# Facial Emotion Recognition with Noisy Multi-task Annotations **Supplementary Material**

## A. Architecture and Implementation Details

The detailed architecture is illustrated in Fig. 1. The encoder $G_Y$ is a modified VGGNet [47] which predicts clean labels $\hat{y}^1, \ldots, \hat{y}^T$, a mean $\mu$, a variance $\sigma$, from which the latent noise vector $y^0$ is sampled as $y^0 \sim \mathcal{N}(\mu, \sigma^2 \boldsymbol{I})$ (Fig. 1 (a)). The decoder $G_X$ (Fig. 1 (b)) takes the concatenation of a Gaussion noise $\tilde{y}^0 \sim \mathcal{N}(0, \boldsymbol{I})$, and all the input noise labels $\tilde{y}^1, \ldots, \tilde{y}^T$ as the input, which is first fed through a linear layer, and then upsampled by deconvolutional blocks to produce the image $\tilde{x}$.

The discriminator $D$ consists of separate streams for each input variable, which is a CNN stream for the input image, and different multilayer perceptron (MLP) streams for the input labels or noise. The marginal scores $S_x, S_{y^0}, \ldots, S_{y^T}$ are computed as a linear transformation of the output features of each stream. In the meantime, the output features of each stream are concatenated into an MLP to produce the joint score $S_{joint}$ for the joint distribution matching. The CNN stream for the image includes several residual blocks and one attention block. Each residual block is a simple residual convolutional block which contains two [convolution, ReLU] blocks and one pooling operation. The attention block is a self-attention CNN block [60] which aims to utilize features from all locations, for modeling long range and multi-level dependencies. The MLP stream is an MLP block which gives the summation of outputs of four separate MLP sub-blocks. See the detailed architecture layouts of residual block, the attention block, and MLP components of $D$ in Fig. 2.

The proposed model is implemented with PyTorch, using ADAM [22] as the optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$). The learning rate are set to 1e-4. $\gamma$ is always set to 1. In each iteration of the alternating optimization schedule, $G_Y$ and $G_X$ are jointly updated once, followed by two consecutive updates of $D$. See analysis of the joint distribution learning weight $\lambda$ in Section C.

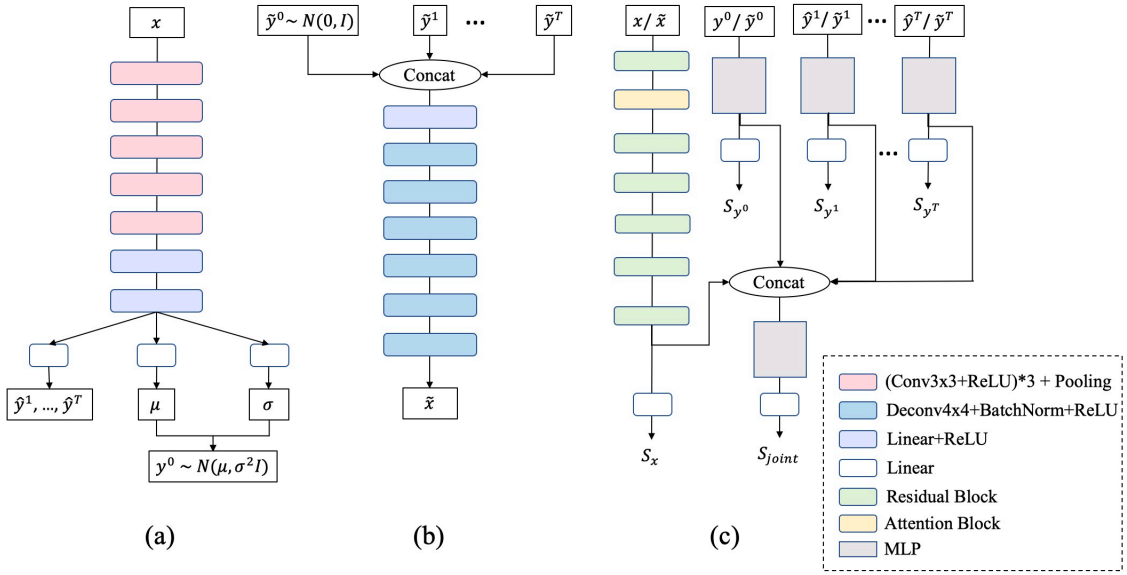



Figure 1. Architecture details of (a) the encoder $G_Y$, (b) the decoder $G_X$, (c) the discriminator $D$.

## B. Confusion Matrix and Visualization

The confusion matrices for the multi-task models on facial emotion datasets in both RAF-base and AffectNet-base cases are shown in Fig. 3. The confusion matrices of the pretrained models which are utilized to create noisy training set are only for reference. Comparing the confusion matrices of the multi-task VGGNet models and the proposed multi-task models, the diagonal values of the proposed models are generally higher than the VGGNet models, showing that the proposed model can achieve more precise predictions by the joint distribution learning from multi-label models. However, the confusion matrices of the three different kinds of models still show similar patterns, suggesting that learning from noisy labels in practical emotion dataset is still a difficult task.

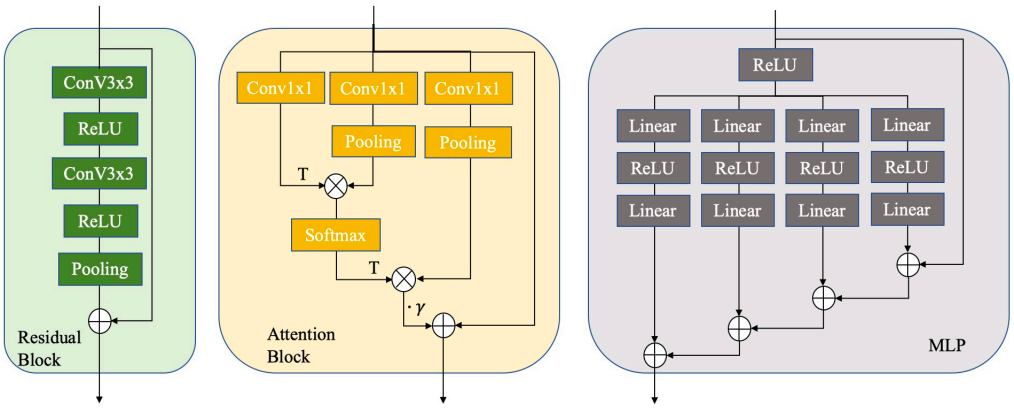



Figure 2. Detailed architecture layouts of components of discriminator $D$: (a) the residual block, (b) the attention block (T denotes transpose and $\gamma$ is a learnable parameter), (c) the MLP block.

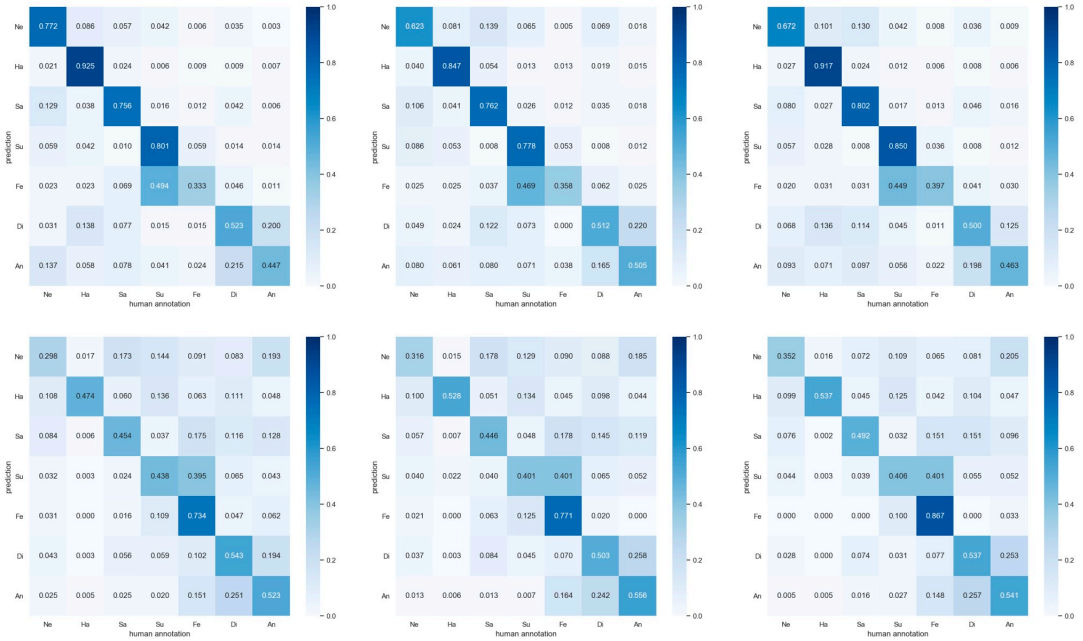



Figure 3. Confusion matrices for the RAF-base (the first row) and AffectNet-base (the second row) cases: (a) the pretrained model to create noisy training set, (b) multi-task VGGNet, (c) the proposed multi-task model trained on the noisy labels.

Images in Fig. 4 are examples of success and failure cases of the proposed multi-task model in RAF-base and AffectNet-base cases. Our proposed model gives correct labels in many cases where the Co-teaching method [16] or multi-task VGGNet gives inconsistent labels with the human annotation, which we believe to be relatively clean in the test set. There are also some failure cases of the proposed model, for example, where sad faces with mouth open are predicted to be happy, or happy faces with extreme activated expressions are predicted to be angry.

## C. Study on Joint Distribution Weight $\lambda$

Fig. 5 illustrates the test accuracy curve of the proposed multi-task model on both RAF-base and AffectNet-base cases with different joint distribution learning weights $\lambda$ varying among 0.2, 0.4, 0.6, 0.8 and 1.0. Since the joint distribution learning serves as a regularizar in the proposed model, the optimal choice of $\lambda$ varies among different datasets depending

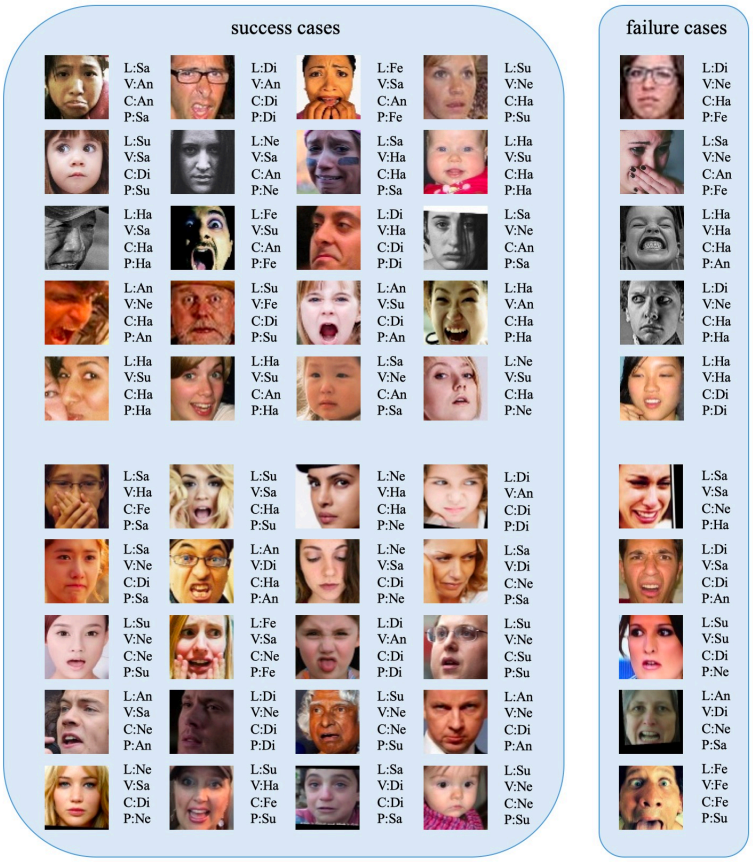
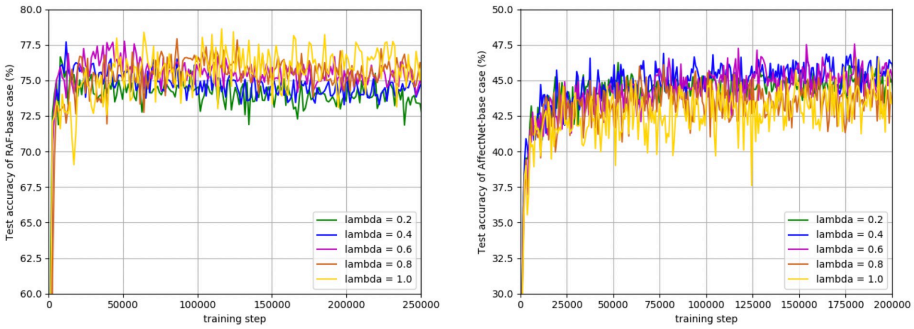



Figure 4. Examples of images and corresponding predictions on RAF-base (the first five rows) and AffectNet-base (the last five rows) cases, with the first four columns where the predictions of our proposed multi-task model are consistent with the human label in the test set, and the last column contains some failure cases of the proposed model. L: human label of test set, V: multi-task VGGNet, C: Co-teaching model [16] trained with only emotion class labels, P: proposed multi-task model.

on the noise intensity. In spite of this, the value of $\lambda$ can be roughly set according to the prior knowledge about the noise intensity. For example, we are given the prior knowledge that the human annotations on AffectNet are relatively worse than those on RAF, as RAF adopts an EM algorithm to achieve a better annotation on about 40 annotations on each sample while AffectNet's annotation is merely from one single annotator. Accordingly, the pretrained model on AffectNet is assumed to offer relatively noisy labels than the one trained on RAF. Using this prior knowledge, we roughly set $\lambda$ to two values, i.e., 0.4 and 1.0, for the model training on AffectNet and RAF respectively. In addition, a general pattern can be observed: the model performance will increase with $\lambda$ increasing in a certain range, and then decrease with $\lambda$ increasing further as the joint distribution learning overweights the task-specific losses (i.e., cross-entropy loss and CCC loss in our case).

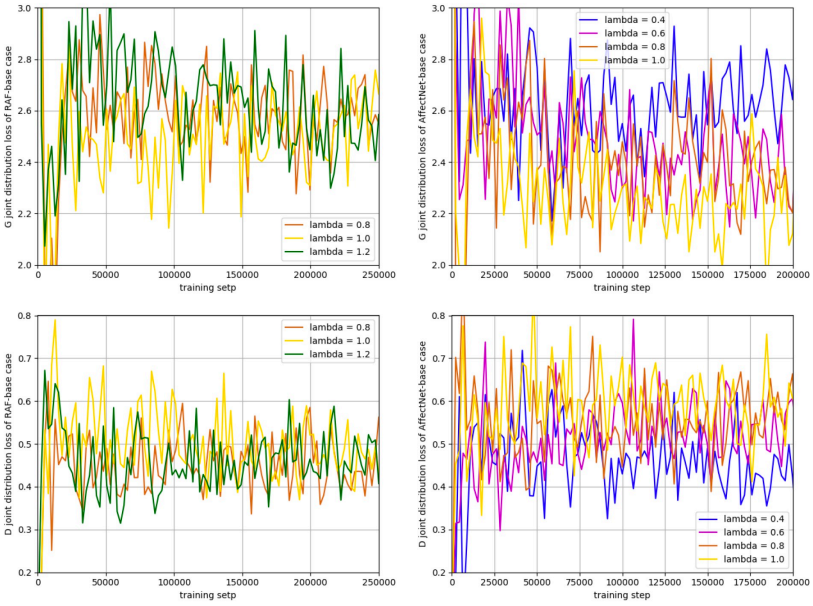
In addition to using the prior knowledge on the noise intensity, we find the training curve of the joint distribution loss can also help us to infer a suitable $\lambda$ for better training of the proposed model. We empirically observe that the optimal models in both RAF-base ($\lambda = 1.0$) and AffectNet-base ($\lambda = 0.4$) cases consistently converge to a value around 2.6 for the generator joint distribution loss, and a value around 0.45 for the discriminator joint distribution loss (see Fig. 6). The RAF-base case is more robust when $\lambda$ is varied in a certain range, because it requires a higher $\lambda$ due to the higher noise intensity in labels. Therefore, this observation can be used as the second strategy to roughly estimate an appropriate $\lambda$ to train the proposed model on new data with noisy multi-task labels.

If the incorrect labels form a specific distribution with a high noise ratio, the joint distribution learning might fail due to the large gap between the noise label distribution and the true label distribution. However, in most practical cases of facial emotion machine labels, the incorrect labels are outliers of the true distribution, and the proposed distribution-to-distribution supervision is based on such assumption, therefore robust to the noise.




(a) (b)

Figure 5. Curve of test accuracy during training with different joint distribution learning weights (i.e. $\lambda$ values) varying among 0.2, 0.4, 0.6, 0.8, 1.0 on (a) RAF-base (b) AffectNet-base cases. The optimal $\lambda$ is selected as 1.0 and 0.4 for RAF-base and AffectNet-base cases respectively.




(a) (b)

Figure 6. Curve of joint distribution loss during training with different joint distribution learning weights (i.e. $\lambda$ values) (a) RAF-base (b) AffectNet-base cases. The first row is the joint distribution loss of generator (i.e., the encoder $G_Y$ and decoder $G_X$), and the second row is the joint distribution loss of the discriminator $D$).