Rethinking Ensemble-Distillation for Semantic Segmentation Based Unsupervised Domain Adaptation

Chen-Hao Chao, Bo-Wun Cheng, and Chun-Yi Lee

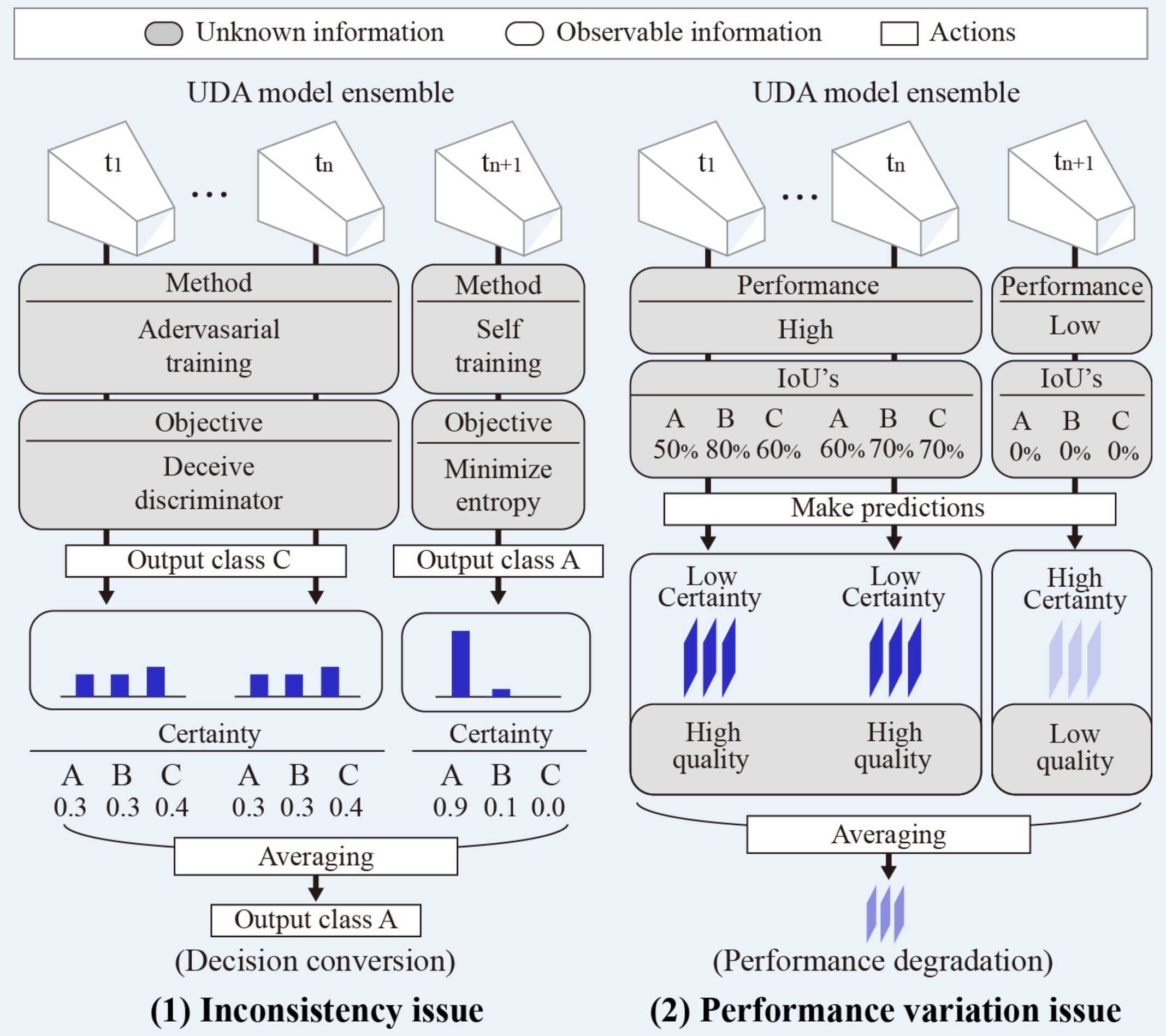
Elsa Lab, Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan

CVPR VIRTUAL JUNE 19-25

Abstract

Recent researches on unsupervised domain adaptation (UDA) have demonstrated that end-to-end ensemble learning frameworks serve as a compelling option for UDA tasks. Nevertheless, these end-to-end ensemble learning methods often lack **flexibility** as any modification to the ensemble requires retraining of their frameworks. To address this problem, we propose a flexible ensemble-distillation framework for performing semantic segmentation based UDA, allowing any arbitrary composition of the members in the ensemble while still maintaining its superior performance. To achieve such flexibility, our framework is designed to be **robust** against the output inconsistency and the performance variation of the members within the ensemble. To examine the effectiveness and the robustness of our method, we perform an extensive set of experiments on both GTA5 to Cityscapes and SYNTHIA to Cityscapes benchmarks to quantitatively inspect the improvements achievable by our method. We further provide detailed analyses to validate that our design choices are practical and beneficial. The experimental evidence validates that the proposed method indeed offer **superior performance, robustness and flexibility** in semantic segmentation based UDA tasks against contemporary baseline methods.

The Issues of Previous Methods



(1) Inconsistency Issue

Since each model in the ensemble can be trained independently using different methods, **the scale** of the output certainty values may **not be consistent** across the ensemble. This may result in a situation that few members' decisions with high certainty values dominate the entire ensemble's output.

(2) Performance Variation Issue

Since the **performance** (either per-class or average accuracy) of each teacher model in the ensemble may **vary substantially**, few under-performing members in the ensemble may cause the quality of the combined prediction to degrade.

Methodology

(1) A Comparison of the Proposed Method and Previous Methods

To address the **inconsistency** and **performance variation** issues, we introduce a new ensemble-distillation framework, and illustrate it in Fig.1(a). The main difference between the proposed method and the previous ones lies in two aspects: **output unification** and **fusion function**.

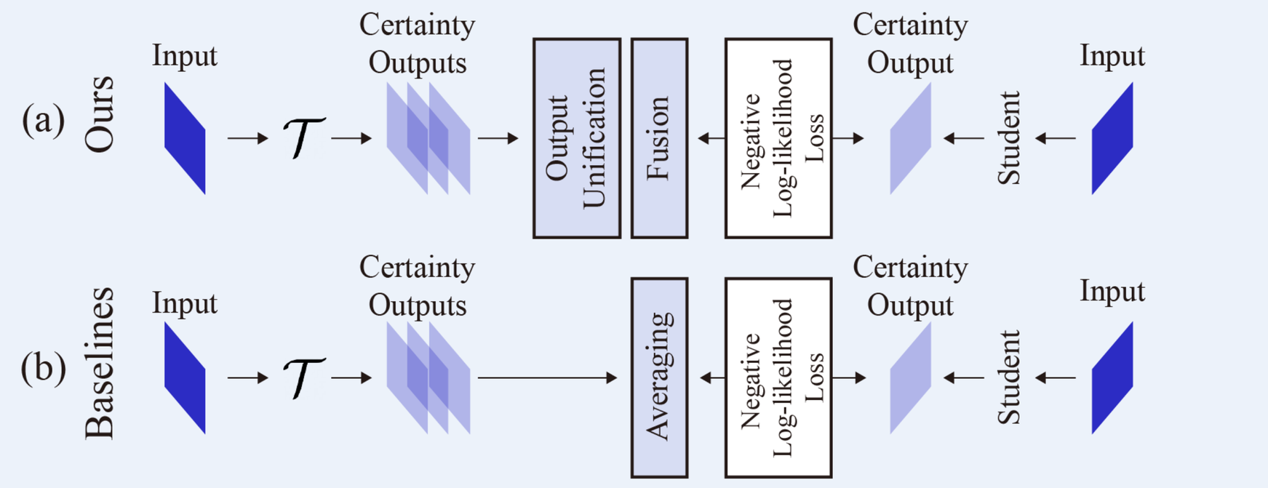


Figure 1: A comparison between (a) the proposed ensemble-distillation framework and (b) the baseline framework.

(2) Output Unification

The soft predictions $\hat{s}$ are unified first:

$$\hat{y}^{(p,c,t)} = \begin{cases} 1, & \text{if } c = \arg\max_{c \in \mathbf{C}} \{\hat{s}^{(p,c,t)}\} \\ 0, & \text{otherwise} \end{cases},$$

where $\mathcal{C}$ is a set of semantic classes. The unification operation converts $\hat{s}$ as **pseudo labels** to ensure that the subsequent fusion function can operate on items with a consistent scale.

(3) Fusion Function

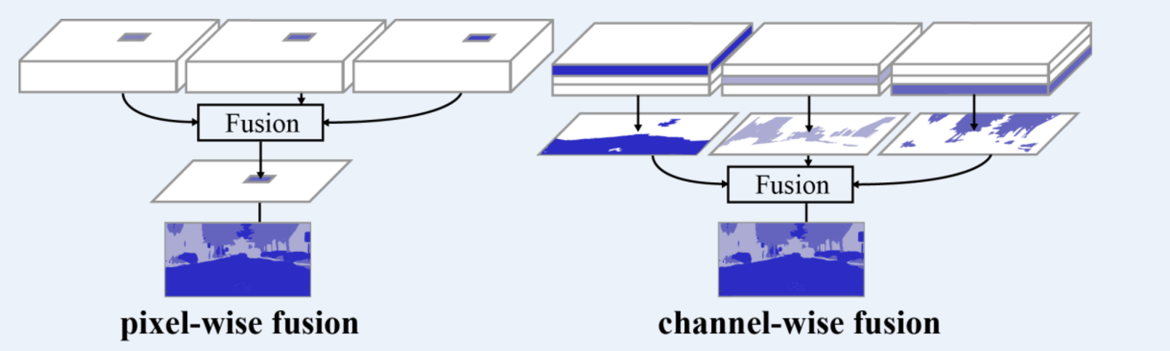


Figure 2: An illustration of the pixel-wise and channel-wise fusions.

The proposed **channel-wise fusion** rearranges the unified outputs according to a **fusion policy** π to construct the fused results. Fig.4. demonstrates the process of the rearrangement.

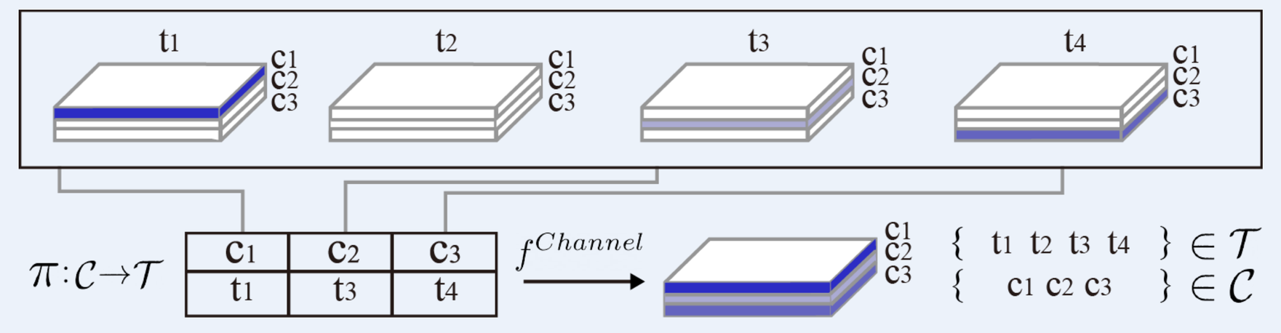


Figure 4: An illustrative example of π used in channel-wise fusion. Here, $\mathcal{C}$ denotes the set of semantic classes and $\mathcal{T}$ represents the set of teacher models in the ensemble.

(4) Policy Selection Strategy

Figure 5: An illustration of the quality of a teacher model's pseudo labels can affect the output certainty values of the student model.

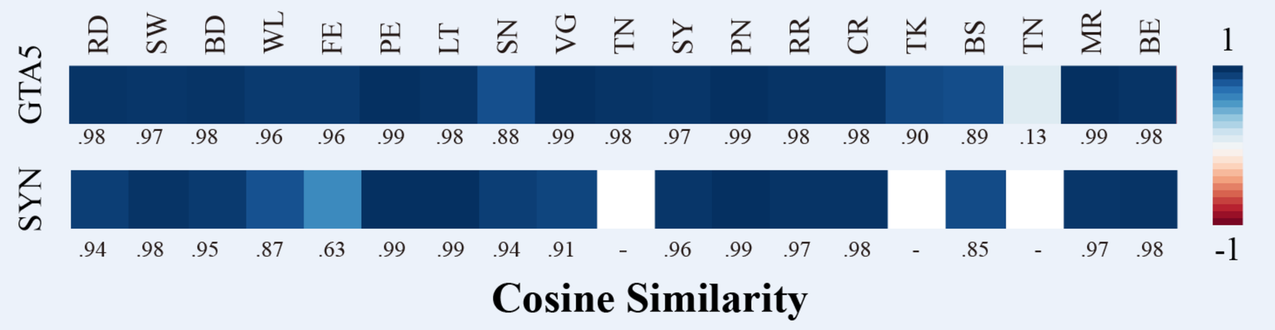


Figure 6: The cosine similarity between the per-class IoU of each $t \in \mathcal{T}$ and the per-class certainty values of the student evaluated on the training set of Cityscapes.

To find an appropriate π, we adopt a certainty-aware policy selection strategy. The strategy is designed based on the observation that the student's output **certainty** values is highly **correlated** to the **quality** of the fused pseudo labels as shown in Fig.5. and Fig.6. Therefore, our framework first performs knowledge distillation on $t \in \mathcal{T}$ and transfers their knowledge to $|\mathcal{T}|$ identical student models using the unified outputs. Then, their output certainty values are measured to obtain the approximated performance of their corresponding teacher models. Finally, the policy that maximizes the student model's output certainty values is selected.

Experimental Results

(1) Performance on the Benchmarks

GTA5 → Cityscapes																				
Method	Road	SideW	Build	Wall	Fence	Pole	Light	Sign	Veg	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	Motor	Bike	mIoU
APODA [20]	85.6	32.8	79.0	29.5	25.5	26.8	34.6	19.9	83.7	40.6	77.9	59.2	28.3	84.6	34.6	49.2	8.0	32.6	39.6	45.9
PatchAlign [21]	92.3	51.9	82.1	29.2	25.1	24.5	33.8	33.0	82.4	32.8	82.2	58.6	27.2	84.3	33.4	46.3	2.2	29.5	32.3	46.5
AdvEnt [22]	89.4	33.1	81.0	26.6	26.8	27.2	33.5	24.7	83.9	36.7	78.8	58.7	30.5	84.8	38.5	44.5	1.7	31.6	32.4	45.5
FDA-MBT [36]	92.5	53.3	82.4	26.5	27.6	36.4	40.6	38.9	82.3	39.8	78.0	62.6	34.4	84.9	34.1	53.1	16.9	27.7	46.4	50.5
PIT [35]	87.5	43.4	78.8	31.2	30.2	36.3	39.9	42.0	79.2	37.1	79.3	65.4	37.5	83.2	46.0	45.6	25.7	23.5	49.9	50.6
CBST [27]	91.8	53.5	80.5	32.7	21.0	34.0	28.9	20.4	83.9	34.2	80.9	53.1	24.0	82.7	30.3	35.9	16.0	25.9	42.8	45.9
MRKLD [28]	91.0	55.4	80.0	33.7	21.4	37.3	32.9	24.5	85.0	34.1	80.8	57.7	24.6	84.1	27.8	30.1	**26.9**	26.0	42.3	47.1
R-MRNet [29]	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
DACS [30]	89.90	39.66	87.87	30.71	39.52	38.52	46.43	52.79	87.98	43.96	**88.76**	67.20	35.78	84.45	45.73	50.19	0.00	27.25	33.96	52.14
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 [43]	92.17	53.12	84.85	24.77	29.76	40.38	40.98	49.35	86.21	42.85	79.74	62.79	35.98	85.72	42.10	44.45	0.26	28.27	51.80	51.34
EnD^2 [50]	92.39	53.84	85.34	24.51	30.53	40.28	42.40	50.28	86.19	43.39	80.55	63.26	36.75	86.15	43.95	43.91	0.20	30.17	53.22	51.96
Ours (Pixel)	92.29	57.34	84.09	36.75	29.17	41.37	48.96	42.26	86.91	39.95	82.81	66.29	37.42	86.94	35.21	48.82	1.48	40.78	53.02	53.26
Ours (Channel)	**94.43**	**60.90**	**88.07**	**39.46**	**41.80**	**43.24**	**49.08**	**56.00**	**88.01**	**45.83**	87.79	**67.58**	**38.05**	**90.08**	**57.64**	**51.90**	0.00	**46.57**	**55.28**	**57.98**

SYNTHIA → Cityscapes																					
Method	Road	SideW	Build	Wall*	Fence*	Pole*	Light	Sign	Veg	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	Motor	Bike	mIoU	mIoU*
APODA [20]	86.4	41.3	79.3	-	-	-	22.6	17.3	80.3	-	81.6	56.9	21.0	84.1	-	49.1	-	24.6	45.7	-	53.1
PatchAlign [21]	82.4	38.0	78.6	8.7	0.6	26.0	3.9	11.1	75.5	-	84.6	53.5	21.6	71.4	-	32.6	-	19.3	31.7	40.0	46.5
AdvEnt [22]	85.6	42.2	79.7	8.7	0.4	25.9	5.4	8.1	80.4	-	84.1	57.9	23.8	73.3	-	36.4	-	14.2	33.0	41.2	48.0
FDA-MBT [36]	79.3	35.0	73.2	-	-	-	19.9	24.0	61.7	-	82.6	61.4	31.1	83.9	-	40.8	-	**38.4**	51.1	-	52.5
PIT [35]	83.1	27.6	81.5	8.9	0.3	21.8	26.4	33.8	76.4	-	78.8	64.2	27.6	79.6	-	31.2	-	31.0	31.3	44.0	51.8
CBST [27]	68.0	29.9	76.3	10.8	1.4	33.9	22.8	29.5	77.6	-	78.3	60.6	28.3	81.6	-	23.5	-	18.8	39.8	42.6	48.9
MRKLD [28]	67.7	32.2	73.9	10.7	1.6	**37.4**	22.2	31.2	80.8	-	80.5	60.8	29.1	82.8	-	25.0	-	19.4	45.3	43.8	50.1
R-MRNet [29]	87.6	41.9	83.1	14.7	1.7	36.2	31.3	19.9	81.6	-	80.6	63.0	21.8	86.2	-	40.7	-	23.6	53.1	47.9	54.9
DACS [30]	80.56	25.12	81.90	21.46	2.85	37.20	22.67	23.99	**83.69**	-	**90.77**	**67.61**	**38.33**	82.92	-	38.90	-	28.49	47.58	48.34	54.81
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	40.06
EnD [43]	85.29	25.47	81.52	15.66	3.94	34.87	30.08	35.41	80.18	-	85.86	59.94	22.78	83.53	-	36.58	-	16.88	52.42	46.90	53.53
EnD^2 [50]	83.88	39.08	81.14	12.65	1.01	41.16	22.91	28.26	82.83	-	84.17	69.54	23.87	87.65	-	41.59	-	21.68	53.26	48.42	55.38
Ours (Pixel)	86.98	44.18	80.95	19.38	1.52	30.47	25.64	30.39	79.92	-	78.84	56.55	27.14	84.47	-	44.52	-	26.03	55.08	48.25	55.43
Ours (Channel)	**88.65**	**46.69**	**83.79**	**22.66**	**4.14**	35.01	**35.93**	**36.16**	82.80	-	81.35	61.61	32.13	**87.93**	-	**52.79**	-	31.95	**57.65**	**52.58**	**59.95**

Table 1: The quantitative results evaluated on the GTA5 to Cityscapes and SYNTHIA to Cityscapes benchmarks. The numbers presented in the middle and the last two columns correspond to per-class IoUs, mIoU, and mIoU*, respectively. The models used in our semantic segmentation based UDA model ensemble $\mathcal{T}$ are highlighted in blue. The setting *'Source Only'* indicates that the student model is trained only with the source domain ground truth annotations.

(2) Robustness

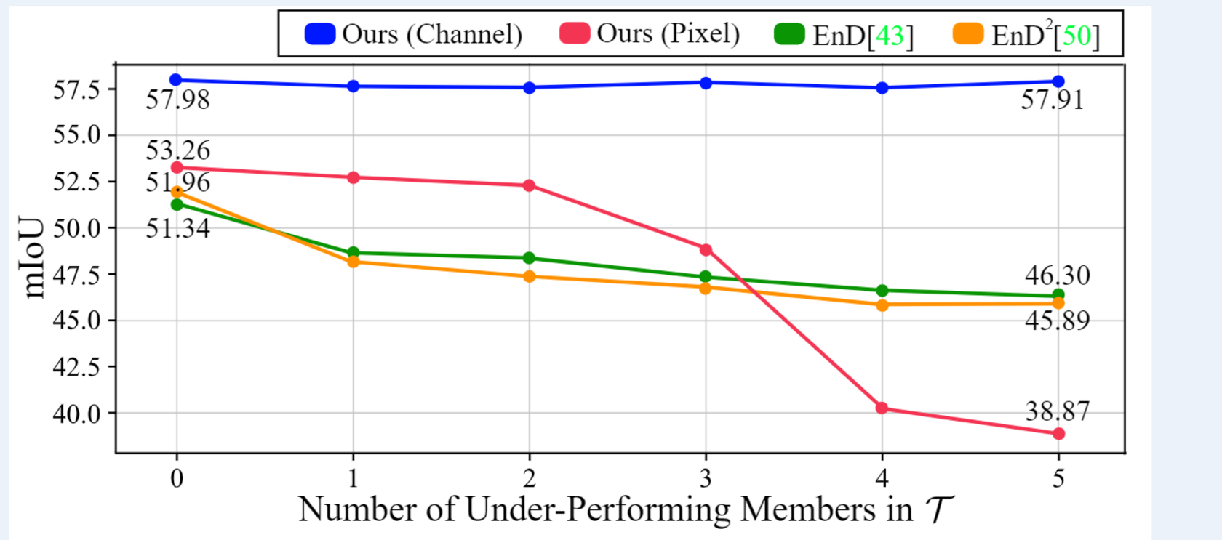


Table 2: The performance comparison of our framework with channel-wise fusion, our framework with pixel-wise fusion, and the baseline methods, with under-performing members (*'Source Only'* in **Table 1**) added to $\mathcal{T}$.

(3) Flexibility

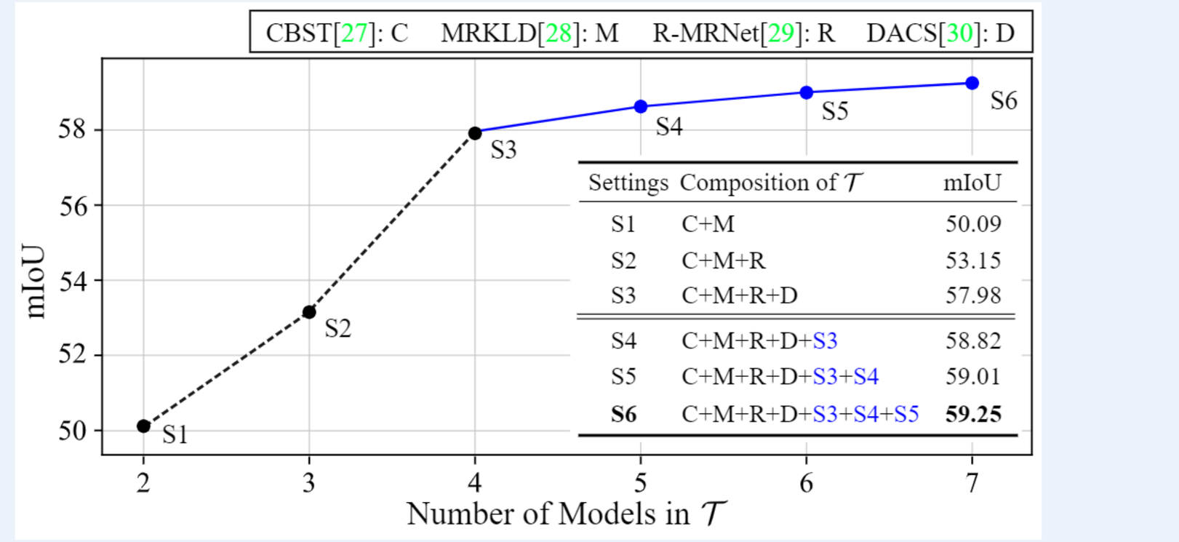


Table 3: The performance of the student models trained with the proposed framework using channel-wise fusion with different compositions of $\mathcal{T}$.

Conclusion

In this paper, we presented a flexible ensemble-distillation framework to address the common pitfalls, i.e., the lack of robustness, of previous methods. We incorporated an output unification operation into the proposed framework to ensure that the fused outputs of the ensemble are free of the influence from the certainty inconsistency among the models in the ensemble. In addition, to tackle the performance variation issue, we proposed a channel-wise fusion function that is robust against this issue. As our framework is able to integrate different types of UDA methods while maintaining its robustness, it therefore pioneers a new direction for future semantic segmentation based UDA researches.

Contributions

We introduce a flexible UDA ensemble-distillation framework which is robust against the inconsistency in the scale of the output certainty values and the performance variations among the members in an ensemble.

Questions?

lance_chao@gapp.nthu.edu.tw
bobcheng15@gapp.nthu.edu.tw
cylee@gapp.nthu.edu.tw

Acknowledgements

TWCC NVIDIA MEDIATEK MOST 科技部 Ministry of Science and Technology