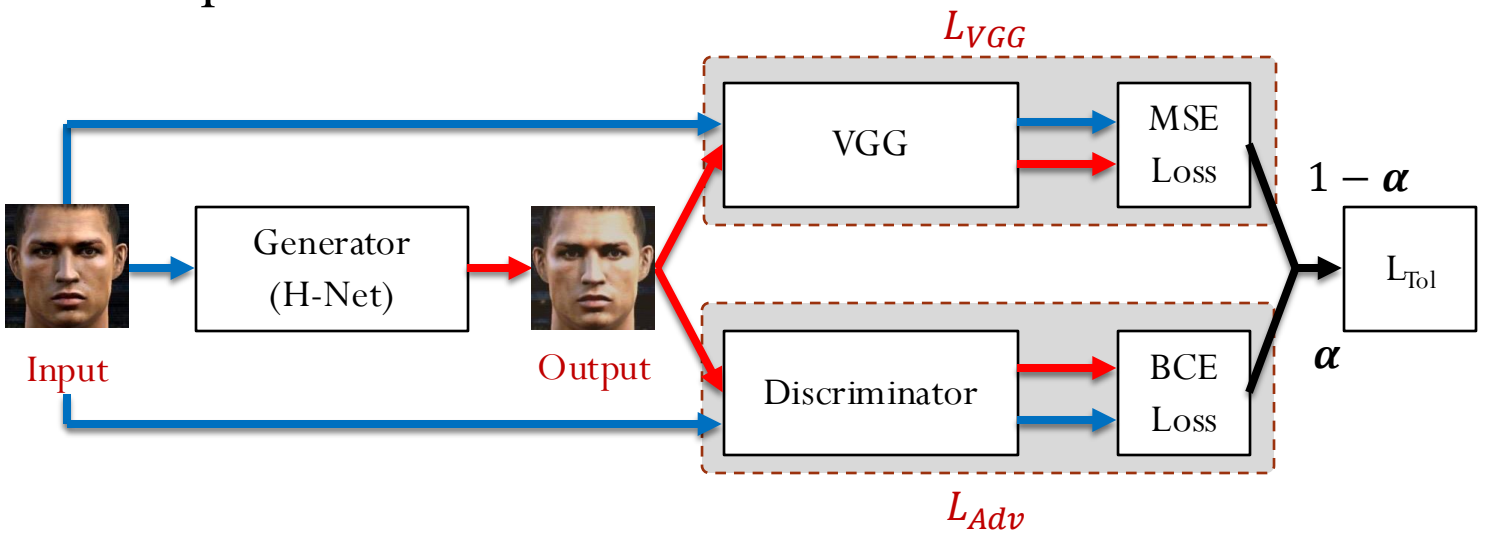
3. Proposed method

3.4. Loss Function

$$L_{Tol} = (1 - \alpha)L_{VGG} + \alpha L_{Adv}$$



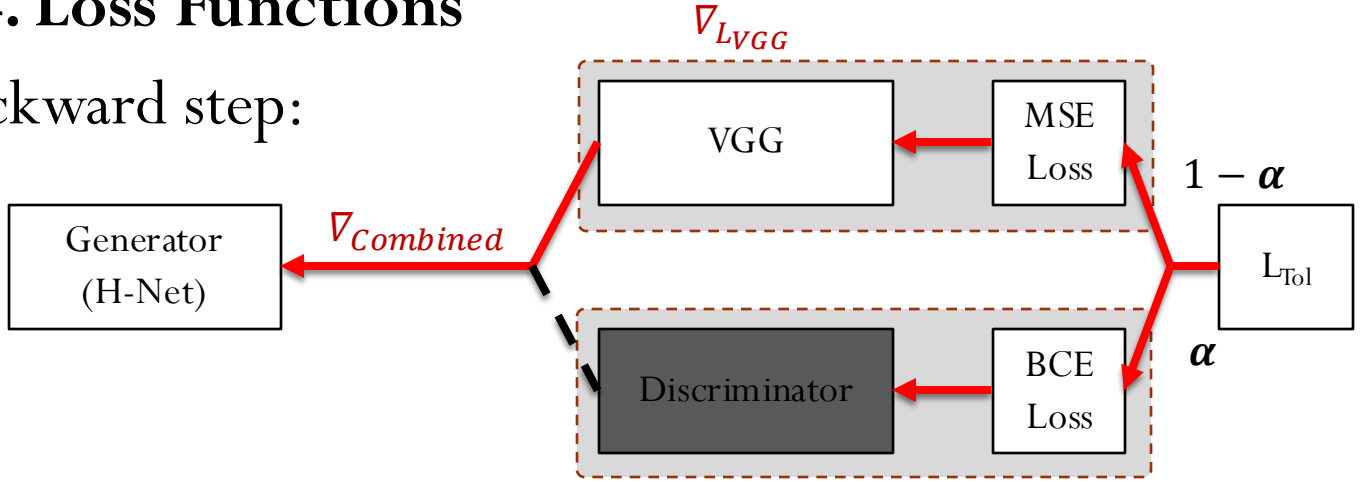
Forward step:



3. Proposed method

3.4. Loss Functions

Backward step:



VGG net is differentiable, however, the discriminator is treated as a **black-box** (not differentiable).

→ Approximating the gradient:

$$\nabla_{Combined} = \frac{(1-\alpha)L_{VGG}+\alpha L_{Adv}}{L_{VGG}} \nabla_{L_{VGG}}$$

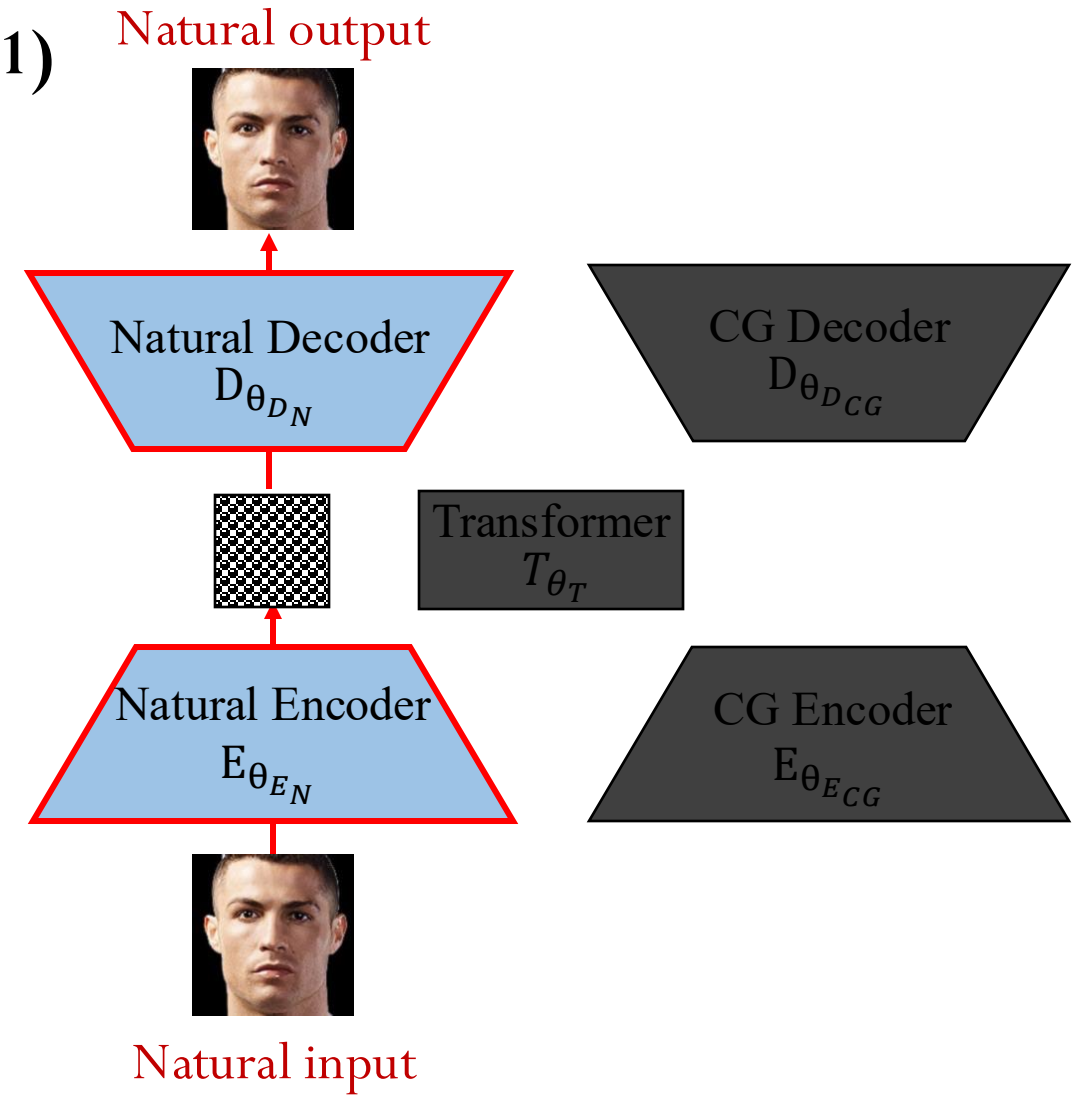
3. Proposed method

3.5. Training

With each batch input, performing **4 steps**:

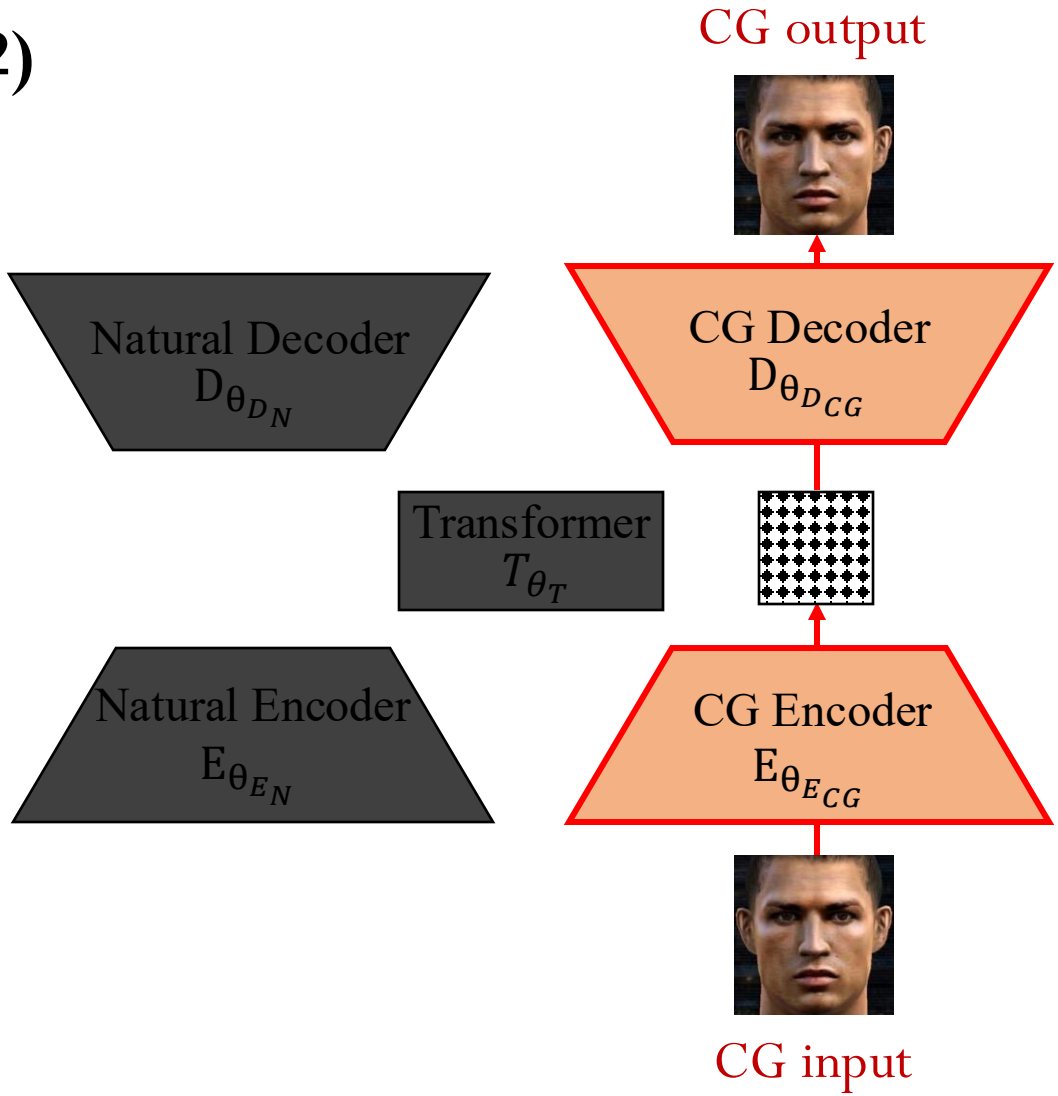
3. Proposed method

3.5. Training (1)



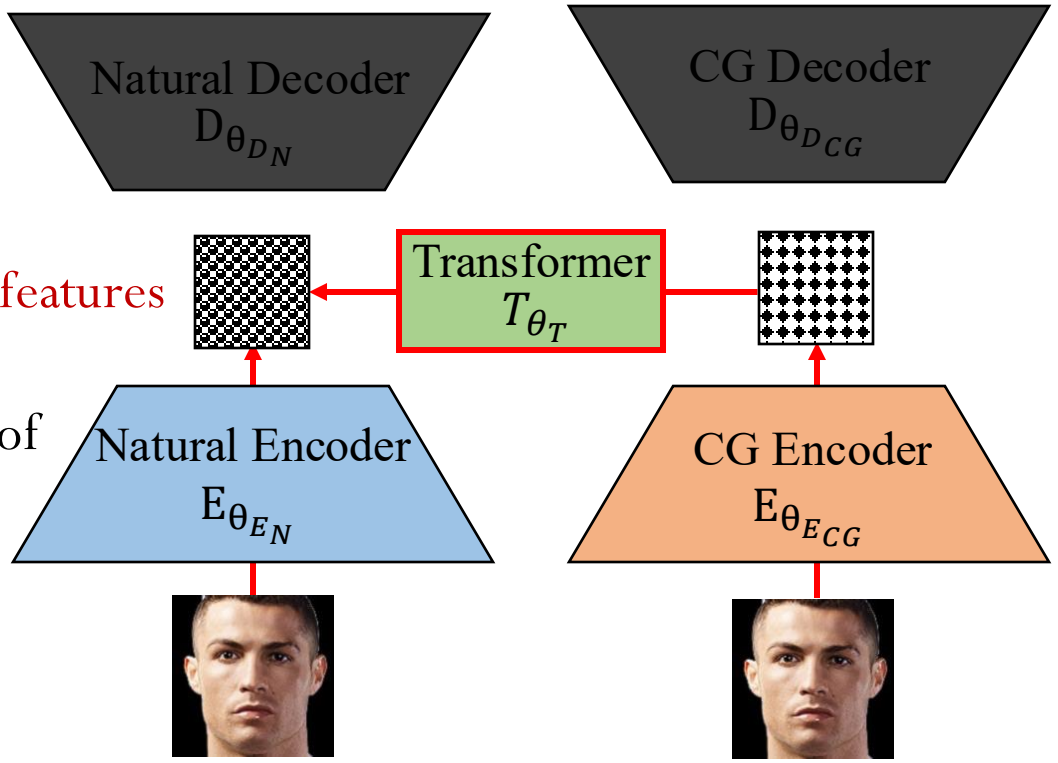
3. Proposed method

3.5. Training (2)



3. Proposed method

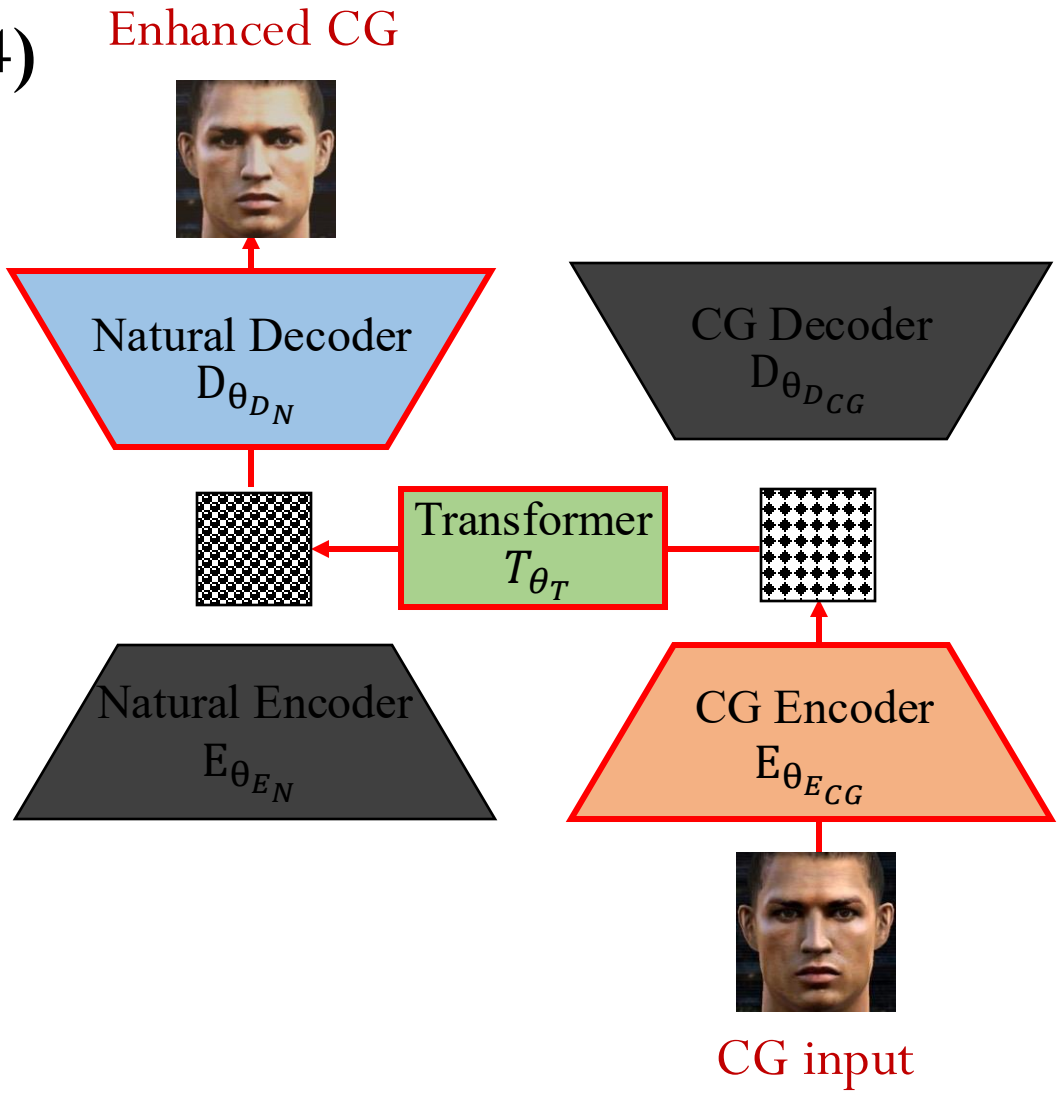
3.5. Training (3)



Note: Freeze parameters of Natural and CG encoders

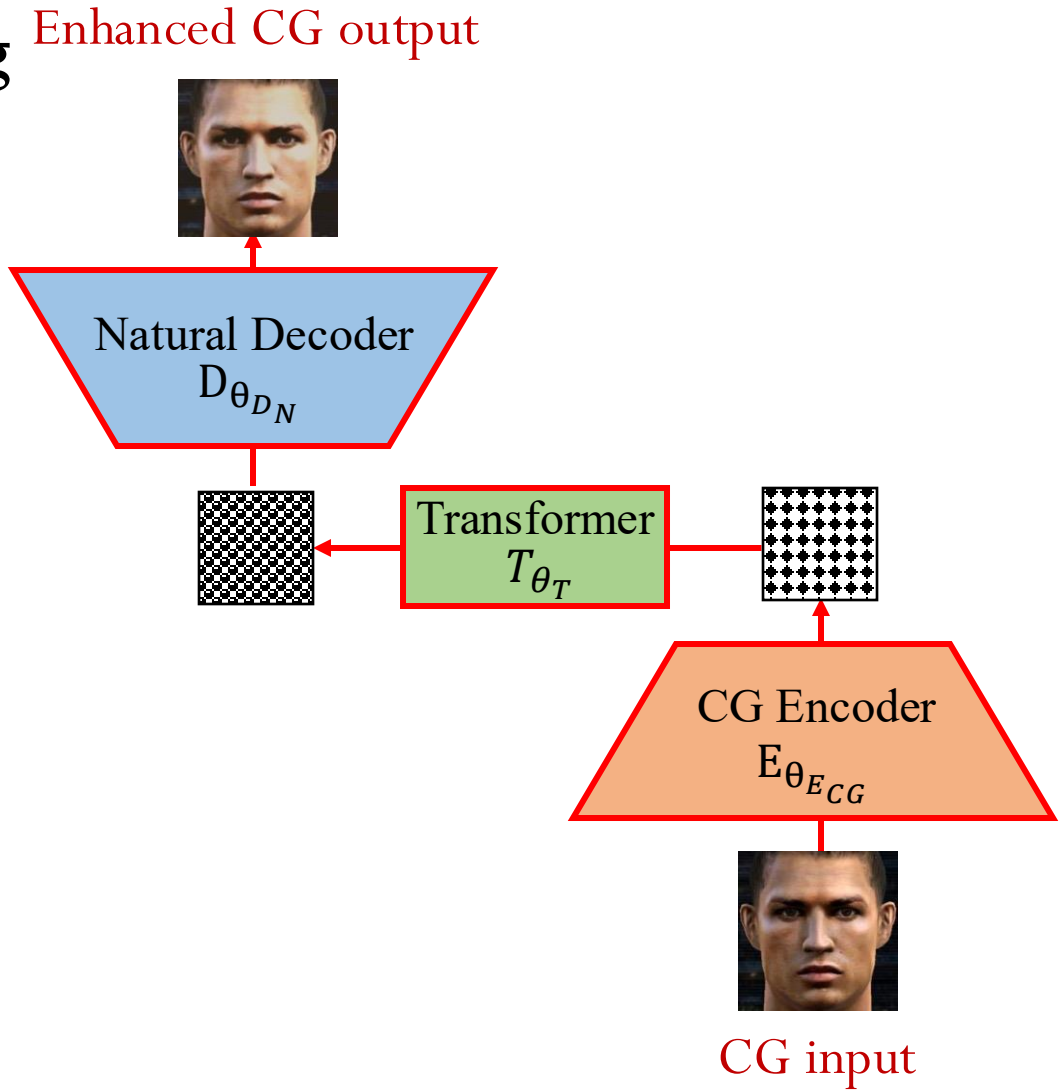
3. Proposed method

3.5. Training (4)



3. Proposed method

3.6. Evaluating



4. Evaluation

4. Evaluation

4.1. Datasets

No	Components	Size	Description
1	Dang Nguyen et al.[1]	CG: 240 Nal: 240	40 **very realistic** CG images collected from Web plus 200 **good quality** CG images extracted from PES 2012 soccer game 240 natural images retrieved from Internet
2	Basel (CG)[2] Caltech99 (Real)[3]	CG: 270 Nal: 270	3D face scans and rendered images from Basel Face Model Natural images from Caltech Faces 1999 dataset
3	MIT (CG)[4] (Grayscale) MS-Celeb-1M (Real)[5]	CG: 3236 Nal: 3236	CG images extracted from MIT CBCL dataset Natural images selected from MS-Celeb-1M cropped version

1. Dang-Nguyen, Duc-Tien, Giulia Boato, and Francesco GB De Natale. "Discrimination between computer generated and natural human faces based on asymmetry information." Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European. IEEE, 2012.
2. Paysan, Pascal, et al. "A 3D face model for pose and illumination invariant face recognition." Advanced Video and Signal Based Surveillance, 2009. AVSS'09. Sixth IEEE International Conference on. Ieee, 2009.
3. Weber, M. "A frontal face dataset collected by Markus Weber at the California Institute of Technology." Available from California Institute of Technology Bg-Caltech http://www. vision. caltech. edu/Image_Datasets/faces/RE ADME(1999).
4. Huang, Jennifer, Bernd Heisele, and Volker Blanz. "Component-based face recognition with 3D morphable models." International conference on audio-and video-based biometric person authentication. Springer, Berlin, Heidelberg, 2003.
5. Guo, Yandong, et al. "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition." European Conference on Computer Vision. Springer International Publishing, 2016.

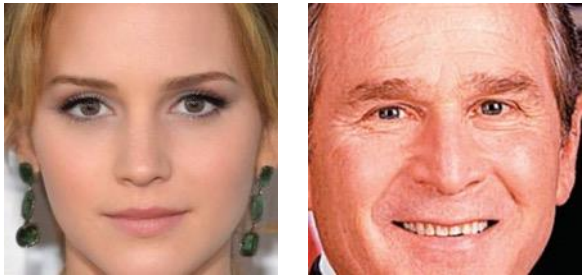
4. Evaluation

4.1. Datasets

Natural images

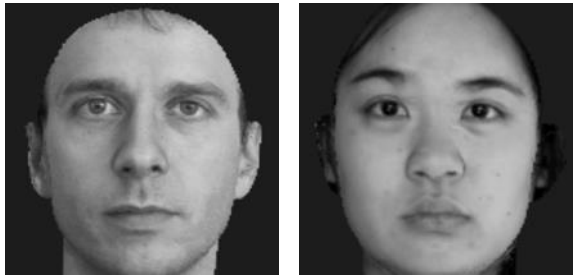
CG images

Dataset 1



Dataset 2

Dataset 3



4. Evaluation

4.2. **Examples:** Dataset 1

Before

After

- Color and brightness of images were normalized by H-Net
- First two images demonstrate good transformation in perception
- Contrast of third image was improved
- Skin color in the last image was whitened a bit undesirably

4. Evaluation

4.2. Examples: Dataset 2 and 3

Before

After

- Brightness of first two image was unexpectedly reduced due to bright background
- Grayscale images were given skin-like color

4. Evaluation

4.3. Measurement

- N: test size (number of evaluated images)
- n_{TP}: number of images correctly classified as CG
- n_{TN}: number of images correctly classified as natural
- n_{FN}: number of CG images misclassified as natural

Accuracy: $\frac{n_{TP}+n_{TN}}{N}$

Detection rate: $\frac{n_{TP}}{n_{TP}+n_{FN}}$

4. Evaluation

4.4. Evaluation Scenarios

There are two scenarios:

1. Attacker knows the training dataset
2. Attacker does not know the training dataset

4. Evaluation

4.4.1. Scenario 1: Attacker knows training dataset

- Train both spoofing detectors and H-Net on dataset 1
- Using H-Net to transform dataset 1, 2, and 3 → dataset 1', 2' & 3'
- Evaluate those spoofing detectors on dataset 1, 2, 3 and 1', 2', 3'

Measurement	Avg. accuracy		Avg. detection rate	
Dataset	w/o H-Net	w/ H-Net	w/o H-Net	w/ H-Net
Wu et al.	80.42	53.51	74.06	18.73
Peng et al.	66.70	28.51	96.50	20.15
Nguyen et al.	44.54	53.60	29.19	46.81

- H-Net successfully worked with Wu's and Peng's detectors
- Nguyen's detector was overfitted

4. Evaluation

4.4.2. Scenario 2: Attacker does not know training dataset

- Train H-Net on dataset 1, spoofing detectors on dataset 2
- Using H-Net to transform dataset 3 → dataset 3'
- Evaluate those spoofing detectors on dataset 3 and 3'

Measurement	Accuracy		Detection rate	
Dataset	**3**	**3'**	**3**	**3'**
Wu et al.	56.38	6.46	100.00	0.19
Peng et al.	92.32	42.57	100.00	0.49
Nguyen et al.	96.72	71.54	99.20	48.89

H-Net successfully worked with all detectors

4. Evaluation

4.4.2. Scenario 2: Attacker does not know training dataset

- Train H-Net on dataset 2, spoofing detectors on dataset 1
- Using H-Net to transform dataset 3 → dataset 3'
- Evaluate those spoofing detectors on dataset 3 and 3'

Measurement	Accuracy		Detection rate	
Dataset	**3**	**3'**	**3**	**3'**
Wu et al.	64.65	82.73	35.51	71.66
Peng et al.	50.20	0.20	100	0.00
Nguyen et al.	32.31	41.27	0.49	18.17

- H-Net successfully worked with Peng's detectors
- Nguyen's detector was overfitted as in scenario 1

5. Conclusion & Future Work

5. Conclusion and Future Work

5.1. Conclusion

- Both H-Net and detectors were affected by training datasets.
- In most cases, over 50% of the CG images transformed using H-Net avoided detection by 3 state-of-the-art spoofing detectors.
- Since the facial features were preserved, facial recognition was unaffected.
- The network can be trained using a black-box discriminator that cannot perform back propagation.
- Raise an alarm about the robustness of spoofing detectors.

5. Conclusion and Future Work

5.2. Future Work

- Trying to attack using local substitute method [4].
- Solving dataset problem (ReplayAttack [5,6], FaceForensics [7]).
- Evaluating with facial recognition system.
- Evaluating with newer spoofing detectors.
- Increasing the capacity of H-Net & dealing with larger images.
- Reducing network size to enable it to work smoothly with video frames in real time.

[4] Papernot, Nicolas, et al. "Practical black-box attacks against machine learning." AsiaCCS 2017.

[5] Chingovska, Ivana, André Anjos, and Sébastien Marcel. "On the effectiveness of local binary patterns in face anti-spoofing." Biometrics Special Interest Group (BIOSIG). IEEE, 2012.

[6] Costa-Pazo, Artur, Sushil Bhattacharjee, Esteban Vazquez-Fernandez, and Sebastien Marcel. "The replay-mobile face presentation-attack database." Biometrics Special Interest Group (BIOSIG). IEEE, 2016.

[7] Rössler, Andreas, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. "FaceForensics: A Large-scale Video Dataset for Forgery Detection in Human Faces." arXiv preprint arXiv:1803.09179 (2018).

Thank you for your attention