Seminar

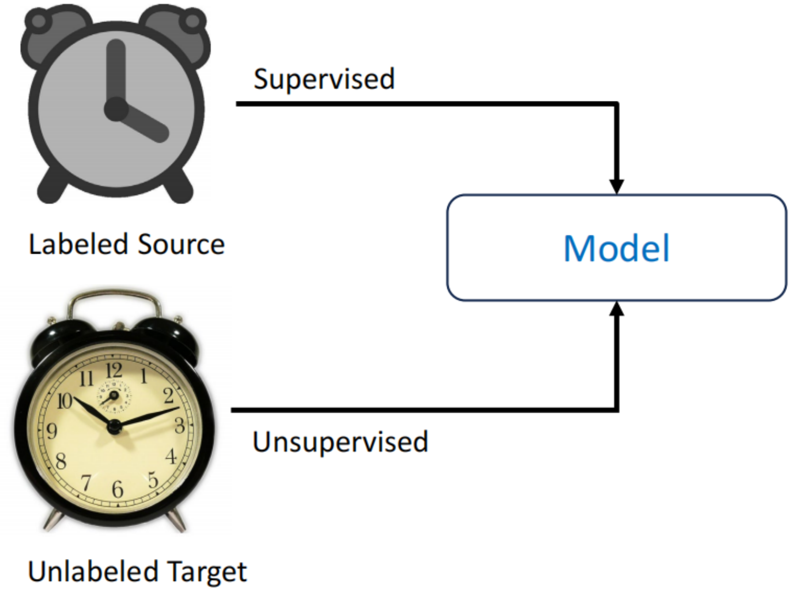
朱顺尧

2025.8.8

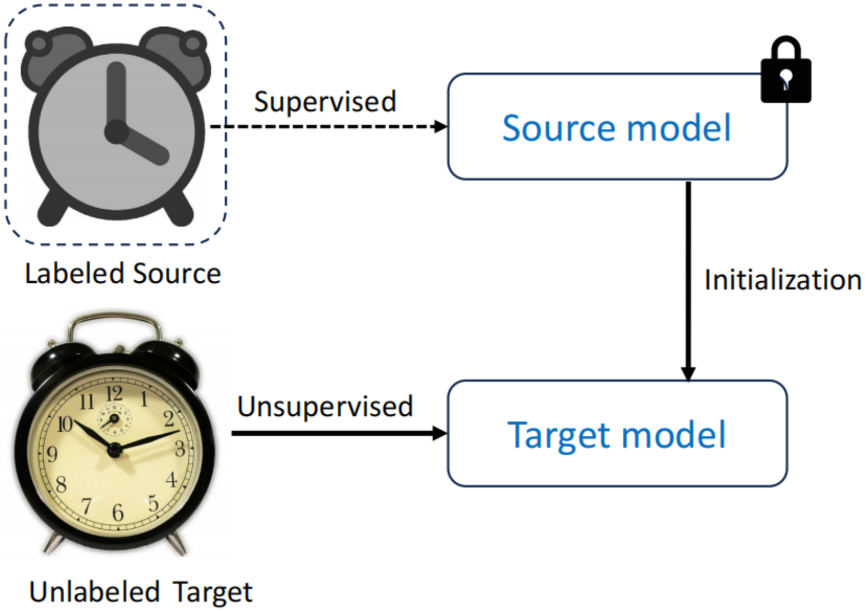
Setting

Domain Adaptation: address distribution shifts across datasets/scenarios

☐ Unsupervised Domain Adaptation (UDA)



☐ Source-Free Domain Adaptation (SFDA)



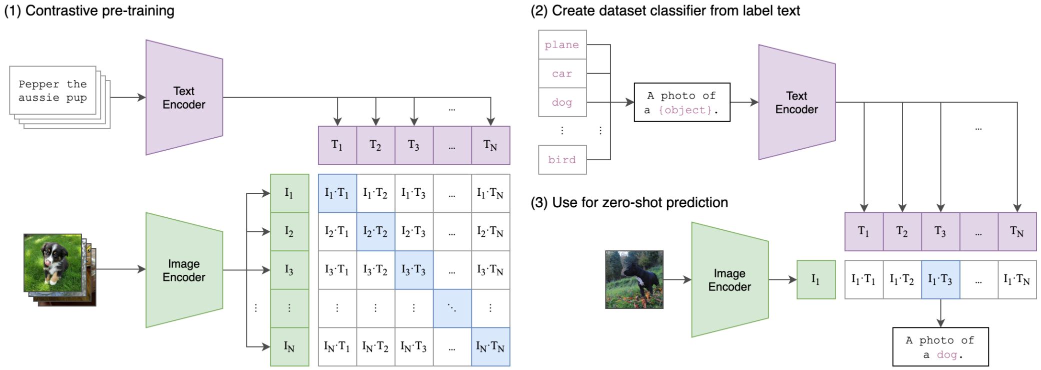
Preserving Clusters in Prompt Learning for Unsupervised Domain Adaptation

Tung-Long Vuong[1], Hoang Phan[2], Vy Vo[1], Anh Bui[1], Thanh-Toan Do[1], Trung Le[1], Dinh Phung[1]
[1]Monash University, [2]New York University
{Tung-Long.Vuong,v.vo,tuananh.bui,toan.do,trunglm,dinh.phung}@monash.edu
hvp2011@nyu.edu

CVPR2025

Base Method

CLIP



$$P(y|\boldsymbol{x},\boldsymbol{p}_k)=\frac{\exp(\langle f_v(\boldsymbol{x}),f_t(\boldsymbol{p}_k)\rangle/\gamma)}{\sum_{k'=1}^{K}\exp(\langle f_v(\boldsymbol{x}),f_t(\boldsymbol{p}_{k'})\rangle/\gamma)}$$

Prompt Learning

A photo of a $[\mathrm{CLASS}_k]$

$$\boldsymbol{P}_{sh}^{k}=[\boldsymbol{v}_1^k|\boldsymbol{v}_2^k|\cdots|\boldsymbol{v}_{M_1}^k]$$

$$\boldsymbol{P}_{S_i}=[\boldsymbol{u}_1^{S_i}|\boldsymbol{u}_2^{S_i}|\cdots|\boldsymbol{u}_{M_2}^{S_i}]$$

$$\boldsymbol{P}_{T}=[\boldsymbol{u}_1^{T}|\boldsymbol{u}_2^{T}|\cdots|\boldsymbol{u}_{M_2}^{T}]$$

$$p_k=[\boldsymbol{P}_{sh}^{k}][\boldsymbol{P}_{S_i}][\mathrm{CLASS}_k]$$

$$p_k=[\boldsymbol{P}_{sh}^{k}][\boldsymbol{P}_{T}][\mathrm{CLASS}_k]$$

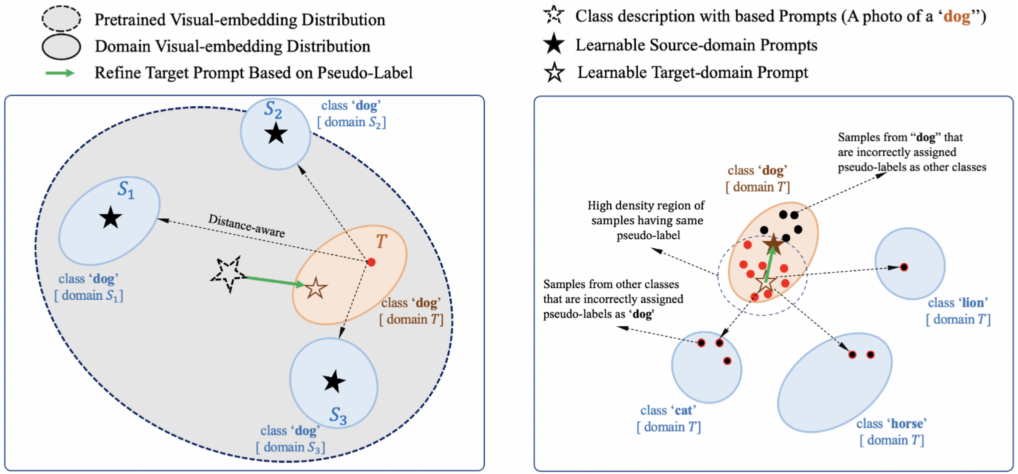
Motivation

Setting	Inference	Ar	Cl	Pr	Rw
	Training on **source** data only				
Zero-shot from	CLIP [34]	71.2	50.4	81.4	82.6
Source-combined	Source prompt	72.2	55.9	82.6	83.3
Multi-source	Average prompt	74.3	57.4	84.5	84.7
	Ar prompt	-	56.0	80.0	82.1
	Cl prompt	70.4	-	81.1	80.4
	Pr prompt	70.2	55.0	-	84.1
	Wr prompt	73.9	55.5	84.5	-
	Supervised training on **target** domain only				
	Target prompt	99.6	91.6	99.3	98.6
	Trainining both source-target domains as Eq. (3)				
	Target prompt	47.6	29.0	53.5	63.5
	Source prompt	74.3	57.2	84.3	84.8

$$[\boldsymbol{P}_T][\text{CLASS}_k]$$

1. Source prompt includes domain-invariant embedding;
2. The transferability varies based on their similarity;
3. The visual embeddings for each class tend to form a single cluster in the embedding space;
4. Due to false pseudo-labels, the trained text embedding acts as a prototype or centroid may fail to represent the entire cluster.

$$\mathcal{L}_{total}(\boldsymbol{P}) = \sum_{i=1}^{N} \mathcal{L}_{S_i}(\boldsymbol{P}_{sh}, \boldsymbol{P}_{S_i}) + \mathcal{L}_T(\boldsymbol{P}_{sh}, \boldsymbol{P}_T)$$



(a) Target prompt refinement.
⟨*Same class from different domains*⟩

(b) Effect of wrong Pseudo-label.
⟨*Different classes from target domain*⟩

Method

Prompt Learning with Source Domains

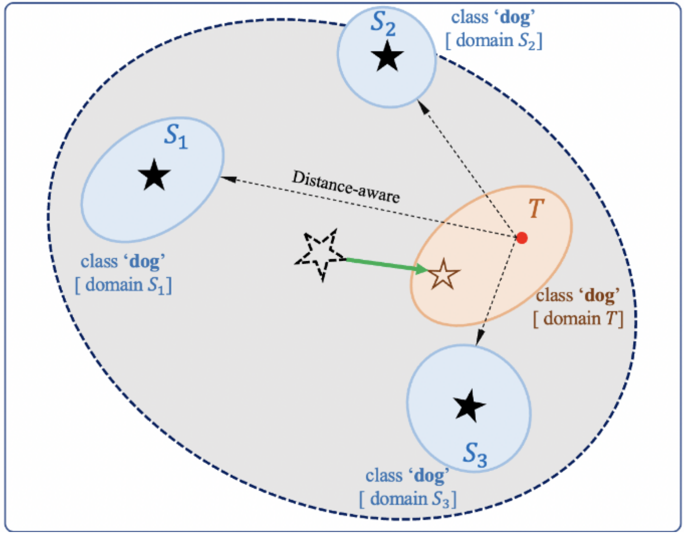
$$\mathcal{L}_{S_i}(\boldsymbol{P}_{sh}, \boldsymbol{P}_{S_i}) = \mathrm{CE}(\boldsymbol{P}_{sh}, \boldsymbol{P}_{S_i}; \boldsymbol{X}_{S_i}, Y_{S_i})$$
$$= -\frac{1}{N_{S_i}} \sum_{j=1}^{N_{S_i}} \log P(y = y_j | \boldsymbol{x}_j, \boldsymbol{P}_{sh}, \boldsymbol{P}_{S_i})$$

$$\mathcal{L}_S = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_{S_i}$$

Method

1. Source prompt includes domain-invariant embedding;
2. The transferability varies based on their similarity

Prompt Learning with Target Domain



$$\hat{y}[k] = \frac{\exp\left(\langle \boldsymbol{z}, \boldsymbol{\tau}^k_{ave}((x))\rangle/\gamma\right)}{\sum_{k'=1}^{K} \exp\left(\langle \boldsymbol{z}, \boldsymbol{\tau}^{k'}_{ave}(x)\rangle/\gamma\right)}$$

$$\boldsymbol{\tau}^k_{ave}(x) = \frac{1}{2}\boldsymbol{\tau}^k_{base} + \frac{w_{k,i}(x)}{2}\sum_{i=1}^{N}\boldsymbol{\tau}^k_{S_i}$$

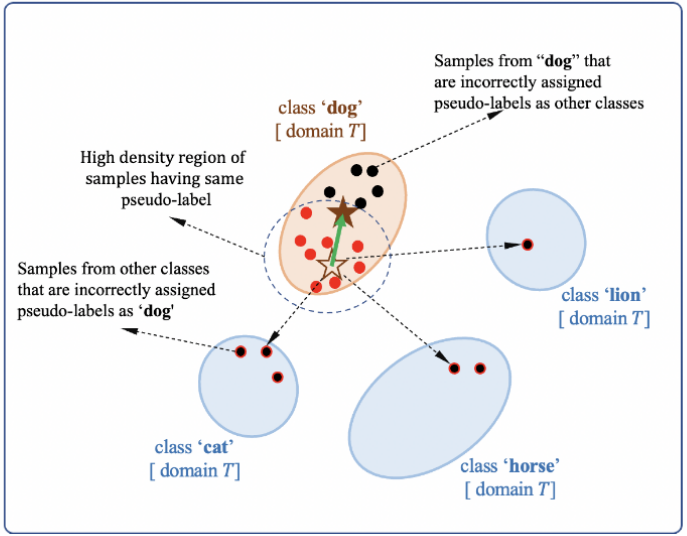
$$w_{i,k}(z) = \frac{\exp\left(\left\|z_{pre} - c^i_k\right\|_2\right)}{\sum_{i'=1}^{N}\exp\left(\left\|z - c^{i'}_k\right\|_2\right)}$$

$$c^i_k = \frac{1}{\sum_{j=1}^{N_{S_i}} \mathbb{I}_{(y^i_j=k)}} \sum_{j=1}^{N_{S_i}} \mathbb{I}_{(y^i_j=k)} \boldsymbol{z}^{pre,i}_j .$$

$$\mathcal{L}_T(\boldsymbol{P}_{sh}, \boldsymbol{P}_T) = \mathrm{CE}_\tau(\boldsymbol{P}_{sh}, \boldsymbol{P}_T; \boldsymbol{X}_T, Y_T)$$
$$= -\frac{1}{N_T}\sum_{j=1}^{N_T}\sum_{k=1}^{K}\hat{y}_j[k]\log P(y=k|\boldsymbol{x}_j, \boldsymbol{P}_{sh}, \boldsymbol{P}_T)$$

Method

Clustering Refined by Optimal Transport



3. The visual embeddings for each class tend to form a single cluster in the embedding space;
4. Due to false pseudo-labels, the trained text embedding acts as a prototype or centroid may fail to represent the entire cluster.

$$\mathcal{L}_{\mathcal{W}} = \mathcal{W}_{d_z}\left(\mathbb{P}_{\tau,\pi}, \mathbb{P}^T\right) \qquad \begin{aligned} \mathbb{P}^T &= \frac{1}{N_T}\sum_{j=1}^{N_T}\delta_{\boldsymbol{z}_j} \\ \mathbb{P}_{\tau,\pi} &= \sum_{k=1}^{K}\pi_k\delta_{\boldsymbol{\tau}_T^k} \end{aligned}$$

$$\min_{\mathcal{T},\pi}\min_{\sigma\in\Sigma_\pi}\mathbb{E}_{z\sim\mathbb{P}^T}\left[d_z\left(z, \tau_T^{\sigma(z)}\right)\right]$$

$$\mathcal{L}_{total}(\boldsymbol{P}) := \mathcal{L}_S + \lambda_T\mathcal{L}_T + \lambda_{\mathcal{W}}\mathcal{L}_{\mathcal{W}}$$

Results

	ImageCLEF				Office-Home				
	→ C	→ I	→ P	Avg	→ Ar	→ Cl	→ Pr	→ Rw	Avg
Zero-Shot									
CLIP [34]	87.9	88.2	78.7	88.1	71.2	50.4	81.4	82.6	71.4
Source Combined									
DAN [10]	93.3	92.2	77.6	87.7	68.5	59.4	79.0	82.5	72.4
DANN [11]	93.7	91.8	77.9	87.8	68.4	59.1	79.5	82.7	72.4
D-CORAL [40]	93.6	91.7	77.1	87.5	68.1	58.6	79.5	82.7	72.2
DAPL [12]	96.0	89.2	76.0	87.1	72.8	51.9	82.6	83.7	72.8
Simple Prompt [4]	93.6	90.6	80.9	88.4	70.7	52.9	82.9	83.9	72.4
PGA [32]	94.2	92.1	78.5	88.2	74.1	53.9	84.4	85.6	74.5
CRPL (Ours)	**94.8**	**94.5**	**81.7**	**90.3**	**76.6**	**60.4**	**86.5**	**86.8**	**77.6**
Multi-Source									
DCTN [51]	95.7	90.3	75.0	87.0	N.A.	N.A.	N.A.	N.A.	N.A.
MDDA [57]	N.A.	N.A.	N.A.	N.A.	66.7	62.3	79.5	79.6	71.0
SIMpIDA [45]	93.3	91.0	77.5	87.3	70.8	56.3	80.2	81.5	72.2
MFSAN [58]	95.4	93.6	79.1	89.4	72.1	62.0	80.3	81.8	74.1
MPA [4]	**97.2**	**96.2**	80.4	91.3	74.8	54.9	86.2	85.7	75.4
MPGA [32]	93.8	95.7	**82.8**	90.8	74.8	56.0	85.2	86.0	75.5
M-CRPL (Ours)	96.2	96.0	82.3	**91.5**	**76.8**	**63.5**	**87.6**	**87.5**	**78.9**

	DomainNet						
	→ Clp	→ Inf	→ Pnt	→ Qdr	→ Rel	→ Skt	Avg
Zero-Shot							
CLIP [34]	61.3	42.0	56.1	10.3	79.3	54.1	50.5
Source Combined							
DANN [11]	45.5	13.1	37.0	**13.2**	48.9	31.8	32.6
MCD [37]	54.3	22.1	45.7	7.6	58.4	43.5	38.5
DAPL [12]	62.4	43.8	59.3	10.6	81.5	54.6	52.0
Simple Prompt [4]	63.1	41.2	57.7	10.0	75.8	55.8	50.6
PGA [32]	65.4	49.0	60.4	11.1	**81.8**	60.6	55.4
CRPL (Ours)	**65.6**	**50.8**	**66.7**	10.6	80.0	**61.1**	**55.8**
Multi-Source							
M³SDA-β [29]	58.6	26.0	52.3	6.3	62.7	49.5	42.6
SImpAl101 [45]	66.4	26.5	56.6	**18.9**	68.0	55.5	48.6
LtC-MSDA [49]	63.1	28.7	56.1	16.3	66.1	53.8	47.4
T-SVDNet [22]	66.1	25.0	54.3	16.5	65.4	54.6	47.0
PFSA [8]	64.5	29.2	57.6	17.2	67.2	55.1	48.5
PTMDA [35]	66.0	28.5	58.4	13.0	63.0	54.1	47.2
MPA [4]	65.2	47.3	62.0	10.2	**82.0**	57.9	54.1
MPGA [32]	67.2	47.8	63.1	11.6	81.7	**61.0**	55.4
M-CRPL (Ours)	**67.6**	**51.4**	**67.0**	11.1	79.3	60.6	**56.2**

Method	Ar→Cl	Ar→Pr	Ar→Rw	Cl→Ar	Cl→Pr	Cl→Rw	Pr→Ar	Pr→Cl	Pr→Rw	Rw→Ar	Rw→Cl	Rw→Pr	Avg
ResNet-50[16]	34.9	50.0	58.0	37.4	41.9	46.2	38.5	31.2	60.4	53.9	41.2	59.9	46.1
DANN [10]	45.6	59.3	70.1	47.0	58.5	60.9	46.1	43.7	68.5	63.2	51.8	76.8	57.6
JAN [25]	45.9	61.2	68.9	50.4	59.7	61.0	45.8	43.4	70.3	63.9	52.4	76.8	58.3
CDAN+E [27]	50.7	70.6	76.0	57.6	70.0	70.0	57.4	50.9	77.3	70.9	56.7	81.6	65.8
BSP+CDAN [5]	52.0	68.6	76.1	58.0	70.3	70.2	58.6	50.2	77.6	72.2	59.3	81.9	66.3
SymNets [54]	47.7	72.9	78.5	64.2	71.3	74.2	63.6	47.6	79.4	73.8	50.8	82.6	67.2
ETD [21]	51.3	71.9	**85.7**	57.6	69.2	73.7	57.8	51.2	79.3	70.2	57.5	82.1	67.3
BNM [6]	52.3	73.9	80.0	63.3	72.9	74.9	61.7	49.5	79.7	70.5	53.6	82.2	67.9
GSDA [17]	**61.3**	76.1	79.4	65.4	73.3	74.3	65.0	53.2	80.0	72.2	**60.6**	83.1	70.3
GVB-GD [7]	57.0	74.7	79.8	64.6	74.1	74.6	65.2	55.1	81.0	74.6	59.7	84.3	70.4
RSDA-MSTN [15]	53.2	77.7	81.3	66.4	74.0	76.5	67.9	53.0	82.0	75.8	57.8	85.4	70.9
SPL [50]	54.5	77.8	81.9	65.1	78.0	81.1	66.0	53.1	82.8	69.9	55.3	**86.0**	71.0
SRDC [41]	52.3	76.3	81.0	69.5	76.2	78.0	68.7	53.8	81.7	**76.3**	57.1	85.0	71.3
DisClusterDA [42]	58.8	77.0	80.8	67.0	74.6	77.1	65.9	**56.3**	81.4	74.2	60.5	83.6	71.4
CLIP [34]	51.6	81.9	82.6	71.9	81.9	82.6	71.9	51.6	82.6	71.9	51.6	81.9	72.0
DAPL [12]	52.7	82.2	84.1	73.9	82.0	**83.8**	73.6	54.6	84.0	73.3	53.4	82.5	73.3
PGA [32]	53.7	83.9	85.0	73.2	83.9	84.6	73.2	53.8	84.1	73.5	53.1	85.3	73.9
CRPL (Ours)	54.7	**84.1**	84.6	**74.3**	**83.2**	83.7	**73.7**	53.4	**84.6**	74.5	55.5	85.5	**74.4**

Ablation

	Inference Prompt	Ar	Cl	Pr	Rw
CPL	τ_T	47.6	29.0	53.5	63.5
	τ_S	74.3	57.2	84.3	84.8
	τ_{avg}	61.2	40.3	73.3	77.8
SPL	τ_T	75.8	62.9	86.6	87.2
	τ_S	73.3	57.4	83.4	84.7
	τ_{avg}	75.8	58.1	83.6	85.3
CPL only	τ_T	47.6	29.0	53.5	63.5
CPL with $\mathcal{L}_{\mathcal{W}}$	τ_T	7.9	4.7	80.4	82.3
SPL only	τ_T	75.8	62.9	86.6	87.2
SPL with $\mathcal{L}_{\mathcal{W}}$	τ_T	76.8	63.5	87.5	87.6

CPL: CLIP zero-shot
SPL: enhanced pseudo-labels derived from the source domains

LW strongly depends on the initial predictions

Distance	Ar	Cl	Pr	Rw	Average
Average	76.0	62.6	87.0	87.5	78.3
cosine	76.6	56.2	88.2	86.7	76.9
L2	**76.8**	**63.5**	**87.5**	**87.6**	**78.9**

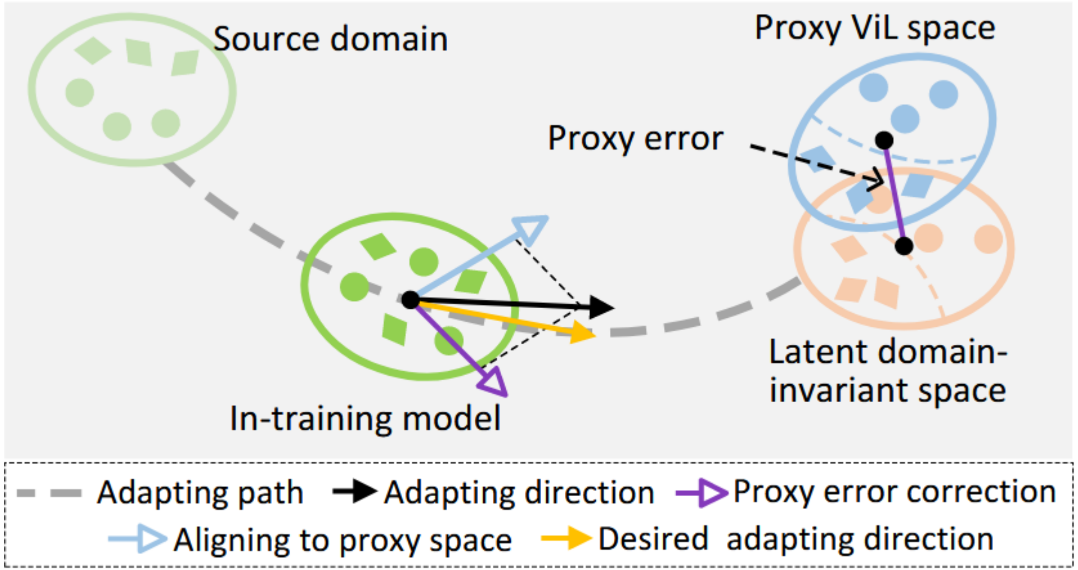
Proxy Denoising for Source-Free Domain Adaptation

Song Tang[1,2], Wenxin Su[1], Mao Ye[*3], Jianwei Zhang[2], and Xiatian Zhu[*4]

[1]University of Shanghai for Science and Technology [2]Universität Hamburg
[3]University of Electronic Science and Technology of China [4]University of Surrey
tangs@usst.edu.cn, {suwenxin43, cvlab.uestc}@gmail.com, xiatian.zhu@surrey.ac.uk

ICLR2025

Motivation

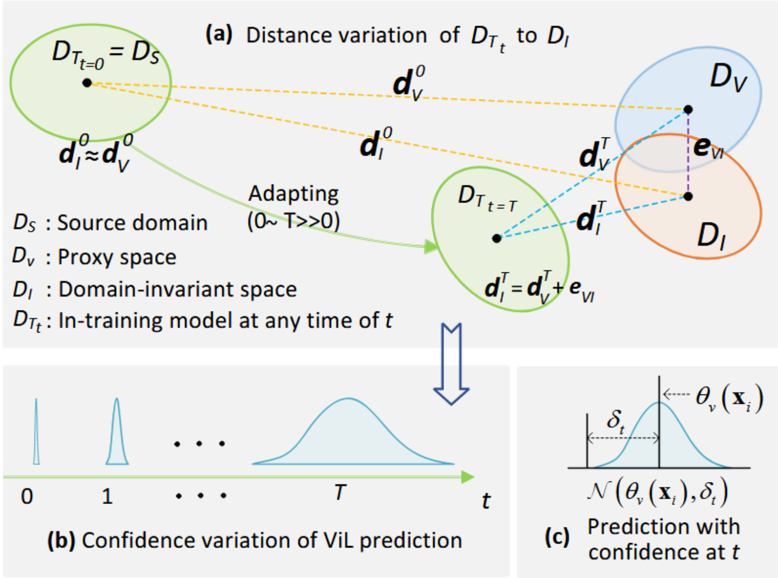


Considering the ViL model/space as a noisy proxy of the latent domain-invariant space, with a need to be denoised.

Exploit the dynamics of domain adaptation process.

Method

Proxy Confidence Theory



$$\delta_t \propto \eta_t$$

Case1: $\boldsymbol{d}_I^0 \approx \boldsymbol{d}_V^0 \gg \boldsymbol{e}_{VI}$

Case2: $\boldsymbol{d}_I^t = \boldsymbol{d}_V^t + \boldsymbol{e}_{VI}$

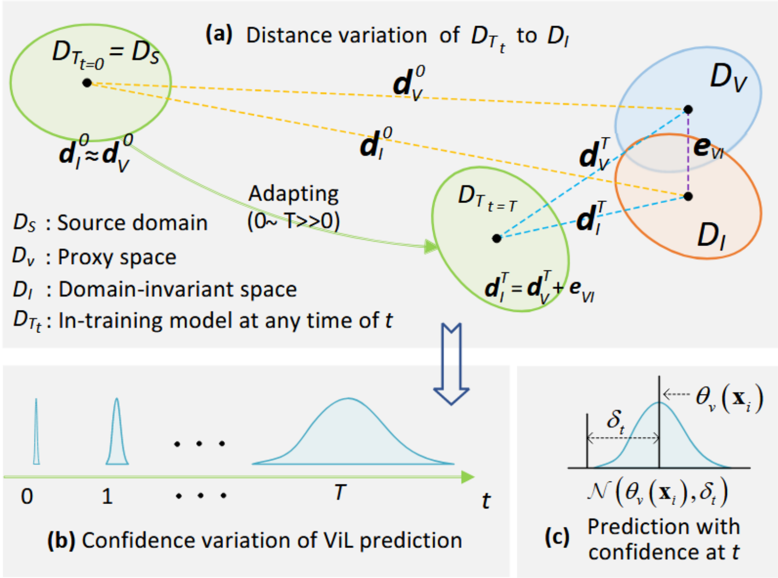
$$\eta_t = \frac{\boldsymbol{d}_I^t}{\boldsymbol{d}_V^t} = \frac{\boldsymbol{d}_V^t + \boldsymbol{e}_{VI}}{\boldsymbol{d}_V^t} = \left(1 + \frac{\boldsymbol{e}_{VI}}{\boldsymbol{d}_V^t}\right)$$

$$\eta_t = \frac{|\boldsymbol{d}_{\mathcal{I}}^t|}{|\boldsymbol{d}_{\mathcal{V}}^t|} = \frac{|\boldsymbol{d}_{\mathcal{V}}^t + \boldsymbol{e}_{\mathcal{V}\mathcal{I}}|}{|\boldsymbol{d}_{\mathcal{V}}^t|} \leq \frac{|\boldsymbol{d}_{\mathcal{V}}^t| + |\boldsymbol{e}_{\mathcal{V}\mathcal{I}}|}{|\boldsymbol{d}_{\mathcal{V}}^t|} = 1 + \frac{|\boldsymbol{e}_{\mathcal{V}\mathcal{I}}|}{|\boldsymbol{d}_{\mathcal{V}}^t|}$$

the impact of errors gradually increases

Method

Proxy Confidence Theory



$$\mathcal{N}(\theta_v(x_i), \delta_t) \Longrightarrow P(G_{P(V)} = True, t)\, P(V)$$

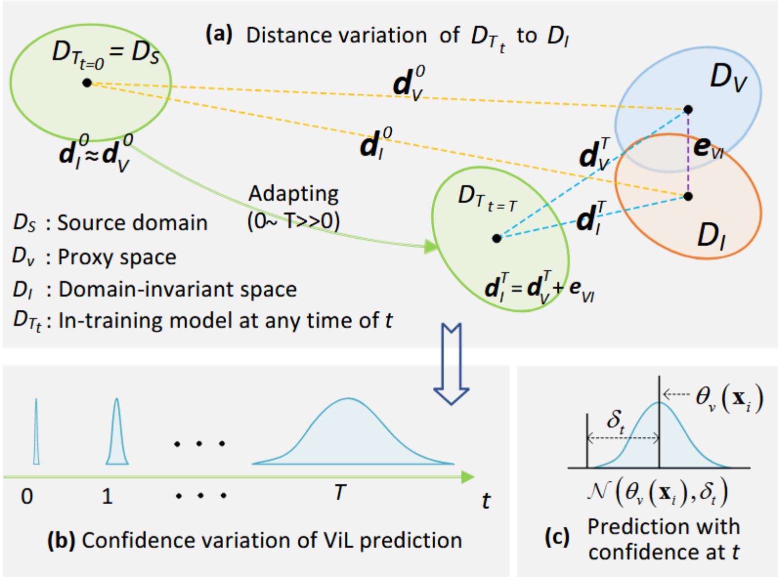
=proxy confidence * probability distribution

$$P(G_{P(V)} = True, t) \propto \frac{P(T_t)}{P(S)}$$

The effect of ViL model’ s prediction error is approximately reflected by the discrepancy between the source domain and the current in-training model

Method

Proxy Confidence Theory



Proof

$$P\left(G_{P(V)} = True, t\right) \propto \frac{Distance(D_{T_t}, D_I)}{Distance(D_S, D_I)} = \frac{d_I^t}{d_S},$$

$$\begin{aligned}
\frac{d_I^t}{d_S} &= \frac{KL\left(P(T_t)||P(I)\right)}{KL\left(P(S)||P(I)\right)} = \frac{\int_{T_t} P(T_t)\log\frac{P(T_t)}{P(I)}dT_t}{\int_S P(S)\log\frac{P(S)}{P(I)}dS} \\
&= \frac{-\int_{T_t} P(T_t)\log P(T_t)dT_t + \int_{T_t} P(T_t)\log P(I)dT_t}{-\int_S P(S)\log P(S)dS + \int_S P(S)\log P(I)dS} \\
&= \frac{H(T_t) + \log P(I)}{H(S) + \log P(I)} \\
&= \frac{H(T_t)}{H(S)},
\end{aligned}$$

$$\frac{H(T_t) + \log P(I)}{H(S) + \log P(I)} = \frac{H(T_t)}{H(S)} \propto \frac{P(T_t)}{P(S)}$$

Denoising mechanism

Theoretical results

$$P\left(G_{P(\mathcal{V})} = True, t\right) \propto \boxed{\frac{P(\mathcal{T}^t)}{P(\mathcal{S})}} \quad (3)$$

Distribution estimated by the current in-training model

Source distribution

Insight: The effect of proxy errors on domain adaptation can be approximately estimated by contrasting the distributions of the source model and the current in-training model

- **Denoising design**

$$\boxed{P\left(G_{P(\mathcal{V})} = True, t\right)} P\left(\mathcal{V}\right)$$

ViL's prediction distribution

$$\log\left(\frac{P(\mathcal{T}^t)}{P(\mathcal{S})} P(\mathcal{V})\right) = \log P(\mathcal{V}) - \left[\log P(\mathcal{S}) - \log P(\mathcal{T}^t)\right]$$

$$\boldsymbol{p}_i' = \operatorname{softmax}\left(\theta_v\left(\boldsymbol{x}_i, \boldsymbol{v}\right) - \omega\left[\theta_s\left(\boldsymbol{x}_i\right) - \theta_t\left(\boldsymbol{x}_i\right)\right]\right) \quad (4)$$

The corrected ViL prediction

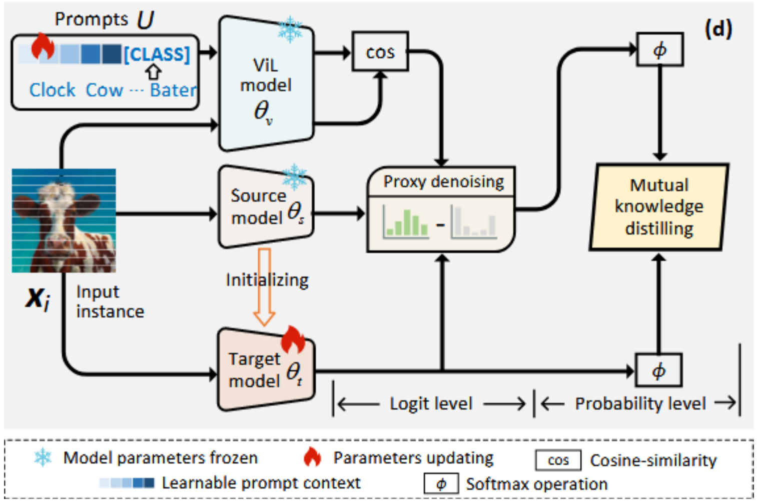
ViL model's logit

Source model's logit

Current in-training model's logit

Method

Proxy denoising



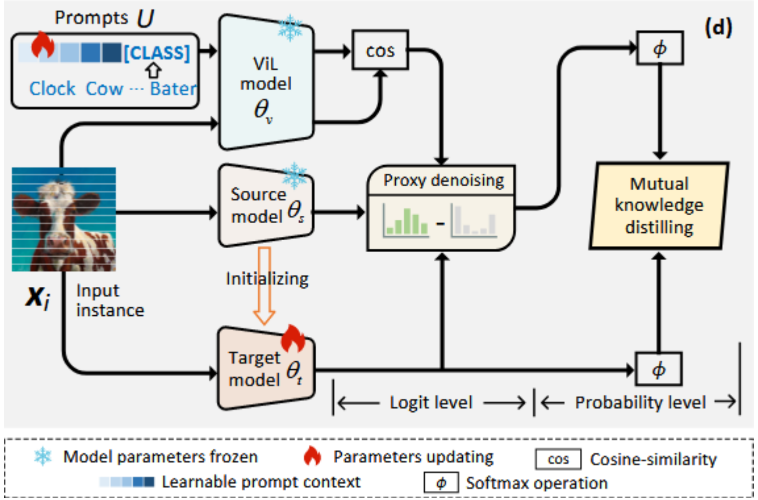
$$\log\left[P\left(G_{P(V)}=True,t\right)P\left(V\right)\right]\propto\log\left(\frac{P(T_t)}{P(S)}P(V)\right)$$
$$=\log P(V)-\left[\log P(S)-\log P(T_t)\right].$$

$$\begin{cases}\boldsymbol{p}_i'=\phi\left(\boldsymbol{l}_i'\right),\ \boldsymbol{l}_i'=\theta_v\left(\boldsymbol{x}_i,\boldsymbol{v}\right)-\omega\Delta_t,\\ \Delta_t=\theta_s\left(\boldsymbol{x}_i\right)-\theta_t\left(\boldsymbol{x}_i\right),\end{cases}$$

$$\boldsymbol{p}_i'=\mathrm{softmax}\left(\theta_v\left(\boldsymbol{x}_i,\boldsymbol{v}\right)-\omega[\theta_s\left(\boldsymbol{x}_i\right)-\theta_t\left(\boldsymbol{x}_i\right)]\right)$$

Method

Mutual knowledge distilling



$$L_{\mathrm{ProDe}} = \min_{\theta_t, \boldsymbol{v}} \alpha \overbrace{\left(-\mathbb{E}_{\boldsymbol{x}_i \in \mathcal{X}_t} \mathbf{MI}\left(\boldsymbol{p}_i', \boldsymbol{p}_i\right) + \gamma \sum_{c=1}^{C} \bar{q}_c \log \bar{q}_c \right)}^{L_{\mathrm{Apt}}} \overbrace{-\beta\, \mathbb{E}_{\boldsymbol{x}_i \in \mathcal{X}_t} \sum_{c=1}^{C} \mathbb{1}\left[c = y_i'\right] \log p_{i,c}}^{L_{\mathrm{Ref}}},$$

Results&Ablation

Method	Office-31	Office-Home	VisDA	DomainNet-126
CLIP-R [33]	71.4	72.1	83.7	72.7
ProDe-R	**90.0**	**82.9**	**89.9**	**81.5**
CLIP-V [33]	79.8	76.1	82.9	76.3
ProDe-V	**92.6**	**86.2**	**91.6**	**85.0**

#	L_{Syn}	L_{Ref}	Office-31	Office-Home	VisDA	Avg.
1	✗	✗	78.6	59.2	49.2	62.3
2	✓	✗	91.8	78.8	90.2	86.9
3	✗	✓	86.5	83.2	90.7	86.8
4	✓	✓	**92.6**	**86.2**	**91.6**	**90.1**
5	ProDe-V w/o pd		90.5	83.9	89.9	88.1
6	ProDe-V w/o source		91.2	84.6	90.2	88.7
7	ProDe-V w/o target		80.1	83.5	90.8	84.8
8	ProDe-V w proba		86.8	83.3	91.3	87.1