JOURNAL OF MECHANICS OF CONTINUA AND MATHEMATICAL SCIENCES
www.journalimcms.org

ISSN (Online) : 2454 -7190 Vol.-15, No.-7, July (2020) pp 352-361 ISSN (Print) 0973-8975

# TRANSFER :- DEEP INDUCTIVE NETWORK FOR FACIAL EMOTION RECOGNITION

**Arpita Gupta[1], Nandhini Swaminathan[2], Ramadoss Balakrishnan[3]**

[1,2,3]National Institute of Technology, Tiruchirappalli, India.

[1]arpitagupta2993@gmail.com, [2]nandhiniiswaminathan@gmail.com, [3]brama@nitt.edu

https://doi.org/10.26782/jmcms.2020.07.00029

## Abstract

*The image-based Facial Emotion Recognition (FER) aims to classify the image into basic emotions being communicated by it. FER is one of the most prominent research areas in computer vision. Most of the existing works are aimed at high-quality images which are collected in the lab environment. These images are very different from the real-life facial emotion that leads to a lack of wild labeled data. Deep learning using transfer learning has shown promising results in computer vision in solving the problem of lack of labeled data. In the recent system, there is a great focus to overcome the lack of data issue in FER. Our paper has utilized the deep residual networks with inductive learning and self-attention module to overcome this problem. We have experimented different pre training settings and datasets for the model, which are Image Net and VGG face dataset (source datasets). The self-attention block is applied for better visual perspective to the model. Our target dataset is FER-2013, a benchmark dataset in FER. Trans FER is a deep residual network based on inductive learning and attention module. Our proposed approach has achieved superior performance than the existing state of the art models in the FER application using transfer learning.*

**Keywords :** Facial Emotion Recognition, Deep Learning, Deep Residual Networks, Transfer Learning, Inductive Learning, Self-Attention.

## I. Introduction

Facial Emotion Recognition (FER) is one of the most relevant fields of research in computer vision. Facial expressions are one of the adequate ways of understanding the emotions a human is trying to convey. Deep learning is providing to be very useful in the field of automated facial expression recognition (AFER). As we can use the deep learning models in the applications of medicine, security, human-computer interactions, robotics and many more. In the early twentieth century, Ekman [III] has recognized the basic six facial emotional expressions (anger, disgust, fear, happiness, sadness, and surprise) among the humans are universal, while neutral is also one of the most important emotion.

Early research in FER considers different types of features from the facial representation. Also, the real-life scenario is way more complicated than the lab

posed expressions; they are very consistent. There are specific datasets collected in the unconstrained environment [VII] [X]. There are specific issues in FER to be addressed, like pose, occlusion, illumination and much more unrelated to facial expressions [XVI]. The advancement of deep learning has helped to overcome the diversity of dataset issue, as deep learning has proved to be very effective in AFER. However, one of the issues in deep learning is the need for training on the vast amount of data, leading to the problem of the need for the large labeled dataset. In the field of FER, there are not many available large labeled datasets [XI]. To overcome this issue, researchers, use the pre trained model on other datasets which could be of the same or different domain. Using the pre training process, the networks could be pre trained supervised once on a large dataset and then trained on the target dataset to get the desired outcome. The source dataset is generally extensive and of better quality than the target dataset. In this process, fine-tuning is done, pre ceding to better convergence on the smaller target datasets.

Still, there are some problems, which need to be addressed before in using fine-tuning for better performance. One of them is the need for annotated data, a difficult and time-consuming process if the source dataset is of different domain. So many researchers prefer using the same domain dataset, as it is helpful to use pre annotated and easy to use for training. Another problem is related to images as the views of cameras, background and quality matter a lot in training [VII]. Another problem is to use the models trained on lab prepared data on the real-time data as many factors change in both.

Focusing on the above-stated problem of lack of labeled dataset, we propose to study the inductive learning facial expression recognition using deep learning models. We have utilized knowledge acquired from different source labeled dataset in the task of FER on target FER2013 dataset [V]. We have employed inductive learning, fine-tuning and the self-attention module in the deep residual network. Experiments on variants of deep residual networks (Res Net) on the source datasets Image Net [I] and VGG datasets [XIII] have been conducted. Due to recent availability of computational power, it is easier to go deeper with the networks which have proved to be useful as deep models outperform the traditional shallow models.

### I.a. Motivation

This work is to overcome the issue of lack of labeled data in the field of FER, which will help in real-time applications of facial emotion detection in different fields like robotics, medicine, security and many more. This study of deep models using transfer learning and attention in FER has yielded comparable and even better results.

### I.b. Organization

The following paper has been organized as follows: Section 2 explains the related works, which are existing studies in the field of transfer learning for the application of FER. Section 3 describes the proposed work TransFER based on deep learning and transfer learning. Section 4 explains the datasets used and the training of the models. Section 5 explains the results achieved, followed by section 5 conclusion.

## II. Related Work

Facial Action Coding System was proposed by a psychologist who explains the micro-expression in humans leading to the categorization of facial expression into six; they are - anger, disgust, fear, happiness, sadness and surprise. The current research is trying to make facial expression recognition automated. Most of these AFER are considering six basic emotions and neutral. Deep learning has proved to itself in the field of computer vision. Some existing studies have used CNN [II] [XVIII], Alex Net [XI] [XII], Inception [X], and their modification in this field. There are existing models based on machine learning technique is like SVM [XVI], PCA [XVI], Bayesian Network, Hidden Markov Model [XIV] and many more.

In FER, there are three feature types they are:- (a)geometric feature, the geometric features are extracted by using the distance between different parts of the face is calculated to find the facial expressions (b) appearance features, using Gabor features, local binary features and others. (c) Hybrid features; in this both the geometric and appearance-based features are used for facial expression recognition. The process of facial expression recognition consists of three steps: - face detection, facial feature extraction and facial expression recognition.

Deep learning has been vastly applied in computer vision, and Res Net has proved to be very successful in FER application because of its capability of extracting features from images. There are no studies employing residual networks with transfer learning and attention module for FER application. Alex Net is a variant of CNN has been used of the FER, they have used two-level fine-tuning using Image Net and Emoti W dataset, but this achieved an accuracy of 54.0% that was a state-of-the-art accuracy till now as our model has achieved superior accuracy [XI]. AU-Aware is an existing study in which the representation of the face is done using convolution and max-pooling layer that uses dense sampling facial patches and greedy filters [VIII].

### II.a. Inductive Learning

Many existing works are based on transfer learning in the field of computer vision, whereas there are some existing approaches which have used transfer learning concept in FER. In one exiting work, they have made use of cross dataset approach for FER applying transfer subspace learning method [XXVII]. Another study uses Kernel Mean Matching for distribution matching of source and target [IX]. In one of the existing studies, they have used SVM with a different kind of features transfer features, Gabor features and distance features for the FER [XVI]. There is only one study employing multi-task learning in facial landmark detection using CNN, a broader field in facial analysis.

## III. Proposed Work

In this paper, we have compared the performance of two networks on two different settings and adding self-attention block to it. The attention block helps in better visual perceptibility to the model [XV].

$$\text{Attention } (Q, I) = \text{SoftMax } (\text{Flatten}(\tanh(Q)) * I \quad (1)$$

The above equation 1 shows the equation for attention where I represent the features which are extracted in prior layers, and Q is the fully connected layer. Our model TransFER makes use of attention module. The following diagram depicts the architecture of the self-attention module.

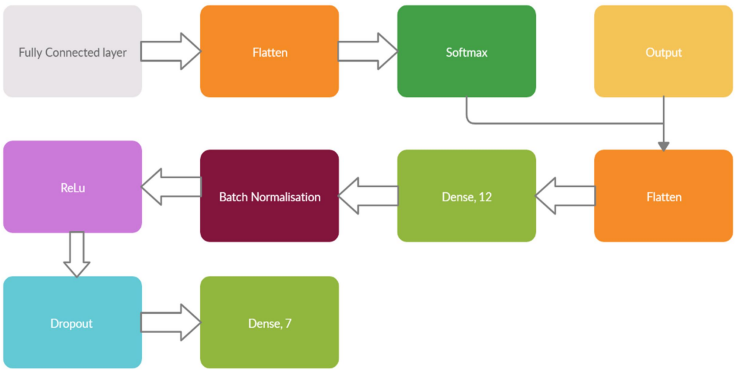


**Fig. 1:** The architecture of Attention block

### III.a. Res Net-50

Res Net was proposed in 2015 and won first prize in the image classification task in ILSVRC (Image Large Scale Visual Recognition Challenge) 2015 on Image Net [VI]. Deep residual network, also known as Res Net, consists of residual connections helping knowledge transfer in between layers. These residual connections are like skip connections, which propagates gradient throughout the layers [IV]. The back propagation feature in Res Net helps in removing the gradient vanishing problem.

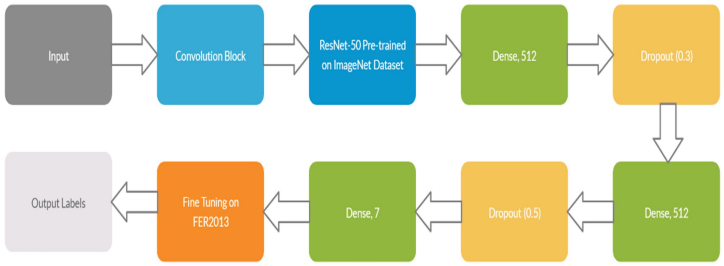


**Fig. 2:** ResNet-50 Pre trained on Image Net dataset

Figure 2 depicts the network architecture of ResNet-50 used in pre training on Image Net and FER2013 as target dataset. We have pre-trained the model on Image Net then using supervised training on FER2013 consisting of seven labels of emotions. We have put dropout layers to overcome the problem of over fitting, very common in transferring knowledge from multiple datasets. We have used Re Lu activation function and categorical cross-entropy loss. The model has a learning rate of 0.1 and is compiled with the SGD optimizer, and categorical cross-entropy loss is

used. The layers 'res5c_branch2b', 'res5c_branch2c' and 'activation_97' of the Res Net 50 model are unfrozen to enable fine tuning on the FER2013 dataset. The fine-tuning is done for 50 epochs. Batch Normalization (from Inception-v2) is after each convolution thereby increasing the speed at which network trains. Tanh activation minimizes the cost of function in the attention module function so as to minimize the cost function faster.

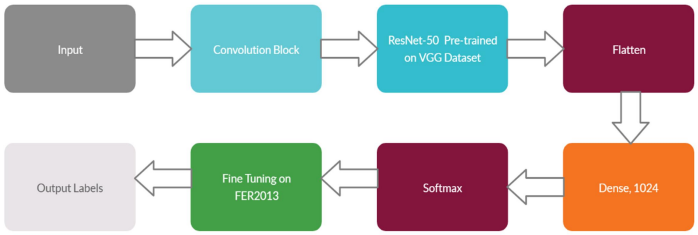


**Fig. 3:** ResNet-50 Pre-trained on VGG dataset

Figure 3 represents the architecture of the Res Net- 50, deep with 50 layers, and it has been pre trained by the VGG face dataset. This network has performed better than the model pre-trained on Image Net dataset as there was more specific task similarity in both the datasets. The model has a learning rate of 0.1, changing dynamically depending on the current epoch index and learning rate. It also possesses an early stop, terminating the training if there is no improvement for a period of 10 epochs and a module to ensure a plateau in accuracy does not occur for a period of 3 epochs by changing the learning rate. The fine-tuning for the model is done for 50 epochs. Batch Normalization (from Inception-v2) is used after each convolution, thereby increasing the speed of network training. Tanh activation is used in the attention module function to minimize the cost function faster.

### III.b. Res Net-152

In figure 3, the model is deeper as we have tested the Res Net with 152 layers. This model is pre trained on the larger dataset Image Net, consisting images of vast categories

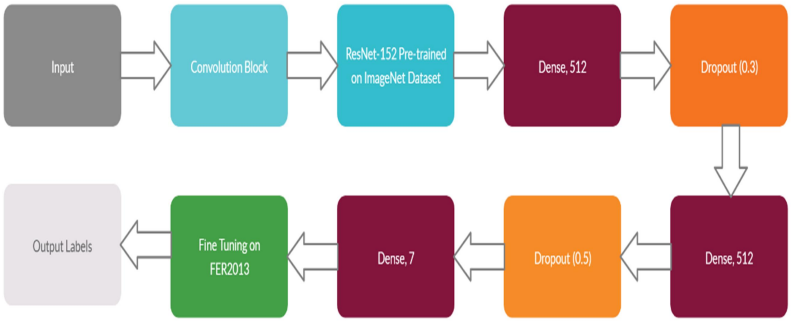


**Fig. 3:** ResNet-152 Pre trained on Image Net dataset

The Res Net 152 model provides a better accuracy due to the increase in a number of deep network layers, trained for thirty epochs, it has a learning rate of 0.1 and is compiled with the SGD optimizer, and categorical cross-entropy loss is used. The layers 'res5c_branch2b', 'res5c_branch2c' and 'activation_97' of the network model are unfrozen so as to enable fine-tuning on the FER2013 dataset. Batch Normalization (from Inception-v2) and tanh activation function in the attention module are used to increase the speed at which network trains.

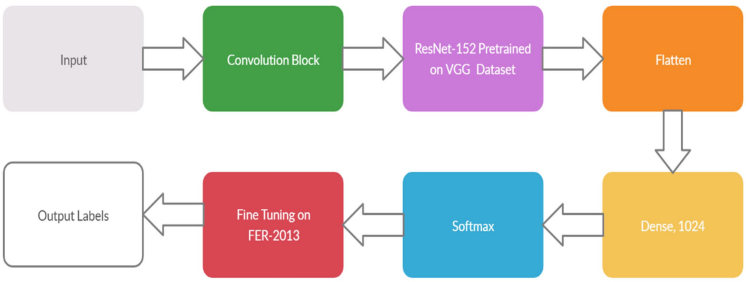


Fig. 4: ResNet-152 Pre trained on VGG dataset

The Res Net- 152, pre trained on the VGG face dataset has a dynamic learning rate module, which depends on the current epoch index and learning rate enables the model to be more efficient. It also possesses an early stop module which stops the training if there's no improvement for a period of 10 epochs. Moreover, changes are made to the learning rate to inhibit the plateauing of accuracy. The fine-tuning for the model is done for 30 epochs. Adam optimizer is used during the compilation of the model as it's faster and reliable while reaching the global minimum.

## IV. Dataset and Training

### IV.a. Pre Training Dataset

We have used two datasets for the TransFER models one of the pre training dataset is Image Net consists of 14 million images labeled into 20,000 categories. Image Net contains labeled dataset leading to better pre training then in an unsupervised manner. Image Net is one of the most significant datasets in pre training. Networks pre training on Image Net has solved the problem of the non-existence of high-quality images with annotations for training the model. The other pre training dataset is VGG face dataset consisting of 2,622 identities, with great result in the network's FER application. Every image in VGG face dataset contains a text file that contains where the image is taken from.

### IV.b. FER-2013 Dataset

FER-2013 dataset is the database that was introduced in Challenges in Representation Learning- ICML- 2013. FER-2013 database images were labeled in seven labels, six basic emotions and neutral. The database consists of 35,887 images; most of them are in wild settings. FER-2016 dataset consists of three categories: -

original training data, public test data and final test data. Original dataset consists of 28709 images; public test data consists of 3589 images, which is for validation and final testing of the model.

We have fine-tuned for preparing the model to perform on the second similar task as the one on which the model was pre trained. This pre training allows the advantage of better feature extraction without the need to train the model more the scratch. In fine-tuning, the output layer of our models is designed to recognize the seven classes then the 1000 classes in Image Net. And then the output layer is trained for lower features value than the existing one. In our study, the pre training on VGG dataset is done comprised of faces images outperforming all the existing models. We have utilized the early stop if there is no improvement in the model. The results are on only 30 epochs in our model training, as after 30 epochs did not show any improvement in the model.

## V. Results and Discussion

The results obtained from our studies of TransFER involving ResNet-50 and ResNet-152 pre trained and using attention module are summarized and compared with the existing models are in table 1. We have used inductive learning in which the model is trained on multiple datasets of a similar task. In our studies, it is proved that the models trained on VGG face dataset have outperformed the existing models. The Res Net with 50 layers using pre training on Image Net and attention module has achieved an accuracy of 33.46% when the target dataset is FER-2013. Moreover, the Res Net with 152 layers pre trained on Image Net with attention module has achieved an accuracy of 35.7%.

**Table 1: Performance Evaluation**

| Model | Source Dataset | Target Dataset | Accuracy |
|---|---|---|---|
| AlexNet [XI] | Image Net | FER-2013 | 54.0 |
| VGG-CNN [XI] | | | 47.0 |
| CNN[VII] | | | 51.8 |
| Mollahosseini[X] | 6 Datasets | | 34.0 |
| CNN+MMD [VII] | RAF | | 52.3 |
| DETN [VII] | | | 52.37 |
| TransFER (Proposed) | | | |
| ResNet-50 + Self-attention | Image Net | | 33.46 |
| ResNet-152 + Self-attention | | | 35.7 |
| ResNet-50 + Self-attention | VGG | | 56.21 |
| ResNet-152 + Self-attention | | | 57.8 |

In our study, usage of another pre training dataset, VGG face dataset, and it has achieved better accuracy with attention module of 56.21%, which is higher than all the existing models trained using transfer learning while the deeper model has outperformed all the networks by achieving superior accuracy of 57.8% with attention module when pre trained on VGG face dataset.

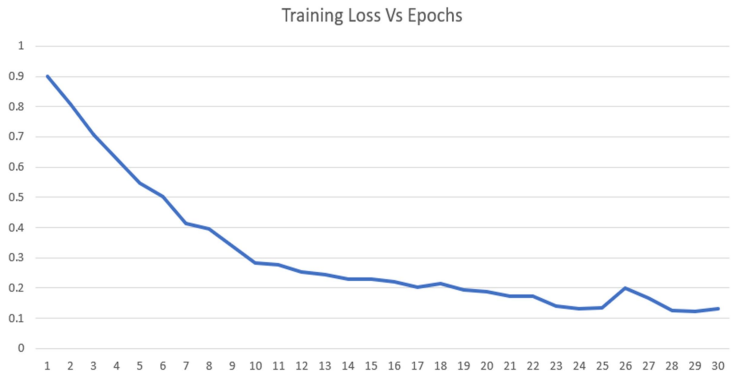


**Fig. 6:** Graph of training loss versus epochs

The graph in figure 6 shows the training loss variation in 30 epochs. The training loss of the classification labels is on the training set of the FER-2013 dataset for the 30 epochs as after that; there is no improvement in the model.

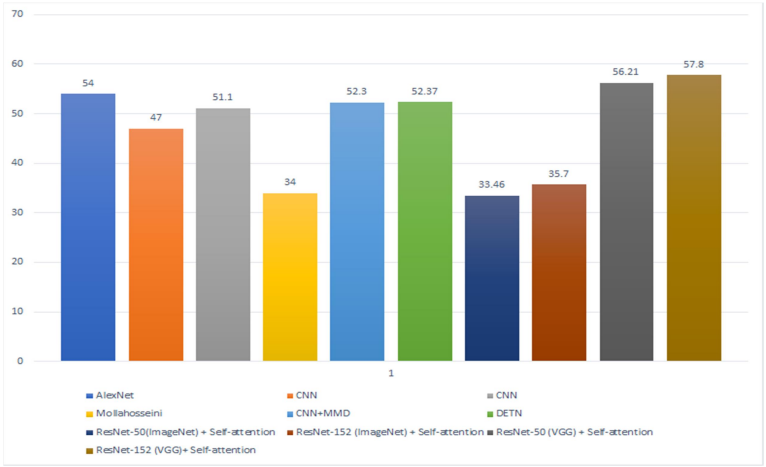


**Fig. 7:** Performance evaluation of Facial Emotion Recognition

Figure 7 summarizes the accuracy achieved by all the existing models based on transfer learning for FER. The Alex Net [XI] and VGG-CNN [XI] network is the one which is pre trained on Image Net. CNN, Mollahosseini [X], CNN-MMD [VII] and DETN [VII] are pre trained on RAF dataset. The above graph shows that our models based on Res Net using pre training and attention block is superior to all the existing models on FER dataset when pre trained on VGG face datasets.

## VI. Conclusion

In this work, the proposed model TransFER has attained performance enhancement to the existing state of art models, while solving the problem of unavailability of the large labeled dataset. Our work presents a model based on the

deep residual network, pre training and self-attention module. Res Net has been pre-trained on Image Net, and VGG faces dataset. We have experimented on two variations of Res Net of 50 and 152 layers. The clear benefit of our model is the regularization stimulated as the model performs well on the related task, also prevents over fitting, whereas another advantage of the model is increased classification accuracy and inductive training. Our model has achieved better accuracy of 56.21% with 50 layers and 57.8% with 152 layers. All experiments show that Res Net pre trained on VGG face dataset with attention model has outperformed existing methods with a clear margin of 5.43%.

## References


I Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. "Imagenet: A large-scale hierarchical image database." In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248-255. Ieee, 2009.

II Devries, Terrance, Kumar Biswaranjan, and Graham W. Taylor. "Multi-task learning of facial landmarks and expression." In *2014 Canadian Conference on Computer and Robot Vision*, pp. 98-103. IEEE, 2014.

III Ekman, Paul, and Wallace V. Friesen. "Constants across cultures in the face and emotion." *Journal of personality and social psychology* 17, no. 2 (1971): 124.

IV Geng, Mengyue, Yaowei Wang, Tao Xiang, and Yonghong Tian. "Deep transfer learning for person re-identification." *ar Xiv preprint arXiv: 1611. 05244* (2016).

V Good fellow, Ian J., Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski et al. "Challenges in representation learning: A report on three machine learning contests." In *International Conference on Neural Information Processing*, pp. 117-124. Springer, Berlin, Heidelberg, 2013.

VI He, Kaiming, 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*, pp. 770-778. 2016.

VII Li, Shan, and Weihong Deng. "Deep emotion transfer network for cross-database facial expression recognition." In *2018 24th International Conference on Pattern Recognition (ICPR)*, pp. 3092-3099. IEEE, 2018.

VIII Liu, Mengyi, Shaoxin Li, Shiguang Shan, and Xilin Chen. "Au-aware deep networks for facial expression recognition." In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pp. 1-6. IEEE, 2013.

IX Miao, Yun-Qian, Rodrigo Araujo, and Mohamed S. Kamel. "Cross-domain facial expression recognition using supervised kernel mean matching." In *2012 11th International Conference on Machine Learning and Applications*, vol. 2, pp. 326-332. IEEE, 2012.

X Mollahosseini, Ali, David Chan, and Mohammad H. Mahoor. "Going deeper into facial expression recognition using deep neural networks." In *2016 IEEE Winter conference on applications of computer vision (WACV)*, pp. 1-10. IEEE, 2016.

XI Ng, Hong-Wei, Viet Dung Nguyen, Vassilios Vonikakis, and Stefan Winkler. "Deep learning for emotion recognition on small datasets using transfer learning." In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pp. 443-449. 2015.

XII Ouellet, Sébastien. "Real-time emotion recognition for gaming using deep convolutional network features." *arXiv preprint arXiv: 1408. 3750* (2014).

XIII Parkhi, Omkar M., Andrea Vedaldi, and Andrew Zisserman. "Deep face recognition." (2015).

XIV Sandbach, Georgia, Stefanos Zafeiriou, Maja Pantic, and Daniel Rueckert. "Recognition of 3D facial expression dynamics." *Image and Vision Computing* 30, no. 10 (2012): 762-773.

XV Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." In *Advances in neural information processing systems*, pp. 5998-6008. 2017.

XVI Xu, Mao, Wei Cheng, Qian Zhao, Li Ma, and Fang Xu. "Facial expression recognition based on transfer learning from deep convolutional networks." In *2015 11th International Conference on Natural Computation (ICNC)*, pp. 702-708. IEEE, 2015.

XVII Yan, Haibin, Marcelo H. Ang, and Aun Neow Poo. "Cross-dataset facial expression recognition." In *2011 IEEE International Conference on Robotics and Automation*, pp. 5985-5990. IEEE, 2011.

XVIII Zhang, Zhanpeng, Ping Luo, Chen-Change Loy, and Xiaoou Tang. "Learning social relation traits from face images." In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3631-3639. 2015.