Facial Expression Recognition by Expression-Specific Representation Swapping

Jie Lei[1(✉)], Zhao Liu[2], Zeyu Zou[2], Tong Li[2], Juan Xu[2], Zunlei Feng[3], and Ronghua Liang[1]

[1] Zhejiang University of Technology, Hangzhou 310023, People's Republic of China
{jasonlei,rhliang}@zjut.edu.cn
[2] Ping An Life Insurance Of China, Ltd.,
Shanghai 200120, People's Republic of China
{liuzhao556,zouzeyu313,litong300,xujuan635}@pingan.com.cn
[3] Zhejiang University, Hangzhou 310027, People's Republic of China
zunleifeng@zju.edu.cn

Abstract. In the field of facial expression recognition (FER), various FER systems have been explored to encode expression information from facial representations. Although significant progress has been made towards improving the expression classification, challenges due to the large variations of individuals and the lack of consistent annotated samples still remain. In this paper, we propose to disentangle facial representations into expression-specific representations and expression-unrelated representations with a representation swapping procedure, called SwER. First, we adopt a variational auto-encoder (VAE) structure to obtain latent vectors (*i.e.*, facial representations) from face images. Next, the representation swapping procedure is introduced for paired face images to disentangle the expression-specific representations from facial representations. Finally, the expression-specific representations and the expression-unrelated representations are jointly learned for facial expression recognition and face comparison tasks, respectively. In this way, better facial representations are obtained by discarding unrelated factors, and the expression-specific representations are more independent. The proposed method has been evaluated on five databases, CK+, Oulu-CASIA, MMI, RAF-DB, and AffectNet. The experimental results demonstrate the superior performance of the proposed method.

Keywords: Facial expression recognition · Representation swapping

1 Introduction

Facial expression is an essential factor in conveying human emotional states and intentions. As a consequence, numerous studies have been conducted on facial

J. Lei and Z. Liu–contributed equally.

I. Farkaš et al. (Eds.): ICANN 2021, LNCS 12892, pp. 80–91, 2021.
https://doi.org/10.1007/978-3-030-86340-1_7

expression recognition for potential use in sociable robots, medical treatment, driver fatigue surveillance, and many other human-computer interaction systems.

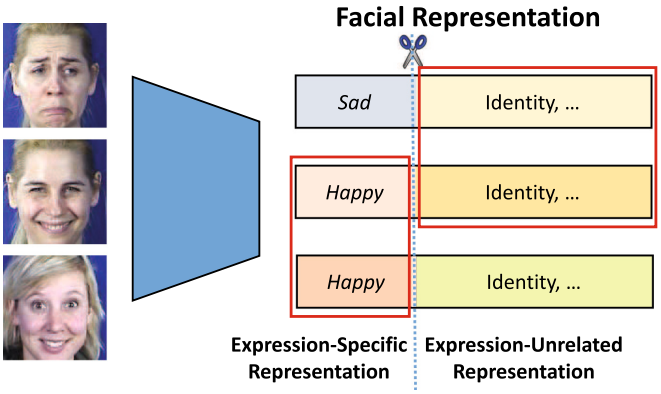


Fig. 1. Facial representation can be disentangled into expression-specific representations (the former part) and expression-unrelated representations (the latter part). The first and the second faces are similar in the identity, while the second and the third faces are similar in the expression.

There has been significant progress towards improving the facial expression classification, from handcrafted feature classification, shallow learning, to deep learning [7]. However, the existing well-constructed FER systems still face two challenges: the large variations of individuals and the lack of consistent annotated samples. There are many expression unrelated variations in face images, such as illumination, head pose, age, gender, and background, *i.e.*, facial expressions may appear different for people with different personalities. These disturbances are nonlinearly confounded with facial expressions and address large intra-class variability, making it hard to learn effective expression-specific representations. Meanwhile, as the subjectivity of human annotators and the ambiguous nature of the expression labels, the annotation inconsistency is widespread and consistent annotated samples are limited.

Researches have shown that people are capable of recognizing facial expressions by comparing a subject's expression with a reference expression [16]. In other words, a facial expression can be disentangled in the image representation space. Inspired by this fact, we introduce a swapping procedure in paired face image representations for expression-specific representation learning. We employ a VAE structure to learn latent vectors as facial representations from face images. The facial representations are divided into two parts (Fig. 1), with the former part for facial expression recognition and the latter part for face comparison. During the joint training process, face image pairs are selected as inputs. In this way, we can make full use of limited but consistent annotated samples extracted from face image sequences. For facial expression recognition, we swap the former

part of the paired image representations to reconstruct the corresponding face images with expected expressions, thus making the former part more specified for expression. For face comparison, the network is further trained based on the differences of the latter part in the representations to predict whether the two input face images share the same identity. As the expression is irrelevant to the identity, the latter part restrains the expression-specific representation, improving the performance of disentangling for the former part in return.

In contrast to the previous methods [16], which focused on introducing well-designed auxiliary blocks or layers to enhance the expression-related representation capability directly, our proposed SwER framework learns the relatively easier facial representations on facial expression datasets and then disentangles more independent expression-specific representations, with jointly learning of facial expression recognition and face comparison tasks.

The major contributions of this paper are two-fold. Firstly, we introduce a representation swapping procedure for disentangling expression-specific representations from face image representations. Secondly, we propose jointly learning of facial expression recognition and face comparison tasks from paired face images, thus taking full advantage of limited consistent annotated samples and improving the disentanglement performance.

2 Related Work

To reduce the impacts of widespread expression-unrelated variations in learning expression-specific representations, several studies have proposed well-designed auxiliary modules to enhance the foundation architecture of deep models. Yao *et al.* [17] proposed HoloNet with three critical considerations in the network design. Li *et al.* [9] proposed an end-to-end trainable Patch-Gated Convolution Neural Network (PG-CNN) that can automatically percept the possible regions of interest on the face. Another area for expression-specific representation learning focuses on facial expression data. Wang *et al.* [15] proposed Self-Cure Network (SCN) to suppress the uncertainties efficiently and prevent deep networks from over-fitting uncertain face images. In [18], the authors proposed an end-to-end trainable LTNet to discover the latent truths with the auxiliary annotations from different datasets. There are other existing works that suggest facial expression recognition could benefit from using a reference image. Yang *et al.* [16] recognized facial expression by learning the residual expressive component in the generative model. Kim *et al.* [5] employed a contrastive representation in the networks to extract the feature level difference between a query face image and a neutral face image. Zhao *et al.* [20] presented a novel peak-piloted deep network (PPDN) that used the peak expression (easy sample) to supervise the non-peak expression (hard sample) of the same type and from the same subject.

The above works focus on directly learning expression-specific representation or expression-specific difference to a reference face image, which is relatively hard for training with a lack of diverse samples for widespread expression unrelated variations. Unlike these works, we propose to learn facial representation at first,

which is relatively easy on limited consistent annotated samples. The expression-specific representation is further disentangled from the facial representation. The recent utility of representation disentangling shows success in learning disassembled object representation from images [4]. Lin *et al.* [10] proposed SPACE to factorize object representations of foreground objects and decompose background segments of complex morphology. Comparing with the object, the expression is implicitly and dispersive in the image. We adopt an auxiliary expression-unrelated task of face comparison to suppress the expression-specific representation on the latter part of the facial representation. In return, the former part can concentrate on learning the expression-specific representation.

3 Proposed Method

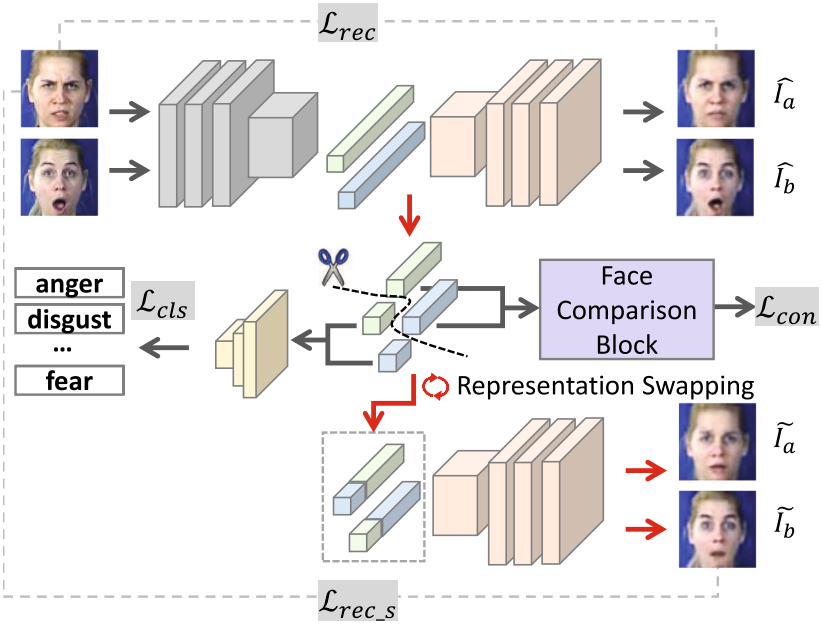


Fig. 2. Framework of the proposed SwER method, which is composed with two reconstruction modules, an expression classification module, and an auxiliary face comparison block.

The framework of our proposed method - SwER is illustrated in Fig. 2, where the network takes a pair of face images as inputs. As shown in Fig. 2, SwER contains three learning processes: the first is learning facial representations from face images; the second is learning expression-specific representations (the former part) disentangled from facial representations; the third is learning to suppress the expression-specific representations on the latter part of facial representations. In this section, we illustrate details of these learning processes.

3.1 Paired Face Images

We take pairs of face images $\langle I_a, I_b \rangle$ as inputs. Specifically, we consider two types of pairs. One is a pair of face images with the same identity and different expressions, the other is a pair of face images with different identities and the same expression. Here, we use $D(I_a, I_b) = 1$ and $E(I_a, I_b) = 1$ to denote $\langle I_a, I_b \rangle$ sharing the same identity or expression, respectively. In Sect. 3.3, we will demonstrate that the supervision information for reconstructed images after expression-specific representation swapping can naturally derived from the inputs, *i.e.*, $\langle I_a, I_b \rangle$.

In the experiments, we sample face images from image sequences. A face image sequence typically begins with a neutral expression and reaches a peak near the middle before returning to the neutral expression. The expression annotations are relatively consistent as frames in the same sequence can be taken as reference images for each other. However, the number of sequences is relatively smaller in comparison to static images. By adopting pairs of images, we can significantly enlarge the number of training samples.

3.2 Facial Representation Learning

A variational auto-encoder structure [14] is exploited to generate a good facial representation from a face image. Without loss of generality, this structure contains an encoder f_E and a decoder f_D. The input face images $\langle I_a, I_b \rangle$ are mapped from image space to the latent representation space by f_E, denoted as $\langle R_a, R_b \rangle$. The latent image representations $\langle R_a, R_b \rangle$ are then mapped back by decoder f_D to reconstruct the image pair. The objective is to simultaneously optimize f_E and f_D for minimizing the reconstruction error:

$$\mathcal{L}_{rec} = \left\| I_a - \widehat{I}_a \right\|_2^2 + \left\| I_b - \widehat{I}_b \right\|_2^2, \tag{1}$$

where $\widehat{I}_a$ and $\widehat{I}_b$ are reconstructed face images. All the input image pairs $\langle I_a, I_b \rangle$ are pre-processed by face detection and face alignment, so the latent representations $\langle R_a, R_b \rangle$ can be referred as facial representations.

3.3 Expression-Specific Representation Swapping

The facial representations $\langle R_a, R_b \rangle$ are divided into two parts: $\left[R_a^E, R_a^U\right]$ and $\left[R_b^E, R_b^U\right]$, respectively. The former parts R_a^E and R_b^E are referred as expression-specific representations. The latter parts R_a^U and R_b^U are expression-unrelated facial representations.

We introduce a swapping procedure to disentangle $\left\langle R_a^E, R_b^E \right\rangle$ from $\langle R_a, R_b \rangle$. After swapping R_a^E and R_b^E, the hybrid latent representations $R_a^{'} = \left[R_b^E, R_a^U\right]$ and $R_b^{'} = \left[R_a^E, R_b^U\right]$ are decoded by f_D and reconstructed as hybrid images $\widetilde{I}_a$ and $\widetilde{I}_b$, respectively.

For pairs $\langle I_a, I_b \rangle$ where $D(I_a, I_b) = 1$ and $E(I_a, I_b) = 0$, the desired reconstruction images for $R_a^{'}$ and $R_b^{'}$ should swap the expression for each other. As we encourage the representation of different expressions to be discriminated, we use $\langle I_b, I_a \rangle$ for supervision:

$$\mathcal{L}_{rec\_s} = \left\| I_b - \widetilde{I}_a \right\|_2^2 + \left\| I_a - \widetilde{I}_b \right\|_2^2. \tag{2}$$

For pairs $\langle I_a, I_b \rangle$ where $D(I_a, I_b) = 0$ and $E(I_a, I_b) = 1$, the desired reconstruction images for $R_a^{'}$ and $R_b^{'}$ should be similar to the inputs. In other words, the expression-specific representation is personality unrelated. We encourage the representation of the same expression to be similar for different people. $\langle I_a, I_b \rangle$ are used for supervision as:

$$\mathcal{L}_{rec\_s} = \left\| I_a - \widetilde{I}_a \right\|_2^2 + \left\| I_b - \widetilde{I}_b \right\|_2^2. \tag{3}$$

Expression-specific representation swapping aims to model the expression factor that affects the appearance of face images. If the expression-specific representation is well disentangled, the change of expression only causes the change of the face on the expression factor, while the other factors are uninfluenced.

$\left\langle R_a^E, R_b^E \right\rangle$ are used for expression classification, the loss function is

$$\mathcal{L}_{cls} = -\sum_{r}^{\{a,b\}} \log\left(\frac{\exp(p^{(k_r)}(R_r^E))}{\sum_i^K \exp(p^{(i)}(R_r^E))}\right), \tag{4}$$

where $p^{(i)}(\cdot)$ is the i-th expression predicted probability of the classifier, K is the total number of facial expression expression classes, and k_a and k_b are the target expressions for I_a and I_b, respectively.

3.4 Auxiliary Face Comparison Block

In further, we introduce an auxiliary face comparison block for an expression unrelated task - face comparison, where a change of expression shall not affect the identity. On one hand, better facial representations are obtained by paying more attention to describing the face. On the other hand, as we use $\left\langle R_a^U, R_b^U \right\rangle$ for the comparison, the expression-specific representations are suppressed on the latter representations. In return, more expression-specific representations are contained in $\left\langle R_a^E, R_b^E \right\rangle$.

Contrastive loss [2] is used for the auxiliary block as:

$$\mathcal{L}_{con} = D(I_a, I_b)d^2 + (1 - D(I_a, I_b)) \max(m - d, 0)^2, \tag{5}$$

where $d = \left\| R_a^U - R_b^U \right\|_2$ is the distance between two face images in the representation space, and m is a threshold for the distance.

3.5 Complete Algorithm

In summary, the total loss $\mathcal{L}$ is a combination of the above modules. The inputs are $\langle I_a, I_b \rangle$, the annotated expression labels $\langle k_a, k_b \rangle$, and $D(I_a, I_b)$. $\langle k_a, k_b \rangle$ and $D(I_a, I_b)$ are used in facial expression classification and face comparison, respectively. The facial representation learning and expression-specific representation swapping take $\langle I_a, I_b \rangle$ as supervision information. The total loss is given as follows:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{rec} + \lambda_2 \mathcal{L}_{rec\_s} + \lambda_3 \mathcal{L}_{cls} + \lambda_4 \mathcal{L}_{con}, \tag{6}$$

where λ_1, λ_2, λ_3, and λ_4 are balanced parameters which are used to control the influence of different learning processes.

4 Experiments

4.1 Datasets and Setting

The proposed SwER approach is evaluated on five public facial expression datasets, inculuding CK+ [11], MMI [13], Oulu-CASIA [19], RAF-DB [8], and AffectNet [12].

CK+ contains 593 video sequences collected from 123 subjects. Among them, 327 video sequences with 118 subjects are labeled as one of seven expressions, *i.e.*, *anger* (AN), *disgust* (DI), *fear* (FE), *happiness* (HA), *sadness* (SA), *surprise* (SU), and *contempt* (CO). Each sequence starts with a peak expression. We chose the first frame as the neutral face (NE) and the last three frames as the expressive face, resulting in 1307 images with 1047 for training and 260 for testing. MMI has 236 sequences with expressions recorded from 30 subjects, where each sequence starts with a neutral face, shifts to a peak expression, and return to a neutral face in the end. In our experiments, for each sequence, the first two images are selected as neutral faces while the middle one-fifth part are chosen as expressive faces. In total, we have 1103 images for training and 399 images for testing. Oulu-CASIA has 480 sequences captured from 80 objects. We use the cropped face images provided by the author, resulting in 29932 images with 21070 images for training and 8862 images for testing. The annotated labels for MMI and Oulu-CASIA are six basic expressions (except for *contempt*) and neutral.

RAF-DB is divided into training and test sets with a size of 12,271 and 3,068, respectively. AffectNet contains more than 400k annotated images. We select 19,239 images for training and 2,518 images for testing, all of which are labeled with six basic expressions and neutral .

For CK+, MMI, Oulu-CASIA, we separate the training set and the testing set by subjects, *i.e.*, the subjects in the two subsets are mutually exclusive. To generate image pairs, we randomly select pairs from the training set on the condition that each sample will be included for at least once. In total, we obtain 24,994, 67,779 and 147,490 pairs for the three datasets, respectively. Since the identities of subjects are not accessible on RAF-DB and AffectNet, we use CK+ for pre-training and conduct fine-tuning on the expression classification module

with their training sets. The face images are pre-processed by face detection and face alignment [3]. The basic variational auto-encoder structure [14] is adopted, with the dimensions for the face representation and expression-specific representation are set as 512 and 64, respectively. We use the Adam optimizer with a learning rate of 0.0001. The parameters $\lambda_{\{1-4\}}$ are empirically chosen from the scales of $\{0.01, 0.1, 0.5, 1, 1.5, 2, 10\}$ and finally set as $\lambda_1 = 1$, $\lambda_2 = 2$, $\lambda_3 = 0.5$, and $\lambda_4 = 0.1$ for the loss function.

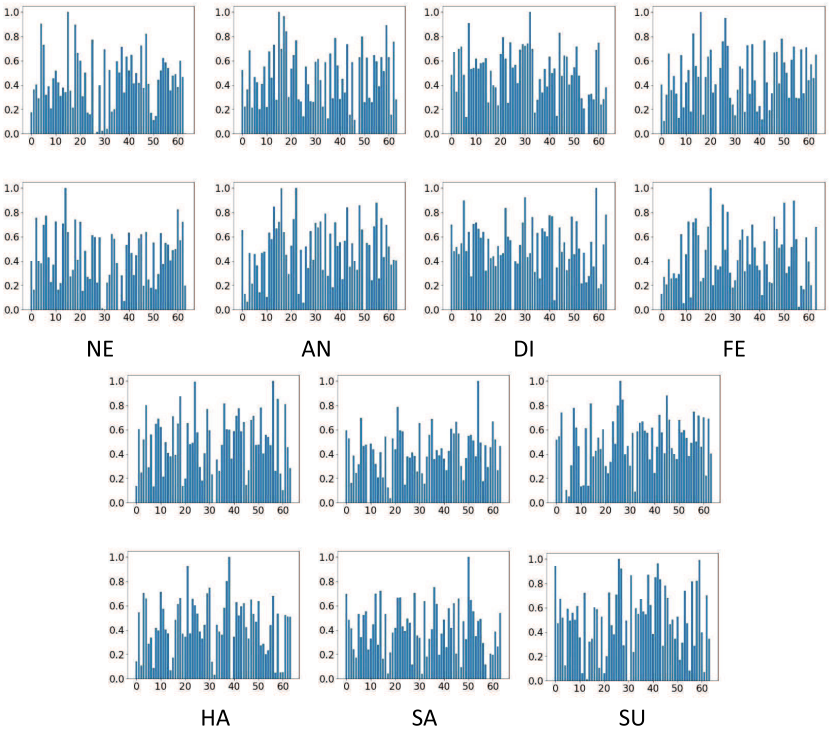


Fig. 3. The generated normalized average seven expression-specific representations of two subjects on CK+. The expression-specific components are similar for the same expression and distinguishable among other expressions for different subjects.

4.2 Results

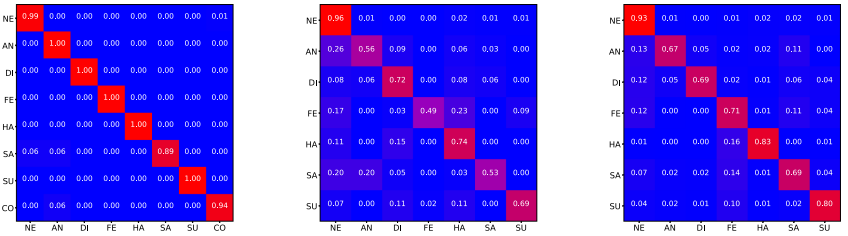
In Fig. 3, we demonstrate an example of the generated expression-specific representations of two subjects on CK+. The average representations for *neutral*, *anger*, *disgust*, *fear*, *happiness*, *sadness*, and *surprise* are displayed, where each histogram is calculated and normalized from all samples with the same expression for the subject. As we can see, the expression-specific components are similar for the same expression and distinguishable among other expressions for different subjects.

Table 1. The average accuracies of expression recognition on CK+, MMI, and Oulu-CASIA, where $\text{SwER}^{-}_{rec\_s}$ and SwER^{-}_{con} are variants of the proposed SwER for ablation studies.

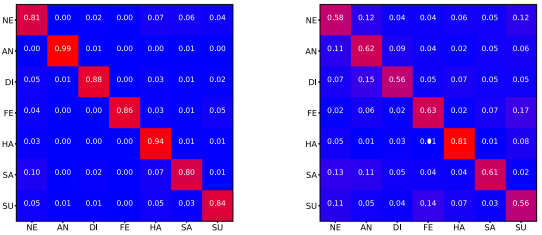
Dataset	CK+	MMI	Oulu-CASIA
FRAME [6]	0.9077	0.5689	0.5971
LTNet [18]	0.9385	0.6065	0.5837
FMPN [1]	0.9731	0.4390	0.5330
SCN [15]	0.9769	0.6717	0.7512
$\text{SwER}^{-}_{rec\_s}$	0.9173	0.5514	0.5349
SwER^{-}_{con}	0.9474	0.6424	0.7257
SwER	**0.9846**	**0.6729**	**0.7708**

Table 2. The average accuracies of expression recognition on RAF-DB and AffectNet.

Dataset	RAF-DB	AffectNet
DLP [8]	0.6874	0.4865
LTNet [18]	0.7864	0.5306
FMPN [1]	0.6610	0.4527
SCN [15]	0.8589	0.5786
SwER	**0.8750**	**0.6250**



(a) CK+ (b) MMI (c) Oulu-CASIA



(d) RAF-DB (e) AffectNet

Fig. 4. Confusion Matrixes on CK+, MMI, Oulu-CASIA, RAF-DB, and AffectNet, where the horizontal axis and vertical axis are predicted label and groundtruth label, respectively.

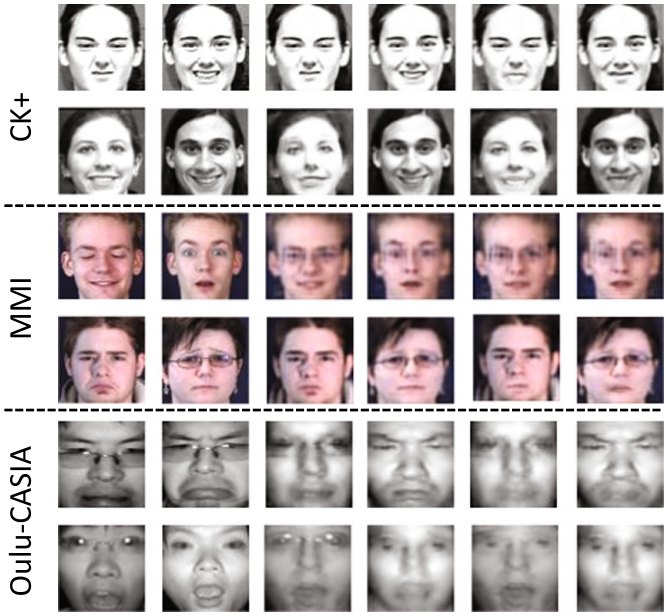


Fig. 5. Face image reconstruction on CK+, MMI, and Oulu-CASIA The first and the second columns are original input face images, the third and the fourth columns are reconstructed images, and the fifth and the sixth columns are reconstructed images after expression-specific representation swapping.

Figure 5 illustrates face image reconstruction on CK+, MMI, and Oulu-CASIA, where the first and the second columns are original input face images, the third and the fourth columns are reconstructed images, and the fifth and the sixth columns are reconstructed images after expression-specific representation swapping. As shown in Fig. 5, the reconstructed face images are similar to the inputs, indicating the facial representation could well describe the face. For the first example of each dataset where two face images share the same identity and have different expressions, the expressions of the reconstructed face images after swapping are similar to the expression in the other input face image. For example, the expressions for the input face images are *disgust* and *happiness* in the first row. After expression-specific representation swapping, the *disgust* face is happier and the *happiness* face is getting disgustting. For the second example of each dataset where two face images share the same expression and have different identity, the reconstructed face images after swapping are similar to the inputs. From these examples, we can conclude that the expression-specific representation is disentangled.

The average accuracies on expression recognition are shown in Table 1 and Table 2. The results are reported as the average of 10 runs. Our SwER method achieves the highest accuracy compared to those of state-of-the-art methods, including FRAME [6], LTNet [18], FMPN [1], and SCN [15]. The confusion

matrixes are also provided in Fig. 4, where the proposed SwER performs very well in recognizing *neutral*, *disgust*, and *happiness*, while *sadness* shows the relatively low recognition rate, which is mostly confused with *neutral*.

4.3 Ablation Study

In SwER, the total loss function is composed of four items. To verify the necessities of the modules of reconstruction with swapping and auxiliary face comparison, we conduct ablation studies by removing $\mathcal{L}_{rec\_s}$ and $\mathcal{L}_{con}$, respectively, denoted as $\text{SwER}^{-}_{rec\_s}$ and SwER^{-}_{con}.

The average accuracies on CK+, MMI, and Oulu-CASIA for $\text{SwER}^{-}_{rec\_s}$ and SwER^{-}_{con} are included in Table 1. It is noticeable SwER achieves the best classification performance than other variants, which demonstrates that the loss of reconstruction with swapping and face comparison can improve the disentanglement performance of expression-specific representations.

5 Conclusion

In this paper, we propose SwER for facial expression recognition by disentangling expression-specific representations from facial representations. SwER is composed with two reconstruction modules, an expression classification module, and an auxiliary face comparison block. The experimental results demonstrate the superior performance of the proposed method over other state-of-the-art methods. Our future work will incorporate the expression-specific representations with temporal information for addressing the issues of AU detection.

Acknowledgements. This work was supported in part by the National Key Research and Development Program of China (No. 2020YFB1707700) and the National Natural Science Foundation of China (No. 62036009).

References

1. Chen, Y., Wang, J., Chen, S., Shi, Z., Cai, J.: Facial motion prior networks for facial expression recognition. In: VCIP (2019)
2. Chopra, S., Hadsell, R., Lecun, Y.: Learning a similarity metric discriminatively, with application to face verification. In: CVPR (2005)
3. Deng, J., Guo, J., Zhou, Y., Yu, J., Kotsia, I., Zafeiriou, S.: RetinaFace: single-stage dense face localisation in the wild. In: arXiv preprint arXiv:1905.00641 (2019)
4. Feng, Z., et al.: One-sample guided object representation disassembling. In: NeurIPS (2020)
5. Kim, Y., Yoo, B., Kwak, Y., Choi, C., Kim, J.: Deep generative-contrastive networks for facial expression recognition. In: CVPR (2017)
6. Kuo, C.M., Lai, S.H., Sarkis, M.: A compact deep learning model for robust facial expression recognition. In: CVPRW (2018)
7. Li, S., Deng, W.: Deep facial expression recognition: a survey. IEEE Trans. Affect. Comput. (99) (2018)

8. Li, S., Deng, W., Du, J.P.: Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: CVPR (2017)
9. Li, Y., Zeng, J., Shan, S., Chen, X.: Patch-gated CNN for occlusion-aware facial expression recognition. In: ICPR (2018)
10. Lin, Z., et al.: SPACE: unsupervised object-oriented scene representation via spatial attention and decomposition. In: ICLR (2020)
11. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Matthews, I.: The extended Cohn-Kanade Dataset (CK+): a complete dataset for action unit and emotion-specified expression. In: CVPRW (2010)
12. Mollahosseini, A., Hasani, B., Mahoor, M.H.: AffectNet: a database for facial expression, valence, and arousal computing in the wild. In: IEEE Transactions on Affective Computing (2017)
13. Pantic, M.V.: Induced disgust, happiness and surprise: an addition to the mmi facial expression database. In: Proceedings 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect (2010)
14. Sohn, K., Yan, X., Lee, H., Arbor, A.: Learning structured output representation using deep conditional generative models. In: NeurIPS (2015)
15. Wang, K., Peng, X., Yang, J., Lu, S., Qiao, Y.: Suppressing uncertainties for large-scale facial expression recognition. In: CVPR (2020)
16. Yang, H., Ciftci, U., Yin, L.: Facial expression recognition by de-expression residue learning. In: CVPR (2018)
17. Yao, A., Cai, D., Hu, P., Wang, S., Chen, Y.: HoloNet: towards robust emotion recognition in the wild. In: ICMI (2016)
18. Zeng, J., Shan, S., Chen, X.: Facial expression recognition with inconsistently annotated datasets. In: ECCV (2018)
19. Zhao, G., Huang, X., Taini, M., Li, S.Z., Pietikälnen, M.: Facial expression recognition from near-infrared videos. In: Image and Vision Computing (2011)
20. Zhao, X., Liang, X., Liu, L., Li, T., Yan, S.: Peak-piloted deep network for facial expression recognition. In: ECCV (2016)