NeurIPS 2022

Unknown-Aware Domain Adversarial Learning for Open-Set Domain Adaptation

JoonHo Jang[1], Byeonghu Na[1], DongHyeok Shin[1], Mingi Ji[1], Kyungwoo Song[2], Il-Chul Moon[1, 3]

1 KAIST
2 University of Seoul
3 summary.ai

Correspondence to: Il-Chul Moon <icmoon@kaist.ac.kr>

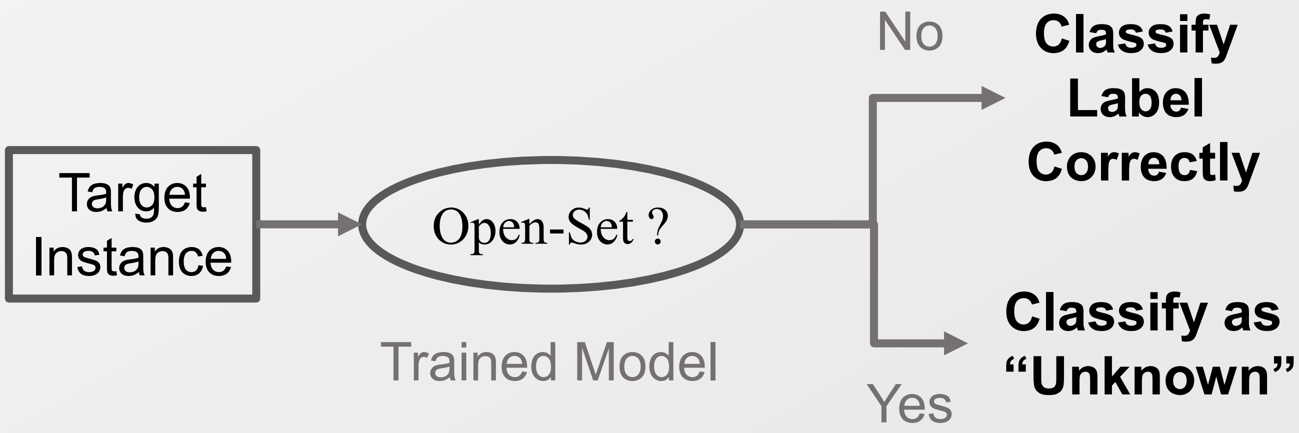
Open-Set Domain Adaptation

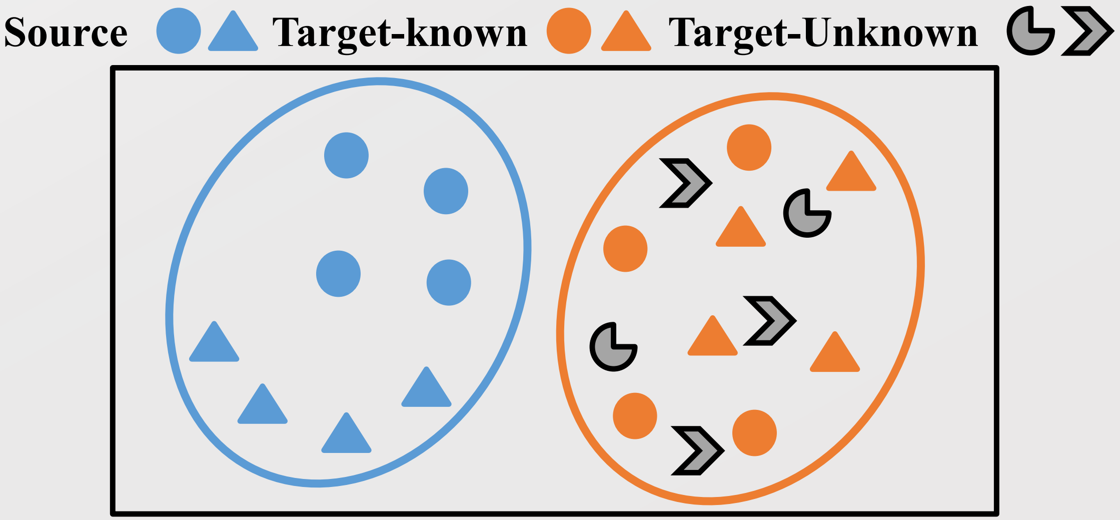
- (Unsupervised) Domain Adaptation
  - We train a model to get high accuracy on the **unlabeled** *target* domain by leveraging the fully labeled source domain knowledge.

- Open-Set Domain Adaptation :
  - However, in a realistic scenario, the target domain may have additional classes called “Open-Set”.
  - Domain Adaptation where the target domain contains unknown classes.

Unknown

- Objective:



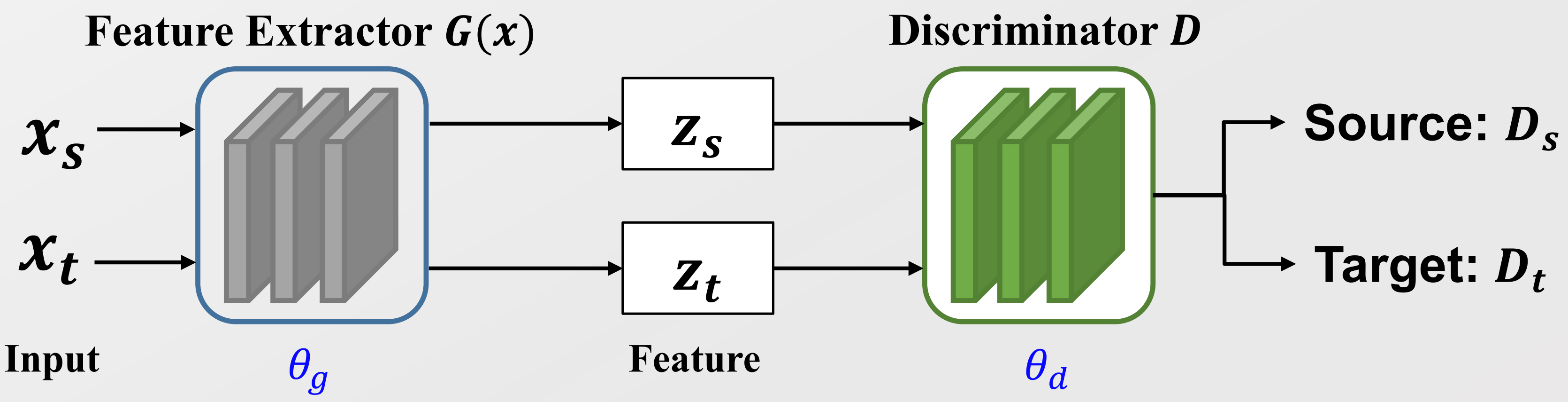


Open-Set Domain Adaptation Setting

Preliminary

- Domain Adaptation
  - Target Classification Error ≤ Source Classification Error + Distribution Matching

- Domain Adversarial Learning
  - The adversarial framework adapts **the feature extractor G** toward indistinguishable feature distributions between the source and the target domain by the ***minimax game*** with **domain discriminator D.**
    - Domain Discriminator: $D\big(G(x)\big) = [D_s\big(G(x)\big), D_t(G(x))]$

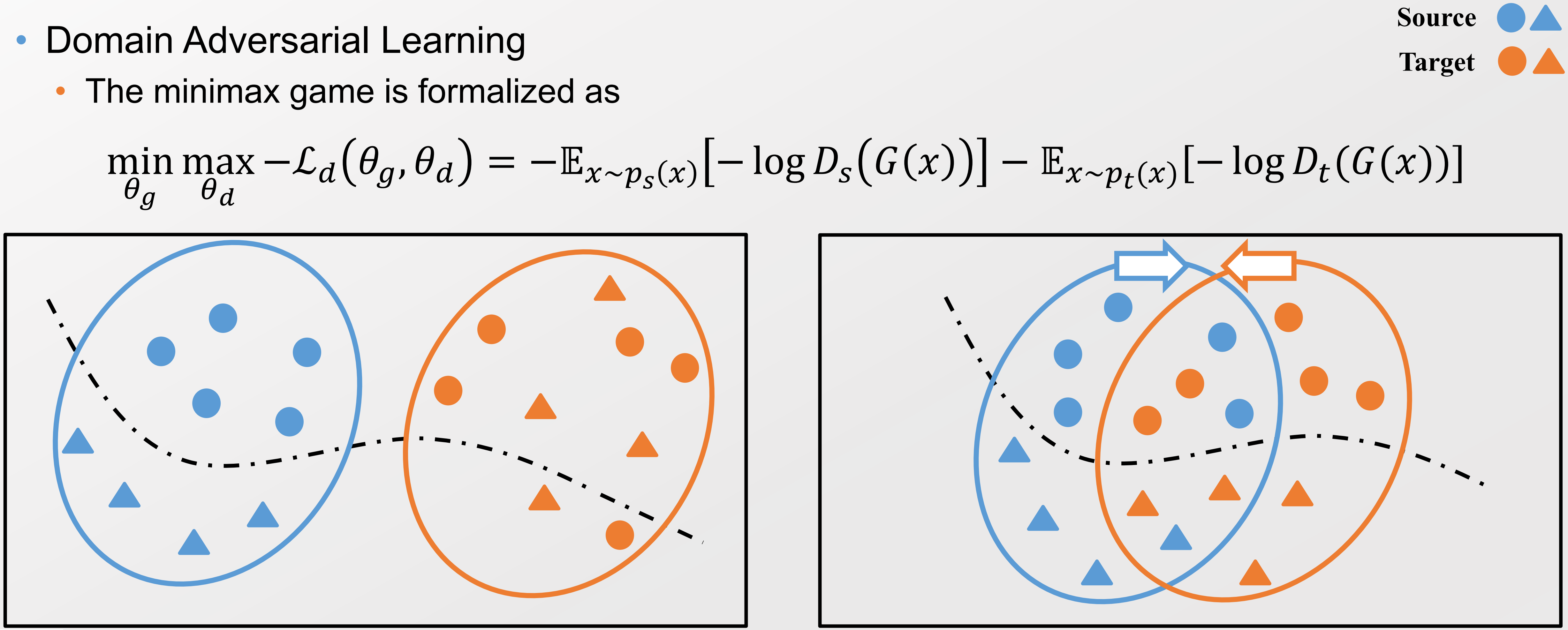


Preliminary

- Domain Adaptation
  - Target Classification Error ≤ Source Classification Error + Distribution Matching

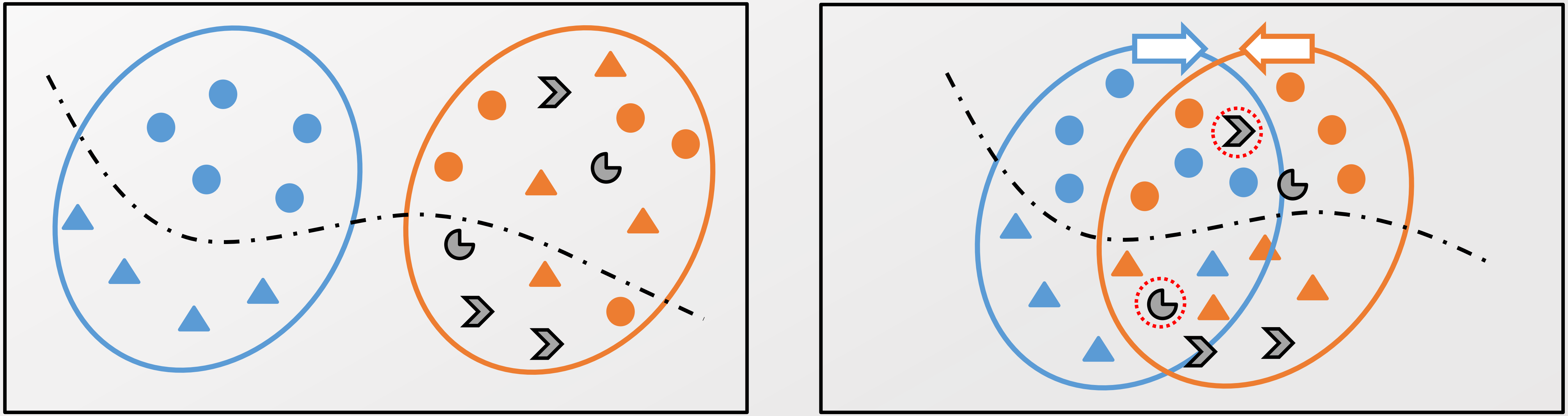
- Domain Adversarial Learning
  - The minimax game is formalized as

$$\min_{\theta_g} \max_{\theta_d} -\mathcal{L}_d(\theta_g, \theta_d) = -\mathbb{E}_{x\sim p_s(x)}[-\log D_s(G(x))] - \mathbb{E}_{x\sim p_t(x)}[-\log D_t(G(x))]$$



Preliminary

- Open-Set Domain Adaptation
  - Target domain contains “Unknown” classes.
- Domain Adversarial Learning to Open-Set Domain Adaptation



It enforces to include the target-*unknown* features in the distribution matching.
→ performance degradation by negative transfer.

Methodology

Domain Adversarial Learning is essential part for feature distribution matching.

For Open-Set Domain Adaptation, the existing approaches of Domain Adversarial Learning is not applicable directly due to the existence of target-*unknown* features.

Domain Adversarial Learning should be designed **simultaneously** to **align** source and target-*known* and to **segregate** target-*unknown* features.

Therefore, we propose **Unknown-Aware Domain Adversarial Learning (UADAL)** for Open-Set Domain Adaptation.

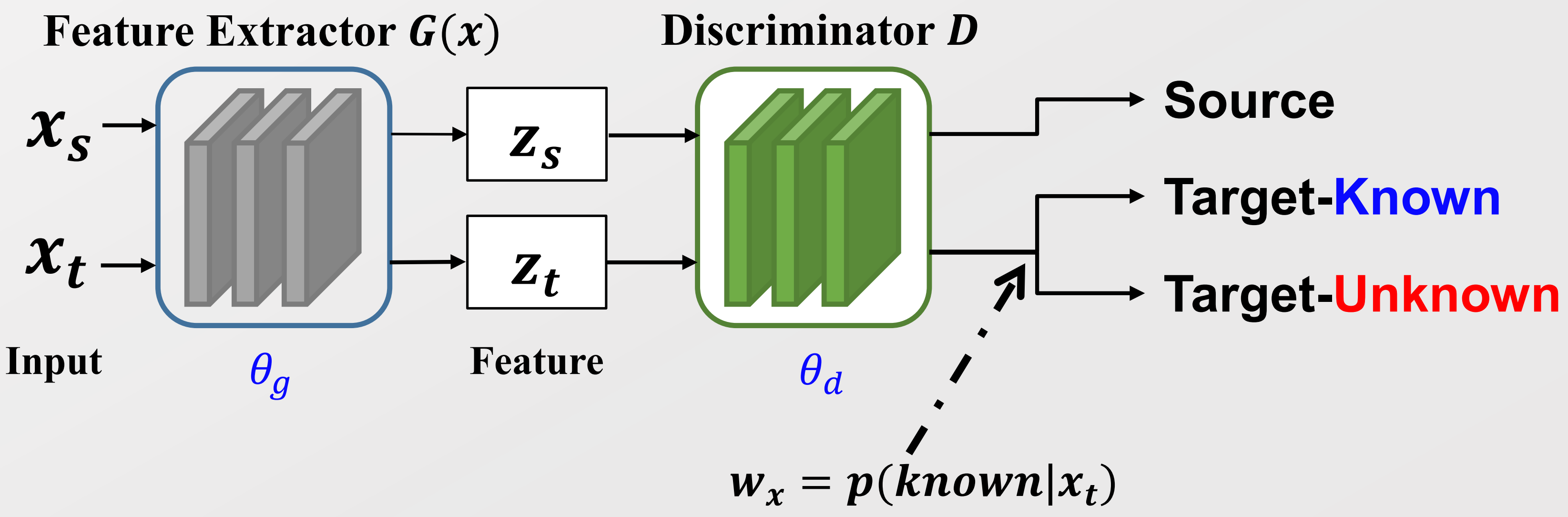
Methodology

- Unknown-Aware Domain Adversarial Learning
  - Domain Discriminator should be able to identify three domain types:
    - Source (s), Target-Known (tk), and Target-Unknown (tu)

$$D(G(x)) = [D_s(G(x)), D_{tk}(G(x)), D_{tu}(G(x))]$$

  - Domain Discrimination Loss

$$\mathcal{L}_d(\theta_g, \theta_d) = \mathbb{E}_{x \sim p_s(x)}[-\log D_s(G(x))] + \mathbb{E}_{x \sim p_t(x)}[-w_x \log D_{tk}(G(x)) - (1 - w_x) \log D_{tu}(G(x))]$$



Methodology

- Unknown-Aware Domain Adversarial Learning
  - Domain Discriminator should be able to identify three domain types:
    - Source (s), Target-Known (tk), and Target-Unknown (tu)

$$D\big(G(x)\big) = \big[D_s\big(G(x)\big), D_{tk}\big(G(x)\big), D_{tu}\big(G(x)\big)\big]$$

  - Domain Discrimination Loss

$$\mathcal{L}_d(\theta_g, \theta_d) = \underbrace{\mathbb{E}_{x\sim p_s(x)}\big[-\log D_s\big(G(x)\big)\big]}_{\mathcal{L}_d^s(\theta_g, \theta_d)} + \underbrace{\mathbb{E}_{x\sim p_t(x)}\big[-w_x \log D_{tk}\big(G(x)\big) - (1-w_x)\log D_{tu}\big(G(x)\big)\big]}_{\text{decompose}\ \mathcal{L}_d^t = \mathcal{L}_d^{tk}(\theta_g, \theta_d) + \mathcal{L}_d^{tu}(\theta_g, \theta_d)}$$

  - Sequential Optimization

$$\min_{\theta_d} \mathcal{L}_D(\theta_g, \theta_d) = \mathcal{L}_d^s(\theta_g, \theta_d) + \mathcal{L}_d^{tk}(\theta_g, \theta_d) + \mathcal{L}_d^{tu}(\theta_g, \theta_d)$$

$$\max_{\theta_g} \mathcal{L}_G(\theta_g, \theta_d) = \mathcal{L}_d^s(\theta_g, \theta_d) + \mathcal{L}_d^{tk}(\theta_g, \theta_d) - \mathcal{L}_d^{tu}(\theta_g, \theta_d)$$

Methodology

- Unknown-Aware Domain Adversarial Learning
  - Sequential Optimization

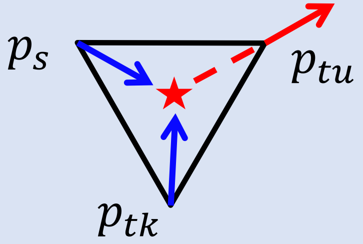
$$\min_{\theta_d} \mathcal{L}_D(\theta_g, \theta_d) = \mathcal{L}_d^s(\theta_g, \theta_d) + \mathcal{L}_d^{tk}(\theta_g, \theta_d) + \mathcal{L}_d^{tu}(\theta_g, \theta_d)$$

$$\boxed{\max_{\theta_g} \mathcal{L}_G(\theta_g, \theta_d) = \mathcal{L}_d^s(\theta_g, \theta_d) + \mathcal{L}_d^{tk}(\theta_g, \theta_d) - \mathcal{L}_d^{tu}(\theta_g, \theta_d)}$$

$$D^*(z) = \left[\frac{p_s(z)}{2p_{avg}(z)}, \frac{\lambda_{tk} p_{tk}(z)}{2p_{avg}(z)}, \frac{\lambda_{tu} p_{tu}(z)}{2p_{avg}(z)}\right], \quad p_{avg}(z) = (p_s(z) + \lambda_{tk} p_{tk}(z) + \lambda_{tu} p_{tu}(z))/2,$$

$$z = G(x) \text{ with fixed } G.$$

$$\min_{\theta_g} -\mathcal{L}_G(\theta_g, \theta_d^*) = D_{KL}(p_s \parallel p_{avg}) + \lambda_{tk} D_{KL}(p_{tk} \parallel p_{avg}) - \lambda_{tu} D_{KL}(p_{tu} \parallel p_{avg}) + C_0$$



Alignment on s and tk.
$p_s \approx p_{tk}$

Segregation on tu.
$p_{tu} \leftrightarrow \{p_{tk}, p_s\}$

[Theorem 3.1.]

- Open-Set Recognition
  - Motivation: Given Decision boundary on known-classes by the source domain,

Target-Known Instances ⟶ Certain Classification Case ⟶ Low Entropy, $\ell_x \downarrow$

Target-Unknown Instances ⟶ Uncertain Classification Case ⟶ High Entropy, $\ell_x \uparrow$

- Posterior Inference

$$\hat{w}_x \coloneqq p(known|\ell_x) = \frac{\lambda_{tk} p(\ell_x|known)}{\lambda_{tk} p(\ell_x|known) + \lambda_{tu} p(\ell_x|unknown)}$$

By fitting Beta Mixture Model

- Open-Set Classification
  - **Classifier $\boldsymbol{C}$** is the extended classifier with the dimensions including the *unknown* class, y_{unk}.

$$\mathcal{L}_{cls}(\theta_g, \theta_c) = \underbrace{\sum_{(x_s, y_s) \in X_s} \mathcal{L}_{CE}(C(G(x_s)), y_s)}_{\text{Source Classification}} + \underbrace{\sum_{x_t \in X_t} \underline{(1 - \hat{w}_x)} \mathcal{L}_{CE}\left(C\left(G(x_t)\right), y_{unk}\right)}_{\text{Target Unknown Classification}} + \underbrace{\sum_{x_t \in X_t} \mathcal{L}_H \left(C(G(x_t))\right)}_{\text{Target Entropy Min.}}$$

$(1 - \hat{w}_x) \uparrow \longrightarrow y_{unk} \uparrow$

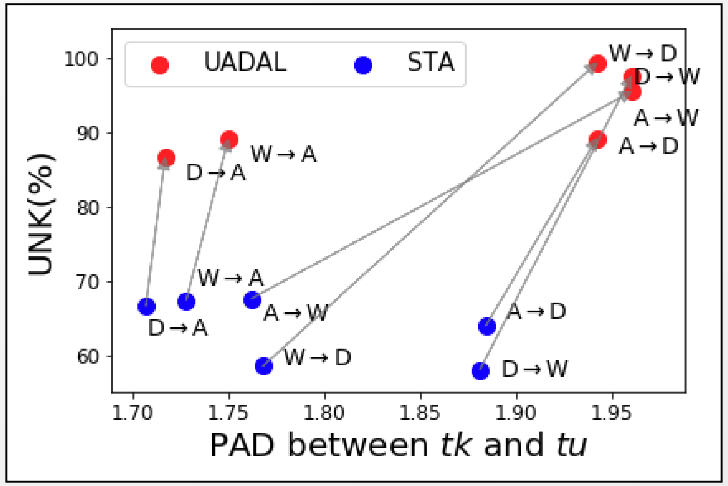
Experimental Part

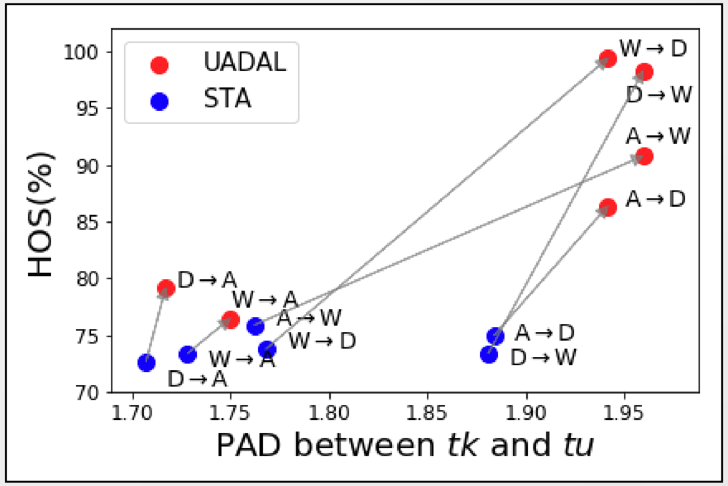
- Experimental Results
  - We conducted the experiments on Office-31 and Office-Home with three backbone networks.
    - In order to show the robustness of the architecture choice.

Backbone (#)/	Model	Office-31							Office-Home												
		A-W	A-D	D-W	W-D	D-A	W-A	Avg.	P-R	P-C	P-A	A-P	A-R	A-C	R-A	R-P	R-C	C-R	C-A	C-P	Avg.
EfficientNet-B0 (5.3M)	DANN	63.2	72.7	92.6	94.8	63.7	57.2	74.0±0.3	35.7	16.5	18.2	34.1	46.3	22.9	40.7	47.8	28.2	12.4	7.5	13.4	27.0±0.3
	CDAN	65.5	73.6	92.4	94.6	64.8	57.9	74.8±0.2	37.9	18.1	20.4	35.6	47.0	24.6	44.1	49.8	30.1	13.5	8.9	15.0	28.8±0.6
	STA	58.3	62.2	81.6	79.6	69.8	67.4	69.8±1.2	59.4	43.6	51.9	53.8	60.6	49.5	58.8	53.5	49.9	53.4	49.5	49.4	52.8±0.2
	OSBP	82.9	87.0	33.8	96.7	27.3	69.9	66.3±2.1	65.0	46.0	58.6	64.2	71.0	54.0	58.3	62.5	50.3	63.7	50.7	55.6	58.3±1.6
	ROS	69.7	80.1	94.7	99.6	73.0	59.2	79.4±0.3	66.9	44.9	53.7	62.5	69.5	50.0	62.0	67.0	52.0	61.2	50.5	54.7	57.9±0.1
	DANCE	68.1	68.8	91.3	85.0	68.5	63.3	74.2±4.0	17.2	47.5	7.2	26.6	19.6	36.6	2.2	19.8	10.9	6.4	4.3	19.0	18.1±2.9
	DCC	87.2	69.1	89.4	94.4	63.5	**76.1**	79.9±2.9	72.2	41.0	56.5	**66.4**	**75.7**	52.8	55.9	71.5	49.9	60.4	48.1	60.8	59.3±1.5
	UADAL	**87.5**	88.3	**97.4**	**96.9**	**74.1**	68.9	85.5±0.5	**75.0**	50.0	62.9	**66.4**	74.1	52.7	**71.5**	72.6	**53.6**	**65.3**	60.8	**63.7**	64.1±0.1
	cUADAL	86.5	**89.1**	97.3	98.0	72.5	71.0	**85.7±0.7**	74.7	**54.4**	**64.2**	66.3	73.9	50.8	71.4	**73.0**	52.4	**65.3**	**61.0**	63.3	**64.2±0.1**
DenseNet-121 (7.9M)	DANN	71.9	72.0	90.2	85.3	73.8	72.3	77.6±0.5	68.8	35.4	48.7	62.6	71.9	45.3	62.8	68.7	45.9	62.2	47.0	54.7	56.2±0.3
	CDAN	69.5	69.8	86.8	84.5	73.8	72.5	76.2±0.2	68.9	39.2	51.9	62.6	71.8	47.1	63.6	68.0	48.7	62.8	49.3	55.2	57.4±0.3
	STA	77.0	68.6	84.0	77.2	76.6	75.1	76.4±1.5	65.6	46.1	58.4	55.8	64.3	50.4	62.6	58.6	51.1	61.0	56.0	55.9	57.1±0.1
	OSBP	81.9	83.0	88.9	96.6	73.1	74.9	83.1±2.2	71.9	46.0	60.3	67.1	72.3	54.5	65.9	71.7	53.7	66.8	59.3	64.1	62.8±0.1
	ROS	67.0	67.8	**97.4**	99.4	77.1	71.8	80.1±1.3	73.0	49.6	59.2	67.8	75.5	52.8	66.4	74.6	54.3	64.8	53.0	57.8	62.4±0.1
	DANCE	69.9	67.8	84.0	82.8	**79.9**	**81.1**	77.6±0.3	51.8	**51.0**	59.7	63.9	58.2	**58.2**	43.4	48.9	55.0	41.3	54.6	60.6	53.9±0.5
	DCC	83.9	80.8	88.4	93.1	79.7	80.4	84.4±1.3	75.1	46.6	58.0	**70.8**	**78.6**	56.6	63.4	75.5	55.8	**71.3**	55.0	63.3	64.2±0.2
	UADAL	**86.0**	82.3	96.7	99.2	77.9	74.2	86.0±0.6	**75.7**	45.5	61.5	70.0	76.9	57.3	71.5	76.1	**60.4**	70.0	**60.1**	67.2	66.0 ±0.2
	cUADAL	85.1	**83.6**	96.4	**99.6**	77.5	75.9	**86.4±0.6**	75.6	48.9	**61.7**	70.0	76.7	57.8	**71.9**	**76.7**	59.1	69.6	**60.1**	**67.5**	**66.3±0.3**
ResNet-50 (25.5M)	DANN	68.1	71.5	86.7	82.5	73.7	72.6	75.9±0.5	69.8	44.6	56.3	65.2	71.0	51.2	65.4	68.4	50.9	66.7	57.6	60.9	60.7±0.2
	CDAN	64.9	66.8	84.3	80.5	72.7	71.0	73.4±1.3	69.7	47.2	58.6	65.1	70.7	52.9	66.0	67.6	52.7	67.1	58.2	61.7	61.4±0.3
	STA*	75.9	75.0	69.8	75.2	73.2	66.1	72.5±0.8	69.5	53.2	61.9	54.0	68.3	55.8	67.1	64.5	54.5	66.8	57.4	60.4	61.1±0.3
	OSBP*	82.7	82.4	97.2	91.1	75.1	73.7	83.7±0.4	73.9	53.2	63.2	65.2	72.9	55.1	66.7	72.3	54.5	70.6	64.3	64.7	64.7±0.2
	PGL*	74.6	72.8	76.5	72.2	69.5	70.1	72.6±1.5	41.6	46.6	47.2	45.6	55.8	29.3	11.4	52.5	0.0	45.6	10.0	36.8	35.2
	ROS*	82.1	82.4	96.0	**99.7**	77.9	77.2	85.9±0.2	74.4	56.3	60.6	69.3	76.5	60.1	68.8	75.7	60.4	68.6	58.9	65.2	66.2±0.3
	DANCE	66.9	70.7	80.0	84.8	65.8	70.2	73.1±1.0	41.2	55.7	54.2	49.8	39.4	53.1	27.5	44.0	48.3	30.2	40.9	45.9	44.2±0.6
	DCC*	87.1	85.5	91.2	87.1	**85.5**	**84.4**	86.8	64.0	52.8	59.5	67.4	**80.6**	52.9	56.0	62.7	**76.9**	67.0	49.8	66.6	64.2
	OSLPP*	89.0	**91.5**	92.3	93.6	79.3	78.7	87.4	74.0	**59.3**	**63.6**	**72.8**	74.3	61.0	67.2	74.4	59.0	70.4	60.9	66.9	67.0
	UADAL	89.1	86.0	97.8	99.5	79.7	76.5	88.1±0.2	**76.9**	56.6	63.0	70.8	77.4	63.2	72.1	**76.8**	60.6	**73.4**	64.2	**69.5**	**68.7±0.2**
	cUADAL	**90.1**	87.9	**98.2**	99.4	80.5	75.1	**88.5±0.3**	76.8	54.6	62.9	71.6	77.5	**63.6**	**72.6**	76.7	59.9	72.6	**65.0**	68.3	68.5±0.1

KAIST

- Experimental Results
  - Correlation analysis between the **PAD on tk and tu** and the evaluation metrics, **HOS** and **UNK**.



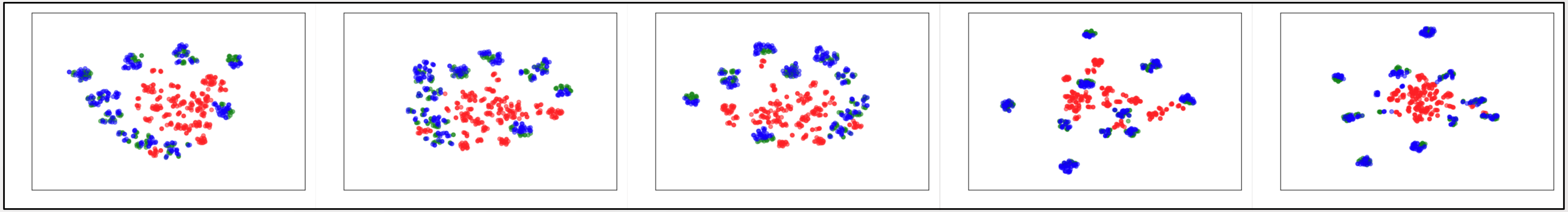


Segregation ↑

⇩

HOS, UNK ↑

  - t-SNE Analysis



DANN OSBP STA DCC UADAL

Conclusion

- We proposed Unknown-Aware Domain Adversarial Learning (UADAL) for Open-Set Domain Adaptation.
  - The first approach to explicitly design the *segregation* of the target-unknown features (tu) in the domain adversarial learning framework for Open-Set Domain Adaptation.

- We design a new domain discrimination loss and formulate the sequential optimization for the unknown-aware feature alignment.
  - By replacing a two-way domain discriminator with the three-way to handle tu information.
  - Providing theoretical analyses on the optimized state of the proposed feature alignment.

- We evaluate UADAL on the benchmark datasets with varying the backbone networks.
  - Empirically, we demonstrated that better feature alignment for OSDA leads to the performances.

Thank you

Contact: adkto8093@kaist.ac.kr