Semantic Learning for Facial Expression Recognition

Yuanling Lv, Guangyu Huang, Yangfan Huang, Yixian Xie

Yuanling Lv: 23020211153954, Guangyu Huang: 23020211153933, Yangfan Huang: 31520211154007, Yixian Xie: 23020211153980

Abstract

In recent years, Facial Expression Recognition (FER) has achieved significant improvement on large-scale labeled data. However, a number of FER methods usually suffer from huge performance drop when dealing with limited labeled data, which is not easy to access in realistic scenarios. In this paper, we introduce a Semantic Learning for Facial Expression Recognition method (SLFER) by using an off-the-shelf module termed Spatial-Semantic Patch Learning (SSPL) of the Facial Attribute Recognition method. SLFER is capable of focusing on the semantic relationship between the different parts of the human face and making full use of this information to explore the intra-class and inter-class similarity of facial expression images. Specifically, the SLFER is composed of three tasks, including an expression classification task, a segmentation task and a classification task, which are jointly trained in a multi-task learning framework. The expression classification task extracts expression features, while the segmentation task and the classification task utilize a facial parsing model to obtain the semantic information of the facial features at a pixel level and image level respectively. Experimental results demonstrate that our method gain good performance on in-the-wild FER database, including RAF-DB and SEFW databases.

Introduction

FER plays an important role in our daily life helping human beings convey emotions and ideas (Darwin 2015), which has attracted extensive attention in the fields of human-computer interaction, security, robot manufacturing, automation, medical care, communication and driving (Zhang et al. 2018). According to psychological studies (Ekman and Friesen 1971), the FER classifies an input facial image into the following seven categories: angry, disgust, fear, happy, sad, surprise and neutral.

With the development of deep learning, FER has achieved satisfactory performance in many aspects and many efforts have been made to improve the accuracy of facial expression recognition. Some methods (Ruan et al. 2020; Mo et al. 2021) consider the disturbances of facial images, implicitly and explicitly disentangle these disturbances from original images respectively. Also, some methods (Ruan et al. 2020; Wang et al. 2020b) take fine-grained discriminative features into consideration by using attention mechanism or decomposing and reconstructing facial feature to model subtle differences between different facial expressions. While other methods (Wang et al. 2020a; Zeng, Shan, and Chen 2018) are designed to solve the noisy labels or inconsistent annotations problems.

Though the methods mentioned gain competing results, they only focus on information expression labels or other attribution labels of databases offered and do not model semantic-aware information, which may benefit the model. Meanwhile, multi-task learning can complement information among different related tasks and acquire a robust performance for recognition. Based on this, we take semantic information into consideration in order to explore the relationship of the regions of facial components and facial expressions at different levels. The method of SSPL (Shu et al. 2021) uses three auxiliary tasks consisting of a Patch Rotation Task(PRT), a Patch Segmentation Task(PST) and a Patch Classification Task(PCT) to jointly learn spatial-semantic relationship in order to deal with limited labeled data. Inspired by this , we use a part of the auxiliary tasks, including Segmentation Task and Classification Task as our basic tasks to build multi-task learning framework, where the semantic relationship of facial images is crucial to classify facial expressions. Since we need to extract the features from the facial details to determine expression categories, it is particularly important to learn fine-grained feature representations. For examples, since we know that the regions of eye, mouth contain rich information to distinguish different expressions, it is useful to locate the mouth region and determine whether the it is "Lip Corner Puller" at a semantic level and guide the model pay attention to the location. We use this proposed model to train on FER annotation datasets and achieve improved performance in the real data prediction.

To be specific , our proposed SLFER method uses the external auxiliary model called the facial parsing model (Yu et al. 2018), to obtain the corresponding proxy semantic labels, thereby supervising the network as it learns the semantic fine-grained feature representations. Specifically, it is a multi-task learning method where three tasks are jointly trained in an end-to-end manner. The ST and CT classify the parts of the human face from pixel level and image level

respectively, and then mine the semantic relationships between the regions of the face at different levels. The expression classification task outputs the final expression prediction classification results. Also, we use SE block (Hu, Shen, and Sun 2018) to build lightweight attention modules into our network, which enables the model to achieve better results on in-the-wild datasets. Our main contributions are summarized as follows:

- We propose the multi-task learning framework SLFER to explore the semantic relationship between the regions of facial components and facial expression. This method effectively uses semantic labels provided by the external model as guidance to obtain different levels of semantic information, and can accurately predict facial expressions under the premise of limited labeled datasets.
- We utilize two auxiliary tasks with two auxiliary tasks, where the goal of the two auxiliary tasks is to mine the intrinsic relationship between the semantic facial parts from different levels (pixel level and image level), allowing the network to extract the semantic-aware fine-grained feature representations more efficiently.

Related Work

Facial Expression Recognition

Deep learning-based methods are attracted more and more attention in FER, which can solve lots of problems and gain significant performances. (Xue, Wang, and Guo 2021) leverage dropout-based transformer and local discriminative patches for FER, which reduce the redundancy of patched and improve the FER performance. Besides, (**?**) explore powerful local patches and the interaction of layers to make network pay more attention to multiple diverse regions. (Wang et al. 2020a) propose the combination of self-attention and relabel to suppress the uncertainties caused by the subjectiveness of expression. (She et al. 2021) performs FER based on auxiliary information to mine the latent distribution in the label space and find the relationship of semantic feature between pairwise. (Ruan et al. 2020; Mo et al. 2021) both consider various disturbing factors and model common ones (such as pose illumination) and potential ones (such as hairstyle) respectively and disentangle these disturbances from facial images to gain facial expression-aware features and achieve excellent results. Besides, based on shard information across different expressions and unique information with expression, (Ruan et al. 2021) design an effective method model the information respectively and gain discriminative feature representations for facial expression.

Although the methods above can deal with many problems and achieve leading performance, they only consider classification task and do not other tasks to help model realize facial expressions deeply . In this work, we propose an effective method using pixel-level and image-level auxiliary tasks to perform improvement since the information multi-tasks learning offers can guide model to focus on the positions of facial components and further to learn rich information regions.

Multi-task Learning

Multi-task learning is inspired by human learning since when learning a new tasks, people often apply the knowledge they have learned on the related tasks. Therefore, Multi-task learning can help model focus on its attention on the features that actually matter because other tasks will provide additional information to reduce the effect of noisy data or disturbances. (Sun 2015) is proposed to use both face identification and verification to develop effective feature representations for reducing intra-personal variants while enlarging inter-personal differences. (Zhang et al. 2016) design a multi-task learning framework for learning fine-grained feature representations by jointly optimizing both classification and similarity constraints.

Since different types of related information can make model more robust, we use sematic-aware tasks as our auxiliary tasks to guide the expression classification , making model acquire enough knowledge to transfer and improve the accuracy of recognition task.

Method

Overview

SLFER consists of four main components, including a backbone network with attention mechanism, an Expression Classification Task (ECT), a Segmentation Task (ST) and a Classification Task (CT). An overview of the SLFER method is shown in Figure 1.

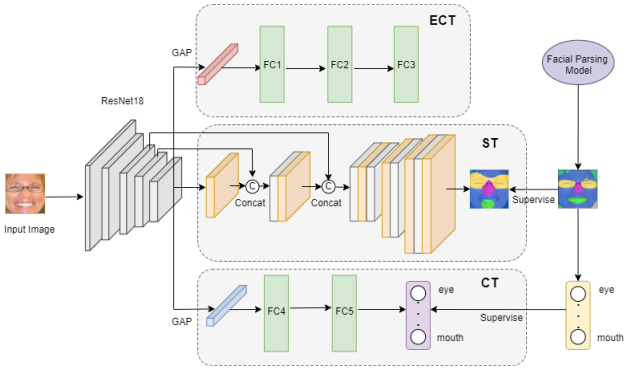


Figure 1: An overview of the SLFER method. SLFER involves three tasks, an Expression Classification Task (ECT), a Segmentation Task (ST) and a Classification Task (CT). ECT extracts expression-related features and predicts an expression label. ST and CT encode the facial semantic information from different levels by using a semantic label generated by an external model . The ST and CT parts are same as PST and PCT from SSPL (Shu et al. 2021)

The facial images are first fed into a ResNet18 backbone network to extract a basic CNN feature, where h, w, and c denote the height, width, and number of channels, respectively. The backbone network includes an attention mechanism SE block. Then, the basic feature is fed into three branches, ECT, ST and CT. For ECT, the expression features are used for expression prediction. Similar to SSPL, ST performs semantic segmentation on the feature maps from

different layers of the pretrained ResNet18 backbone and classifies the pixels of the output feature maps into semantic classes, where the semantic labels are produced by an external facial parsing model named BiSeNet (Yu et al. 2018) . According to the number of pixels each semantic label takes up, we can estimate whether a semantic class exists or not, thus gaining the facial component labels, which are used to predict the dominant facial component classes by CT. Finally, ECT, ST and CT are jointly trained in a multi-task learning network. For the testing stage, ST and CT are removed such that the facial expressions are only predicted by the ResNet18 backbone and ECT.

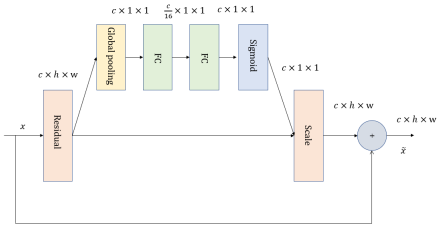


Figure 2: The network architecture of the attention mechanism (Hu, Shen, and Sun 2018)

Segmentation Task (ST)

In this part, we introduce segmentation task to predict semantic labels of pixels. The original facial image is fed to the backbone of ResNet-18 to extract features at different levels.

ST performs semantic segmentation on the facial image and assigns a semantic label to each pixel. Specifically, in order to make full use of multi-level information, the output feature map of the 4th residual block of ResNet-18 is fed into the upper sampling layer, and then concatenated with the output of 3rd residual block . In this way, the concatenated feature map contains multi-level information, and then is sent into a convolution layer, batch normalization, ReLU and an upper sampling layer. Finally, three pairs of upsampling layers and convolution layers are used to classify each pixel of the final feature map by batch normalization and ReLU.

Given J semantic classes and the class prediction probabilities for the k-th pixel as $h_s = [h_{k1}, \ldots, h_{kJ}]$, The loss of ST is defined as:

$$L_S = \frac{1}{K}\sum_{k=1}^{K}(\sum_{j=1}^{J}(-q_{kj}log(h_{kj}))) \quad (2)$$

where K is the total number of pixels in facial images

Classification Task(CT)

The architecture of CT uses two FC layers to predict the number of facial component classes. Specifically, we first send the a facial image to the pretrained ResNet-18 backbone to obtain a feature map $F_s \in R^{c'*w'*h'}$, which c,w, and h, represent the channel, width, and height of the feature map. Then input F_s into GAP layer to get a feature map f_s. Finally, the feature map f_s is input into two fully connected layers to predict the facial component label. Since not every facial component make the same contribution for facial expression recognition, we only choose the top n dominant facial components in each input image and label them as 1, while for the rest of other components we label them 0 as y_j. The loss of CT adopts the binary cross-entropy:

$$L_C = \sum_{j=1}^{J}(y_j log(x_j) + (1-y_j)log(1-x_j)) \quad (3)$$

where x_j is the output prediction probability of the j-th facial component.

Expression Classification Task(ECT)

For ECT, the expression-related features are obtained by employing a global average pooling (GAP) layer for the basic CNN feature and then fed into three FC layers to predict the expression categories. Given semantic classes, we denote the prediction probability as and train ECT by minimizing the cross-entropy loss, which can be formulated as:

$$L_{EC} = \sum_{k=1}^{K}(-p_k log(r_k)) \quad (4)$$

where p_k=1 when k is equal to the ground-truth label and p_k=0 otherwise.

Joint Loss Function

The joint loss function for the SLFER is defined as:

$$L_{SLFER} = L_{EC} + \lambda_1 L_S + \lambda_2 L_C \quad (5)$$

where λ_1 and λ_2 denote the regularization parameters. By optimizing the joint loss function, SLFER is capable of focusing on the facial semantic information and extracting discriminative expression-related features.

Experiments

In this section, we first describe two public datasets and both of them are in-the-wild datasets. We then conduct ablation studies to show the importance of the key components of our methods with qualitative and quantitative results. Finally, we compare our methods with state-of-the Art FER methods.

Datasets

RAF-DB The Real-world Affective Face Dataset(RAF-DB) (Li and Deng 2018) is an in-the-wild database, which contains 15339 images labeled with six basic facial expressions and one neutral expression, which are divided into 12271 and 3068 images for training and testing.

SFEW The Static Facial Expression in the Wild(SFEW) (Dhall et al. 2011) dataset is built by selecting the frames from AFEW (Dhall et al. 2012) database, containing unconstrained facial expressions. We use SFEW2.0 (Dhall et al. 2015) to conduct our experiments which contains 958 images for training, 436 images for validation. Each image is annotated with one of the seven expressions.

Table 1: Ablation studies for three key modules of our SLFER method on the RAF-DB and SFEW databases.

	Accuracy(%)	
	RAF-DB	SFEW
Baseline	86.93	52.98
CT	87.77	55.73
ST	87.87	54.35
CT+ST	88.01	59.63
Proposed	**88.33**	**59.86**

Implementation Details

In our experiments, all the facial images are randomly cropped to the size of 224*224. We also center crop the input images to the size of 224*224 at test process. Similar to (Wang et al. 2020b,a), our method is implemented with ResNet18 (He et al. 2016) as backbone which is pretrained on the MS-Celeb-1M face database (Guo et al. 2016). In the training task, the number of key facial components set to 4 in CT and we aggregate the semantic labels using externally-trained facial parsing model (BiSeNet) from 19 categories to 8 categories (background, skin, eye, ear, nose, mouth, neck and hair). At the test process, we simply use ResNet18 (trained with two auxiliary tasks before) and three Fully-Conneted (FC) layers for classification, which do not use the auxiliary tasks.

All experiments are implemented by Pytorch and run on 1080Ti GPU for 40 epochs , and batch size for both datasets is set to 16. Adam optimizer (Kingma and Ba 2014) with initial learning rate of 0.0001 and β_1 = 0.5, β_2 = 0.999 is applied to our method. The learning rate decays by 0.1 after 10,18,25 and 32 epochs.

Ablation studies

We conduct ablation studies to evaluate the influence of different auxiliary tasks (i.e., ST, and CT) and crucial parameters (including balance parameters of ST loss $\lambda_1$and CT loss λ_2, and the number of key facial components) of the proposed method on the final performance.

Influence of different auxiliary tasks. We evaluate five variants of the proposed method, including: (1) the baseline method that is based on the ResNet18 pretrained on the MS-Celeb-1M face database. (2) the method (denoted as "CT") that only adopts CT as the auxiliary task; (3) the method (denoted as "ST") that uses ST as the auxiliary task; (4) the method (denoted as "ST+CT") that uses ST and CT as the auxiliary tasks; (5) the method (denoted as "Proposed") that uses ST and CT as the auxiliary tasks with attention mechanism.

From Table 1, our proposed SLFE method gain higher accuracy than baseline method on the RAF-DB and SFEW. Compared with CT+ST, our method also improves the performance. Our method achieved the best performance of all variants when we use CT and ST as auxiliary tasks with attention mechanism, which shows the importance of modeling the semantic relationship can be beneficial for the FER task.

Table 2: Performance comparisons among different methods on several FER databases. The best results are boldfaced.

Method	Accuracy(%)	
	RAF-DB	SFEW
DLP-CNN (Li and Deng 2018)	84.13	51.05
IPA2LT (Zeng et al. 2018)	86.77	58.29
RAN (Wang et al. 2020b)	86.90	56.40
SCN (Wang et al. 2020a)	87.03	-
DDL (Ruan et al. 2020)	87.71	59.86
D^3Net (Mo et al. 2021)	88.79	**62.16**
FDRL (Ruan et al. 2021)	89.47	**62.16**
TransFER (Xue et al. 2021)	**90.91**	-
Baseline	86.93	52.98
Proposed	88.33	59.86

Influence of balance parameters of ST loss $\lambda_1$and CT lossλ_2. We evaluate the recognition performance of the proposed method with the different values of balance parameters of ST loss λ_1 and CT loss λ_2. To be specific, we first fix λ_2= 0.5 and set λ_1 0, 0.001, 0.1, 0.5, 1, respectively. Experimental results are given in Table 3 (a). We can know for the table that our method achieves the best performance when the value ofλ_1 is set to 0.5. Similar to λ_2, when test λ_2, we fix λ_1= 0.5 and λ_2 is set from 0 to 1. We can observe from the Table3(b) that when the value of λ_2 is 0.5, our method can gain better results. In the following, we set the values of both λ_1 and λ_2 to 0.5 in our proposed method.

Influence of the number of dominant facial components n. We evaluate the influence of the number of dominant facial components in CT on the final performance. The experimental results are given in Figure 3 . Our proposed method obtains the best results when the value of n is set to 4. When the value of n is set too large, there are only a few pixels can be chosen as dominant facial components. Also, when n is too small, some important facial components are ignored. Both cases may introduce error and lead to performance degradation.

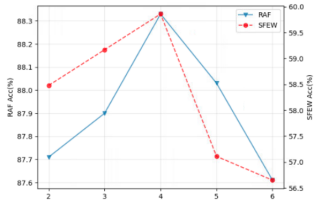


Figure 3: Ablation studies for the number of dominant facial components on RAF-DB and SFEW databases.

Visualization

2D feature visualization. We use t-SNE to visualize the expression features extracted by the baseline method and the proposed FDRL method on the 2D space, respectively, as shown in Figure 4. We can observe that baseline does not

(a) Baseline

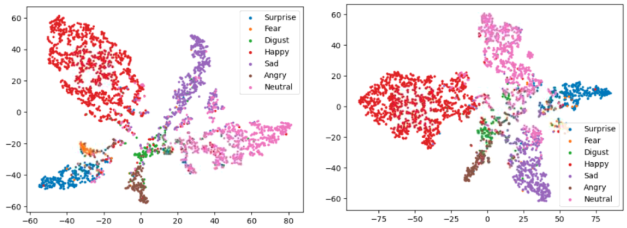
λ_1	Accuracy(%) RAF-DB	SFEW
0	87.54	57.56
0.001	87.90	57.79
0.1	88.07	58.52
0.5	**88.33**	**59.86**
1	88.13	57.79

(b) SLFER

λ_2	Accuracy(%) RAF-DB	SFEW
0	87.74	55.96
0.001	87.84	56.42
0.1	87.77	57.79
0.5	**88.33**	**59.86**
1	87.64	56.19

Table 3: Ablation studies for different values of $\lambda_1$and λ_2 representing the balance parameters for ST loss and CT loss respectively.

do a good job of making homogeneous expressions get close to each other in order to gain a compact expression representation. In contrast, the features extracted from our proposed method can effectively reduce intra-class differences and enhance inter-class separability for different expressions. Especially, compared with baseline, Happy and Sad can learn more compact representations for SLFER.



(a) baseline (b) SLFER

Figure 4: Visualization of the expression features using t-SNE Features are extracted from the RAF-DB database.

Figure 5: Semantic masks generated by externally-trained facial parsing model on RAF-CB database.

Comparison with State-of-the-Art Methods

In this section, we compare the proposed method with seven state-of-art methods on two public databases. Table 1 show the experiment results on in-the-wild databases(RAF-DB, SFEW).

In Table 2, among all the competing FER methods , IPA2LT and SCN is used to deal with the noisy and inconsistent labels ,and DDL and D^3Net both consider various disturbing factors modeling common ones and potential

Table 4: Performance comparisons among different methods on several FER databases. The best results are boldfaced.

Method	Accuracy(%) RAF-DB	SFEW
DLP-CNN (Li and Deng 2018)	84.13	51.05
IPA2LT (Zeng et al. 2018)	86.77	58.29
RAN (Wang et al. 2020b)	86.90	56.40
SCN (Wang et al. 2020a)	87.03	-
DDL (Ruan et al. 2020)	87.71	59.86
D^3Net (Mo et al. 2021)	88.79	**62.16**
FDRL (Ruan et al. 2021)	89.47	**62.16**
TransFER (Xue et al. 2021)	**90.91**	-
Baseline	86.93	52.98
Proposed	88.33	59.86

ones. DLP-CNN (Li and Deng 2018) is proposed to use a LP-Loss to alleviate intra-class variations. Recently TransFER improve the model performance by using an effective combination of Vision Transformer(ViT) (Farzaneh and Qi 2021) and Dropout to learn rich relations-aware local representations. As we can see , though the proposed method dose not outperform some competing FER methods, such as D3Net, FDRL and TransFER, our proposed method also gains good performance. The methods above do not consider that by using some auxiliary tasks can help model realize the facial expression deeply. In contract, SLFER is developed to focus on auxiliary tasks to gain better results by using the semantic information.

Conclusion

In this paper, we have presented a SLFER method, mainly consisting of three tasks (i.e, ECT, ST and CT) for effective FER, especially ST and CT as our auxiliary tasks. Driven by the attention mechanism that is employed in the backbone network, ECT is able to extract more discriminative expression-related features. ST and CT encode the facial semantic information from pixel level and image level respectively, which enables the target ECT task to pay attention to the critical facial parts related to the corresponding expressions and thus makes the extracted features robust to various disturbanc. Extensive experiments on the RAF-DB and the SFEW databases have demonstrated the effectiveness of SLFER.

References

Darwin, C. 2015. *The expression of the emotions in man and animals*. University of Chicago press.

Dhall, A.; Goecke, R.; Lucey, S.; and Gedeon, T. 2011. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, 2106–2112. IEEE.

Dhall, A.; Goecke, R.; Lucey, S.; and Gedeon, T. 2012. Collecting large, richly annotated facial-expression databases from movies. *IEEE multimedia*, 19(03): 34–41.

Dhall, A.; Ramana Murthy, O.; Goecke, R.; Joshi, J.; and Gedeon, T. 2015. Video and image based emotion recognition challenges in the wild: Emotiw 2015. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, 423–426.

Ekman, P.; and Friesen, W. V. 1971. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2): 124.

Farzaneh, A. H.; and Qi, X. 2021. Facial expression recognition in the wild via deep attentive center loss. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2402–2411.

Guo, Y.; Zhang, L.; Hu, Y.; He, X.; and Gao, J. 2016. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European conference on computer vision*, 87–102. Springer.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Li, S.; and Deng, W. 2018. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Transactions on Image Processing*, 28(1): 356–370.

Mo, R.; Yan, Y.; Xue, J.-H.; Chen, S.; and Wang, H. 2021. D^3Net: Dual-Branch Disturbance Disentangling Network for Facial Expression Recognition. In *Proceedings of the 29th ACM International Conference on Multimedia*, 779–787.

Ruan, D.; Yan, Y.; Chen, S.; Xue, J.-H.; and Wang, H. 2020. Deep disturbance-disentangled learning for facial expression recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*, 2833–2841.

Ruan, D.; Yan, Y.; Lai, S.; Chai, Z.; Shen, C.; and Wang, H. 2021. Feature Decomposition and Reconstruction Learning for Effective Facial Expression Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7660–7669.

She, J.; Hu, Y.; Shi, H.; Wang, J.; Shen, Q.; and Mei, T. 2021. Dive into Ambiguity: Latent Distribution Mining and Pairwise Uncertainty Estimation for Facial Expression Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6248–6257.

Shu, Y.; Yan, Y.; Chen, S.; Xue, J.-H.; Shen, C.; and Wang, H. 2021. Learning Spatial-Semantic Relationship for Facial Attribute Recognition With Limited Labeled Data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11916–11925.

Sun, Y. 2015. *Deep learning face representation by joint identification-verification*. The Chinese University of Hong Kong (Hong Kong).

Wang, K.; Peng, X.; Yang, J.; Lu, S.; and Qiao, Y. 2020a. Suppressing uncertainties for large-scale facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6897–6906.

Wang, K.; Peng, X.; Yang, J.; Meng, D.; and Qiao, Y. 2020b. Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Transactions on Image Processing*, 29: 4057–4069.

Xue, F.; Wang, Q.; and Guo, G. 2021. TransFER: Learning Relation-aware Facial Expression Representations with Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3601–3610.

Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; and Sang, N. 2018. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, 325–341.

Zeng, J.; Shan, S.; and Chen, X. 2018. Facial expression recognition with inconsistently annotated datasets. In *Proceedings of the European conference on computer vision (ECCV)*, 222–237.

Zhang, F.; Zhang, T.; Mao, Q.; and Xu, C. 2018. Joint pose and expression modeling for facial expression recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3359–3368.

Zhang, X.; Zhou, F.; Lin, Y.; and Zhang, S. 2016. Embedding label structures for fine-grained feature representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1114–1123.