Appendix: *Continual Distillation of Teachers from Different Domains*

Nicolas Michel[1,4] Maorong Wang[2] Jiangpeng He[3,†] Toshihiko Yamasaki[1,†]
[1]The University of Tokyo [2]National Institute of Informatics
[3]Indiana University Bloomington [4]Japanese-French Laboratory of Informatics
{nicolas, yamasaki}@cvm.t.u-tokyo.ac.jp, maorong@nii.ac.jp, jhe2@iu.edu

A. Experimental Setup

A.1. Implementation Details

For training, we start from pre-trained weights and use the Adam optimizer with a learning rate of 0.0001 for 3 epochs. As students start from pre-trained weights, we observe negligible improvement when using more epochs. Images are resized to 224×224 to fit the size used during pre-training. We use random horizontal flips as augmentations and use data normalization. We use a batch size of 64. Regarding the distillation, we use the KL-divergence with a temperature of 10. Experiments are conducted with ViT architecture, namely the ViT-B/16. We use ViT-base as teachers for CIFAR20 and DomainNet. For Digits, we use a ViT-tiny as a teacher. Every teacher is initialised from pre-trained weights and trained for 50 epochs with Adam optimizer and a learning rate of 0.0001. We use the same architectures for students in the main draft and additionally experiment with ViT-tiny. For all results, 3 seeds are used.

For DomainNet, since domains are unbalanced, we oversample or undersample them so that each task has the same number of steps. The base task length is determined by the Internal Data (ID) size, and the External Data (ED) is sampled accordingly. For example, if ID is larger than ED, we oversample from ED.

A.2. Metric

For all experiments, we report the domain-wise Final Accuracy ($\%$) of the student at the end of the training sequence.

B. Additional Discussions

Feasibility of obtaining unknown ED. In CD, a question naturally arises as to whether obtaining ED unknown to FMs is feasible, since such models are trained on vast and diverse data. We argue that obtaining unknown data is feasible in two highly plausible scenarios: **Temporal Gap:** Any data uploaded to public repositories *after* the FM's training cutoff is guaranteed to be unknown. **Private/Synthetic Data:** In industrial settings, private proprietary datasets or synthetically generated data serve as excellent ED.

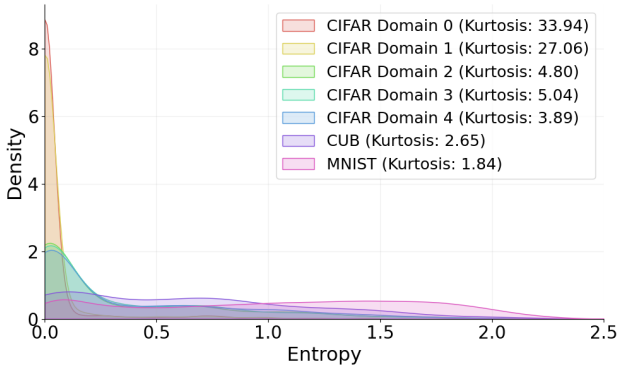

Figure B.1. Entropy distribution of the teacher (ViT-Base) trained on domains 0 and 1 of **CIFAR20** for various datasets and domains.

Assessing ED quality. As presented throughout this paper, ED selection is key in Continual Distillation. While quantifying semantic similarity is hard without teacher data, we propose using the teacher's own predictive uncertainty, by measuring the entropy, as a proxy. In Figure B.1, we show the entropy distribution of a teacher trained on two domains of CIFAR20, for various domains of CIFAR20 as well as CUB and MNIST. We observe that **the more the external data is "unrelated", the more "flat" the entropy distribution becomes**. To select adequate ED without knowing the training distribution, a user can 1) filter out samples based on an entropy threshold, 2) use the 4th-order moment (kurtosis) of the entropy distribution to quantify "flatness". As shown in Fig. B.1, lower flatness correlates with higher UKT potential. Additionally, in Figure B.2, the entropy distribution varies widely across domains, partially explaining the mitigated results observed on this dataset.

[†]Equal supervision.

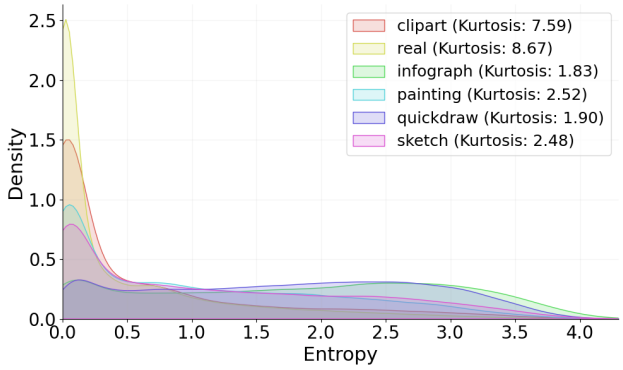

Figure B.2. Entropy distribution of the teacher (ViT-Base) trained on domains Real and Clipart of Domainnet for various **DomainNet** domains.

C. Additional Experiments

C.1. Additional Metrics

Forgetting For each method, we measure forgetting as the drop in performance on the learned domains after training from new teachers. Formally, for a student trained to imitate a teacher of index t, the forgetting for domain d is computed as:

$$F_d = \max_{i<t} A_d^{(i)} - A_d^{(t)}$$

where $A_d^{(i)}$ is the accuracy on domain d after distillation from teacher t. The overall forgetting is averaged across all domains:

$$F = \frac{1}{D}\sum_{d=1}^{D} F_d$$

This metric captures the extent to which a method forgets previously learned knowledge when adapting to new tasks. The results are presented in Tables C.1, C.2 and C.3. It can be observed that our method leads to competitive forgetting in all scenarios.

Accuracy Curves We report per-domain accuracy curves during training with related external data in Figures D.1 to Figure D.4.

C.2. Additional Architecures

We report results with additional architectures. Namely, we experimented with a ViT-tiny as a student instead of the ViT-base version in the main manuscript. Similarly, we experimented with larger models as teacher using CLIP-base teachers (ViT-L/14). Results are presented in Table C.5.

Table C.1. Forgetting (%, lower is better) of the student at the end of training on CIFAR20 for 4 scenarios. Internal Data Only (D0), Related External Data (D4), CUB as ED, and MNIST as ED. The number of runs is set to 3.

CIFAR20 - Internal Data Only

Method	D0	D1	D2	D3	D4 ✗	Avg. (0-3)
KL-divergence	0	11.95	2.98	0	-	3.73
DKD [CVPR'22]	0	**2.38**	3.47	0	-	**1.46**
LS [CVPR'24]	0	10.42	**1.88**	0	-	3.08
MDS [ICLR'25]	0	11.95	2.98	0	-	3.23
Self-Distillation	0	16.02	2.47	0	-	4.12

CIFAR20 + Related External Data

Method	D0	D1	D2	D3	D4	Avg. (0-3)
KL-divergence	0.52	38.15	30.26	0	-	17.23
DKD [CVPR'22]	0.13	25.15	16.25	0	-	10.38
LS [CVPR'24]	0.62	39.23	32.08	0	-	17.98
MDS [ICLR'25]	0.29	37.74	31.00	0	-	17.26
Self-Distillation	0	25.33	7.93	0	-	8.32
SE2D (ours)	0	**16.10**	**3.67**	0	-	**4.44**

CIFAR20 + CUB

Method	D0	D1	D2	D3	CUB	Avg. (0-3)
KL-divergence	0.93	30.68	17.82	0	-	12.36
DKD [CVPR'22]	0.89	**5.37**	3.86	0	-	**2.03**
LS [CVPR'24]	2.54	32.28	20.66	0	-	11.87
MDS [ICLR'25]	0.61	30.79	16.26	0	-	11.41
Self-Distillation	0.81	27.88	**2.48**	0	-	7.29
SE2D (ours)	**0.40**	21.61	6.81	0	-	7.21

CIFAR20 + MNIST

Method	D0	D1	D2	D3	MNIST	Avg. (0-3)
KL-divergence	0	15.20	7.40	0	-	5.15
DKD [CVPR'22]	0	**2.74**	**0.57**	0	-	**0.83**
LS [CVPR'24]	1.57	14.58	9.50	0	-	5.41
MDS [ICLR'25]	0.97	15.34	7.67	0	-	5.00
Self-Distillation	0	9.69	1.02	0	-	2.18
SE2D (ours)	0.85	9.88	2.85	0	-	4.43

Algorithm 1 Continual Distillation with KL divergence and SGD.

Require: Sequence of teachers $\{\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_T\}$, student model $\mathcal{S}_\theta$, distillation dataset $\mathcal{D}^{\mathcal{S}}$

for $t = 1$ to N **do**
  for $x \in \mathcal{D}^{\mathcal{S}}$ **do**
    Obtain teacher predictions $p_t(x) = \mathcal{T}_t(x)$
    Student predictions $q_\theta(x) = \mathcal{S}_\theta(x)$
    Distillation loss: $\mathcal{L}_t = \mathrm{KL}\big(p_t(x) \,\|\, q_\theta(x)\big)$
    Update student parameters $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_t$
  end for
end for
return Trained student model $\mathcal{S}_\theta$

C.3. Additional Sequences

As presented in the main paper, DomainNet is particularly challenging. Therefore, we conducted additional experiments where challenging domains have either been removed

Algorithm 2 Overview of SE2D training algorithm with Continual Distillation.

Require: Sequence of teachers $\{\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_T\}$, student model $\mathcal{S}_\theta$, distillation dataset $\mathcal{D}^{\mathcal{S}} = \mathcal{D}_i \cup \mathcal{D}_e$

for $t = 1$ to N **do**
  if $t = 1$ **then**
    for $x \in \mathcal{D}^{\mathcal{S}}$ **do**
      Obtain teacher predictions $p_t(x) = \mathcal{T}_t(x)$
      Compute student predictions $q_\theta(x) = \mathcal{S}_\theta(x)$
      Compute distillation loss: $\mathcal{L}_t = \mathrm{KL}\big(p_t(x) \,\|\, q_\theta(x)\big)$
      Update student parameters $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_t$
    end for
  else
    Load previous student checkpoint $\mathcal{S}_\theta^{t-1}$
    for $(x_e, x_i) \in (\mathcal{D}_e, \mathcal{D}_i)$ **do**
      Compute student predictions on internal data: $q_\theta(x_i) = \mathcal{S}_\theta(x_i)$
      Compute student predictions on external data: $q_\theta(x_e) = \mathcal{S}_\theta(x_e)$
      Obtain previous student predictions on external data: $p_{t-1}(x_e) = \mathcal{S}_\theta^{t-1}(x_e)$
      Obtain current teacher predictions on all data: $p_t^{\text{teacher}}((x_e, x_i)) = \mathcal{T}_t((x_e, x_i))$
      Compute distillation loss on external data from previous student: $\mathcal{L}_{\text{student}} = \mathrm{KL}\big(p_{t-1}(x_e) \,\|\, q_\theta(x_e)\big)$
      Compute distillation loss on all data from teacher: $\mathcal{L}_{\text{teacher}} = \mathrm{KL}\big(p_t^{\text{teacher}}((x_e, x_i)) \,\|\, q_\theta((x_e, x_i))\big)$
      Compute total loss: $\mathcal{L}_t = \mathcal{L}_{\text{student}} + \mathcal{L}_{\text{teacher}}$
      Update student parameters $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_t$
    end for
  end if
end for
return Trained student model $\mathcal{S}_\theta$

Table C.2. Forgetting (%, lower is better) of the student at the end of training on Digits for 2 scenarios. Internal Data Only (D0), Related External Data (D4). The number of runs is set to 3.

Digits - Internal Data Only

Method	MNIST	SVHN	MNIST-M	USPS	KMNIST ✗	Avg.
KL-divergence	0.23	3.34	16.20	0	-	4.94
DKD [CVPR'22]	0.47	**0.53**	11.73	0	-	**3.12**
LS [CVPR'24]	**0.10**	4.09	12.88	0	-	4.40
MDS [ICLR'25]	0.23	3.34	16.20	0	-	4.94
Self-Distillation	0.13	4.06	**7.46**	0	-	3.54

Digits + Related External Data

Method	MNIST	SVHN	MNIST-M	USPS	KMNIST	Avg.
KL-divergence	0.24	40.68	38.41	0	-	19.17
DKD [CVPR'22]	0.84	12.00	50.77	0	-	15.87
LS [CVPR'24]	0.27	40.50	36.59	0	-	19.22
MDS [ICLR'25]	0.25	40.69	38.35	0	-	19.16
Self-Distillation	**0.10**	16.33	6.35	0	-	5.58
SE2D (ours)	0.13	**10.37**	**4.90**	0	-	**3.73**

or used as external data. Such results are presented in Table C.6. Despite SE2D's lower overall performance compared to Self-Distillation, the disparity between the two methods diminishes in less complex scenarios, where SE2D simultaneously expands its lead over the baseline.

D. Algorithms

To provide a clear overview of our training methodology, we present the Continual Distillation procedure in Algorithm 1. Furthermore, a detailed description of our proposed SE2D approach is provided in Algorithm 2.

Table C.3. Forgetting (%, lower is better) on DomainNet for 2 scenarios: Internal Data Only (Seq -1) and Related External Data (Seq 0). Averaged across 3 runs.

DomainNet - Internal Data Only							
Method	Clipart	Infograph	Painting	Quickdraw	Real	Sketch ✗	Avg.
KL-divergence	4.63	9.91	13.70	20.31	0	-	12.14
DKD [CVPR'22]	3.85	10.34	13.89	22.38	0	-	12.61
LS [CVPR'24]	3.50	11.48	15.39	23.73	0	-	13.52
MDS [ICLR'25]	5.20	10.70	15.69	21.46	0	-	13.26
Checkpoint	**1.27**	**5.86**	**11.30**	20.68	0	-	**9.78**
SE2D (ours)	4.60	9.58	13.49	**20.27**	0	-	11.98

DomainNet + Related External Data							
Method	Clipart	Infograph	Painting	Quickdraw	Real	Sketch ✗	Avg.
KL-divergence	5.56	11.89	18.07	29.12	0	-	16.16
DKD [CVPR'22]	4.64	12.15	18.31	29.69	0	-	16.20
LS [CVPR'24]	5.40	13.70	20.98	32.44	0	-	18.13
MDS [ICLR'25]	6.25	13.03	19.68	29.62	0	-	17.15
Checkpoint	**1.70**	**6.63**	**14.54**	**29.43**	0	-	**13.08**
SE2D (ours)	3.42	6.91	15.54	30.86	0	-	14.18

Table C.4. Performances (%, higher is better) of the student at the end of training on CIFAR20 for 2 scenarios with a ViT-tiny. Internal Data Only (D0), Related External Data (D4). The number of runs is set to 3. Average and standard deviations are reported.

CIFAR20 - Internal Data Only						
Method	D0	D1	D2	D3	D4 ✗	Avg. (0-3)
$\mathcal{T}_{best}$ (upper bound)	97.75	95.80	96.70	95.75	-	96.5
KL-divergence	96.27 ± 0.21	37.85 ± 0.78	52.28 ± 0.42	49.48 ± 1.28	-	58.97 ± 0.67
DKD [CVPR'22]	95.40 ± 1.05	33.27 ± 1.60	45.38 ± 0.90	40.45 ± 1.34	-	53.62 ± 1.22
LS [CVPR'24]	96.25 ± 0.85	$\mathbf{38.95 \pm 2.17}$	51.58 ± 2.47	$\mathbf{50.15 \pm 4.40}$	-	$\mathbf{59.23 \pm 2.47}$
MDS [ICLR'25]	94.45 ± 0.44	35.08 ± 0.88	45.35 ± 0.41	41.98 ± 3.10	-	54.22 ± 1.21
Self-Distillation	$\mathbf{96.27 \pm 0.35}$	37.93 ± 1.22	$\mathbf{51.82 \pm 0.53}$	46.43 ± 0.97	-	58.11 ± 0.77

CIFAR20 + Related External Data						
Method	D0	D1	D2	D3	D4	Avg. (0-3)
$\mathcal{T}_{best}$ (upper bound)	97.75	95.80	96.70	95.75	-	96.5
KL-divergence	96.36 ± 0.33	46.54 ± 1.17	55.86 ± 1.10	75.96 ± 1.76	-	68.68 ± 1.09
DKD [CVPR'22]	95.18 ± 0.46	40.77 ± 0.95	50.92 ± 1.07	60.69 ± 1.32	-	61.89 ± 0.95
LS [CVPR'24]	96.45 ± 0.26	46.80 ± 0.74	55.08 ± 0.61	$\mathbf{76.49 \pm 1.50}$	-	68.7 ± 0.78
Self-Distillation	$\mathbf{97.26 \pm 0.25}$	55.42 ± 0.31	$\mathbf{61.62 \pm 0.70}$	68.74 ± 1.46	-	70.76 ± 0.68
SE2D (ours)	96.77 ± 0.25	$\mathbf{62.33 \pm 1.19}$	59.45 ± 1.36	65.82 ± 2.03	-	$\mathbf{71.09 \pm 1.21}$

Table C.5. Domain Accuracy (%, higher is better) on DomainNet with CLIP-based teachers. Mean and standard deviation are reported.

DomainNet - Internal Data Only							
Method	Clipart	Infograph	Painting	Quickdraw	Real	Sketch ✗	Avg.
$\mathcal{T}_{best}$ (upper bound)	86.33	53.89	79.74	69.62	89.86	-	76.23
KL-divergence	78.33 ± 0.23	19.00 ± 0.30	39.25 ± 0.81	35.36 ± 0.49	51.51 ± 0.74	-	44.69 ± 0.32
DKD [CVPR'22]	78.55 ± 0.03	18.89 ± 0.41	39.69 ± 0.19	34.77 ± 0.38	52.45 ± 0.49	-	44.87 ± 0.06
LS [CVPR'24]	74.39 ± 0.39	15.54 ± 0.66	33.92 ± 0.56	20.09 ± 1.11	46.62 ± 0.76	-	38.11 ± 0.64
MDS [ICLR'25]	75.51 ± 0.10	17.52 ± 0.20	35.67 ± 1.39	26.22 ± 0.70	47.71 ± 0.69	-	40.53 ± 0.27
Self-Distillation	**79.99 ± 0.55**	**21.72 ± 0.80**	**43.65 ± 0.57**	26.52 ± 1.42	**57.21 ± 0.13**	-	**45.82 ± 0.61**

DomainNet + Related External Data							
Method	Clipart	Infograph	Painting	Quickdraw	Real	Sketch ✗	Avg.
$\mathcal{T}_{best}$ (upper bound)	86.33	53.89	79.74	69.62	89.86	-	76.23
KL-divergence	78.22 ± 0.39	19.98 ± 0.17	41.59 ± 0.55	**42.83 ± 0.37**	52.62 ± 0.40	-	47.05 ± 0.22
DKD [CVPR'22]	78.17 ± 0.11	20.01 ± 0.19	41.78 ± 0.39	42.69 ± 0.84	52.37 ± 0.24	-	47.00 ± 0.01
LS [CVPR'24]	74.64 ± 0.13	16.63 ± 0.31	37.09 ± 0.09	26.78 ± 0.35	47.45 ± 0.35	-	40.52 ± 0.17
MDS [ICLR'25]	75.21 ± 0.59	18.55 ± 0.45	39.32 ± 0.13	31.84 ± 0.73	49.32 ± 0.39	-	42.85 ± 0.39
Self-Distillation	**79.47 ± 0.50**	**22.84 ± 0.60**	**47.51 ± 1.33**	30.15 ± 0.77	**58.51 ± 1.37**	-	**47.69 ± 0.84**
SE2D (ours)	78.52 ± 0.10	21.34 ± 0.45	47.12 ± 0.35	29.18 ± 0.30	58.01 ± 0.61	-	46.83 ± 0.31

Table C.6. Accuracy per domain (%). Grey : Internal Data; Blue : Domain known by the teacher (active); Red : External Data (ED); White : Ignored. Avg computed on Internal Data + Active domains.

DomainNet - Sequence 1 (Quickdraw is used as ED and Infograph is ignored)							
Method	Clipart	Infograph	Painting	Quickdraw	Real	Sketch	Avg.
KL-divergence	74.66 ± 0.14	-	33.01 ± 0.21	-	46.17 ± 0.62	48.68 ± 0.15	50.63
DKD [CVPR'22]	76.18 ± 0.20	-	36.20 ± 0.49	-	49.70 ± 0.62	54.85 ± 0.32	54.23
MDS [ICLR'25]	70.76 ± 0.40	-	30.01 ± 0.78	-	43.52 ± 1.09	51.56 ± 0.68	48.96
Self-Distillation	**80.43 ± 0.10**	-	**47.47 ± 0.71**	-	**61.10 ± 0.59**	58.15 ± 0.31	**61.79**
SE2D (ours)	76.89 ± 0.25	-	38.72 ± 0.49	-	52.83 ± 0.20	**58.84 ± 0.25**	56.82

DomainNet - Sequence 2 (Infograph is used as ED)							
Method	Clipart	Infograph	Painting	Quickdraw	Real	Sketch	Avg.
KL-divergence	75.78 ± 0.14	-	34.61 ± 0.52	16.59 ± 0.22	49.27 ± 0.43	52.43 ± 0.44	45.73
DKD [CVPR'22]	76.73 ± 0.33	-	36.48 ± 0.68	17.53 ± 0.44	51.84 ± 0.55	56.83 ± 0.34	47.88
MDS [ICLR'25]	75.28 ± 0.17	-	35.43 ± 0.92	16.47 ± 0.48	50.74 ± 0.88	56.65 ± 0.37	46.91
Self-Distillation	**80.45 ± 0.06**	-	48.48 ± 0.41	**20.84 ± 0.89**	**64.79 ± 0.40**	**59.04 ± 0.24**	**54.72**
SE2D (ours)	78.08 ± 0.20	-	**49.02 ± 0.36**	18.50 ± 0.40	63.29 ± 0.32	58.99 ± 0.22	53.58

DomainNet - Sequence 3 (Sketch is used as ED, Infograph and Quickdraw are ignored)							
Method	Clipart	Infograph	Painting	Quickdraw	Real	Sketch	Avg.
KL-divergence	75.75 ± 0.35	-	42.44 ± 0.67	-	65.15 ± 0.53	-	61.11
DKD [CVPR'22]	76.25 ± 0.23	-	45.52 ± 0.72	-	70.50 ± 0.12	-	64.09
MDS [ICLR'25]	75.25 ± 0.07	-	44.74 ± 0.94	-	70.11 ± 0.35	-	63.37
Self-Distillation	**79.54 ± 0.26**	-	**59.61 ± 0.30**	-	**71.48 ± 0.26**	-	**70.21**
SE2D (ours)	78.04 ± 0.15	-	59.36 ± 0.68	-	70.61 ± 0.40	-	69.34

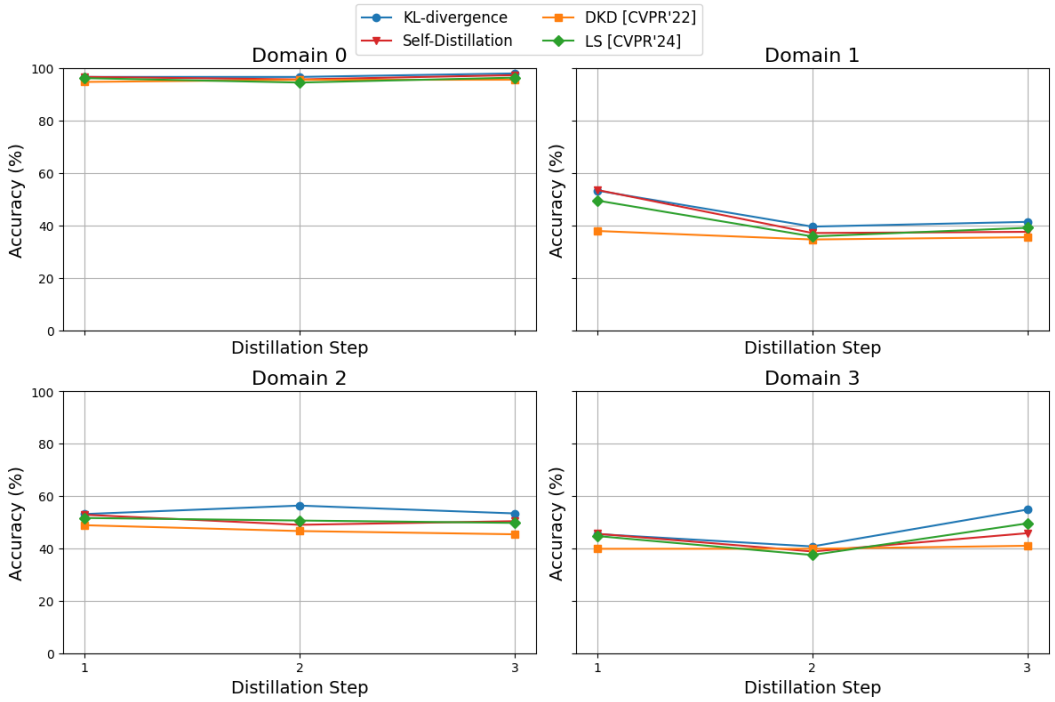

Figure D.1. Accuracy of the student on all domains at all steps for the considered methods on CIFAR20, training with Internal Data only.

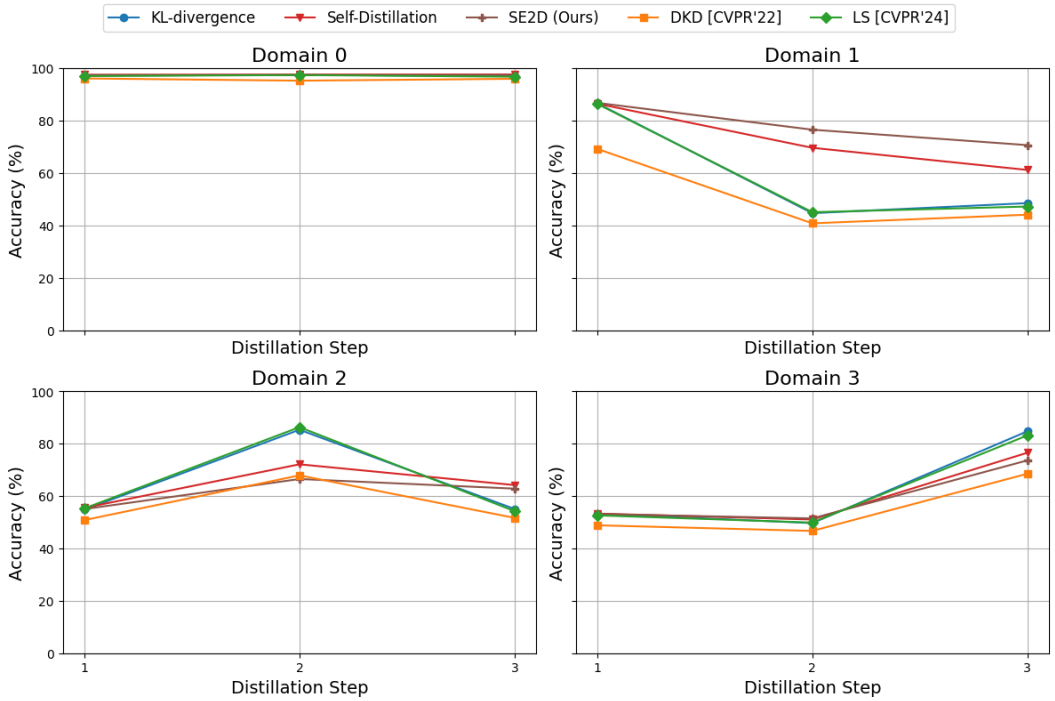

Figure D.2. Accuracy of the student on all domains at all steps for the considered methods on CIFAR20, training with External Data.

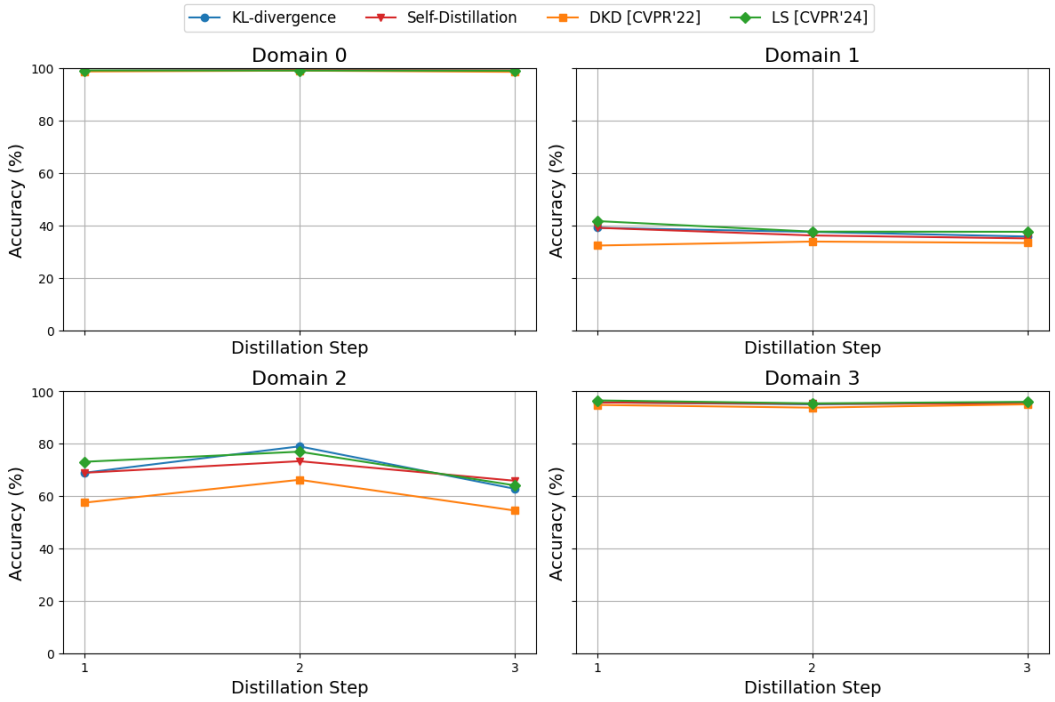

Figure D.3. Accuracy of the student on all domains at all steps for the considered methods on Digits, training with Internal Data only.

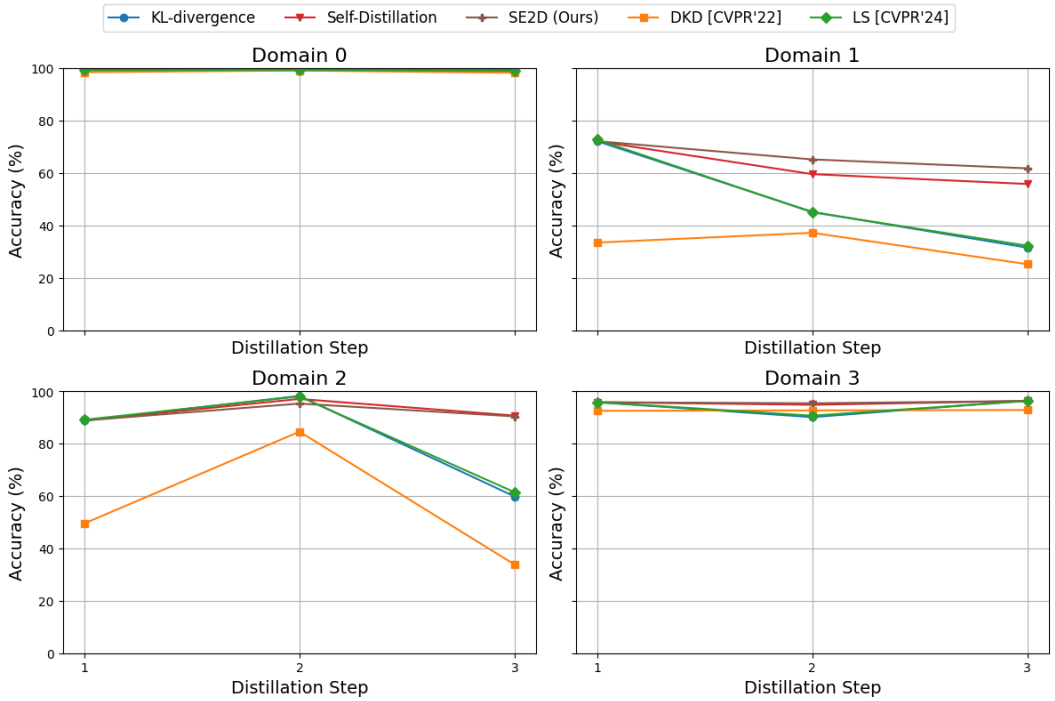

Figure D.4. Accuracy of the student on all domains at all steps for the considered methods on Digits, training with External Data.