I The influences of hyper-parameter λ for our loss

As shown in Tab.5, the network is difficult to converge if λ is too small (concentrated loss takes a dominant role) and there will be a big performance degradation if λ is too large (the unimodal loss takes a dominant role). Within a long reasonable range, our method performs stably.

Table 5. The influences of hyper-parameter λ for our loss.

Threshold λ	1e-1	1e1	1e2	1e3	2e3	1e4
MORPH II	NaN	1.92	1.88	1.86	1.88	3.24
AFLW2000	NaN	4.21	4.26	4.13	4.11	6.12
BIWI	NaN	3.67	3.71	3.57	3.61	5.75

II Demonstration for softmax+mean & variance loss superior to unimodal+mean & variance loss

The Mean-Variance loss [22] can be formulated as

$$\begin{aligned} L_{m-v} &= L_s + \lambda_1 L_m + \lambda_2 L_v \qquad (21) \\ &= \frac{1}{N}\sum_{i=1}^{N} -logp_{i,y_i} + \frac{\lambda_1}{2}(\hat{y}_i - y_i)^2 + \lambda_2 v_i, \qquad (22) \end{aligned}$$

where L_s is the softmax loss.

We first demonstrate that it is hard for the network to directly optimize mean loss and variance loss without softmax loss. Based on Eq. 2 and derivation process in [22], the gradient of L_m w.r.t. $z_{i,j}$ can be computed as

$$\frac{\partial L_m}{\partial z_{i,j}} = \frac{\hat{y}_i - y_i}{N} p_{i,j}(j - \hat{y}_i). \qquad (23)$$

According to the Eq. 23 , as analyzed in [22], for an estimated distribution with mean value $\hat{y}_i$, if $\hat{y}_i < y_i$, the network will be updated to increase the probabilities of the classes $j(j > \hat{y}_i)$ via their negative gradients, and decrease the probability of those classes $j(j < \hat{y}_i)$ via their positive gradients. In this way, the mean value of the estimated distribution will be increased, and becomes closer to y_i.

The gradient of L_v w.r.t. $z_{i,j}$ can be computed as

$$\frac{\partial L_v}{\partial z_{i,j}} = \frac{1}{N} p_{i,j}((j - \hat{y}_i)^2 - v_i). \qquad (24)$$

The gradient in Eq. 24 has the following properties:

$$j \in (\hat{y}_i - \sqrt{v_i}, \hat{y}_i + \sqrt{v_i}), \frac{\partial L_v}{\partial z_{i,j}} < 0, \qquad (25)$$

and

$$j \in [1, \hat{y}_i - \sqrt{v_i}) \cup (\hat{y}_i + \sqrt{v_i}, C], \frac{\partial L_v}{\partial z_{i,j}} > 0. \qquad (26)$$

As analyzed in [22], Eq. 25 shows that, the network will be updated to increase the probabilities of the classes j close to $\hat{y}_i(j \in (\hat{y}_i - \sqrt{v_i}, \hat{y}_i + \sqrt{v_i}))$ via their negative gradients. On the contrary, Eq. 26 shows that the network will be updated to decrease the probabilities of the classes j far away from $\hat{y}_i(j \in [1, \hat{y}_i - \sqrt{v_i}) \cup (m_i + \sqrt{v_i}, C])$ via their positive gradients.

Base on analysis above, it can be observed that

$$\begin{aligned} &\text{if} \quad j \in (\hat{y}_i - \sqrt{v_i}, \hat{y}_i) \quad \text{and} \quad \hat{y}_i < y_i \\ &\text{then} \quad \frac{\partial L_m}{\partial z_{i,j}} > 0 \quad , \quad \frac{\partial L_v}{\partial z_{i,j}} < 0. \end{aligned} \qquad (27)$$

In the case of Eq. 27, let $|\frac{\partial L_v}{\partial z_{i,j}}| > |\frac{\partial L_m}{\partial z_{i,j}}|$ (λ_1 and λ_2 are omitted for demonstration)

$$\begin{aligned} &\Rightarrow v_i - (j-\hat{y}_i)^2 > (\hat{y}_i - y_i)(j-\hat{y}_i), \\ &\Rightarrow v_i > (j-y_i)(j-\hat{y}_i). \end{aligned} \tag{28}$$

According to the Eq. 28, when $v_i > (j-y_i)(j-\hat{y}_i)$, the absolute value of $\frac{\partial L_v}{\partial z_{i,j}}$ is larger than that of $\frac{\partial L_m}{\partial z_{i,j}}$. Consequently, the network will be updated to increase the probabilities of the classes $j(j \in (\hat{y}_i - \sqrt{v_i}, \hat{y}_i)$ which are far from the ground-truth y_i. That is to say, when large fluctuation appears at the early stage of training [22] which meets the such condition, the probabilities of the classes far from the ground-truth y_i will be increased and it is against principle I. It accounts for that it is hard to optimize the network with the mean & variance loss only. A typical example corresponding to this condition is given in Fig. 7.

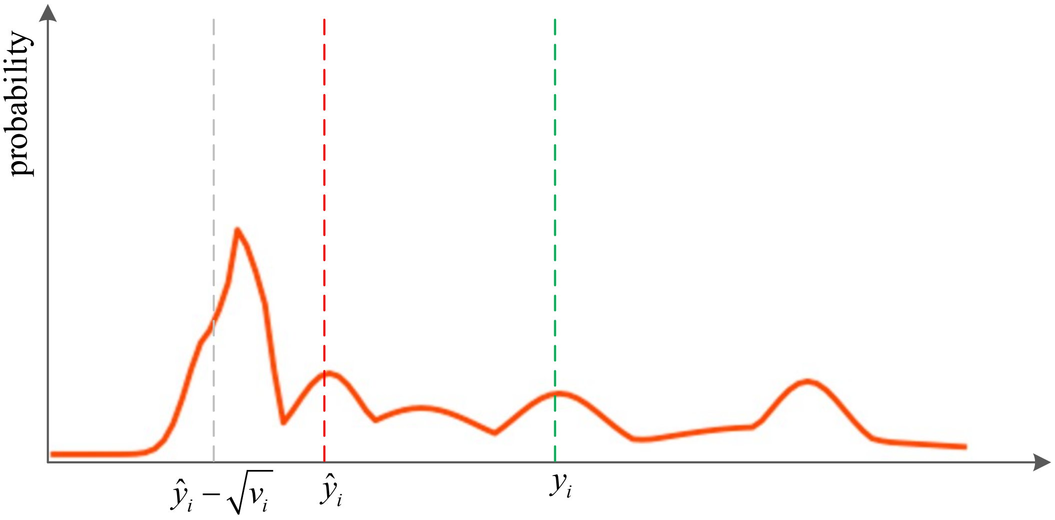

Figure 7. A typical example of distribution at the early stage of training.

When adding the softmax loss, as we all know, the gradient of L_s w.r.t. $z_{i,j}$ can be computed as

$$\frac{\partial L_s}{\partial z_{i,j}} = p_{i,j} - y_{i,j} \tag{29}$$

where $y_{i,j}$ is the indicator whether the instance i belongs to class j. If instance i belongs to class j, $y_{i,j} = 1$, otherwise, $y_{i,j} = 0$. According to the Eq. 29, it can be seen that the network will always be updated to increase the probability of the class y_i via their negative gradients. It accounts for that softmax loss can promote the network to converge with the mean & variance loss.

When adding the unimodal loss, the gradient of L_{uni} w.r.t. $z_{i,j}$ can be computed as

$$\frac{\partial L_{uni}}{\partial z_{i,j}} = \frac{\partial L_{uni}}{\partial p_{i,j}} \frac{\partial p_{i,j}}{\partial z_{i,j}} = \begin{cases} p_{i,j}(1-p_{i,j}), & (p_{i,j} - p_{i,j+1}) * \text{sign}[j - y_i] < 0 \\ 0, & (p_{i,j} - p_{i,j+1}) * \text{sign}[j - y_i] >= 0, \end{cases} \tag{30}$$

the gradient of L_{uni} w.r.t. $z_{i,j+1}$ can be computed as

$$\frac{\partial L_{uni}}{\partial z_{i,j+1}} = \frac{\partial L_{uni}}{\partial p_{i,j+1}} \frac{\partial p_{i,j+1}}{\partial z_{i,j+1}} = \begin{cases} -p_{i,j+1}(1-p_{i,j+1}), & (p_{i,j} - p_{i,j+1}) * \text{sign}[j - y_i] < 0 \\ 0, & (p_{i,j} - p_{i,j+1}) * \text{sign}[j - y_i] >= 0. \end{cases} \tag{31}$$

According to the Eq. 30 and 31, it can be seen that our unimodal loss aims at correcting the ordinal relationship when two neighboring probabilities are ranked by mistake instead of directly maxmizing the probability of the class y_i like softmax loss. So, compared with the combination of softmax and mean loss & variance loss, the combination of unimodal and mean loss & variance loss gets the poorer performance.

III Ablation study about single unimodal loss and single concentrated loss.

Without the unimodal loss, the network is difficult to converge. The MeanVariance loss encounters the similar problem without softmax loss as mentioned by the paper [22] in Sec. 3.2. Therefore, in ablation as shown in Tab. 6 of main script, only our concentrated loss along with softmax loss or unimodal loss and MeanVariance loss along with softmax loss or unimodal loss are campared. Without the concentrated loss, although the network can converge. However, with only the unimodal loss, different distributions (shown as the three samples in Fig. 8) may have the same loss value, but very distinct shapes, which may make the network converge to a bad local minimum. Its age estimation error is larger than 5 on MORPH II, which verifies this.

Table 6. The results for the single unimodal loss and single concentrated loss.

Combinations		Benchmarks		
Auxiliary	Primary	MORPH II	AFLW2000	BIWI
Softmax	Concentrated	1.92	4.25	3.61
Unimodal	Concentrated	1.86	4.13	3.57
Softmax	Mean & Variance	2.01	4.36	4.01
Unimodal	Mean & Variance	3.30	4.53	4.39
Unimodal	-	≥ 5	≥ 7	≥ 7
-	Concentrated	NaN	NaN	NaN

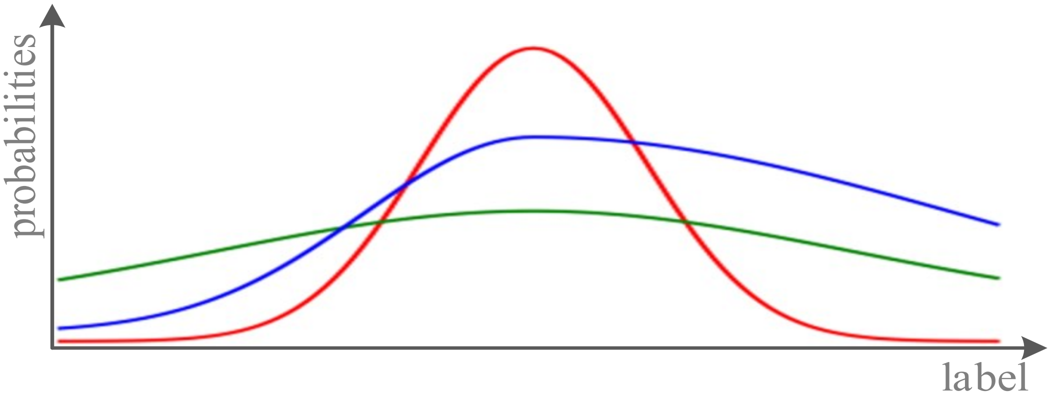

Figure 8. Typical distribution examples.