Revisiting Source-Free Domain Adaptation: a New Perspective via Uncertainty Control

Gezheng Xu*, Hui Guo*, Li Yi, Charles Ling, Boyu Wang†, Grace Y. Yi† University of Western Ontario, Tiktok

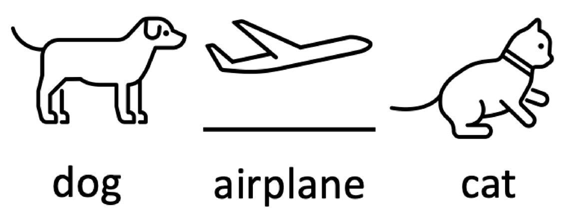
Western UNIVERSITY · CANADA

VECTOR INSTITUTE | INSTITUT VECTEUR

Problem Setup

- **Source-Free Domain Adaptation (SFDA):** adapting a well-trained source model to a target domain with unlabeled target data
- **Data and model:**

Source data: $\{x_i^S, y_i^S\}_{i=1}^{N_S}$ **Target data:** $\{x_i^T\}_{i=1}^{N_T}$



⇨ **Source Model: h_S** → **Target Model: h_T**

➢ **Source data:** $\mathcal{D}_\mathrm{S} \triangleq \{\mathrm{x}_i^\mathrm{S}, \mathrm{y}_i^\mathrm{S}\}_{i=1}^{N_\mathrm{S}}$ from source domain distribution ⇒ *Source model*: $h_\mathrm{S} : \mathcal{X} \to \mathcal{Y}$

➢ **Target data:** $\mathcal{D}_\mathrm{T} \triangleq \{\mathrm{x}_i^\mathrm{T}\}_{i=1}^{N_\mathrm{T}}$ from target domain distribution

➢ **Goal:** learn a *target model* $h_\mathrm{T} : \mathcal{X} \to \mathcal{Y}$ that predicts labels in the target domain by adapting h_S on $\mathcal{D}_\mathrm{T}$

[Notation: $\boldsymbol{\theta}/\boldsymbol{\theta}_\mathrm{S}/\boldsymbol{\theta}_\mathrm{T}$ – generic/source/target model parameters]

Motivation

- **Contrastive learning-based methods for SFDA:**

$\mathcal{S}_{\boldsymbol{\theta}} : \mathcal{X} \times \mathcal{X} \to [0, 1]$: similarity measure between two instances e.g., cosine similarity between features, $\langle f(\mathrm{x};\boldsymbol{\theta}), f(\mathrm{x}';\boldsymbol{\theta})\rangle$

$$\mathcal{R}_{\text{basic}}(\boldsymbol{\theta}) = \mathbb{E}_{P_\mathrm{X}^\mathrm{T}}\left[-\mathbb{E}_{P^+}\left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathrm{X}^+;\mathrm{X})\right\} + \mathbb{E}_{P^-}\left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathrm{X}^-;\mathrm{X})\right\}\right],$$

maximize the similarity between positive pairs $(\mathrm{x}, \mathrm{x}^+)$
x^+: belong to the same class as x (e.g., κ-nearest neighbours of x)

minimize the similarity between negative pairs $(\mathrm{x}, \mathrm{x}^-)$
x^-: belong to a different class than x (e.g., $\mathcal{B}\backslash\{\mathrm{x}\}$ in a mini-batch $\mathcal{B}$)

➢ **Challenge 1:** The empirical negative distribution (P^-) deviates from the true distribution due to the inclusion of false negatives

$$\mathcal{R}_\mathrm{x}^-(\boldsymbol{\theta}; P^-, \delta) = \sup_{Q^- \in \Gamma_\delta(P^-)} \left[\mathbb{E}_{Q^-}\left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathrm{X}^-;\mathrm{x})\right\}\right]$$

$\Gamma_\delta(P^-) = \{Q^- \in \mathcal{P}_p(\mathcal{X}) : d(Q^-, P^-) \le \delta\}$
uncertainty set: distributions obtained by perturbing P^-

➢ **Challenge 2:** Supervisory information from positive examples may not fully align with the ground truth

$$\mathcal{R}_\mathrm{x}^+(p; \mathrm{x}^+, \delta) \triangleq \sup_{q^+ \in \Gamma_\delta(p^+)} \langle q^+, -p\rangle$$

$p \triangleq f(\mathrm{x};\boldsymbol{\theta}) \in \Delta^{K-1}$: predicted probabilities for x
$p^+ \triangleq f(\mathrm{x}^+;\boldsymbol{\theta}) \in \Delta^{K-1}$: predicted probabilities for x^+

Empirical Results

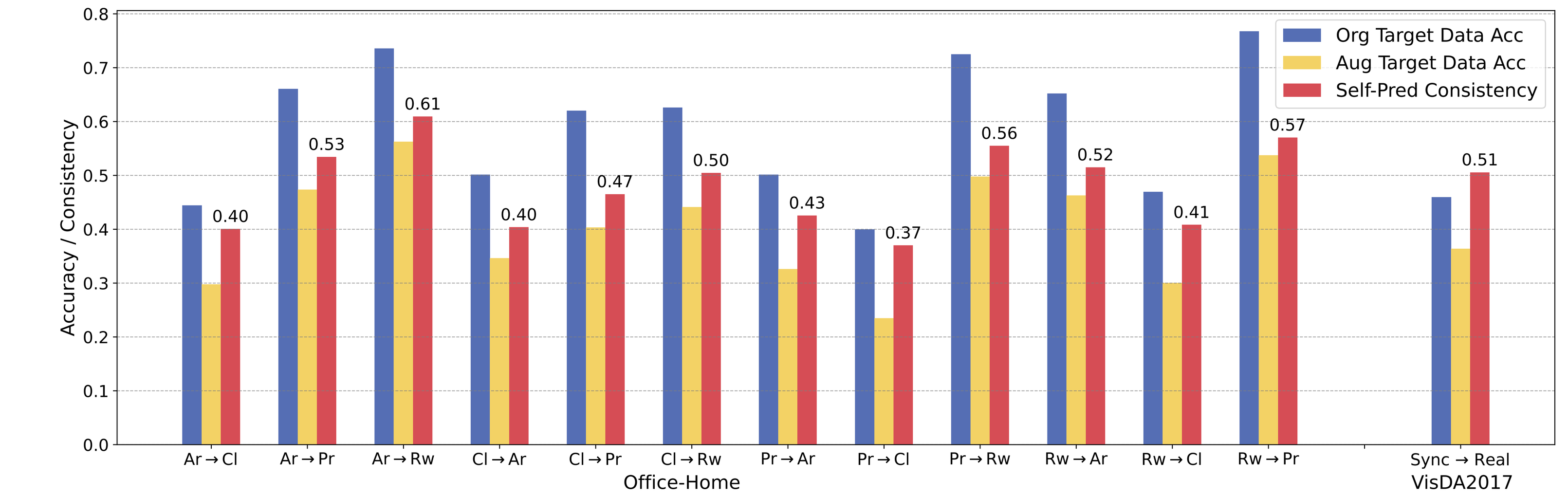


Figure 1: Inconsistency between the prediction results between the anchor image and its augmented view by source model

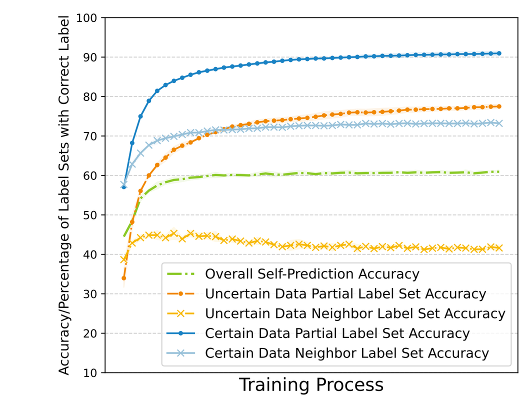


Figure 2: Positive supervision uncertainty.

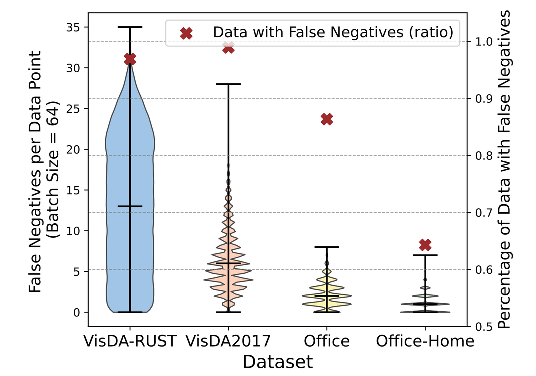


Figure 3: Negative sampling uncertainty.

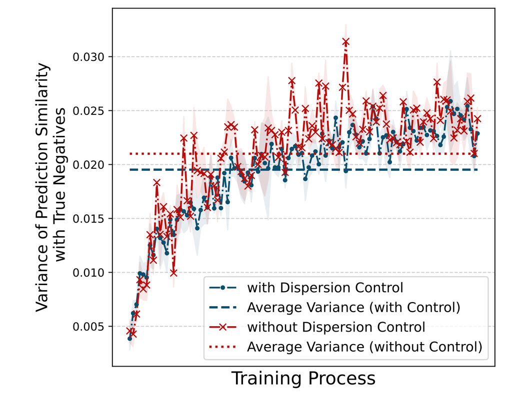


Figure 4: Similarity dispersion control.

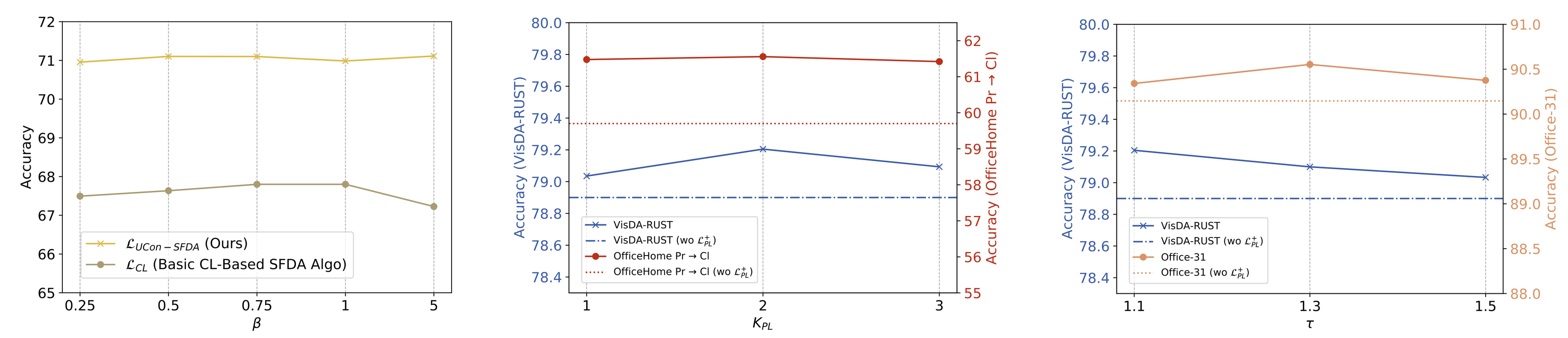


Figure 5: Hyperparameter sensitivity analysis.

Table 1: Hyperparameter values across different datasets. "Orig. λ_DC", "Orig. K_PL", and "Orig. τ" refer to the original values used in our paper, which are selected following the general hyper-parameter tuning pipeline in the literature. Other hyperparameters are directly calculated with theory-motivated hyperparameter determination approaches, where "Init." and "Final" indicate the first and the last training epochs, respectively.

Metric	Office31	OfficeHome	OfficeHome (partial set)	VisDA2017	VisDA-RUST	DomainNet126
Orig. λ_DC	1.000	0.500	1.000	1.000	0.500	0.500
New λ_DC	0.390	0.520	0.476	0.494	0.461	0.553
Orig. K_PL	2.000	2.000	2.000	1.000	2.000	2.000
Init. k_0 **(Averaged)**	1.320	1.535	1.513	1.341	1.348	1.644
Final k_0 **(Averaged)**	1.003	1.028	1.003	1.008	1.020	1.079
Orig. τ	1.300	1.100	1.100	1.100	1.100	1.100
Init. τ_t	1.308	1.265	1.238	1.790	1.674	1.232
Final τ_t	1.056	1.090	1.042	1.260	1.368	1.092
τ_s **(10th percentile)**	2.037	1.230	1.268	1.164	1.163	1.264

UCon-SFDA: Theoretical Analysis and Methodology

Uncertainty Control algorithm for SFDA (UCon-SFDA)
[Negative Sampling Uncertainty and Dispersion Control | Positive Supervision Uncertainty and Partial Labeling]

❖ **Conventional Contrastive Loss:** $\mathcal{L}_\mathrm{CL} \triangleq \mathcal{L}_\mathrm{CL}^+ + \lambda_\mathrm{CL}^-\mathcal{L}_\mathrm{CL}^-$

➢ $\mathcal{L}_\mathrm{CL}^+ = \frac{1}{N_\mathrm{T}}\sum_{i=1}^{N_\mathrm{T}}\{-\sum_{\mathrm{x}_i^+\in\mathcal{C}_i}\mathcal{S}_{\boldsymbol{\theta}}(\mathrm{x}_i^+;\mathrm{x}_i)\}$: positive samples are the κ-nearest neighbours in the feature space

➢ $\mathcal{L}_\mathrm{CL}^- = \frac{1}{N_\mathrm{T}}\sum_{i=1}^{N_\mathrm{T}}\sum_{\mathrm{x}_i^-\in\mathcal{B}\backslash\{\mathrm{x}_i\}}\mathcal{S}_{\boldsymbol{\theta}}(\mathrm{x}_i^-;\mathrm{x}_i)$: negative samples are the remaining data points in the same mini-batch $\mathcal{B}$

❖ **Dispersion Control via Data Augmentation Alignment:**

Addressing Challenge 1:
Theorem (*Informal.*)

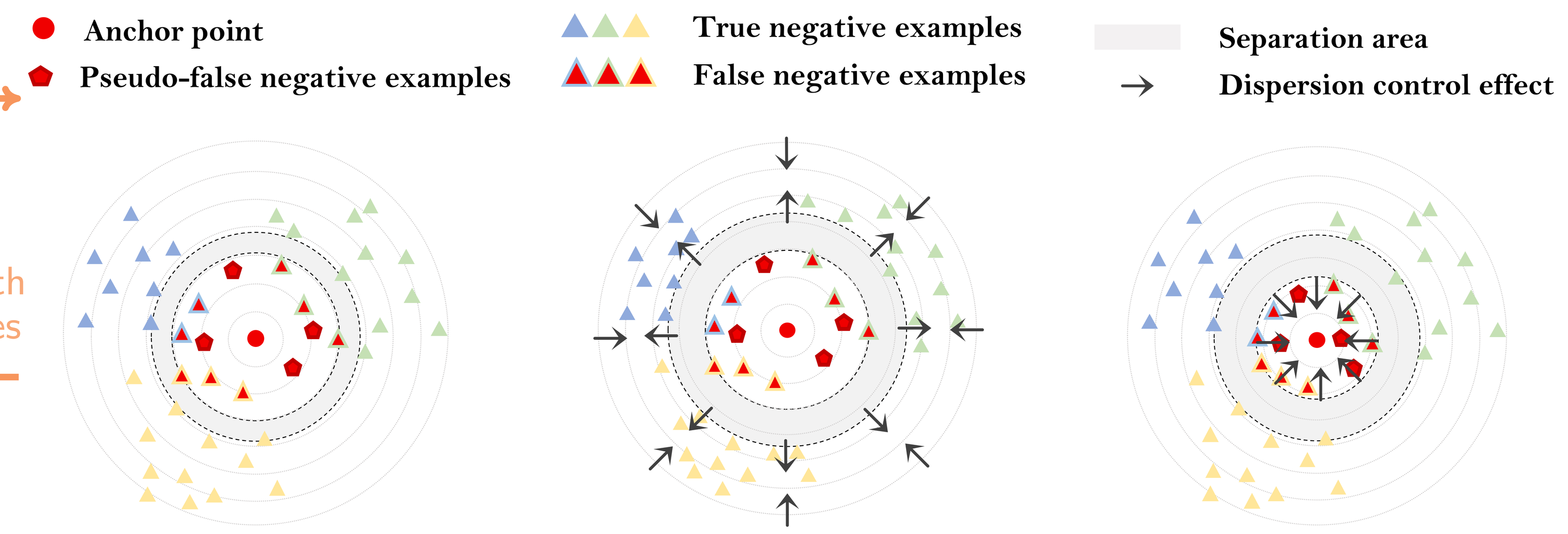
$$\mathcal{R}_\mathrm{x}^-(\boldsymbol{\theta}; P_{\text{train}}^-, \delta, \epsilon) \le \underbrace{\mathbb{E}_{P^*}\left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathrm{X}^-;\mathrm{x})\right\}}_{\text{similar to the traditional negative sample loss } \mathcal{L}_\mathrm{CL}^-} + \underbrace{\mathcal{V}_d\left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathrm{X}^-;\mathrm{x})\right\}}_{\text{dispersion term (depending on } d\text{) e.g., variance}} + O(\delta).$$

dispersion control effect

dispersion control with pseudo-false negatives

➢ Negative uncertainty control loss:

$$\mathcal{L}_\mathrm{UCon}^- \triangleq \lambda_\mathrm{CL}^-\mathcal{L}_\mathrm{CL}^- + \lambda_\mathrm{DC}\mathcal{L}_\mathrm{DC}^- \triangleq \lambda_\mathrm{CL}^-\mathcal{L}_\mathrm{CL}^- + \lambda_\mathrm{DC}\left\{-\frac{1}{N_\mathrm{T}}\sum_{i=1}^{N_\mathrm{T}} \mathrm{d}_{\boldsymbol{\theta}}\left(\mathtt{AUG}(\mathrm{x}_i), \mathrm{x}_i\right)\right\}$$



❖ **Supervision Relaxation by Partial Label Training:**

Addressing Challenge 2:
Theorem (*Informal.*) Minimizing $\mathcal{R}_\mathrm{x}^+(p;\mathrm{x}^+,\delta)$ w.r.t. p yields $p^{\star(j)} = \frac{1}{k_0}\mathbf{1}(j \le k_0)$ for some $k_0 \in [K]$, where $p^{\star(j)}$ corresponds to the index of $p^{+(j)}$, the jth largest element of $p^+ \triangleq (p_1^+, \dots, p_K^+)^\top$. In particular, if the difference between $p^{+(1)}$ and $p^{+(2)}$ is large enough, then $k_0 = 1$.

➢ Positive uncertainty control loss:

$$\mathcal{L}_\mathrm{UCon}^+ \triangleq \mathcal{L}_\mathrm{CL}^+ + \lambda_\mathrm{PL}\mathcal{L}_\mathrm{PL}^+ \triangleq \mathcal{L}_\mathrm{CL}^+ + \lambda_\mathrm{PL}\cdot\frac{1}{N_\mathrm{T}}\sum_{i=1}^{N_\mathrm{T}}\sum_{\mathrm{y}_{k,i}\in\mathcal{Y}_{\mathrm{PL},i}}\mathbb{1}_{\{\mathrm{x}_i\in\mathcal{U}\}}\ell_\mathrm{CE}(\mathrm{y}_{k,i}, f(\mathrm{x}_i;\boldsymbol{\theta}))$$

$\mathcal{Y}_{\mathrm{PL},i}$: partial label set for x_i
– historical TOP-K_PL predicted labels for x_i
– selection criterion: $f(\mathrm{x}_i;\boldsymbol{\theta})^{(1)}/f(\mathrm{x}_i;\boldsymbol{\theta})^{(2)} \le \tau$ for some $\tau > 1$

Contributions

1. We theoretically analyze **two key sources of uncertainty** in contrastive learning-based SFDA methods, identifying two types of worst-case risks under a unified DRO framework.
2. Based on our theoretical result, we design a novel **Uncertainty Control algorithm for SFDA (UCon-SFDA)** to minimize the negative effects of uncertainty in negative sample selection while leveraging the uncertain information from positive supervision.
3. We conduct **extensive experiments** to demonstrate the effectiveness of the proposed method.