Mixed Samples as Probes for Unsupervised Model Selection in Domain Adaptation

Dapeng Hu, Jian Liang, Jun Hao Liew,
Chuhui Xue, Song Bai, Xinchao Wang

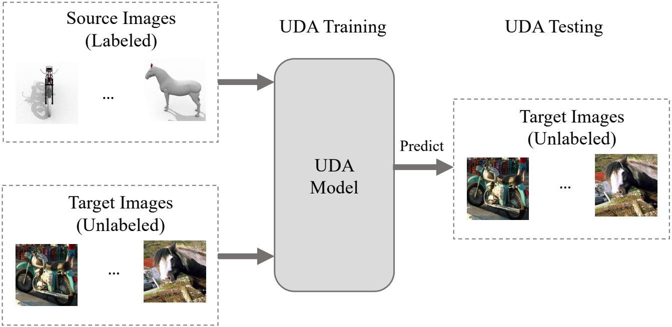
National University of Singapore ByteDance Inc.
CRIPAC & MAIS, Institute of Automation, Chinese Academy of Sciences

November 10, 2023

Outline

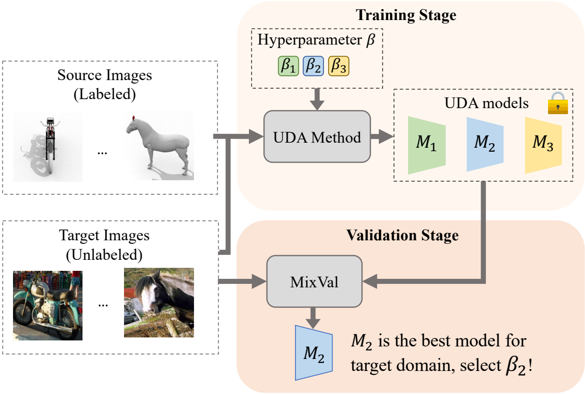
Unsupervised domain adaptation (UDA)

- **Source domain** $\mathcal{D}_s$: n_s labeled samples $\{x_s^i, y_s^i\}_{i=1}^{n_s}$ from $P_S(X,Y)$.
- **Target domain** $\mathcal{D}_t$: n_t unlabeled samples $\{x_t^i\}_{i=1}^{n_t}$ from $P_T(X,Y)$.
- **Goal**: Use $\mathcal{D}_s$ and $\mathcal{D}_t$ during training (transductive) to learn a good UDA model to predict $\{y_t^i\}_{i=1}^{n_t}$ under **domain shift** ($P_S \neq P_T$).



Model selection: a critical problem in UDA

- Without proper hyperparameters, state-of-the-art UDA models can underperform the baseline source-trained model on target data.
- **Model selection**: to select the model with the minimal target risk.
- **Challenges**: no labeled target data; domain distribution shifts.



Previous model selection methods

- **Source-based methods**: use labeled source data for validation.
  - Risk of the labeled source validation set: **SourceVal**.
  - Re-weight the source risk via importance weighting: **IWCV** and **DEV**.
  - Source risk of the model trained in a reversed UDA task: **RV**.

- **Target-only methods**: measure properties of target predictions.
  - Low mean entropy: **Entropy**.
  - Low mean entropy and high category diversity: **InfoMax**.
  - High neighborhood consistency: **SND**.
  - Low category correlation: **Corr-C**.

- Target-only methods are simpler and more versatile compared to source-based methods.

Outline

Our motivation

- Two limitations with existing target-only validation methods: selection via raw predictions; incomplete evaluation of target data structure.

Table: Comparison of assumptions considered in validation methods using target data.

Validation Method	*neighborhood consistency*	no prior of *class diversity*	*low-density separation*
Entropy	✗	✓	✓
InfoMax	✗	✗	✓
SND	✓	✓	✗
Corr-C	✗	✗	✓
MixVal (Ours)	✓	✓	✓

- Our MixVal leverages mixed target samples to directly and comprehensively probe the learned target structure.

MixVal: mixed samples as probes for validation

- **Step 1: Generation of mixed samples** $\{(x^i_{\mathrm{mix}}, y^i_{\mathrm{mix}})\}_{i=1}^{n_{\mathrm{t}}}$. To obtain mixed sample $(x^i_{\mathrm{mix}}, y^i_{\mathrm{mix}})$, we perform *mixup* with a pair of target samples $x^i_{\mathrm{t}}, x^j_{\mathrm{t}}$ and their predicted pseudo labels $\hat{y}^i_{\mathrm{t}}, \hat{y}^j_{\mathrm{t}}$.

$$x_{\mathrm{mix}} = \lambda * x^i_{\mathrm{t}} + (1-\lambda) * x^j_{\mathrm{t}}, \qquad y_{\mathrm{mix}} = \lambda * \hat{y}^i_{\mathrm{t}} + (1-\lambda) * \hat{y}^j_{\mathrm{t}}.$$

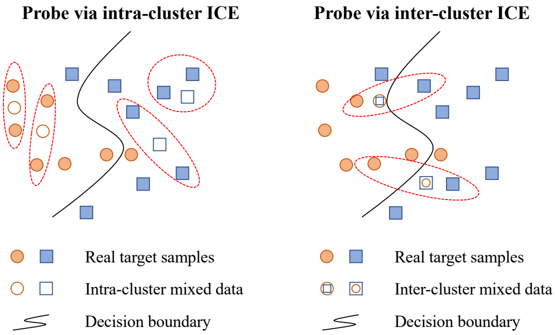
- **Step 2: Interpolation consistency evaluation (ICE)**. With a given UDA model, we conduct inference for all mixed samples $\{x^i_{\mathrm{mix}}\}_{i=1}^{n_{\mathrm{t}}}$, resulting in predicted labels $\{\hat{y}^i_{\mathrm{mix}}\}_{i=1}^{n_{\mathrm{t}}}$ and evaluate the consistency between the mixed labels and predicted labels.

$$\mathrm{ICE} = \mathrm{Accuracy}(\{\mathrm{y}^{\mathrm{i}}_{\mathrm{mix}}\}_{\mathrm{i=1}}^{\mathrm{n_t}}, \{\hat{\mathrm{y}}^{\mathrm{i}}_{\mathrm{mix}}\}_{\mathrm{i=1}}^{\mathrm{n_t}}).$$

MixVal: mixed samples as probes for validation

- MixVal uses two types of probes: intra-cluster mixed samples to probe the neighborhood consistency property and inter-cluster mixed samples to probe the low-density separation property.
- **Metric**: the average ranking of both ICE scores and higher is better.

$$\mathrm{MixVal} = \frac{1}{2}(\mathrm{Ranking}(\mathrm{ICE}_{\mathrm{intra}}) + \mathrm{Ranking}(\mathrm{ICE}_{\mathrm{inter}})).$$



Outline

Setup: UDA methods and hyperparameter settings

UDA method	UDA Type	Hyperparameter	Search Space	Default Value
ATDOC (CVPR'2021)	closed-set self-training	loss coefficient λ	$\{0.02, 0.05, 0.1, 0.2, 0.5, 1.0, 2.0\}$	0.2
BNM (CVPR'2020)	closed-set output regularization	loss coefficient λ	$\{0.02, 0.05, 0.1, 0.2, 0.5, 1.0, 2.0\}$	1.0
CDAN (NeurIPS'2018)	closed-set feature alignment	loss coefficient λ	$\{0.05, 0.1, 0.2, 0.5, 1.0, 2.0, 5.0\}$	1.0
MCC (ECCV'2020)	closed-set output regularization	temperature T	$\{1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0\}$	2.5
MDD (ICML'2019)	closed-set output alignment	margin factor γ	$\{0.5, 1.0, 2.0, 3.0, 4.0, 5.0, 6.0\}$	4.0
SAFN (ICCV'2019)	closed/partial-set feature regularization	loss coefficient λ	$\{0.002, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2\}$	0.05
PADA (ECCV'2018)	partial-set feature alignment	loss coefficient λ	$\{0.05, 0.1, 0.2, 0.5, 1.0, 2.0, 5.0\}$	1.0
DANCE (NeurIPS'2020)	open-partial-set self-supervision	loss coefficient η	$\{0.02, 0.05, 0.1, 0.2, 0.5, 1.0, 2.0\}$	0.05
SHOT (ICML'2020)	source-free hypothesis transfer	loss coefficient β	$\{0.03, 0.05, 0.1, 0.3, 0.5, 1.0, 3.0\}$	0.3
DMRL (ECCV'2020)	closed-set *mixup* training	loss coefficient λ	$\{0.1, 0.2, 0.5, 1.0, 2.0, 5.0, 10.0\}$	2.0
AdaptSeg (CVPR'2018)	closed-set output alignment	loss coefficient λ	$\{0.0001, 0.0003, 0.001, 0.003, 0.01, 0.03\}$	0.0002
AdvEnt (CVPR'2019)	closed-set output alignment	loss coefficient λ	$\{0.0001, 0.0003, 0.001, 0.003, 0.01, 0.03\}$	0.001

Results: closed-set UDA on *Office-Home*

- MixVal outperforms all baselines on this standard benchmark.

Method	ATDOC (CVPR'2021)					BNM (CVPR'2020)					CDAN (NeurIPS'2018)				
	→Ar	→Cl	→Pr	→Re	avg	→Ar	→Cl	→Pr	→Re	avg	→Ar	→Cl	→Pr	→Re	avg
SourceVal	66.63	52.54	78.57	76.61	68.59	62.44	50.74	77.53	74.76	66.37	55.00	42.65	69.50	68.81	58.99
IWCV	67.97	54.03	78.31	79.26	69.89	66.56	48.16	74.09	73.28	65.52	61.31	41.24	67.17	71.93	60.41
DEV	67.39	54.23	77.78	79.39	69.70	65.76	56.39	73.92	77.59	68.41	67.23	57.04	68.76	76.91	67.49
RV	**68.68**	56.13	**78.93**	**79.64**	**70.85**	**68.25**	**56.75**	**78.08**	**78.67**	**70.44**	67.66	56.74	76.01	77.68	69.52
Entropy	63.67	55.83	76.54	78.36	68.60	66.28	54.49	74.15	77.64	68.14	67.66	57.56	76.37	77.45	69.76
InfoMax	63.67	55.63	77.61	78.36	68.82	66.28	54.49	74.15	77.64	68.14	67.66	57.56	76.37	77.45	69.76
SND	63.67	55.63	76.54	77.54	68.34	66.28	54.49	74.15	77.64	68.14	**67.94**	57.56	**76.96**	77.68	**70.04**
Corr-C	63.51	50.39	73.89	73.88	65.42	58.10	45.37	68.97	70.59	60.76	53.84	41.21	64.96	67.65	56.91
MixVal	66.47	**56.87**	78.14	79.20	70.17	67.36	56.18	76.10	78.12	69.44	67.71	**57.78**	76.89	**77.76**	70.03
Worst	62.89	50.39	73.89	73.88	65.26	58.10	45.37	68.96	70.59	60.75	53.80	41.21	64.78	67.65	56.86
Best	68.97	58.35	80.27	80.58	72.04	68.93	57.51	78.43	79.57	71.11	68.19	57.90	77.44	78.19	70.43

Method	MCC (ECCV'2020)					MDD (ICML'2019)					SAFN (ICCV'2019)					*Home*
	→Ar	→Cl	→Pr	→Re	avg	→Ar	→Cl	→Pr	→Re	avg	→Ar	→Cl	→Pr	→Re	avg	AVG
SourceVal	66.57	56.53	79.55	80.90	70.89	62.53	54.43	75.27	75.55	66.94	63.54	51.34	73.66	74.54	65.77	66.26
IWC	68.69	58.93	80.37	80.08	72.02	64.20	56.50	73.78	74.28	67.19	64.31	52.36	72.31	74.29	65.82	66.81
DEV	68.81	58.07	78.54	80.10	71.38	64.42	56.94	76.85	75.94	68.54	63.15	50.47	71.20	74.54	64.84	68.39
RV	**70.40**	58.80	80.63	80.39	**72.56**	**66.57**	55.75	76.60	76.90	68.96	64.31	50.13	73.77	74.93	65.78	69.68
Entropy	69.29	**59.33**	**80.63**	**80.96**	72.55	66.54	57.63	77.27	**77.45**	69.72	59.85	46.41	72.51	73.18	62.99	68.63
InfoMax	66.58	58.48	79.12	80.81	71.25	66.54	57.74	77.27	**77.45**	**69.75**	64.56	49.71	73.77	73.18	65.31	68.84
SND	69.05	55.61	79.72	79.10	70.87	51.34	38.01	**77.61**	68.46	58.86	57.90	46.41	67.04	68.18	59.88	66.02
Corr-C	69.05	55.61	79.72	79.10	70.87	47.79	31.69	63.40	60.63	50.88	62.66	46.41	68.83	68.18	61.52	61.06
MixVal	69.79	59.24	80.47	80.74	**72.56**	65.73	**58.01**	77.36	76.91	69.50	**65.98**	**53.14**	**74.76**	**75.40**	**67.32**	**69.84**
Worst	62.72	54.63	76.19	78.19	67.93	47.79	31.69	63.40	60.63	50.88	57.90	46.41	67.04	68.18	59.88	60.26
Best	70.68	59.95	80.93	81.02	73.14	66.75	58.36	77.61	77.45	70.04	66.59	53.14	74.90	75.57	67.55	70.72

Results: closed-set UDA on *Office-31* and *VisDA*

- MixVal shows larger advantages on the large-scale benchmark *VisDA*.

Method	ATDOC (CVPR'2021)					BNM (CVPR'2020)					CDAN (NeurIPS'2018)				
	→A	→D	→W	avg	T→V	→A	→D	→W	avg	T→V	→A	→D	→W	avg	T→V
SourceVal	72.56	88.96	**87.80**	83.11	67.79	72.92	**90.36**	**89.43**	**84.24**	70.51	63.90	91.16	**89.06**	81.37	64.50
IWCV	72.56	86.14	86.54	81.75	67.79	72.92	85.54	**89.43**	82.63	**76.94**	63.90	69.08	58.74	63.91	64.50
DEV	72.56	86.14	86.54	81.75	70.34	72.92	85.54	**89.43**	82.63	**76.94**	63.90	91.16	88.30	81.12	64.50
RV	**74.93**	89.96	87.23	84.04	**77.37**	70.71	88.55	**89.43**	82.90	74.58	**73.27**	91.16	88.30	84.24	76.02
Entropy	73.29	86.14	**87.80**	82.41	62.85	72.67	85.54	83.14	80.45	58.36	71.62	91.16	**89.06**	83.95	**80.46**
InfoMax	73.29	86.14	**87.80**	82.41	76.49	70.52	85.54	83.14	79.73	58.36	71.62	91.16	88.30	83.69	**80.46**
SND	73.29	**92.37**	**87.80**	**84.49**	**77.37**	74.44	85.54	83.14	81.04	69.65	71.55	92.37	88.55	84.16	**80.46**
Corr-C	71.05	90.96	84.40	82.14	67.79	67.16	84.34	78.99	76.83	70.51	58.29	67.67	59.62	61.86	64.50
MixVal	73.61	90.96	86.54	83.70	**77.37**	**74.97**	86.48	87.00	82.81	74.51	72.73	**92.64**	**89.06**	**84.81**	**80.46**
Worst	71.05	86.14	84.40	80.53	62.85	67.16	84.34	78.99	76.83	23.08	58.29	67.67	57.11	61.02	64.50
Best	75.31	92.37	87.80	85.16	77.37	75.52	90.36	89.43	85.10	76.94	73.38	92.77	89.06	85.07	80.46

Method	MCC (ECCV'2020)					MDD (ICML'2019)					SAFN (ICCV'2019)					*Office*	*VisDA*
	→A	→D	→W	avg	T→V	→A	→D	→W	avg	T→V	→A	→D	→W	avg	T→V	AVG	AVG
SourceVal	73.11	90.96	91.07	85.05	80.46	75.72	91.06	86.23	84.34	72.25	69.20	83.73	87.17	80.03	70.71	83.02	71.04
IWCV	73.11	91.16	88.55	84.27	81.48	75.49	91.16	89.18	85.28	72.25	69.32	86.55	80.38	78.75	66.33	79.43	71.55
DEV	72.70	89.16	93.08	84.98	81.48	75.65	91.16	89.18	85.33	72.25	68.21	86.55	80.38	78.38	66.33	82.36	71.97
RV	73.97	89.06	93.08	85.37	**82.22**	74.46	**92.57**	86.79	84.61	77.23	68.69	90.83	87.17	82.23	66.33	83.90	75.62
Entropy	73.93	90.56	**93.46**	85.98	**82.22**	76.31	**92.57**	90.82	86.57	**78.95**	68.23	**91.57**	85.66	81.82	70.20	83.53	72.17
InfoMax	73.93	89.16	88.55	83.88	81.48	**76.50**	**92.57**	90.82	**86.63**	**78.95**	68.23	**91.57**	**87.42**	**82.41**	70.20	83.13	74.32
SND	73.93	**91.97**	**93.46**	**86.45**	69.35	**76.50**	92.17	90.82	86.50	**78.95**	68.23	89.96	85.66	81.28	58.15	83.99	72.32
Corr-C	73.93	91.37	**93.46**	86.25	69.35	74.25	91.57	85.66	83.83	72.25	68.39	86.75	80.38	78.51	62.52	78.24	67.82
MixVal	**74.09**	91.77	93.21	86.36	81.48	75.97	91.77	**91.74**	86.49	**78.95**	**69.61**	89.96	86.83	82.13	**74.41**	**84.39**	**77.86**
Worst	70.56	86.75	87.17	81.49	69.35	73.06	87.35	85.66	82.02	72.25	67.27	83.73	80.38	77.13	58.15	76.50	58.36
Best	74.42	91.97	93.46	86.62	82.23	76.52	92.57	92.20	87.10	78.95	70.06	91.57	87.42	83.02	75.30	85.34	78.54

Results: closed-set UDA on *DomainNet*

- Notably, on a recent large-scale benchmark *DomainNet*, MixVal consistently and significantly outperforms all target-only model selection baselines, including Entropy and SND.

Method	CDAN (NeurIPS'2018) → C	→ P	→ R	→ S	avg	BNM (CVPR'2020) → C	→ P	→ R	→ S	avg	ATDOC (CVPR'2021) → C	→ P	→ R	→ S	avg
Entropy	**67.09**	**65.80**	**74.42**	**59.34**	**66.66**	63.36	64.28	74.31	48.69	62.66	63.75	61.85	**79.60**	52.17	64.34
InfoMax	**67.09**	**65.80**	**74.42**	**59.34**	**66.66**	67.05	64.28	74.31	55.67	65.33	63.75	61.85	**79.60**	52.17	64.34
SND	**67.09**	64.68	**74.42**	**59.34**	66.38	56.56	54.50	74.31	42.37	56.93	63.75	61.85	**79.60**	47.00	63.05
Corr-C	57.35	62.88	**74.42**	54.63	62.32	59.75	63.41	77.62	42.37	60.79	59.98	62.27	74.42	53.69	62.59
MixVal	**67.09**	**65.80**	**74.42**	**59.34**	**66.66**	**67.84**	**66.40**	**78.68**	**58.49**	**67.85**	**68.94**	**68.44**	**79.60**	**61.73**	**69.68**
Worst	57.35	60.76	73.44	51.41	60.74	55.79	54.50	74.31	42.37	56.74	59.98	61.85	74.42	47.00	60.81
Best	67.09	65.80	74.44	59.34	66.66	67.86	66.50	78.68	58.49	67.88	70.30	68.44	80.38	62.23	70.34

Results: partial-set UDA on *Office-Home*

- MixVal consistently outperforms all other validation baselines. The 'class diversity' prior proves to be detrimental due to the label shift between the two domains.

Method	PADA (ECCV'2018)					SAFN (ICCV'2019)				
	$\to$ Ar	$\to$ Cl	$\to$ Pr	$\to$ Re	avg	$\to$ Ar	$\to$ Cl	$\to$ Pr	$\to$ Re	avg
SourceVal	57.21	41.90	64.48	71.89	58.87	66.82	54.71	74.41	76.48	68.11
IWCV	59.65	50.51	66.84	72.96	62.49	69.36	53.91	71.78	76.38	67.86
DEV	66.88	49.29	72.40	70.46	64.76	69.36	54.94	73.95	76.06	68.58
RV	57.79	40.87	63.87	70.83	58.34	68.98	52.74	72.83	77.14	67.92
Entropy	60.08	46.51	53.16	62.47	55.56	**71.75**	55.62	76.36	76.59	70.08
InfoMax	60.08	**51.40**	60.20	66.67	59.59	63.67	51.74	69.64	73.62	64.67
SND	**67.80**	50.71	59.46	67.13	61.27	**71.75**	51.74	76.36	78.36	69.55
Corr-C	61.34	45.65	54.90	62.25	56.04	71.23	55.70	76.94	**79.13**	70.75
MixVal	67.68	51.01	**72.94**	**78.64**	**67.57**	71.70	**57.91**	**77.08**	78.94	**71.41**
Worst	56.29	39.76	50.49	59.31	51.46	62.48	49.91	68.50	73.62	63.63
Best	69.33	55.86	74.55	79.59	69.83	73.37	58.09	77.35	79.33	72.03

Results: open-partial-set UDA and source-free UDA

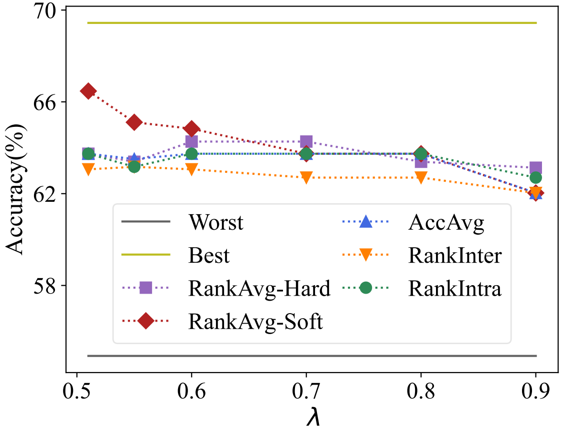
- We are among the first to evaluate validation baselines on open-partial-set UDA (DANCE) and source-free UDA (SHOT). MixVal consistently outperforms all other target-only baselines.

Table: HOS (%) of open-partial-set UDA and accuracy (%) of source-free UDA.

Method	DANCE (NeurIPS'2020)				*Home*	SHOT (ICML'2020)			*Office*	*VisDA*
	→Ar	→Cl	→Pr	→Re	avg	→A	→D	→W	avg	T→V
Entrop	32.00	39.48	27.52	38.08	34.27	71.67	90.76	88.68	83.70	82.65
InfoMax	32.00	39.48	27.52	38.01	34.25	71.67	90.76	88.68	83.70	82.65
SND	15.05	4.33	23.75	16.79	14.98	71.67	90.76	88.68	83.70	82.65
Corr-C	29.60	4.33	23.75	16.79	18.62	71.58	90.76	90.19	84.18	82.65
MixVal	**71.54**	**52.90**	**78.61**	**65.01**	**67.01**	**72.04**	**92.37**	**92.32**	**85.58**	**83.12**
Worst	15.05	4.33	15.17	16.79	12.84	71.56	90.76	88.68	83.67	80.57
Best	77.01	66.29	78.81	69.81	72.98	75.06	94.78	93.33	87.72	83.12

Analysis: ablation of probing in MixVal

- Observations from the PADA results on *Office-Home*: 1) A value of λ near 0.5 is better than a value of λ near 1. 2) Hard pseudo labels are simpler and more stable than soft ones. 3) Both types of probes are effective, and their combination enhances stability.



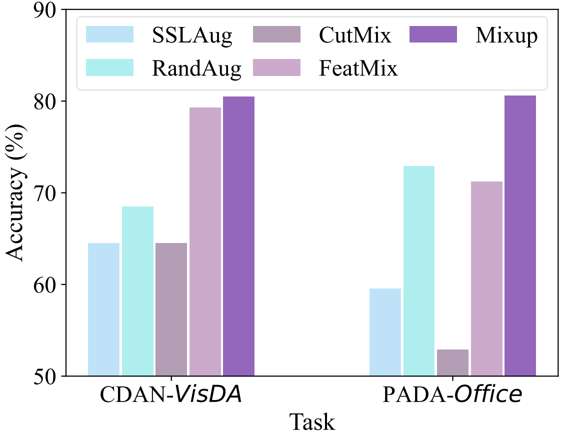
Analysis: further analysis of probing in MixVal

- When a metric inspired by linear discriminant analysis (LDA) is used to measure the raw predictions, which involves minimizing intra-cluster variance while maximizing inter-cluster variance, we observe that this LDA metric outperforms other measurements but falls short of MixVal's performance, which adopts a direct probing with mixed samples.

Method	PADA (ECCV'2018)					SAFN (ICCV'2019)					*Home*
	$\rightarrow$ Ar	$\rightarrow$ Cl	$\rightarrow$ Pr	$\rightarrow$ Re	avg	$\rightarrow$ Ar	$\rightarrow$ Cl	$\rightarrow$ Pr	$\rightarrow$ Re	avg	AVG
SourceVal	57.21	41.90	64.48	71.89	58.87	66.82	54.71	74.41	76.48	68.11	63.49
Entropy	60.08	46.51	53.16	62.47	55.56	**71.75**	55.62	76.36	76.59	70.08	62.82
InfoMax	60.08	**51.40**	60.20	66.67	59.59	63.67	51.74	69.64	73.62	64.67	62.13
SND	**67.80**	50.71	59.46	67.13	61.27	**71.75**	51.74	76.36	78.36	69.55	65.41
Corr-C	61.34	45.65	54.90	62.25	56.04	71.23	55.70	76.94	**79.13**	70.75	63.40
LDA	64.52	46.51	69.47	72.67	63.29	**71.75**	54.39	73.93	77.10	69.29	66.29
MixVal	67.68	51.01	**72.94**	**78.64**	**67.57**	71.70	**57.91**	**77.08**	78.94	**71.41**	**69.49**
Worst	56.29	39.76	50.49	59.31	51.46	62.48	49.91	68.50	73.62	63.63	57.55
Best	69.33	55.86	74.55	79.59	69.83	73.37	58.09	77.35	79.33	72.03	70.93

Analysis: influence of consistency evaluation

- The use of image-level *mixup* in MixVal yields superior results compared to other strategies, including instance-based augmentations like RandAug and SSLAug, as well as *mixup* at different levels such as CutMix and FeatMix.



Analysis: attack of mixup-based UDA method

- Even when particularly using MixVal to validate the UDA method DMRL, which incorporates mixup-based consistency training within the target domain, MixVal continues to demonstrate superior performance.

Method	DMRL (ECCV'2020)				*Home*	DMRL (ECCV'2020)			*Office*
	→Ar	→Cl	→Pr	→Re	avg	→A	→D	→W	avg
Entropy	58.14	50.25	69.06	71.37	62.20	63.67	80.52	**86.67**	76.95
InfoMax	58.14	**50.75**	69.06	71.37	62.33	63.67	80.52	**86.67**	76.95
SND	57.74	49.96	**69.28**	71.42	62.10	61.84	**84.14**	**86.67**	77.55
Corr-C	58.29	49.46	68.67	71.73	62.04	60.23	77.51	81.13	72.95
MixVal	**59.13**	50.41	**69.28**	**72.07**	**62.72**	**64.89**	82.93	**86.67**	**78.16**
Worst	57.71	48.99	67.78	70.72	61.30	60.23	76.31	81.13	72.55
Best	59.20	50.75	69.28	72.24	62.87	65.30	84.14	86.67	78.70

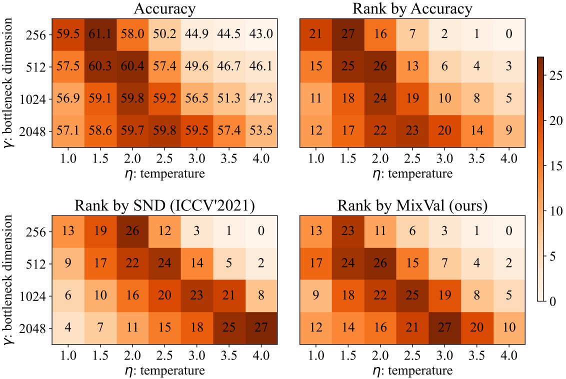
Analysis: two-hyperparameter validation

- The addition of a hyperparameter for the bottleneck dimension in MCC and MDD ranging from $\{256, 512, 1024, 2048\}$ is considered. MixVal maintains its leading performance when evaluated on *Office-Home*.

Method	MDD (ICML'2019)					MCC (ECCV'2020)				
	Ar → Cl	Cl → Pr	Pr → Re	Re → Ar	avg	Ar → Cl	Cl → Pr	Pr → Re	Re → Ar	avg
SourceVal	55.99	73.15	78.77	69.39	69.33	57.91	76.84	81.13	72.89	72.19
IWCV	37.89	72.92	80.42	58.43	62.42	46.09	77.74	80.68	**74.45**	69.74
DEV	52.60	72.11	53.36	67.70	61.44	59.47	76.84	81.94	74.08	73.08
RV	**57.59**	72.25	**80.83**	70.79	70.37	59.13	76.84	82.03	71.98	72.50
Entropy	57.21	**73.19**	80.06	**72.31**	70.69	59.75	77.77	82.37	74.33	73.56
InfoMax	**57.59**	72.92	80.06	**72.31**	**70.72**	59.70	**78.73**	**82.58**	70.33	72.84
SND	38.10	56.45	70.03	65.10	57.42	53.49	74.97	77.25	74.12	69.96
Corr-C	30.17	44.74	57.15	50.76	45.71	44.90	56.75	74.32	67.61	60.90
MixVal	55.99	72.63	80.27	72.12	70.25	**60.08**	78.52	81.95	74.43	**73.75**
Worst	30.17	39.81	53.36	50.76	43.53	43.02	56.75	73.47	67.24	60.12
Best	57.59	73.35	80.93	72.52	71.10	61.10	78.94	83.04	75.36	74.61

Analysis: two-hyperparameter validation

- Qualitative comparisons between SND and MixVal when validating MCC on Ar→Cl, with γ as the bottleneck dimension and η as the temperature. MixVal scores demonstrate a strong correlation with actual accuracy, whereas SND scores exhibit noticeable deviation.



Analysis: two-hyperparameter validation

- The addition of a hyperparameter for the training iteration in semantic segmentation UDA methods ranging from 10k to 30k with the 2k as the step. MixVal maintains the leading performance.

Table: Segmentation mIoU (%) on *GTAV* → *Cityscapes*.

Method	AdaptSegt (CVPR'2018)	AdvEnt (CVPR'2019)
SourceVal	39.52	39.08
Entropy	39.47	38.41
SND	40.69	**40.02**
MixVal	**42.20**	**40.02**
Worst	33.84	33.06
Best	42.20	41.78

Analysis: influence of backbone

- When ViT is used as the backbone, MixVal maintains its superior performance, while validation baselines including Entropy, InfoMax, and Corr-C collapse.

Table: ViT accuracy (%) on R $\rightarrow$S.

Method	BNM (CVPR'2020)
Entropy	28.21
InfoMax	28.21
SND	52.42
Corr-C	28.21
MixVal	**54.78**
Worst	28.21
Best	55.16

Summary: take home message

- Unsupervised model selection is significant for real-world deployment of UDA methods yet overlooked by mainstream studies.
- We propose a novel and simple target-only validation method MixVal, which uses two types of mixed target samples to directly probe the target structure, considering both neighborhood consistency and low-density separation.
- Our comprehensive experiments highlight MixVal's exceptional stability and superior performance in UDA model selection compared to existing baselines.

Thank you!

- Code is available at `https://github.com/LHXXHB/MixVal`.

- If you require any further information, feel free to contact me.

Email: lhxxhb15@gmail.com