A. Supplementary Theoretical Results

Lemma 1 *Suppose $S^f(\boldsymbol{x})$ fulfills the Tsybakov condition on instance-label dependence for constants C_1, $\lambda_1 > 0$, and $t_0 \in (0, m]$. Let $\kappa^f(h, \boldsymbol{x}, \bar{\boldsymbol{y}}) := \hat{\bar{S}}^f_{\bar{\boldsymbol{y}}}(\boldsymbol{x}) / \hat{\bar{S}}^f_{\boldsymbol{y}^*}(\boldsymbol{x})$. We define $\epsilon := \max_{\boldsymbol{x},\boldsymbol{z}} \left[|\hat{\bar{S}}^f_{\boldsymbol{z}}(\boldsymbol{x}) - \bar{S}^f_{\boldsymbol{z}}(\boldsymbol{x})|, |\hat{\bar{S}}^l_{\boldsymbol{z}}(\boldsymbol{x}) - \bar{S}^l_{\boldsymbol{z}}(\boldsymbol{x})|, |\bar{S}^l_{\boldsymbol{z}}(\boldsymbol{x}) - S^l_{\boldsymbol{z}}(\boldsymbol{x})| \right]$ and $\tau := \min_i T_{ii}$. We analyze two cases:*

(1) If $\bar{\boldsymbol{y}}$ is corrected by $\kappa^f(h, \boldsymbol{x}, \bar{\boldsymbol{y}})$ with the threshold $\hat{\delta}$, let $\delta_1 = \min \left[\frac{\tau S^f_{\boldsymbol{b}_{\boldsymbol{x}}} + \sum_{l_j \in \bar{\boldsymbol{y}}} \sum_{i \neq j} T_{ij} \mathbb{P}(l_i|\boldsymbol{x})}{\hat{\bar{S}}^f_{\boldsymbol{y}^*}} \right]$ and $\rho_1 := |\hat{\delta} - \delta_1|$. Assume that $\epsilon \leq t_0 \tau - \rho_1 m$. Then, $\mathbb{P}[\bar{\boldsymbol{y}}_{new} = h^*(\boldsymbol{x}), \bar{\boldsymbol{y}}\ is\ flipped]$ is at least $1 - C_1[\max(\epsilon, \rho_1)]^{\lambda_1} - \mathbb{P}[\boldsymbol{a}_{\boldsymbol{x}} \neq \{\boldsymbol{y}^*, \bar{\boldsymbol{y}}\}]$.

(2) If $\bar{\boldsymbol{y}}$ is not corrected by $\kappa^f(h, \boldsymbol{x}, \bar{\boldsymbol{y}})$ with the threshold $\hat{\delta}$, let $\delta_2 = \max \left[\frac{\hat{\bar{S}}^f_{\bar{\boldsymbol{y}}}(\boldsymbol{x})}{\tau S^f_{\boldsymbol{b}_{\boldsymbol{x}}}(\boldsymbol{x}) + \sum_{l_j \in \boldsymbol{y}^*} \sum_{i \neq j} T_{ij} \mathbb{P}(l_i|\boldsymbol{x})} \right]$ and $\rho_2 := |\hat{\delta} - \delta_2|$. Assume that $\epsilon \leq \frac{t_0 \delta_2^2 \tau - \rho_2 m - \rho_2^2 m}{\delta_2^2}$. Then, $\mathbb{P}[\bar{\boldsymbol{y}}_{new} = h^*(\boldsymbol{x}), \bar{\boldsymbol{y}}\ is\ accepted]$ is at least $1 - C_1[\max(\epsilon, \rho_2)]^{\lambda_1} - \mathbb{P}[\boldsymbol{a}_{\boldsymbol{x}} \neq \{\boldsymbol{y}^*, \bar{\boldsymbol{y}}\}]$.

Lemma 1 claims that, even though with noisy multiple labels, there is a guaranteed success rate to make proper label corrections by instance-label dependency.

B. Proofs of Theoretical Results

B.1. Proof of Theorem 1

Proof 1 *For the case (1),*

$$\mathbb{P}[\bar{\boldsymbol{y}}_{new} \neq h^*(\boldsymbol{x}), \bar{\boldsymbol{y}}\ is\ flipped] = \mathbb{P}\left[\bar{\boldsymbol{y}}_{new} \neq h^*(\boldsymbol{x}), \frac{\hat{\bar{S}}_{\bar{\boldsymbol{y}}}(\boldsymbol{x})}{\hat{\bar{S}}_{\boldsymbol{y}^*}(\boldsymbol{x})} < \hat{\delta} \right] \tag{1}$$

$$\leq \mathbb{P}\left[h^*(\boldsymbol{x}) = \bar{\boldsymbol{y}}, \frac{\hat{\bar{S}}_{\bar{\boldsymbol{y}}}(\boldsymbol{x})}{\hat{\bar{S}}_{\boldsymbol{y}^*}(\boldsymbol{x})} < \hat{\delta} \right] + \mathbb{P}[\boldsymbol{a}_{\boldsymbol{x}} \neq \{\boldsymbol{y}^*, \bar{\boldsymbol{y}}\}]. \tag{2}$$

For the first term,

$$\mathbb{P}\left[h^*(\boldsymbol{x}) = \bar{\boldsymbol{y}}, \frac{\hat{\bar{S}}_{\bar{\boldsymbol{y}}}(\boldsymbol{x})}{\hat{\bar{S}}_{\boldsymbol{y}^*}(\boldsymbol{x})} < \hat{\delta} \right] = \mathbb{P}\left[S^f_{\bar{\boldsymbol{y}}}(\boldsymbol{x}) > S^f_{\boldsymbol{b}_{\boldsymbol{x}}}(\boldsymbol{x}), \hat{\bar{S}}_{\bar{\boldsymbol{y}}} < \hat{\delta} \hat{\bar{S}}_{\boldsymbol{y}^*}(\boldsymbol{x}) \right] \tag{3}$$

$$= \mathbb{P}\left[S_{\bar{\boldsymbol{y}}}(\boldsymbol{x}) - S^l_{\bar{\boldsymbol{y}}}(\boldsymbol{x}) > S_{\boldsymbol{b}_{\boldsymbol{x}}}(\boldsymbol{x}) - S^l_{\boldsymbol{b}_{\boldsymbol{x}}}(\boldsymbol{x}), \hat{\bar{S}}^f_{\bar{\boldsymbol{y}}}(\boldsymbol{x}) + \hat{\bar{S}}^l_{\bar{\boldsymbol{y}}}(\boldsymbol{x}) < \hat{\delta} \hat{\bar{S}}_{\boldsymbol{y}^*}(\boldsymbol{x}) \right] \tag{4}$$

$$\leq \mathbb{P}\left[S_{\bar{\boldsymbol{y}}}(\boldsymbol{x}) \geq S_{\boldsymbol{b}_{\boldsymbol{x}}}(\boldsymbol{x}), \bar{S}^f_{\bar{\boldsymbol{y}}}(\boldsymbol{x}) + \hat{\bar{S}}^l_{\bar{\boldsymbol{y}}}(\boldsymbol{x}) < \hat{\delta} \hat{\bar{S}}_{\boldsymbol{y}^*}(\boldsymbol{x}) + \epsilon \right] \tag{5}$$

$$\leq \mathbb{P}\left[S_{\bar{\boldsymbol{y}}}(\boldsymbol{x}) \geq S_{\boldsymbol{b}_{\boldsymbol{x}}}(\boldsymbol{x}), \bar{S}^f_{\bar{\boldsymbol{y}}}(\boldsymbol{x}) + S^l_{\bar{\boldsymbol{y}}}(\boldsymbol{x}) < \hat{\delta} \hat{\bar{S}}_{\boldsymbol{y}^*}(\boldsymbol{x}) + 3\epsilon \right] \tag{6}$$

$$\leq \mathbb{P}\left[S_{\bar{\boldsymbol{y}}}(\boldsymbol{x}) \geq S_{\boldsymbol{b}_{\boldsymbol{x}}}(\boldsymbol{x}), S_{\bar{\boldsymbol{y}}}(\boldsymbol{x}) < \frac{\delta \hat{\bar{S}}_{\boldsymbol{y}^*}(\boldsymbol{x}) - \sum_{l_j \in \bar{\boldsymbol{y}}} \sum_{i \neq j} T_{ij} \mathbb{P}(l_i|\boldsymbol{x})}{\tau} + \frac{3\epsilon + m\rho_1}{\tau} \right]. \tag{7}$$

If $\delta = \min \left[\frac{\tau S_{\boldsymbol{b}_{\boldsymbol{x}}}(\boldsymbol{x}) + \sum_{l_j \in \bar{\boldsymbol{y}}} \sum_{i \neq j} T_{ij} \mathbb{P}(l_i|\boldsymbol{x})}{\hat{\bar{S}}_{\boldsymbol{y}^*}(\boldsymbol{x})} \right]$, we have $\mathbb{P}\left[h^*(\boldsymbol{x}) = \bar{\boldsymbol{y}}, \frac{\hat{\bar{S}}_{\bar{\boldsymbol{y}}}(\boldsymbol{x})}{\hat{\bar{S}}_{\boldsymbol{y}^*}(\boldsymbol{x})} < \hat{\delta} \right] \leq C_2[O(\max(\epsilon, \rho_1))]^{\lambda_2}$. Therefore, for the case (1),

$$\mathbb{P}[\bar{\boldsymbol{y}}_{new} = h^*(\boldsymbol{x}), \bar{\boldsymbol{y}}\ is\ flipped] \geq 1 - C_2[O(\max(\epsilon, \rho_1))]^{\lambda_2} - \mathbb{P}[\boldsymbol{a}_{\boldsymbol{x}} \neq \{\boldsymbol{y}^*, \bar{\boldsymbol{y}}\}]. \tag{8}$$

For the case (2),

$$\mathbb{P}[\bar{\boldsymbol{y}}_{new} \neq h^*(\boldsymbol{x}), \bar{\boldsymbol{y}}\ is\ accepted] = \mathbb{P}\left[\bar{\boldsymbol{y}}_{new} \neq h^*(\boldsymbol{x}), \frac{\hat{\bar{S}}_{\bar{\boldsymbol{y}}}(\boldsymbol{x})}{\hat{\bar{S}}_{\boldsymbol{y}^*}(\boldsymbol{x})} \geq \hat{\delta} \right] \tag{9}$$

$$\leq \mathbb{P}\left[S^f_{\boldsymbol{y}^*}(\boldsymbol{x}) > S^f_{\boldsymbol{b}_{\boldsymbol{x}}}(\boldsymbol{x}), \hat{\bar{S}}_{\boldsymbol{y}^*} \leq \hat{\bar{S}}_{\bar{\boldsymbol{y}}}(\boldsymbol{x}) / \hat{\delta} \right] + \mathbb{P}[\boldsymbol{a}_{\boldsymbol{x}} \neq \{\boldsymbol{y}^*, \bar{\boldsymbol{y}}\}]. \tag{10}$$

For the first term,

$$\mathbb{P}\left[S^f_{\boldsymbol{y}^*}(\boldsymbol{x}) > S^f_{\boldsymbol{b}_{\boldsymbol{x}}}(\boldsymbol{x}), \hat{\bar{S}}_{\boldsymbol{y}^*} \leq \hat{\bar{S}}_{\bar{\boldsymbol{y}}}(\boldsymbol{x})/\hat{\delta}\right] \tag{11}$$

$$\leq \mathbb{P}\left[S_{\boldsymbol{y}^*}(\boldsymbol{x}) - S^l_{\boldsymbol{y}^*}(\boldsymbol{x}) > S_{\boldsymbol{b}_{\boldsymbol{x}}}(\boldsymbol{x}) - S^l_{\boldsymbol{b}_{\boldsymbol{x}}}(\boldsymbol{x}), \bar{S}^f_{\boldsymbol{y}^*}(\boldsymbol{x}) + \hat{\bar{S}}^l_{\boldsymbol{y}^*}(\boldsymbol{x}) \leq \hat{\bar{S}}_{\bar{\boldsymbol{y}}}(\boldsymbol{x})/\hat{\delta} + \epsilon\right] \tag{12}$$

$$\leq \mathbb{P}\left[S_{\boldsymbol{y}^*}(\boldsymbol{x}) \geq S_{\boldsymbol{b}_{\boldsymbol{x}}}(\boldsymbol{x}), S_{\boldsymbol{y}^*}(\boldsymbol{x}) \leq \frac{\hat{\bar{S}}_{\bar{\boldsymbol{y}}}(\boldsymbol{x})/\hat{\delta} - \sum_{l_j \in \boldsymbol{y}^*}\sum_{i \neq j} T_{ij}\mathbb{P}(l_i|\boldsymbol{x})}{\tau} + \frac{3\epsilon}{\tau}\right] \tag{13}$$

$$\leq \mathbb{P}\left[S_{\boldsymbol{y}^*}(\boldsymbol{x}) \geq S_{\boldsymbol{b}_{\boldsymbol{x}}}(\boldsymbol{x}), S_{\boldsymbol{y}^*}(\boldsymbol{x}) \leq \frac{\hat{\bar{S}}_{\bar{\boldsymbol{y}}}(\boldsymbol{x})/\delta - \sum_{l_j \in \boldsymbol{y}^*}\sum_{i \neq j} T_{ij}\mathbb{P}(l_i|\boldsymbol{x})}{\tau} + \frac{3\epsilon}{\tau} + \frac{\frac{\rho_2 \hat{\bar{S}}_{\bar{\boldsymbol{y}}}(\boldsymbol{x})}{\delta(\delta-\rho_2)}}{\tau}\right]. \tag{14}$$

If $\delta = \max\left[\frac{\hat{\bar{S}}_{\bar{\boldsymbol{y}}}(\boldsymbol{x})}{\tau S_{\boldsymbol{b}_{\boldsymbol{x}}}(\boldsymbol{x}) + \sum_{l_j \in \boldsymbol{y}^*}\sum_{i \neq j} T_{ij}\mathbb{P}(l_i|\boldsymbol{x})}\right]$, *we have*

$$\mathbb{P}\left[S^f_{\boldsymbol{y}^*}(\boldsymbol{x}) > S^f_{\boldsymbol{b}_{\boldsymbol{x}}}(\boldsymbol{x}), \hat{\bar{S}}_{\boldsymbol{y}^*} \leq \hat{\bar{S}}_{\bar{\boldsymbol{y}}}(\boldsymbol{x})/\hat{\delta}\right] \tag{15}$$

$$\leq \mathbb{P}\left[S_{\boldsymbol{b}_{\boldsymbol{x}}}(\boldsymbol{x}) \leq S_{\boldsymbol{y}^*}(\boldsymbol{x}) \leq S_{\boldsymbol{b}_{\boldsymbol{x}}} + \frac{3\epsilon}{\tau} + \frac{\rho_2 m}{\delta^2 \tau} + \frac{\rho_2^2 m}{\delta^2 \tau}\right] \leq C_2[O(\max(\epsilon, \rho_2))]^{\lambda_2}. \tag{16}$$

Therefore, we have

$$\mathbb{P}[\bar{\boldsymbol{y}}_{new} = h^*(\boldsymbol{x}), \bar{\boldsymbol{y}}\ is\ accepted] \geq 1 - C_2[O(\max(\epsilon, \rho_2))]^{\lambda_2} - \mathbb{P}[\boldsymbol{a}_{\boldsymbol{x}} \neq \{\boldsymbol{y}^*, \bar{\boldsymbol{y}}\}]. \tag{17}$$

B.2. Proofs of Lemma 1 and Corollary 1

We first prove Lemma 1. Lemma 1 uses the similar proof skill of Theorem 3 of [65]. We extend it into multi-label classification.

Proof 2 *For the case (1),*

$$\mathbb{P}[\bar{\boldsymbol{y}}_{new} \neq h^*(\boldsymbol{x}), \bar{\boldsymbol{y}}\ is\ flipped] = \mathbb{P}\left[\bar{\boldsymbol{y}}_{new} \neq h^*(\boldsymbol{x}), \frac{\hat{\bar{S}}^f_{\bar{\boldsymbol{y}}}(\boldsymbol{x})}{\hat{\bar{S}}^f_{\boldsymbol{y}^*}(\boldsymbol{x})} < \hat{\delta}\right] \tag{18}$$

$$= \mathbb{P}\left[\bar{\boldsymbol{y}}_{new} = \boldsymbol{y}^* \neq h^*(\boldsymbol{x}) = \bar{\boldsymbol{y}}, \frac{\hat{\bar{S}}^f_{\bar{\boldsymbol{y}}}(\boldsymbol{x})}{\hat{\bar{S}}^f_{\boldsymbol{y}^*}(\boldsymbol{x})} < \hat{\delta}\right] + \mathbb{P}\left[\bar{\boldsymbol{y}}_{new} = \boldsymbol{y}^* \neq h^*(\boldsymbol{x}) = \boldsymbol{a}_{\boldsymbol{x}} \neq \bar{\boldsymbol{y}}, \frac{\hat{\bar{S}}^f_{\bar{\boldsymbol{y}}}(\boldsymbol{x})}{\hat{\bar{S}}^f_{\boldsymbol{y}^*}(\boldsymbol{x})} < \hat{\delta}\right] \tag{19}$$

$$\leq \mathbb{P}\left[h^*(\boldsymbol{x}) = \bar{\boldsymbol{y}}, \frac{\hat{\bar{S}}^f_{\bar{\boldsymbol{y}}}(\boldsymbol{x})}{\hat{\bar{S}}^f_{\boldsymbol{y}^*}(\boldsymbol{x})} < \hat{\delta}\right] + \mathbb{P}\left[\bar{\boldsymbol{y}}_{new} = \boldsymbol{y}^* \neq h^*(\boldsymbol{x}) = \boldsymbol{a}_{\boldsymbol{x}} \neq \bar{\boldsymbol{y}}, \frac{\hat{\bar{S}}^f_{\bar{\boldsymbol{y}}}(\boldsymbol{x})}{\hat{\bar{S}}^f_{\boldsymbol{y}^*}(\boldsymbol{x})} < \hat{\delta}\right] \tag{20}$$

$$\leq \mathbb{P}\left[h^*(\boldsymbol{x}) = \bar{\boldsymbol{y}}, \frac{\hat{\bar{S}}^f_{\bar{\boldsymbol{y}}}(\boldsymbol{x})}{\hat{\bar{S}}^f_{\boldsymbol{y}^*}(\boldsymbol{x})} < \hat{\delta}\right] + \mathbb{P}[\boldsymbol{a}_{\boldsymbol{x}} \neq \{\boldsymbol{y}^*, \bar{\boldsymbol{y}}\}]. \tag{21}$$

For the first term, we have

$$\mathbb{P}\left[h^*(\boldsymbol{x}) = \bar{\boldsymbol{y}}, \frac{\hat{\bar{S}}^f_{\bar{\boldsymbol{y}}}(\boldsymbol{x})}{\hat{\bar{S}}^f_{\boldsymbol{y}^*}(\boldsymbol{x})} < \hat{\delta}\right] = \mathbb{P}\left[S^f_{\bar{\boldsymbol{y}}}(\boldsymbol{x}) > S^f_{\boldsymbol{b}_{\boldsymbol{x}}}(\boldsymbol{x}), \hat{\bar{S}}^f_{\bar{\boldsymbol{y}}}(\boldsymbol{x}) < \hat{\delta}\hat{\bar{S}}^f_{\boldsymbol{y}^*}(\boldsymbol{x})\right] \tag{22}$$

$$\leq \mathbb{P}\left[S^f_{\bar{\boldsymbol{y}}}(\boldsymbol{x}) > S^f_{\boldsymbol{b}_{\boldsymbol{x}}}(\boldsymbol{x}), \bar{S}^f_{\bar{\boldsymbol{y}}}(\boldsymbol{x}) < \hat{\delta}\hat{\bar{S}}^f_{\boldsymbol{y}^*}(\boldsymbol{x}) + \epsilon\right] \tag{23}$$

$$= \mathbb{P}\left[S^f_{\bar{\boldsymbol{y}}}(\boldsymbol{x}) > S^f_{\boldsymbol{b}_{\boldsymbol{x}}}(\boldsymbol{x}), \sum_{l_j \in \bar{\boldsymbol{y}}} T_{jj}\mathbb{P}(l_j|\boldsymbol{x}) + \sum_{l_j \in \bar{\boldsymbol{y}}}\sum_{i \neq j} T_{ij}\mathbb{P}(l_i|\boldsymbol{x}) < \hat{\delta}\hat{\bar{S}}^f_{\boldsymbol{y}^*}(\boldsymbol{x}) + \epsilon\right] \tag{24}$$

$$\leq \mathbb{P}\left[S^f_{\bar{\boldsymbol{y}}}(\boldsymbol{x}) > S^f_{\boldsymbol{b}_{\boldsymbol{x}}}(\boldsymbol{x}), S^f_{\bar{\boldsymbol{y}}}(\boldsymbol{x}) < \frac{\delta\hat{\bar{S}}^f_{\boldsymbol{y}^*}(\boldsymbol{x}) - \sum_{l_j \in \bar{\boldsymbol{y}}}\sum_{i \neq j} T_{ij}\mathbb{P}(l_i|\boldsymbol{x})}{\tau} + \frac{\epsilon + \rho_1}{\tau}\right]. \tag{25}$$

If $\delta = \min\left[\frac{\tau S^f_{\boldsymbol{b_x}} + \sum_{l_j \in \bar{\boldsymbol{y}}} \sum_{i \neq j} T_{ij} \mathbb{P}(l_i|\boldsymbol{x})}{\hat{\tilde{S}}^f_{\boldsymbol{y}^*}}\right]$, *we have*

$$\mathbb{P}\left[h^*(\boldsymbol{x}) = \bar{\boldsymbol{y}}, \frac{\hat{\tilde{S}}^f_{\bar{\boldsymbol{y}}}(\boldsymbol{x})}{\hat{\tilde{S}}^f_{\boldsymbol{y}^*}(\boldsymbol{x})} < \delta\right] \leq \mathbb{P}\left[S^f_{\boldsymbol{b_x}} < S^f_{\bar{\boldsymbol{y}}}(\boldsymbol{x}) < S^f_{\boldsymbol{b_x}} + \frac{\epsilon + \rho_1}{\tau}\right] = C_1[O(\max(\epsilon, \rho_1))]^{\lambda_1}. \tag{26}$$

Therefore,

$$\mathbb{P}[\bar{\boldsymbol{y}}_{new} = h^*(\boldsymbol{x}), \bar{\boldsymbol{y}}\ is\ flipped] \geq 1 - C_1[\max(\epsilon, \rho_1)]^{\lambda_1} - \mathbb{P}[\boldsymbol{a_x} \neq \{\boldsymbol{y}^*, \bar{\boldsymbol{y}}\}]. \tag{27}$$

The case (2) shares the similar proof with the case (1). Specifically,

$$\mathbb{P}[\bar{\boldsymbol{y}}_{new} \neq h^*(\boldsymbol{x}), \bar{\boldsymbol{y}}\ is\ accepted] = \mathbb{P}\left[\bar{\boldsymbol{y}}_{new} \neq h^*(\boldsymbol{x}), \frac{\hat{\tilde{S}}^f_{\bar{\boldsymbol{y}}}(\boldsymbol{x})}{\hat{\tilde{S}}^f_{\boldsymbol{y}^*}(\boldsymbol{x})} \geq \hat{\delta}\right] \tag{28}$$

$$\leq \mathbb{P}\left[S^f_{\boldsymbol{y}^*}(\boldsymbol{x}) > S^f_{\boldsymbol{b_x}}(\boldsymbol{x}), \hat{\tilde{S}}^f_{\boldsymbol{y}^*} \leq \hat{\tilde{S}}^f_{\bar{\boldsymbol{y}}}(\boldsymbol{x})/\hat{\delta}\right] + \mathbb{P}[\boldsymbol{a_x} \neq \{\boldsymbol{y}^*, \bar{\boldsymbol{y}}\}] \tag{29}$$

For the first term, we have

$$\mathbb{P}\left[S^f_{\boldsymbol{y}^*}(\boldsymbol{x}) > S^f_{\boldsymbol{b_x}}(\boldsymbol{x}), \hat{\tilde{S}}^f_{\boldsymbol{y}^*} \leq \hat{\tilde{S}}^f_{\bar{\boldsymbol{y}}}(\boldsymbol{x})/\hat{\delta}\right] \tag{30}$$

$$\leq \mathbb{P}\left[S^f_{\boldsymbol{b_x}}(\boldsymbol{x}) < S^f_{\boldsymbol{y}^*}(\boldsymbol{x}) \leq \frac{\hat{\tilde{S}}^f_{\bar{\boldsymbol{y}}}(\boldsymbol{x})/\hat{\delta} - \sum_{l_j \in \boldsymbol{y}^*} \sum_{i \neq j} T_{ij} \mathbb{P}(l_i|\boldsymbol{x})}{\tau} + \frac{\epsilon}{\tau}\right] \tag{31}$$

$$\leq \mathbb{P}\left[S^f_{\boldsymbol{b_x}}(\boldsymbol{x}) < S^f_{\boldsymbol{y}^*}(\boldsymbol{x}) \leq \frac{\hat{\tilde{S}}^f_{\bar{\boldsymbol{y}}}(\boldsymbol{x})/(\delta - \rho_2) - \sum_{l_j \in \boldsymbol{y}^*} \sum_{i \neq j} T_{ij} \mathbb{P}(l_i|\boldsymbol{x})}{\tau} + \frac{\epsilon}{\tau}\right] \tag{32}$$

$$= \mathbb{P}\left[0 < S^f_{\boldsymbol{y}^*}(\boldsymbol{x}) - S^f_{\boldsymbol{b_x}}(\boldsymbol{x}) < \frac{\hat{\tilde{S}}^f_{\bar{\boldsymbol{y}}}(\boldsymbol{x})/\delta - \sum_{l_j \in \boldsymbol{y}^*} \sum_{i \neq j} T_{ij} \mathbb{P}(l_i|\boldsymbol{x})}{\tau} - S^f_{\boldsymbol{b_x}}(\boldsymbol{x}) + \frac{\epsilon}{\tau} + \frac{\frac{\rho_2 \hat{\tilde{S}}^f_{\bar{\boldsymbol{y}}}(\boldsymbol{x})}{\delta(\delta - \rho_2)}}{\tau}\right] \tag{33}$$

If $\delta = \max\left[\frac{\hat{\tilde{S}}^f_{\bar{\boldsymbol{y}}}(\boldsymbol{x})}{\tau S^f_{\boldsymbol{b_x}}(\boldsymbol{x}) + \sum_{l_j \in \boldsymbol{y}^*} \sum_{i \neq j} T_{ij} \mathbb{P}(l_i|\boldsymbol{x})}\right]$, *we have*

$$\mathbb{P}\left[S^f_{\boldsymbol{y}^*}(\boldsymbol{x}) > S^f_{\boldsymbol{b_x}}(\boldsymbol{x}), \hat{\tilde{S}}^f_{\boldsymbol{y}^*} \leq \hat{\tilde{S}}^f_{\bar{\boldsymbol{y}}}(\boldsymbol{x})/\hat{\delta}\right] \tag{34}$$

$$\leq \mathbb{P}\left[0 < S^f_{\boldsymbol{y}^*}(\boldsymbol{x}) - S^f_{\boldsymbol{b_x}}(\boldsymbol{x}) \leq \frac{\epsilon}{\tau} + \frac{\frac{\rho_2 \hat{\tilde{S}}^f_{\bar{\boldsymbol{y}}}(\boldsymbol{x})}{\delta(\delta - \rho_2)}}{\tau}\right] \tag{35}$$

$$\leq \mathbb{P}\left[0 \leq S^f_{\boldsymbol{y}^*}(\boldsymbol{x}) - S^f_{\boldsymbol{b_x}}(\boldsymbol{x}) \leq \frac{\epsilon}{\tau} + \frac{\rho_2 \hat{\tilde{S}}^f_{\bar{\boldsymbol{y}}}(\boldsymbol{x})}{\delta^2 \tau} + \frac{\rho_2 O(\rho_2) \hat{\tilde{S}}^f_{\bar{\boldsymbol{y}}}(\boldsymbol{x})}{\delta^2 \tau}\right] \tag{36}$$

$$\leq \mathbb{P}\left[0 \leq S^f_{\boldsymbol{y}^*}(\boldsymbol{x}) - S^f_{\boldsymbol{b_x}}(\boldsymbol{x}) \leq \frac{\epsilon}{\tau} + \frac{\rho_2 m}{\delta^2 \tau} + \frac{\rho_2^2 m}{\delta^2 \tau}\right]. \tag{37}$$

Here, since $\epsilon \leq \frac{t_0 \delta^2 \tau - \rho_2 m - \rho_2^2 m}{\delta^2}$, *we have* $\frac{\epsilon}{\tau} + \frac{\rho_2 m}{\delta^2 \tau} + \frac{\rho_2^2 m}{\delta^2 \tau} \leq t_0$. *Therefore,*

$$\mathbb{P}[\bar{\boldsymbol{y}}_{new} = h^*(\boldsymbol{x}), \bar{\boldsymbol{y}}\ is\ accepted] \geq 1 - C_1[O(\max(\epsilon, \rho_2))] - \mathbb{P}[\boldsymbol{a_x} \neq \{\boldsymbol{y}^*, \bar{\boldsymbol{y}}\}]. \tag{38}$$

The proof of Lemma 1 is completed. Combining Lemma 1 and Theorem 1, Corollary 1 can be achieved.

C. Related Literature

C.1. Procedure of ADDGCN

ADDGCN is the preparation technology of our HLC. We detail ADDGCN [57] as follows.

SAM. Given an example $(\boldsymbol{x}, \boldsymbol{y})$, we feed $\boldsymbol{x}$ into a deep network and obtain its corresponding feature map $\boldsymbol{x}'$. SAM first calculates label-specific activation maps $\mathbf{M} = [\boldsymbol{m}_1, \ldots, \boldsymbol{m}_q]$ by using class-activation-mapping [66]. Then, $\mathbf{M}$ is used to convert the feature map $\boldsymbol{x}'$ into the content-aware class-label representations $\mathbf{C} = [\boldsymbol{c}_1, \ldots, \boldsymbol{c}_q]$. Let $[q] = \{1, \ldots, q\}$. Mathematically, for $k \in [q]$, we have $\boldsymbol{c}_k = \boldsymbol{m}_k^\top \boldsymbol{x}'$. That is, $\boldsymbol{c}_k$ selectively aggregate features related to its specific class label k.

GCNM. With the content-aware class-label representations $\mathbf{C}$ achieved by SAM, GCNM is introduced to adaptively transform their coherent correlation for multi-label classification. Specifically, GCNM consists of two parts: a static GCN and a dynamic GCN. The representations $\mathbf{C}$ are taken by GCNM as input node features and sequentially fed into the static GCN and dynamic GCN.

The single layer of the static GCN is defined as $\mathbf{H} = \text{LReLU}(\mathbf{A}^s\mathbf{C}\mathbf{W}^s)$, where $\mathbf{A}^s$ denotes the correlation matrix shared for all instances, $\mathbf{W}^s$ denotes state-update weights, and $\text{LReLU}(\cdot)$ denotes the LeakyReLU activation function [53]. Besides, $\mathbf{A}^s$ and $\mathbf{W}^s$ are randomly initialized and learned by gradient decent during training. The dynamic GCN transforms $\mathbf{H}$. Its correlation matrix $\mathbf{A}^d$ is constructed dynamically dependent on input features $\mathbf{H}$. Namely, each examples have different $\mathbf{A}^d$. Formally, the output of the dynamic GCN is formulated as $\mathbf{Z} = \text{LReLU}(\mathbf{A}^d\mathbf{H}\mathbf{W}^d)$, where $\mathbf{W}^d$ are state-update weights. Later, we use $\mathbf{A}^d(\boldsymbol{x})$ to denote the correlation matrix of $\boldsymbol{x}$, where $\mathbf{A}^d(\boldsymbol{x})_{jk} = \hat{\mathbb{P}}(l_k|l_j, \boldsymbol{x})$ for any $j, k \in [q]$.

Classification and Loss. The label-specific activation map $\mathbf{M} = [\boldsymbol{m}_1, \ldots, \boldsymbol{m}_q]$ and final category representation $\mathbf{Z} = [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_q]$ are employed simultaneously for multi-label classification. Specifically, we use global spatial pooling on $\mathbf{M}$ to obtain a score vector $\mathbf{s}^m = [s_1^m, \ldots, s_q^m]$. Besides, each category representation $\mathbf{Z}$ is put into a binary classifier to obtain another score vector $\mathbf{s}^z = [s_1^z, \ldots, s_q^z]$. We simply average two score vectors to predict more reliable results. The aggregated score vector is denoted as $\mathbf{s} = [s_1, \ldots, s_q] = [(s_1^m + s_1^z)/2, \ldots, (s_q^m + s_q^z)/2]$. The Sigmoid activation function $\sigma(\cdot)$ is then used on $\mathbf{s}$ for probabilistic interpretation. That is to say, $\sigma(\mathbf{s}) = [\sigma(s_1), \ldots, \sigma(s_q)] = [\hat{\mathbb{P}}(l_1|\boldsymbol{x}), \ldots, \hat{\mathbb{P}}(l_q|\boldsymbol{x})]$. The binary cross-entropy loss is exploited for the updates of all weights, *i.e.*, $\mathcal{L} = \sum_{l_i \in \boldsymbol{y}} \log(\sigma(s_i))$.

Given a multi-label example $(\boldsymbol{x}, \bar{\boldsymbol{y}})$, for the two dependences $\hat{S}^f$ and $\hat{S}^l$, based on $\sigma(\mathbf{s})$ and $\mathbf{A}^d(\boldsymbol{x})$ achieved by learning with multiple *noisy* labels, they can be estimated as

$$\hat{\tilde{S}}_{\boldsymbol{z}}^f(\boldsymbol{x}) = \sum_{\{\bar{\boldsymbol{Y}}=\boldsymbol{z}, l_i \in \boldsymbol{z}\}} \hat{\mathbb{P}}(l_i|\boldsymbol{x}) \quad \text{and} \quad \hat{\tilde{S}}_{\boldsymbol{z}}^l(\boldsymbol{x}) := \sum_{\{\bar{\boldsymbol{Y}}=\boldsymbol{z}, l_i, l_j \in \boldsymbol{z}\}} \frac{1}{2}\left[\hat{\mathbb{P}}(l_j|l_i, \boldsymbol{x}) + \hat{\mathbb{P}}(l_i|l_j, \boldsymbol{x})\right]. \tag{39}$$

C.2. Related Literature on Multi-Class Classification with Noisy Labels

Multi-class classification with noisy labels can date back to three decades ago [1], and keeps vibrant in recent years [11]. There is a large body of recent works that include but do not limit to the estimation of the noise transition matrix [33, 12, 37, 49, 70, 24, 61, 58], confident sample selection [45, 54, 55, 47, 15, 29, 38, 31], robust loss function design [26, 67, 28, 8], implicit/explicit regularization [14, 25, 27, 20, 16], and the integration of diverse techniques [30, 19, 21, 32]. We refer readers to [39, 11] for comprehensive review on multi-class classification with noisy labels.

In addition, the methods belonging to *label correction* have attracted much attention in multi-class classification with noisy labels [40, 65, 62]. Generally speaking, this kind of methods relies the prediction of a classifier trained on the noisy dataset, which recalibrates labels to the mislabeled data. Benefiting from the memorization effect of deep networks [2], the prediction is a good indicator to determine the clean label of mislabeled data. The dataset after label correction is then less noisy, which brings better generalization. However, few label-correction methods are investigated for multi-label classification with noisy labels, which is much more challenging than multi-class classification with noisy labels [23].

C.3. Related Literature on Multi-Label Classification with Clean & Noisy Labels

We briefly review works on multi-label classification with clean labels. If there is no confusion, we directly state multi-label classification. Multi-label classification has been studied for many years [60, 22, 59, 18, 6, 50]. In consideration of the increasing needs of today's big data, lots of methods based on deep learning are proposed [69, 36, 9, 63, 56, 3, 68, 43, 46, 10, 4]. In addition to the above works, some works [5, 57] claim that the label dependence can be used to enhance the learning of the instance-label dependence. They exploit graph convolutional networks to capture the -label dependence and inject

the captured information into multi-label classification, following promising classification performance. Recently, imperfect training data make us consider the side-effect of noisy labels in multi-label classification. Till now, there are relatively few methods specifically targeting this realistic problem. More advanced methods need to be excavated.

Normally, these methods perform an overall model adjustment to combat noisy labels. However, these methods highly rely on additional information except for provided training data with noisy labels. For example, partial methods [13, 42, 41, 34] learn an overall transition between noisy and clean labels to handle noisy labels, where a small dataset with clean labels is relied to guide the transition learning. Partial methods [64] introduce overall semantics-based regularization on training data to relieve the model's overfitting to noisy labels, where semantic label embeddings are injected with large-scale predefined word embeddings [35, 7]. Although the additional information is helpful, in many actual scenarios, it is luxurious or not feasible at all. Without the additional information, these methods become weak in multi-label classification with noisy labels [64], which greatly limits their practical applications [23].

C.4. Setting Difference between Multi-Label Classification with Noisy Labels and Partial Multi-Label Learning

It should be noted that the problem settings of multi-label classification with noisy labels and partial multi-label learning [51, 17] are different. Partial multi-label learning deals with the problem where each instance is assigned with a candidate label set, which contains multiple relevant labels and some irrelevant labels. The size of the candidate label set is usually much smaller than the size of label space. We need to detect the relevant labels for training. However, for our problem, there is no small candidate label set for reference, where we can only observe the whole label space. Intuitively, the methods in partial multi-label learning could be applied to multi-label classification with noisy labels. That is, we can identify some clean labels from noisy labels for training. However, this paradigm is inefficient, since only fractional labels are considered. Additionally, it is rather hard to accurately determine the number of identified labels for each instance.

D. Supplementary Experimental Settings

D.1. The Details of Baselines

In the main paper, we consider three types of baselines in experiments. Here, we detail the baselines.

1. Type-I baselines are designed for multi-label classification without considering noisy labels, which include

- CSRA [69] proposes simple and effective residual attention for multi-label learning. CSRA generates class-specific features for different labels by using spatial attention scores, and then combines them with the class-agnostic average pooling features.

- ADDGCN [57] proposes to exploit a semantic attention module and a GCN module for multi-label classification. As we discussed in the main paper, ADDGCN is the preparation technology of our HLC.

2. Type-I baselines are designed for multi-class classification with noisy labels, which include

- APL [26] combines two mutually reinforcing robust loss functions. For this baseline, we employ its combination of normalized BCE and MAE for comparison. The trade-off hyperparameter for the combinations of NBCE and MAE is set to 1.

- CDR [48] handles multi-class noisy labels using network pruning. A parameter judgment criteria is proposed to distinguish the critical/non-critical parameters for memorizing clean labels. The non-critical ones are forbidden to update, which mitigates the overfitting to mislabeled data.

- JOINT [40] shares a similar philosophy compared with our method, *i.e.*, label correction. It uses a joint optimization framework to handle noisy labels. The pseudo labels are generated dynamically by using the network's prediction to improve robustness. Meanwhile, regularizations about the class prior and entropy of prediction probabilities are used. In experiments, we utilize the hard-label version of JOINT [40].

3. Type-III baselines are designed for multi-label classification with noisy labels, which include:

- WSIC [13] consists of a clean net and a residual net. The aim is to learn a mapping from feature space to clean label space and a residual mapping from feature space to the residual between clean labels and noisy labels respectively. For fair comparison with our method, we only provide noisy training examples to WSIC.

- CCMN [52] establishes unbiased estimators with error bounds for solving the problem of multi-label learning with noisy labels, and further prove that the estimators are consistent with commonly used multi-label loss functions under some conditions.

4. The simple baseline that trains deep models on multi-label noisy datasets directly:

- BCE [60] uses the binary cross-entropy loss to train deep models in noisy datasets, without considering the side-effect of mislabeled data for generalization.

D.2. The Details of the Label Transition Matrix

In this paper, we consider both symmetric and pairflip cases for the generation of noisy labels. Specifically, if the overall noise rate is ϱ, the label transition matrix for symmetric cases are defined as

$$\text{Sym. } \varrho: \quad T := \begin{bmatrix} 1-\varrho & \frac{\varrho}{q-1} & \dots & \frac{\varrho}{q-1} & \frac{\varrho}{q-1} \\ \frac{\varrho}{q-1} & 1-\varrho & \frac{\varrho}{q-1} & \dots & \frac{\varrho}{q-1} \\ \vdots & & \ddots & & \vdots \\ \frac{\varrho}{q-1} & \dots & \frac{\varrho}{q-1} & 1-\varrho & \frac{\varrho}{q-1} \\ \frac{\varrho}{q-1} & \frac{\varrho}{q-1} & \dots & \frac{\varrho}{q-1} & 1-\varrho \end{bmatrix}_{q\times q}. \tag{40}$$

The label transition matrix for pariflip cases are defined as

$$\text{Pair. } \varrho: \quad T = \begin{bmatrix} 1-\varrho & \varrho & \dots & 0 & 0 \\ 0 & 1-\varrho & \varrho & \dots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & \dots & 0 & 1-\varrho & \varrho \\ \varrho & 0 & \dots & 0 & 1-\varrho \end{bmatrix}_{q\times q}. \tag{41}$$

E. Supplementary Experimental Results

In the main paper, we report results based on the performance of the last epoch during training, as did in [11, 44, 45, 19]. Here, to make comparison more comprehensive, we report results on noisy Pascal-VOC 2007 based on the best performance achieved during training. The results are provided in Table 1. Due to the memorization effect of deep networks [2], the networks would first memorize clean training data and then noisy training data. Therefore, in the early training, all methods could achieve good performance. We compared HLC with other advanced methods. Specifically, for mAP, although HLC does not always achieve the best results like the results in the main paper, the results are still competitive. For OF1 and CF1, HLC outperforms the other methods consistently.

It is worth mentioning that, the results in the main paper are much lower than the results in Table 1 in some cases. The experimental phenomenon means that one method severely overfits training data with incorrect labels as training progresses, which is pessimistic. Therefore, we should strive to design more robust methods to address the problem of multi-label classification with noisy labels. In this paper, we try and give a potential method, which outperforms baselines clearly. More efforts are expected to be put in by the community.

References

[1] Dana Angluin and Philip Laird. Learning from noisy examples. *Machine Learning*, 2(4):343–370, 1988. 4

[2] Devansh Arpit, Stanisław Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *ICML*, pages 233–242, 2017. 4, 6

[3] Tianshui Chen, Liang Lin, Xiaolu Hui, Riquan Chen, and Hefeng Wu. Knowledge-guided multi-label few-shot learning for general image recognition. *Transactions on Pattern Analysis and Machine Intelligence*, 2020. 4

[4] Tianshui Chen, Muxin Xu, Xiaolu Hui, Hefeng Wu, and Liang Lin. Learning semantic-specific graph representation for multi-label image recognition. In *ICCV*, pages 522–531, 2019. 4

[5] Zhao-Min Chen, Xiu-Shen Wei, Xin Jin, and Yanwen Guo. Multi-label image recognition with joint class-aware map disentangling and label correlation embedding. In *ICME*, pages 622–627, 2019. 4

Table 1. Comparisons with advanced methods on noisy Pascal-VOC 2007. The mean and standard deviation of the best results (%) during training are presented.

Metrics	Methods / Noise	Sym. 30%	Sym. 40%	Sym. 50%	Pair. 20%	Pair. 30%	Pair. 40%
mAP ↑	BCE	82.01±0.61	80.50±0.62	76.80±0.31	80.97±0.24	75.95±1.12	65.54±2.67
	CSRA	83.15±0.08	80.39±1.17	77.93±2.73	82.36±0.35	76.02±1.58	65.38±1.61
	ADDGCN	81.70±0.96	80.29±0.44	74.22±2.86	80.33±1.50	74.92±2.64	63.11±1.80
	APL	82.13±1.44	79.92±0.63	76.68±2.47	82.20±0.09	76.02±1.80	66.92±2.09
	CDR	82.35±1.17	78.33±1.04	77.01±1.61	81.00±0.20	76.37±1.04	66.21±2.35
	JOINT	82.12±0.55	81.00±0.39	76.84±1.12	81.33±0.60	76.77±0.55	66.50±1.86
	WSIC	82.17±0.19	78.14±1.06	77.25±0.90	81.06±1.06	75.22±1.37	65.88±2.80
	CCMN	81.80±0.73	80.20±1.10	76.77±1.73	82.27±0.41	76.03±1.39	66.93±2.03
	HLC†	82.40±0.17	81.19±1.22	78.04±0.29	82.30±0.61	76.40±1.82	67.61±2.12
OF1 ↑	BCE	70.97±0.65	62.99±0.75	55.43±1.80	75.95±0.77	71.26±0.88	63.33±2.74
	CSRA	73.52±1.06	65.21±0.93	52.84±2.11	78.02±0.92	72.66±1.75	62.77±3.06
	ADDGCN	71.11±1.16	63.05±1.84	48.62±2.31	74.29±1.72	67.83±0.98	59.12±2.73
	APL	71.10±0.50	61.44±0.88	51.77±2.84	74.50±1.29	68.04±1.98	63.30±1.60
	CDR	71.65±1.63	63.06±1.50	54.83±2.26	76.88±2.16	72.06±1.90	62.89±3.17
	JOINT	74.08±1.12	70.22±2.31	65.27±2.66	77.82±1.01	72.56±0.71	65.82±1.73
	WSIC	71.05±0.16	63.86±1.00	52.88±2.27	76.05±1.10	70.39±1.16	60.88±1.37
	CCMN	72.33±0.18	65.44±1.26	57.29±1.10	77.19±0.11	72.04±0.50	62.05±1.18
	HLC†	76.33±0.19	74.83±1.29	72.11±3.06	78.05±0.13	73.88±1.90	66.32±0.30
CF1 ↑	BCE	68.33±0.92	59.63±1.29	49.77±2.81	73.17±1.46	66.82±2.96	57.19±2.32
	CSRA	70.59±1.26	62.33±1.60	48.15±2.90	75.06±0.77	68.72±1.63	56.25±3.28
	ADDGCN	67.83±0.64	59.75±1.06	46.72±3.50	71.33±0.65	64.02±0.65	55.82±4.91
	APL	67.33±1.85	59.11±2.02	47.86±3.13	74.80±0.77	66.92±2.84	57.02±1.90
	CDR	68.03±1.62	60.02±1.17	48.94±2.65	73.77±1.04	67.06±1.84	57.38±2.10
	JOINT	71.17±0.29	66.11±1.59	57.93±1.82	75.25±0.73	70.01±1.99	56.28±2.19
	WSIC	68.11±0.52	60.39±1.14	46.25±4.74	74.02±1.26	67.09±2.84	55.76±3.66
	CCMN	68.58±0.44	64.82±2.17	54.82±1.06	74.15±0.92	67.79±2.33	58.06±2.37
	HLC†	73.11±0.91	72.08±1.16	68.31±0.77	76.00±0.71	71.07±1.95	61.52±2.28

[6] Tejas Chheda, Purujit Goyal, Trang Tran, Dhruvesh Patel, Michael Boratko, Shib Sankar Dasgupta, and Andrew McCallum. Box embeddings: An open-source library for representation learning using geometric structures. *arXiv preprint arXiv:2109.04997*, 2021. 4

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 5

[8] Lei Feng, Senlin Shu, Zhuoyi Lin, Fengmao Lv, Li Li, and Bo An. Can cross entropy loss be robust to label noise? In *IJCAI*, pages 2206–2212, 2021. 4

[9] Bin-Bin Gao and Hong-Yu Zhou. Learning to discover multi-class attentional regions for multi-label image recognition. *IEEE Transactions on Image Processing*, 30:5920–5932, 2021. 4

[10] Hao Guo, Kang Zheng, Xiaochuan Fan, Hongkai Yu, and Song Wang. Visual attention consistency under image transforms for multi-label image classification. In *CVPR*, pages 729–739, 2019. 4

[11] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *NeurIPS*, pages 8527–8537, 2018. 4, 6

[12] Dan Hendrycks, Mantas Mazeika, Duncan Wilson, and Kevin Gimpel. Using trusted data to train deep networks on labels corrupted by severe noise. In *NeurIPS*, 2018. 4

[13] Mengying Hu, Hu Han, Shiguang Shan, and Xilin Chen. Weakly supervised image classification through noise regularization. In *CVPR*, pages 11517–11525, 2019. 5

[14] Wei Hu, Zhiyuan Li, and Dingli Yu. Simple and effective regularization methods for training on noisily labeled data with generalization guarantee. In *ICLR*, 2020. 4

[15] Jinchi Huang, Lie Qu, Rongfei Jia, and Binqiang Zhao. O2u-net: A simple noisy label detection approach for deep neural networks. In *ICCV*, pages 3326–3334, 2019. 4

[16] Lu Jiang, Di Huang, Mason Liu, and Weilong Yang. Beyond synthetic noise: Deep learning on controlled noisy labels. In *ICML*, pages 4804–4815, 2020. 4

[17] Kaustav Kundu and Joseph Tighe. Exploiting weakly supervised visual patterns to learn from partial annotations. In *NeurIPS*, pages 561–572, 2020. 5

[18] Cheng Li, Bingyu Wang, Virgil Pavlu, and Javed Aslam. Conditional bernoulli mixtures for multi-label classification. In *ICML*, pages 2482–2491, 2016. 4

[19] Junnan Li, Richard Socher, and Steven C.H. Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *ICLR*, 2020. 4, 6

[20] Jingling Li, Mozhi Zhang, Keyulu Xu, John Dickerson, and Jimmy Ba. How does a neural network's architecture impact its robustness to noisy labels? In *NeurIPS*, 2021. 4

[21] Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. In *NeurIPS*, 2020. 4

[22] Weiwei Liu, Ivor W Tsang, and Klaus-Robert Müller. An easy-to-hard learning paradigm for multiple classes and multiple labels. *Journal of Machine Learning Research*, 18, 2017. 4

[23] Weiwei Liu, Haobo Wang, Xiaobo Shen, and Ivor Tsang. The emerging trends of multi-label learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 4, 5

[24] Yang Liu. Identifiability of label noise transition matrix. *arXiv preprint arXiv:2202.02016*, 2022. 4

[25] Michal Lukasik, Srinadh Bhojanapalli, Aditya Menon, and Sanjiv Kumar. Does label smoothing mitigate label noise? In *ICML*, pages 6448–6458, 2020. 4

[26] Xingjun Ma, Hanxun Huang, Yisen Wang, Simone Romano, Sarah Erfani, and James Bailey. Normalized loss functions for deep learning with noisy labels. In *ICML*, pages 6543–6553, 2020. 4, 5

[27] Xingjun Ma, Yisen Wang, Michael E Houle, Shuo Zhou, Sarah M Erfani, Shu-Tao Xia, Sudanthi Wijewickrema, and James Bailey. Dimensionality-driven learning with noisy labels. In *ICML*, pages 3361–3370, 2018. 4

[28] Aditya Krishna Menon, Ankit Singh Rawat, Sashank J Reddi, and Sanjiv Kumar. Can gradient clipping mitigate label noise? In *ICLR*, 2019. 4

[29] Baharan Mirzasoleiman, Kaidi Cao, and Jure Leskovec. Coresets for robust training of neural networks against noisy labels. In *NeurIPS*, 2020. 4

[30] Duc Tam Nguyen, Chaithanya Kumar Mummadi, Thi Phuong Nhung Ngo, Thi Hoai Phuong Nguyen, Laura Beggel, and Thomas Brox. Self: Learning to filter noisy labels with self-ensembling. In *ICLR*, 2020. 4

[31] Curtis G Northcutt, Tailin Wu, and Isaac L Chuang. Learning with confident examples: Rank pruning for robust classification with noisy labels. In *UAI*, 2017. 4

[32] Diego Ortego, Eric Arazo, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Multi-objective interpolation training for robustness to label noise. In *CVPR*, pages 6606–6615, 2021. 4

[33] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *CVPR*, pages 1944–1952, 2017. 4

[34] Cosmin Octavian Pene, Amirmasoud Ghiassi, Taraneh Younesian, Robert Birke, and Lydia Y Chen. Multi-label gold asymmetric loss correction with single-label regulators. *arXiv preprint arXiv:2108.02032*, 2021. 5

[35] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014. 5

[36] Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In *ICCV*, 2021. 4

[37] Jun Shu, Qian Zhao, Zengben Xu, and Deyu Meng. Meta transition adaptation for robust deep learning with noisy labels. *arXiv preprint arXiv:2006.05697*, 2020. 4

[38] Hwanjun Song, Minseok Kim, and Jae-Gil Lee. Selfie: Refurbishing unclean samples for robust deep learning. In *ICML*, pages 5907–5915, 2019. 4

[39] Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. 4

[40] Daiki Tanaka, Daiki Ikami, Toshihiko Yamasaki, and Kiyoharu Aizawa. Joint optimization framework for learning with noisy labels. In *CVPR*, 2018. 4, 5

[41] Arash Vahdat. Toward robustness against label noise in training deep discriminative neural networks. In *NeurIPS*, pages 5596–5605, 2017. 5

[42] Andreas Veit, Neil Alldrin, Gal Chechik, Ivan Krasin, Abhinav Gupta, and Serge Belongie. Learning from noisy large-scale datasets with minimal supervision. In *CVPR*, 2017. 5

[43] Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu. Cnn-rnn: A unified framework for multi-label image classification. In *CVPR*, pages 2285–2294, 2016. 4

[44] Yisen Wang, Weiyang Liu, Xingjun Ma, James Bailey, Hongyuan Zha, Le Song, and Shu-Tao Xia. Iterative learning with open-set noisy labels. In *CVPR*, pages 8688–8696, 2018. 6

[45] Hongxin Wei, Lei Feng, Xiangyu Chen, and Bo An. Combating noisy labels by agreement: A joint training method with co-regularization. In *CVPR*, pages 13726–13735, 2020. 4, 6

[46] Yunchao Wei, Wei Xia, Min Lin, Junshi Huang, Bingbing Ni, Jian Dong, Yao Zhao, and Shuicheng Yan. Hcp: A flexible cnn framework for multi-label image classification. *Transactions on Pattern Analysis and Machine Intelligence*, 38(9):1901–1907, 2015. 4

[47] Pengxiang Wu, Songzhu Zheng, Mayank Goswami, Dimitris Metaxas, and Chao Chen. A topological filter for learning with label noise. In *NeurIPS*, 2020. 4

[48] Xiaobo Xia, Tongliang Liu, Bo Han, Chen Gong, Nannan Wang, Zongyuan Ge, and Yi Chang. Robust early-learning: Hindering the memorization of noisy labels. In *ICLR*, 2021. 5

[49] Xiaobo Xia, Tongliang Liu, Nannan Wang, Bo Han, Chen Gong, Gang Niu, and Masashi Sugiyama. Are anchor points really indispensable in label-noise learning? In *NeurIPS*, pages 6835–6846, 2019. 4

[50] Ming-Kun Xie and Sheng-Jun Huang. Multi-label learning with pairwise relevance ordering. In *NeurIPS*, 2021. 4

[51] Ming-Kun Xie and Sheng-Jun Huang. Partial multi-label learning with noisy label identification. *Transactions on Pattern Analysis and Machine Intelligence*, 2021. 5

[52] Ming-Kun Xie and Sheng-Jun Huang. Ccmn: A general framework for learning with class-conditional multi-label noise. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 6

[53] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*, 2015. 4

[54] Quanming Yao, Hansi Yang, Bo Han, Gang Niu, and James Tin-Yau Kwok. Searching to exploit memorization effect in learning with noisy labels. In *ICML*, pages 10789–10798, 2020. 4

[55] Yazhou Yao, Zeren Sun, Chuanyi Zhang, Fumin Shen, Qi Wu, Jian Zhang, and Zhenmin Tang. Jo-src: A contrastive approach for combating noisy labels. In *CVPR*, pages 5192–5201, 2021. 4

[56] Vacit Oguz Yazici, Abel Gonzalez-Garcia, Arnau Ramisa, Bartlomiej Twardowski, and Joost van de Weijer. Orderless recurrent models for multi-label classification. In *CVPR*, 2020. 4

[57] Jin Ye, Junjun He, Xiaojiang Peng, Wenhao Wu, and Yu Qiao. Attention-driven dynamic graph convolutional network for multi-label image recognition. In *ECCV*, pages 649–665, 2020. 4, 5

[58] Mingyuan Zhang, Jane Lee, and Shivani Agarwal. Learning from noisy labels with no change to the training process. In *ICML*, pages 12468–12478, 2021. 4

[59] Min-Ling Zhang and Lei Wu. Lift: Multi-label learning with label-specific features. *Transactions on Pattern Analysis and Machine Intelligence*, 37(1):107–120, 2014. 4

[60] Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819–1837, 2013. 4, 6

[61] Yivan Zhang, Gang Niu, and Masashi Sugiyama. Learning noise transition matrix from only noisy labels via total variation regularization. In *ICML*, 2021. 4

[62] Yikai Zhang, Songzhu Zheng, Pengxiang Wu, Mayank Goswami, and Chao Chen. Learning with feature-dependent label noise: A progressive approach. In *ICLR*, 2021. 4

[63] Jiawei Zhao, Ke Yan, Yifan Zhao, Xiaowei Guo, Feiyue Huang, and Jia Li. Transformer-based dual relation graph for multi-label image recognition. In *ICCV*, pages 163–172, 2021. 4

[64] Wenting Zhao and Carla Gomes. Evaluating multi-label classifiers with noisy labels. *arXiv preprint arXiv:2102.08427*, 2021. 5

[65] Songzhu Zheng, Pengxiang Wu, Aman Goswami, Mayank Goswami, Dimitris Metaxas, and Chao Chen. Error-bounded correction of noisy labels. In *ICML*, pages 11447–11457, 2020. 2, 4

[66] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, pages 2921–2929, 2016. 4

[67] Xiong Zhou, Xianming Liu, Junjun Jiang, Xin Gao, and Xiangyang Ji. Asymmetric loss functions for learning with noisy labels. In *ICML*, pages 12846–12856, 2021. 4

[68] Feng Zhu, Hongsheng Li, Wanli Ouyang, Nenghai Yu, and Xiaogang Wang. Learning spatial regularization with image-level supervisions for multi-label image classification. In *CVPR*, pages 5513–5522, 2017. 4

[69] Ke Zhu and Jianxin Wu. Residual attention: A simple but effective method for multi-label recognition. In *ICCV*, pages 184–193, 2021. 4, 5

[70] Zhaowei Zhu, Tongliang Liu, and Yang Liu. A second-order approach to learning with instance-dependent label noise. In *CVPR*, pages 10113–10123, 2021. 4