THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

DEPARTMENT OF Electronic and Information Engineering

Deep Multi-task Learning for Facial Expression Recognition and Synthesis Based on Selective Feature Sharing

ICPR 2020

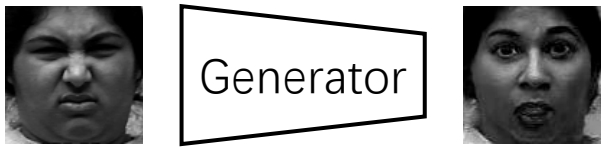
Rui Zhao, Tianshan Liu, Jun Xiao, Daniel P.K. Lun, and Kin-Man Lam

Department of Electronic and Information Engineering
The Hong Kong Polytechnic University

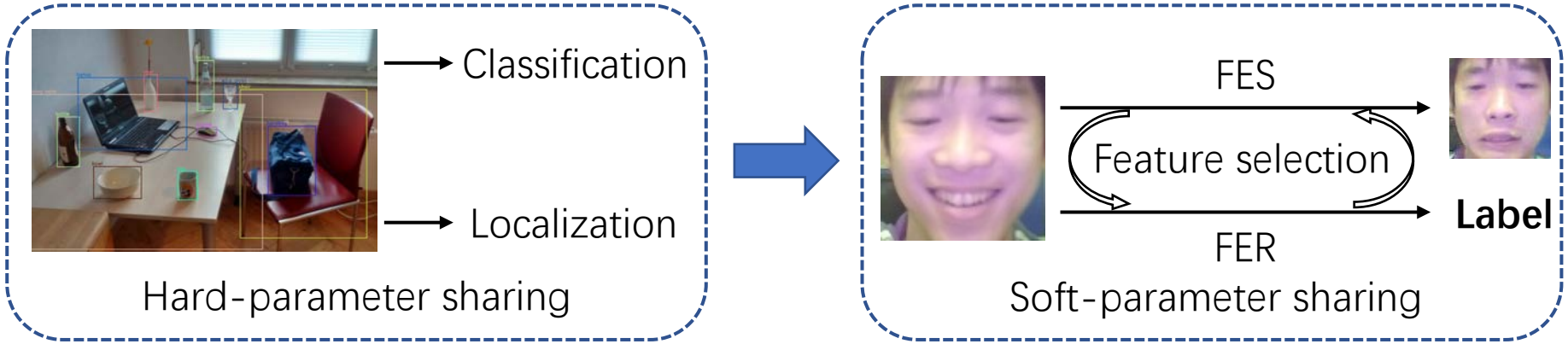
Motivations

- Facial expression synthesis (FES)
  - ❑ GAN: facial observation → latent code → facial observation

    The generator naturally captures strong semantics of facial expressions



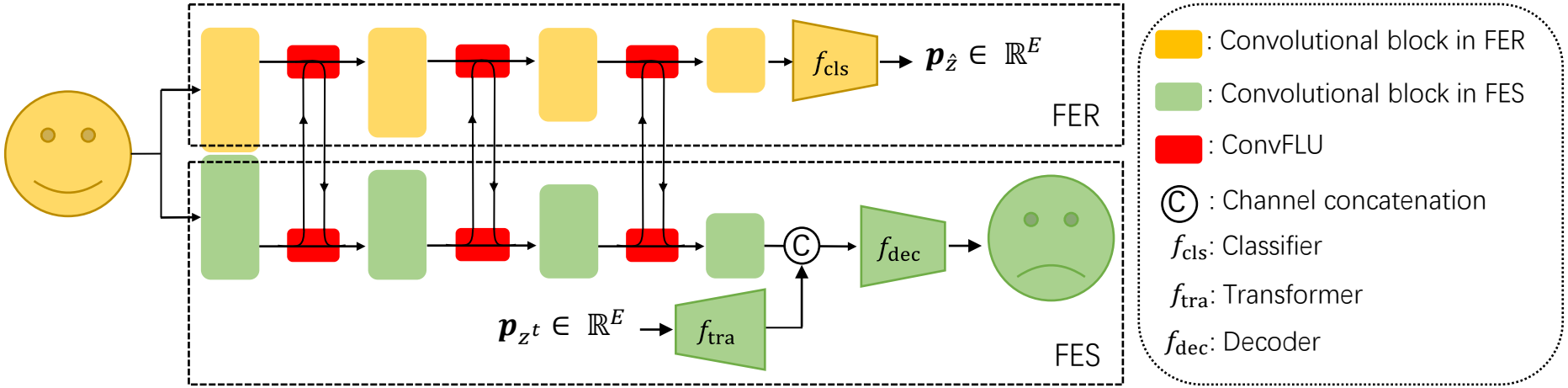
- Regularize the interaction between different tasks
  - ❑ Current multi-task networks adopt a simple hard-parameter sharing strategy:



DEPARTMENT OF Electronic and Information Engineering

Main Idea

- We propose a novel multi-task network, with convolutional feature leaky units, to selectively transfer the beneficial features between FER and FES.
- We employ the FES branch to enlarge and balance the training dataset for further improving the generalization ability.



DEPARTMENT OF Electronic and Information Engineering

Methodology

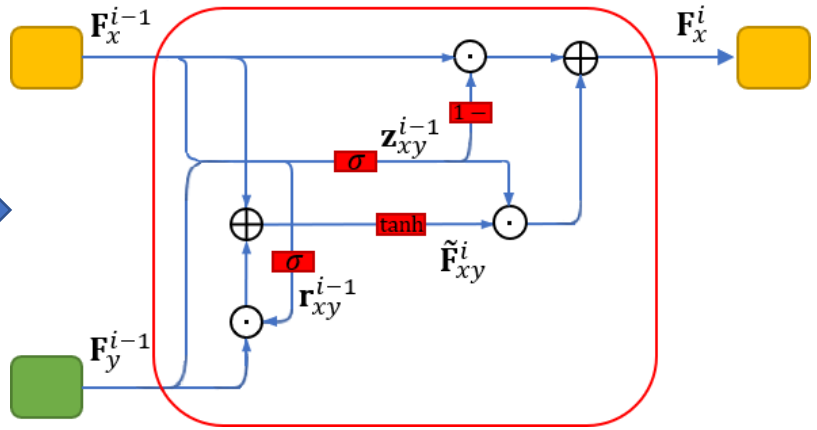
- Convolutional Feature Leaky Unit

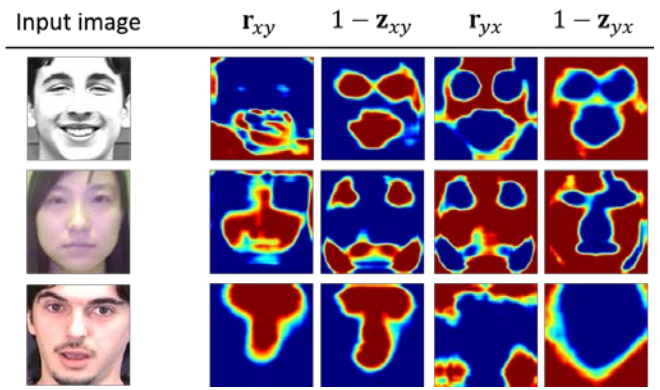
$$\mathbf{r}_{xy}^{i-1} = \sigma(\mathbf{W}_{\mathbf{r}}^{i-1} * [\mathbf{F}_x^{i-1}, \mathbf{F}_y^{i-1}]),$$

$$\widetilde{\mathbf{F}}_{xy}^{i} = \tanh(\mathbf{W}^{i-1} * (\mathbf{r}_{xy}^{i-1} \odot \mathbf{F}_y^{i-1}) + \mathbf{U}^{i-1} * \mathbf{F}_x^{i-1}),$$

$$\mathbf{z}_{xy}^{i-1} = \sigma(\mathbf{W}_{\mathbf{z}}^{i-1} * [\mathbf{F}_x^{i-1}, \mathbf{F}_y^{i-1}]),$$

$$\mathbf{F}_x^{i} = (1 - \mathbf{z}_{xy}^{i-1}) \odot \mathbf{F}_x^{i-1} + \mathbf{z}_{xy}^{i-1} \odot \widetilde{\mathbf{F}}_{xy}^{i}.$$





x : Task of facial expression recognition

y : Task of facial expression synthesis

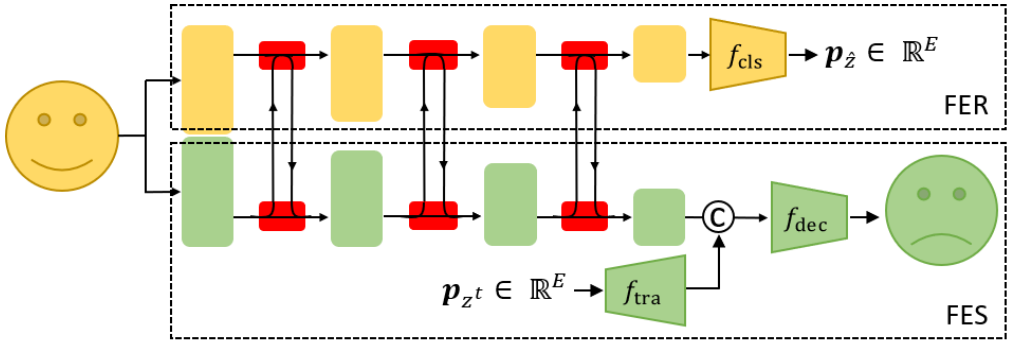
$\mathbf{r}_{xy}^{i-1}$: Leaky gate determines the knowledge transfer.

$\mathbf{z}_{xy}^{i-1}$: Memory gate determines the knowledge preservation

DEPARTMENT OF Electronic and Information Engineering

Methodology

- Learning Criteria



❑ FER → Accurate classification:

$$\mathcal{L}_{\text{cls}} = \frac{1}{n}\sum_{i=1}^{n} -\log\left(\frac{\exp(\boldsymbol{p}_i[z_i])}{\sum_j \exp(\boldsymbol{p}_i[j])}\right)$$

❑ FES → Photo-realistic facial images with the expected expressions

$$\mathcal{L}_{\text{GAN}} = \mathbb{E}_{z^t,\boldsymbol{I}^t}[\log D(z^t, \boldsymbol{I}^t)] + \mathbb{E}_{z^t,\boldsymbol{I}}[\log(1 - D(z^t, G(z^t, \boldsymbol{I})))]$$ → Photo realistic

$$\mathcal{L}_{\text{rec}} = \frac{1}{n}\sum_{i=1}^{n} ||G(z^t, \boldsymbol{I}_i) - \boldsymbol{I}_i^t||_2^2$$ → Image content

$$\mathcal{L}_{\text{cyc}} = \frac{1}{n}\sum_{i=1}^{n} ||G(z_i, G(z^t, \boldsymbol{I}_i)) - \boldsymbol{I}_i||_2^2$$ → Cycle consistency

$$\mathcal{L}_{\text{idt}} = \frac{1}{n}\sum_{i=1}^{n} ||f_{\text{LiCNN}}(G(z^t, \boldsymbol{I}_i)) - f_{\text{LiCNN}}(\boldsymbol{I}_i^t)||_2^2$$ → Identity preservation

Experiments

- Datasets
  - ❑ Extended Cohn-Kanade (CK+)
  - ❑ Oulu-CASIA (Oulu)
  - ❑ MMI

THE NUMBER OF VIDEO SEQUENCES IN CK+, OULU-CASIA, AND MMI, BASED ON DIFFERENT EMOTION LABELS.

Database	An	Co	Di	Fe	Ha	Sa	Su	Total
CK+	45	18	59	25	69	28	83	327
Oulu-CASIA	80	-	80	80	80	80	80	480
MMI	33	-	32	28	42	32	41	208

- Settings:
  - ❑ The three peak-intensity facial images are selected
  - ❑ Ten-fold cross-validation strategy, based on the subject identity, is adopted.

➢ [CK+]: T. Kanade, et al. , “Comprehensive database for facial expression analysis,” in IEEE International Conference FG, 2000.
➢ [Oulu]: G. Zhao, et al. , “Facial expression recognition from near-infrared videos,” Image Vision Computation, 2011.
➢ [MMI]: M. Pantic , et al. , “Web-based database for facial expression analysis,” in ICME, 2005

Experiments

- Recognition Results for FER

Methods	Pre-train	Setting	Accuracy (%)
LBP-TOP [35]	✗	Image sequence	88.99
HOG 3D [36]	✗	Image sequence	91.44
3DCNN [37]	✗	Image sequence	85.9
IACNN [25]	✓	Single image	95.37
DTAGN [26]	✓	Image sequence	97.25
IPA2LT [38]	✗	Single image	91.67
DeRL [27]	✓	Single image	97.30
LBVCNN [28]	✓	Image sequence	97.38
DMT-CNN [10]	✓	Single image	97.55
FERSNet	✗	Single image	97.35
FERSNet (BU-4DFE)	✓	Single image	**97.85**

CK+

Methods	Pre-train	Setting	Accuracy (%)
LBP-TOP [35]	✗	Image sequence	68.13
HOG 3D [36]	✗	Image sequence	70.63
STM-Explet [40]	✗	Image sequence	74.59
DTAGN [26]	✓	Image sequence	81.46
IPA2LT [38]	✗	Single image	61.02
DeRL [27]	✓	Single image	88.0
LBVCNN [28]	✓	Image sequence	82.41
DMT-CNN [10]	✓	Single image	87.5
ExprGAN [13]	✓	Single image	84.72
FERSNet	✗	Single image	83.47
FERSNet (BU-4DFE)	✓	Single image	**89.23**

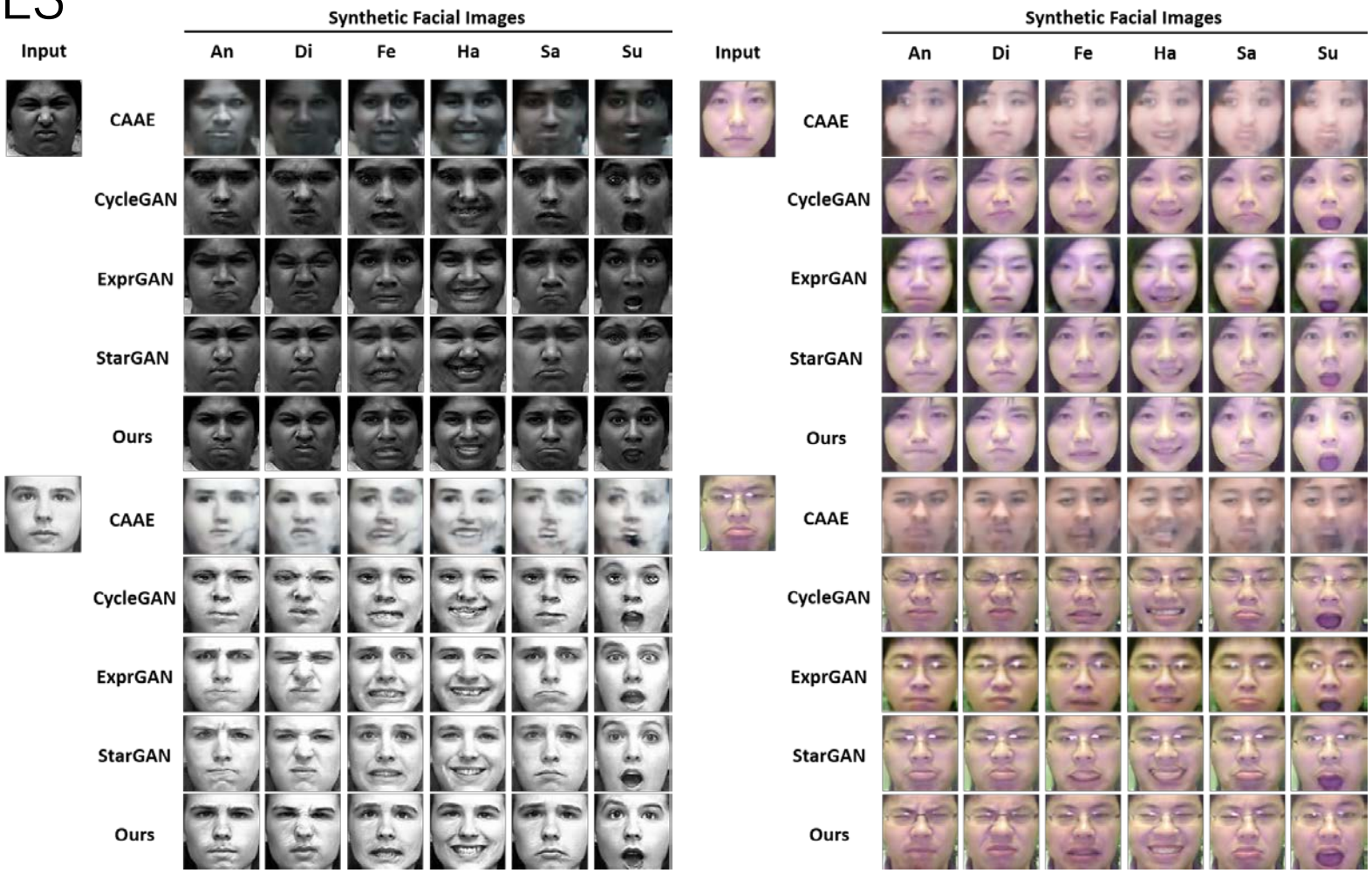
Oulu

Methods	Pre-train	Setting	Accuracy (%)
LBP-TOP [35]	✗	Image sequence	59.51
HOG 3D [36]	✗	Image sequence	60.89
STM-Explet [40]	✗	Image sequence	75.12
DTAGN [26]	✓	Image sequence	70.24
IACNN [25]	✓	Single image	71.55
DeRL [27]	✓	Single image	73.23
LBVCNN [28]	✓	Image sequence	**76.28**
FERSNet	✗	Single image	71.31
FERSNet (BU-4DFE)	✓	Single image	75.32

MMI

Experiments

- Results of FES



DEPARTMENT OF
Electronic and Information Engineering

Experiments

- Quantitative Results on FES

We employ a standard FER model to recognize the synthetic facial images from different generative models.

THE RECOGNITION ACCURACY (%) ON THE SYNTHETIC FACIAL IMAGES PRODUCED BY THE DIFFERENT METHODS. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**.

Method	CAAE	CycleGAN	ExprGAN	StarGAN	Ours
CK+	79.41	88.89	95.41	96.43	**97.04**
Oulu-CASIA	46.18	74.44	80.07	79.51	**81.52**

- Ablation Study

❖ FERSNet w/o MTL: single-task network
❖ FERSNet w/o ConvFLU: hard-parameter sharing
❖ FERSNet w. FES-DA: using FES for data augmentation

THE RECOGNITION RESULTS (%) FOR ABLATION STUDY. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**.

Method	CK+	Oulu-CASIA	MMI
FERSNet w/o MTL	94.70	73.33	63.78
FERSNet w/o ConvFLU	95.21	77.92	69.07
FERSNet (original)	97.35	83.47	71.31
FERSNet w/ FES-DA	**97.75**	**87.64**	**73.87**

Conclusions

- We proposed a novel multi-task learning strategy to tackle both FER and FES problems simultaneously in a network.
- We designed a convolutional feature leaky unit to transfer only the beneficial features between the FER and FES tasks, while filtering out the harmful or useless information.
- We conducted extensive experiments to evaluate the proposed framework on both the FER and FES tasks. The results demonstrated that our proposed method achieved state-of-the-art performance on those commonly used facial benchmarks.

Thank you!