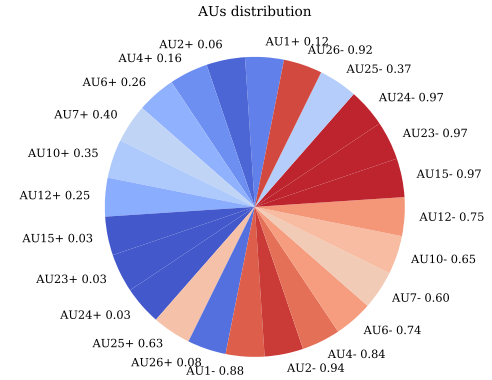
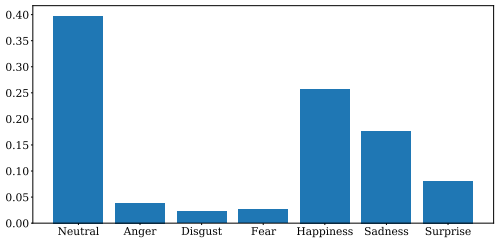
Appendices

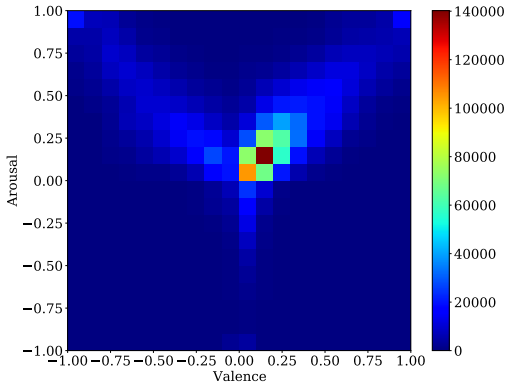
A. Data Distribution



(a) The data distribution for the AU subset.



(b) The data distribution for the EXPR subset.



(c) The data distribution for the VA subset.

Figure 6: The data distributions of the Aff-wild2 dataset.

The p_c for class reweighting depends on the data distributions. From AU1 to AU26, $\mathbf{p} = [7.7, 24.7, 5.3, 2.9, 1.5, 1.9, 3, 32.3, 32.3, 32.3, 0.59, 11.5]$ to alleviate the unbalanced data problem. For the EXPR subset, $\mathbf{p} = [0.02, 0.2, 0.33, 0.24, 0.03, 0.05, 0.1]$ in Equation 5.

B. Multitask Balancing

Our multitask balancing algorithm is given as follows:

Algorithm 1 Balancing Multitask Weights

Input

The multitask model f.
The training set $\mathcal{D}_{train}$ and the validation set $\mathcal{D}_{val}$.
Tasks set $\mathcal{T} = \{AU, EXPR, VA\}$.
The number of tasks $n(\mathcal{T}) = 3$.
The number of epochs with no performance improvement $\mathcal{M} = \{m^{AU}, m^{EXPR}, m^{VA}$.
The weights for all tasks $\Lambda = \{\lambda^{AU}, \lambda^{EXPR}, \lambda^{VA}\}$.
The number of training epochs M.

procedure
  while $i < n(\mathcal{T})$ **do**
    $m^i = 1$
    $\lambda^i = \frac{1}{n(\mathcal{T})}$
  while i_epoch $< M$ **do**
    Optimize f on $\mathcal{D}_{train}$.
    Evaluate f on $\mathcal{D}_{val}$.
    while $i < n(\mathcal{T})$ **do**
      $Val^i \leftarrow$ validation performance of the i^{th} task.
      if Val^i is improved **then**
        $m^i \leftarrow 1$
      else
        $m^i \leftarrow m^i + 1$
    while $i < n(\mathcal{T})$ **do**
      $\lambda^i = max(1, \log_2(m^i))$
    while $i < n(\mathcal{T})$ **do**
      $\lambda^i = \frac{\lambda^i}{\sum_i \lambda^i}$

Experiments	T	AU	EXPR	VA		Total Emotion
				Valence	Arousal	
w/o ba.	1	0.6307	0.5620	0.3902	0.5344	2.1173
w/o ba.	5	0.6487	**0.5802**	0.4221	**0.5585**	2.2095
w/ ba.	1	0.6632	0.5541	0.4202	0.5192	2.1527
w/ ba.	5	**0.6808**	0.5779	**0.4423**	0.5455	**2.2465**

Table 5: Experiment results with the teacher models using visual modality only. "w/ ba." means we apply Algorithm 1 for multitask training. T is the number of models in an ensemble. Total emotion metric is the sum of all metrics of the three emotion tasks.

The main idea of this algorithm is to increase the weight of certain task if this task has not been improved on the validation set for a number of epochs. Once this task has been improved, the weight of the loss function for this task is set to its initial value.

We conducted ablation studies on the effect of Algorithm 1. The model we used is the EMENet-V trained on original dataset. From the experiment results in Table 5, we notice that the EXPR and arousal metrics are better without multitask balancing algorithm, but the total emotion metric is

Methods	AU1	AU2	AU4	AU6	AU7	AU10	AU12	AU15	AU23	AU24	AU25	AU26	Avg.
Tea	0.380	0.302	0.427	0.407	0.472	0.422	0.352	0.307	0.378	0.394	0.466	0.401	0.392
Stu1	0.326	0.264	0.366	**0.381**	**0.459**	**0.412**	0.355	**0.256**	0.289	0.256	**0.458**	0.359	**0.348**
Stu2	0.320	0.261	0.368	0.384	0.465	0.419	0.366	**0.256**	0.280	0.251	0.467	0.352	0.349
Stu3	**0.308**	**0.255**	**0.359**	0.388	0.469	0.427	0.383	0.262	**0.276**	**0.233**	0.474	**0.342**	**0.348**
TS [12]	0.380	0.301	0.415	0.406	0.472	0.421	**0.351**	0.306	0.377	0.381	0.465	0.400	0.390
MC [11]	0.356	0.274	0.419	0.411	0.473	0.416	**0.351**	0.311	0.345	0.384	0.463	0.382	0.382

Table 6: The NLL values for 12 action units, which are evaluated on the validation set of the Aff-wild set. We compare our single teacher models and single student models with other methods, *i.e.*, TS (temperature scaling [12]) and MC (Monte-Carol Dropout [11]). The model architecture used in this comparison is EMENet-VA.

Methods	EXPR NLL	Valence RMSE	Arousal RMSE
Tea	1.060	0.416	0.240
Stu1	**0.825**	0.397	0.233
Stu2	0.846	0.392	0.231
Stu3	0.858	**0.383**	**0.230**
TS [12]	0.955	-	-
MC [11]	0.994	0.416	0.237

Table 7: The NLL values for EXPR recognition and the RMSE values for valence and arousal prediction. Metrics are evaluated on the validation set. TS optimizes temperature for lower NLL on a held-out validation set, which is not beneficial for RMSE in regression tasks. Therefore, we only compare our models with TS for EXPR task.

better when using multitask balancing algorithm. We value the performance of each emotion tasks equally. The Algorithm 1 was used in all other experiments for better total emotion metric.

C. In-domain Uncertainty for EMENet-VA

We show the AU uncertainty performance using the EMENet-VA in Table 6. We also compare our single teacher models and single student models with Temerature Scaling (TS) and Monte-Carol (MC) Dropout. Similar to the results in Table 3 (EMENet-V), the models using our algorithm achieved the lowest NLL value, compared with TS and MC Dropout. The lowest AVg. NLL value is 0.348 for EMENet-VA, while for EMENet-V, the lowest average NLL value is 0.381. We find that when using audio features with visual features, the uncertainty (NLL) of facial actions can be improved by about 8.7%.

Table 7 shows the uncertainty performance for the EXPR task, valence and arousal detection. We evaluated NLL for classification tasks and RMSE for regression tasks. Comparing Table 7 with Table 4, we find that incorporating audio features with visual features, it improved the EXPR NLL by 5.2%. However, it failed to improve the valence RMSE, and barely had an influence on the arousal RMSE.