2024 하계 세미나

Prompt Learning in Domain Adaptation

Sogang University
Vision & Display Systems Lab, Dept. of Electronic Engineering

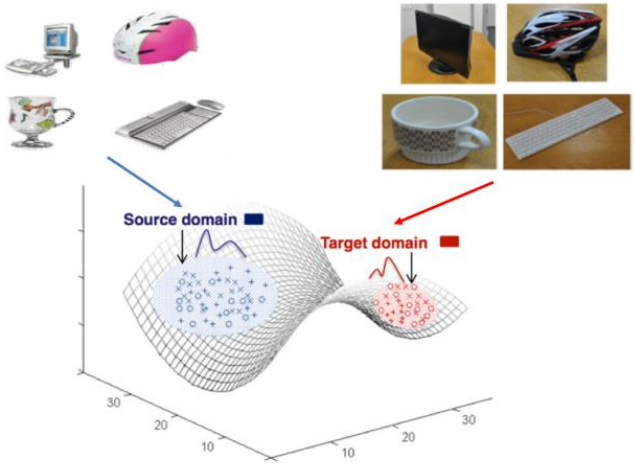
Presented By
박지원

Outline

- Background
  - Domain adaptation
  - Prompt Learning in Vision-Language
- Learning Domain-Aware Detection Head with Prompt
  - NeurIPS 2023
- Source-Free Domain Adaptation with Frozen Multimodal Foundation Model
  - CVPR 2024

Background

- Domain Adaptation
  - Concept
    - 특정 domain에서 학습된 모델을 다른 domain 으로 adapt 하려는 것
      - Source domain data: 모델이 학습하는 데이터
      - Target domain data: source data 로 학습한 모델이 적응하고자 하는 데이터
    - Domain 간의 domain gap 을 극복하고 source domain 에서 학습된 모델을 target domain 에 효과적으로 적응하기 위한 방법론 연구
      - Domain gap: source domain 과 target domain 의 분포 상의 차이



Background

- Prompt Learning in Vision-Language Model
  - Prompt learning (tuning) 이란
    - Pretrained vision-language model (VLM) 을 특정 downstream task에 맞게 조정하는 기법
    - 모델의 전체 파라미터를 튜닝하는 대신 입력 텍스트에 learnable textual prompt 를 붙여 학습
  - Context Optimization (CoOp) [1)]
    - Hand-crafted prompt 를 learnable continuous tokens 로 대체하여 prompt engineering 을 자동화

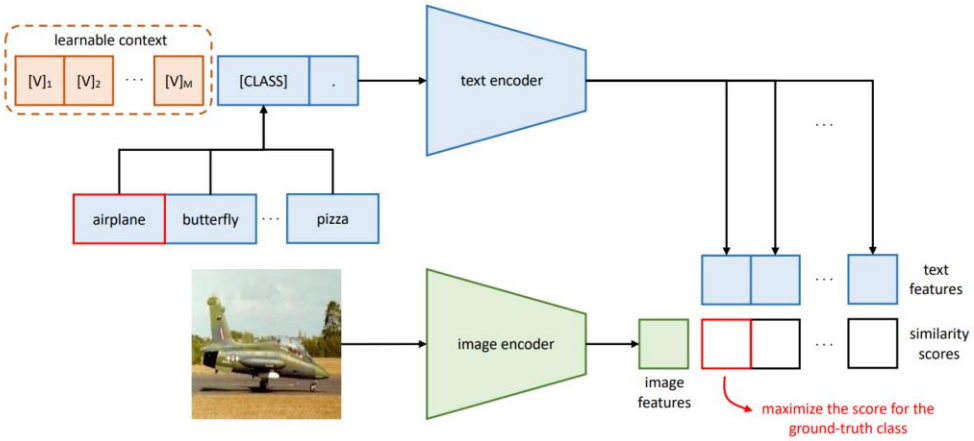
Caltech101	Prompt	Accuracy
	a [CLASS].	82.68
	a photo of [CLASS].	80.81
	a photo of a [CLASS].	86.29
	$[V]_1$ $[V]_2$ ... $[V]_M$ [CLASS].	**91.83**

Prompt engineering vs. context optimization (CoOP)

Learnable prompt t_i: $\mathbf{t}_i = [\mathbf{v}_1][\mathbf{v}_2]...[\mathbf{v}_M][\mathbf{c}_i]$

M learnable tokens $\mathbb{V} = [\mathbf{v}_1][\mathbf{v}_2]...[\mathbf{v}_M]$

$\mathbf{c}_i = e(\text{"class-i"})$



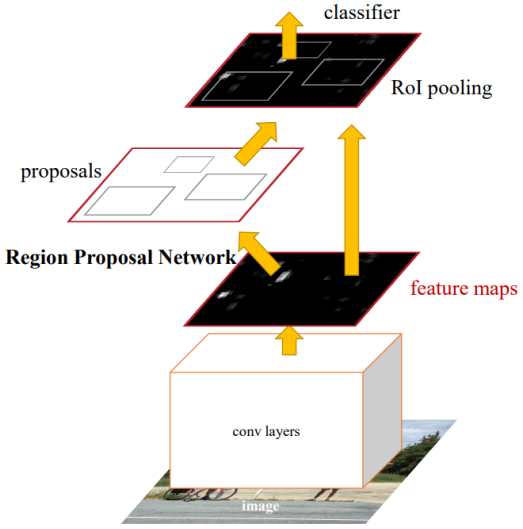
Overview of context optimization (CoOP)

Li, Zhang, et al. “Learning Domain-Aware Detection Head with Prompt.”

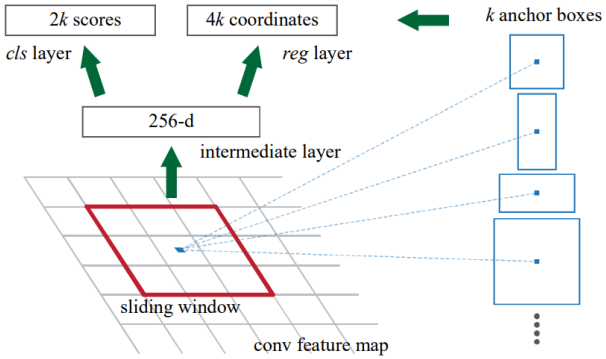
37th Conference on Neural Information Processing Systems (NeurIPS), 2023.

Background

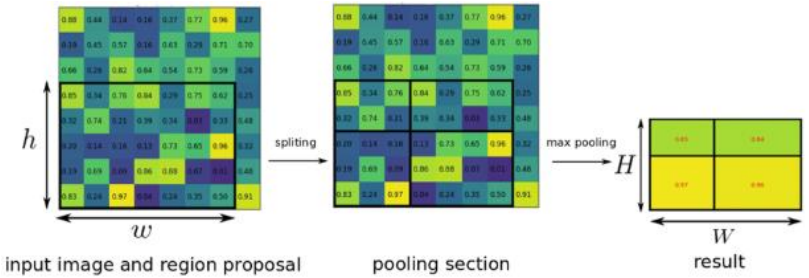
- Object detection
  - 개념
    - 이미지/비디오에서 object의 식별하고 분류하는 컴퓨터 비전 기술
  - Faster R-CNN
    - Region Proposal Network (RPN)
      - Background / foreground 구분, bounding box 결정
    - RoI(Region of Interest) pooling
      - 다양한 크기의 region proposals 로부터 고정된 크기의 feature map 얻음



Faster R-CNN 구조



RPN



RoI pooling

Introduction

- Motivation
  - 기존 domain adaptive object detection 방법론의 한계점
    - Backbone 의 domain bias 를 줄이는 것에 집중, detection head 의 domain bias 는 해결하지 못함
  - Vision-language model (VLM)
    - 다양한 도메인에 대한 downstream task에서 generalization 성능이 뛰어남
    - Domain-related prompt 사용으로 domain-aware detection head 생성 가능

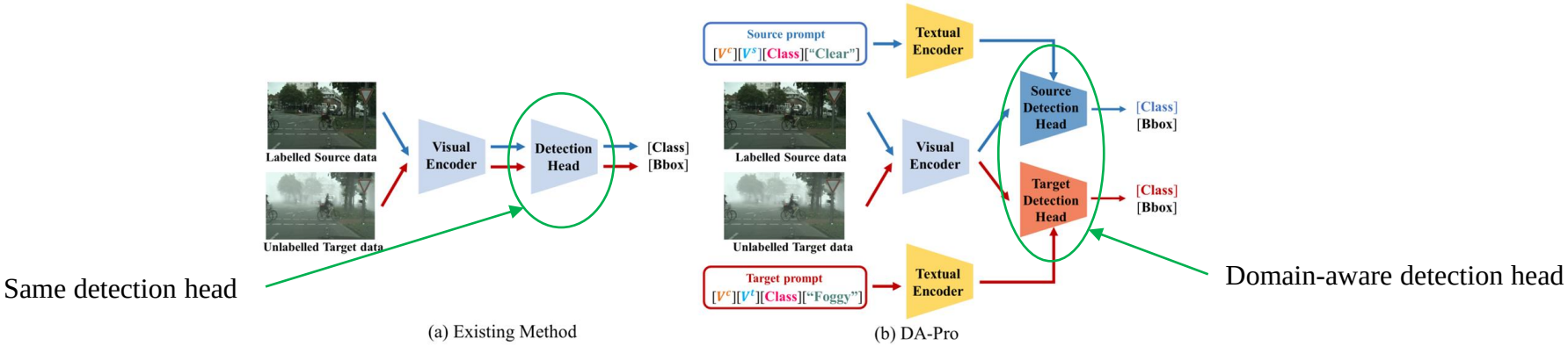


Figure 1: (a) Existing methods focus on reducing the domain bias of the detection backbone by inferring a discriminative visual encoder across domains, ignoring the domain bias in the detection head. (b) The proposed DA-Pro consists of a VLM-based backbone and a domain-aware detection head obtained by learning domain-adaptive prompt.

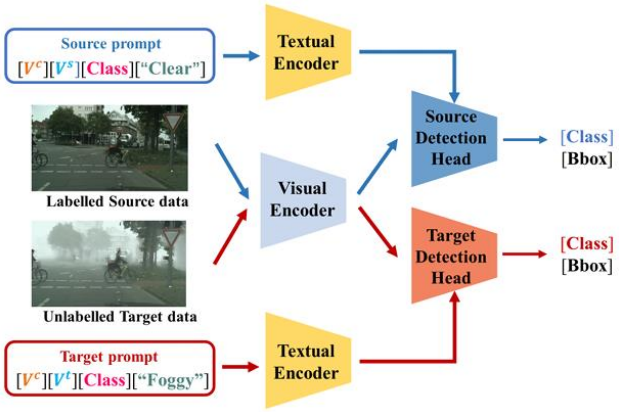
Method

- Domain-Adaptive Prompt
  - Dynamic textual class embedding by feeding different prompts
    - Source-related prompt → source-related detection head
    - Target-related prompt → target-related detection head

$c_i = e(\text{"class-i"})$

$dl_d = e(\text{"domain-d"})$

$e(\cdot)$: text tokenizer which maps the text description into vectors



Overview

Domain-invariant tokens (*M* learnable tokens)

Tokenized vector of the *i*-th class

$$\mathbf{t}_i^d = [\mathbf{v}_1^c][\mathbf{v}_2^c]\dots[\mathbf{v}_M^c][\mathbf{v}_1^d][\mathbf{v}_2^d]\dots[\mathbf{v}_N^d][\mathbf{c}_i][\mathbf{dl}_d]$$

Domain-specific tokens (*N* learnable tokens)

Domain-related prompts

EXAMPLE) Cityscapes(source) → FoggyCityscapes(target)

$t_i^s = [v_1^c][v_2^c]\dots[v_M^c]\ [v_1^s][v_2^s]\dots[v_N^s]\ [e(\text{"car"})]\ [e(\text{"clear"})]$

$t_i^t = [v_1^c][v_2^c]\dots[v_M^c]\ [v_1^t][v_2^t]\dots[v_N^t]\ [e(\text{"car"})]\ [e(\text{"foggy"})]$

Method

- Domain-Adaptive Prompt for object detection
  - Domain-invariant prompt
    - Detection head가 두 domain 의 입력 이미지에 대해 모두 classification 성능이 높도록 학습
  - Domain-specific prompt
    - Detection head 가 각 domain에 대해 높은 domain confidence 를 갖도록 학습
  - Domain adaptive detection head의 predict probability
    - image region boxes $R = \{r_j\}_{j=1}^{N_r}$ r(·) : class-agnostic RPN
    - Visual encoder $F = \{f_j\}_{j=1}^{N_r}$

visual embedding *textual embedding*

$$p(\hat{y} = y|\mathbf{r}_j, d, D) = \frac{exp(s(f(\mathbf{r}_j), g(\mathbf{t}_y^d))/\tau)}{\sum_{k \in D} \sum_{i=1}^{K} exp(s(f(\mathbf{r}_j), g(\mathbf{t}_i^k))/\tau)},$$

g(·) : text encoder
s(·) : cosine similarity
Class y
Domain d

Method

- Domain-Adaptive Prompt for object detection
  - Source domain cross entropy loss
    - $D \in \{\{s\},\{t\},\{s,t\}\}$
    - Source prompt $\{t_i^s\}_{i=1}^K$, Target prompt $\{t_i^t\}_{i=1}^K$
    - Both prompts $\{t_i^s\}_{i=1}^K \cup \{t_i^t\}_{i=1}^K$

$$\mathcal{L}_{s,D} = \mathbb{E}_{\mathcal{X}^s}[-\frac{1}{N_r}\sum_{j=1}^{N_r}\log p(\hat{y} = y_j^s|\mathbf{r}_j^s, s, D)]$$

$$\mathcal{L}_d^{inv} = \mathcal{L}_{s,\{s\}} + \mathcal{L}_{s,\{t\}} \qquad \mathcal{L}_d^{spc} = \mathcal{L}_{s,\{s,t\}}$$

Source objective function: $\mathcal{L}_s = \mathcal{L}_s^{inv} + \mathcal{L}_s^{spc}.$

Method

- Domain-Adaptive Prompt for object detection
  - Target domain cross entropy loss
    - Label이 없는 target 이미지에 대해 pseudo label 생성 $y_j^t = \arg\max_y p(\hat{y} = y|\mathbf{r}_j^t)$
    - $D \in \{\{t\}, \{s,t\}\}$
    - Target prompt $\{t_i^t\}_{i=1}^K$, Both prompts $\{t_i^s\}_{i=1}^K \cup \{t_i^t\}_{i=1}^K$

$$\mathcal{L}_{t,\mathcal{D}} = \mathbb{E}_{\mathcal{X}^t}[-\frac{1}{N_r}\sum_{j=1}^{N_r}\mathbf{I}(p(\hat{y} = y_j^t|\mathbf{r}_j^t, t, \{t\}) \geq \tau)\log p(\hat{y} = y_j^t|\mathbf{r}_j^t, t, \mathcal{D})]$$

$$\mathcal{L}_{ent} = \mathbb{E}_{\mathcal{X}^t}[-\frac{1}{N_r}\sum_{j=1}^{N_r}\mathbf{I}(p(\hat{y} = y_j^t|\mathbf{r}_j^t, t, \{t\}) \geq \tau)p(\hat{y} = y_j^t|\mathbf{r}_j^t, t, \{t\})\log p(\hat{y} = y_j^t|\mathbf{r}_j^t, t, \{t\})]$$

Target objective function: $\mathcal{L}_t = \mathcal{L}_{t,\{t\}} + \mathcal{L}_{t,\{s,t\}} + \mathcal{L}_{ent}$

Final objective function: $\mathcal{L} = \mathcal{L}_s + \lambda\mathcal{L}_t.$

Experiments

- Adaptation tasks
  - Cross-Weather, Cross-Fov, Sim-to-Real comparison results

Table 1: Comparison with existing methods on three adaptation tasks, for Cross-Weather adaptation Cityscapes→Foggy Cityscapes (C→F), Cross-Fov adaptation KITTI→Cityscapes (K→C) and Sim-to-Real adaptation SIM10K→Cityscapes (S→C). mAP: mean Average Precision (%).

	C→F									K→C	S→ C
Methods	Person	Rider	Car	Truck	Bus	Train	Motor	Bicycle	mAP	mAP	mAP
DA-Faster [3]	29.2	40.4	43.4	19.7	38.3	28.5	23.7	32.7	32.0	41.9	38.2
VDD [38]	33.4	44.0	51.7	33.9	52.0	34.7	34.2	36.8	40.0	-	-
DSS [37]	42.9	51.2	53.6	33.6	49.2	18.9	36.2	41.8	40.9	42.7	44.5
MeGA [36]	37.7	49.0	52.4	25.4	49.2	46.9	34.5	39.0	41.8	43.0	44.8
SCAN [21]	41.7	43.9	57.3	28.7	48.6	48.7	31.0	37.3	42.1	45.8	52.6
TIA [44]	52.1	38.1	49.7	37.7	34.8	46.3	48.6	31.1	42.3	44.0	-
SIGMA [22]	44.0	43.9	60.3	31.6	50.4	51.5	31.7	40.6	44.2	45.8	53.7
AT [24]	**56.3**	51.9	64.2	38.5	45.5	55.1	**54.3**	35.0	50.9	-	-
Baseline	51.8	59.0	67.4	36.8	59.5	50.6	39.7	55.9	52.6	59.5	60.8
DA-Pro	55.4	**62.9**	**70.9**	**40.3**	**63.4**	**54.0**	42.3	**58.0**	**55.9**	**61.4**	**62.9**

Experiments

- Ablation studies
  - Comparison on prompt design
    - CoOp-style predefined prompt
    - Only domain invariant tokens , only domain specific tokens
    - Prompt length (number of M, N)

Table 2: Ablation studies (%) on Cross-Weather adaptation scenario Cityscapes→Foggy Cityscapes. AP50 evaluates mAP on detection boxes with IoU ≥ 0.5, and ≥ 0.75 for AP75. AP averages AP50 to AP95 with step 5.

CoOp-style learnable prompt →

Prompt Design	M	N	Prompt Ensemble	AP	AP50	AP75
A photo of a [class][domain]				28.5	52.6	28.7
$[\mathbf{v}_1^c][\mathbf{v}_2^c]...[\mathbf{v}_M^c][\mathbf{c}_i]$	16	0		28.9	53.0	28.5
$[\mathbf{v}_1^c][\mathbf{v}_2^c]...[\mathbf{v}_M^c][\mathbf{c}_i][\mathbf{dl}_d]$	16	0		29.2	53.8	29.3
$[\mathbf{v}_1^d][\mathbf{v}_2^d]...[\mathbf{v}_N^d][\mathbf{c}_i][\mathbf{dl}_d]$	0	16		28.9	53.1	28.7
$[\mathbf{v}_1^c][\mathbf{v}_2^c]...[\mathbf{v}_M^c][\mathbf{v}_1^d][\mathbf{v}_2^d]...[\mathbf{v}_N^d][\mathbf{c}_i][\mathbf{dl}_d]$	8	8		31.2	55.5	30.5
$[\mathbf{v}_1^c][\mathbf{v}_2^c]...[\mathbf{v}_M^c][\mathbf{v}_1^d][\mathbf{v}_2^d]...[\mathbf{v}_N^d][\mathbf{c}_i][\mathbf{dl}_d]$	8	8	✓	**31.9**	**55.9**	**32.0**
$[\mathbf{v}_1^c][\mathbf{v}_2^c]...[\mathbf{v}_M^c][\mathbf{v}_1^d][\mathbf{v}_2^d]...[\mathbf{v}_N^d][\mathbf{c}_i][\mathbf{dl}_d]$	4	4	✓	31.1	55.0	30.2
$[\mathbf{v}_1^c][\mathbf{v}_2^c]...[\mathbf{v}_M^c][\mathbf{v}_1^d][\mathbf{v}_2^d]...[\mathbf{v}_N^d][\mathbf{c}_i][\mathbf{dl}_d]$	12	12	✓	31.4	55.4	31.3
$[\mathbf{v}_1^c][\mathbf{v}_2^c]...[\mathbf{v}_M^c][\mathbf{v}_1^d][\mathbf{v}_2^d]...[\mathbf{v}_N^d][\mathbf{c}_i][\mathbf{dl}_d]$	16	16	✓	31.3	55.3	30.6

Experiments

- Ablation studies
  - Comparison on loss function

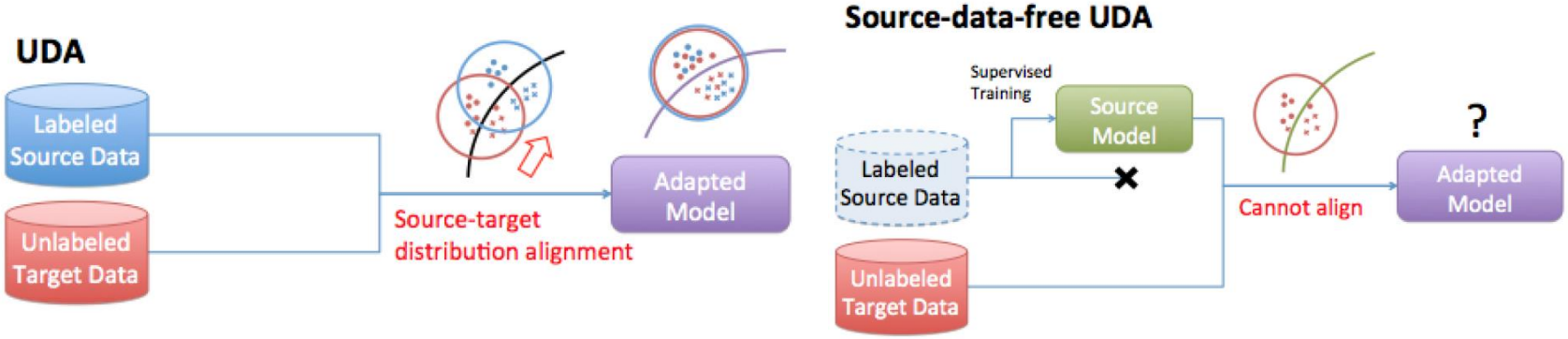
Table 3: The influence (%) of loss design on Cross-Weather adaptation scenario Cityscapes→Foggy Cityscapes. Total number of tokens is set to 16. - stands for the prompt is not compatible with the loss. Without the historical prompt ensemble strategy.

Prompt Design	$\mathcal{L}_{s,\{s\}} + \mathcal{L}_{t,\{t\}}$	$\mathcal{L}_{s,\{t\}}$	$\mathcal{L}_{s,\{s,t\}} + \mathcal{L}_{t,\{s,t\}}$	$\mathcal{L}_{ent}$	$\mathcal{L}_{t,\{s\}}$	mAP
A photo of a [class][domain]	-	-	-	-	-	52.6
$[\mathbf{v}_1^c][\mathbf{v}_2^c]...[\mathbf{v}_M^c][\mathbf{c}_i]$	✓	-	-	-	-	53.0
$[\mathbf{v}_1^c][\mathbf{v}_2^c]...[\mathbf{v}_M^c][\mathbf{c}_i][\mathbf{dl}_d]$	✓					53.5
$[\mathbf{v}_1^c][\mathbf{v}_2^c]...[\mathbf{v}_M^c][\mathbf{c}_i][\mathbf{dl}_d]$	✓	✓				53.8
$[\mathbf{v}_1^c][\mathbf{v}_2^c]...[\mathbf{v}_M^c][\mathbf{c}_i][\mathbf{dl}_d]$	✓	✓	✓			53.8
$[\mathbf{v}_1^c][\mathbf{v}_2^c]...[\mathbf{v}_M^c][\mathbf{c}_i][\mathbf{dl}_d]$	✓	✓	✓	✓		53.7
$[\mathbf{v}_1^d][\mathbf{v}_2^d]...[\mathbf{v}_N^d][\mathbf{c}_i][\mathbf{dl}_d]$	✓					52.9
$[\mathbf{v}_1^d][\mathbf{v}_2^d]...[\mathbf{v}_N^d][\mathbf{c}_i][\mathbf{dl}_d]$	✓	✓				53.1
$[\mathbf{v}_1^d][\mathbf{v}_2^d]...[\mathbf{v}_N^d][\mathbf{c}_i][\mathbf{dl}_d]$	✓	✓	✓			52.7
$[\mathbf{v}_1^d][\mathbf{v}_2^d]...[\mathbf{v}_N^d][\mathbf{c}_i][\mathbf{dl}_d]$	✓	✓	✓	✓		52.9
$[\mathbf{v}_1^c][\mathbf{v}_2^c]...[\mathbf{v}_M^c][\mathbf{v}_1^d][\mathbf{v}_2^d]...[\mathbf{v}_N^d][\mathbf{c}_i][\mathbf{dl}_d]$	✓					54.4
$[\mathbf{v}_1^c][\mathbf{v}_2^c]...[\mathbf{v}_M^c][\mathbf{v}_1^d][\mathbf{v}_2^d]...[\mathbf{v}_N^d][\mathbf{c}_i][\mathbf{dl}_d]$	✓	✓				54.8
$[\mathbf{v}_1^c][\mathbf{v}_2^c]...[\mathbf{v}_M^c][\mathbf{v}_1^d][\mathbf{v}_2^d]...[\mathbf{v}_N^d][\mathbf{c}_i][\mathbf{dl}_d]$	✓	✓	✓			55.2
$[\mathbf{v}_1^c][\mathbf{v}_2^c]...[\mathbf{v}_M^c][\mathbf{v}_1^d][\mathbf{v}_2^d]...[\mathbf{v}_N^d][\mathbf{c}_i][\mathbf{dl}_d]$	✓	✓	✓	✓		55.5
$[\mathbf{v}_1^c][\mathbf{v}_2^c]...[\mathbf{v}_M^c][\mathbf{v}_1^d][\mathbf{v}_2^d]...[\mathbf{v}_N^d][\mathbf{c}_i][\mathbf{dl}_d]$	✓	✓	✓	✓	✓	53.4

Tang, Su, et al. “Source-Free Domain Adaptation with Frozen Multimodal Foundation model” The IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR), 2024.

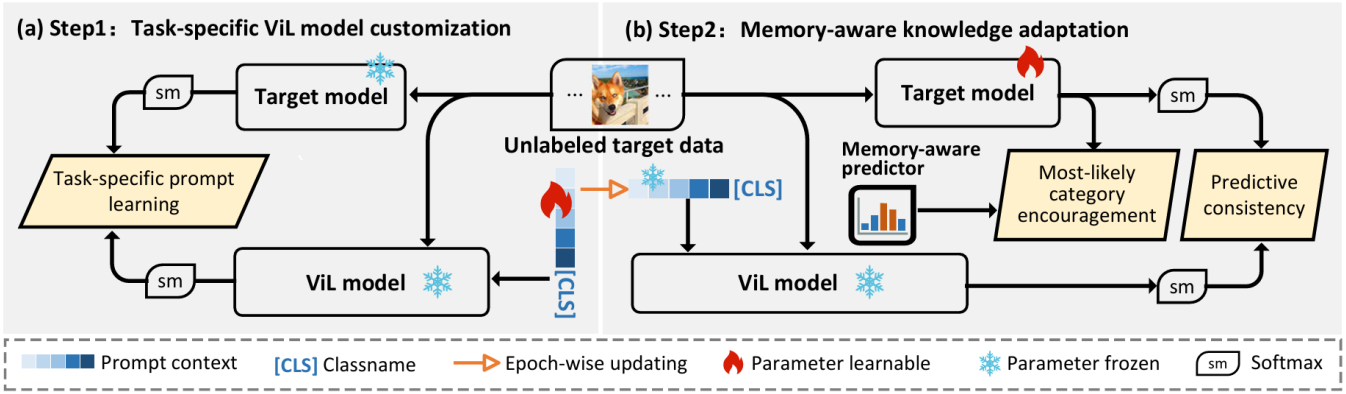
Background

- Source-free Unsupervised Domain Adaptation
  - Unsupervised Domain Adaptation (UDA)
    - 타겟 도메인의 데이터가 라벨 없이도 task 를 구행할 수 있도록 학습
  - Source-free UDA
    - Source model 과 라벨이 없는 target data 를 통해 target domain 에 adapting 하는 방법론



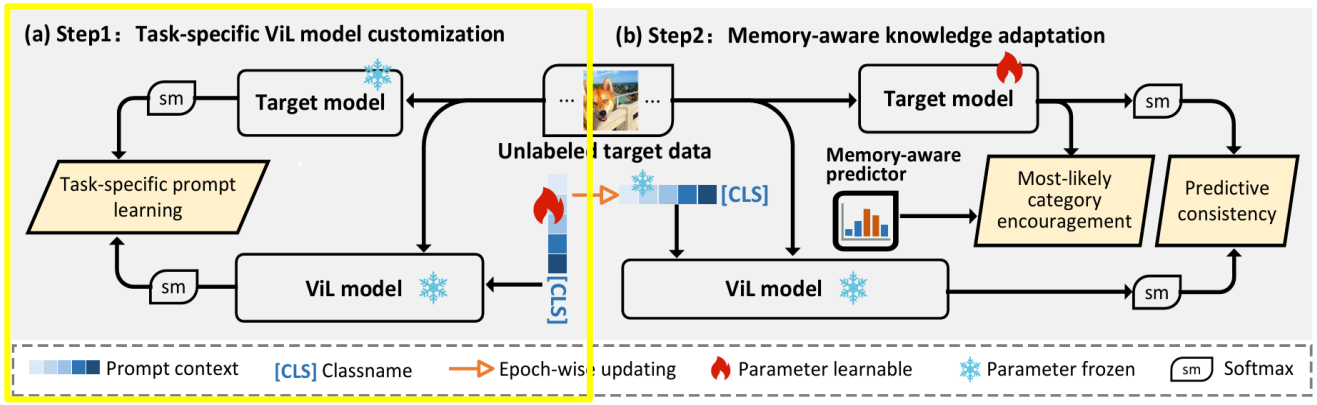
Introduction

- Motivation
  - 기존의 UDA methods과 한계점
    - Pseudo source domain 구축하여 기존의 UDA 방법론(ex. contrastive learning) 활용
    - Source model 또는 target 데이터로부터 extra supervision 구축
    - 하지만 domain distribution의 차이로 인해 pseudo-labeling 과 auxiliary supervision에서 error 발생
  - CLIP과 같은 multi-modal 기반 모델 활용하는 방법 제안
    - Unsupervised prompt learning 통해 task-specific information 활용
    - Customized VIL 모델로부터 targe model knowledge distillation



Method

- Task-Specific ViL Model Customization
  - Prompt learning part
    - ViL model의 모든 파라미터는 frozen, 각 클래스에 부여된 prompt 만 learnable
    - Target model과 ViL model의 prediction으로 mutual information 계산하여 학습
      - KL divergence 보다 lower optimization bound를 가지므로 deeper alignment 가능

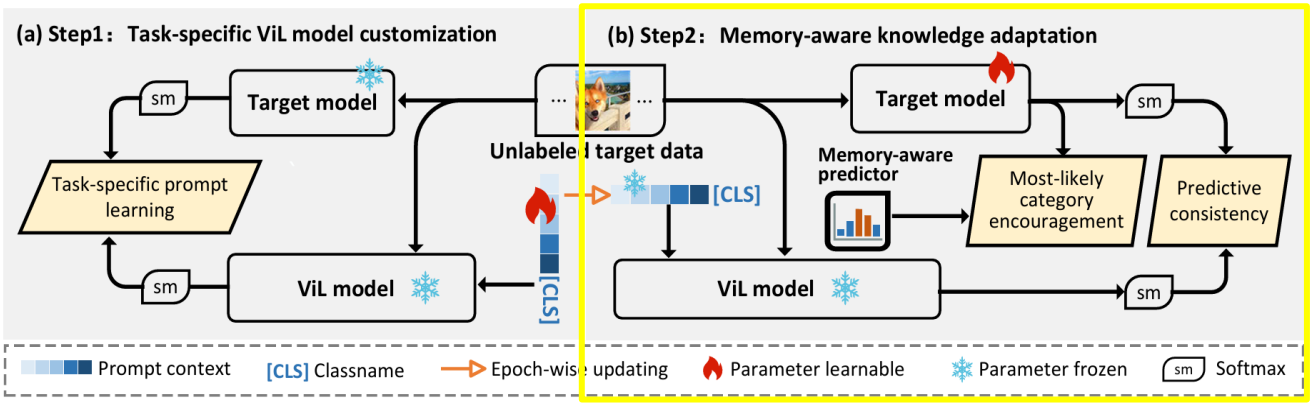


Mutual information loss $L_{\mathrm{TSC}} = -\min_{\boldsymbol{v}} \mathbb{E}_{\boldsymbol{x}_i \in \mathcal{X}_t} \mathrm{I}\left(\theta_t\left(\boldsymbol{x}_i\right), \theta_v\left(\boldsymbol{x}_i, \boldsymbol{v}\right)\right)$ v: prompt context

Target model ViL model

Method

- Memory-Aware Knowledge Adaptation
  - Most-likely category encouragement – memory aware predictor
    - Target model과 ViL model의 collective knowledge 를 모두 활용
    - Target model 의 prediction $\{p_i\}_{i=1}^{n}$ 과 ViL model 의 prediction $\{p'_i\}_{i=1}^{n}$ 을 prediction bank 에 저장
      - 이때 target model 의 prediction은 iteration 단위, ViL model의 prediction 은 epoch 단위 update
      - 이를 통해 customized ViL model 의 안정성을 유지하면서 동시에 task-specific dynamics 활용 가능



Historical prediction fusion process

$$\bar{p}_i = \omega\, p_i + (1-\omega)\, p'_i.$$

Method

- Memory-Aware Knowledge Adaptation
  - Most-likely category encouragement – category attention calibration
    - Top-N most probable categories 를 $M_i = \{m_k\}_{k=1}^{N}$ 로 지정
    - Target domain 샘플 x_i 에 대한 target model 의 logit 을 l_i 라고 할 때 regularization loss 정의

$$L_{\mathrm{MCE}} = \min_{\theta t} \mathbb{E}_{\boldsymbol{x}_i \in \mathcal{X}_t} \log \frac{\exp\left(a_i/\tau\right)}{\sum_{j \neq \mathcal{M}_i} \exp\left(b_i \cdot \boldsymbol{l}_{i,j}/\tau\right)} \qquad a_i = \prod_{k=1}^{N} \boldsymbol{l}_{i,m_k}, \quad b_i = \sum_{k=1}^{N} \boldsymbol{l}_{i,m_k}$$

$l_{i,a}$: a-th element of l_i

$$L_{\mathrm{PC}} = \min_{\theta_t} \left[-\mathbb{E}_{\boldsymbol{x}_i \in \mathcal{X}_t} \mathrm{I}\left(\theta_t\left(\boldsymbol{x}_i\right), \theta_v\left(\boldsymbol{x}_i, \boldsymbol{v}*\right)\right) + \alpha L_B\right]$$ Conventional consistency loss, $L_{\mathrm{B}} = \mathrm{KL}(\bar{\boldsymbol{q}} \,||\, \frac{\mathbf{1}}{C})$

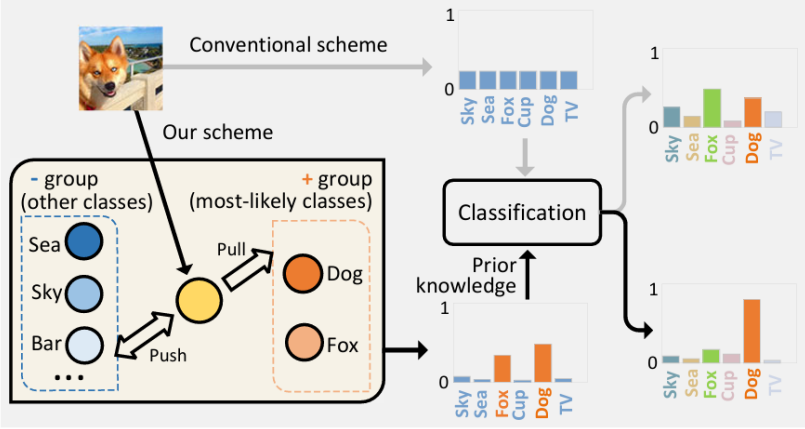


Figure 3. Illustration of most-likely category encouragement. In contrast to the conventional approach that assigns equal importance to all categories (depicted by the gray line), our approach (represented by the black line) introduces additional supervision by incorporating extra knowledge about the two most likely categories.

Experiments

- Comparison on Closed-set SFDA setting

Table 1. Closed-set SFDA on **Office-31** (%)

Method	Venue	A→D	A→W	D→A	D→W	W→A	W→D	Avg.
Source	–	79.1	76.6	59.9	95.5	61.4	98.8	78.6
SHOT [25]	ICML20	93.7	91.1	74.2	98.2	74.6	**100.**	88.6
NRC [49]	NIPS21	96.0	90.8	75.3	99.0	75.0	**100.**	89.4
GKD [38]	IROS21	94.6	91.6	75.1	98.7	75.1	**100.**	89.2
HCL [12]	NIPS21	94.7	92.5	75.9	98.2	77.7	**100.**	89.8
AaD [50]	NIPS22	96.4	92.1	75.0	**99.1**	76.5	**100.**	89.9
AdaCon [2]	CVPR22	87.7	83.1	73.7	91.3	77.6	72.8	81.0
CoWA [20]	ICML22	94.4	95.2	76.2	98.5	77.6	99.8	90.3
SCLM [40]	NN22	95.8	90.0	75.5	98.9	75.5	99.8	89.4
ELR [51]	ICLR23	93.8	93.3	76.2	98.0	76.9	**100.**	89.6
PLUE [26]	CVPR23	89.2	88.4	72.8	97.1	69.6	97.9	85.8
TPDS [41]	IJCV23	97.1	94.5	75.7	98.7	75.5	99.8	90.2
DIFO-C-RN	–	93.6	92.1	78.5	95.7	78.8	97.0	89.3
DIFO-C-B32	–	**97.2**	**95.5**	**83.0**	97.2	**83.2**	98.8	**92.5**

Table 2. Closed-set SFDA on **Office-Home** and **VisDA** (%). **SF** and **M** means source-free and multimodal, respectively; the full results on **VisDA** are in `Supplementary`.

Method	Venue	SF	M	Office-Home													VisDA
				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.	Sy→Re
Source	–	–	-	43.7	67.0	73.9	49.9	60.1	62.5	51.7	40.9	72.6	64.2	46.3	78.1	59.2	49.2
DAPL-RN [9]	TNNLS23	✗	✓	54.1	84.3	84.8	74.4	83.7	85.0	74.5	54.6	84.8	75.2	54.7	83.8	74.5	86.9
PADCLIP-RN [18]	ICCV23	✗	✓	57.5	84.0	83.8	77.8	85.5	84.7	76.3	59.2	85.4	78.1	60.2	86.7	76.6	88.5
ADCLIP-RN [36]	ICCVW23	✗	✓	55.4	85.2	85.6	76.1	85.8	86.2	76.7	56.1	85.4	76.8	56.1	85.5	75.9	87.7
SHOT [25]	ICML20	✓	✗	56.7	77.9	80.6	68.0	78.0	79.4	67.9	54.5	82.3	74.2	58.6	84.5	71.9	82.7
NRC [49]	NIPS21	✓	✗	57.7	80.3	82.0	68.1	79.8	78.6	65.3	56.4	83.0	71.0	58.6	85.6	72.2	85.9
GKD [38]	IROS21	✓	✗	56.5	78.2	81.8	68.7	78.9	79.1	67.6	54.8	82.6	74.4	58.5	84.8	72.2	83.0
AaD [50]	NIPS22	✓	✗	59.3	79.3	82.1	68.9	79.8	79.5	67.2	57.4	83.1	72.1	58.5	85.4	72.7	88.0
AdaCon [2]	CVPR22	✓	✗	47.2	75.1	75.5	60.7	73.3	73.2	60.2	45.2	76.6	65.6	48.3	79.1	65.0	86.8
CoWA [20]	ICML22	✓	✗	56.9	78.4	81.0	69.1	80.0	79.9	67.7	57.2	82.4	72.8	60.5	84.5	72.5	86.9
SCLM [40]	NN22	✓	✗	58.2	80.3	81.5	69.3	79.0	80.7	69.0	56.8	82.7	74.7	60.6	85.0	73.0	85.3
ELR [51]	ICLR23	✓	✗	58.4	78.7	81.5	69.2	79.5	79.3	66.3	58.0	82.6	73.4	59.8	85.1	72.6	85.8
PLUE [26]	CVPR23	✓	✗	49.1	73.5	78.2	62.9	73.5	74.5	62.2	48.3	78.6	68.6	51.8	81.5	66.9	88.3
TPDS [41]	IJCV23	✓	✗	59.3	80.3	82.1	70.6	79.4	80.9	69.8	56.8	82.1	74.5	61.2	85.3	73.5	87.6
DIFO-C-RN	–	✓	✓	62.6	87.5	87.1	79.5	87.9	87.4	78.3	63.4	88.1	80.0	63.3	87.7	79.4	88.8
DIFO-C-B32	–	✓	✓	**70.6**	**90.6**	**88.8**	**82.5**	**90.6**	**88.8**	**80.9**	**70.1**	**88.9**	**83.4**	**70.5**	**91.2**	**83.1**	**90.3**

Experiments

- Comparison to CLIP based prediction results
  - Original CLIP model 과의 성능 비교

Table 4. Results (%) of CLIP and Source+CLIP on the four evaluation datasets. The backbone of CLIP image-encoder in CLP-C-RN and CLP-C-B32 are the same as **DIFO**-C-RN and **DIFO**-C-B32, respectively. The full results are provided in `Supplementary`.

Method	Venue	Office-31				Office-Home					VisDA	DomainNet-126				
		→A	→D	→W	→Avg.	→Ar	→Cl	→Pr	→Rw	→Avg.	Sy→Re	→C	→P	→R	→S	→Avg.
CLIP-RN [31]	ICML21	73.1	73.9	67.0	71.4	72.5	51.9	81.5	82.5	72.1	83.7	67.9	70.2	87.1	65.4	72.7
Source+CLIP-RN	–	76.3	90.4	84.0	83.6	75.4	57.4	84.4	85.7	75.7	82.0	71.8	71.4	87.3	66.5	74.3
DIFO-C-RN	–	**78.6**	**95.3**	**93.9**	**89.3**	**79.3**	**63.1**	**87.7**	**87.5**	**79.4**	**88.8**	**74.5**	**74.2**	**88.5**	**69.7**	**76.7**
CLIP-B32 [31]	ICML21	76.0	82.7	80.6	79.8	74.6	59.8	84.3	85.5	76.1	82.9	74.7	73.5	85.7	71.2	76.3
Source+CLIP-B32	–	78.5	93.0	89.6	87.0	78.9	62.5	86.1	87.7	78.8	82.0	76.8	73.7	86.0	70.8	76.8
DIFO-C-B32	–	**83.1**	**98.0**	**96.4**	**92.5**	**82.3**	**70.4**	**90.8**	**88.8**	**83.1**	**90.3**	**80.4**	**76.9**	**87.3**	**75.3**	**80.0**

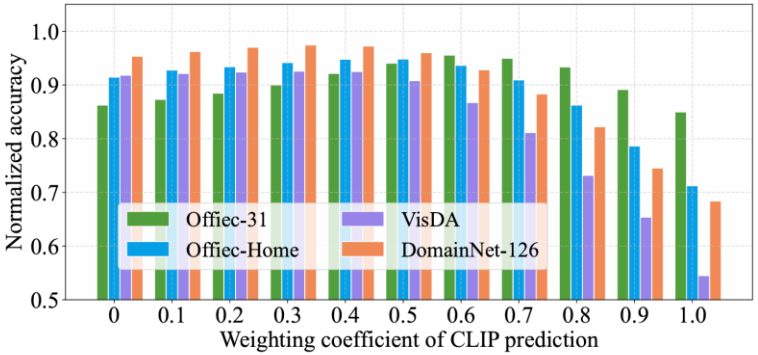


Figure 4. The performance of the scheme directly weighting the source model and CLIP-B32. All results are normalized by corresponding DIFO-C-B32 accuracies for a clear view.

Experiments

- Comparison on Partial-set and Open-set SFDA settings & ablation study

Table 5. Partial-set SFDA and Open-set SFDA on **Office-Home** (%). The full results are provided in `Supplementary`.

Partial-set SFDA	Venue	Avg.	Open-set SFDA	Venue	Avg.
Source	–	62.8	Source	–	46.6
SHOT [25]	ICML20	79.3	SHOT [25]	ICML20	72.8
HCL [12]	NIPS21	79.6	HCL [12]	NIPS21	72.6
CoWA [20]	ICML22	83.2	CoWA [20]	ICML22	73.2
AaD [50]	NIPS22	79.7	AaD [50]	NIPS22	71.8
CRS [52]	CVPR23	80.6	CRS [52]	CVPR23	73.2
DIFO-C-B32	–	**85.6**	**DIFO**-C-B32	–	**75.9**

Table 6. Classification results of ablation study (%) on **Office-31** **Office-Home** and **VisDA**.

L_{TSC}	L_{MCE}	L_{PC}	**Office-31**	**Office-Home**	**VisDA**	Avg.
✗	✗	✗	78.6	59.2	49.2	62.3
✓	✗	✗	82.4	77.4	84.4	81.4
✗	✓	✗	82.1	76.5	88.6	82.4
✓	✓	✗	87.0	80.0	88.3	85.1
✓	✓	✓	**92.5**	**83.1**	**90.3**	**88.6**
DIFO-C-B32 w/ KL			90.4	81.5	89.0	87.0
DIFO-C-B32 w/ CLIP			90.7	81.1	88.8	86.8
DIFO-C-B32 w/o $\boldsymbol{p}_i'$			89.8	73.5	87.0	83.4
DIFO-C-B32 w/o $\boldsymbol{p}_i$			88.9	82.2	88.9	86.7

Experiments

- Feature distribution visualization

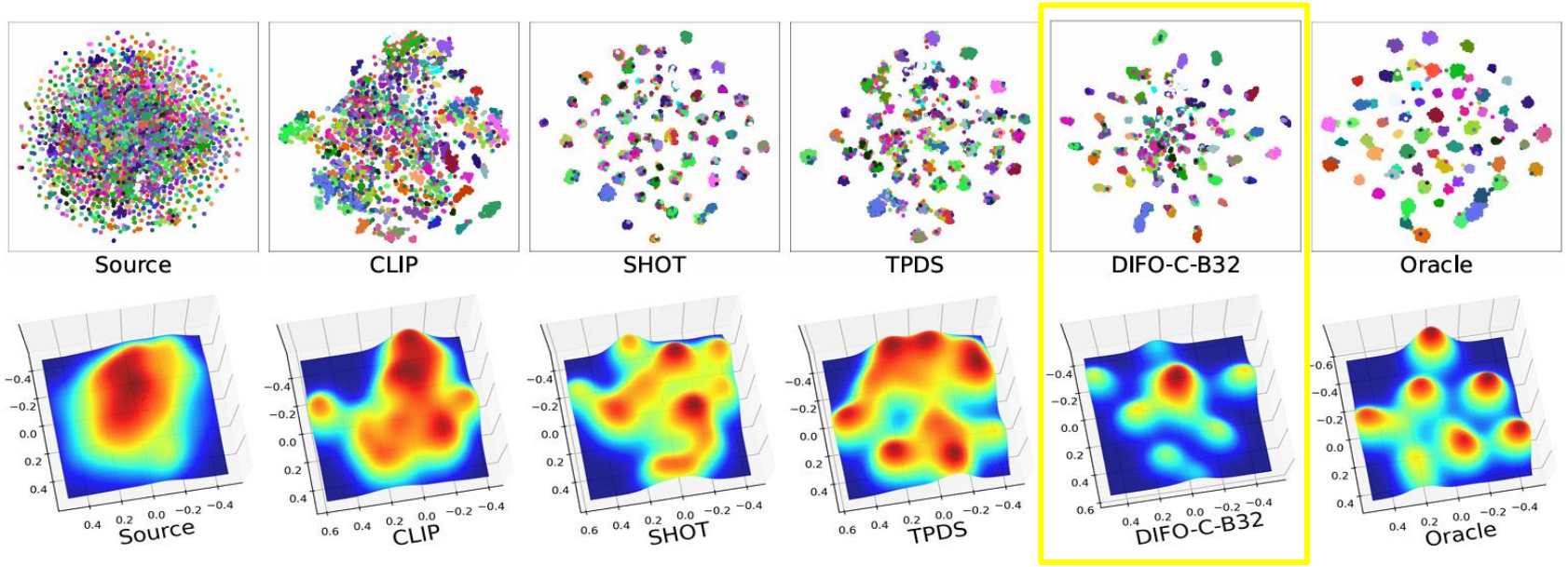


Figure 5. Feature distribution visualization comparison on transfer task Ar→Cl in Office-Home. Oracle is trained on target domain Cl using the ground-truth labels. Different colors stand for different categories. **Top**: t-SNE feature distribution over 65 categories. **Bottom**: The corresponding 3D density charts. For easy view, the first 10 categories were used in this plot.

감사합니다