Continual Domain Adversarial Adaptation via Double-Head Discriminators

AISTATS 2024

Yan Shen, Zhanghexuan Ji, Chunwei Ma, Mingchen Gao

Background: Domain Shifts

Learner:
$h_\theta(x) \rightarrow y$

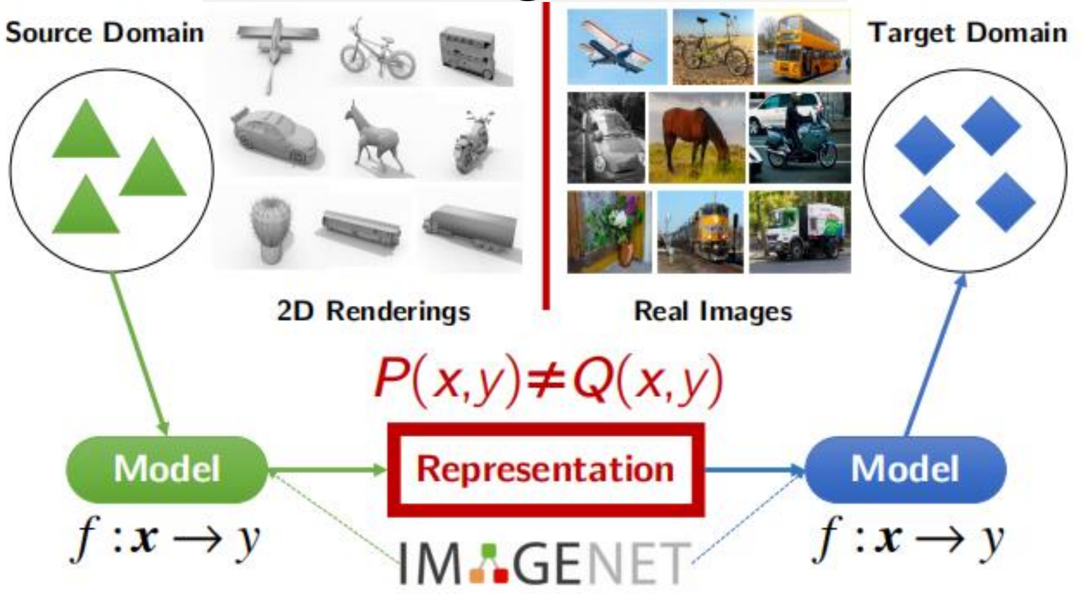
minimize

Task Risks
$\epsilon_P(h) = \mathbb{E}_{(x,y)\sim P}[l(h_\theta(x), y)]$

Source Distribution:
$\hat{S} = \{(x_i^s, y_i^s)\}_{i=1}^n$

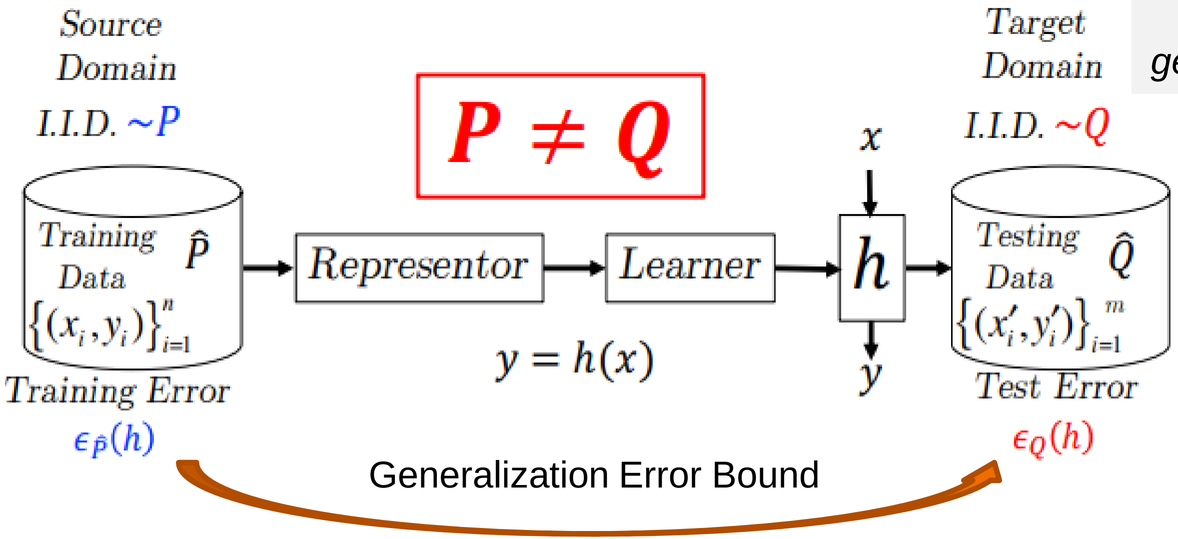
Target Distribution:
$\hat{T} = \{(x_i^t, y_i^t)\}_{i=1}^n$

Challenges:



Background: Transfer Learning

- Machine Learning across domains of different distributions $P \neq Q$
- Shifted Domains are Independently and Differently Distributed

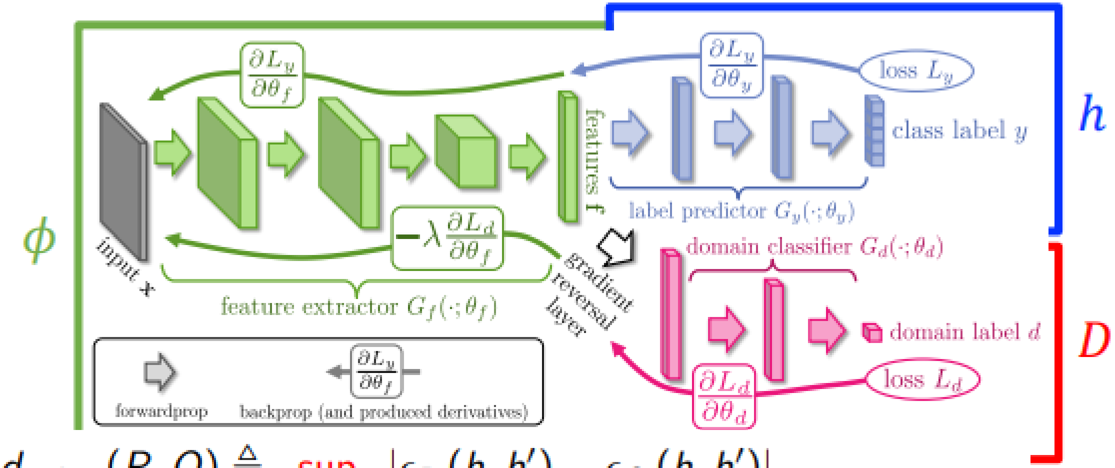


Challenges:

How to effective bound the generalization error on target domains

Background: Existing Methods

- Adversarial Learning



$$d_{\mathcal{H}\Delta\mathcal{H}}(P,Q) \triangleq \sup_{h,h'\in\mathcal{H}} \left|\epsilon_P(h,h') - \epsilon_Q(h,h')\right|$$

$$= \sup_{\delta\in\mathcal{H}\Delta\mathcal{H}} \left|\mathbb{E}_P[\delta(\mathbf{x})\neq 0] - \mathbb{E}_Q[\delta(\mathbf{x})\neq 0]\right|$$

$$\leqslant \sup_{D\in\mathcal{H}_D} \left|\mathbb{E}_P[D(\mathbf{x})=1] + \mathbb{E}_Q[D(\mathbf{x})=0]\right|$$

Theorem

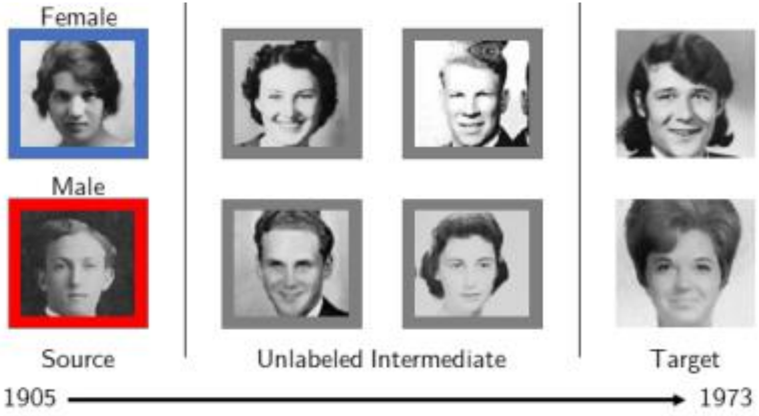
Let $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}\times\mathcal{Y}}$ be a hypothesis set with $\mathcal{Y} = \{1,2\ \ldots k\}$ and $\mathcal{H} \subseteq \mathcal{Y}$ be the corresponding
Y-valued classifier class. For every scoring function $f \in \mathcal{F}$,

$$err_Q(f) \leq err_P(f) + d_{\mathcal{H}\Delta\mathcal{H}}(P,Q) + \lambda$$

$$\lambda = \min_{f^*\in\mathcal{H}}\{err_P(f^*) + err_Q(f^*)\}$$

Our work: Continual Domain Adaptation

❑ Challenges of Continual Domain Adaptation



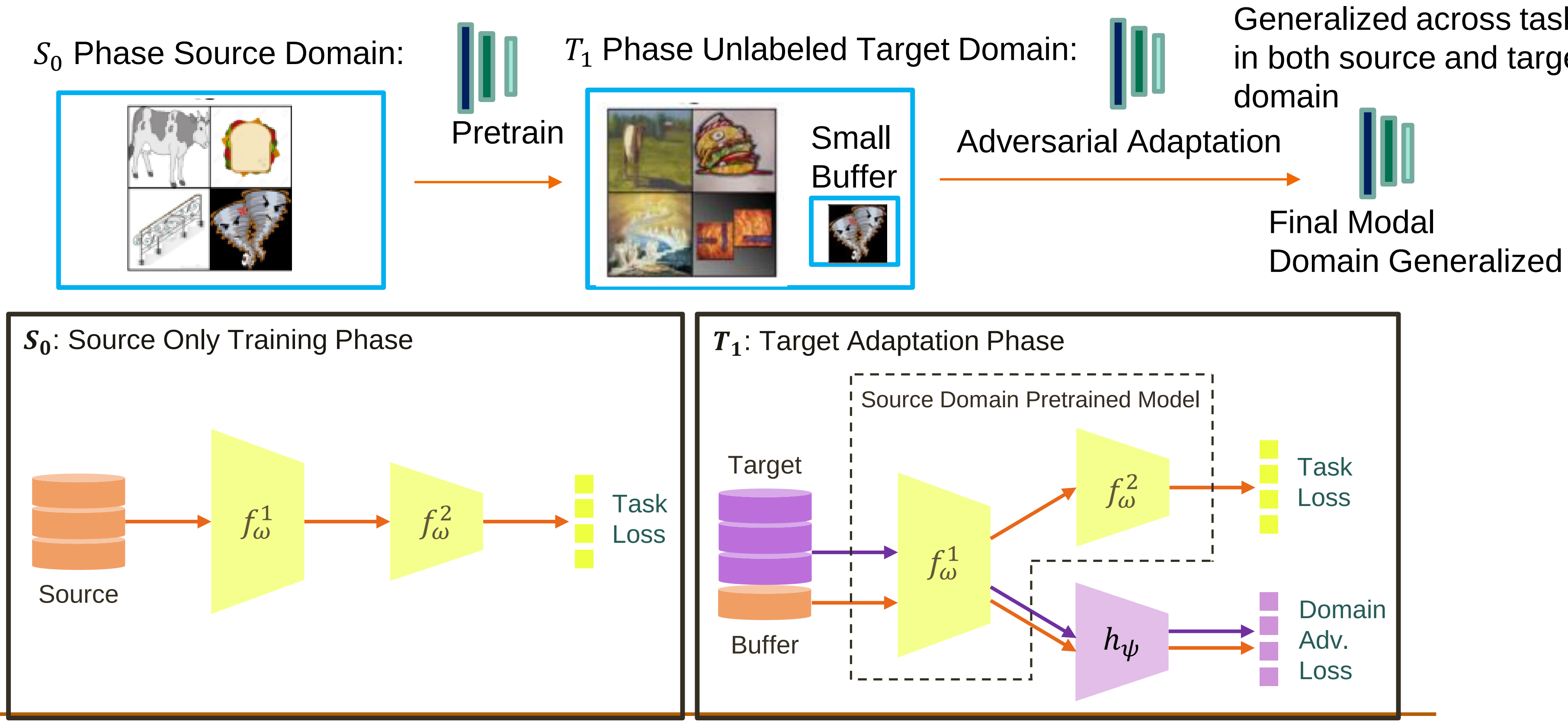
Challenges

The domain data exists in an sequential form. Only online data is accessible. Sequential Learning would result catastrophic forgetting phenomenon.

- Limitations of previous research:
  - Self-supervised learning methods requires intermediate domain to be close enough
  - Unlike supervised continual learning, buffering a small set of previous samples works poorly
  - Unsupervised Learning only on current data would cause catastrophic forgetting

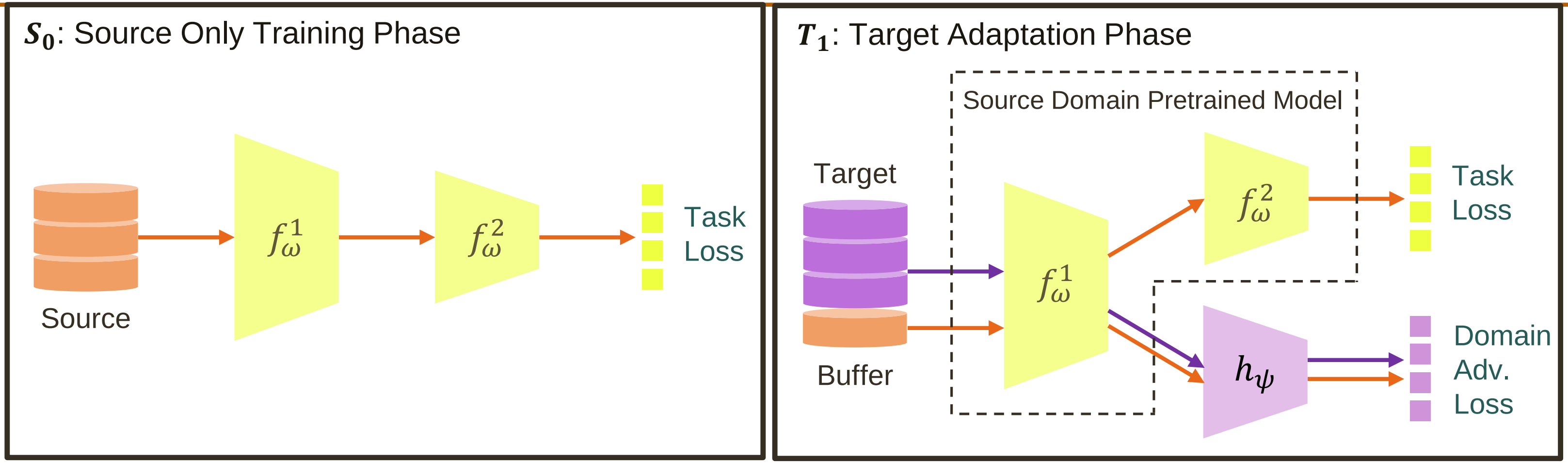
Yan Shen, et al. " Continual Domain Adversarial Adaptation via Double-Head Discriminators." AISTATS. 2024.

Our work: Problem Settings



Yan Shen, et al. " Continual Domain Adversarial Adaptation via Double-Head Discriminators." AISTATS. 2024.

Our work: Unique Challenge in Continual Adv Adaptation



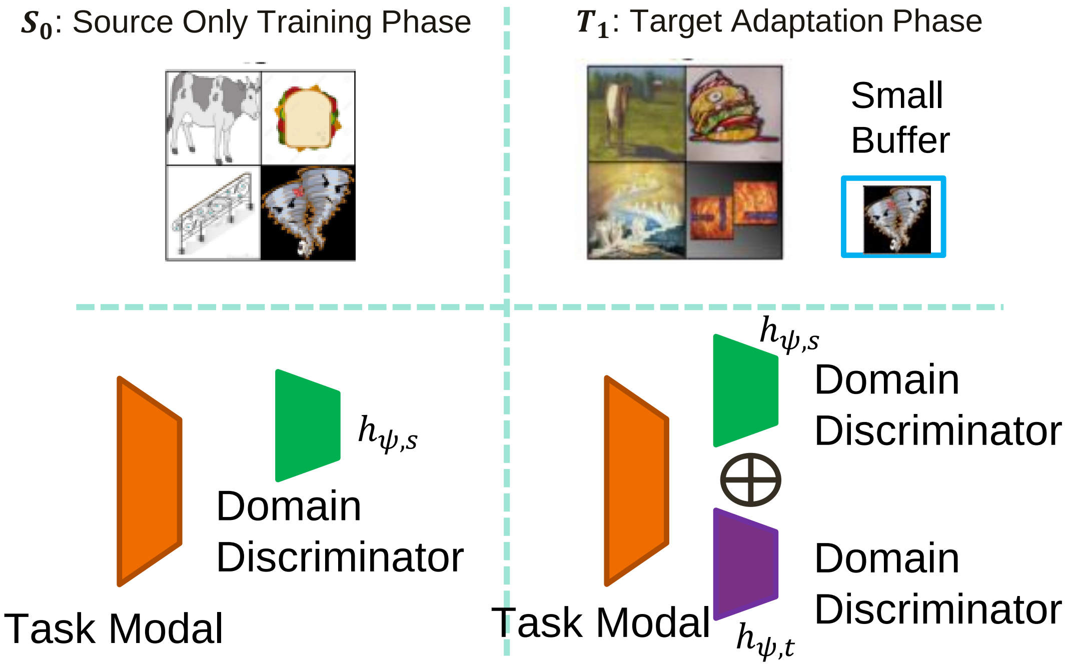
Theorem

Let $\mathcal{F}$ *be a hypothesis space with VC dimensions d, if* $\mathcal{S}'$ *are samples of size m from* $\mathcal{S}$ *and* $\mathcal{T}'$ *be samples of size n from* $\mathcal{T}$ *respectively and* $\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}',\mathcal{T}')$ *is the empirical* $\mathcal{H}$*-divergence between samples, then for any* $\delta \in (0,1)$ *,with probability at least* $1-\delta$

$$d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}',\mathcal{T}') \leq \hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}',\mathcal{T}') + 2\sqrt{\frac{dlog2m + \log\left(\frac{2}{\delta}\right)}{m}} + 2\sqrt{\frac{dlog2n + \log\left(\frac{2}{\delta}\right)}{n}}$$

Yan Shen, et al. " Continual Domain Adversarial Adaptation via Double-Head Discriminators." AISTATS. 2024.

Our work: Double-Head Continual Adv Adaptation

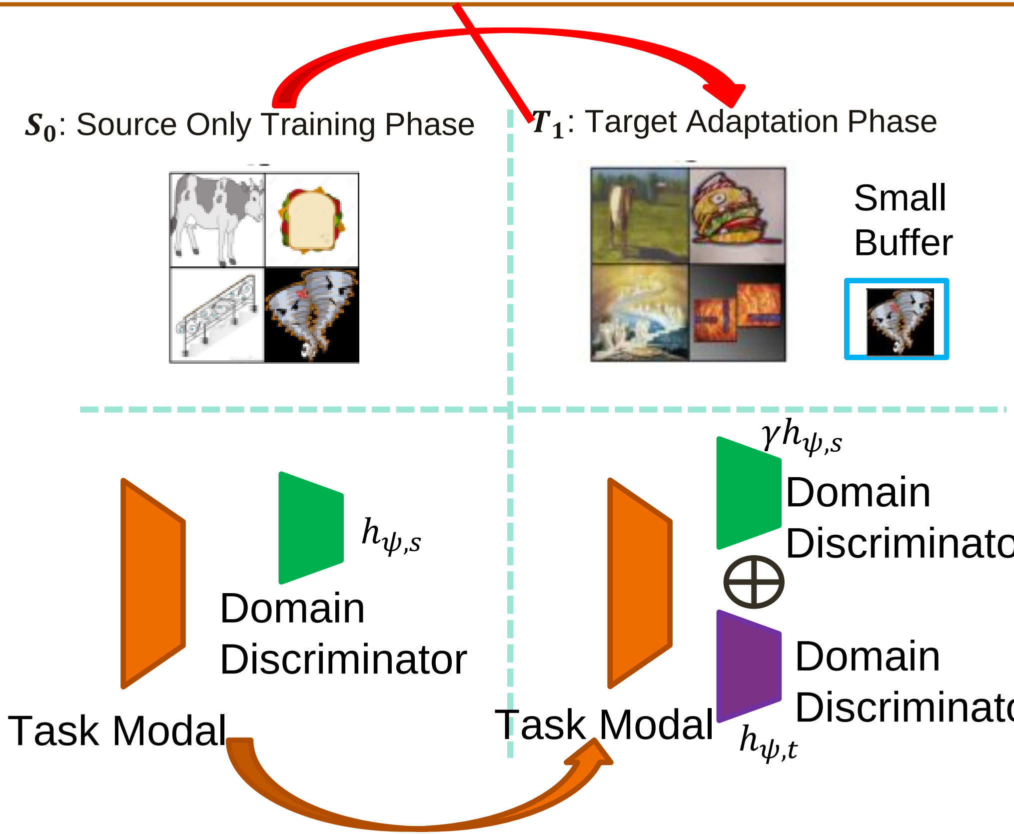


Intuitive Idea

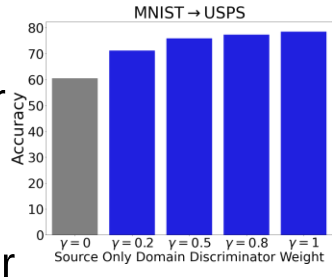
Using two domain discriminators, one $h_{\psi,s}$ is trained at $\boldsymbol{S_0}$ and the other $h_{\psi,t}$ is trained at $\boldsymbol{T_1}$

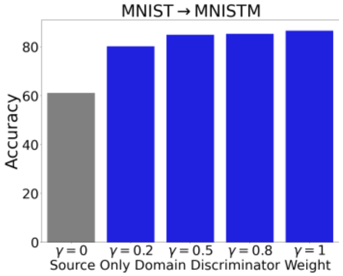
Yan Shen, et al. " Continual Domain Adversarial Adaptation via Double-Head Discriminators." AISTATS. 2024.

Our work: Double-Head Continual Adv Adaptation



- ❑ **Using two domain discriminators is better than one domain discriminator for continual adversarial adaptation**

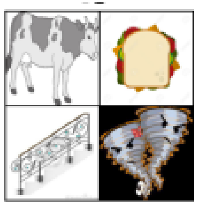


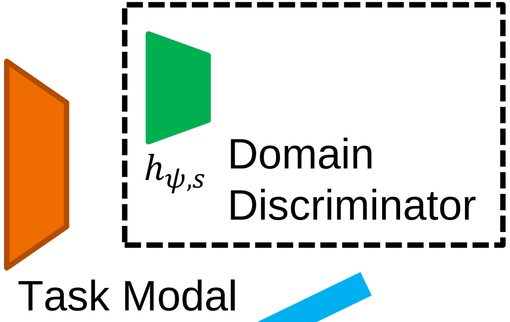


Yan Shen, et al. " Continual Domain Adversarial Adaptation via Double-Head Discriminators." AISTATS. 2024.

Our work: Single Domain Discriminator Learning on S_0

$\boldsymbol{S_0}$: Source Only Training Phase



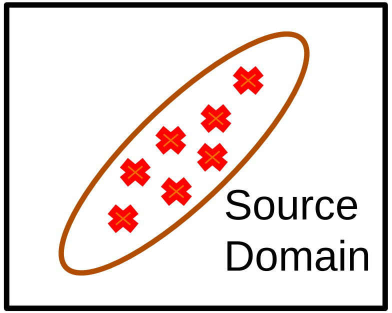


$h_{\psi,s}$ *Learning Objectives:*

$$\widehat{d}_{\mathcal{H}\Delta\mathcal{H}} \triangleq \boxed{\mathbb{E}_{x_i^s\in S_0} D\left(\delta\left(h_{\psi,s}\left(f_\omega^1(x_i^s)\right)\right)\right)} - \mathbb{E}_{x_i^s\notin S_0} D\left(\delta\left(h_{\psi,s}\left(f_\omega^1(x_i^s)\right)\right)\right)$$

Only Source Domain data
$x_i^s \in S_0$ is accessible in $\boldsymbol{S_0}$

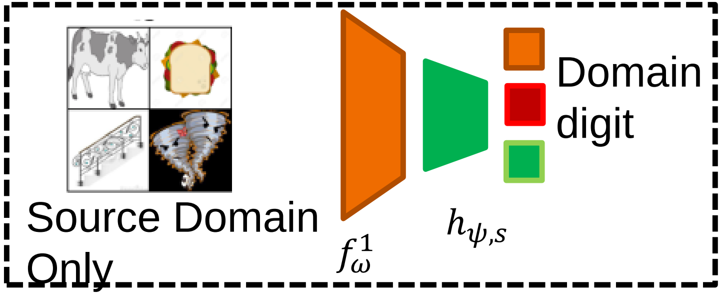
$h_{\psi,s}$ *is a source-only domain discriminator that is trained to determine how possible a data lies in source domain*



Yan Shen, et al. " Continual Domain Adversarial Adaptation via Double-Head Discriminators." AISTATS. 2024.

Our work: Single Domain Discriminator Learning on S_0

- **One-Class Learning on Source-only domain discriminator**
- **H-Regularization Loss in Binary domain digit**
- **MDD Loss in Multi-class domain digit**



MDD Learning Objectives $h_{\psi,s}(\cdot)$ *is a vector output function :*

$$\widehat{d_{\mathcal{H}\Delta\mathcal{H}}} \triangleq \mathbb{E}_{x_i^s \in S_0} softmax\left(h_{\psi,s}\left(f_\omega^1(x_i^s), \mathrm{argmax}_c f_\omega^2 \cdot\right.\right.$$

H-Regularization Learning Objectives $h_{\psi,s}(\cdot)$ *is scalar output function :*

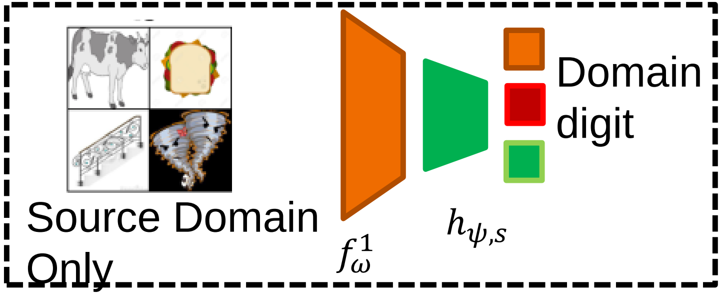
$$\widehat{d_{\mathcal{H}\Delta\mathcal{H}}} \triangleq \mathbb{E}_{x_i^s \in S_0} sigmoid\left(h_{\psi,s}\left(f_\omega^1(x_i^s)\right)\right) + \lambda ||\nabla_{\psi,s}\, h_{\psi,s}\left(f_\omega^1(x_i^s)\right)||_2^n$$

Yan Shen, et al. " Continual Domain Adversarial Adaptation via Double-Head Discriminators." AISTATS. 2024.

Our work: Single Domain Discriminator Learning on S_0

- ❑ **One-Class Learning on Source-only domain discriminator**
  - **H-Regularization Loss in Binary domain digit**
  - **MDD Loss in Multi-class domain digit**

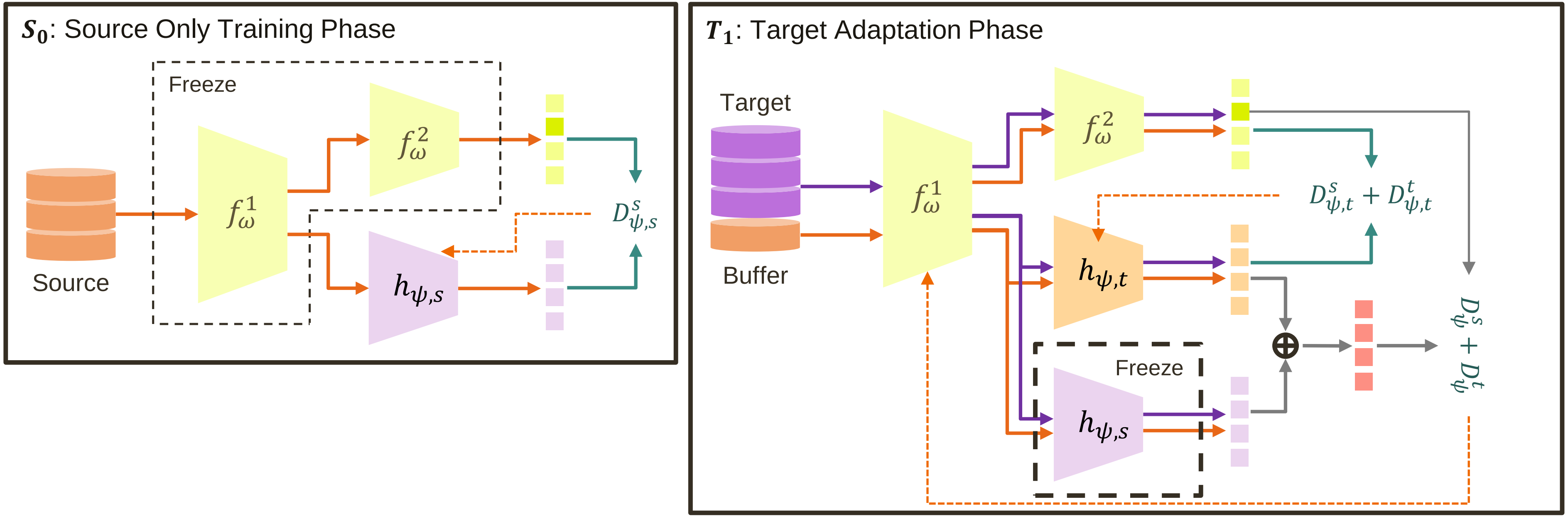
MDD is better than H-Reg



$h_{\psi,s}$ Learning	MDD	H-Reg DANN	H-Reg CDAN
$MNIST \rightarrow USPS$	**78.1**	69.1	73.4
$MNIST \rightarrow MNISTM$	**87.3**	78.1	80.3
$MNIST \rightarrow SVHN$	**45.8**	37.5	40.8

Yan Shen, et al. " Continual Domain Adversarial Adaptation via Double-Head Discriminators." AISTATS. 2024.

Our work: Our Algorithm



Yan Shen, et al. " Continual Domain Adversarial Adaptation via Double-Head Discriminators." AISTATS. 2024.

Our work: Our Algorithm

Phase 1 – Source Only training phase

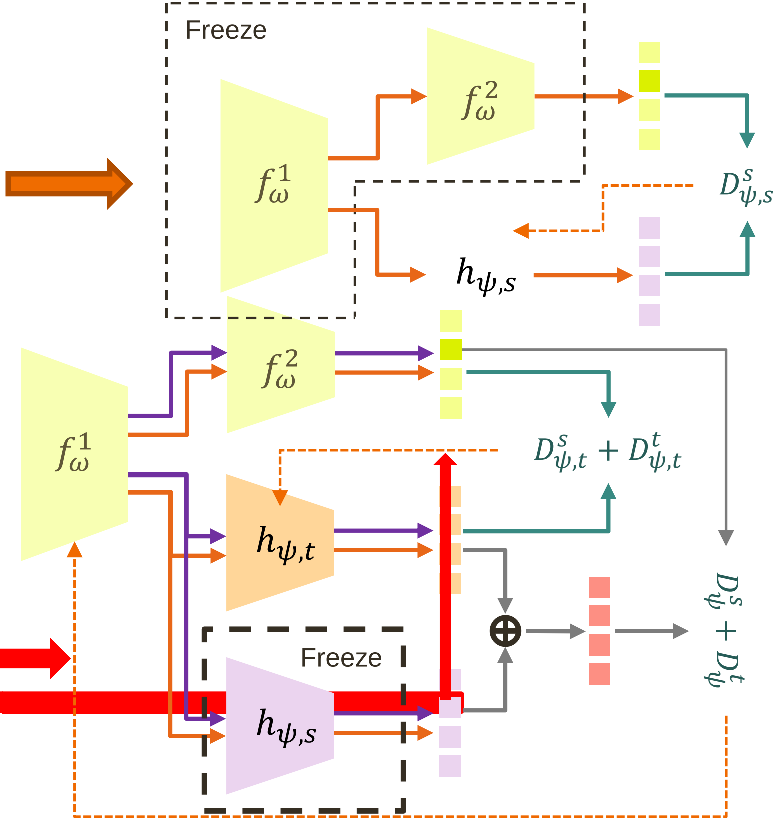
---

procedure Task Model Training Phase ▷ Train Task Model on Source Domain
procedure Source Only Domain Classifier Training Phase
  for $t \in \{1, \dots, t_2\}$ do
    for $(x_1, \dots, x_K) \overset{K}{\sim} S_0$ do
      $d'(x) \rightarrow \arg\max_c f_\omega(x_i, c) \quad \forall x \in \{x_1 \dots x_K\}$ ▷ Get Pseudo Domain Label from Task Model
      $D = \frac{1}{K}\sum_{i=i}^{K} -\log(\mathrm{softmax}(h_{\psi,s}(f_\omega^1(x_i)), d'(x_i)))$
      $\psi_s \rightarrow SGD(D, \psi_s)$ ▷ Train on Source Only Domain Classifier
    end for
  end for
end procedure

Phase 3 – Unlabeled Target Adaptation Phase with Memory Reply

---

Initialization:Target Adaptation Phase multi-class Domain Classifier $h_{\psi,t}$
procedure Target Phase
  for $t \in \{1, \dots, t_3\}$ do
    for $\{(x_1^s, y_1^s), \dots, (x_K^s, y_K^s)\} \overset{K}{\sim} \mathcal{M}$, $(x_1^t \dots, x_K^t) \overset{K}{\sim} T_1$ do
      $L = \frac{1}{K}\sum_{i=1}^{K} \ell(f_\omega(x_i^s), y_i^s)$
      $d'(x) \rightarrow \arg\max_c f_\omega(x, c) \quad \forall x \in \{x_1^s \dots x_K^s, x_1^t \dots x_K^t\}$
      $D_{\psi,t} = \frac{1}{K}\sum_{i=1}^{K} -\log(\mathrm{softmax}(h_{\psi,t}(f_\omega^1(x_i^s)), d'(x_i^s))) - \log(1 - \mathrm{softmax}(h_{\psi,t}(f_\omega^1(x_i^t)), d'(x_i^t)))$
      $D_\psi = \frac{1}{K}\sum_{i=1}^{K} -\log(\mathrm{softmax}(h_{\psi,s}(f_\omega^1(x_i^s)) + h_{\psi,t}(f_\omega^1(x_i^s)), d'(x_i^s))) - \log(1 - \mathrm{softmax}(h_{\psi,s}(f_\omega^1(x_i^t)) + h_{\psi,t}(f_\omega^1(x_i^t)), d'(x_i^t)))$
      $\omega \rightarrow SGD(L - \beta D_\psi, \omega)$
      $\psi_t \rightarrow SGD(D_{\psi,t}, \psi_t)$
    end for
  end for
end procedure



Yan Shen, et al. " Continual Domain Adversarial Adaptation via Double-Head Discriminators." AISTATS. 2024.

Our work: Theoretical Analysis of Our Algorithm

Theorem

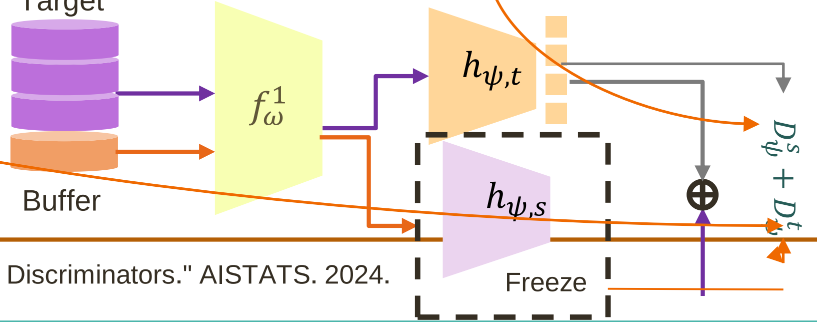
Let $f_0 \in \mathcal{F}$ *be a fixed hypothesis space that maps from* $\mathcal{X} \times \mathcal{Y} \rightarrow R$ *which satisfies that* $\rho_{f_0}(x^s, h_f) \geq \epsilon_s$ *for source domain data* x^s *and* $\rho_{f_0}(x^t, h_f) \leq \epsilon_t$ *for target domain data* x^t*. if* $x_i^s \in S$ *are i.i.d samples of size m from* $\mathcal{S}$ *and* $x_i^t \in T$ *be samples of size n from* $\mathcal{T}$ *respectively, then for any* $\delta \in (0,1)$ *with probability at least* $1 - 2\delta$, we have the following generalization error bound for $\mathcal{H}$-divergence based adversarial loss function

$$\mathbb{E}_{\mathbf{x}^s \in \mathcal{S}}[\log(\frac{e^{\rho_{f'}(\mathbf{x}^s,h_f)+\rho_{f_0}(\mathbf{x}^s,h_f)}}{1+e^{\rho_{f'}(\mathbf{x}^s,h_f)+\rho_{f'}(\mathbf{x}^s,h_f)}})] + \mathbb{E}_{\mathbf{x}^t \in \mathcal{T}}[\log(\frac{1}{1+e^{\rho_{f'}(\mathbf{x}^t,h_f)+\rho_{f_0}(\mathbf{x}^t,h_f)}})]$$

$$\leq \frac{1}{m}\sum_{i=1}^{m}\log(\frac{e^{\rho_{f'}(\mathbf{x}_i^s,h_f)+\rho_{f_0}(\mathbf{x}_i^s,h_f)}}{1+e^{\rho_{f'}(\mathbf{x}_i^s,h_f)+\rho_{f'}(\mathbf{x}_i^s,h_f)}}) + \frac{1}{n}\sum_{i=1}^{n}\log(\frac{1}{1+e^{\rho_{f'}(\mathbf{x}_i^t,h_f)+\rho_{f'}(\mathbf{x}_i^t,h_f)}})$$

$$+\max\{\frac{2}{(e^{\epsilon_s}-1)\lambda_s^+ + 1}, \frac{2}{(e^{\epsilon_s}-1)\lambda_s^- + 1}\}\Re_{m,\mathcal{D}_s}(\mathcal{G}_s) + \max\{\frac{2e^{\epsilon_t}}{(1-\lambda_t^+)e^{\epsilon_t}+\lambda_t^+}, \frac{2e^{\epsilon_t}}{(1-\lambda_t^-)e^{\epsilon_t}+\lambda_t^-}\}\Re_{n,\mathcal{D}_t}(\mathcal{G}_t) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}} + \sqrt{\frac{\log\frac{1}{\delta}}{2n}}$$

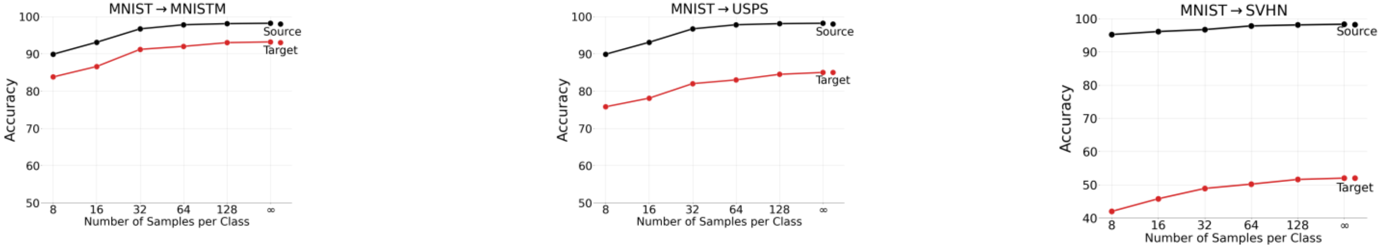
Error reduced Empirical Estimation of $\mathcal{H}$-divergence based adversarial loss from source domain side



Yan Shen, et al. " Continual Domain Adversarial Adaptation via Double-Head Discriminators." AISTATS. 2024.

Our work: Experiment Results

❑Ablation study



Our proposed algorithm entails minimal performance loss from smaller buffer size

A smaller learning rate with mild epochs leads to better performance of source-only domain discriminator training

Yan Shen, et al. " Continual Domain Adversarial Adaptation via Double-Head Discriminators." AISTATS. 2024.

Our work: Experiment Results

❑Comparing with existing continual domain adaptation methods on Office-31

Methods	Office-31 Target Domain Adaptations $A \to W$	$D \to W$	$W \to D$	$A \to D$	$D \to A$	$W \to A$
NLL-OT(Asano et al., 2019)	85.5	95.1	98.7	88.8	64.6	66.7
NLL-KL(Zhang et al., 2021)	86.8	94.8	98.7	89.4	65.1	67.1
HD-SHOT(Liang et al., 2020)	83.1	95.1	98.1	86.5	66.1	68.9
SD-SHOT(Liang et al., 2020)	83.7	95.3	97.1	89.2	67.9	71.1
DINE(Liang et al., 2022)	86.8	96.2	98.6	91.6	72.2	73.3
Ours	92.6	97.3	99.2	92.0	73.9	73.8
Ours+KD	**93.8**	**98.4**	**100.0**	**93.8**	**74.0**	**75.6**
Ours+SL	93.2	97.7	100.0	92.5	73.9	74.4
i.i.d-adv	94.5	98.4	100.0	93.5	74.6	74.2

Table 3: Office-31 Target Domain Adaptation

Methods	Office-31 Source Domain Forgetting $A \to W$	$D \to W$	$W \to D$	$A \to D$	$D \to A$	$W \to A$
NLL-OT(Asano et al., 2019)	4.53	3.14	2.73	4.30	6.17	5.11
NLL-KL(Zhang et al., 2021)	4.37	2.99	2.48	4.02	5.94	4.99
HD-SHOT(Liang et al., 2020)	5.12	4.01	3.98	4.87	7.80	5.56
SD-SHOT(Liang et al., 2020)	5.31	4.54	4.03	4.85	7.88	5.72
DINE(Liang et al., 2022)	3.81	2.16	1.50	3.32	5.08	3.98
Ours	**1.97**	**1.03**	**0.98**	**1.55**	**3.72**	**2.96**

Table 4: Office-31 Source Domain Forgetting

- With a final stage of SSL fine-tuning, our proposed methods achieve over 2% performance increase over these strong baselines
- By employing continual adversarial adaptation methods, we effectively addressed catastrophic forgetting by learning a domain generalized model

Yan Shen, et al. " Continual Domain Adversarial Adaptation via Double-Head Discriminators." AISTATS. 2024.

Our work: Experiment Results

❑Comparing with existing continual domain adaptation methods on Office-home

	Office-home Target Domain Adaptations											
Methods	$Ar \to Cl$	$Ar \to Pr$	$Ar \to Re$	$Cl \to Ar$	$Cl \to Pr$	$Cl \to Re$	$Pr \to Ar$	$Pr \to Cl$	$Pr \to Re$	$Re \to Ar$	$Re \to Cl$	$Re \to Pr$
NLL-OT(Asano et al., 2019)	49.1	71.7	77.3	60.2	68.7	73.1	57.0	46.5	76.8	67.0	52.3	79.5
NLL-KL(Zhang et al., 2021)	49.0	71.5	77.1	59.0	68.7	72.9	56.4	46.9	76.6	66.2	52.3	79.1
HD-SHOT(Liang et al., 2020)	48.6	72.8	77.0	60.7	70.0	73.2	56.6	47.0	76.7	67.5	52.6	80.2
SD-SHOT(Liang et al., 2020)	50.1	75.0	78.8	63.2	72.9	76.4	60.0	48.0	79.4	69.2	54.2	81.6
DINE(Liang et al., 2022)	52.2	78.4	81.3	65.3	76.6	78.7	62.7	49.6	82.2	69.8	55.8	84.2
Ours	53.8	78.8	81.9	66.4	77.8	77.9	63.0	52.9	83.2	72.0	59.4	84.9
Ours+KD	**54.8**	**81.1**	**84.0**	**67.5**	**79.0**	**80.5**	**65.1**	**53.8**	**84.5**	**73.2**	**60.0**	**86.7**
Ours+SL	54.0	79.2	82.4	66.8	78.3	79.0	63.7	53.2	83.2	72.8	59.4	85.8
i.i.d-adv	54.9	79.0	82.8	67.0	78.7	78.1	63.6	54.2	83.8	72.9	60.8	85.8

Table 1: Comparison of Target Domain Adaptation Performance on Office-home.

	Office-home Source Domain Forgetting											
Methods	$Ar \to Cl$	$Ar \to Pr$	$Ar \to Re$	$Cl \to Ar$	$Cl \to Pr$	$Cl \to Re$	$Pr \to Ar$	$Pr \to Cl$	$Pr \to Re$	$Re \to Ar$	$Re \to Cl$	$Re \to Pr$
NLL-OT(Asano et al., 2019)	10.91	7.64	7.31	12.73	13.18	11.13	7.29	7.72	6.19	7.07	7.28	5.35
NLL-KL(Zhang et al., 2021)	10.93	7.66	7.34	13.01	13.05	10.98	7.27	7.50	6.03	6.97	7.26	5.46
HD-SHOT(Liang et al., 2020)	11.10	9.69	8.06	14.99	15.02	12.06	7.57	7.86	6.58	7.22	7.92	6.02
SD-SHOT(Liang et al., 2020)	11.21	8.93	7.89	15.24	15.55	12.25	7.75	7.93	6.72	7.22	8.13	6.05
DINE(Liang et al., 2022)	9.67	6.66	6.26	9.29	10.02	9.76	6.13	5.92	5.82	6.19	6.05	4.93
Ours	**4.52**	**3.95**	**3.53**	**5.12**	**4.83**	**4.69**	**1.93**	**2.05**	**1.89**	**2.12**	**3.13**	**1.43**

Table 2: Comparison of Source Domain Forgetting Performance on Office-home.

Yan Shen, et al. " Continual Domain Adversarial Adaptation via Double-Head Discriminators." AISTATS. 2024.

Thank you for listening!

Q & A