

Facial Emotion Recognition with Deep Neural Networks

By Ameya Anand Kamat, Arghadeep Ghosh and Shankar Ram Vasudevan

5 May, 2022



Facial Emotion Recognition (FER)

- Facial expressions play an important role in daily human-human communication.
- Automatic facial expression analysis is an important area of artificial intelligence with several potential applications.
- FER faces many challenges in the form of illumination variation, occlusions, variant poses, identity bias, insufficient qualitative data, etc.



Our Goals

- **Our goals in this project is to read and understand the paper,**

Region Attention Networks for Pose and Occlusion Robust Facial Expression Recognition by Kai Wang, Xiaojiang Peng, Jianfei Yang, Debin Meng, and Yu Qiao, Senior Member, IEEE

- **Build an effective neural network that is able to accurately determine the emotion expressed in a given face image.**
- **Ensure that the network is able to overcome aforementioned challenges such as occlusions and variant poses.**

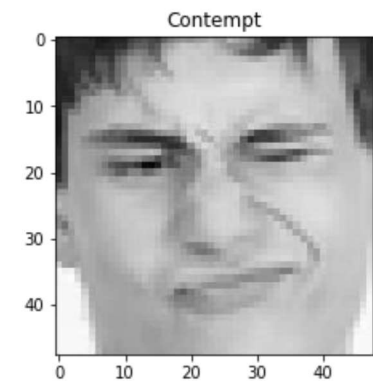
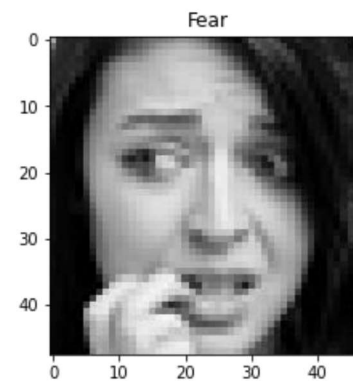
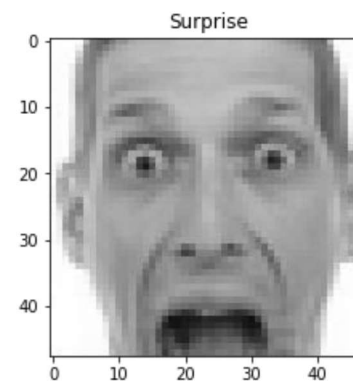
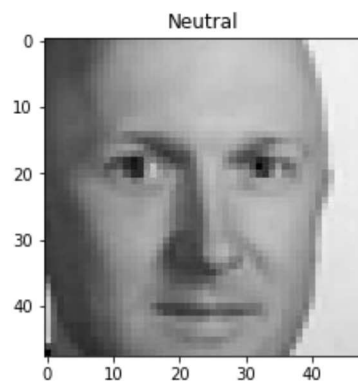
FERPlus

The FERPlus 2013 Dataset contains 35710 48×48 face images out of which we have taken 32137 Training Images and 3173 Testing Images.



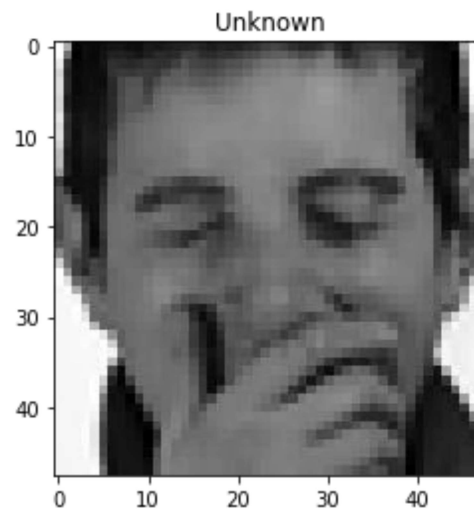
These images are classified into 9 categories.

Eight Emotions...

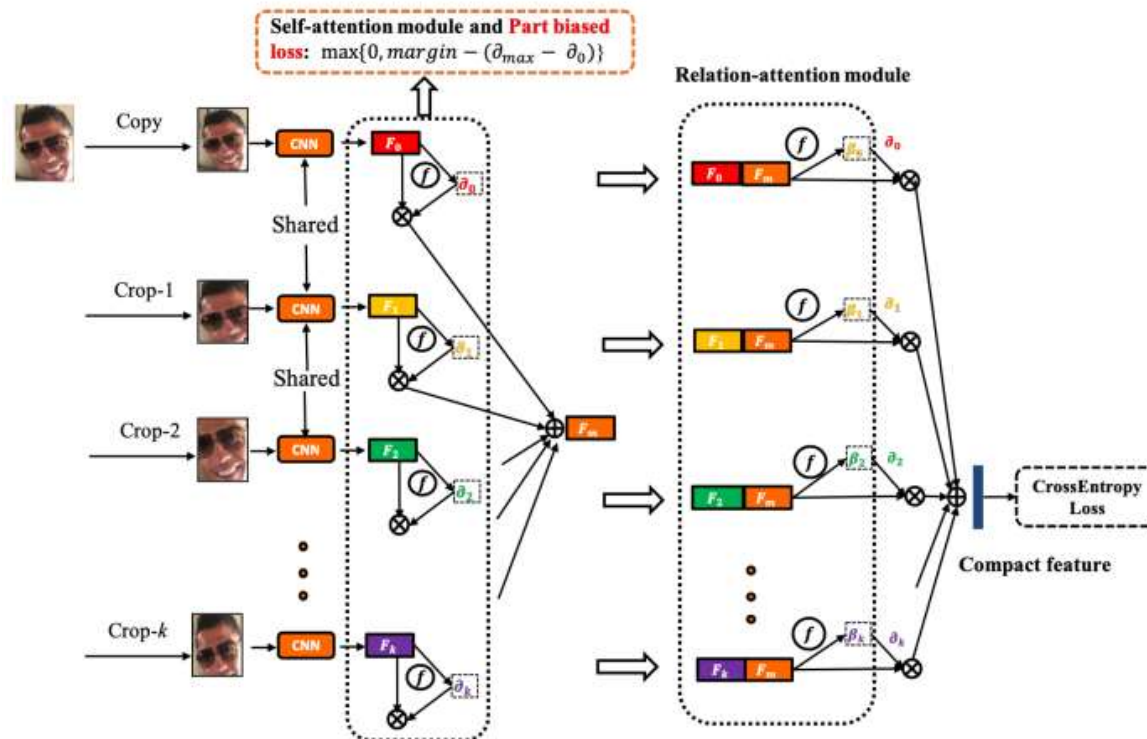


...and Unknown

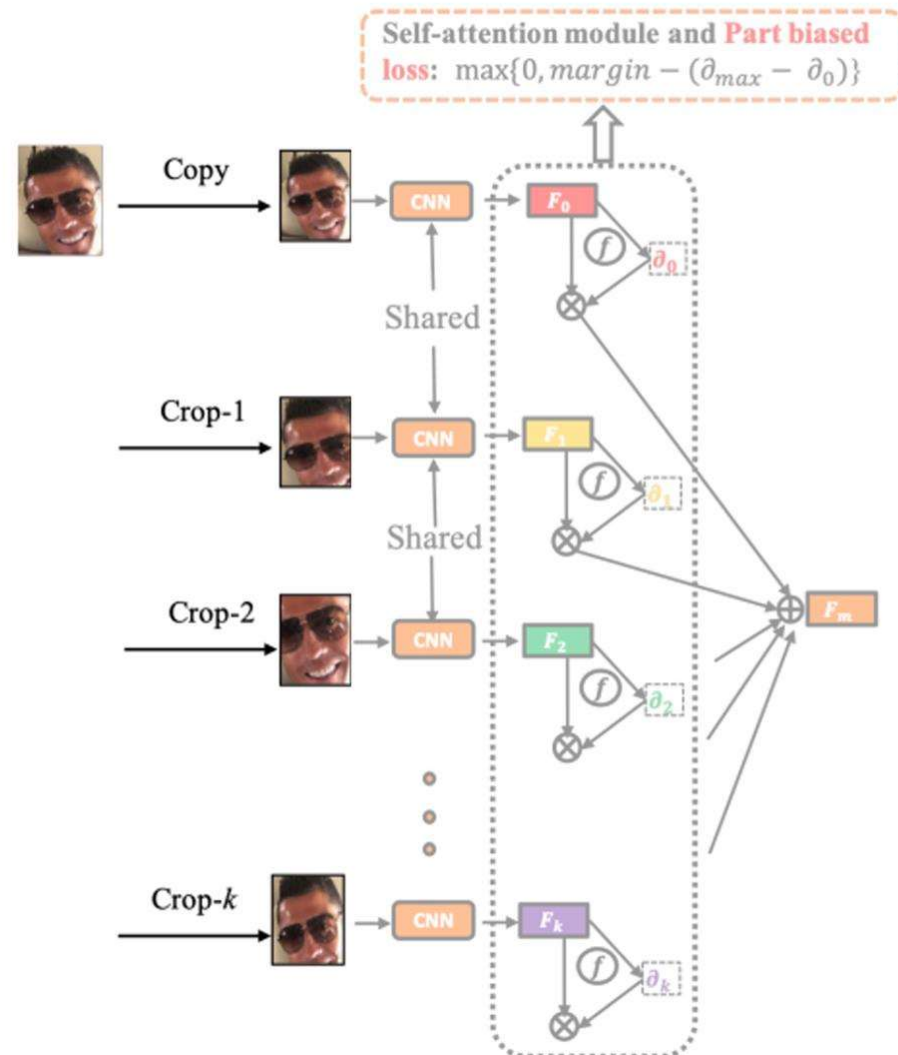
For images that don't fit any of the above descriptions



The Region Attention Network Framework

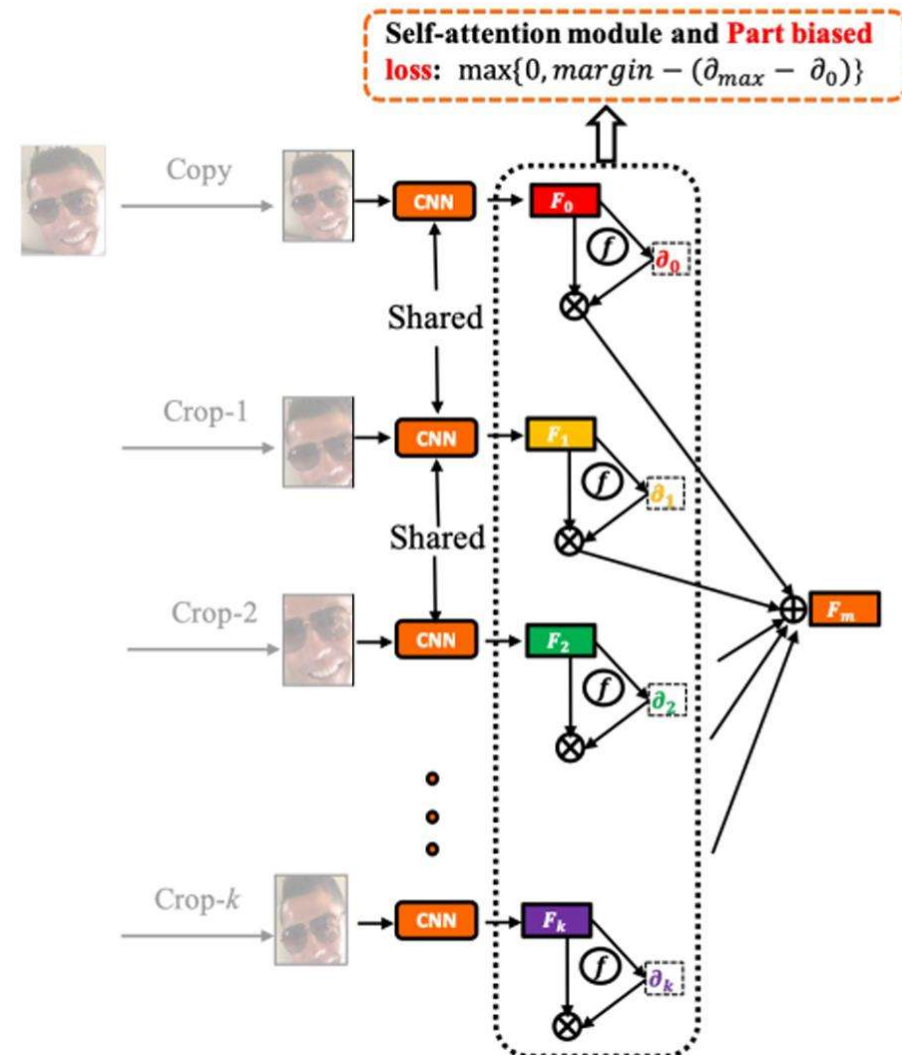


- The image is cropped into k sub-images to capture the importance of facial regions.
- The cropped regions allow for robust classification in occlusion and pose-variant FER.



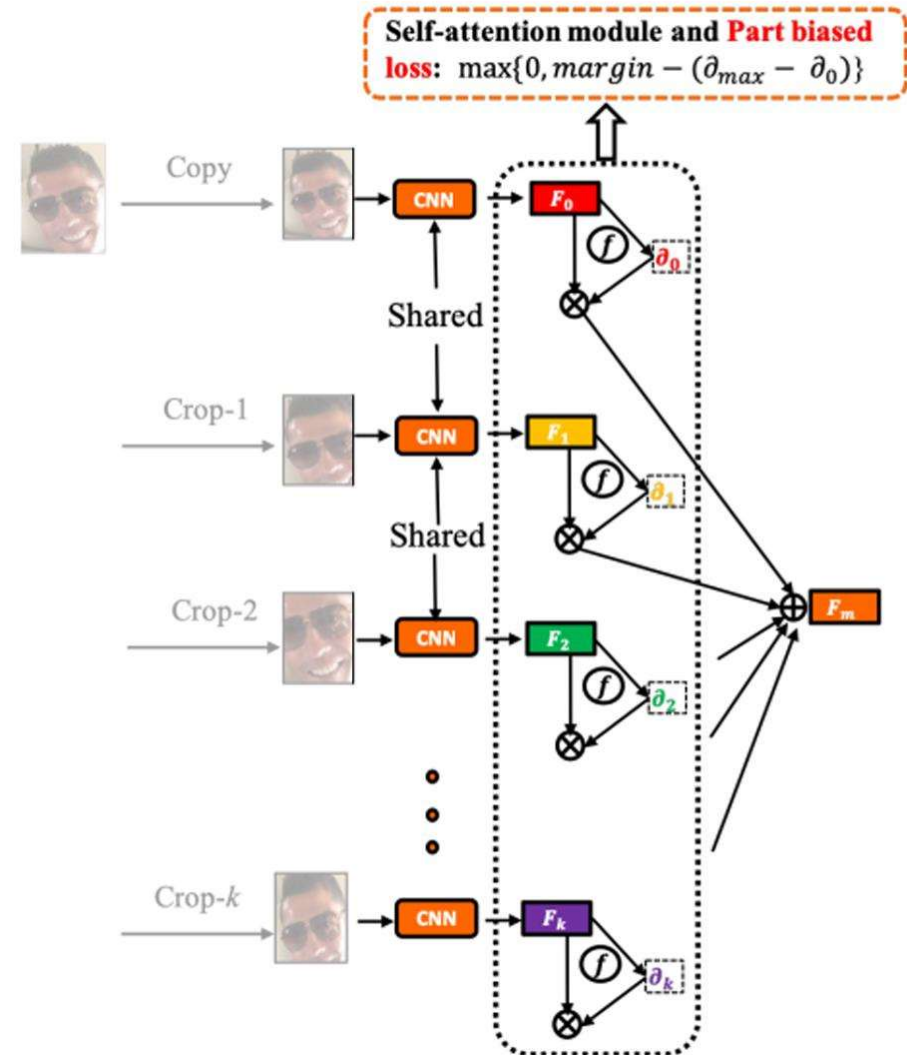
- The cropped images are passed through a backbone CNN in order to obtain an encoded feature F_i vector for each crop.
- The self-attention module applies a fully connected layer and a sigmoid function to estimate the weightage of each feature given by,

$$\mu_i = f(F_i^\top \mathbf{q}^0)$$



- The features F_i are combined in order to summarize all region vectors by taking a weighted average.
- The obtained vector is:

$$F_m = \frac{1}{\sum_{i=0}^n \mu_i} \sum_{i=0}^n \mu_i F_i$$



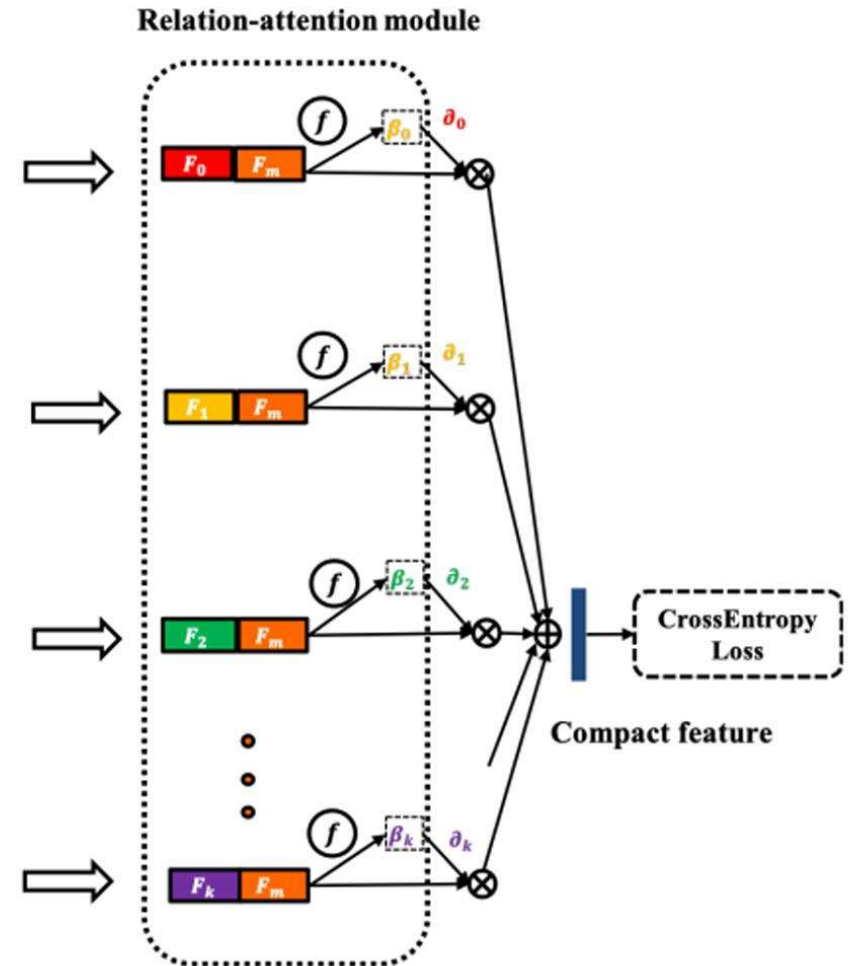
Region Biased Loss



- The Self-Attention module generates a Region Biased loss adding a constraint that enforces one of the attention weights from facial crops to be larger than the original face image by a margin.
- The margin used for our network is 0.1

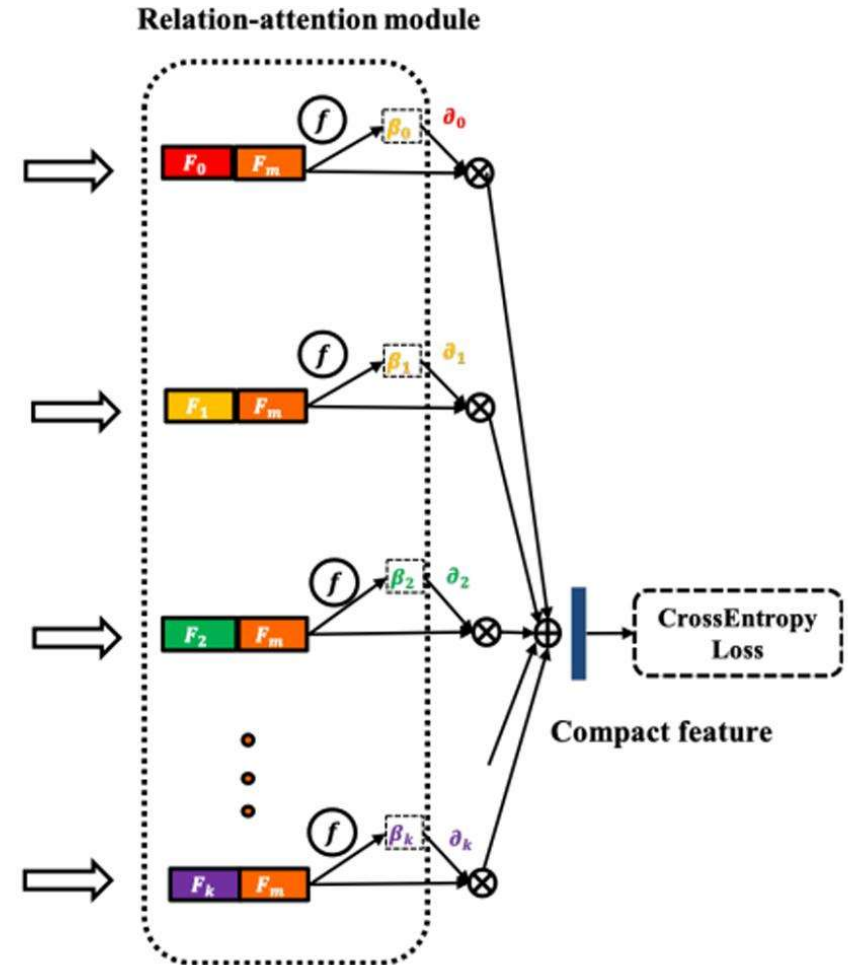
- The second part of the network, Region-attention module takes each feature vector \mathbf{q} and F_i summary vector F_m
- The new attention weights are formulated as

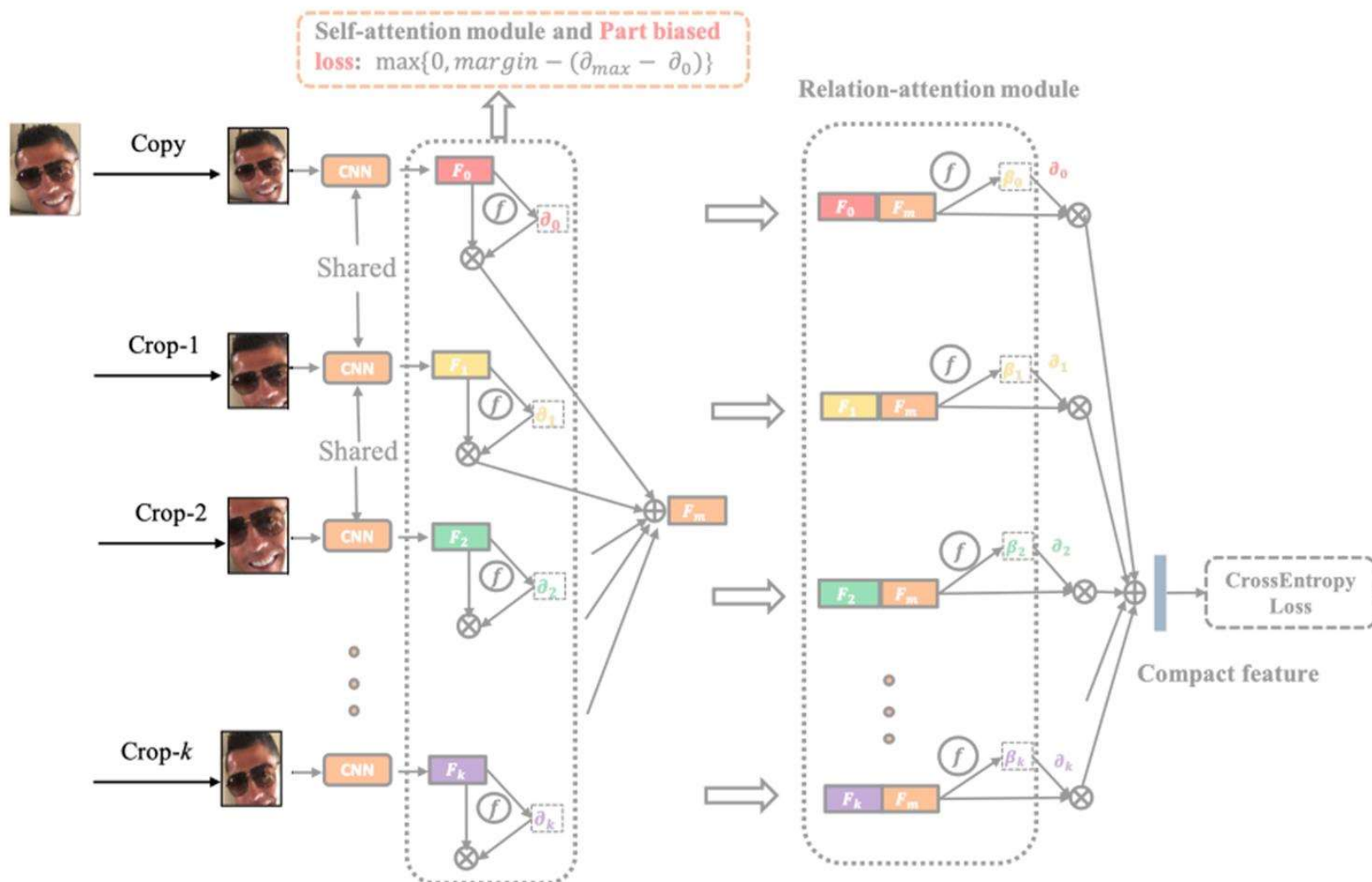
$$\nu_i = f([F_i : F_m]^\top \mathbf{q}^1)$$

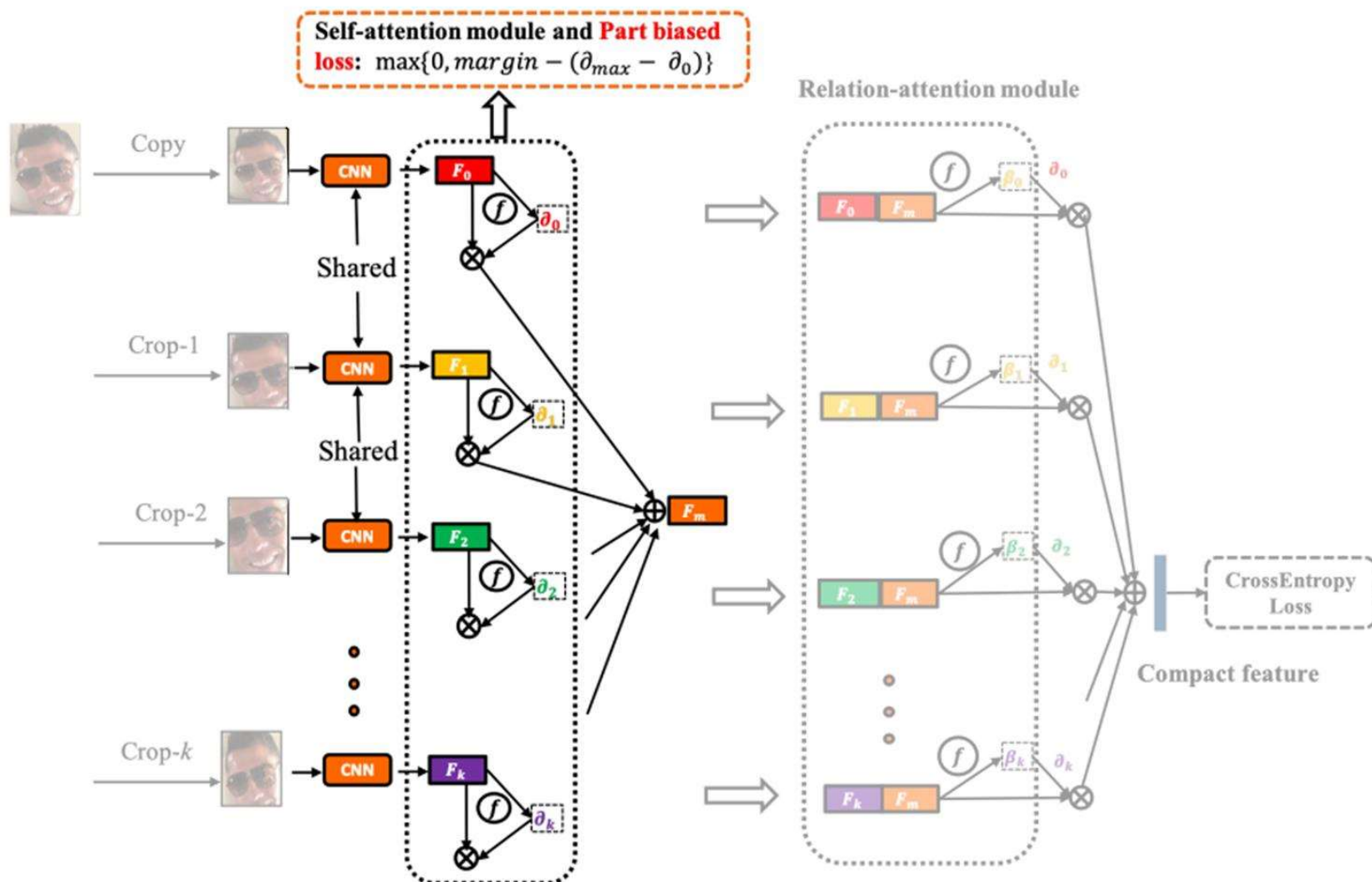


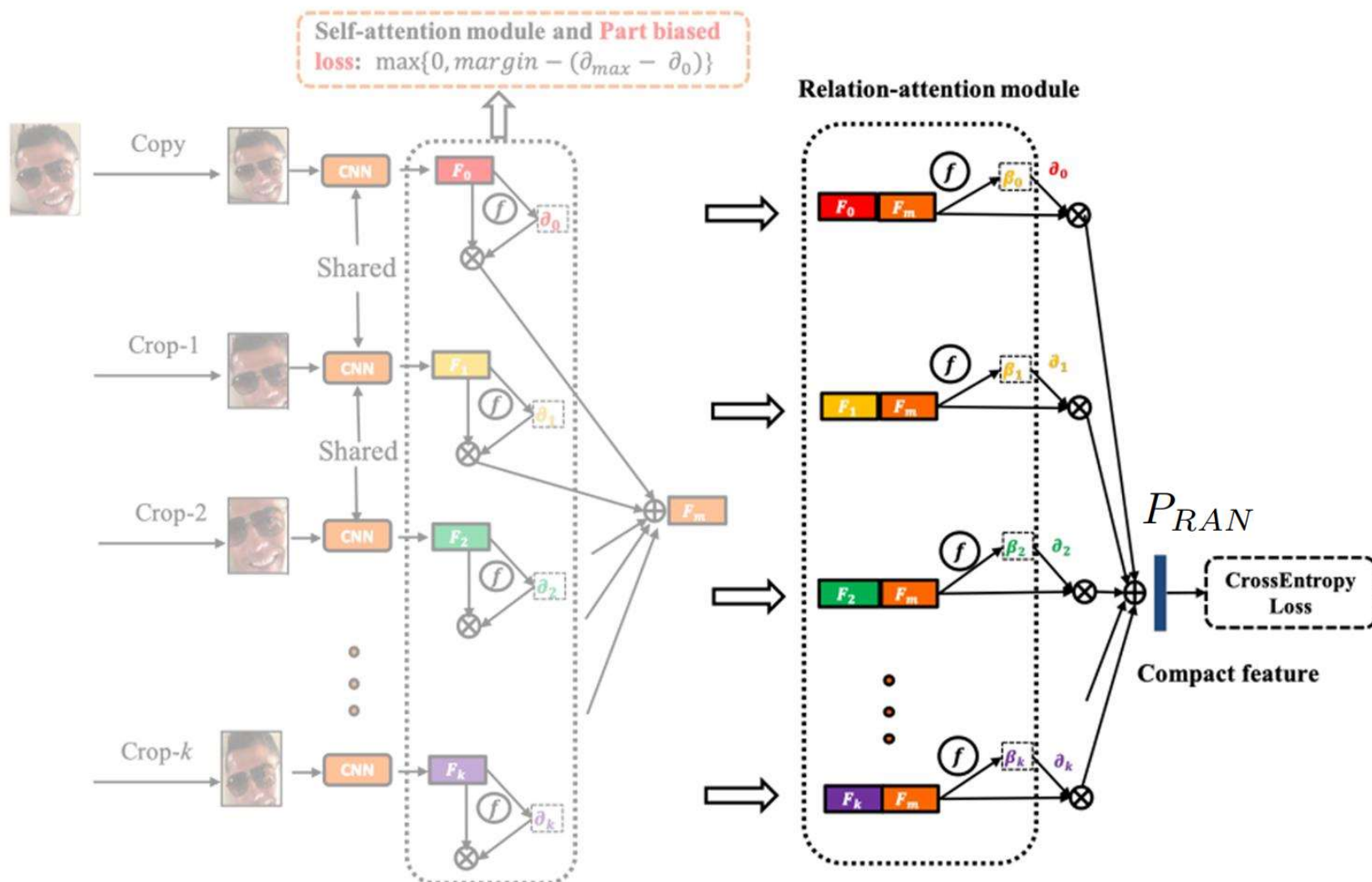
- Finally, all the region information and the coarse global information is combined to form the final output vector.
- This new vector P_{RAN} is given by

$$P_{RAN} = \frac{1}{\sum_{i=0}^n \mu_i \nu_i} \sum_{i=0}^n \mu_i \nu_i [F_i : F_m].$$

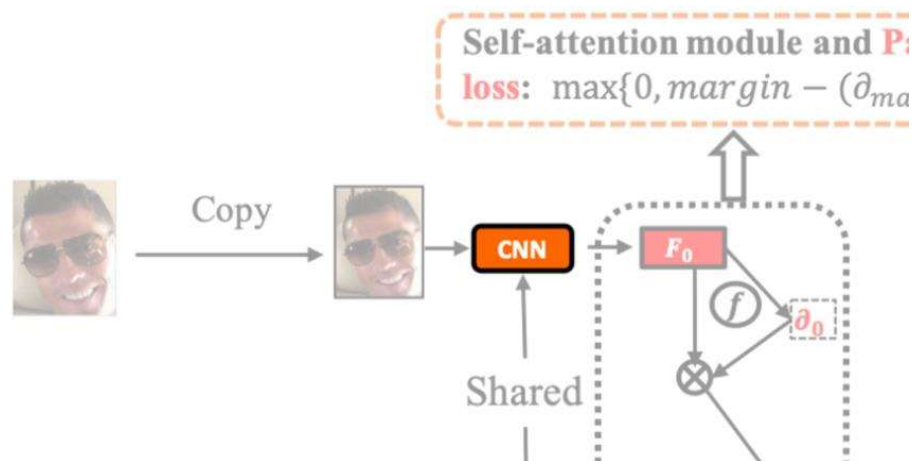






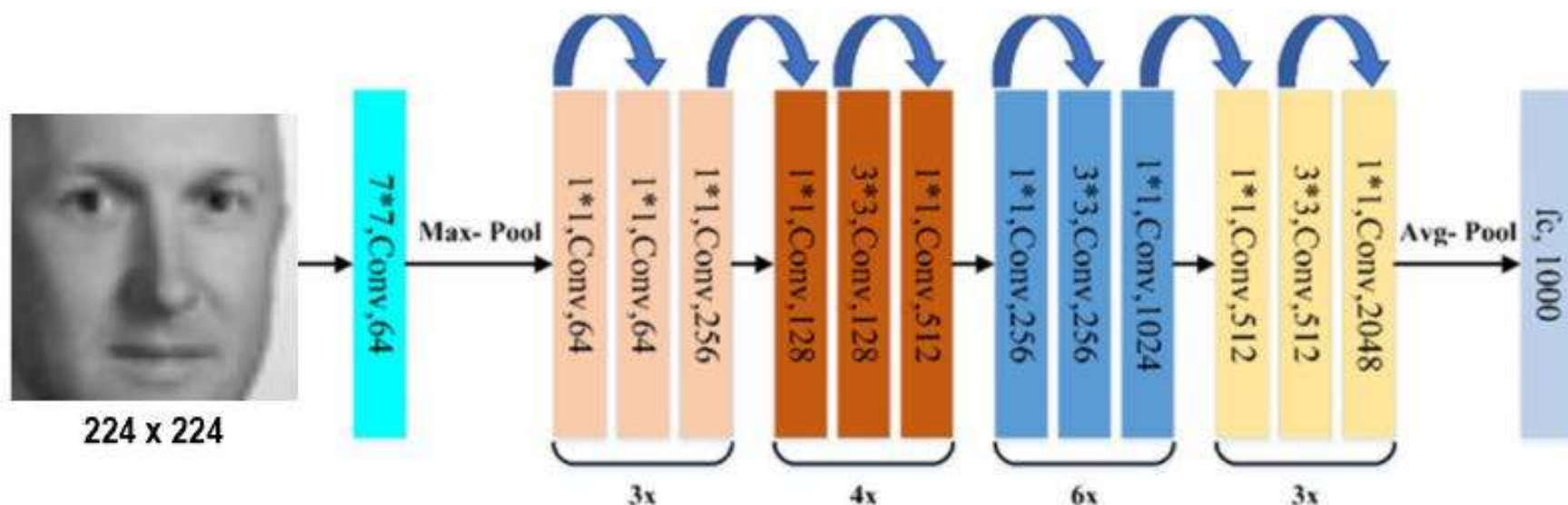


The Base CNN

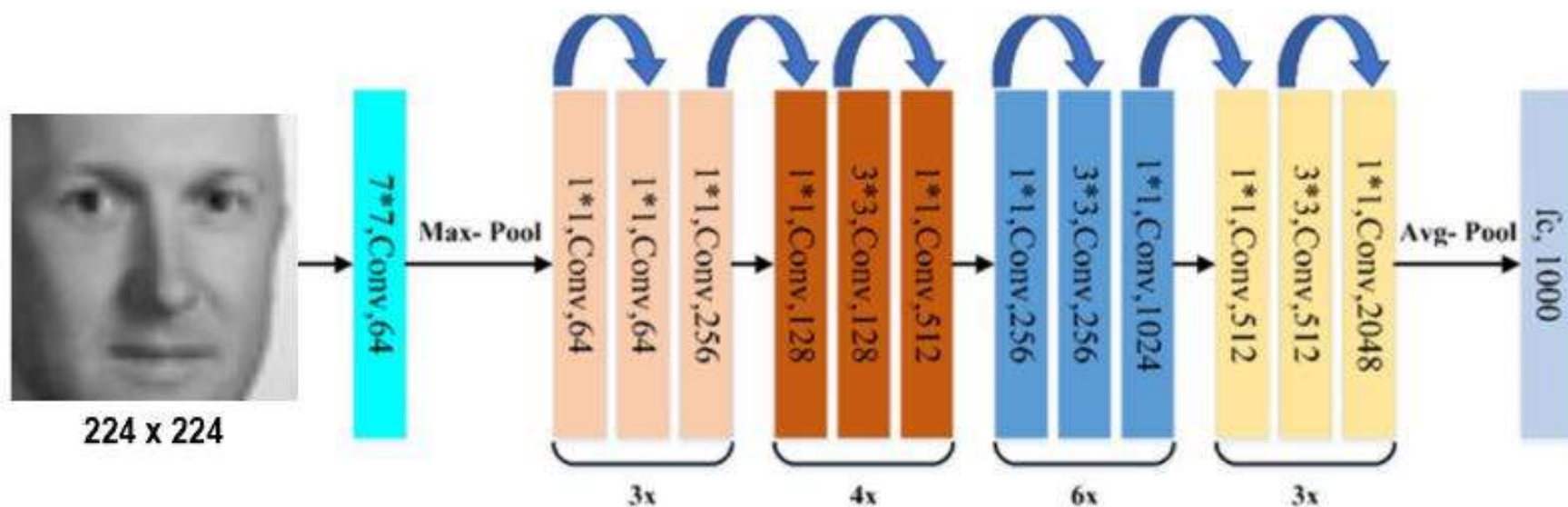


The backbone CNN is used to generate the feature vectors from each of the cropped regions.

The CNN Architecture we have chosen is iResNet50.



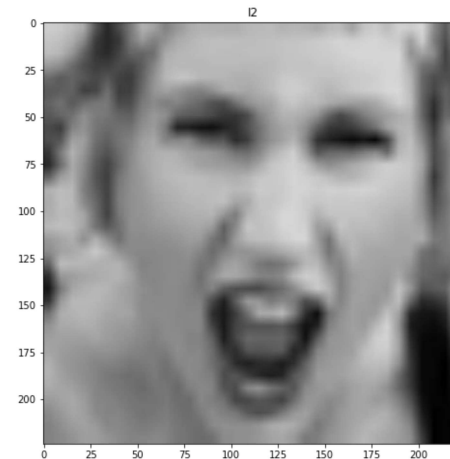
- Residual Networks (ResNets) are widely used for facial image classification problems. They use “Skip-Connections”, or shortcuts to jump over layers.
- Adding skip connections helps avoid vanishing gradients, and mitigates the accuracy saturation problem allowing us to train deep networks effectively.



- The Architecture chosen by us, IResNet50 is a residual network containing 50 layers. The architecture including the types of layers and the skip connections are illustrated in the image above.
- IResNet50 works on inputs of size 224x224.

Preparing the Dataset

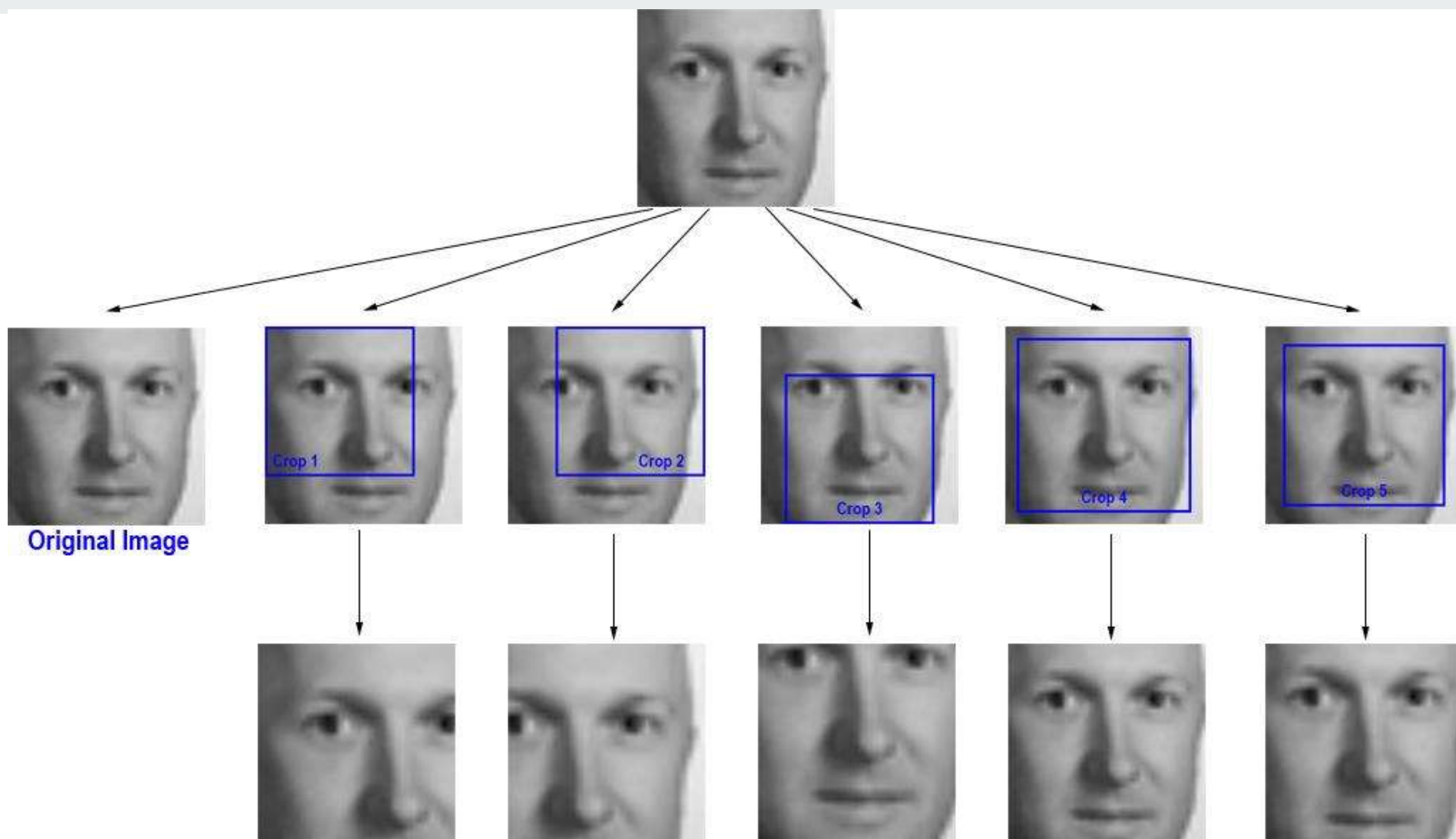
We have used Dlib 68-point facial landmark detector in order to locate the faces in complex scenes. Further, the image is scaled to size 224x224 to be passed through IResNet50.





Region Cropping Module

- The first step of our model involves selecting appropriate regions to be passed to our network for feature extraction.
- The training procedure allows the model to adaptively capture the importance of each region.
- We have chosen a methodology which selects and crops five fixed regions from the image along with the original image.

















Evaluating our model

- The entire model is trained on the FERPlus Training data for 70 epochs.
- The model has been evaluated on the aforementioned testing data.

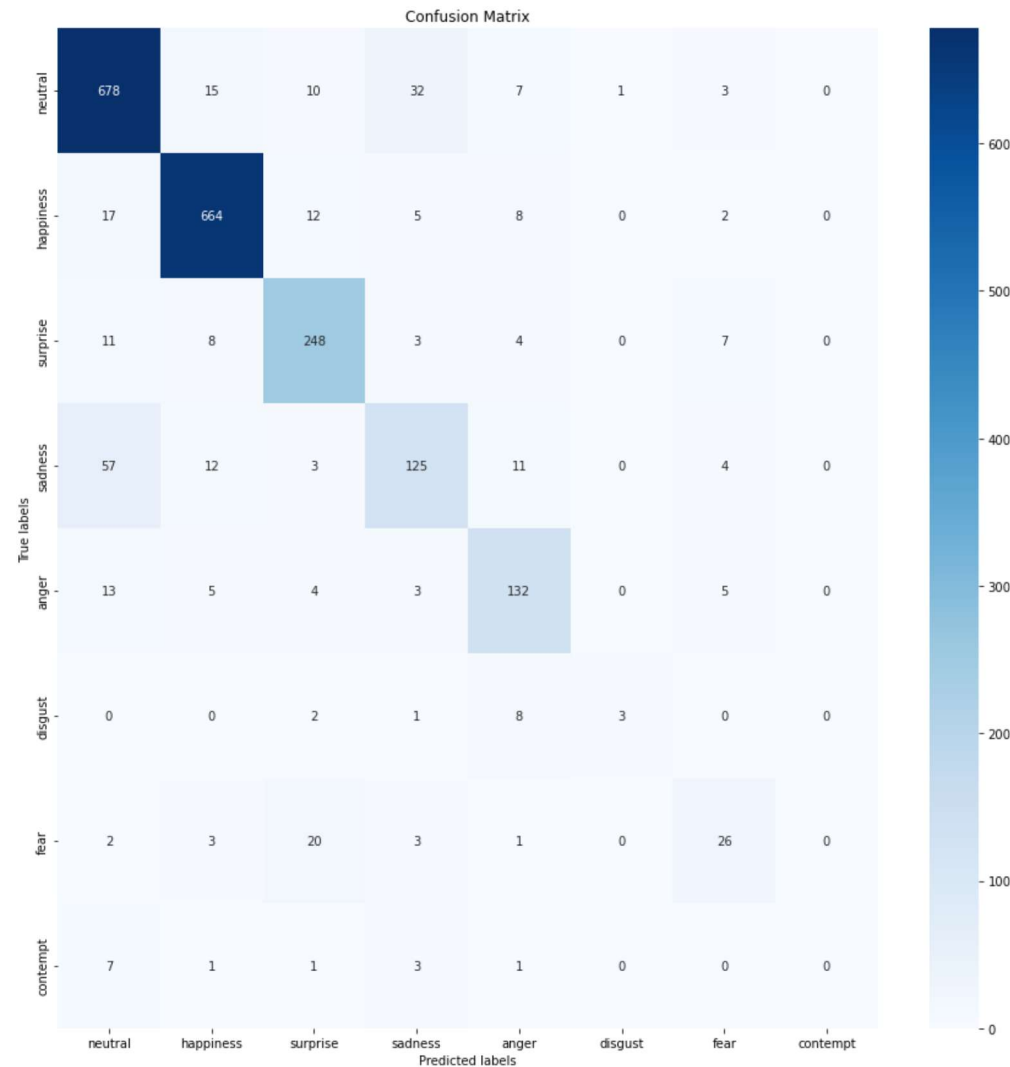
Some of the results from the testing procedure have been displayed.

Test Image	Cropped Images	Model Prediction	Actual Classification
		Happiness	Happiness
		Anger	Anger
		Happiness	Contempt

Test Image	Cropped Images	Model Prediction	Actual Classification
		<p>Surprise</p>	<p>Surprise</p>
		<p>Happiness</p>	<p>Happiness</p>
		<p>Sadness</p>	<p>Surprise</p>

Our Results

- Upon running the network over the entire Test Dataset we have obtained an accuracy of 86%.
- The confusion matrix illustrates the classifications of our model.
- Note: The 9th Category has been eliminated by Dlib 68-point facial landmark detector due to the lack of clear faces.



Occlusion and Pose Variant Robustness

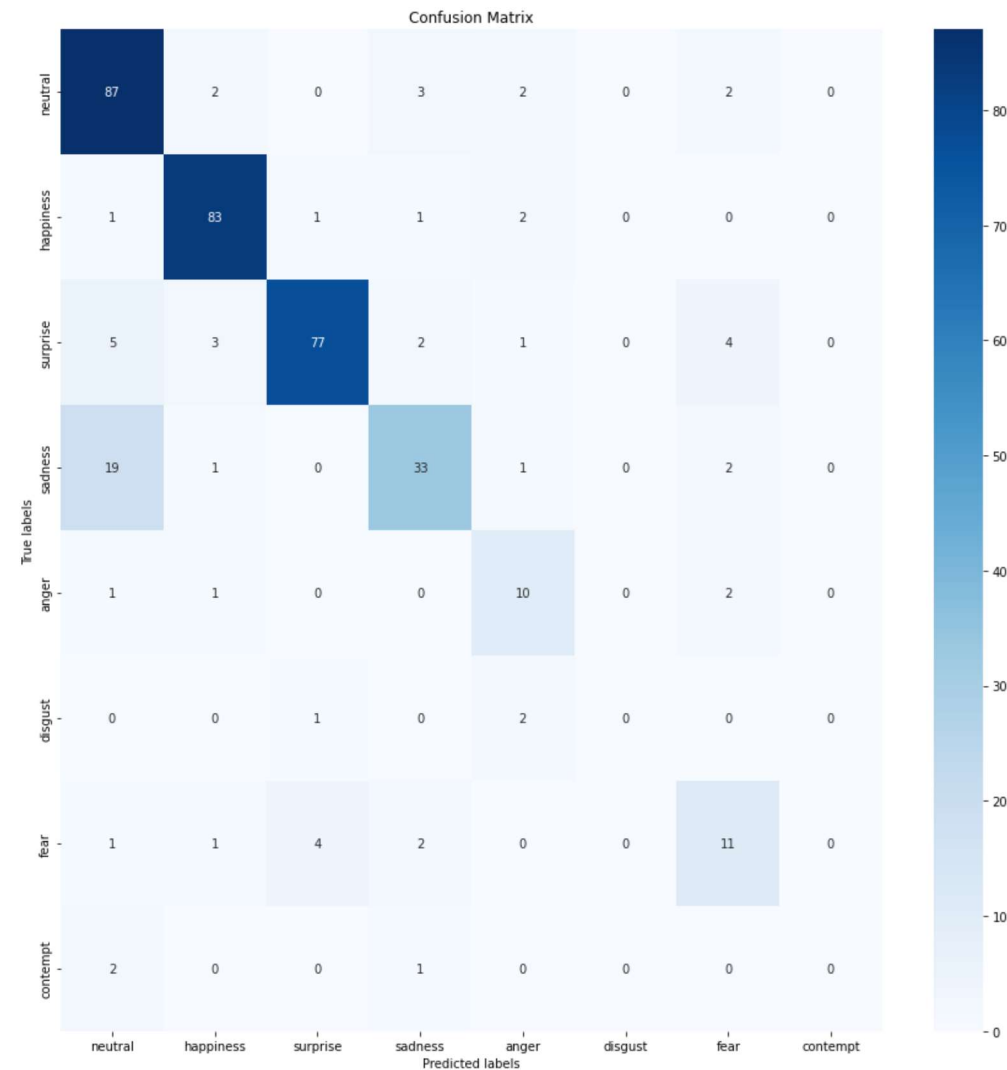
- To verify the robustness of our network a subset of the training images are selected as Pose variant or Occluded Images.



- The Testing procedure is run on this subset separately to measure performance.

Occlusion and Pose Variant Results

- Upon running the network over the Occlusion and Pose Variant Test Dataset we have obtained an accuracy of 81%.
- The confusion matrix illustrates the classifications of our model over this test set.





Thank You!