A. DaFKD Without Uploading the Discriminator

The privacy of local generators can be protected by using secure aggregation. The privacy of the global generator can be protected by only outputting features instead of the original data, as shown in Figure 7, which is elaborated in FEDGEN[40] and mentioned in Section 3.2.

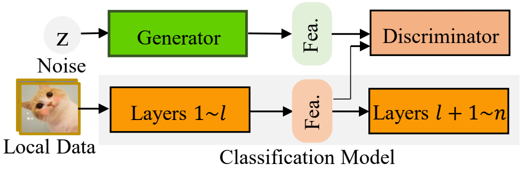

Figure 7. Feature generator. The discriminator and the generator learn the feature map instead of the original dataset. Similarly, the distillation dataset also includes the feature map.

B. DaFKD Without Uploading the Discriminator

In fact, the discriminator and the correlation factors are not necessarily visible to the server to protect the privacy of clients. More specifically, all clients can use the same generator to produce pseudo distillation data locally. Then, each client k inputs the distillation data x_i to the discriminator θ_k^d to produce correlation factors $f(\theta_{k,d}, x_i)$ and input the distillation data to the classification model w^k to produce soft predictions $s_{k,i}$. To enable the domain-aware federated distillation, each client k multiplies the correlation factors $f(\theta_{k,d}, x_i)$ to the corresponding soft predictions s obtaining $f(\theta_{k,d}, x_i)s_{k,i}$ and transmits it to the server. At the same time, the server aggregates $f(\theta_{k,d}, x_i)$ from all clients in a privacy-preserving manner by using differential privacy or homomorphic encryption to obtain $\sum_{k=1}^{K_t} f(\theta_{k,d}, x_i)$. After receiving multiplied soft predictions $\alpha_{k,i}s_{k,i}$ from all clients and the aggregated $\sum_{k=1}^{K_t} f(\theta_{k,d}, x_i)$, the server normalizes the multiplied soft predictions getting $\frac{\alpha_{k,i}s_{k,i}}{\sum_{k=1}^{K_t} f(\theta_{k,d},x_i)}$. To enable distillation, the server uses the same random seed as each client is adopted to produce the pseudo data x_i and inputs it to the global model $\hat{w}$ obtaining $s(\hat{w}; x_i)$. Finally, the server implements the ensemble distillation using (8), i.e.,

$$w_{t+1} = \arg\min_{\hat{w}_{t+1}} \mathcal{L}_{KD}(\hat{w}_{t+1}) = \frac{1}{\hat{D}^g} \sum_{x_i \in \hat{\mathbb{D}}^g} KL\left(\sum_{k=1}^{K_t} \hat{\alpha}_{k,i} \cdot s(w_t^k; x_i), s(\hat{w}_{t+1}; x_i)\right).$$

C. Proof of Theorem 1

Theorem 1 *Denote the data distribution of each client k by $p_k(x)$, the data distribution of all clients by $p(x)$, and the pseudo data distribution of the generator by $p_g(x)$. If the Algorithm 1 trains the discriminator θ_k^d and the global generator θ^g to the optima for the loss function (5), then the pseudo data distribution of the generator is $p_g^*(x) = p(x)$, and the discriminator outputs $f^*(\theta_k^d; x) = \frac{p_k(x)}{p_k(x)+p(x)}$ for each client $k = 1 \ldots K$.*
Proof: To analyze the distribution fitted by the global generator and multiple discriminators, we formally present the overall adversarial loss function including the generator and all discriminators as:

$$\max_{\theta^g} \min_{\theta_1^d, \cdots, \theta_K^d} \mathcal{L}_{adv}(\theta_1^d, \cdots, \theta_K^d) = -\frac{1}{K} \sum_{k=1}^{K} \Big[\mathbb{E}_{x \sim p_k(x)} \log f(\theta_k^d; x) + \mathbb{E}_{z \sim p_z(z)} \log\left(1 - f(\theta_k^d; g(\theta^g; z))\right)\Big], \qquad (13)$$

where $p_k(x)$ is the data distribution of client k. Given the fixed generator θ^g, considering the distribution of generated data as $p_g(x)$, we have

$$\begin{aligned}\min_{\theta_1^d,\cdots,\theta_K^d} \mathcal{L}_{adv}(\theta_1^d,\cdots,\theta_K^d) &= -\frac{1}{K}\sum_{k=1}^{K}\Big[\mathbb{E}_{x\sim p_k(x)}\text{log } f(\theta_k^d;x) + \mathbb{E}_{x\sim p_g(x)}\text{log }(1-f(\theta_k^d;x))\Big] \\ &= -\frac{1}{K}\sum_{k=1}^{K}\Big[\int_x p_k(x)\text{log } f(\theta_k^d;x)dx + \int_x p_g(x)\text{log }(1-f(\theta_k^d;x))dx\Big] \\ &= -\frac{1}{K}\sum_{k=1}^{K}\Big[\int_x p_k(x)\text{log } f(\theta_k^d;x) + p_g(x)\text{log }(1-f(\theta_k^d;x))dx\Big].\end{aligned} \tag{14}$$

Obviously, the equation (16) achieves the minima when

$$f^*(\theta_k^d;x) = \frac{p_k(x)}{p_k(x)+p_g(x)},\ \forall k = 1,\cdots,K. \tag{15}$$

Now, to solve the optimal generator, we bring (15) back to (13) and obtain

$$\begin{aligned}\max_{\theta^g}\mathcal{L}_{adv}(\theta^g) &= -\frac{1}{K}\sum_{k=1}^{K}\Big[\mathbb{E}_{x\sim p_k(x)}\text{log }\frac{p_k(x)}{p_k(x)+p_g(x)} + \mathbb{E}_{x\sim p_g(x)}\text{log }\frac{p_g(x)}{p_k(x)+p_g(x)}\Big] \\ &= -\frac{1}{K}\sum_{k=1}^{K}\Big[\int_x p_k(x)\text{log }\frac{p_k(x)}{p_k(x)+p_g(x)}dx + \int_x p_g(x)\text{log }\frac{p_g(x)}{p_k(x)+p_g(x)}dx\Big] \\ &= -\frac{1}{K}\sum_{k=1}^{K}\Big[\int_x p_k(x)\text{log }\frac{p_k(x)}{p_k(x)+p_g(x)} + p_g(x)\text{log }\frac{p_g(x)}{p_k(x)+p_g(x)}dx\Big] \\ &= -\int_x \frac{1}{K}\sum_{k=1}^{K}\Big[p_k(x)\text{log }\frac{p_k(x)}{p_k(x)+p_g(x)} + p_g(x)\text{log }\frac{p_g(x)}{p_k(x)+p_g(x)}\Big]dx \\ &= \text{log } 4 - \frac{1}{K}\sum_{k=1}^{K}\text{JSD}(p_k(x)||p_g(x)),\end{aligned} \tag{16}$$

where JSD denotes the Jensen-Shannon Divergence. Since the centroid defined as the average sum of a finite set of probability distributions is the minimizer of Jensen-Shannon divergences between a probability distribution and the prescribed set of distributions, we can derive the formulation of optimal $p_g(x)$ as $p_g^*(x) = \frac{1}{K}\sum_{k=1}^{K}p_k(x)$, which completes the proof.

D. Proof of Theorem 2

Theorem 2 *Denote the empirical distribution of activation from each client k by $\hat{p}_k$ and the empirical distribution of global dataset by $\hat{p} = \frac{1}{K}\sum_{k=1}^{K}\hat{p}_k$. Then, given the constants $0 < \delta \leq 1$ and $\sigma > 0$, with the probability at least $1-\delta$, the expected generalization error $\mathcal{L}_p(\sum_{k=1}^{K}\hat{\alpha}_k(x)h_{\hat{p}_k})$ of domain-aware ensemble model is:*

$$\begin{aligned}&\mathcal{L}_p(\sum_{k=1}^{K}\hat{\alpha}_k(x)h_{\hat{p}_k}) \\ &\leq (K+1)\mathcal{L}_{\hat{p}}(h_{\hat{p}}) + (K+1)\sqrt{\frac{\sigma^2 log\frac{2K}{\delta}}{2m}}.\end{aligned} \tag{17}$$

Proof: We seek to establish the relationship between $\mathcal{L}_p(\frac{1}{K}\sum_{k=1}^{K}\hat{\alpha}_k h_{\hat{p}_k})$ and $\mathcal{L}_{\hat{p}}(h_{\hat{p}})$. Considering that the convexity of the loss function in terms of the prediction, we have

$$\mathcal{L}_p(\sum_{k=1}^{K}\hat{\alpha}_k(x)h_{\hat{p}_k}) = \int_x p(x)L(\sum_{k=1}^{K}\hat{\alpha}_k(x)h_{\hat{p}_k}(x))dx \leq \int_x p(x)\big[\sum_{k=1}^{K}\hat{\alpha}_k(x)L(h_{\hat{p}_k}(x))\big]dx. \tag{18}$$

Considering the optimal discriminator $f^*(\theta_k^d;x) = \frac{p_k(x)}{p_k(x)+p(x)}$ where $p(x) = \frac{1}{K}\sum_{k=1}^K p_k(x)$, we have

$$\begin{aligned}
\hat{\alpha}_k(x) &= \frac{f(\theta_k^d;x)}{\sum_{k=1}^K f(\theta_k^d;x)} = \frac{\frac{p_k(x)}{p_k(x)+p(x)}}{\sum_{k=1}^K \frac{p_k(x)}{p_k(x)+p(x)}} \\
&= \frac{p_k(x)}{(p_k(x)+p(x))\sum_{i=1}^K \frac{p_i(x)}{p_i(x)+p(x)}} \\
&\leq \frac{p_k(x)}{\frac{p_k(x)+p(x)}{\max\{p_1(x),\cdots,p_K(x)\}+p(x)}\sum_{i=1}^K p_i(x)} \\
&\leq \frac{p_k(x)}{\frac{p(x)}{Kp(x)+p(x)}\sum_{i=1}^K p_i(x)} \\
&= (K+1)\frac{p_k(x)}{\sum_{i=1}^K p_i(x)} \\
&= \frac{(K+1)}{K}\cdot\frac{p_k(x)}{p(x)}.
\end{aligned} \tag{19}$$

Bringing the bound of $\hat{\alpha}$ in (19) back to (18) derives:

$$\begin{aligned}
\mathcal{L}_p(\sum_{k=1}^K \hat{\alpha}_k(x)h_{\hat{p}_k}) &\leq \int_x p(x)\big[\sum_{k=1}^K \frac{(K+1)}{K}\cdot\frac{p_k(x)}{p(x)}L(h_{\hat{p}_k}(x))\big]dx \\
&= \frac{(K+1)}{K}\sum_{k=1}^K \int_x p_k(x)L(h_{\hat{p}_k}(x))dx \\
&= \frac{(K+1)}{K}\sum_{k=1}^K \mathcal{L}_{p_k}(h_{\hat{p}_k}).
\end{aligned} \tag{20}$$

Next, we bound the $\mathcal{L}_{p_k}(h_{\hat{p}_k})$ with its empirical counterpart $\mathcal{L}_{\hat{p}_k}(h_{\hat{p}_k})$ through Hoeffding inequality. Without losing the generality, we consider the simplified case where the size of samples in all clients are equal, i.e., $D_1 = D_2 = \ldots = D_K = m$. Then, a simple application of the Hoeffding's inequality gives:

$$P(|\mathcal{L}_{p_k}(h_{\hat{p}_k}) - \mathcal{L}_{\hat{p}_k}(h_{\hat{p}_k})| \geq \epsilon) \leq 2\text{exp}(-\frac{2m\epsilon^2}{\sigma^2}), \tag{21}$$

where $\epsilon > 0$ and $\sigma > 0$ are the constants. Thereby, with probability at least $1 - \frac{\delta}{K}$, we have:

$$\mathcal{L}_{p_k}(h_{\hat{p}_k}) \leq \mathcal{L}_{\hat{p}_k}(h_{\hat{p}_k}) + \sqrt{\frac{\sigma^2\text{log}\frac{2K}{\delta}}{2m}}. \tag{22}$$

For all K devices, we have

$$\begin{aligned}
&P\Big[\bigcap_{k=1}^K \big(\mathcal{L}_{p_k}(h_{\hat{p}_k}) \leq \mathcal{L}_{\hat{p}_k}(h_{\hat{p}_k}) + \sqrt{\frac{\sigma^2\text{log}\frac{2K}{\delta}}{2m}}\big)\Big] \\
&= 1 - P\Big[\bigcup_{k=1}^K \big(\mathcal{L}_{p_k}(h_{\hat{p}_k}) \geq \mathcal{L}_{\hat{p}_k}(h_{\hat{p}_k}) + \sqrt{\frac{\sigma^2\text{log}\frac{2K}{\delta}}{2m}}\big)\Big] \\
&\geq 1 - \sum_{k=1}^K P\Big[\big(\mathcal{L}_{p_k}(h_{\hat{p}_k}) \geq \mathcal{L}_{\hat{p}_k}(h_{\hat{p}_k}) + \sqrt{\frac{\sigma^2\text{log}\frac{2K}{\delta}}{2m}}\big)\Big] \\
&\geq 1 - \delta.
\end{aligned} \tag{23}$$

Putting (22) back to (20) derives:

$$\mathcal{L}_p(\sum_{k=1}^{K} \hat{\alpha}_k(x) h_{\hat{p}_k}) \leq \frac{(K+1)}{K} \sum_{k=1}^{K} \Big(\mathcal{L}_{\hat{p}_k}(h_{\hat{p}_k}) + \sqrt{\frac{\sigma^2 \log \frac{2K}{\delta}}{2m}} \Big). \tag{24}$$

Considering that $h_{\hat{p}_k}$ minimizes the loss function over the distribution $\hat{p}_k$ of training dataset $\mathbb{D}_k$, $\mathcal{L}_{\hat{p}_k}(h_{\hat{p}_k}) \leq \mathcal{L}_{\hat{p}_k}(h_{\hat{p}})$ can be easily obtained. According to the definition that $\hat{p} = \frac{1}{K} \sum_{k=1}^{K} \hat{p}_k$, we can derive

$$\frac{1}{K} \sum_{k=1}^{K} \mathcal{L}_{\hat{p}_k}(h_{\hat{p}_k}) \leq \frac{1}{K} \sum_{k=1}^{K} \mathcal{L}_{\hat{p}_k}(h_{\hat{p}}) = \mathcal{L}_{\hat{p}}(h_{\hat{p}}). \tag{25}$$

Thereby, the following inequality holds with probability at least $1 - \delta$:

$$\begin{aligned} \mathcal{L}_p(\sum_{k=1}^{K} \hat{\alpha}_k(x) h_{\hat{p}_k}) &\leq \frac{(K+1)}{K} \sum_{k=1}^{K} \mathcal{L}_{\hat{p}_k}(h_{\hat{p}_k}) + (K+1) \sqrt{\frac{\sigma^2 \log \frac{2K}{\delta}}{2m}} \\ &\leq (K+1) \mathcal{L}_{\hat{p}}(h_{\hat{p}}) + (K+1) \sqrt{\frac{\sigma^2 \log \frac{2K}{\delta}}{2m}}. \end{aligned} \tag{26}$$