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

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, Hong Kong

Email: {rick10.zhao, tianshan.liu, jun.xiao}@connect.polyu.hk, {enpklun, enkmlam}@polyu.edu.hk

Abstract—**Multi-task learning is an effective learning strategy for deep-learning-based facial expression recognition tasks. However, most existing methods take into limited consideration the feature selection, when transferring information between different tasks, which may lead to task interference when training the multi-task networks. To address this problem, we propose a novel selective feature-sharing method, and establish a multi-task network for facial expression recognition and facial expression synthesis. The proposed method can effectively transfer beneficial features between different tasks, while filtering out useless and harmful information. Moreover, we employ the facial expression synthesis task to enlarge and balance the training dataset to further enhance the generalization ability of the proposed method. Experimental results show that the proposed method achieves state-of-the-art performance on those commonly used facial expression recognition benchmarks, which makes it a potential solution to real-world facial expression recognition problems.**

I. INTRODUCTION

Facial expression is one of the most effective and natural approaches for human beings to express their emotions. Identifying and synthesizing facial expressions enable current human-computer interaction systems to better understand and simulate human behaviors. In spite of their great effectiveness, the facial expression recognition algorithms, based on deep neural networks, are highly reliant on large amounts of training samples with clean labels. The annotation of a large-scale database is not only time consuming, but also impractical. To overcome this limitation, a number of studies [1]–[3] have focused on multi-task learning for facial expression recognition (FER), which aims to train the models under the regularization from some auxiliary tasks. With the regularization effect, multi-task learning (MTL) becomes an effective learning strategy to tackle the overfitting problem, resulting from insufficient training samples. Moreover, learning multiple tasks simultaneously in a network can improve the performance by transferring beneficial knowledge to the main task from other relevant auxiliary tasks. Specifically, in FER, single-task networks are able to learn some very discriminative features with respect to facial expressions. However, the learnt features do not take the nuisance factors, such as subject identity, head pose, and illumination, into sufficient consideration, which results in poor generalization when applied to practical applications.

Therefore, MTL contributes significantly to a more robust solution with better generalization for FER tasks.

There are two main underlying problems in current MTL-based FER algorithms, i.e. the design of the auxiliary tasks, and the building of the connections between different tasks. In terms of the auxiliary task design, numerous studies have been proposed for setting the tasks, such as expression-related hidden unit detection [4], identity classification [5], and landmark localization [6], as the auxiliary tasks. In this paper, we propose to learn the facial expression synthesis (FES) as the auxiliary task for FER. FES aims to synthesize a facial expression image based on a guiding expression label. We employ a patch-based conditional generative adversarial network (cGAN) [7] to learn the FES task. In addition to the high correlation with FER, FES can also generate extra training samples and balance the training dataset, which can greatly enhance the performance of a deep facial expression recognition framework.

On the other hand, establishing task interaction is another important factor when building multi-task networks, because the interaction will directly affect the information flow between different tasks. Conventional algorithms apply the hard parameter-sharing approach, which shares the feature maps at the bottom layers of a network and separates different branches for different tasks at the top layers, such as the two-head structure in object detection for classification and localization [8]. In spite of its simplicity, the hard parameter-sharing approach lacks the ability in differentiating helpful and harmful information between tasks. To address this issue, we propose a novel multi-task network for FER and FES, namely facial expression recognition and synthesis network (FERSNet), with a soft parameter-sharing mechanism, which contributes to effectively selecting useful features from different tasks and different layers. Therefore, the main contributions of this paper can be summarized as follows:

- We propose a novel multi-task convolutional neural network, with the convolutional feature leaky unit, to selectively transfer the beneficial features between the facial expression recognition task and the facial expression synthesis task.
- We employ the facial expression synthesis branch to enlarge and balance the training dataset for further improv-
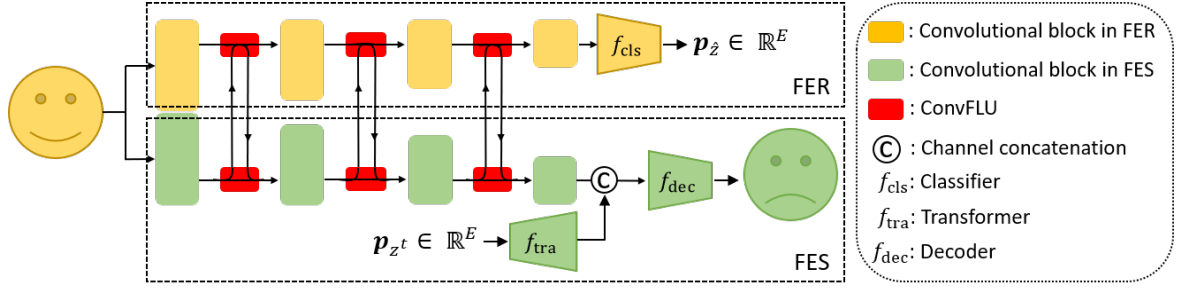
Fig. 1. The proposed two-stream multi-task network for FER and FES with the convolutional feature leaky units (ConvFLU). $\boldsymbol{p}_{\hat{z}}$ is the output probability vector indicating the predicted expression label $\hat{z}$ of the input image, and $\boldsymbol{p}_{z^t}$ is a one-hot vector, which controls the expression label z^t of the synthetic facial image. E is the number of the expression classes.

ing the generalization ability of the proposed algorithm.
- We conduct experiments to demonstrate that the proposed multi-task network achieves promising performance in recognizing and synthesizing facial expression images.

II. RELATED WORK

1) Multi-task learning for FER: Multi-task learning for FER has been widely studied over the past few decades. Previous works on multi-task-based FER attempt to combine FER with other facial image analysis tasks, in order to obtain a more robust representation of facial expressions in the feature space. Meng et al. [1] proposed a two-stream network to extract identity-invariant expression features for the emotion classification. Pons and Masip [2] suggested that jointly learning a model for FER and facial action units detection can significantly improve the FER performance. Moreover, Ranjan et al. [3] proposed a multi-branch network to solve diverse facial image analysis tasks simultaneously. Zhang et al. [5] proposed a multi-signal CNN under the supervision of the FER and face verification tasks, which forces the model to learn more discriminative features with respect to facial expressions. Ming et al. [9] proposed a multi-task network with the dynamic weights for the FER and face recognition tasks to enhance the model performance. However, the above-mentioned studies have not considered feature selection when sharing information between different tasks, which may greatly degrade their performance, because useless, or even harmful information, is transferred.

2) Facial expression synthesis: Facial expression synthesis is another widely studied topic in the field of facial image analysis. With the development of generative adversarial network (GAN) [10], the facial expression edit/synthesis has achieved appealing performance. Zhang and Song [11] proposed a conditional adversarial autoencoder (CAAE) for synthesizing the facial images with different expressions and ages. ExprGAN [12] adopted the conditional GAN strategy to produce facial images with different expression intensities. Moreover, Choi et al. [13] proposed StarGAN to achieve multi-domain image-to-image translation for facial image synthesis. In addition, the geometric-guided methods [14]–[16] employed the shape-aware supervision from the facial landmarks for expression editing, which achieved state-of-the-art performance in facial

expression transference. However, limited existing work employs the soft parameter-sharing strategy to enhance the quality of the synthetic facial images with the FER regularization.

3) Feature selection mechanism: The selection mechanism plays an important role in multi-task learning, which has been widely applied to natural language processing (NLP). Ruder et al. [17] presented Sluice Networks, in which a subspace combination approach was proposed to determine the information flow between different tasks. Moreover, Xiao et al. [18] took advantage of gated recurrent unit (GRU) and proposed a leaky unit with the property of remembering and forgetting information, which achieved state-of-the-art performance in text classification. However, different from the NLP tasks, the facial image analysis tasks usually fuse the local features to form a global representation of the image. Therefore, in this paper, we propose a convolutional feature leaky unit (ConvFLU) to perform feature selection between different tasks and layers.

III. THE PROPOSED METHOD

The proposed FERSNet solves the FER and FES tasks in parallel with the reliable knowledge transference. Fig. 1 illustrates the pipeline of FERSNet. The branch at the top aims to recognize the expression of the input facial image, while the branch at the bottom aims to generate a new facial expression image with the same identity as the input, based on the target-expression label. The two branches are connected by a set of soft parameter-sharing blocks, i.e. ConvFLUs. In this section, we first introduce the proposed multi-task framework for the FER and FES tasks, and then we present the details of ConvFLU for selective feature sharing. Finally, we show the learning strategy for the proposed multi-task network to solve the FER and FES problems simultaneously.

A. Framework

As presented in Fig. 1, the proposed two-stream network consists of two branches for FER and FES, respectively. The first four convolutional blocks in the FER and FES branches are connected with ConvFLUs, which aim to extract discriminative features for the FER and FES tasks, respectively. Based on the extracted FER features, the classifier f_{cls} predicts the emotion label $\hat{z}$ of the input facial image. In the FES branch,
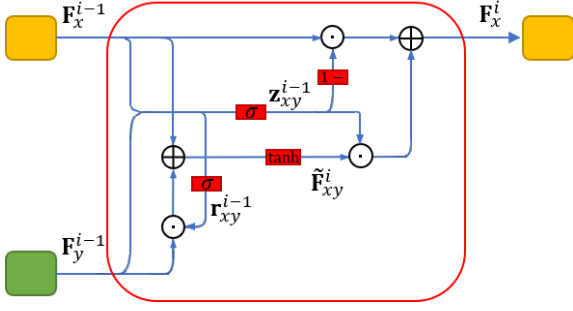
Fig. 2. The proposed ConvFLU in FERSNet. $\odot$ and $\oplus$ represent the element-wise multiplication and addition, respectively. σ and "tanh" refer to the sigmoid function and the hyperbolic tangent function, respectively.
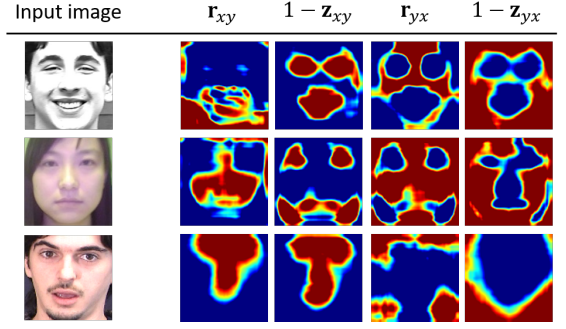

Fig. 3. The visualization of the gate values in ConvFLU. The blue and red values represent the low and high response values, respectively. Task x and y are the FER and the FES task in FERSNet, respectively.

another input variable z^t, which controls the expression of the synthetic image, is fed to the transformer f_{tra}. f_{tra} encodes the information of the target-expression label to produce a feature map with the same size as the extracted FES features. The two feature maps are fused by channel concatenation. Finally, the decoder f_{dec} reconstructs the image based on the fused features to mislead a patch-based discriminator [7].

In summary, the proposed FERSNet takes a facial image and a target-expression label as inputs to predict the expression label of the input image and synthesize another facial image with the expected (target) expression.

B. Convolutional Feature Leaky Unit

Inspired by GRU in [19], we propose a convolutional feature leaky unit, which inherits the effective property of remembering and forgetting features. We revise the original structure of GRU and employ it to filter out useless and harmful features, when transferring information between the different tasks. The structure of the proposed ConvFLU is illustrated in Fig. 2. We consider transferring the features from task y to task x in the $(i-1)$-st layer. It is worth noting that there are two ConvFLUs in one transference block, as shown in Fig. 1. Therefore, the information flow is bidirectional between FER and FES.

The leaky gate $\mathbf{r}_{xy}^{i-1}$ for transferring information from task y to task x is defined as follows:

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

where $\mathbf{W}_{\mathbf{r}}^{i-1}$ represents the trainable convolutional kernels for $\mathbf{r}_{xy}^{i-1}$. "[ ]" and "$*$" denote the concatenation and the convolution operation, respectively. $\mathbf{F}_x^{i-1}$ and $\mathbf{F}_y^{i-1}$ are the input feature maps from task x and task y in the $(i-1)$-st layer, respectively. Then, a new feature map is generated based on $\mathbf{r}_{xy}^{i-1}$ as follows:

$$\widetilde{\mathbf{F}}_{xy}^i = \tanh(\mathbf{W}^{i-1} * (\mathbf{r}_{xy}^{i-1} \odot \mathbf{F}_y^{i-1}) + \mathbf{U}^{i-1} * \mathbf{F}_x^{i-1}), \tag{2}$$

where $\mathbf{U}^{i-1}$ and $\mathbf{W}^{i-1}$ are the trainable convolutional kernels for combining $\mathbf{F}_x^{i-1}$ and $\mathbf{F}_y^{i-1}$. $\odot$ denotes the element-wise multiplication. It is clear that $\mathbf{r}_{xy}^{i-1}$ controls the information leakage from task y to task x. Moreover, we further consider a memory gate $\mathbf{z}_{xy}^{i-1}$, which determines the information that

should be remembered from the previous feature map. $\mathbf{z}_{xy}^{i-1}$ is defined as follows:

$$\mathbf{z}_{xy}^{i-1} = \sigma(\mathbf{W}_{\mathbf{z}}^{i-1} * [\mathbf{F}_x^{i-1}, \mathbf{F}_y^{i-1}]), \tag{3}$$

where $\mathbf{W}_{\mathbf{z}}^{i-1}$ is the trainable convolutional kernels for generating $\mathbf{z}_{xy}^{i-1}$. Thus, the final feature map $\mathbf{F}_x^i$, containing both the information from $\mathbf{F}_x^{i-1}$ and $\mathbf{F}_y^{i-1}$, is computed as follows:

$$\mathbf{F}_x^i = (1 - \mathbf{z}_{xy}^{i-1}) \odot \mathbf{F}_x^{i-1} + \mathbf{z}_{xy}^{i-1} \odot \widetilde{\mathbf{F}}_{xy}^i. \tag{4}$$

With the leaky gate and the memory gate, ConvFLU is able to select the beneficial features for FER and FES from each other. If the values in the leaky gates $\mathbf{r}_{xy}$ are close to 1, the branch tends to utilize more information from the other task. Similarly, if the values in the memory gates $\mathbf{z}_{xy}$ are close to 0, the branch tends to preserve its information for the corresponding task. We further visualize the leakage gate and the memory gate in Fig. 3 to better illustrate the selective feature-sharing strategy. It can be seen from the figure that the FER task (task x) mainly focuses on the mouth and the eye regions to predict the expression label, while the FES task (task y) tends to preserve the identity information, when synthesizing a new expression. Therefore, the proposed ConvFLU can be regarded as an attention mechanism, based on task correlation, for performing selective feature sharing.

C. Learning for FER and FES

In this paper, we solve FER and FES simultaneously with the proposed FERSNet. For the FER task, the classifier in the top branch consists of fully connected layers, under a standard cross entropy loss for training, which is defined as follows:

$$\mathcal{L}_{\text{cls}} = \frac{1}{n} \sum_{i=1}^{n} -\log \left(\frac{\exp(\boldsymbol{p}_i[z_i])}{\sum_j \exp(\boldsymbol{p}_i[j])} \right), \tag{5}$$

where $\boldsymbol{p}_i$ denotes the output probability vector of the i-th training sample, z_i represents its corresponding ground-truth label, and n denotes the total number of training samples.

In terms of the FES task in the bottom branch, we consider the loss from a patch-based discriminator [7], denoted as

$\mathcal{L}_{\text{GAN}}$, and also adopt the reconstruction loss, $\mathcal{L}_{\text{rec}}$, to synthesize visually pleasant facial images. Specifically, the GAN loss is defined as follows:

$$\mathcal{L}_{\text{GAN}} = \mathbb{E}_{z^t, I^t}[\log D(z^t, I^t)] + \\ \mathbb{E}_{z^t, I}[\log(1 - D(z^t, G(z^t, I)))], \quad (6)$$

where I^t and I denote the real target and the original facial image, respectively. z^t represents the target-expression label. G and D refer to the generator and the discriminator in the network, respectively. Specifically, the generator consists of all the modules in FERSNet for producing the synthetic face. As for the reconstruction loss, we define it as the mean squared error between the target and the synthetic facial image as follows:

$$\mathcal{L}_{\text{rec}} = \frac{1}{n} \sum_{i=1}^{n} ||G(z^t, I_i) - I_i^t||_2^2. \quad (7)$$

In addition, we follow the learning strategy in [14]–[16], and employ the cycle consistency loss and the identity preserving loss to further improve the synthesis accuracy. The cycle consistency loss $\mathcal{L}_{\text{cyc}}$ guarantees the consistency between the source image and the cycle-reconstructed image, which is defined as:

$$\mathcal{L}_{\text{cyc}} = \frac{1}{n} \sum_{i=1}^{n} ||G(z_i, G(z^t, I_i)) - I_i||_2^2. \quad (8)$$

In other words, if we transform the synthetic face back to the original expression z_i via FERSNet, the same facial image as I_i should be obtained. Moreover, the identity preserving loss $\mathcal{L}_{\text{idt}}$ is defined as the distance between the features extracted from the original face and the synthetic face with the pre-trained model-B of the Light CNN [20], which is formulated as follows:

$$\mathcal{L}_{\text{idt}} = \frac{1}{n} \sum_{i=1}^{n} ||f_{\text{LiCNN}}(G(z^t, I_i)) - f_{\text{LiCNN}}(I_i^t)||_2^2, \quad (9)$$

where f_{LiCNN} denotes the pre-trained Light CNN for feature extraction. As the Light CNN aims to recognize the identity information based on the facial images, it can extract the most prominent features for identity discrimination.

Therefore, the overall loss function for training FERSNet is computed as follows:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_1 \mathcal{L}_{\text{GAN}} + \lambda_2 \mathcal{L}_{\text{rec}} + \lambda_3 \mathcal{L}_{\text{cyc}} + \lambda_4 \mathcal{L}_{\text{idt}}, \quad (10)$$

where λ_1, λ_2, λ_3, and λ_4 are the hyperparameters, controlling the trade-off between the different loss terms.

IV. Experiments

In this section, we present the implementation details and the experiment settings for evaluating FERSNet. We also compare FERSNet with other state-of-the-art methods for facial expression recognition and synthesis.

A. Implementation Details

As shown in Fig. 1, the FER branch employs four convolutional blocks. Each convolutional block consists of two convolutional layers, two batch normalization [21] layers, two ReLU [22] activation layers, and one max pooling layer. For the FES branch, we employ eights convolutional blocks. The first four blocks are established with the same structure as that of the FER branch. The next four blocks form the decoder f_{dec} in Fig. 1, in which we replace the max pooling layer with the deconvolutional layer to upscale the feature maps for reconstructing the target image. The classifier f_{cls} consists of two fully connected layers, which produces a E-dimensional logit, indicating the predicted emotion probability. The transformer f_{tra} makes an inverse mapping to generate a feature map from the target E-dimensional logit, and thus it consists of fully connected layers and deconvolutional layers. In the network, all the convolutional kernels are of size 3×3, with padding 1 and stride 1, except for those in ConvFLUs, where we utilize convolutional filters of size 1×1, with padding 0 and stride 1. The max pooling layers and the deconvolutional layers consist of 2×2 kernels with stride 2 for rescaling the feature maps.

In the training phase, we randomly select a target label z_i^t with the corresponding facial image I_i^t for each training sample. We apply face alignment to all the facial images, based on the method [23]. Each aligned face is resized to 110×110. We then randomly crop a 96×96 region from the aligned facial images, and apply random mirroring and random rotation at $\{-15°, -10°, -5°, 0°, 5°, 10°, 15°\}$ to the cropped images to obtain the final training samples. For each testing image, we apply the same face alignment method [23], and resize the aligned face to 110×110. We adopt the center-crop approach to generate the final testing sample with size 96×96 from each testing image. It is worth noting that the above-mentioned preprocessing and augmentation approaches are commonly used in those methods [24]–[27] compared in our experiments.

We implement FERSNet with PyTorch [28]. During training, we set the batch size to 64, and the learning rate is set to decrease from 10^{-3} to 10^{-5} within 500 epochs based on the cosine annealing strategy [29]. We adopt the Adam [30] optimizer to minimize the objective function, defined in Eq. (10), with λ_1, λ_2, and λ_3 empirically set to 0.3, 1, and 0.5, respectively. λ_4 is set to 0.1 at the beginning, and is gradually increased to 0.5 during the training process, which also follows the design in [14]. We train the network on a Nvidia GEFORCE GTX 1080 Ti GPU, and it takes about 5 hours to train up one FERSNet model.

B. Evaluation on FER

As the proposed FERSNet aims to recognize and synthesize facial expression images, the training dataset is required to contain both the expression and the identity information. Therefore, we employ three commonly used facial expression benchmarks, which are the the Extended Cohn-Kanade dataset (CK+) [31], the Oulu-CASIA NIR&VIS facial expression database (Oulu-CASIA) [32], and the MMI facial expression

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

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 [24]	✓	Single image	95.37
DTAGN [25]	✓	Image sequence	97.25
IPA2LT [38]	✗	Single image	91.67
DeRL [26]	✓	Single image	97.30
LBVCNN [27]	✓	Image sequence	97.38
DMT-CNN [9]	✓	Single image	97.55
FERSNet	✗	Single image	97.35
FERSNet (BU-4DFE)	✓	Single image	**97.85**

database (MMI) [33], to evaluate the emotion recognition performance of FERSNet. In addition, we mainly consider six standard facial expressions, i.e. anger (An), disgust (Di), fear (Fe), happiness (Ha), sadness (Sa) and surprise (Su) in our experiments, because these six expressions are universal among humans, irrespective of their age, gender and race [34]. Moreover, one additional emotion class, i.e. contempt (Co), is included, when evaluating FERSNet on CK+. The number of video sequences in each database is summarized in terms of the expression labels in Table I.

The **Extended Cohn-Kanade** (CK+) dataset [31] consists of 593 video sequences collected from 123 subjects. We use the video sequences with the provided seven expression labels, and select the last three peak frames as the emotional faces, which results in 981 images in total. We further split the images into 10 folds based on the identity, and perform the 10-fold identity-independent cross-validation to train FERSNet using 90% of the samples and test its performance using the remaining 10% of the samples. The final recognition results are obtained by averaging the accuracy over the 10 runs.

The comparison results on CK+ are listed in Table II. As DeRL [26] was pre-trained on the BU-4DFE dataset [39], we established two versions of FERSNet, which are trained from scratch and pre-trained on BU-4DFE, to make the comparison with DeRL fair and reasonable. The BU-4DFE dataset [39] contains 60, 600 images from 101 subjects with the six standard emotion labels. We follow the settings in Sec. IV-A to pre-train FERSNet on BU-4DFE. In addition, the original DMT-CNN is designed for recognizing the eight facial expressions, i.e the original seven expressions and the neutral faces, in CK+. To make a fair comparison, we established and re-trained the DMT-CNN model, based on the settings in [9], to recognize the the original seven expressions in CK+. It

can be observed from the table that the proposed FERSNet outperforms all the other methods on CK+. Compared to DeRL, which is also a "GAN + Classifier" method for FER, FERSNet obtains an accuracy improvement of about 0.55%.

The **Oulu-CASIA** [32] database consists of video sequences under three different illumination conditions. In our experiments, we only use 480 video sequences, taken from 80 subjects under the strong illumination condition. There are six emotion labels in Oulu-CASIA, i.e. anger (An), disgust (Di), fear (Fe), happiness (Ha), sadness (Sa) and surprise (Su). For each video sequence, we select the last three peak frames with the provided emotion label to form the dataset, which results in 1, 440 images in total. Similar to CK+, the 10-fold identity-independent cross-validation is performed to evaluate FERSNet on Oulu-CASIA.

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 [25]	✓	Image sequence	81.46
IPA2LT [38]	✗	Single image	61.02
DeRL [26]	✓	Single image	88.0
LBVCNN [27]	✓	Image sequence	82.41
DMT-CNN [9]	✓	Single image	87.5
ExprGAN [12]	✓	Single image	84.72
FERSNet	✗	Single image	83.47
FERSNet (BU-4DFE)	✓	Single image	**89.23**

The comparison results on Oulu-CASIA are summarized in Table III. FERSNet pre-trained on BU-4DFE outperforms all the competitors, and it surpasses DeRL by about 1.2%. In addition, we observe that with the pre-training on BU-4DFE, FERSNet acquires a larger accuracy improvement of about 6% on Oulu-CASIA, and the FERSNet model trained from scratch achieves an accuracy of 83.47%.

The **MMI** database [33] consists of 236 video sequences, recorded from 31 subjects. Each video sequence is labelled as one of the six standard emotions. We select 208 video sequences with the frontal-view faces. Because the label is provided for each sequence and the peak expression face mainly appears in the middle, we further select the three frames in the middle of each sequence, which results in 624 images in total. We also follow the 10-fold identity-independent cross-validation strategy to evaluate the performance of FERSNet on MMI. The final accuracy is averaged over the 10 runs on MMI. We present the comparison results in Table IV.

It can be seen from the table that the sequence-based method, i.e. LBVCNN [27], achieves higher accuracy than the proposed FERSNet. Those sequence-based methods employ the temporal information, while FERSNet only considers the spatial information from a static image. We achieve about 75.3% on MMI, which is very close to the sequence-based methods and still surpasses DeLR by about 2%.

Fig. 4. The visual results of the facial expression synthesis task on CK+ and Oulu-CASIA, based on the different generative methods.

(a) CK+

(b) Oulu-CASIA

TABLE IV
THE RECOGNITION ACCURACY OF THE SIX EXPRESSIONS ON MMI BASED
ON DIFFERENT METHODS. THE BEST RESULTS ARE HIGHLIGHTED IN
BOLD.

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 [25]	✓	Image sequence	70.24
IACNN [24]	✓	Single image	71.55
DeRL [26]	✓	Single image	73.23
LBVCNN [27]	✓	Image sequence	**76.28**
FERSNet	✗	Single image	71.31
FERSNet (BU-4DFE)	✓	Single image	75.32

In the above experiments, FERSNet consistently outperforms DeLR on CK+, Oulu-CASIA, and MMI, which shows the effectiveness of jointly learning for FER and FES. Compared to the de-expression strategy in DeRL [26], the proposed facial expression editing (synthesizing) strategy is a more general case for transferring expression information, and thus FERSNet produces better results.

C. Evaluation on FES

We evaluate the performance of FERSNet on synthesizing facial images with the expected expression. In the FES experiments, we mainly consider the six basic emotions, and we present the visual results of FERSNet with the other generative models, including StarGAN [13], ExprGAN [12], CycleGAN [41] and CAAE [11] on CK+ [31] and Oulu-CASIA [32]. Among them, ExprGAN and StarGAN are the conditional generative frameworks, which are specially designed for facial expression synthesis. It is worth noting that the geometric-guided methods [14]–[16] are not included in the comparison, because they obviously use much stronger supervision than the proposed algorithm. The qualitative results are presented in Fig. 4, which clearly shows that CAAE and CycleGAN fail to produce satisfactory facial images, as they create the images with visible distortions. ExprGAN and StarGAN generally produce comparable results with our proposed method. Nevertheless, the proposed FERSNet can better maintain the color consistency on Oulu-CASIA. In addition, it can be observed from Fig. 4 that FERSNet exhibits a better ability to generate the eye and the mouth regions, and preserve the identity information. This is because the leaky gate and the memory gate in ConvFLU can more effectively transfer and

memorize the related features for the FES task. In summary, FERSNet synthesizes more visually pleasant images with less artefacts and blurs.

To further validate FERSNet on the FES task, we follow the evaluation approach in StarGAN [13] and ExprGAN [12], and establish the quantitative comparison. Specifically, we train the above-mentioned FES methods on the training set, and then perform the expression synthesis on the unseen testing set. The synthetic facial images are fed to an expression recognition network. A higher recognition accuracy on the synthetic facial images indicates more realistic expression synthesis, because the generated images lie in the same manifold of natural expressions. It is worth noting that the expression recognition network is independently trained on the original training set, containing real facial images only. Then, we employ this network to recognize the synthetic images from the different generative models. The results are summarized in Table V. We also adopt the 10-fold identity-independent cross-validation strategy in this experiment, and the final recognition results are obtained by averaging the accuracy over the 10 runs. As

TABLE V
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**

shown in the table, the proposed FERSNet achieves the highest accuracy, which demonstrates its superiority in synthesizing more realistic facial expression images.

D. Ablation Study

In order to make a comprehensive analysis on the proposed method, we present the ablation study to investigate the effectiveness of the different novel designs in FERSNet. Specifically, we establish the FERSNet model without the FES regularization and without ConvFLU, respectively. We further employ the FES branch as a data augmentation approach for the FES task. The results are listed in Table VI.

TABLE VI
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**

The FERSNet model, trained without FES regularization, becomes a single-task network, which does not adopt the multi-task learning strategy (FERSNet w/o MTL). The FERSNet model without ConvFLU (FERSNet w/o ConvFLU) becomes a hard parameter-sharing multi-task network, which does not obtain the ability to select beneficial information

when transferring features. FERSNet with FES data augmentation (FERSNet w/ FES-DA) is the original FERSNet further fine-tuned on the synthetic samples produced by the FES branch. In other words, we employ the FES branch as a data augmentation approach for fine-tuning the network. As listed in Table I, MMI and CK+ are highly imbalanced datasets, and therefore we employ the FES branch to enlarge and balance the number of samples in each emotion class. For each training sample, we synthesize all the possible expression images from it to make the training set completely balanced. In Table VI, we present the results of the "FERSNet w/ FES-DA" model using 24K images for training. It is worth noting that all the methods in Table VI are based on the original FERSNet, without pre-training on BU-4DFE. To make a fair comparison, we enlarge the kernel size and network depth in "FERSNet w/o MTL" and "FERSNet w/o ConvFLU" to make their model capacity equal to or larger than the original FERSNet.

It is obvious from the table that jointly learning with FES contributes significantly to the FER performance, as the original FERSNet outperforms "FERSNet w/o MTL" by about 3%, 10%, and 8% on CK+, Oulu-CASIA, and MMI, respectively. In addition, the proposed soft parameter-sharing strategy can further enhance the performance of MTL, because, without ConvFLU, the performance will be degraded by more than 2% on the three datasets. More importantly, the FES branch can serve as a data augmentation method for FER. The original FERSNet, fine-tuned on the synthetic samples, acquires a better generalization ability for FER. With the FES data augmentation, the model can further obtain an accuracy gain of about 0.4%, 4%, and 2.5% on CK+, Oulu-CASIA, and MMI, respectively.

V. CONCLUSION

In this paper, we have proposed a multi-task network, namely FERSNet, for facial expression recognition (FER) and facial expression synthesis (FES). FERSNet aims to solve the two tasks in parallel with the proposed convolutional feature leaky units (ConvFLU). ConvFLU adopts a soft parameter-sharing strategy, in order to filter out the useless and harmful features, when transferring information between FER and FES. Moreover, we further employ the FES branch for data augmentation to enlarge and balance the training dataset. This augmentation approach contributes to a better generalization of FERSNet for recognizing facial expressions in real-world applications. We evaluate the proposed method on three commonly used benchmarks. The experimental results have demonstrated that the proposed method achieves state-of-the-art performance, which makes it a potential solution to practical facial image analysis problems.

REFERENCES

[1] Z. Meng, P. Liu, J. Cai, S. Han, and Y. Tong, "Identity-aware convolutional neural network for facial expression recognition," in *FG*, May 2017, pp. 558–565.

[2] G. Pons and D. Masip, "Multi-task, multi-label and multi-domain learning with residual convolutional networks for emotion recognition," *Arxiv*, 2018. [Online]. Available: http://arxiv.org/abs/1802.06664

[3] R. Ranjan, S. Sankaranarayanan, C. D. Castillo, and R. Chellappa, "An all-in-one convolutional neural network for face analysis," in *FG*, May 2017, pp. 17–24.

[4] S. Reed, K. Sohn, Y. Zhang, and H. Lee, "Learning to disentangle factors of variation with manifold interaction," in *ICML*, 2014, pp. 1431–1439. [Online]. Available: http://dl.acm.org/citation.cfm?id=3044805.3045052

[5] K. Zhang, Y. Huang, Y. Du, and L. Wang, "Facial expression recognition based on deep evolutional spatial-temporal networks," *IEEE Transactions on Image Processing*, vol. 26, no. 9, pp. 4193–4203, Sep. 2017.

[6] T. Devries, K. Biswaranjan, and G. W. Taylor, "Multi-task learning of facial landmarks and expression," in *Canadian Conference on Computer and Robot Vision*, May 2014, pp. 98–103.

[7] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *CVPR*, 2017, pp. 1125–1134.

[8] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *NIPS*, 2015.

[9] Z. Ming, J. Xia, M. Luqman, J.-C. Burie, and K. Zhao, "Dynamic multi-task learning for face recognition with facial expression," in *ICCV Workshop*, Nov. 2019.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS*, 2014, pp. 2672–2680.

[11] Z. Zhang, Y. Song, and H. Qi, "Age progression/regression by conditional adversarial autoencoder," in *CVPR*, 2017, pp. 5810–5818.

[12] H. Ding, K. Sricharan, and R. Chellappa, "Exprgan: Facial expression editing with controllable expression intensity," in *AAAI*, 2018.

[13] Y. Choi, M. Choi, M. Kim, J. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2018, pp. 8789–8797.

[14] L. Song, Z. Lu, R. He, Z. Sun, and T. Tan, "Geometry guided adversarial facial expression synthesis," in *ACM Multimedia*, 10 2018, pp. 627–635.

[15] A. Lindt, P. V. A. Barros, H. Siqueira, and S. Wermter, "Facial expression editing with continuous emotion labels," in *FG*, 2019, pp. 1–8.

[16] F. Qiao, N. Yao, Z. Jiao, Z. Li, H. Chen, and H. Wang, "Geometry-contrastive gan for facial expression transfer," in *ArXiv*, 2018.

[17] S. Ruder, J. Bingel, I. Augenstein, and A. Søgaard, "Sluice networks: Learning what to share between loosely related tasks," *ArXiv*, 2017.

[18] L. Xiao, H. Zhang, W. Chen, Y. Wang, and Y. Jin, "Learning what to share: Leaky multi-task network for text classification," in *International Conference on Computational Linguistics*, Aug. 2018, pp. 2055–2065.

[19] K. Cho and et al., "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *EMNLP*, Oct. 2014, pp. 1724–1734.

[20] X. Wu, R. He, and Z. Sun, "A lightened cnn for deep face representation," *IEEE Transactions on Information Forensics and Security*, vol. abs/1511.02683, 2018.

[21] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *ICML*, 2015, pp. 448–456.

[22] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *ICML*, 2013, p. 3.

[23] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks)," in *ICCV*, 2017.

[24] C. Zhang, P. Wang, K. Chen, and J. Kämäräinen, "Identity-aware convolutional neural networks for facial expression recognition," *Journal of Systems Engineering and Electronics*, vol. 28, no. 4, pp. 784–792, Aug 2017.

[25] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *ICCV*, December 2015.

[26] H. Yang, U. Ciftci, and L. Yin, "Facial expression recognition by de-expression residue learning," in *CVPR*, June 2018, pp. 2168–2177.

[27] S. Kumawat, M. Verma, and S. Raman, "LBVCNN: local binary volume convolutional neural network for facial expression recognition from image sequences," in *CVPRW*, vol. abs/1904.07647, 2019.

[28] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," in *NIPS Workshop*, 2017.

[29] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," in *ICLR*, 2017.

[30] D. P. and J. B. Kingma, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

[31] T. Kanade, J. F. Cohn, and Yingli Tian, "Comprehensive database for facial expression analysis," in *IEEE International Conference on Automatic Face and Gesture Recognition*, March 2000, pp. 46–53.

[32] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. Pietikälnen, "Facial expression recognition from near-infrared videos," *Image Vision Comput.*, vol. 29, no. 9, pp. 607–619, Aug. 2011.

[33] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *ICME*, July 2005, pp. 5 pp.–.

[34] P. Ekman, E. Rosenberg, and E. , "What the face reveals: Basic and applied studies of spontaneous expression using the facial action coding system (facs)," *Oxford University Press*, 01 1997.

[35] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 6, pp. 915–928, 2007.

[36] A. Klaser, M. Marszalek, and C. Schmid, "A Spatio-Temporal Descriptor Based on 3D-Gradients," in *BMVC*, Sep. 2008, pp. 275:1–10.

[37] M. Liu, S. Li, S. Shan, R. Wang, and X. Chen, "Deeply learning deformable facial action parts model for dynamic expression analysis," in *ACCV*, D. Cremers, I. Reid, H. Saito, and M.-H. Yang, Eds., 2015, pp. 143–157.

[38] J. Zeng, S. Shan, and X. Chen, "Facial expression recognition with inconsistently annotated datasets," in *ECCV*, September 2018.

[39] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, and Peng Liu, "A high-resolution spontaneous 3d dynamic facial expression database," in *FG*, April 2013, pp. 1–6.

[40] M. Liu, S. Shan, R. Wang, and X. Chen, "Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition," in *CVPR*, June 2014, pp. 1749–1756.

[41] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networkss," in *ICCV*, 2017.