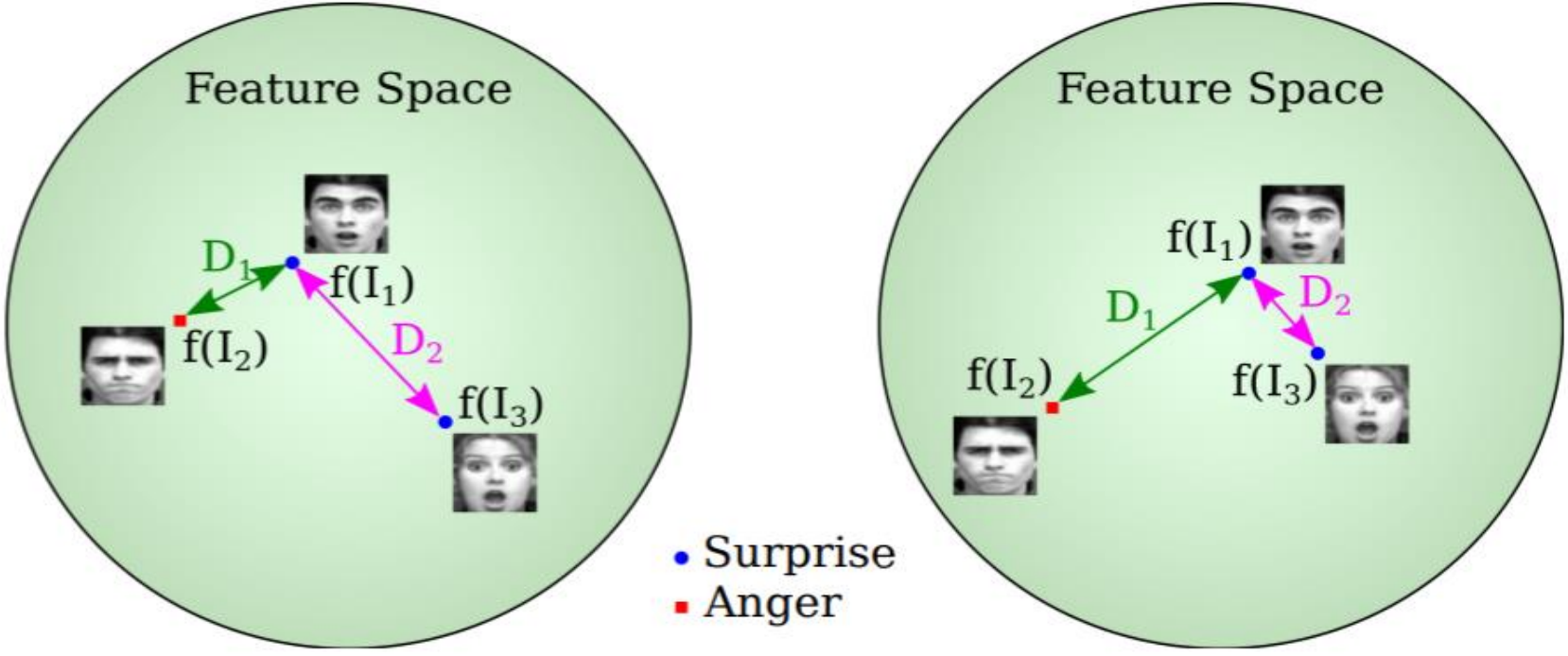
Facial Expression Recognition By Using a Disentangled Identity-Invariant Expression Representation

ICPR 2020

Kamran Ali and Charles E. Hughes

Synthetic Reality Lab, CS/CECS, University of Central Florida, USA



Problem

Objective

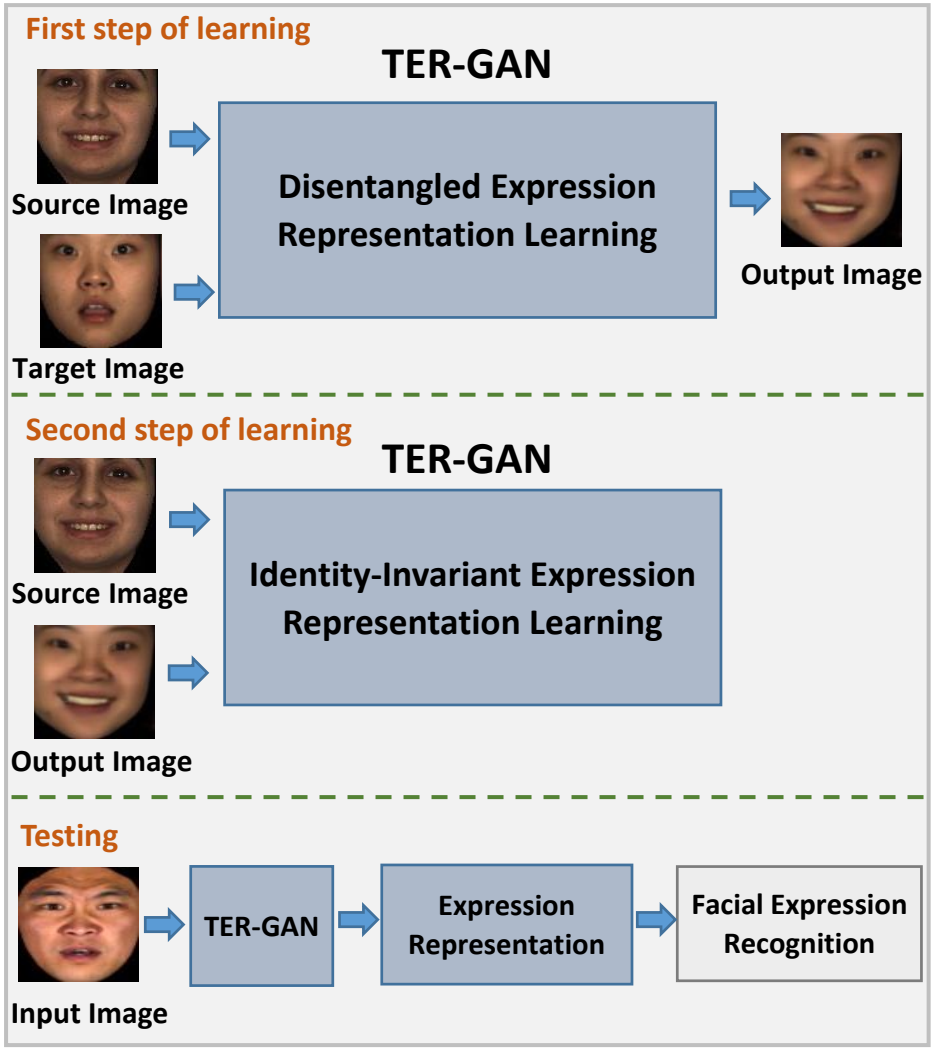
Image taken from the IACNN paper by Zibo et al. FG 2017

TER-GAN

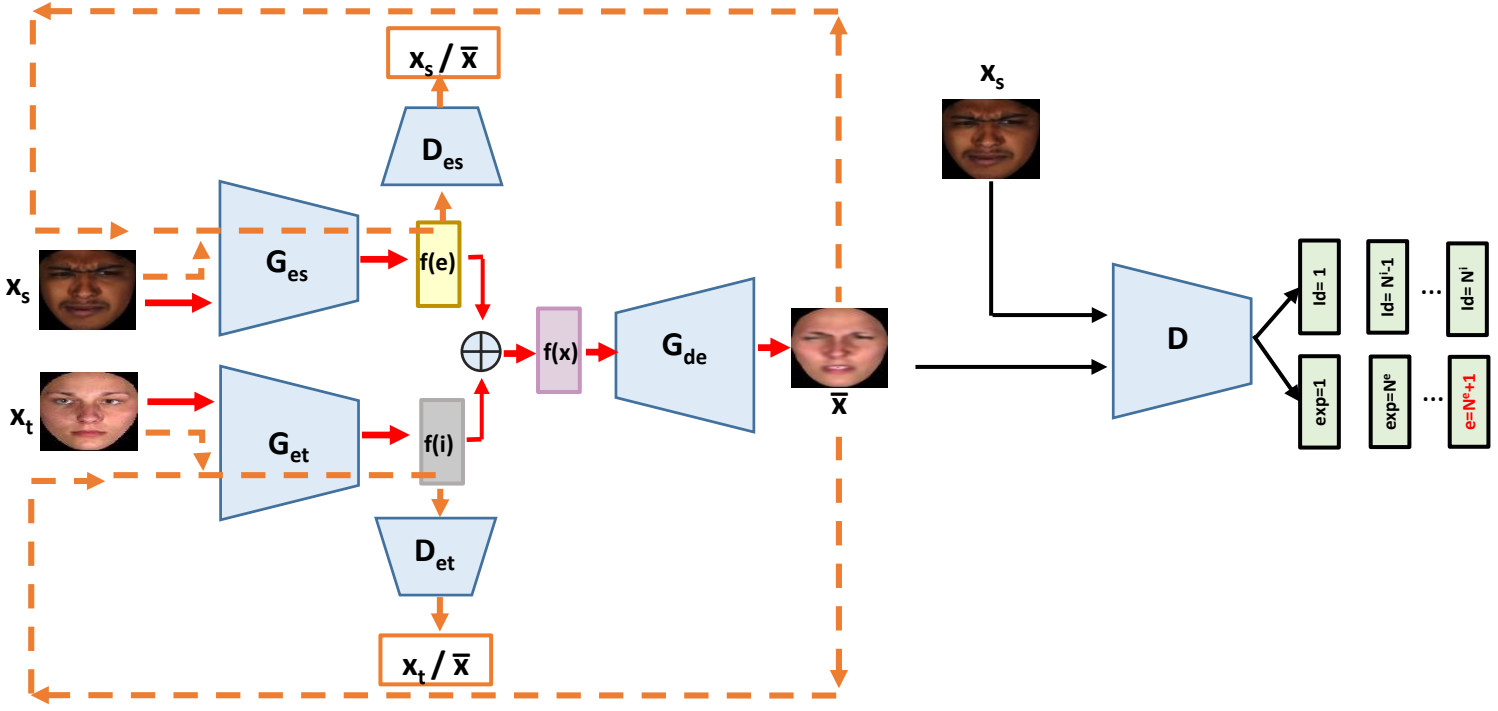
Transfer-based Expression Recognition Generative Adversarial Network (TER-GAN)

- Disentangle expression features from the identity information
- Identity-invariant expression representation learning for FER

TER-GAN Overall Framework



Architecture of TER-GAN



TER-GAN

Discriminator: The main objective of D is three-fold:

- To classify between real and fake images
- To categorize facial expressions
- To recognize the identities of expression images

$$\max_{D} \mathcal{L}_{\mathcal{D}}(D,G) = E_{\substack{x_s,y_s \sim p_s(x_s,y_s), \\ x_t,y_t \sim p_t(x_t,y_t)}} [\log(D^e_{y^e_s}(x_s)) + \log(D^i_{y^i_t}(x_t))] +$$

$$E_{\substack{x_s,y_s \sim p_s(x_s,y_s), \\ x_t,y_t \sim p_t(x_t,y_t)}} [\log(D^e_{N^e+1}(G(x_s,x_t)))]$$

TER-GAN

Generator: The goal of G is to:

- Disentangle expression features from the source image
- Extract identity features from the target image
- Synthesize an expression image to fool D to classify it to the expression of the source image and the identity of the target image

$$\max_{G} \mathcal{L}_{\mathcal{G}}(D, G) = E_{\substack{x_s, y_s \sim p_s(x_s, y_s), \\ x_t, y_t \sim p_t(x_t, y_t)}} [\log(D^e_{y^e_s}(G(x_s, x_t)) + \log(D^i_{y^i_t}(G(x_s, x_t))]$$

Adversarial Expression Consistency Loss

Encoder G_{es}: The goal of **G_{es}** is to:

- Learn an identity-invariant expression embedding
- A discriminator D_{es} is trained on top of expression embedding

$$\min_{G_{es}} \max_{D_{es}} \mathcal{L}_{D_{es}} = E_{x_s \sim p_d(x_s)} \mathcal{L}(1, D_{es}(G_{es}(x_s))) + E_{\bar{x} \sim p_{\bar{x}}(\bar{x})} \mathcal{L}(2, D_{es}(G_{es}(\bar{x})))$$

The total TER-GAN loss is given by:

$$\min_{G_{es},G_{et}} \max_{D_{et}} \mathcal{L}_{TER-GAN} = \lambda_1 \mathcal{L}_{adv} + \lambda_2 \mathcal{L}_{D_{et}} + \lambda_3 \mathcal{L}_{D_{es}} + \lambda_4 \mathcal{L}_{pixel}$$

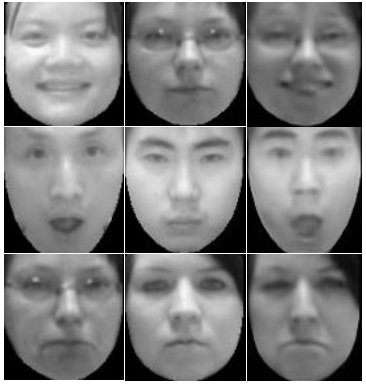
BU-3DFE

Source Image | Target Image | Generated Image

CK+

Source Image | Target Image | Generated Image

OC



Source Image | Target Image | Generated Image

MMI

Source Image | Target Image | Generated Image

Experimental Results

CK+

Method	Setting	Accuracy
STM-Explet	Dynamic	94.19
STCAM	Dynamic	99.08
IACNN	Static	95.37
DeRL	Static	97.30
CNN(baseline)	Static	91.64
TER-GAN without L_{Des}	Static	94.93
TER-GAN(Ours)	Static	**98.47**

Experimental Results

Oulu-CASIA

Method	Setting	Accuracy
STM-Explet	Dynamic	74.59
STCAM	Dynamic	91.25
PPDN	Static	84.59
DeRL	Static	88.0
CNN(baseline)	Static	73.14
TER-GAN without L_{Des}	Static	82.74
TER-GAN(Ours)	Static	**89.14**

Experimental Results

BU-3DFE

Method	Setting	Accuracy
Berretti et al.[1]	3D	77.54
Yang et al.[2]	3D	84.80
Lo et al.[3]	2D+3D	86.32
DeRL	Static	84.17
CNN(baseline)	Static	75.63
TER-GAN without L_{Des}	Static	81.25
TER-GAN(Ours)	Static	**84.83**

Experimental Results

MMI

Method	Setting	Accuracy
STM-Explet	Dynamic	75.12
STCAM	Dynamic	82.21
IACNN	Static	71.55
DeRL	Static	73.23
CNN(baseline)	Static	60.37
TER-GAN without L_{Des}	Static	67.81
TER-GAN(Ours)	Static	**74.69**

Experimental Results

Cross dataset validation

Method	Setting	Accuracy
CNN(baseline)	Static	75.63
TER-GAN(Ours) train on BU-4DFE, test on BU-3DFE	Static	73.97

References

[1]. S. Berretti, A. Del Bimbo, P. Pala, B. B. Amor, and M. Daoudi, “A set of selected sift features for 3d facial expression recognition,” in 2010 20th International Conference on Pattern Recognition. IEEE, 2010, pp. 4125–4128.

[2]. X. Yang, D. Huang, Y. Wang, and L. Chen, “Automatic 3d facial expression recognition using geometric scattering representation,” in 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), vol. 1. IEEE, 2015, pp. 1–6.

[3]. H. Li, H. Ding, D. Huang, Y. Wang, X. Zhao, J.-M. Morvan, and L. Chen, “An efficient multimodal 2d+ 3d feature-based approach to automatic facial expression recognition,” Computer Vision and Image Understanding, vol. 140, pp. 83–92, 2015.

Thank You!

Questions?