Facial Expression Recognition using Residual Masking Network

Luan Pham, Huynh Vu, Tuan Anh Tran

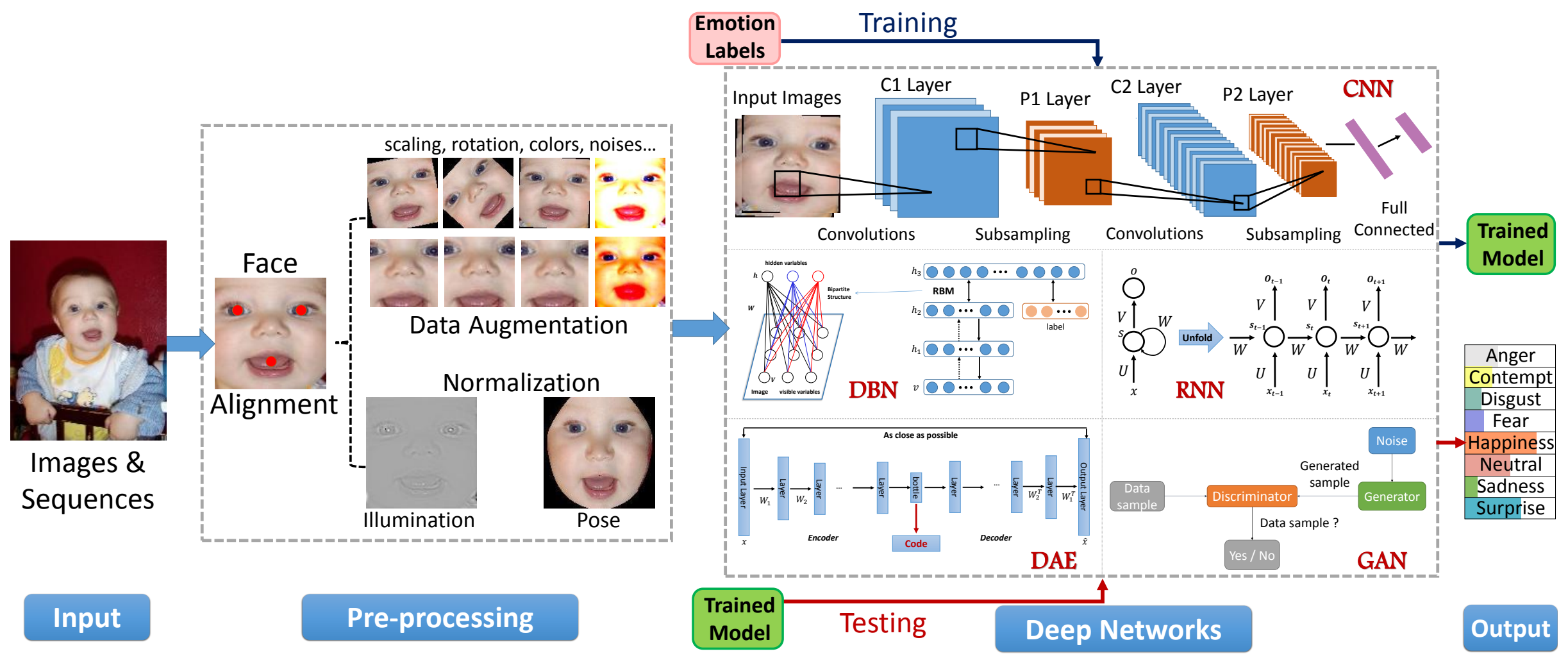
Research & Development - Cinnamon AI
Faculty of Computer Science and Engineering - HCMUT

INTRODUCTION

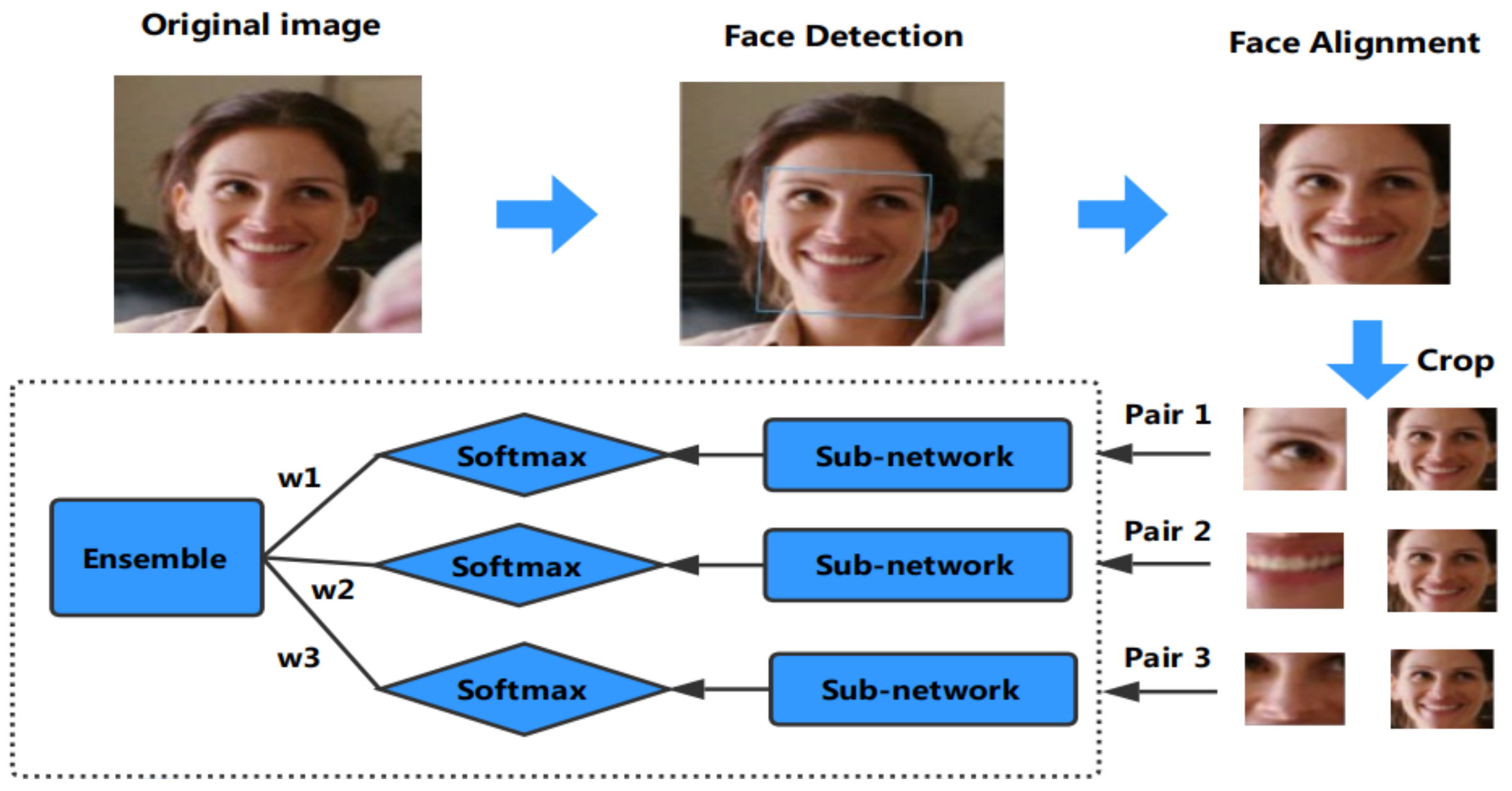
Automatic facial expression recognition (FER) has gained much attention due to its applications in human-computer interaction. Among the approaches to improve FER tasks, this paper focuses on deep architecture with the attention mechanism. We propose a novel Masking Idea to boost the performance of CNN in facial expression task. It uses a segmentation network to refine feature maps, enabling the network to focus on relevant information to make correct decisions. In experiments, we combine the ubiquitous Deep Residual Network and U-net like architecture to produce a Residual Masking Network. The proposed method hold competitive accuracy on the well-known FER2013 and private VEMO datasets.

RELATED WORKS

1. The general pipeline of deep facial expression recognition systems. [1]



2. Multi-region ensemble convolutional neural network for facial expression recognition [2]



REFERENCES

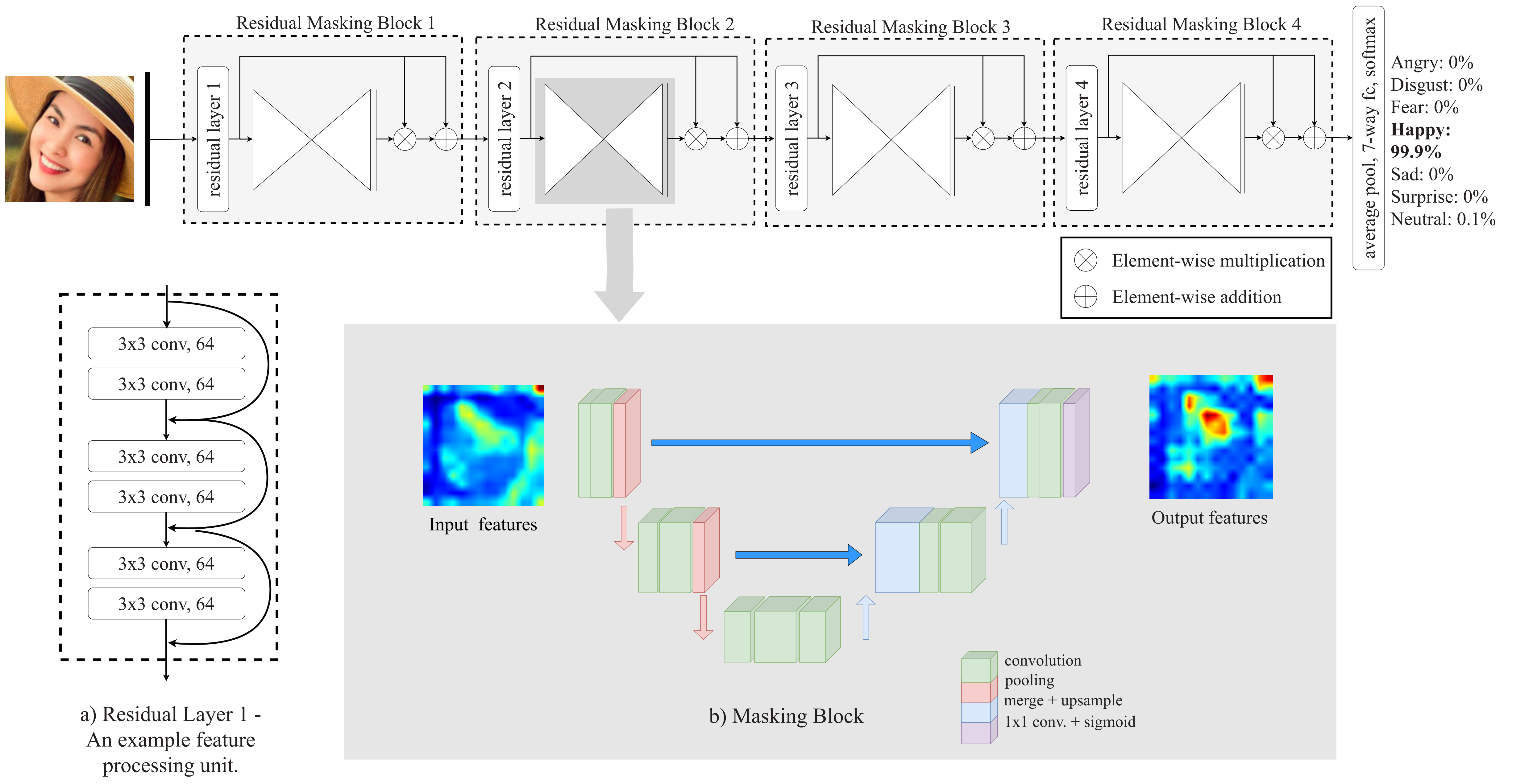
[1] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *arXiv preprint arXiv:1804.08348*, 2018.

[2] Yingruo Fan, Jacqueline CK Lam, and Victor OK Li. Multi-region ensemble convolutional neural network for facial expression recognition. In *International Conference on Artificial Neural Networks*, pages 84–94. Springer, 2018.

RESIDUAL MASKING NETWORK

The main flow of the proposed method is the Residual Masking Network illustrated in the Figure below. This network contains four main Residual Masking Blocks. Each Residual Masking Block, which operates on different feature sizes, contains a Residual Layer and a Masking Block.

An input image of size 224 × 224 will go through the first 3 × 3 convolutional layer with stride 2 before passing a 2 × 2 max-pooling layer, reducing its spatial size to 56 × 56. Next, the feature maps obtained after the previous pooling layer are transformed by the following four Residual Masking Blocks with generated features maps of four spatial sizes, including $56\times56, 28\times28, 14\times14$, and 7×7. The network ends with an average pooling layer and a 7-way fully-connected layer with softmax to produce outputs corresponding to seven facial expression states (6 emotions and one neutral state).



a) Residual Layer 1 - An example feature processing unit.

b) Masking Block

EXPERIMENTAL RESULTS

Evaluation results on FER2013

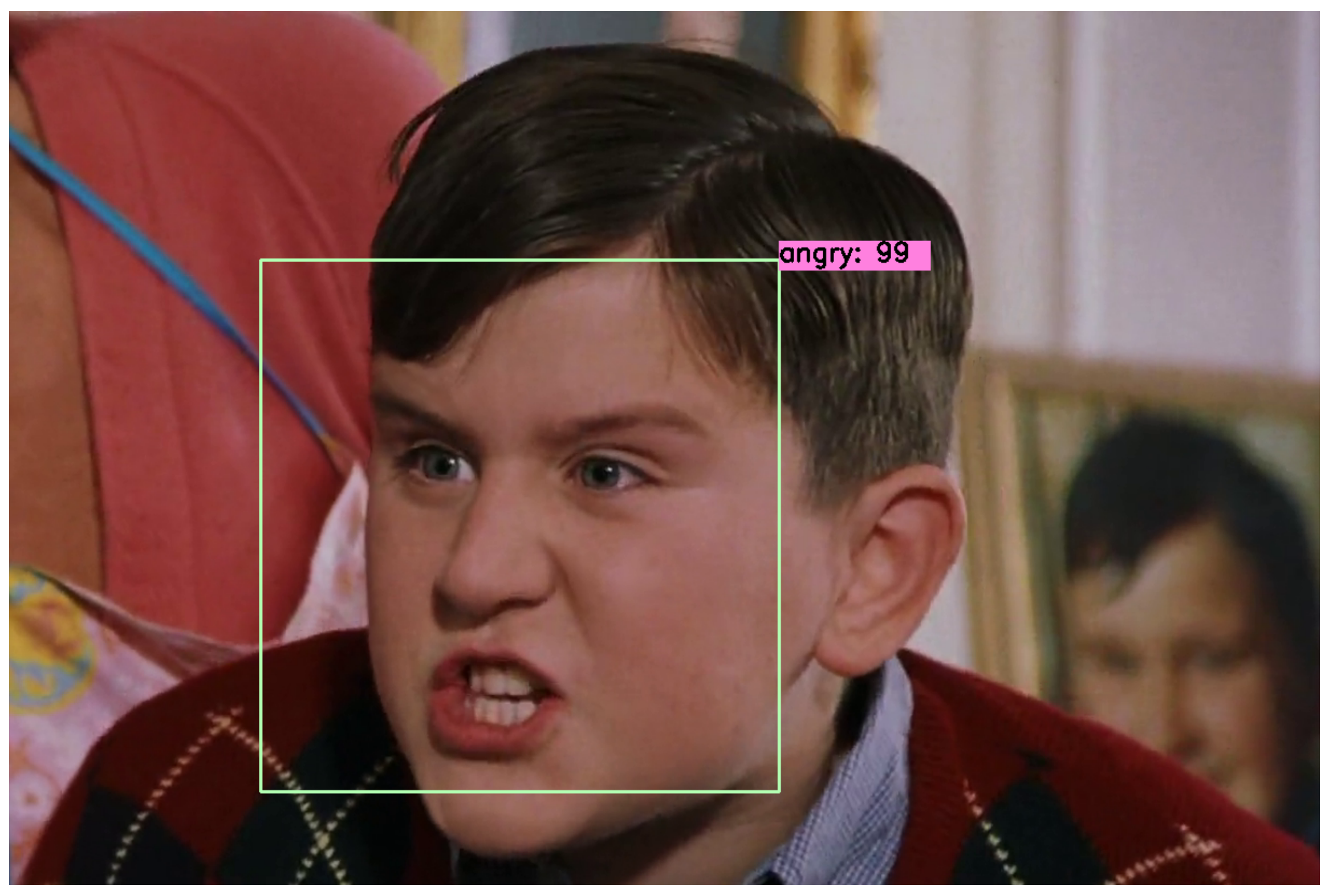
Model	Accuracy (%)
VGG19	70.80
Resnet18	72.90
DenseNet121	73.17
Inception_V3	72.72
Efficientnet_B2B	70.80
ResMaskingNet	**74.14**

Evaluation results on VEMO

Mô hình	Accuracy (%)
DenseNet121	59.95
ResAttNet56	60.82
Resnet18	63.94
Resnet34	64.84
ResMaskingNet	**65.95**

(*): Citations could be found in the paper.

Demo images



Dudley being angry - Harry Potter movie

A Vietnamese actress being sad.