Single-Image Facial Expression Recognition Using Deep 3D Re-Centralization

Zhipeng Bao
Tsinghua University, China
bzp15@mails.tsinghua.edu.cn

Shaodi You
Data61-CSIRO, Australia
shaodi.you@data61.csiro.au

Lin Gu
National Institute of Infomatics, Japan
ling@nii.ac.jp

Zhenglu Yang
Nankai University, China
yangzl@nankai.edu.cn

Abstract

Facial expression recognition (FER) aims to encode expression information from faces. Previous studies often hold the assumption that human subjects should properly face the camera. Such a laboratory-controlled condition, however, is too rigid for in-wide applications. To tackle this issue, we propose a single image facial expression recognition method that is robust to face orientation and light conditions. We achieved this by proposing a novel face re-centralization method by reconstructing a 3D face model from a single image. We then propose a novel end-to-end deep neural network that utilizes both re-centralized 3D model and landmarks for FER task. A comprehensive evaluation on three real-world datasets illustrates that the proposed model outperforms the state-of-the-art techniques in both large-scale and small-scale datasets. The superiority of our model on effectiveness and robustness is also demonstrated in both laboratory conditions and wild images.

Ground Truth	**Surprise**
Our Method	**Surprise**
(Li et al. 2017)	Fear
(Jung et al. 2015)	Surprise
(Zhang et al. 2017)	Fear
(Fabian et al. 2016)	Neutral

(a)

Ground Truth	**Surprise**
Our Method	**Surprise**
(Li et al. 2017)	Surprise
(Jung et al. 2015)	Fear
(Zhang et al. 2017)	Sad
(Fabian et al. 2016)	Fear

(b)

Figure 1. Facial images in challenging conditions and the classification results: (a) an image with side face and (b) an image with dark shadow.

1. Introduction

Facial expression recognition (FER) is an important computer vision task. According to the subjects of the problem, the tasks can be divided into two categories: image sequence-based and single image-based problems [8, 20]. The single image-based approaches provide the basic model for facial expression recognition problems, and they can be extended to sequence-based approaches with a few modifications.

Most of the existing works are based on the assumption that the given faces should be in the right orientation and lighting conditions (i.e., right front of the camera and in normal lighting conditions). These methods work well for standard datasets in laboratory conditions like CK+ [15], JAFFE [16] and OULU-CASIA [29].

However, the facial images collected in real practice are usually in various orientations and lighting conditions. Different orientations can affect the extraction of key facial features; likewise, the shadow caused by different lighting conditions can also cause some confusions. Owing to these deviations, the previous models, especially the neural network-based ones, have resulted in inaccurate findings. Figure 1 shows two sample images under challenging conditions and the classification results of some popular methods.

To tackle this problem, we propose to use 3D face reconstruction to re-centralize the facial images for FER task. It is based on the understanding that facial expression is highly related to face geometry (facial muscle movement) while 3D geometry is invariant from shading and orientation. As illustrated in Figure 2, our method could robustly reconstruct the 3D face from a single image under various orientation and lighting condition. After that, we re-align the face such that it is rightly facing the camera and then generate the shading. In such a case, the facial expression would be much easier to infer as the shading could efficiently depict the geometric features.

As shown in Figure 6, we have proposed a full pipeline for single-image FER tasks which integrates our 3D re-centralization sub-network and 2D facial landmarks sub-network into a multi-modal deep neural network. The facial landmarks characterize the key points in the face, and the 3D sub-network takes advantage of the best perspective and the geometry of the image for FER. Our experiments on three widely used datasets show the superior performance of the proposed approach among state-of-the-art in terms of both accuracy and robustness.

In this paper, our contributions are three-fold:

- We propose a novel 3D facial reconstruction method to centralize the single still face image. This process can significantly reduce the influence of orientations and shadows for a wide range of FER tasks.
- We propose a novel end-to-end neural networks for single image-based FER. This method incorporates the re-aligned 3D facial geometry and landmark features to achieve robust expression detection.
- The experimental results on three widely used datasets demonstrate that our model outperforms the existing state-of-the-art approaches in terms of accuracy and robustness.

2. Related Work

2.1. DNN for FER Problems

In recent years, well-known deep neural network (NN) algorithms such as convolution neural networks (CNNs), recurrent neural networks (RNNs) and the long short-term memory (LSTM) models have gained popularity in FER tasks [25, 10, 21]. With the development of deep learning, complicated deep neural networks (DNNs) have further improved the performance, especially for large datasets. [8] first applies DNNs in combination with landmark features. Mollahosseini *et al.* use three inception structures in convolution for FER to extract the deeper features of the images [17]. Zhao *et al.* propose a Peak-Piloted network, which is proven to be robust and effective for sequence-based FER problems [30]. The work of [11] considers the context information and their method works well for processing wild facial images. Zhu *et al.* improve the traditional DNN model by combining RNN and CNN to carry out the classification [32]. [24] supposes the facial expression can be seen as the expressive component and the neutral component, and they designed a de-expression residue learning method to separate these two components. Zhang *et al.* consider the influence of poses and perspectives and propose an adversarial network for FER [27]. [19] provides a novel method to compare the similarity among different facial expressions. Jia *et al.* [7] apply local low-rank label correlations in their networks.

2.2. Facial Landmarks

Facial landmarks are the key points in a human face. These points include the contour of the face and the positions of the eyes, mouth, nose, and eyebrows. Numerous studies have been conducted on extracting facial landmark points, and several robust methods have been proposed [22, 9]. Many approaches for FER tasks based on landmark features have also gained great achievements.

For example, [3] builds a system only based on landmark points. They use both geometric and shading features generated from landmark points to conduct the classification. This approach has achieved high accuracy for specific datasets. However, it is not robust to classify the expressions merely based on landmark points because of the limited information they can provide. These points are generally taken as a secondary channel to provide auxiliary information for FER tasks in other situations [8, 23].

2.3. 3D Facial Features for FER Systems

Although 3D features have been applied to FER systems for many years, most of the previous works use either 3D landmarks or 3D coordinates to carry out the classification [2]. [27] also takes advantage of the 3D model, but they use the 3D model to generate multi-poses and multi-perspectives facial projection to improve the FER.

Different from the previous approaches, the proposed method uses the 3D model to re-align the facial images, which can reduce the influence of orientation and lighting, thus improve the performance of expression classifiers.

3. Single Image 3D Facial Re-centralization

As mentioned above, in-wild facial images may be of various positions and perspective. They may also contain different degrees of facial shadows. Figure 3 lists several examples of real-world facial images.

Noted that the 3D geometry, which is highly related to the expressions, is invariant from shadows and orientations. We propose a pipeline to reconstruct and align the 3D face model from a single image through the following three steps: 3D model generation, face re-centralization and shading generation. The whole process is demonstrated in Figure 2.

3.1. 3D Model Generation

The foundation of the whole process is to build a reliable 3D model from one single image. We adopt the method

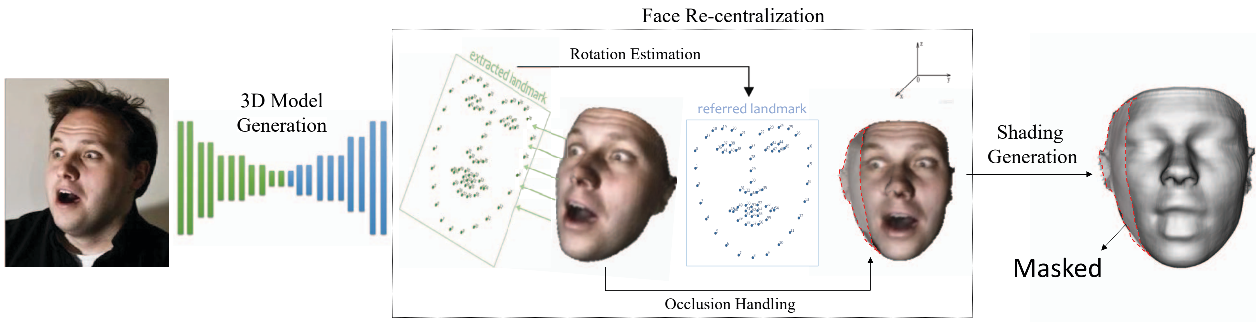


Figure 2. The whole process of 3D Face Re-centralization.

Figure 3. Potential problems in real-world facial expression recognition: different orientations (the first two rows) and different shadows (the last row).

initially proposed by Feng *et al*. [4] in our study. This model applies a Position map Regression Network (PRN) with an encoder-decoder architecture to reconstruct the 3D face. This network incorporates the context information by tracking the facial landmarks. As shown in Figure 2, this step delivers a position-color point cloud from a single image. We adopt a pre-trained model on 300-WLP dataset [31] in this paper.

3.2. Face Re-centralization

We re-centralize the 3D face by tracking the 3D landmarks. Assume $\mathbf{P}$ records the positions of landmark points in the generated face and $\mathbf{P}^*$ denotes the locations of referred landmarks of centralized face. Then the problem is to find a Matrix $\mathbf{R}$ so that:

$$\mathbf{R} = \arg\min_{\Omega} \|\Omega\mathbf{P} - \mathbf{P}^*\|. \tag{1}$$

Since the problem is also essentially a 3D rotation problem and $\mathbf{R}$ is the rotation matrix, $\mathbf{R}$ would be an orthogonal matrix. Then this problem becomes an orthogonal Procrustes problem [5] and could be solved by singular value decomposition (SVD) of $\mathbf{P}^*\mathbf{P}^{\mathbf{T}}$.

The solution of this equation requires at least six pairs of corresponding points. In our method, sixty-eight pairs of points extracted by [22, 9] are used.

A potential concern for the re-centralization process is the extravagant distortions when images are of extreme orientations. However, in practice, these images are also challenging for state-of-the-art FER methods. Compared with the original 2D image, the re-aligned face model can provide more reliable information for expression recognition even though it is not completely precise. We show some examples in Figure 4.

Above model-based 3D reconstruction and re-centralization provides us the occluded face part. However, we do not use it in the FER task, because it still lacks reliable information through hallucination. Specifically, before the 3D rotation, we check the 3D vertices first. For the points in the point cloud, if several points share the same x, y coordinates, we only keep the outer-most two points and mask the others. Then the generated parts, which are invisible from the original image, will be discarded.

3.3. Shading Generation

The re-centralized 3D model is represented by a 3D point cloud. A straightforward approach is to build a 3D cube based on the point cloud. However, introducing the 3D cube requires considerable memory and calculation time, which poses a heavy burden for real-time expression recognition.

Therefore, we generate 2D shading image to represent 3D geometry to support our FER network. Based on Lambertian reflection [1], the shading image of each point on the given lighting can be modeled as:

$$I = \mathbf{N} \cdot \mathbf{L}, \tag{2}$$

where I is the shading image density, N is the surface normal, and L represents incoming light direction. To reduce the influence of lighting, we relight the face with a standard L.

Now we estimate the normals $\mathbf{N}$ of the given 3D face. For each point in the centralized point, we select n nearest

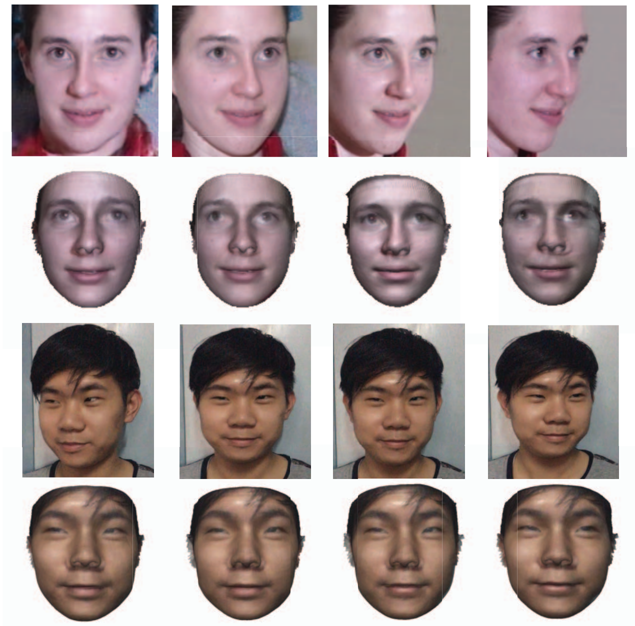


Figure 4. 3D face re-centralization better aligns the face and preserves the facial expression: images in the first two rows are selected from existing database, those in the last two rows are collected by us. The occluded part are also visualized for illustration but will not be used in the estimation.

points of it. Then we fit a plane with the minimum error based on them, the normal of that plane is considered as the normal of the selected point. In our experiments, we set n as 10.

Once we obtained the normals, we relight the face with $L = [1,\ 0,\ 0]$, which is right facing the front face. Then we can generate a shading $(N \cdot L)$ map [18] that effectively reflects geometric facial features as shown in Fig 5.

4. Facial Expression Recognition

Based on the proposed face re-centralization method, we propose a novel end-to-end deep model for single image facial expression recognition. As shown in Figure 6, our model integrates the re-aligned 3D facial geometry and facial landmarks to recognize facial expression from a single image.

4.1. 3D Re-centralization Sub-network

Given the generated shading image of the size 100×100 , we extract the features with 3 convolution layers of the filter numbers as 256, 128 and 64. The kernel size of each convolution layer is 3, and the stride is 2. Then two fully-connected layers are applied to encode the 3D geometric feature vector into 1×128.

Figure 5. Shading image examples that effectively represent facial features without influence of shadow.

4.2. 2D Landmark Sub-network

Since facial landmarks convey much information on expression [3], we incorporate 2D Landmark information for expression recognition.

We determine the face contour by [22] and locate the facial key points by [9]. Following [3], we then combine them to generate Sixty-eight landmark points recorded in total as follows:

$$LP = [x_1, y_1; x_2, y_2; \cdots, x_{68}, y_{68}] \,. \tag{3}$$

With the extracted points, we normalize their coordinates by max-min normalization with the position of the nose as the origin point. We then measure the distances between each pair of points as

$$D = [d_{1,2}, d_{1,3}, \cdots, d_{67,68}] \,. \tag{4}$$

In the formula, $d_{i,j}$ means the normalized distance between the i^{th} and the j^{th} landmark points.

After that, we encode the landmark features via four fully connected layers whose hidden dimensions are 1024, 512, 256 and 128. The final dimension of this path is 128.

4.3. Network Structure

Given an input image, we at first detect the human face with Haar cascades [13] to crop it and resize into 100×100. Then we encode the details with ResNet [6] structure. We

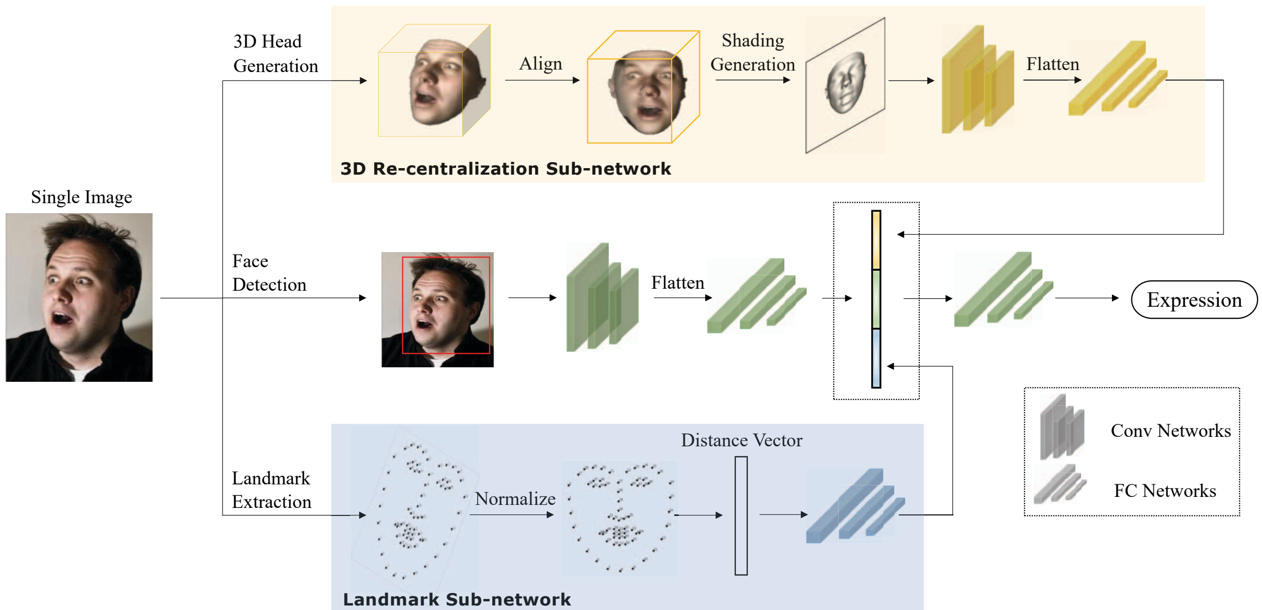


Figure 6. Overall architecture of the proposed model.

Dataset	Orientation	Lighting	SU	FE	DI	HA	SA	AN	NE	CO	Total
RAF-subset(train)	Y	Y	735	167	472	3157	1263	418	1481	-	7693
RAF-subset(test)	Y	Y	195	34	97	791	299	99	410	-	1925
OULU-CASIA	N	Y	240	240	240	240	240	240	-	-	1440
CK+	N	N	83	25	59	69	28	45	-	18	327

Table 1. Details of RAF-subset, OULU-CASIA and CK+ dataset. “SU”, “FE”, “DI”, “HA”, “SA”, “AN”, “NE” and “CO” stand for surprise, fear, disgust, happiness, sadness, anger, neutral and contempt expressions.

apply a separate convolution layer followed by one identity block and one convolution block to extract the global features of the detected 2D face. There are 7 convolution layers in total, and the number of filters for each layer are 128, 64, 64, 128, 64, 64, and 256. This is followed by a max-pooling layer with the pooling size 4. We further encode the 2D global feature by adding three fully connected layers whose hidden layers are 1024, 1024 and 128. The shape of the bottleneck output feature is also 128.

Finally, we fuse the above ResNet bottleneck feature with 3D geometry and landmark features by concatenating them together into a $[384, 1]$ vector. The concatenated features are passed into two fully connected layers. The hidden layers of these two layers are 512 and 128. Finally, we classify the expression with a softmax layer.

5. Experiments

5.1. Implementation Details

In our experiments, the batch size is 128. The optimizer is $Adam$ with $learning\_rate$ initialized to 0.01. The $dropout\_rate$ is 0.4 and we apply an early stopping method with training patience as 10. To attain a more robust model, we introduce Gaussian noise in the training process. The variance of Gaussian noise is set to 0.5. We use categorical cross entropy as loss function in our experiments if not specially declared. Besides, all the subnets have not been pre-trained before fusion.

5.2. Experiment Setting

Datasets We evaluate our experiements on RAF [12],OULU-CASIA [29] and CK+ dataset [15].

RAF dataset [12] contains many challenging cases, particularly including large face orientation and shadow. Notice that we did a clean up for RAF dataset that we remove invalid label (images containing more than 2 major faces) and tiny faces (the facial region is smaller than 50 pixels). For clarity, we name the clean dataset as RAF-subset.

OULU-CASIA dataset [29] contains image sequences in laboratory conditions, the lighting of which is standard but not perfect. For each image sequence, we pick the first, third and fifth images in reverse order in our experiments. The reason is that these images contain the peaks of the expressions and also differ from one another. CK+

Model	SU	FE	DI	HA	SA	AN	NE	Acc	C-Acc
Li-2017	60.51	44.12	42.27	91.02	67.22	63.64	87.56	78.85	65.19
Jung-2015	73.85	41.76	37.11	92.65	72.24	69.70	80.73	80.14	66.78
Zhang-2017	69.74	38.24	31.96	91.78	72.91	57.58	72.93	76.92	64.01
Fabian-2016	56.41	29.41	24.74	90.52	47.83	49.5	80.49	71.83	54.13
INC-2016	62.05	26.47	37.11	86.85	66.22	55.56	55.37	69.28	55.66
Zeng-2018	67.69	47.06	42.27	92.41	68.90	61.62	83.41	79.47	66.19
Our Model	75.38	41.18	38.14	94.31	71.91	70.71	83.17	**81.60**	**67.83**
Our Model(FL)	73.33	64.71	57.73	92.92	72.58	75.76	81.95	**82.29**	**74.14**

Table 2. Results on RAF-subset dataset: "SU", "FE", "DI", "HA", "SA", "AN", "NE" stand for the accuracy for "Surprise", "Fear", "Disgust", "Happiness", "Sadness", "Anger" and , "Neutral" expressions in the dataset. "Acc" is the accuracy for this dataset and "C-Acc" is the class accuracy. "Our Model(FL)" means our model with focal loss. All the values in the table are percentages (%).

Figure 7. Representative results from RAF dataset.

dataset [15] is relatively small, and it is used to test the robustness of our model in the small dataset. Following previous works [8, 25], we apply the 10-fold cross validation method on the last two datasets.

Table 1 lists the details of all the three datasets.

Methods to compare we compare with six most recent and well-performing methods: Li-2017 [12] is the baseline model for RAF dataset. Jung-2015 [8], Zhang-2017 [28] designs the special fusion network for landmark points and combine them with deep neural networks. Fabian-2016 [3] uses landmark points to calculate the geometric and shading features of the face. INC-2016 [17] introduces inception modules for FER. Zeng-2018 [26] proposes a multi-database learning framework for FER tasks. Zhang *et al.*'s method [27] requires multi-input, is not compared. Yang *et al.*'s method does not provide code and does not have sufficient detail for implementation, and therefore is not compared [24].

During our experiments, the landmark points used the compared approaches are extracted by the same method as ours. We retrain all the compared models based on the original articles or the released codes. A single image is treated as a sequence with only one frame for the sequence-based models (Jung-2015 [8] and Zhang-2017 [28]). Besides, we do not apply data augmentation or cross-database learning for all the models.

Evaluation metric We evaluate the performance by the classification accuracy on individual dataset that describes the overall performance of the compared models. We denote it as **Accuracy** and reported in percentages (%). We also report **Class Accuracy** on average accuracy for all the classes as RAF-subset and CK+ suffer class imbalance.

5.3. Results

5.3.1 RAF-subset dataset

RAF-subset contains many cases with large face orientation or strong shadow. The representative results are shown in Figure 7. As shown in Table 2, our model reaches the best performance in terms of both Accuracy and Class Accuracy.

We also provide the confusion matrix of our model in RAF dataset in Table 4. Interestingly, we find that "fair" is

Model	SU	FE	DI	HA	SA	AN	CO	Acc	C-Acc
Li-2017	100.00	56.00	93.22	97.10	57.14	82.22	55.56	86.24	77.32
Jung-2015	100.00	92.00	89.83	98.55	75.00	80.00	50.00	89.60	83.63
Zhang-2017	97.59	92.00	96.61	100.00	60.71	91.11	44.44	90.52	83.21
Fabian-2016	100.00	84.00	96.61	98.55	85.71	86.67	55.56	92.35	86.73
INC-2016	95.18	48.00	88.14	97.10	60.71	75.56	50.00	82.57	73.53
Zeng-2018	95.18	48.00	89.83	94.20	60.71	88.89	50.00	84.10	75.26
Our Model	100.00	84.00	94.92	100.00	85.71	97.78	83.33	**95.41**	**92.25**

Table 3. Results on CK+ dataset. "CO" stands for "Contempt", "Acc" is the accuracy and "C-Acc" is the class accuracy. All the values in the table are percentages (%).

	SU	FE	DI	HA	SA	AN	NE
SU	**75.4**	0.0	2.1	4.6	2.6	3.6	11.8
FE	23.5	**41.2**	0.0	0.0	14.7	11.8	8.8
DI	1.0	1.0	**37.1**	15.5	9.3	13.4	22.7
HA	0.5	0.0	0.6	**94.3**	1.0	0.3	3.3
SA	1.0	0.3	5.7	6.4	**71.9**	2.0	12.7
AN	5.1	0.0	10.1	7.1	1.0	**70.7**	6.1
NE	1.7	0.2	3.9	4.9	5.9	0.2	**83.2**

Table 4. Confusion matrix of our model on RAF-subset dataset. All the values are reported in percentage (%).

most likely to be misclassified as "surprise".

Focal loss Since the number of each category in the dataset is imbalanced (Table 1), there is a gab between Accuracy and Class Accuracy. To relieve the imbalanced class, we introduce focal loss [14]. The normal categorical cross entropy loss has the following form:

$$CE(p_t) = -log(p_t), \tag{5}$$

where p_t is a function of y and p. $y \in \{\pm 1\}$ indicates the ground-truth class, and $p \in [0,\ 1]$ is the model's estimated probability for certain class $y = 1$:

$$p_t = \begin{cases} p & if\ y = 1, \\ 1 - p & otherwise. \end{cases} \tag{6}$$

Then the focal loss for categorical cross entropy is:

$$FL(p_t) = -\alpha_t (1 - p_t)^{\gamma} log(p_t). \tag{7}$$

In the formula, α_t is the balanced coefficient, which has a similar definition as p_t:

$$\alpha_t = \begin{cases} \alpha & if\ y = 1, \\ 1 - \alpha & otherwise. \end{cases} \tag{8}$$

and γ controls the rate of weight conduction. In our experiment, α is set as 0.25 and γ is set as 2.

When trained with the above focal loss, the performance, particularly the class accuracy has been significantly improved as reported in Table 2.

Model	SU	FE	DI	HA	SA	AN	Acc
Li-2017	49.58	95.00	72.92	85.83	72.08	66.25	73.61
Jung-2015	45.83	76.67	61.25	97.50	62.92	56.67	66.81
Zhang-2017	87.08	93.33	61.67	92.08	68.33	90.83	82.22
Fabian-2016	68.75	92.92	65.00	91.67	72.50	79.17	78.33
INC-2016	50.00	80.83	61.25	52.50	62.08	70.00	62.78
Zeng-2018	66.25	61.67	82.92	55.83	62.92	52.08	61.02
Our Model	87.50	98.33	72.08	89.17	83.33	78.75	**84.86**

Table 5. Accuracy on OULU-CASIA dataset. All the values in the table are percentages (%).

5.3.2 OULU-CASIA Dataset

OULU-CASIA dataset is collected in laboratory conditions with face properly facing the camera, but the lighting condition is not satisfactory which results in facial shadows.

The experimental result on OULU-CASIA dataset is shown in Table 5. As OULU-CASIA dataset is a balanced dataset, C-Acc has a same value as Acc and we only report Acc in Table 5.

5.3.3 CK+ dataset

CK+ dataset is a small dataset in proper laboratory situations. We also report the performance in Table 3 and demonstrates the superior performance of the proposed model.

5.4. Ablation Study

To analyze the effect of the individual component in our proposed method, we conduct an ablation study. Since the features from 3D re-cenralization sub-net and landmark sub-net are fused into the main network, we could choose to connect the main network to specific sub-net or not. In this section, we name **3D** for the 3D re-cenralization sub-net and **Landmark** for landmark sub-net. When neither **3D** or **Landmark** is fused, we call the rest as **Base**. Thus, we also evaluate these four types of networks and reported their performance in Table 6.

From the result, we can see that both re-centralized 3D geometry features and landmark features could help to improve the FER performances. Combing two features together could further boost the performance significantly.

Model	RAF-s		OULU	CK+	
	Acc	C-Acc	Acc	Acc	C-Acc
Base	78.69	64.96	76.11	85.93	77.89
Base+Landmark	79.62	68.37	76.60	93.88	90.78
Base+3D	79.73	67.80	77.36	92.04	88.57
Whole Model	81.60	67.83	84.86	95.41	92.25

Table 6. Ablation study of individual component on three datasets. "Acc" stands for accuracy and "C-Acc" stands for class accuracy. All the values in the table are percentages (%).

The only exception is for RAF-subset dataset where the class accuracy falls when including 3D geometric features. We believe it is due to the imbalanced class distribution as discussed in RAF-subset dataset. The ablation analysis validates the effect of our proposed 3D re-centralization and landmark sub-networks.

5.5. Time Consumption

We test and compare the whole executive time (**including pre-processing time**) with 2 Titan X GPU among all the models. The per-frame process time is reported in Table 7. As can be seen, the proposed method only takes slightly more time than existing methods but achieves much more accurate result.

Model	Time(s)
Our Model	0.123
Li-2017	0.084
Jung-2015	0.086
Zhang-2019	0.090
Fabian-2016	0.089
INC-2016	0.097

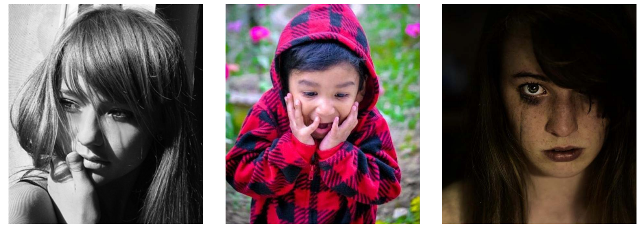
Table 7. Time consumption analysis.

5.6. Failure Cases

Figure 8 lists three representative failure cases. These images contain a huge part of occlusions in the front face, which will cause serious distortion of 3D reconstruction and further influence the shading generation. However, as shown in Figure 8, these images would also confuse alternative existing methods.

6. Conclusions and Future Work

This study proposes a novel system for single image Facial Expression Recognition (FER). To reduce the influence of orientations and shadows, we propose a novel approach to reconstruct and re-centralize a 3D facial model from a single image. Re-aligned 3D facial geometry and landmarks are then integrated into the proposed network for the robust FER. The experiments on three datasets demonstrate that the proposed model obtains state-of-the-art performance.



Ground Truth	SA	Ground Truth	SU	Ground Truth	SA
Our Method	NE	Our Method	FE	Our Method	NE
Li-2017	DI	Li-2017	NE	Li-2017	FE
Jung-2015	FE	Jung-2015	DI	Jung-2015	FE
Zhang-2017	NE	Zhang-2017	SA	Zhang-2017	NE
Fabian-2016	NE	Fabian-2016	NE	Fabian-2016	DI
Zeng-2018	DI	Zeng-2018	HA	Zeng-2018	HA

Figure 8. Failure cases from RAF dataset.

For future work, we consider optimizing our model in two directions. We will extend our model in image sequence-based emotion classification, which is more practical in the real world. We will also commit to further enhancing the robustness of 3D face alignment for extreme orientation cases.

References

[1] Ronen Basri and David W Jacobs. Lambertian reflectance and linear subspaces. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (2):218–233, 2003.

[2] Hui Chen, Jiangdong Li, Fengjun Zhang, Yang Li, and Hongan Wang. 3d model-based continuous emotion recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1836–1845, 2015.

[3] C Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5562–5570, 2016.

[4] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. *arXiv preprint arXiv:1803.07835*, 2018.

[5] John C Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[7] Xiuyi Jia, Xiang Zheng, Weiwei Li, Changqing Zhang, and Zechao Li. Facial emotion distribution learning by exploiting low-rank label correlations locally. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, June 2019.

[8] Heechul Jung, Sihaeng Lee, Junho Yim, Sunjeong Park, and Junmo Kim. Joint fine-tuning in deep neural networks for fa-

cial expression recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2983–2991, 2015.

[9] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1867–1874, 2014.

[10] Daijin Kim and Jaewon Sung. Facial expression recognition. In *Automated Face Analysis: Emerging Technologies and Research*, pages 255–317. IGI Global, 2009.

[11] Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. Emotion recognition in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

[12] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2584–2593, 2017.

[13] Rainer Lienhart, Alexander Kuranov, and Vadim Pisarevsky. Empirical analysis of detection cascades of boosted classifiers for rapid object detection. In *Joint Pattern Recognition Symposium*, pages 297–304. Springer, 2003.

[14] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *arXiv preprint arXiv:1708.02002*, 2017.

[15] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *IEEE conference on computer vision and pattern recognition Workshops (CVPRW)*, pages 94–101, 2010.

[16] Michael J Lyons, Shigeru Akamatsu, Miyuki Kamachi, Jiro Gyoba, and Julien Budynek. The japanese female facial expression (jaffe) database. In *Proceedings of International Conference on Automatic Face and Gesture Recognition*, pages 14–16, 1998.

[17] Ali Mollahosseini, David Chan, and Mohammad H Mahoor. Going deeper in facial expression recognition using deep neural networks. In *WACV*, pages 1–10, 2016.

[18] Richard Szeliski. *Computer vision: algorithms and applications*. Springer Science & Business Media, 2010.

[19] Raviteja Vemulapalli and Aseem Agarwala. A compact embedding for facial expression similarity. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, June 2019.

[20] Ziheng Wang, Shangfei Wang, and Qiang Ji. Capturing complex spatio-temporal relations among facial muscles for facial expression recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3422–3429, 2013.

[21] Martin Wöllmer, Angeliki Metallinou, Florian Eyben, Björn Schuller, and Shrikanth Narayanan. Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling. In *Proceedings of INTERSPEECH*, pages 2362–2365, 2010.

[22] Xuehan Xiong and Fernando De la Torre. Supervised descent method and its applications to face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 532–539, 2013.

[23] Jingwei Yan, Wenming Zheng, Zhen Cui, Chuangao Tang, Tong Zhang, and Yuan Zong. Multi-cue fusion for emotion recognition in the wild. *Neurocomputing*, 2018.

[24] Huiyuan Yang, Umur Ciftci, and Lijun Yin. Facial expression recognition by de-expression residue learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2168–2177, 2018.

[25] Zhiding Yu and Cha Zhang. Image based static facial expression recognition with multiple deep network learning. In *Proceedings of ACM International Conference on Multimodal Interaction*, pages 435–442, 2015.

[26] Jiabei Zeng, Shiguang Shan, and Xilin Chen. Facial expression recognition with inconsistently annotated datasets. In *Preceedings of European Conference on Computer Vision*, pages 222–37, 2018.

[27] Feifei Zhang, Tianzhu Zhang, Qirong Mao, and Changsheng Xu. Joint pose and expression modeling for facial expression recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, June 2018.

[28] Kaihao Zhang, Yongzhen Huang, Yong Du, and Liang Wang. Facial expression recognition based on deep evolutional spatial-temporal networks. *IEEE Transactions on Image Processing*, 26(9):4193–4203, 2017.

[29] Guoying Zhao, Xiaohua Huang, Matti Taini, Stan Z Li, and Matti PietikäInen. Facial expression recognition from near-infrared videos. *Image and Vision Computing*, 29(9):607–619, 2011.

[30] Xiangyun Zhao, Xiaodan Liang, Luoqi Liu, Teng Li, Yugang Han, Nuno Vasconcelos, and Shuicheng Yan. Peak-piloted deep network for facial expression recognition. In *Preceedings of European Conference on Computer Vision*, pages 425–442, 2016.

[31] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 146–155, 2016.

[32] Xinge Zhu, Liang Li, Weigang Zhang, Tianrong Rao, Min Xu, Qingming Huang, and Dong Xu. Dependency exploitation: a unified cnn-rnn approach for visual emotion recognition. In *Proceedings of the International Joint Conferences on Artificial Intelligence*, 2017.