A Proofs

Section A.1 presents the lemmas used to prove the main results. Section A.2 presents the main results and their detailed proofs.

A.1 Preliminaries

Lemma A.1 ([53], Theorem 26.5)**.** *With probability of at least $1-\delta$, for all $h \in \mathcal{H}$,*

$$|\epsilon_u(h) - \widehat{\epsilon}_{\widehat{u}}(h)| \leq 2\mathcal{R}_H^l(\widehat{u}) + 4c\sqrt{\frac{\ln(4/\delta)}{N}}.$$

Lemma A.2 ([53], Theorem 26.5)**.** *With probability of at least $1-\delta$, for all $h \in \mathcal{H}$,*

$$\epsilon_u(\widehat{h}_u) - \epsilon_u(h_u^*) \leq 2\mathcal{R}_H^l(\widehat{u}) + 5c\sqrt{\frac{\ln(8/\delta)}{N}}.$$

Lemma A.3 ([44], Talagrand's contraction lemma)**.** *For any L-Lipschitz loss function $l(\cdot,\cdot)$, we obtain,*

$$\mathcal{R}(l \circ \mathcal{H} \circ \widehat{u}) \leq L\mathcal{R}(\mathcal{H} \circ \widehat{u}),$$

where $\mathcal{R}_{\mathcal{H}}^l(\widehat{u}) = \frac{1}{2N}\mathbb{E}_{\boldsymbol{\sigma}\sim\{\pm1\}^{2N}}\left[\sup_{h\in\mathcal{H}}\sum_{x\in\widehat{u}}\sigma_i h(x_i)\right]$.

Lemma A.4 ([18], Theorem 1)**.** *Let $\mathcal{H}$ be the class of real-valued networks of depth D over the domain $\mathcal{X}$. Assume the Frobenius norm of the weight matrices are at most $M_1, \ldots, M_d$. Let the activation function be 1-Lipschitz, positive-homogeneous, and applied element-wise (such as the ReLU). Then,*

$$\mathbb{E}_{\sigma\in\{\pm1\}^N}\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^{N}\sigma_i h(x_i)\right] \leq \sqrt{N}B(\sqrt{2D\ln 2}+1)\prod_{i=1}^{D}M_i.$$

A.2 Main Results

Theorem 4.1. *For any $x \in \mathcal{X}$, $h \in \mathcal{H}$ and $a \in \mathcal{A}$, assume we have $|l(h(x), a(x))| \leq c$. Then, with probability at least $1-\delta$, we can derive the following general bound,*

$$|\mathcal{D}_R(p\|q) - \widehat{\mathcal{D}}_R(\widehat{p}\|\widehat{q})| \leq 2\mathcal{D}_{\mathrm{L}}(p\|q) + 2\mathcal{R}_{\mathcal{H}}^l(\widehat{p}) + 2\mathcal{R}_{\mathcal{H}}^l(\widehat{q}) + 12c\sqrt{\frac{\ln(8/\delta)}{N}}.$$

Proof. We observe that

$$\begin{aligned}
&|\mathcal{D}_{\mathrm{R}}(p\|q) - \widehat{\mathcal{D}}_{\mathrm{R}}(\widehat{p}\|\widehat{q})| \\
=&\left||\epsilon_q(h_u^*) - \epsilon_p(h_u^*)| - |\widehat{\epsilon}_{\widehat{q}}(\widehat{h}_u) - \widehat{\epsilon}_{\widehat{p}}(\widehat{h}_u)|\right| \\
\leq&\left|\epsilon_q(h_u^*) - \epsilon_p(h_u^*,) - \widehat{\epsilon}_{\widehat{q}}(\widehat{h}_u) + \widehat{\epsilon}_{\widehat{p}}(\widehat{h}_u)\right| \\
=&\left|\epsilon_q(h_u^*) - \epsilon_p(h_u^*) + \widehat{\epsilon}_{\widehat{p}}(\widehat{h}_u) - \epsilon_p(\widehat{h}_u) + \epsilon_p(\widehat{h}_u) - \epsilon_q(\widehat{h}_u) + \epsilon_q(\widehat{h}_u) - \widehat{\epsilon}_{\widehat{q}}(\widehat{h}_u)\right| \\
\leq&|\epsilon_q(h_u^*) - \epsilon_p(h_u^*)| + \left|\epsilon_p(\widehat{h}_u) - \epsilon_q(\widehat{h}_u)\right| + \left|\widehat{\epsilon}_{\widehat{p}}(\widehat{h}_u) - \epsilon_p(\widehat{h}_u)\right| + \left|\epsilon_q(\widehat{h}_u) - \widehat{\epsilon}_{\widehat{q}}(\widehat{h}_u)\right| \\
\leq&2\mathcal{D}_{\mathrm{L}}(p\|q) + \left|\widehat{\epsilon}_{\widehat{p}}(\widehat{h}_u) - \epsilon_p(\widehat{h}_u)\right| + \left|\epsilon_q(\widehat{h}_u) - \widehat{\epsilon}_{\widehat{q}}(\widehat{h}_u)\right|.
\end{aligned} \tag{7}$$

The first two inequalities are owing to the triangle inequality, and the third inequality is due to the definition of L-divergence Eq.(5). We complete the proof by applying Lemma A.1 to bound $\left|\widehat{\epsilon}_{\widehat{p}}(\widehat{h}_u) - \epsilon_p(\widehat{h}_u)\right|$ and $\left|\epsilon_q(\widehat{h}_u) - \widehat{\epsilon}_{\widehat{q}}(\widehat{h}_u)\right|$ in Eq.(7). □

Corollary 4.2. *Following the conditions of Theorem 4.1, the upper bound of* $\sqrt{\text{Var}\left[\widehat{\mathcal{D}}_R(\widehat{p}\|\widehat{q})\right]}$ *is*

$$\frac{34c}{\sqrt{N}} + \sqrt{2\pi}\left(2\mathcal{D}_{\text{L}}(p\|q) + 2\mathcal{R}^l_{\mathcal{H}}(\widehat{p}) + 2\mathcal{R}^l_{\mathcal{H}}(\widehat{q})\right).$$

Proof. For convenience, we assume $\mathcal{C} = 2\mathcal{D}_{\text{L}}(p\|q) + 2\mathcal{R}^l_{\mathcal{H}}(\widehat{p}) + 2\mathcal{R}^l_{\mathcal{H}}(\widehat{q})$. Following to proof of Corollary 1 in [62], we have □

$$\begin{aligned}
\text{Var}\left[\widehat{\mathcal{D}}_{\text{R}}(\widehat{p}\|\widehat{q})\right] \leq & \mathbb{E}\left[\left(\mathcal{D}_{\text{R}}(p\|q) - \widehat{\mathcal{D}}_{\text{R}}(\widehat{p}\|\widehat{q})\right)^2\right] \\
= & \int_{t=0}^{\infty} \Pr\left[\left|\mathcal{D}_{\text{R}}(p\|q) - \widehat{\mathcal{D}}_{\text{R}}(\widehat{p}\|\widehat{q})\right| \geq \sqrt{t}\right] dt \\
\overset{s=\sqrt{t}}{=} & \int_{s=0}^{\infty} \Pr\left[\left|\mathcal{D}_{\text{R}}(p\|q) - \widehat{\mathcal{D}}_{\text{R}}(\widehat{p}\|\widehat{q})\right| \geq s\right] 2sds \\
\leq & \int_{s=0}^{\mathcal{C}} 2sds + \int_{s=0}^{\infty} \Pr\left[\left|\widehat{\mathcal{D}}_{\text{R}}(\widehat{p}\|\widehat{q}) - \mathcal{D}_{\text{R}}(p\|q)\right| \geq \mathcal{C} + s\right] 2(\mathcal{C}+s)ds \\
\leq & \mathcal{C}^2 + 16\int_{s=0}^{\infty} (\mathcal{C}+s) e^{\frac{-Ns^2}{144c^2}} ds \\
= & \mathcal{C}^2 + 16\mathcal{C}\int_{s=0}^{\infty} e^{\frac{-Ns^2}{144c^2}} ds + 16\int_{s=0}^{\infty} se^{\frac{-Ns^2}{144c^2}} ds \\
\overset{t=\frac{s}{c}\sqrt{\frac{N}{144}}}{=} & \mathcal{C}^2 + 16\mathcal{C}c\sqrt{\frac{a}{N}}\int_{t=0}^{\infty} e^{-t^2} dt + \frac{2304c^2}{N}\int_{s=0}^{\infty} te^{-t^2} dt \\
= & \mathcal{C}^2 + 96\mathcal{C}c\sqrt{\frac{\pi}{N}} + \frac{1152c^2}{N} \\
\leq & \left(\frac{34c}{\sqrt{N}} + \mathcal{C}\sqrt{2\pi}\right)^2,
\end{aligned} \tag{8}$$

where the third inequality is due to

$$\Pr\left[\left|\mathcal{D}_{\text{R}}(p\|q) - \widehat{\mathcal{D}}_{\text{R}}(\widehat{p}\|\widehat{q})\right| \geq \mathcal{C} + s\right] \leq 8e^{\frac{-Ns^2}{144c^2}}, \tag{9}$$

obtained from the results of Theorem 4.1.

Proposition 4.3. *Based on the conditions of Theorem 4.1, we assume* $\mathcal{H}$ *is the class of real-valued networks of depth* D *over the domain* $\mathcal{X}$*. Let the Frobenius norm of the weight matrices be at most* $M_1, \ldots, M_D$*, the activation function be 1-Lipschitz, positive-homogeneous and applied element-wise (such as the ReLU). Then, with probability at least* $1-\delta$*, we have,*

$$|\mathcal{D}_{\text{R}}(p\|q) - \widehat{\mathcal{D}}_{\text{R}}(\widehat{p}\|\widehat{q})| \leq 2\mathcal{D}_{\text{L}}(p\|q) + \frac{4LB(\sqrt{2D\ln 2}+1)\prod_{i=1}^{D} M_i}{\sqrt{N}} + 12c\sqrt{\frac{\ln(8/\delta)}{N}}.$$

Proof. We complete the proof by applying Lemma A.3 and Lemma A.4 to bound the Rademacher complexity of deep neural networks in Theorem 4.1. □

Corollary 4.4. *Following the conditions of Proposition 4.3, with probability at least* $1-\delta$*, we have the following conditional bounds if* $\mathcal{D}_{\text{L}}(p\|q) \leq \left|\epsilon_u(h^*_u) - \epsilon_u(\widehat{h}_u)\right|$*,*

$$|\mathcal{D}_{\text{R}}(p\|q) - \widehat{\mathcal{D}}_{\text{R}}(\widehat{p}\|\widehat{q})| \leq \frac{8LB(\sqrt{2D\ln 2}+1)\prod_{i=1}^{D} M_i}{\sqrt{N}} + 22c\sqrt{\frac{\ln(16/\delta)}{N}}.$$

Proof. Following the proof of Theorem 4.1, we have

$$
\begin{aligned}
&|\mathcal{D}_{\mathrm{R}}(p\|q) - \widehat{\mathcal{D}}_{\mathrm{R}}(\widehat{p}\|\widehat{q})| \\
\leq& \left|\epsilon_q(h_u^*) - \epsilon_p(h_u^*) + \widehat{\epsilon}_{\widehat{p}}(\widehat{h}_u) - \epsilon_p(\widehat{h}_u)\right| + \left|\widehat{\epsilon}_{\widehat{p}}(\widehat{h}_u) - \epsilon_p(\widehat{h}_u)\right| + \left|\epsilon_q(\widehat{h}_u) - \widehat{\epsilon}_{\widehat{q}}(\widehat{h}_u)\right| \\
=& \max\left(\left|\epsilon_q(h^*) - \epsilon_p(h^*) + \epsilon_p(\widehat{h}) - \epsilon_q(\widehat{h})\right|, \left|(\epsilon_q(h^*) + \epsilon_p(h^*)) - \left(\epsilon_p(\widehat{h}) - \epsilon_q(\widehat{h})\right)\right|\right) \\
&+ \left|\widehat{\epsilon}_{\widehat{p}}(\widehat{h}_u) - \epsilon_p(\widehat{h}_u)\right| + \left|\epsilon_q(\widehat{h}_u) - \widehat{\epsilon}_{\widehat{q}}(\widehat{h}_u)\right| \\
\leq& 2\max\left(D_{\mathrm{L}}(p\|q), \left|\epsilon_u(h^*) - \epsilon_u(\widehat{h})\right|\right) + \left|\widehat{\epsilon}_{\widehat{p}}(\widehat{h}_u) - \epsilon_p(\widehat{h}_u)\right| + \left|\epsilon_q(\widehat{h}_u) - \widehat{\epsilon}_{\widehat{q}}(\widehat{h}_u)\right| \\
\leq& 2\left|\epsilon_u(h^*) - \epsilon_u(\widehat{h})\right| + \left|\widehat{\epsilon}_{\widehat{p}}(\widehat{h}_u) - \epsilon_p(\widehat{h}_u)\right| + \left|\epsilon_q(\widehat{h}_u) - \widehat{\epsilon}_{\widehat{q}}(\widehat{h}_u)\right|.
\end{aligned}
\tag{10}
$$

The first two inequalities are owing to the triangle inequality, and the third inequality is due to the given condition $\mathcal{D}_{\mathrm{L}}(p\|q) \leq \left|\epsilon_u(h_u^*) - \epsilon_u(\widehat{h}_u)\right|$. We complete the proof by applying Lemma A.2 to bound $\left|\epsilon_u(h^*) - \epsilon_u(\widehat{h})\right|$ and Lemma A.1 to bound $\left|\widehat{\epsilon}_{\widehat{p}}(\widehat{h}_u) - \epsilon_p(\widehat{h}_u)\right|$ and $\left|\epsilon_q(\widehat{h}_u) - \widehat{\epsilon}_{\widehat{q}}(\widehat{h}_u)\right|$ in Eq.(7). □

Corollary 4.5. *Following the conditions of Proposition 4.3, as $N \to \infty$, we have,*

$$\widehat{\mathcal{D}}_{\mathrm{R}}(\widehat{p}\|\widehat{q}) \leq 2\mathcal{D}_{\mathrm{L}}(p\|q) + \mathcal{D}_{\mathrm{R}}(p\|q).$$

Proof. Based on the result on Proposition 4.3, for any $\delta \in (0, 1)$, we know that

$$\frac{4LB(\sqrt{2D\ln 2}+1)\prod_{i=1}^{D} M_i}{\sqrt{N}} + 12c\sqrt{\frac{\ln(8/\delta)}{N}} \to 0, \tag{11}$$

when $N \to \infty$. We complete the proof by applying the triangle inequality. □

B Algorithm Procedure

B.1 Training Procedure of R-Div

Algorithm 1 Estimate model-oriented distribution discrepancy by R-divergence

Input: two datasets $\widehat{p}$ and $\widehat{q}$ and a learning model $\mathcal{T}$ with hypothesis space $\mathcal{H}$ and loss function l
Generate the merged dataset: $\widehat{u} = \widehat{p} \cup \widehat{q}$
Learn the minimum hypothesis on the mixed data:

$$\widehat{h}_u \in \arg\min_{h\in\mathcal{H}} \widehat{\epsilon}_{\widehat{u}}(h)$$

Evaluate the empirical risks: $\widehat{\epsilon}_{\widehat{p}}(\widehat{h}_{\widehat{u}})$ and $\widehat{\epsilon}_{\widehat{q}}(\widehat{h}_{\widehat{u}})$
Estimate the R-divergence as the discrepancy:

$$\widehat{\mathcal{D}}_{\mathrm{R}}(\widehat{p}\|\widehat{q}) = |\widehat{\epsilon}_{\widehat{p}}(\widehat{h}_{\widehat{u}}) - \widehat{\epsilon}_{\widehat{q}}(\widehat{h}_{\widehat{u}})|$$

Output: empirical estimator $\widehat{\mathcal{D}}_{\mathrm{R}}(\widehat{p}\|\widehat{q})$

B.2 Calculation Procedure of the Average Test Power

Algorithm 2 Calculate the average test power of R-divergence

Input: datasets $\widehat{p}$ and $\widehat{q}$, empirical estimator $\widehat{\mathcal{D}}_{\mathrm{R}}(\cdot\|\cdot)$, α, K, Z
for $k = 1$ **to** K **do**
  for $z = 1$ **to** Z **do**
    if $z == 1$ **then**
      $(\widehat{p}^1, \widehat{q}^1) = (\widehat{p}, \widehat{q})$
    else
      Generate $(\widehat{p}^z, \widehat{q}^z)$ by uniformly randomly swapping samples between $\widehat{p}$ and $\widehat{q}$
    end if
    Calculate $\widehat{\mathcal{D}}_{\mathrm{R}}(\widehat{p}^z\|\widehat{q}^z)$
  end for
  Obtain $\mathcal{G} = \{\widehat{\mathcal{D}}_{\mathrm{R}}(\widehat{p}^z\|\widehat{q}^z)\}_{z=1}^{Z}$
  $r_k = \left(\widehat{\mathcal{D}}_{\mathrm{R}}(\widehat{p}\|\widehat{q}) \text{ is in the top } \alpha\text{-quantile among } \mathcal{G}\right)$
end for
Output: average test power $\frac{\sum_{k=1}^{K} r_k}{K}$

Table 7: Method Comparison.

Methods	Discrepancy	Remarks
ME	$\sqrt{\frac{1}{J}\sum_{j=1}^{J}\left(\mu_p(T_j) - \mu_q(T_j)\right)^2}$	I:$\mu_p(t) = \int k(x,t)dp(x)$ II:$k(x,t) = \exp(-\|x-y\|^2/\gamma^2)$ III:T is drawn from an absolutely continuous distribution.
SCF	$\sqrt{\frac{1}{2J}\sum_{j=1}^{2}\left(z_p^{\sin}(T_j) - z_p^{\sin}(T_j)\right)^2 + \left(z_p^{\cos}(T_j) - z_p^{\cos}(T_j)\right)^2}$	I:$z_p^{\sin}(t) = \int \kappa(x)\sin(x^T t)dp(x)$ II:$z_p^{\cos}(t) = \int \kappa(x)\cos(x^T t)dp(x)$ III:$\kappa(x)$ is the Fourier transform of a kernel.
C2STS-S	$\frac{1}{\lvert\hat{p}_{\mathrm{te}}\rvert}\sum_{x\in\hat{p}_{\mathrm{te}}} f_w(x) - \frac{1}{\lvert\hat{q}_{\mathrm{te}}\rvert}\sum_{x'\in\hat{q}_{\mathrm{te}}} f_w(x')$	I : $\hat{p}_{\mathrm{tr}}, \hat{p}_{\mathrm{te}} \sim p$, $\hat{q}_{\mathrm{tr}}, \hat{q}_{\mathrm{te}} \sim q$ II: f_w is a binary classifier to separate $\hat{p}_{\mathrm{tr}}$ and $\hat{q}_{\mathrm{tr}}$ III: Samples from p and q are labeled with 0 and 1, respectively. IV: $g_w(x,y) = \mathbb{I}\left[\mathbb{I}\left(f_w(x) > \frac{1}{2}\right) = y\right]$
C2ST-L	$\frac{1}{\lvert\hat{p}_{\mathrm{te}}\rvert+\lvert\hat{q}_{\mathrm{te}}\rvert}\sum_{x\in\hat{p}_{\mathrm{te}}} g_w(x,0) + \sum_{x'\in\hat{q}_{\mathrm{te}}} g_w(x',1)$	
MMD-O	$\sqrt{\mathbb{E}\left[k(x,x') + k(y,y') - 2k(x,y)\right]}$, $x, x \sim p, y, y' \sim q$	I:$k(x,y)$ is a simple kernal.
MMD-D	$\sqrt{\mathbb{E}\left[k_w(x,x') + k_w(y,y') - 2k_w(x,y)\right]}$, $x, x \sim p, y, y' \sim q$	I:$k_w(x,y) = [(1-\epsilon)k_1(\phi_w(x),\phi_w(y)) + \epsilon]k_2(x,y)$ II:ϕ_w is deep network with parameters w III:$k_1(x,y) = \exp(-\|x-y\|^2/\gamma_1^2)$ IV:$k_2(x,y) = \exp(-\|x-y\|^2/\gamma_2^2)$
H-Div	$\phi(\epsilon_u(h_u^*) - \epsilon_p(h_p^*), \epsilon_u(h_u^*) - \epsilon_q(h_q^*))$	I:$h_u^* \in \arg\min_{h\in\mathcal{H}} \epsilon_u(h)$ II:$h_q^* \in \arg\min_{h\in\mathcal{H}} \epsilon_q(h)$ III:$\epsilon_u(h) = \mathbb{E}_{x\sim u} l(h(x), a(x))$ IV:$\epsilon_q(h) = \mathbb{E}_{x\sim q} l(h(x), a(x))$ V:$\phi(\theta,\lambda) = \frac{\theta+\lambda}{2}$ or $\max(\theta,\lambda)$
R-Div	$\lvert\epsilon_p(h_u^*) - \epsilon_q(h_u^*)\rvert$	

C Compared Methods

Mean embedding (ME) [11] and smooth characteristic functions (SCF) [30] are the state-of-the-art methods using differences in Gaussian mean embeddings at a set of optimized points and frequencies, respectively. Classifier two-sample tests, including C2STS-S [43] and C2ST-L [10], apply the classification accuracy of a binary classifier to distinguish between the two distributions. The binary classifier treats samples from one dataset as positive and the other dataset as negative. These two methods assume that the binary classifier cannot distinguish these two kinds of samples if their distributions are identical. Differently, C2STS-S and C2ST-L apply the test error and the test error gap to evaluate the discrepancy, respectively. MMD-O [20] measures the maximum mean discrepancy (MMD) with a Gaussian Kernel [19], and MMD-D [41] improves the performance of MMD-O by replacing the Gaussian Kernel with a learnable deep kernel. H-Divergence (H-Div) [62] learns optimal hypotheses for the mixture distribution and each individual distribution for the specific model, assuming that the expected risk of training data on the mixture distribution is higher than that on each individual distribution if the two distributions are identical. The equations of these compared methods are presented in Figure 7.

D Additional Experimental Results

D.1 Benchmark Dataset

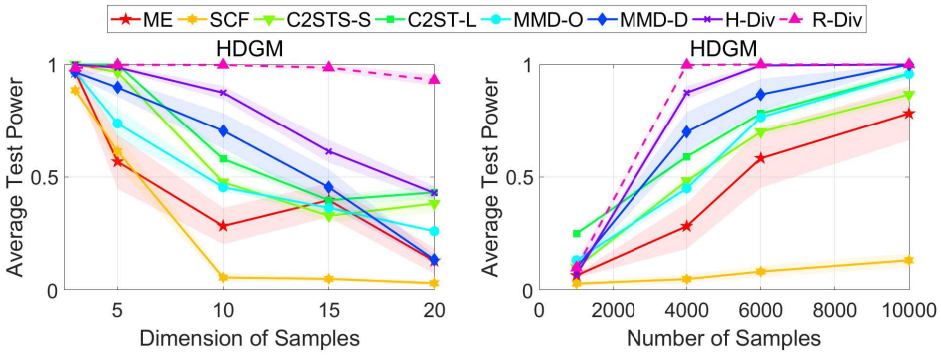


Figure 3: The average test power on HDGM with the significant level $\alpha = 0.05$. Left panel: results with the same sample size (4,000) and different feature dimensions. Right panel: results with the same feature dimensions (10) and different sample sizes.

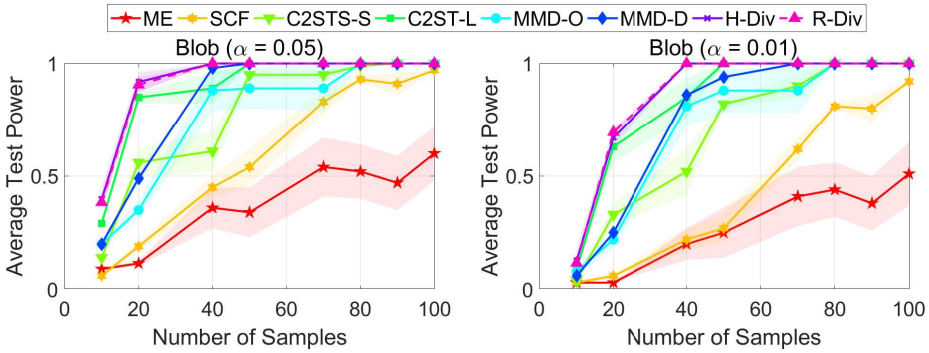


Figure 4: The average test power on Blob at the significant levels $\alpha = 0.05$ and $\alpha = 0.01$.

Table 8: The average test power $\pm$ standard error with the significant level $\alpha = 0.05$ on MNIST. N represents the number of samples in each given dataset, and boldface values represent the relatively better discrepancy estimation.

N	200	400	600	800	1000	Avg.
ME	0.414±0.050	0.921±0.032	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**	0.867
SCF	0.107±0.018	0.152±0.021	0.294±0.008	0.317±0.017	0.346±0.019	0.243
C2STS-S	0.193±0.037	0.646±0.039	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**	0.768
C2ST-L	0.234±0.031	0.706±0.047	0.977±0.012	**1.000±0.000**	**1.000±0.000**	0.783
MMD-O	0.188±0.010	0.363±0.017	0.619±0.021	0.797±0.015	0.894±0.016	0.572
MMD-D	0.555±0.044	0.996±0.004	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**	0.910
H-Div	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**	**1.000**
R-Div	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**	**1.000**

D.2 PACS Dataset

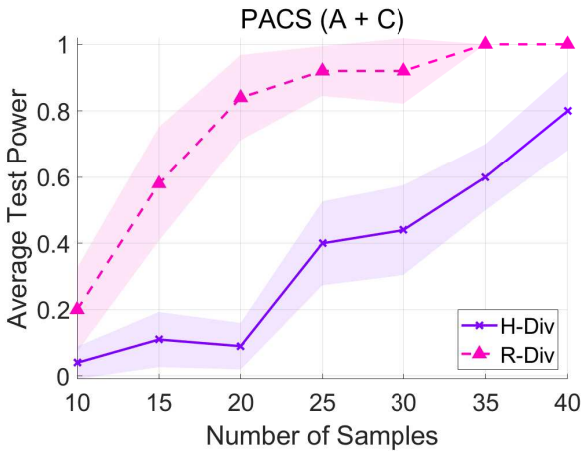
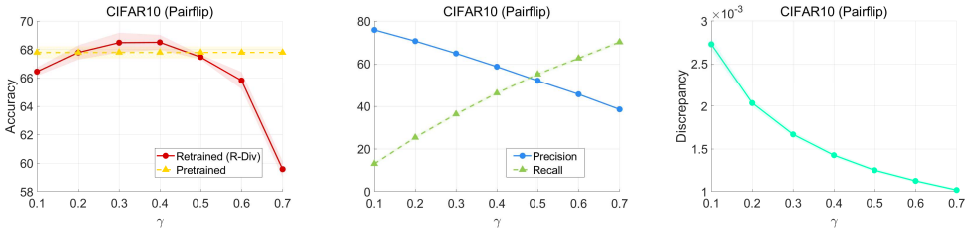


Figure 5: The average test power at the significant level $\alpha = 0.05$ on the art painting and cartoon domains of PACS.

D.3 Learning with Noisy Labels



(a) Accuracy (b) Detection (c) Discrepancy

Figure 6: Results on CIFAR10 with pair flipping. All values are averaged over five trials. **Left:** Classification accuracy of pretrained and retrained networks. **Middle:** Precision and recall rates of detecting clean and noisy samples. **Right:** Discrepancy between predicted clean and noisy samples.