Supplementary Material — Towards Reliable Model Selection for Unsupervised Domain Adaptation: An Empirical Study and A Certified Baseline

Dapeng Hu[1*] **Mi Luo**[3] **Jian Liang**[4,5†] **Chuan-Sheng Foo**[2,1]
[1]Centre for Frontier AI Research, A*STAR, Singapore
[2]Institute for Infocomm Research, A*STAR, Singapore
[3]National University of Singapore
[4]NLPR & MAIS, Institute of Automation, Chinese Academy of Sciences
[5]School of Artificial Intelligence, University of Chinese Academy of Sciences

A Proof of Proposition 1

We first prove the first inequality using Jensen's inequality, which states that for a real-valued, convex function φ with its domain as a subset of $\mathbb{R}$ and numbers $t_1, \ldots, t_n$ in its domain, the inequality $\varphi\left(\frac{1}{n}\sum_{i=1}^{n} t_i\right) \leq \frac{1}{n}\sum_{i=1}^{n}\varphi(t_i)$ holds. Given that $-\log$ is convex, and assuming $m > 1$ with candidate models having different parameter weights θ, resulting in distinct discriminative mappings of $f(x, \theta)$, we can strictly obtain $l(\frac{1}{m}\sum_{i=1}^{m} f(x, \theta_i), y) < \frac{1}{m}\sum_{i=1}^{m} l(f(x, \theta_i), y)$ without the equal situation. Next, we leverage the property of inequalities to prove the second inequality. Here, θ_{worst} denotes the worst candidate model, i.e., the model with the largest loss. For any other candidate model θ_i, we have $l(f(x, \theta_i), y) < l(f(x, \theta_{\text{worst}}), y)$. This ensures that $\frac{1}{m}\sum_{i=1}^{m} l(f(x, \theta_i), y) < \frac{1}{m}\sum_{i=1}^{m} l(f(x, \theta_{\text{worst}}), y)$, or explicitly, $\frac{1}{m}\sum_{i=1}^{m} l(f(x, \theta_i), y) < l(f(x, \theta_{\text{worst}}), y)$. Substituting the NLL loss with any strongly convex loss function would still uphold the proposition.

B Model Selection Baselines

Let $\{p_{\text{t}}^i\}_{i=1}^{n_{\text{t}}}$ represent the output probability vectors of all n_{t} target samples, and let $P \in \mathbb{R}^{n_{\text{t}} \times C}$ denote the total probability matrix. We introduce the respective computation involved in the existing model selection approaches.

Source risk. The SourceRisk approach [1] utilizes a held-out validation set from the source domain to select the model θ_k that performs best on this set as the final decision. However, this method has limited effectiveness in scenarios with severe domain shifts between the source and target domains. Additionally, it introduces additional hyperparameters for dataset splitting, which can further complicate the model selection process.

Importance-weighted source risk. Directly taking source risk as target risk is unreliable due to domain distribution shifts between domains. To address this challenge, [2] propose Importance-Weighted Cross Validation (IWCV), which re-weights the source risk using a source-target density ratio estimated in the input space. [3] further enhance IWCV by introducing Deep Embedded Validation (DEV), which estimates the density ratio in the feature space using a domain discriminator and controls the variance. Both IWCV and DEV rely on the importance weighting technique [4], which assumes that the target distribution is included in the source distribution [2], making the weighting unreliable in scenarios with severe covariate shift and label shift. In addition, both IWCV

*This work was completed while Dapeng (lhxxhb15@gmail.com) was a scientist at A*STAR.
†Corresponding author: Jian Liang (liangjian92@gmail.com)

and DEV involve hyperparameters and extra model training during the density ratio estimation process.

Reversed source risk. Building upon the concept of reverse cross-validation [5], [6] propose a novel Reverse Validation approach (RV). This method first conducts source-to-target adaptation to obtain a UDA model, which enables the acquisition of pseudo labels for the target unlabeled data. Subsequently, Reverse Validation performs a reversed adaptation from the pseudo-labeled target to the source and utilizes the source risk in this reversed adaptation task for validation. Reverse Validation relies on the symmetry between domains and cannot handle label shifts. Additionally, this approach involves hyperparameters for dataset splitting.

Entropy. [7] propose using the mean Shannon's Entropy of target-domain predictions as a validation metric, prioritizing predictions with high certainty. The underlying intuition is that a good decision boundary should avoid crossing high-density regions in the target structure [8, 9]. Lower Entropy scores indicate better model performance for this metric.

$$\text{Entropy} = -\frac{1}{n_\text{t}} \sum_{i=1}^{n_\text{t}} \sum_{j=1}^{C} P_{ij} \log P_{ij}$$

Information maximization. The Entropy score only considers sample-wise certainty, which can be misleading when high-certainty predictions are biased towards a small fraction of classes [10]. To address this challenge, [11] utilize input-output mutual information maximization (InfoMax) [12] as a validation metric. In contrast to Entropy, InfoMax includes an additional class-balance regularization by encouraging the averaged prediction $\bar{p} = \frac{1}{n_\text{t}} \sum_{i=1}^{n_\text{t}} P_{ij}, \quad \bar{p} \in \mathbb{R}^C$ to be even. Higher InfoMax scores indicate better model performance according to this metric.

$$\text{InfoMax} = -\sum_{j=1}^{C} \bar{p} \log \bar{p} + \frac{1}{n_\text{t}} \sum_{i=1}^{n_\text{t}} \sum_{j=1}^{C} P_{ij} \log P_{ij}$$

Neighborhood consistency. [10] introduce Soft Neighborhood Density (SND), a novel metric that focuses on the property of neighborhood consistency. SND leverages softmax predictions as features and constructs a sample-to-sample similarity matrix. This matrix is transformed into a probabilistic distribution using the softmax function: $S = \text{softmax}(PP^T/\tau), \quad S \in \mathbb{R}^{n_\text{t} \times n_\text{t}}$. Here, τ is a small temperature parameter that sharpens the distribution, enabling the difference between nearby and distant samples. SND favors high neighborhood consistency by prioritizing samples whose predictions are similar to other samples within the same neighborhood, resulting in higher SND scores.

$$\text{SND} = -\frac{1}{n_\text{t}} \sum_{i=1}^{n_\text{t}} \sum_{j=1}^{n_\text{t}} S_{ij} \log S_{ij}$$

Class correlation. [13] introduce Corr-C, a class correlation-based metric that evaluates both class diversity and prediction certainty. Corr-C calculates the cosine similarity between the class correlation matrix and an identity matrix. Lower Corr-C scores are indicative of better model performance based on this metric.

$$\text{Corr-C} = \frac{\text{sum}(\text{diag}(P^T P))}{\|P^T P\|_\text{F}}$$

We can generally classify model selection baselines into two categories: source domain-based methods, including SourceRisk, IWCV, DEV, and RV, and target domain-specific methods, encompassing Entropy, InfoMax, SND, and Corr-C. Recent model selection studies [10, 11, 13] predominantly align with the target domain-specific approach. This trend arises because access to source data restricts UDA to closed-set UDA and often involves additional model training, making the validation process even more complex than UDA model training. In contrast, target domain-specific methods are more straightforward and effective [10]. EnsV, our proposed method, also falls within the category of target domain-specific methods, but fortunately with enhanced reliability due to a theoretical guarantee designed to avert worst-case model selection scenarios.

C Hyperparameter Configurations

In our main experiments, we adopt the setting of previous studies [3, 10] by tuning a single hyper-parameter for various UDA methods. The comprehensive hyperparameter settings can be found in Table 1.

Table 1: Hyperparameter settings for all considered UDA methods. The settings are partially based on [10], with an expanded search space size from 5 to 7 and the inclusion of additional UDA methods across diverse UDA scenarios.

UDA method	UDA type	Hyperparameter	Search space	Default value
ATDOC [14]	CDA self-training	loss coefficient λ	$\{0.02, 0.05, 0.1,$ $0.2, 0.5, 1.0, 2.0\}$	0.2
BNM [15]	CDA output regularization	loss coefficient λ	$\{0.02, 0.05, 0.1,$ $0.2, 0.5, 1.0, 2.0\}$	1.0
CDAN [16]	CDA feature alignment	loss coefficient λ	$\{0.05, 0.1, 0.2,$ $0.5, 1.0, 2.0, 5.0\}$	1.0
MCC [17]	CDA output regularization	temperature T	$\{1.0, 1.5, 2.0,$ $2.5, 3.0, 3.5, 4.0\}$	2.5
MDD [18]	CDA output alignment	margin factor γ	$\{0.5, 1.0, 2.0,$ $3.0, 4.0, 5.0, 6.0\}$	4.0
SAFN [19]	CDA/PDA feature regularization	loss coefficient λ	$\{0.002, 0.005, 0.01,$ $0.02, 0.05, 0.1, 0.2\}$	0.05
PADA [20]	PDA feature alignment	loss coefficient λ	$\{0.05, 0.1, 0.2,$ $0.5, 1.0, 2.0, 5.0\}$	1.0
DANCE [21]	OPDA self-supervision	loss coefficient η	$\{0.02, 0.05, 0.1,$ $0.2, 0.5, 1.0, 2.0\}$	0.05
SHOT [22]	white-box SFUDA hypothesis transfer	loss coefficient β	$\{0.03, 0.05, 0.1,$ $0.3, 0.5, 1.0, 3.0\}$	0.3
DINE [14]	black-box SFUDA knowledge distillation	loss coefficient β	$\{0.05, 0.1, 0.2,$ $0.5, 1.0, 2.0, 5.0\}$	1.0
AdaptSeg [23]	segmentation output alignment	loss coefficient λ	$\{0.0001, 0.0003, 0.001,$ $0.003, 0.01, 0.03\}$	0.0002
AdvEnt [24]	segmentation output alignment	loss coefficient λ	$\{0.0001, 0.0003, 0.001,$ $0.003, 0.01, 0.03\}$	0.001

D Full Model Selection Results

For a comprehensive study, we further consider the parameter weight-based ensemble [25] as our role model, and the EnsV variant based on this role model is denoted as 'EnsV-W'. While the parameter weight-based ensemble also shows competitiveness, it requires all candidate models to share the same architecture and lacks a theoretical guarantee of the ensemble performance. Thus, we recommend the simple and generic prediction-based ensemble, i.e., the default 'EnsV'.

In our experiments, we perform hyperparameter selection for both classification and segmentation tasks. For open-partial-set UDA experiments, we utilize the H-score (%) [26, 27] metric, which combines the accuracy of known classes and unknown samples. For semantic segmentation tasks, we employ the mean intersection-over-union (mIoU) (%) [23, 24] metric. As for other classification tasks, we adopt the accuracy (%) metric. Kindly refer to Table 2 to Table 17 for the complete model selection results.

Table 2: Validation accuracy (%) of a closed-set UDA method ATDOC [14] on *Office-Home*.

Method	Ar → Cl	Ar → Pr	Ar → Re	Cl → Ar	Cl → Pr	Cl → Re	Pr → Ar	Pr → Cl	Pr → Re	Re → Ar	Re → Cl	Re → Pr	avg
SourceRisk [1]	51.41	**77.31**	78.17	**66.87**	74.36	75.60	61.85	48.04	76.06	71.16	58.14	**84.05**	68.59
IWCV [2]	55.88	76.57	78.88	66.25	74.50	78.33	65.60	48.04	80.58	**72.06**	58.14	83.87	69.89
DEV [3]	51.41	76.55	78.88	66.25	74.36	77.67	64.77	51.29	81.62	71.16	**59.98**	82.43	69.70
RV [6]	56.38	76.12	80.01	66.25	76.80	78.33	67.82	55.62	80.58	71.98	56.40	83.87	70.85
Entropy [7]	55.88	74.14	78.88	59.25	74.52	77.67	64.19	54.39	78.54	67.57	57.23	80.96	68.60
InfoMax [11]	55.88	74.14	78.88	59.25	77.74	77.67	64.19	54.39	78.54	67.57	56.61	80.96	68.82
SND [10]	55.88	74.14	78.88	59.25	74.52	75.21	64.19	54.39	78.54	67.57	56.61	80.96	68.34
Corr-C [13]	51.41	72.00	76.04	59.37	69.36	69.54	61.85	48.04	76.06	69.30	51.71	80.31	65.42
EnsV-W	**57.85**	76.57	**81.04**	66.25	**79.48**	**78.52**	**67.94**	55.62	**82.17**	71.9	59.24	84.03	71.72
EnsV	**57.85**	76.57	80.54	66.25	78.82	**78.52**	**67.94**	**57.07**	**82.17**	71.9	59.24	84.03	**71.74**
Worst	51.41	72.00	76.04	59.25	69.36	69.54	61.85	48.04	76.06	67.57	51.71	80.31	65.26
Best	58.01	77.31	81.04	66.91	79.48	78.52	67.94	57.07	82.17	72.06	59.98	84.03	72.04

Table 3: Validation accuracy (%) of a closed-set UDA method BNM [15] on *Office-Home*.

Method	Ar → Cl	Ar → Pr	Ar → Re	Cl → Ar	Cl → Pr	Cl → Re	Pr → Ar	Pr → Cl	Pr → Re	Re → Ar	Re → Cl	Re → Pr	avg
SourceRisk [1]	56.93	**77.00**	77.74	57.64	**73.33**	69.36	56.45	42.38	77.19	73.22	52.90	82.26	66.37
IWCV [2]	46.46	**77.00**	79.30	63.86	61.34	62.54	63.95	42.38	78.01	71.86	55.65	83.92	65.52
DEV [3]	57.75	71.62	79.30	57.64	67.90	75.46	**66.21**	**54.04**	78.01	**73.42**	57.37	82.25	68.41
RV [6]	**58.67**	**77.00**	79.30	65.68	**73.33**	75.46	65.64	52.05	**81.25**	**73.42**	**59.54**	83.92	70.44
Entropy [7]	53.40	67.04	78.04	63.41	71.44	73.93	63.58	52.69	80.95	71.86	57.37	**83.96**	68.14
InfoMax [11]	53.40	67.04	78.04	63.41	71.44	73.93	63.58	52.69	80.95	71.86	57.37	**83.96**	68.14
SND [10]	53.40	67.04	78.04	63.41	71.44	73.93	63.58	52.69	80.95	71.86	57.37	**83.96**	68.14
Corr-C [13]	46.46	67.04	74.82	49.73	61.34	62.54	56.45	42.38	74.41	68.11	47.26	78.51	60.76
EnsV-W	**58.67**	**77.00**	**80.61**	**66.21**	**73.33**	**76.75**	**66.21**	53.93	**81.25**	**73.42**	57.59	83.92	70.74
EnsV	**58.67**	**77.00**	**80.61**	**66.21**	**73.33**	**76.75**	**66.21**	53.93	**81.25**	**73.42**	**59.54**	83.92	**70.90**
Worst	46.46	67.04	74.82	49.73	61.34	62.54	56.45	42.38	74.41	68.11	47.26	78.51	60.75
Best	58.67	77.00	80.61	67.16	74.16	76.75	66.21	54.04	81.36	73.42	59.82	84.12	71.11

Table 4: Validation accuracy (%) of a closed-set UDA method CDAN [16] on *Office-Home*.

Method	Ar → Cl	Ar → Pr	Ar → Re	Cl → Ar	Cl → Pr	Cl → Re	Pr → Ar	Pr → Cl	Pr → Re	Re → Ar	Re → Cl	Re → Pr	avg
SourceRisk [1]	43.41	62.51	75.51	43.96	61.59	57.70	53.75	37.50	73.22	67.28	47.01	84.39	58.99
IWCV [2]	43.14	62.51	77.81	44.71	54.58	56.14	65.14	37.50	**81.85**	74.08	43.02	84.39	60.41
DEV [3]	57.16	71.75	77.81	62.46	55.64	71.08	65.14	56.54	**81.85**	74.08	57.43	78.89	67.49
RV [6]	57.16	71.75	77.78	63.62	72.92	73.40	65.14	54.50	**81.85**	**74.21**	**58.56**	83.37	69.52
Entropy [7]	**57.55**	72.43	77.74	63.62	72.92	73.40	**65.27**	**56.66**	81.20	74.08	58.47	83.76	69.76
InfoMax [11]	**57.55**	72.43	77.74	63.62	72.92	73.40	**65.27**	**56.66**	81.20	74.08	58.47	83.76	69.76
SND [10]	**57.55**	72.43	77.78	**64.61**	**73.73**	73.40	65.14	**56.66**	**81.85**	74.08	58.47	**84.73**	70.04
Corr-C [13]	43.14	63.05	73.61	43.96	54.58	56.12	51.75	37.50	73.22	65.80	43.00	77.25	56.91
EnsV-W	57.18	73.30	77.78	63.37	73.89	73.38	65.14	55.44	81.36	73.88	**58.56**	84.39	69.81
EnsV	**57.55**	**73.71**	**78.33**	**64.61**	**73.73**	**74.39**	65.14	56.56	**81.85**	73.88	**58.56**	**84.73**	**70.25**
Worst	43.14	62.51	73.61	43.96	54.58	56.12	51.63	37.50	73.22	65.80	43.00	77.25	56.86
Best	57.55	73.71	78.33	64.61	73.89	74.39	65.76	56.66	81.85	74.21	59.50	84.73	70.43

Table 5: Validation accuracy (%) of a closed-set UDA method MCC [17] on *Office-Home*.

Method	Ar → Cl	Ar → Pr	Ar → Re	Cl → Ar	Cl → Pr	Cl → Re	Pr → Ar	Pr → Cl	Pr → Re	Re → Ar	Re → Cl	Re → Pr	avg
SourceRisk [1]	57.23	78.19	**81.75**	60.65	76.50	**78.79**	64.15	53.15	82.17	**74.91**	59.20	83.96	70.89
IWCV [2]	**60.02**	78.15	81.34	68.73	**78.51**	77.85	64.15	**57.85**	81.04	73.18	58.92	84.46	72.02
DEV [3]	57.16	78.15	81.34	**69.10**	73.01	76.80	64.15	**57.85**	82.17	73.18	59.20	84.46	71.38
RV [6]	59.34	**78.53**	80.70	**69.10**	77.83	78.22	**67.20**	**57.85**	82.24	**74.91**	59.20	**85.54**	72.56
Entropy [7]	59.31	**78.53**	81.59	66.87	77.83	**78.79**	**67.20**	**57.85**	**82.51**	73.79	60.82	**85.54**	72.55
InfoMax [11]	**60.02**	74.66	**81.75**	64.98	78.24	78.49	64.15	54.52	82.19	70.62	60.89	84.46	71.25
SND [10]	53.56	77.43	79.46	67.28	76.48	76.80	65.06	54.34	81.04	74.82	58.92	85.24	70.87
Corr-C [13]	53.56	77.43	79.46	67.28	76.48	76.80	65.06	54.34	81.04	74.82	58.92	85.24	70.87
EnsV-W	59.31	77.86	81.59	**69.10**	**78.51**	**78.79**	66.87	**57.85**	82.19	73.79	**61.35**	85.22	**72.70**
EnsV	59.31	77.86	81.59	**69.10**	77.83	**78.79**	66.87	**57.85**	82.19	73.79	**61.35**	85.22	72.65
Worst	53.56	73.44	79.25	60.65	73.01	75.76	59.74	53.15	79.55	67.78	57.18	82.11	67.93
Best	60.02	78.53	81.75	69.22	78.51	78.79	67.90	58.49	82.51	74.91	61.35	85.74	73.14

Table 6: Validation accuracy (%) of a closed-set UDA method MDD [18] on *Office-Home*.

Method	Ar → Cl	Ar → Pr	Ar → Re	Cl → Ar	Cl → Pr	Cl → Re	Pr → Ar	Pr → Cl	Pr → Re	Re → Ar	Re → Cl	Re → Pr	avg
SourceRisk [1]	54.85	73.35	77.05	58.76	69.95	72.23	60.03	51.02	77.36	68.81	57.42	82.50	66.94
IWCV [2]	56.40	69.52	76.59	58.76	67.40	69.43	61.89	56.43	76.82	71.94	56.68	**84.43**	67.19
DEV [3]	57.71	**75.42**	77.05	58.76	**72.99**	70.51	**63.95**	56.43	80.26	70.54	56.68	82.14	68.54
RV [6]	**58.05**	**75.42**	76.59	63.54	69.95	**73.74**	**63.95**	51.02	**80.38**	**72.23**	58.17	**84.43**	68.96
Entropy [7]	57.73	74.54	**78.22**	**64.07**	**72.99**	**73.74**	**63.95**	55.85	**80.38**	71.61	59.31	84.28	69.72
InfoMax [11]	**58.05**	74.54	**78.22**	**64.07**	**72.99**	**73.74**	**63.95**	55.85	**80.38**	71.61	59.31	84.28	**69.75**
SND [10]	**58.05**	**75.42**	77.05	44.99	**72.99**	48.06	37.08	21.60	80.26	71.94	34.39	**84.43**	58.86
Corr-C [13]	39.08	59.74	69.61	44.99	54.58	48.06	37.08	21.60	64.22	61.31	34.39	75.87	50.88
EnsV-W	54.89	**75.42**	78.01	61.89	**72.99**	72.23	63.08	56.43	79.66	**72.23**	**60.02**	83.96	69.23
EnsV	56.40	**75.42**	77.05	**64.07**	**72.99**	72.23	63.08	**57.02**	80.26	**72.23**	**60.02**	**84.43**	69.60
Worst	39.08	59.74	69.61	44.99	54.58	48.06	37.08	21.60	64.22	61.31	34.39	75.87	50.88
Best	58.05	75.42	78.22	64.07	72.99	73.74	63.95	57.02	80.38	72.23	60.02	84.43	70.04

Table 7: Validation accuracy (%) of a closed-set UDA method SAFN [19] on *Office-Home*.

Method	Ar → Cl	Ar → Pr	Ar → Re	Cl → Ar	Cl → Pr	Cl → Re	Pr → Ar	Pr → Cl	Pr → Re	Re → Ar	Re → Cl	Re → Pr	avg
SourceRisk [1]	50.78	69.72	76.06	59.66	70.29	69.86	60.90	46.07	**77.71**	70.05	**57.16**	80.96	65.77
IWCV [2]	50.24	69.72	77.28	62.63	67.24	69.86	58.84	49.69	75.72	**71.45**	**57.16**	79.97	65.82
DEV [3]	51.07	69.72	76.64	59.66	67.24	71.26	58.84	49.69	75.72	70.95	50.65	76.64	64.84
RV [6]	51.07	71.41	76.64	62.63	68.44	70.44	58.84	44.49	**77.71**	**71.45**	54.82	**81.46**	65.78
Entropy [7]	45.93	69.72	75.49	55.29	67.22	68.35	54.26	43.30	75.69	70.00	49.99	80.60	62.99
InfoMax [11]	50.47	69.72	75.49	62.46	**70.98**	68.35	61.23	43.30	75.69	70.00	55.37	80.60	65.31
SND [10]	45.93	64.36	70.60	55.29	60.13	62.50	54.26	43.30	71.43	64.15	49.99	76.64	59.88
Corr-C [13]	45.93	69.72	70.60	55.29	60.13	62.50	61.23	43.30	71.43	**71.45**	49.99	76.64	61.52
EnsV-W	**51.73**	72.07	76.64	**64.65**	**70.98**	71.26	**63.66**	**50.52**	77.48	70.99	**57.16**	**81.46**	**67.38**
EnsV	51.07	**72.27**	**77.30**	63.58	70.29	**71.70**	62.71	49.69	**77.71**	**71.45**	55.78	80.96	67.04
Worst	45.93	64.36	70.60	55.29	60.13	62.50	54.26	43.30	71.43	64.15	49.99	76.64	59.88
Best	51.73	72.27	77.30	64.65	70.98	71.70	63.66	50.52	77.71	71.45	57.16	81.46	67.55

Table 8: Validation accuracy (%) of closed-set UDA methods on *Office-31*.

Method	ATDOC [14]					BNM [15]					CDAN [16]				
	A → D	A → W	D → A	W → A	avg	A → D	A → W	D → A	W → A	avg	A → D	A → W	D → A	W → A	avg
SourceRisk [1]	88.96	**87.80**	73.65	71.46	80.47	**90.36**	**89.43**	73.13	72.70	81.41	91.16	**89.06**	66.33	61.46	77.00
IWCV [2]	86.14	86.54	73.65	71.46	79.45	85.54	**89.43**	73.13	72.70	80.20	69.08	58.74	66.33	61.46	63.90
DEV [3]	86.14	86.54	73.65	71.46	79.45	85.54	**89.43**	73.13	72.70	80.20	91.16	88.30	66.33	61.46	76.81
RV [6]	89.96	87.23	74.28	**75.58**	81.76	88.55	**89.43**	74.90	66.52	79.85	91.16	88.30	**76.18**	**70.36**	81.50
Entropy [7]	86.14	**87.80**	73.87	72.70	80.13	85.54	83.14	71.07	74.26	78.50	91.16	**89.06**	72.88	**70.36**	80.87
InfoMax [11]	86.14	**87.80**	73.87	72.70	80.13	85.54	83.14	71.07	69.97	77.43	91.16	88.30	72.88	**70.36**	80.68
SND [10]	**92.37**	**87.80**	73.87	72.70	81.69	85.54	83.14	74.62	74.26	79.39	92.37	88.55	72.88	70.22	81.01
Corr-C [13]	90.96	84.40	71.88	70.22	79.37	84.34	78.99	67.80	66.52	74.41	67.67	59.62	58.15	58.43	60.97
EnsV-W	**92.37**	**87.80**	**74.65**	75.01	**82.46**	88.55	**89.43**	**75.43**	**75.29**	82.18	**92.77**	88.55	**76.18**	70.22	**81.93**
EnsV	90.96	**87.80**	**74.65**	75.01	82.11	**90.36**	**89.43**	**75.43**	74.30	**82.38**	**92.77**	88.55	**76.18**	70.22	**81.93**
Worst	86.14	84.40	71.88	70.22	78.16	84.34	78.99	67.80	66.52	74.41	67.67	57.11	58.15	58.43	60.34
Best	92.37	87.80	75.04	75.58	82.70	90.36	89.43	75.75	75.29	82.71	92.77	89.06	76.18	70.57	82.15

Table 9: Validation accuracy (%) of closed-set UDA methods on *Office-31*.

Method	MCC [17]					MDD [18]					SAFN [19]				
	A → D	A → W	D → A	W → A	avg	A → D	A → W	D → A	W → A	avg	A → D	A → W	D → A	W → A	avg
SourceRisk [1]	90.96	91.07	73.33	72.89	82.06	91.06	86.23	76.68	74.76	82.18	83.73	87.17	68.96	69.44	77.33
IWCV [2]	91.16	88.55	73.33	72.89	81.48	91.16	89.18	76.68	74.30	82.83	86.55	80.38	68.96	**69.68**	76.39
DEV [3]	89.16	93.08	73.33	72.06	81.91	91.16	89.18	76.68	74.62	82.91	86.55	80.38	68.96	67.45	75.84
RV [6]	89.06	93.08	74.42	73.52	82.52	**92.57**	86.79	73.91	**74.97**	82.07	90.83	87.17	68.76	68.62	78.85
Entropy [7]	90.56	**93.46**	**74.83**	73.02	82.97	**92.57**	**90.82**	**78.03**	74.58	84.00	**91.57**	85.66	67.20	69.26	78.42
InfoMax [11]	89.16	88.55	74.16	**73.70**	81.39	**92.57**	**90.82**	**78.03**	**74.97**	**84.10**	**91.57**	**87.42**	67.20	69.26	78.86
SND [10]	**91.97**	**93.46**	**74.83**	73.02	**83.32**	92.17	**90.82**	**78.03**	**74.97**	84.00	89.96	85.66	67.20	69.26	78.02
Corr-C [13]	91.37	**93.46**	**74.83**	73.02	83.17	91.57	85.66	73.91	74.58	81.43	86.75	80.38	67.09	**69.68**	75.98
EnsV-W	90.56	91.07	74.16	**73.70**	82.37	**92.57**	**90.82**	77.53	74.30	83.80	**91.57**	87.17	**70.22**	69.12	**79.52**
EnsV	90.56	91.45	73.80	**73.70**	82.38	**92.57**	**90.82**	77.53	74.30	83.80	90.96	87.17	**70.22**	69.12	79.37
Worst	86.75	87.17	71.18	69.93	78.76	87.35	85.66	73.91	72.20	79.78	83.73	80.38	67.09	67.45	74.66
Best	91.97	93.46	74.83	74.01	83.57	92.57	92.20	78.03	75.01	84.45	91.57	87.42	70.43	69.68	79.78

Table 10: Validation accuracy (%) of a closed-set UDA method CDAN [16] on *DomainNet-126*.

Method	C → S	P → C	P → R	R → C	R → P	R → S	S → P	avg
Entropy [7]	**58.04**	**64.78**	74.42	**69.39**	**68.65**	**60.63**	**62.94**	**65.55**
InfoMax [11]	**58.04**	**64.78**	74.42	**69.39**	**68.65**	**60.63**	**62.94**	**65.55**
SND [10]	**58.04**	**64.78**	74.42	**69.39**	**68.65**	**60.63**	60.70	65.23
Corr-C [13]	**58.04**	57.73	74.42	56.98	65.07	51.23	60.70	60.60
EnsV-W	55.15	60.98	73.86	60.99	65.07	55.50	60.27	61.69
EnsV	56.73	64.67	**74.44**	67.08	67.97	58.12	62.57	64.51
Worst	51.59	57.73	73.44	56.98	63.06	51.23	58.46	58.93
Best	58.04	64.78	74.44	69.39	68.65	60.63	62.94	65.55

Table 11: Validation accuracy (%) of a closed-set UDA method BNM [15] on *DomainNet-126*.

Method	C → S	P → C	P → R	R → C	R → P	R → S	S → P	avg
Entropy [7]	56.42	61.57	74.31	65.15	65.15	40.95	63.42	61.00
InfoMax [11]	56.42	68.95	74.31	65.15	65.15	54.93	63.42	64.05
SND [10]	43.78	61.57	74.31	51.55	54.40	40.95	54.59	54.45
Corr-C [13]	43.78	60.03	**77.62**	59.47	67.19	40.95	59.64	58.38
EnsV-W	**58.48**	68.42	**77.62**	66.05	**67.79**	**57.65**	**64.34**	65.76
EnsV	57.73	**69.63**	**77.62**	**66.10**	**67.79**	**57.65**	**64.34**	**65.84**
Worst	43.78	60.03	74.31	51.55	54.40	40.95	54.59	54.23
Best	58.48	69.63	78.68	66.10	67.79	58.50	65.20	66.34

Table 12: Validation accuracy (%) of a closed-set UDA method ATDOC [14] on *DomainNet-126*.

Method	C → S	P → C	P → R	R → C	R → P	R → S	S → P	avg
Entropy [7]	46.43	65.98	79.60	61.52	64.24	57.92	59.46	62.16
InfoMax [11]	46.43	65.98	79.60	61.52	64.24	57.92	59.46	62.16
SND [10]	46.43	65.98	79.60	61.52	64.24	47.58	59.46	60.69
Corr-C [13]	54.71	60.63	74.42	59.33	64.58	52.66	59.95	60.90
EnsV-W	**63.12**	69.57	78.33	67.93	69.32	60.85	66.33	67.92
EnsV	62.11	**71.14**	**80.01**	**69.45**	**69.79**	**61.35**	**67.10**	**68.71**
Worst	46.43	60.63	74.42	59.33	64.24	47.58	59.46	58.87
Best	63.12	71.14	80.38	69.45	69.79	61.35	67.10	68.90

Table 13: Validation accuracy (%) of a partial-set UDA method PADA [20] on *Office-Home*.

Method	Ar → Cl	Ar → Pr	Ar → Re	Cl → Ar	Cl → Pr	Cl → Re	Pr → Ar	Pr → Cl	Pr → Re	Re → Ar	Re → Cl	Re → Pr	avg
SourceRisk [1]	45.03	68.85	81.89	43.25	46.83	57.26	57.21	36.42	76.53	71.26	44.30	77.76	58.87
IWCV [2]	**55.58**	65.10	84.54	51.42	**61.29**	53.01	56.93	35.16	81.34	70.52	**60.78**	74.12	62.49
DEV [3]	54.81	**78.15**	78.02	**58.13**	**61.29**	50.14	67.86	35.16	83.21	74.66	57.91	77.76	64.76
RV [6]	43.22	65.10	81.89	42.70	48.74	52.79	57.21	35.16	77.80	73.46	44.30	77.76	58.34
Entropy [7]	40.12	40.11	55.94	52.43	37.25	50.14	57.30	47.22	81.34	70.52	52.18	82.13	55.56
InfoMax [11]	54.81	69.24	78.02	52.43	37.25	50.14	57.30	47.22	71.84	70.52	52.18	74.12	59.59
SND [10]	40.12	40.11	55.94	**58.13**	56.13	64.11	70.62	**51.22**	81.34	74.66	**60.78**	82.13	61.27
Corr-C [13]	40.12	40.11	55.94	54.18	46.89	53.01	58.59	38.93	77.80	71.26	57.91	77.70	56.04
EnsV-W	**55.58**	77.25	86.14	**58.13**	60.17	**67.86**	**73.00**	37.97	**84.04**	76.77	57.91	83.75	**68.21**
EnsV	54.81	69.24	**86.53**	**58.13**	56.13	64.11	70.62	**51.22**	**84.04**	**76.86**	**60.78**	**84.20**	68.06
Worst	40.12	40.11	55.94	41.41	37.25	50.14	56.93	34.87	71.84	70.52	44.24	74.12	51.46
Best	55.58	78.15	86.53	58.13	61.29	68.19	73.00	51.22	84.04	76.86	60.78	84.20	69.83

Table 14: Validation accuracy (%) of a partial-set UDA method SAFN [19] on *Office-Home*.

Method	Ar → Cl	Ar → Pr	Ar → Re	Cl → Ar	Cl → Pr	Cl → Re	Pr → Ar	Pr → Cl	Pr → Re	Re → Ar	Re → Cl	Re → Pr	avg
SourceRisk [1]	**59.40**	77.14	81.34	63.97	67.00	71.29	65.60	46.21	76.81	70.89	58.51	79.10	68.11
IWCV [2]	52.24	74.45	**82.16**	70.98	62.41	70.18	63.45	53.49	76.81	73.65	56.00	78.49	67.86
DEV [3]	55.22	74.45	80.07	70.98	67.00	71.29	63.45	51.70	76.81	73.65	57.91	80.39	68.58
RV [6]	53.67	71.60	81.34	67.58	67.00	73.27	65.70	48.54	76.81	73.65	56.00	79.89	67.92
Entropy [7]	58.93	74.90	80.73	70.98	**74.12**	69.80	70.16	50.09	79.24	74.10	57.85	80.06	70.08
InfoMax [11]	51.82	67.62	76.97	64.65	65.77	69.80	59.69	50.09	74.10	66.67	53.31	75.52	64.67
SND [10]	51.82	74.90	80.73	70.98	**74.12**	**75.10**	70.16	50.09	79.24	74.10	53.31	80.06	69.55
Corr-C [13]	**59.40**	**77.20**	**82.16**	67.58	72.89	**75.10**	70.16	**55.70**	80.12	**75.94**	52.00	**80.73**	70.75
EnsV-W	**59.40**	**77.20**	**82.16**	**71.72**	72.89	74.82	**72.45**	**55.70**	**80.73**	**75.94**	**59.16**	**80.73**	**71.91**
EnsV	55.22	76.30	81.28	67.58	70.31	74.05	70.16	54.63	80.12	75.21	58.51	80.39	70.31
Worst	51.52	67.62	76.97	61.07	62.35	69.80	59.69	46.21	74.10	66.67	52.00	75.52	63.63
Best	59.40	77.20	82.16	71.72	74.12	75.10	72.45	55.70	80.73	75.94	59.16	80.73	72.03

Table 15: H-score [26, 27] (%) of an open-partial-set UDA method DANCE [21] on *Office-Home*.

Method	Ar → Cl	Ar → Pr	Ar → Re	Cl → Ar	Cl → Pr	Cl → Re	Pr → Ar	Pr → Cl	Pr → Re	Re → Ar	Re → Cl	Re → Pr	avg
Entropy [7]	38.29	26.08	36.51	32.92	17.10	32.19	37.69	46.40	45.53	25.39	33.75	39.37	34.27
InfoMax [11]	38.29	26.08	36.51	32.92	17.10	32.19	37.69	46.40	45.33	25.39	33.75	39.37	34.25
SND [10]	1.00	0.00	12.73	0.00	42.84	1.95	19.77	11.99	35.69	25.39	0.00	28.40	14.98
Corr-C [13]	1.00	0.00	12.73	0.00	42.84	1.95	19.77	11.99	35.69	69.02	0.00	28.40	18.62
EnsV-W	**67.00**	75.15	**66.57**	67.87	67.35	59.05	66.41	**62.59**	**69.40**	59.86	**67.54**	73.40	66.85
EnsV	38.40	**76.96**	**66.57**	**71.76**	**75.17**	**69.99**	**77.42**	48.15	**69.40**	**81.84**	**67.54**	**84.31**	**68.96**
Worst	1.00	0.00	12.73	0.00	17.10	1.95	19.77	11.99	35.69	25.39	0.00	28.40	12.84
Best	67.00	76.96	66.57	71.76	75.17	69.99	77.42	64.32	72.87	81.84	67.54	84.31	72.98

Table 16: Validation accuracy (%) of a white-box source-free UDA method SHOT [22] on *Office-Home*.

Method	Ar → Cl	Ar → Pr	Ar → Re	Cl → Ar	Cl → Pr	Cl → Re	Pr → Ar	Pr → Cl	Pr → Re	Re → Ar	Re → Cl	Re → Pr	avg
Entropy [7]	49.14	76.17	79.23	60.57	73.94	74.00	60.69	48.66	79.73	68.89	53.56	81.93	67.21
InfoMax [11]	49.14	76.17	79.23	60.57	73.94	74.00	60.69	48.66	79.73	68.89	53.56	81.93	67.21
SND [10]	49.14	76.17	79.23	60.57	76.59	74.00	64.28	**54.55**	79.73	68.89	58.81	81.93	68.66
Corr-C [13]	55.60	76.66	79.83	67.04	76.59	76.86	66.63	**54.55**	80.74	**73.71**	58.81	**84.61**	70.97
EnsV-W	**56.36**	**77.81**	**81.36**	**68.27**	**78.78**	**78.91**	65.80	54.52	**82.01**	73.01	**59.45**	**84.61**	71.74
EnsV	**56.36**	**77.81**	**81.36**	**68.27**	**78.78**	**78.91**	**67.12**	54.52	**82.01**	73.34	**59.45**	**84.61**	**71.88**
Worst	49.14	76.17	79.23	60.57	73.94	74.00	60.69	48.66	79.73	68.89	53.56	81.93	67.21
Best	56.36	77.95	81.36	68.27	79.05	78.91	67.33	55.33	82.01	73.88	59.54	84.66	72.05

Table 17: Validation accuracy (%) of a white-box source-free UDA method SHOT [22] on *Office-31*.

Method	A → D	A → W	D → A	W → A	avg
Entropy [7]	90.76	88.68	71.21	72.13	80.69
InfoMax [11]	90.76	88.68	71.21	72.13	80.69
SND [10]	90.76	88.68	71.21	72.13	80.69
Corr-C [13]	90.76	90.19	71.21	71.96	81.03
EnsV-W	**94.78**	**91.82**	**75.15**	**74.55**	**84.08**
EnsV	**94.78**	**91.82**	**75.15**	**74.55**	**84.08**
Worst	90.76	88.68	71.21	71.92	80.64
Best	94.78	93.33	75.58	74.55	84.56

References

[1] Ganin, Y., V. Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*. 2015.

[2] Sugiyama, M., M. Krauledat, K.-R. Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 2007.

[3] You, K., X. Wang, M. Long, et al. Towards accurate model selection in deep unsupervised domain adaptation. In *International Conference on Machine Learning*. 2019.

[4] Cortes, C., M. Mohri, M. Riley, et al. Sample selection bias correction theory. In *Algorithmic Learning Theory*. 2008.

[5] Zhong, E., W. Fan, Q. Yang, et al. Cross validation framework to choose amongst models and datasets for transfer learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. 2010.

[6] Ganin, Y., E. Ustinova, H. Ajakan, et al. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 2016.

[7] Morerio, P., J. Cavazza, V. Murino. Minimal-entropy correlation alignment for unsupervised deep domain adaptation. *arXiv preprint arXiv:1711.10288*, 2017.

[8] Grandvalet, Y., Y. Bengio. Semi-supervised learning by entropy minimization. In *Advances in Neural Information Processing Systems*. 2004.

[9] Chapelle, O., A. Zien. Semi-supervised classification by low density separation. In *International Workshop on Artificial Intelligence and Statistics*. 2005.

[10] Saito, K., D. Kim, P. Teterwak, et al. Tune it the right way: Unsupervised validation of domain adaptation via soft neighborhood density. In *IEEE International Conference on Computer Vision*. 2021.

[11] Musgrave, K., S. Belongie, S.-N. Lim. Benchmarking validation methods for unsupervised domain adaptation. *arXiv preprint arXiv:2208.07360*, 2022.

[12] Bridle, J., A. Heading, D. MacKay. Unsupervised classifiers, mutual information and'phantom targets. In *Advances in Neural Information Processing Systems*. 1991.

[13] Tu, W., W. Deng, T. Gedeon, et al. Assessing model out-of-distribution generalization with softmax prediction probability baselines and a correlation method, 2023.

[14] Liang, J., D. Hu, J. Feng. Domain adaptation with auxiliary target domain-oriented classifier. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2021.

[15] Cui, S., S. Wang, J. Zhuo, et al. Towards discriminability and diversity: Batch nuclear-norm maximization under label insufficient situations. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2020.

[16] Long, M., Z. Cao, J. Wang, et al. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*. 2018.

[17] Jin, Y., X. Wang, M. Long, et al. Minimum class confusion for versatile domain adaptation. In *European Conference on Computer Vision*. 2020.

[18] Zhang, Y., T. Liu, M. Long, et al. Bridging theory and algorithm for domain adaptation. In *International Conference on Machine Learning*. 2019.

[19] Xu, R., G. Li, J. Yang, et al. Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation. In *IEEE International Conference on Computer Vision*. 2019.

[20] Cao, Z., L. Ma, M. Long, et al. Partial adversarial domain adaptation. In *European Conference on Computer Vision*. 2018.

[21] Saito, K., D. Kim, S. Sclaroff, et al. Universal domain adaptation through self supervision. In *Advances in Neural Information Processing Systems*. 2020.

[22] Liang, J., D. Hu, J. Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International Conference on Machine Learning*. 2020.

[23] Tsai, Y.-H., W.-C. Hung, S. Schulter, et al. Learning to adapt structured output space for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2018.

[24] Vu, T.-H., H. Jain, M. Bucher, et al. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2019.

[25] Wortsman, M., G. Ilharco, S. Y. Gadre, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning*. 2022.

[26] Fu, B., Z. Cao, M. Long, et al. Learning to detect open classes for universal domain adaptation. In *European Conference on Computer Vision*. 2020.

[27] Bucci, S., M. R. Loghmani, T. Tommasi. On the effectiveness of image rotation for open set domain adaptation. In *European Conference on Computer Vision*. 2020.