電機資訊學院 2022 作 BRAIN PLUS HAND 實作專題競賽

EECS006

Rainbow UDA: Combining Domain Adaptive Models for Semantic Segmentation Tasks

Huang-Ru Liao (廖凰汝), Tzu-Wen Wang (王子文)

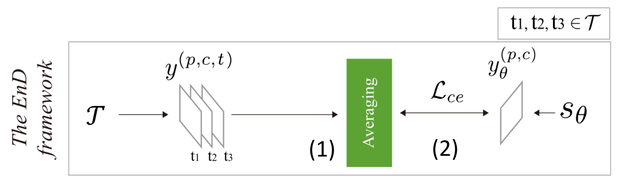
Abstract

In this work, we propose Rainbow UDA, a new framework designed to address the drawbacks of the previous ensemble-distillation frameworks when combining multiple unsupervised domain adaptation (UDA) models for semantic segmentation tasks (UDA-SST). Such drawbacks are mainly resulted from overlooking the magnitudes of the output certainties of different members in an ensemble as well as their individual performances in the target domain, causing the distillation process to suffer from certainty inconsistency and performance variation issues. These issues may hinder the effectiveness of an ensemble that includes members with either biased certainty distributions or have poor performance in the target domain. To mitigate such a deficiency, Rainbow UDA introduces two operations: the unification and the combinatorial fusion operations, to address the above two issues.

Motivation

The Previous Ensemble Distillation (EnD) framework

- $\mathcal{T}$ is a teacher ensemble, C is a given set of semantic classes, P is the set of pixels in an image.
- $y^{(p,c,t)}$: output certainty predictions from each $p \in P, c \in C, t \in \mathcal{T}$
- S_θ: a student model parameterized by θ.



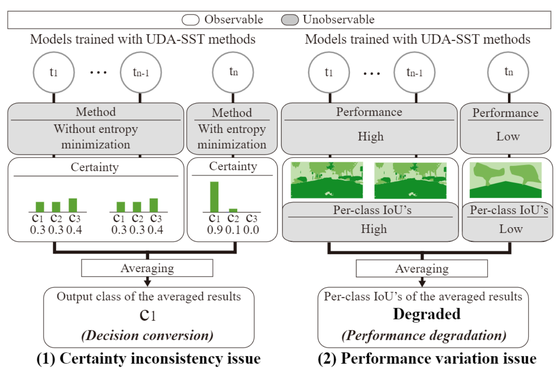
(1) Derive the averaged output certainties from $\mathcal{T}$ using the average operator

$$\mathbb{A}(y^{(p,c,t)}) = \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}} y^{(p,c,t)}.$$

(2) Minimize the cross-entropy loss between the averaged output certainties from $\mathcal{T}$ and S_θ

The Certainty Inconsistency and Performance Variation Issues

An illustration of the two issues of the average operation adopted in the previous ensemble distillation framework. For the first issue, directly averaging the output certainties of them together may cause the final prediction of the ensemble to be dominated by the teacher trained with the entropy minimization method, rather than the majority of all the models. For the second, assume that there is an under-performing model. The performance of the ensemble may degrade if the average operation is adopted.



Experiment Result

Quantitative Results

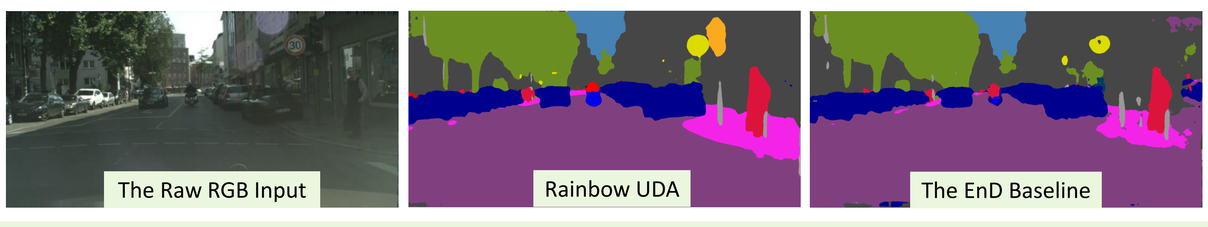
GTA5 → Cityscapes (G→C)

Method	Road	SideW	Build	Wall	Fence	Pole	Light	Sign	Veg	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	Motor	Bike	mIoU
CBST (Zou et al. 2018)	90.7	52.9	76.6	30.4	21.8	42.4	32.6	15.9	84.5	31.0	56.8	62.7	28.4	80.1	20.9	21.2	25.6	37.5	37.5	44.7
MRKLD (Zou et al. 2019)	89.1	44.7	75.5	30.7	27.4	44.5	42.0	28.7	84.3	40.0	77.3	66.7	36.4	79.0	34.2	30.6	3.9	48.0	42.8	48.7
CAG_UDA (Zhang et al. 2019)	90.4	51.6	83.8	34.2	27.8	38.4	25.3	48.4	85.4	38.2	78.1	58.6	34.6	84.7	21.9	42.7	**41.1**	29.3	37.2	50.2
R-MRNet (Zheng and Yang 2020a)	90.4	31.2	85.1	36.9	25.6	37.5	48.8	48.5	85.3	34.8	81.1	64.4	36.8	86.3	34.9	52.2	1.7	29.0	44.6	50.3
SAC (Araslanov and Roth 2021)	90.4	53.9	86.6	42.4	27.3	45.1	48.5	42.7	87.4	40.1	86.1	67.5	29.7	88.5	49.1	54.6	9.8	26.6	45.3	53.8
DACS (Tranheden et al. 2021)	94.1	58.6	87.4	30.7	39.1	37.9	44.2	51.4	87.2	43.7	87.6	66.3	38.3	89.0	55.8	50.4	0.0	35.4	51.3	55.2
ProDA (Zhang et al. 2021)	87.8	56.0	79.7	**46.3**	**44.8**	45.6	53.5	53.5	88.6	45.2	82.1	70.7	**39.2**	88.8	45.5	59.4	1.0	48.9	56.4	57.5
Source Only	57.40	21.43	56.80	8.93	22.14	32.38	34.62	24.90	78.98	15.92	63.71	55.55	13.83	58.11	21.99	29.78	2.36	28.41	33.98	34.80
EnD (Hinton, Vinyals, and Dean 2015)	91.97	46.18	85.44	26.61	30.96	42.99	46.86	33.44	88.01	44.60	86.11	68.87	38.03	88.80	36.97	45.68	0.00	49.05	54.03	52.87
Rainbow UDA	94.31	60.84	**88.42**	34.88	45.89	41.93	50.41	**53.79**	88.72	**49.59**	**89.43**	**71.34**	37.62	**90.39**	**53.63**	54.56	30.33	**50.62**	**58.16**	**60.26**

SYNTHIA → Cityscapes (S→C)

Method	Road	SideW	Build	Wall	Fence	Pole	Light	Sign	Veg	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	Motor	Bike	mIoU
CBST (Zou et al. 2018)	66.0	33.4	79.6	12.0	0.9	33.2	23.3	34.9	72.8	-	83.6	58.5	22.4	81.9	-	16.5	-	23.4	29.0	42.0
MRKLD (Zou et al. 2019)	75.8	39.8	80.5	6.8	0.6	37.0	23.6	33.0	74.6	-	84.9	60.2	24.1	80.2	-	12.6	-	15.2	30.8	42.5
CAG_UDA (Zhang et al. 2019)	82.6	42.4	83.8	5.5	0.1	33.6	15.9	23.4	83.8	-	82.7	69.5	32.2	80.9	-	13.8	-	37.3	46.7	45.9
DACS (Tranheden et al. 2021)	75.8	20.3	71.9	7.6	0.2	34.3	14.7	44.7	85.7	-	90.8	75.6	29.6	86.8	-	40.8	-	35.1	52.6	47.9
R-MRNet (Zheng and Yang 2020a)	85.9	45.6	84.6	12.8	1.0	34.6	29.6	20.9	81.9	-	83.4	68.2	24.6	88.5	-	35.5	-	30.1	52.3	48.7
SAC (Araslanov and Roth 2021)	84.7	45.8	**87.2**	24.1	1.3	41.9	45.7	33.6	87.9	-	**90.3**	69.2	29.0	**90.4**	-	41.6	-	42.5	50.7	54.1
ProDA (Zhang et al. 2021)	84.4	42.3	86.6	26.0	0.2	40.4	**54.9**	42.3	87.1	-	89.0	**78.5**	**30.2**	89.1	-	45.8	-	**47.3**	44.1	55.5
Source Only	25.40	15.55	59.70	18.07	0.66	26.35	19.36	30.22	72.50	-	74.28	48.11	13.67	74.62	-	36.94	-	13.92	36.45	35.36
EnD (Hinton, Vinyals, and Dean 2015)	89.76	45.97	81.28	1.61	0.08	25.55	8.92	27.02	84.08	-	84.38	53.46	0.84	81.95	-	26.62	-	12.33	51.05	42.18
Rainbow UDA	89.50	50.73	86.21	**32.65**	**1.65**	41.88	52.69	**39.80**	**88.69**	-	89.22	72.27	25.00	89.49	-	55.11	-	37.95	50.39	**56.45**

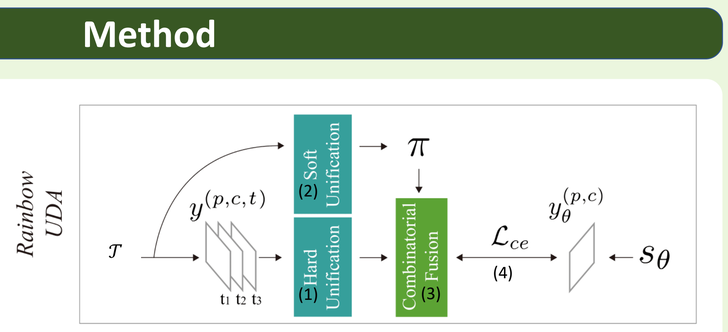
• The table reports the quantitative results evaluated on the GTA5 → Cityscapes (G→C) and SYNTHIA → Cityscapes (S→C) benchmarks. The models used in $\mathcal{T}$ are presented in the first seven rows for both G→C and S→C. The setting *Source Only* corresponds to the student model trained only with the source domain ground truth annotations.



Method

Framework overview

(1) Convert $y^{(p,c,t)}$ to one-hot pseudo labels using *hard unification (HU)*
(2) Perform *soft unification (SU)* operation on each $t \in \mathcal{T}$ to determine a fusion policy π.



(3) Fuse the one-hot pseudo labels based on π to generate the fused predictions.
(4) Train a student model s_θ using the images from the target domain dataset and the corresponding fused predictions to distill the knowledge from $\mathcal{T}$ to s_θ.

Unification Operation

- Soft Unification $\mathbb{S}(\cdot)$ is formulated as:

$$\mathbb{S}(y^{(p,c,t)}) = \hat{y}^{(p,c,t')}, t' \in \mathcal{T}',$$
$$\mathcal{T}' \leftarrow \underset{\forall t\in\mathcal{T}}{\text{minimize}}\, \mathcal{L}_{ce}(y^{(p,c,t)}, \mathbb{H}(y^{(p,c,t)}))$$

- Hard unification (HU) $\mathbb{H}(\cdot)$ is formulated as:

$$\mathbb{H}(y^{(p,c,t)}) = \begin{cases} 1, & \text{if } c = \arg\max_{c'\in\mathcal{C}} \{y^{(p,c',t)}\} \\ 0, & \text{otherwise} \end{cases}$$

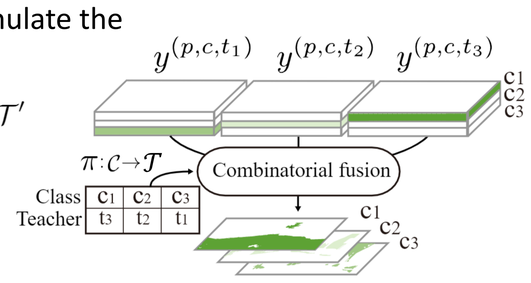
Combinatorial Fusion Operation

The combinational fusion operation consists of two components: the fusion policy and the resolving operator.

Fusion Policy π

- Fuses the predictions of different $t \in T$ from the perspectives of segmentation channels.
- The unified outputs $\mathbb{H}\left(y^{(p,c,t)}\right)$ are fused relying on the fusion policy π.
- In our experiment, per-class IoU on the training dataset has positive correlation with average per-class certainty: $\frac{1}{|\mathcal{A}^{(c,t')}|}\sum_{p\in\mathcal{A}^{(c,t')}} \hat{y}^{(p,c,t')}$
- $\mathcal{A}^{(c,\pi(c))} = \{p \,|\, p \in \mathcal{P}, \mathbb{H}(y^{(p,c,\pi(c))}) = y_\pi^{(p,c)} = 1\}$ represents the set of pixels that cover the segmentation map of a class $c \in C$ predicted by a certain $t = \pi(c) \in \mathcal{T}$.
- Based on the above observation, we formulate the fusion policy π as:

$$\pi(c) = \arg\max_{t'\in\mathcal{T}'} \frac{1}{|\mathcal{A}^{(c,t')}|}\sum_{p\in\mathcal{A}^{(c,t)}} \hat{y}^{(p,c,t')}, t' \in \mathcal{T}'$$

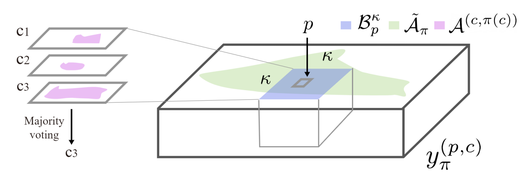


- The fused prediction is given by $y_\pi^{(p,c)} \triangleq \mathbb{H}\left(y^{(p,c,\pi(c))}\right)$

Resolving Operator

- Let the area where $A^{(c,\pi(c))}$ for different $c \in C$ overlap be denoted as $\tilde{A}_\pi$

$$\tilde{A}_\pi \triangleq \bigcup_{\substack{c_1\neq c_2,\\ c_1,c_2\in\mathcal{C}}} \left(\mathcal{A}^{(c_1,\pi(c_1))} \cap \mathcal{A}^{(c_2,\pi(c_2))}\right)$$



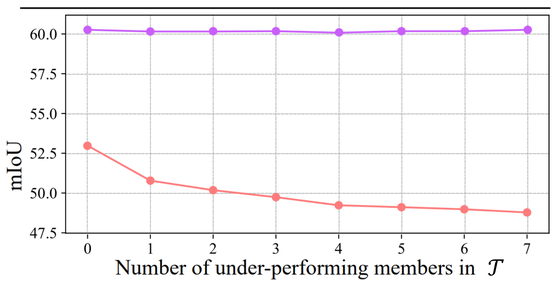
- Since $\{A^{(c,\pi(c))} | c \in C\}$ are not necessarily mutually exclusive, $\tilde{A}_\pi$ is not guaranteed to be in one-hot form. Thus we employ a spatially-aware resolving operator $\mathbb{R}(\cdot) : [0,1]^{|\mathcal{P}|\times|\mathcal{C}|}$ that assigns a class label for each pixel under $\tilde{A}_\pi$ using majority voting over a fix-sized kernel, defined as:

$$\mathbb{R}^\kappa(y_\pi^{(p,c)}) = \begin{cases} \mathbb{V}^\kappa(y_\pi^{(p,c)}), & \text{if } p \in \tilde{\mathcal{A}}^\pi \\ y_\pi^{(p,c)}, & \text{otherwise} \end{cases}$$

, where $\mathbb{V}^\kappa(\cdot)$ denotes the majority voting kernel of size, and κ is a hyper-parameter.

The Robustness to the Performance Variation Issue

Setting	+0	+1	+2	+3	+4	+5	+6	+7
EnD	52.87	50.78	50.18	49.74	49.23	49.11	48.98	48.78
Rainbow UDA	**60.26**	**60.07**	**60.09**	**60.11**	**60.03**	**60.16**	**60.12**	**60.18**



Impacts of the Composition of $\mathcal{T}$

(a) CBST (Zou et al. 2018)
(b) MRKLD (Zou et al. 2019)
(c) CAG_UDA (Zhang et al. 2019)
(d) R-MRNet (Zheng and Yang 2020a)
(e) SAC (Araslanov and Roth 2021)
(f) DACS (Tranheden et al. 2021)
(g) ProDA (Zhang et al. 2021)

Setting	Composition of $\mathcal{T}$	mIoU
S1	a+b	50.09
S2	a+b+c	53.27
S3	a+b+c+d	53.43
S4	a+b+c+d+e	56.24
S5	a+b+c+d+e+f	58.04
S6	a+b+c+d+e+f+g	60.26