Exploring Continual Distillation of Teachers from Different Domains

Anonymous Author(s)
Affiliation
Address
email

Abstract

With Foundation Models (FM) training costs rising to unprecedented heights, Continual Learning (CL) is a particularly compelling training paradigm that helps minimize the training cost of FM by incorporating new data incrementally instead of re-training from scratch. Instead of learning from a sequence of data suffering from domain shift, we propose **Continual Distillation (CD)**, a new paradigm where a single student model learns continuously from a sequence of teachers. Similar to CL, re-distilling from all teachers when introducing a new teacher is unsustainable, or even impossible when depending on a third-party provider. Therefore, when learning from one teacher, other teachers are considered unavailable. We show that CD naturally suffers from catastrophic forgetting, as knowledge distilled from the earlier teachers is forgotten when learning from subsequent teachers. Moreover, we find that the choice of the distillation data plays a central role, and even data unrelated to the teacher's original training domain can serve as an effective medium of knowledge transfer. This property has significant implications for FMs, where the original training dataset is often unavailable, undisclosed, or prohibitively large to reuse. While CD alleviates dependence on the original data, it also introduces security concerns, as a student may inadvertently absorb undesired knowledge without conscious control. Our study establishes CD as a new direction for leveraging foundation models in a resource-constrained environment.

1 Introduction

Foundation Models (FM) [6, 18, 7, 25, 1] have become the backbone of the modern deep learning community, offering remarkable generalization capability across diverse downstream tasks. Nevertheless, their unprecedented scale makes them expensive to train, costly to store, demanding to run inference on, and in many cases only accessible via restricted APIs. Consequently, as a natural paradigm for efficient training, Continual Learning (CL) [2, 23] has recently drawn significant attention. The core idea is to continuously train the model on a sequence of data, without forgetting previously acquired knowledge [8]. In this study, we introduce a complementary paradigm tailored to FMs, which we term **Continual Distillation (CD)**. In CD, instead of learning from a sequence of datasets, we propose to learn from a sequence of models. Precisely, a single student learns sequentially from multiple teachers without retaining access to earlier ones. The rise of foundation models makes this paradigm particularly salient, as, similar to CL, distilling from all teachers whenever a new teacher is introduced is unsustainable. For example, it is estimated that storing 10B parameters requires 38GB [4] and FM are commonly above 100B parameters. Training lightweight or task-specific models through distillation is therefore a practical necessity. Yet a major difficulty arises: the original training data of a foundation model is typically unavailable, undisclosed, or tremendously large to reuse. Thus, distillation must occur on a different dataset, likely representing a different domain. Such distillation is often referred to as Data-Free Distillation [30].

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

Eventually, when training a student in a Continual Distillation manner, we make several observations. Firstly, similar to Continual Learning, we show that the student suffers from catastrophic forgetting: knowledge distilled from earlier teachers gradually diminishes as the student adapts to subsequent ones. Moreover, we find that the domain overlap between teacher training data and distillation data plays a crucial role in which domain is learned, or forgotten. Surprisingly, even data completely unrelated to the teacher's training domain can serve as an effective medium of knowledge transfer. This property has profound implications: it alleviates dependence on the original dataset, yet also introduces risks, as a student may unknowingly absorb undesired knowledge. Since such transfer occurs independently of the distillation data itself, the changes are difficult to monitor; for example, if a teacher encodes harmful biases or unsafe behaviors, we found that these can be silently inherited by the student even when the distillation data appears benign. Our contributions are as follows:

- We introduce the paradigm of Continual Distillation, motivated by the practical challenges when previous teacher models are no longer accessible.
- We provide an empirical study showing that in CD, the student suffers from forgetting, as knowledge distilled from earlier teachers diminishes when learning from subsequent ones.
- We demonstrate that the choice of distillation data is crucial: even out-of-domain or unrelated data can effectively support knowledge transfer, revealing both opportunities for flexible distillation and risks of unintentionally propagating undesirable knowledge.

2 Related Work

Continual Learning. Continual Learning (CL) [2, 23] focuses on enabling models to learn from a sequence of tasks or data distributions while retaining previously acquired knowledge. In the era of Foundation Models, CL has become increasingly relevant because it allows computationally efficient updates of machine learning models given new data, leveraging existing representations to adapt to emerging tasks without retraining from scratch or accessing all previous data. While Knowledge Distillation is often employed in CL to mitigate forgetting [21], we instead investigate the phenomenon of forgetting within distillation itself, treating it as the central challenge of our study.

Knowledge Distillation. Knowledge Distillation (KD) [11] is a technique that allows a smaller student model to replicate the behavior of a larger, more capable teacher model. Traditionally, KD assumes that both the teacher and student are trained on the same dataset, and has been widely used for model compression [26, 33] and transfer learning [32]. With the development of the KD community, Data-Free KD (DFKD) [30] has emerged as a crucial technique for scenarios where the teacher's original training data is unavailable. Representative DFKD techniques include deep inversion, generative model-assisted sampling, or using out-of-domain data to approximate the teacher's behavior. While effective, these approaches often introduce distributional shifts between the distillation data and the teacher's original training data, which can negatively affect the quality of the distilled knowledge.

3 Continual Distillation

3.1 Definition

We define Continual Distillation (CD) as the process of distilling the knowledge from a **sequence of teacher models** continuously into **one student model**, on a **fixed dataset**. When distilling from one teacher to the student, other teachers are considered unavailable. The distillation process from a given teacher to the student is analogous to a task in standard CL. Formally, given a sequence of teachers $\{\mathcal{T}_0, \mathcal{T}_2, \ldots, \mathcal{T}_N\}$, each trained on a dataset $\mathcal{D}_t^{\mathcal{T}}$, the student $\mathcal{S}$ is optimized to minimize a distillation loss $\mathcal{L}_{dist}$, with respect to $\mathcal{T}_t$ on a distillation dataset $\mathcal{D}^{\mathcal{S}}$. Importantly, we only consider a distillation, no extra task-specific or label-dependent loss is considered. We present the overall procedure in Figure 1. We note $\mathbb{D}_{\mathcal{S}}(\mathcal{T}_{\mathcal{D}^{\mathcal{T}}}, \mathcal{D}^{\mathcal{S}})$ the operation of distilling from teacher $\mathcal{T}$ trained with $\mathcal{D}^{\mathcal{T}}$ to student $\mathcal{S}$ on distillation dataset $\mathcal{D}^{\mathcal{S}}$. We further omit the $\mathcal{S}$ from the notation for simplicity.

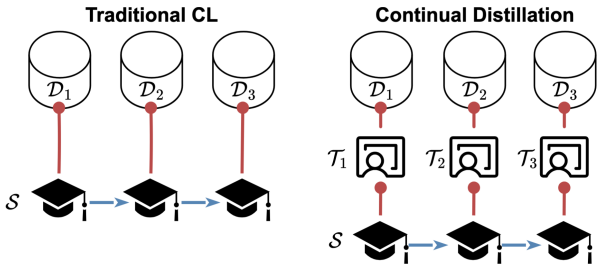

Figure 1: Continual Distillation. A student is trained on a teacher sequence instead of a data sequence.

3.2 Problem Scenarios

Traditional Knowledge Distillation assumes that the teacher training datasets are available for distillation. In CD, not only are such datasets considered unavailable, but dataset domains might differ from one teacher training to another. In other words, $\mathcal{D}_t^{\mathcal{T}}$, $\mathcal{D}_{t'}^{\mathcal{T}}$ and $\mathcal{D}^{\mathcal{S}}$ may cover partially or totally different domains. Therefore, we identify various scenarios.

Identical Teacher Domains (ITD) In this case, all teachers are trained on datasets stemming from the same underlying domain, i.e., $\mathcal{D}_t^{\mathcal{T}} = \mathcal{D}_{t'}^{\mathcal{T}} = \mathcal{D}^{\mathcal{T}}$. From this point, the domain of the distillation dataset can overlap to various degrees with $\mathcal{D}^{\mathcal{T}}$. In this study, we consider $\mathcal{D}^{\mathcal{T}}$ to be unavailable, therefore we study the case where $\mathcal{D}^{\mathcal{T}} \bigcap \mathcal{D}^{\mathcal{S}} = \emptyset$.

Exclusive Teacher Domains (ETD) Here, each teacher is trained on a dataset drawn from a disjoint domain, i.e., $\mathcal{D}_t^{\mathcal{T}} \cap \mathcal{D}_{t'}^{\mathcal{T}} = \emptyset$. We study both the case where $\exists t$ s.t. $\mathcal{D}_t^{\mathcal{T}} = \mathcal{D}^{\mathcal{S}}$ and $\mathcal{D}_t^{\mathcal{T}} \bigcap \mathcal{D}^{\mathcal{S}} = \emptyset, \forall t$.

Partially Exclusive Teacher Domains (PETD) A more realistic setting occurs when domains overlap only partially, i.e., $\mathcal{D}_t^{\mathcal{T}} \cap \mathcal{D}_{t'}^{\mathcal{T}} \neq \emptyset$. Precisely, we study the case where all teachers share a common domain but are additionally trained on exclusive domains. In this context, we study DC in two cases. (1) The student is trained on the shared domain $\mathcal{D}_0^{\mathcal{T}} \bigcap \mathcal{D}_1^{\mathcal{T}} = \cdots = \mathcal{D}_{N-1}^{\mathcal{T}} \bigcap \mathcal{D}_N^{\mathcal{T}} = \mathcal{D}^{\mathcal{S}}$ and (2) The student is trained on an external domain $\mathcal{D}_t^{\mathcal{T}} \bigcap \mathcal{D}^{\mathcal{S}} = \emptyset, \forall t$.

4 Experiments and Findings

4.1 Experimental Setup

Datasets We experiment with publicly available vision and language models, pretrained on **ImageNet-1000** [5], **English Wikipedia**, and **BookCorpus** [34]. For distillation, we consider **CIFAR-100** [13] and **CLINC** [15]. For domain-specific teachers, we consider **CIFAR-20** [27], a variation of CIFAR-100 using the 20 superclasses instead of the 100 fine-grained classes. Since in CIFAR-100, each superclass is composed of 5 sub-classes, using the superclasses allows for defining 5 different domains where each domain is images from different subclasses but identical super-class. This setup is traditionally used in Domain Incremental Learning [27]. More details and experiments on additional datasets such as **MNIST** [16] and **SVHN** [22] can be found in the appendix.

Continual Distillation Scenarios Details We consider all scenarios presented in Section 3.2, with KL-divergence [14] as loss. For **ITD**, the sequence of teachers is ViT-large, Swin-base, and DeiT-base, and the student is a ViT-base, all pre-trained on ImageNet. The distillation occurs on CIFAR-100, with predictions corresponding to ImageNet classes. See Table 1 as an example. For **ETD**, each teacher is a ViT-base fine-tuned on a specific domain of CIFAR-20, ranging from 0 to 4, for 10 epochs such that it reaches above 90% accuracy on each domain. An example can be found in Table 2. The student is the same architecture, and the distillation is done on the domain 0. In the case of **PETD**, the architecture is also a ViT-base for all models, and teachers are trained on overlapping domains of CIFAR-20. The distillation is done either on the domain that is shared across teachers or partially shared. Examples of such sequences can be seen in Table 3. For text experiments, we use pretrained RoBERTa-base, BART-base-uncased, and BERT-base-uncased models, and the student is the untrained BERT-base model trained from scratch, as shown in Table 4.

Table 1: Identical Teacher Domains. KL-divergence of the student model predictions compared to the teacher models predictions, on test sets, after being distilled from teachers $\{\mathcal{T}^{ViT}, \mathcal{T}^{Swin}, \mathcal{T}^{DEIT}\}$ on $\mathcal{D}_{\text{CIFAR100}}$.

Teacher	$\mathcal{T}^{ViT}$	$\mathcal{T}^{Swin}$	$\mathcal{T}^{DEIT}$
After $\mathbb{D}(\mathcal{T}_{\text{IN}}^{ViT}, \mathcal{D}_{\text{CIFAR100}})$	**0.2444**	1.5089	1.6760
After $\mathbb{D}(\mathcal{T}_{\text{IN}}^{Swin}, \mathcal{D}_{\text{CIFAR100}})$	1.2574	**0.2596**	1.1173
After $\mathbb{D}(\mathcal{T}_{\text{IN}}^{DEIT}, \mathcal{D}_{\text{CIFAR100}})$	1.3660	0.9335	**0.2982**

Table 2: Exclusive Teacher Domains, CIFAR-20. Accuracy (%) of the student model on test sets, domain-wise, after being distilled from teachers.

Domain	0	1	2	3	4
After $\mathbb{D}(\mathcal{T}_0, \mathcal{D}_0)$	**94.55**	35.95	46.25	42.75	41.10
After $\mathbb{D}(\mathcal{T}_1, \mathcal{D}_0)$	42.05	**90.10**	38.30	44.85	49.05
After $\mathbb{D}(\mathcal{T}_2, \mathcal{D}_0)$	55.50	42.55	**89.20**	48.05	40.30
After $\mathbb{D}(\mathcal{T}_3, \mathcal{D}_0)$	43.45	44.75	47.55	**89.65**	44.55
After $\mathbb{D}(\mathcal{T}_4, \mathcal{D}_0)$	49.00	42.25	30.40	40.15	**89.30**

4.2 Experimental Results and Discussion

Student Forgetting One way to evaluate forgetting in our setup is to measure the evolution of the similarity between the student and all teachers. Similar to standard CL, we observe forgetting from the student when distilled from new teachers. For example, in Table 1 and Table 4, when measuring the similarity in output distributions of student-teacher pairs through KL-divergence and Maximum Mean Discrepency (MMD) [10] with an RBF kernel, it can be seen that the KL-divergence systematically increases with previous teachers, while being minimal with the current teacher. Likewise, MMD captures the same trend; we believe such observations to be expected when learning from new models, as output distributions shift over time when distilling from later models.

Table 3: Partially Exclusive Teacher Domains, CIFAR-20. Accuracy (%) of the student model on test sets, domain-wise, for various domain overlaps.

Domain	0	1	2	3	4
After $\mathbb{D}(\mathcal{T}_{01}, \mathcal{D}_{04})$	**96.35**	77.15	48.95	48.45	58.65
After $\mathbb{D}(\mathcal{T}_{02}, \mathcal{D}_{04})$	**96.35**	43.3	80.1	46.55	48.75
After $\mathbb{D}(\mathcal{T}_{03}, \mathcal{D}_{04})$	**95.7**	49.4	57.6	77.8	53.95
After $\mathbb{D}(\mathcal{T}_{01}, \mathcal{D}_{034})$	**94.60**	85.20	51.85	57.15	58.05
After $\mathbb{D}(\mathcal{T}_{02}, \mathcal{D}_{034})$	**94.60**	44.00	83.55	51.55	47.95
After $\mathbb{D}(\mathcal{T}_{012}, \mathcal{D}_{014})$	**92.70**	**93.15**	68.35	53.80	55.85
After $\mathbb{D}(\mathcal{T}_{013}, \mathcal{D}_{014})$	**92.45**	**92.95**	48.95	72.55	62.65

Domain	0	1	2	3	4
After $\mathbb{D}(\mathcal{T}_{01}, \mathcal{D}_0)$	**93.25**	37.80	45.30	36.90	42.15
After $\mathbb{D}(\mathcal{T}_{02}, \mathcal{D}_0)$	**93.95**	31.60	42.20	35.25	35.05
After $\mathbb{D}(\mathcal{T}_{03}, \mathcal{D}_0)$	**92.10**	32.00	39.05	33.00	31.05
After $\mathbb{D}(\mathcal{T}_{04}, \mathcal{D}_0)$	**91.05**	30.20	41.00	33.00	32.35

Unseen Domain Forgetting Another way to consider forgetting is to study student performances domain-wise. If the student and teacher show strong performances in the same domains, we assume that domain-related knowledge is shared. First, we consider domain-exclusive teachers, distilled on domain 0, as presented in Table 2. Two observations can be made: (1) the student forgets what it has learned from $\mathcal{T}_0$ as the performances drop on domain 0; (2) on subsequent tasks, the student learns and then forgets knowledge corresponding to domains it has never encountered (1 to 4). We call this phenomenon *Unseen Domain Forgetting*. Interestingly, such a phenomenon does not occur when the student distillation domain is included in the teacher domain. For example, in Table 3, right side, teachers all share domain 0, which is the domain used for distillation. Here, there is no learning or forgetting in unseen domains. However, when distillation also occurs on domains unknown to the teacher, unseen domain forgetting occurs. For example, Table 3 left side, where the student is distilled on domains that partially include teachers shared domains (e.g. 0), but also external domains (e.g. 4).

Relation Between Domains Overlap and Unseen Domains Performances In our experiments, we found that this unseen domain forgetting behaviour occurs to a varying degree depending on the overlap between teacher domains and student domains. To show this more clearly, we compute the average accuracy on domains unseen by the student, but known by the teacher (underlined values in tables). Notably, we report such values with regard to the amount of distillation data that fall outside the teacher domain in Figure 2. For example, we say that there is 33% of data outside teacher domain for $\mathbb{D}(\mathcal{T}_{012}, \mathcal{D}_{014})$ and report performances on domain 2. We observe that performances on unseen domains increase proportionally to the amount of data outside of teacher domains used for distillation. Eventually, when the original dataset is not available, it is impossible to guarantee that the distillation domain is fully included in the teacher domains; particularly when leveraging generated data.

Table 4: Partially Exclusive Teacher Domains, text datasets. Maximum Mean Discrepency (MMD) of the student's feature representation compared to the teacher models' representation on test set, after being distilled from teachers $\{\mathcal{T}^{RoBERTa}, \mathcal{T}^{BART}, \mathcal{T}^{BERT}\}$ on $\mathcal{D}_{\text{CLINC}}$. Details about the teachers' training dataset are shown in appendix.

Teacher	$\mathcal{T}^{RoBERTa}$	$\mathcal{T}^{BART}$	$\mathcal{T}^{BERT}$
After $\mathbb{D}(\mathcal{T}^{RoBERTa}_{\text{5-Dataset}}, \mathcal{D}_{\text{CLINC}})$	**0.1917**	0.3821	0.3676
After $\mathbb{D}(\mathcal{T}^{BART}_{\text{2-Dataset}}, \mathcal{D}_{\text{CLINC}})$	0.4721	**0.0104**	0.4151
After $\mathbb{D}(\mathcal{T}^{BERT}_{\text{2-Dataset}}, \mathcal{D}_{\text{CLINC}})$	0.2943	0.2601	**0.1367**

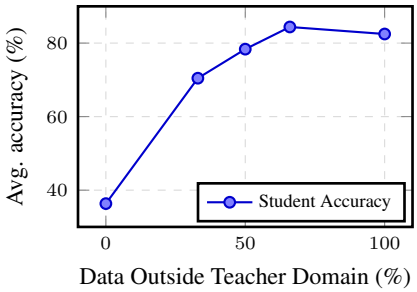

Figure 2: Average accuracy on unseen domains w.r.t. the percentage of student data outside teacher domains.

4.3 Experiments with external data

We experimented in two scenarios, one where the teachers are trained on a purely digit dataset only, and the other where we also consider Japanese hiragana to be a different domain of digits. The result are presented in Table 5 and Table 6. Similar to the results presented in the main draft, it is clear that some domain forgetting can still be observed, even though it is not as pronounced, especially for MNIST-M. We believe this is a direct consequence of MNIST and MNIST-M having strong domain similarity. Nonetheless, as shown in Table 5, performances still decrease from 62.97 to 44.97 on SVHN while increasing from 86.72 to 96.64 on MNIST-M.

Table 5: Partially Exclusive Teacher Domains, Digits datasets. Accuracy (%) of the student model on test sets, domain-wise, when training with teachers $\{\mathcal{T}_{\text{MNIST,SVHN}}, \mathcal{T}_{\text{MNIST,MNIST-M}}\}$ on $\{\mathcal{D}_{\text{MNIST,KMNIST}}\}$.

Domain	MNIST	SVHN	MNIST-M	KMNIST
After $\mathbb{D}(\mathcal{T}_{\text{MNIST,SVHN}}, \mathcal{D}_{\text{MNIST-M,KMNIST}})$	**99.22**	<u>62.97</u>	86.72	7.25
After $\mathbb{D}(\mathcal{T}_{\text{MNIST,MNIST-M}}, \mathcal{D}_{\text{MNIST-M,KMNIST}})$	**99.37**	44.97	<u>96.64</u>	7.95

Table 6: Partially Exclusive Teacher Domains, iDigits with KMNIST as training dataset, Fashion-MNIST as distillation dataset. Accuracy (%) of the student model on test sets, domain-wise, when training with teachers $\{\mathcal{T}_{\text{MNIST,SVHN}}, \mathcal{T}_{\text{MNIST,MNIST-M}}, \mathcal{T}_{\text{MNIST,KMNIST}}\}$ on $\{\mathcal{D}_{\text{MNIST,Fashion-MNIST}}\}$.

Domain	MNIST	SVHN	MNIST-M	KMNIST	Fashion-MNIST
After $\mathbb{D}(\mathcal{T}_{\text{MNIST,SVHN}}, \mathcal{D}_{\text{MNIST,Fashion-MNIST}})$	**98.91**	<u>79.01</u>	88.08	7.34	15.82
After $\mathbb{D}(\mathcal{T}_{\text{MNIST,MNIST-M}}, \mathcal{D}_{\text{MNIST,Fashion-MNIST}})$	**98.93**	45.37	<u>97.71</u>	7.60	6.58
After $\mathbb{D}(\mathcal{T}_{\text{MNIST,KMNIST}}, \mathcal{D}_{\text{MNIST,Fashion-MNIST}})$	**99.19**	17.68	61.22	<u>67.48</u>	17.26

5 Conclusion

In this study, we introduced Continual Distillation, a paradigm for learning a single student from a sequence of teachers, where only one teacher is available at a time. We observe forgetting of previous teachers and note that the student may gain or lose knowledge on domains unseen during distillation, but known to the teachers. Empirically, the fraction of data outside the teachers' domains strongly influences student absorption of unknown knowledge. This behavior can be beneficial or detrimental, depending on whether the acquired knowledge aligns with desired performance goals. Naturally, this study remains limited by the length of its experiments. Future work will investigate larger-scale datasets and architectures.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Zhiyuan Chen and Bing Liu. Lifelong machine learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 12(3):1–207, 2018.

[3] Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. Deep learning for classical japanese literature. *arXiv preprint arXiv:1812.01718*, 2018.

[4] Ankita De, Edward Wang, Rohan Varma, Anjali Sridhar, and Kartikay Khandelwal. Scaling multimodal foundation models in torchmultimodal with pytorch distributed. https://pytorch.org/blog/scaling-multimodal-foundation-models-in-torchmultimodal-with-pytorch-distributed, 2025.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[8] Robert M French. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135, 1999.

[9] Yaroslav Ganin, Evgeniya Ustinova, Hubert Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. In *International Conference on Machine Learning*, 2016.

[10] Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. A kernel method for the two-sample-problem. *Advances in Neural Information Processing Systems*, 19, 2006.

[11] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[12] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[13] Alex Krizhevsky. Learning multiple layers of features from tiny images. *University of Toronto*, 05 2012.

[14] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[15] Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1311–1316, 2019.

[16] Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

[17] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, 2020.

[18] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[19] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[20] Joel Mackenzie, Rodger Benham, Matthias Petri, Johanne R Trippas, J Shane Culpepper, and Alistair Moffat. Cc-news-en: A large english news corpus. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 3077–3084, 2020.

[21] Nicolas Michel, Maorong Wang, Ling Xiao, and Toshihiko Yamasaki. Rethinking momentum knowledge distillation in online continual learning. In *Forty-first International Conference on Machine Learning*, 2024.

[22] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, page 7. Granada, 2011.

[23] German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural networks*, 113:54–71, 2019.

[24] Joshua Peterson, Stephan Meylan, and David Bourgin. Openwebtext. https://github.com/jcpeterson/openwebtext, 2019.

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763, 2021.

[26] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

[27] Stefan Stojanov, Samarth Mishra, Ngoc Anh Thai, Nikhil Dhanda, Ahmad Humayun, Chen Yu, Linda B Smith, and James M Rehg. Incremental object learning from contiguous views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8777–8786, 2019.

[28] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357, 2021.

[29] Trieu H Trinh and Quoc V Le. A simple method for commonsense reasoning. *arXiv preprint arXiv:1806.02847*, 2018.

[30] Yuzheng Wang, Dingkang Yang, Zhaoyu Chen, Yang Liu, Siao Liu, Wenqiang Zhang, Lihua Zhang, and Lizhe Qi. De-confounded data-free knowledge distillation for handling distribution shifts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12615–12625, 2024.

[31] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[32] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4133–4141, 2017.

[33] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11953–11962, 2022.

[34] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *The IEEE International Conference on Computer Vision*, 2015.

Appendix

A Related Work

Continual Learning. Continual Learning (CL) [2, 23] focuses on enabling models to learn from a sequence of tasks or data distributions while retaining previously acquired knowledge. In the era of Foundation Models, CL has become increasingly relevant because it allows computationally efficient updates of machine learning models given new data, leveraging existing representations to adapt to emerging tasks without retraining from scratch or accessing all previous data. While Knowledge Distillation is often employed in CL to mitigate forgetting [21], we instead investigate the phenomenon of forgetting within distillation itself, treating it as the central challenge of our study.

Knowledge Distillation. Knowledge Distillation (KD) [11] is a technique that allows a smaller student model to replicate the behavior of a larger, more capable teacher model. Traditionally, KD assumes that both the teacher and student are trained on the same dataset, and has been widely used for model compression [26, 33] and transfer learning [32]. With the development of the KD community, Data-Free KD (DFKD) [30] has emerged as a crucial technique for scenarios where the teacher's original training data is unavailable. Representative DFKD techniques include deep inversion, generative model-assisted sampling, or using out-of-domain data to approximate the teacher's behavior. While effective, these approaches often introduce distributional shifts between the distillation data and the teacher's original training data, which can negatively affect the quality of the distilled knowledge.

B Additional Results

B.1 Results on Digits datasets

In order to further confirm the results presented in the main draft, we experimented on additional computer vision datasets that all share the same classes (the 10-digit classification), but from different domains. Namely, we trained two teachers on:

- MNIST [16], the standard 10 digits classification dataset; containing 60,000 $28x28$ train images and 10,000 test images;
- KMNIST [3], a dataset containing the same format and number of images as MNIST. However, the 10 classes are Japanese Hiragana characters instead of digits. We prefer to experiment with such a character as distillation datasets only since the classes are different; however, we equally experiment with this dataset as a different domain.
- MNIST-M [9], which consists of the original MNIST digit images that we blended with random color patches extracted from CIFAR10.
- SVHN [22], which consists of real-world images of house numbers collected from Google Street View. It contains 73,257 $32x32$ training images and 26,032 test images;
- Fashion-MNIST [31], which has the same format as MNIST, containing various instances of clothing. Similar to KMNIST, we prefer to experiment with such a dataset as a distillation dataset only, as the objects are different from digits.

B.2 Same Teacher and Student Domains

In this section, we provide some additional results on CIFAR-20. Notably, we report the accuracy of the student on CIFAR-20 when distilling from a sequence of domain-exclusive teachers, with a changing distillation dataset. This is similar to training in a traditional Domain Incremental Learning scenario, but the labels are the predictions of the teachers rather than actual labels. Such results are reported in Table 7. Naturally, we observe forgetting in this setup.

Table 7: Accuracy (%) of the student model on test sets, domain-wise, after being distilled from teachers $\{\mathcal{T}_0, \cdots, \mathcal{T}_4\}$ on $\{\mathcal{D}_0, \cdots, \mathcal{D}_4\}$.

Domain	0	1	2	3	4
After $\mathbb{D}(\mathcal{T}_0, \mathcal{D}_0)$	**94.55**	35.95	46.25	42.75	41.10
After $\mathbb{D}(\mathcal{T}_1, \mathcal{D}_1)$	53.25	**93.85**	38.55	46.00	50.15
After $\mathbb{D}(\mathcal{T}_2, \mathcal{D}_2)$	48.90	50.80	**92.30**	49.30	41.15
After $\mathbb{D}(\mathcal{T}_3, \mathcal{D}_3)$	42.30	54.85	60.95	**93.70**	45.35
After $\mathbb{D}(\mathcal{T}_4, \mathcal{D}_4)$	47.90	51.85	41.60	56.60	**94.90**

B.3 Additional PETD results

We report additional results regarding the PETD scenario. Notably, when for a sequence of teachers $\{\mathcal{T}_{01}, \mathcal{T}_{02}, \mathcal{T}_{03}\}$ with distillation on domain 4. Such results were used for computing the 100% case of Figure 2. We report these results in Table 8.

Table 8: Accuracy (%) of the student model on test sets, domain-wise, after being distilled from teachers $\{T_{01}, \cdots, T_{03}\}$ on $\{\mathcal{D}_4\}$.

Domain	0	1	2	3	4
After $\mathbb{D}(T_{01}, \mathcal{D}_4)$	**83.25**	**82.30**	45.90	51.80	58.75
After $\mathbb{D}(T_{02}, \mathcal{D}_4)$	**85.95**	43.10	**81.95**	50.50	49.80
After $\mathbb{D}(T_{03}, \mathcal{D}_4)$	**82.25**	43.85	49.15	**79.10**	55.25

B.4 Identical Teacher Domain with Opposite Sequence

We report the KL divergence between the student and the teachers when distilling on CIFAR-100, similar to the results presented in Table 1. To show that such results are not sequence-dependent, we reversed the order of the teachers and show the results in Table 9. A similar conclusion can be made, the student forgets what it has learned from previous teachers.

C Additional Experimental Details

Pre-Trained Models For vision models pre-trained on ImageNet we consider ViT-base [7], ViT-large, Swin-base [19], DeiT-base [28]. When training domain-specific teachers on CIFAR-20 domains, we start from a ViT-base pre-trained on ImageNet, which we train for 10 epochs using the Adam optimizer with a learning rate of 0.0001, on the considered domain. For Pre-trained language models, we use BART-base-uncased [17], BERT-base-uncased [6], and RoBERTa-base [18]. Following the official weights, both BART-base-uncased and BERT-base-uncased models are pre-trained on 2 datasets, including BookCorpus [34] and English Wikipedia. RoBERTa-base model is pretrained with 5 datasets, including BookCorpus [34], English Wikipedia, OpenWebText [24], Stories [29], and CC-News [20]

Implementation Details For training on CIFAR-100 and CIFAR-20, we use the Adam [12] optimizer with a learning rate of 0.0001 for 5 epochs. Images are resized to 224x224 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 1. For the training of student BERT-base model, we use the Adam optimizer with a learning rate of 2×10^{-5} for 5 epochs. In the distillation, we compute the loss function using the Maximum Mean Discrepency (MMD) with an RBF kernel.

Table 9: Identical Teacher Domains, vision datasets. KL-divergence of the student model predictions compared to the teacher models predictions, on test sets, after being distilled from teachers $\{\mathcal{T}^{DeiT}, \mathcal{T}^{Swin}, \mathcal{T}^{ViT}\}$ on $\mathcal{D}_{\text{CIFAR100}}$. The sequence is reversed compared to the main draft.

Teacher	$\mathcal{T}^{DeiT}$	$\mathcal{T}^{Swin}$	$\mathcal{T}^{ViT}$
After $\mathbb{D}(\mathcal{T}_{\text{IN}}^{DeiT}, \mathcal{D}_{\text{CIFAR100}})$	**0.3627**	1.0615	1.3921
After $\mathbb{D}(\mathcal{T}_{\text{IN}}^{Swin}, \mathcal{D}_{\text{CIFAR100}})$	1.0130	**0.2582**	1.3773
After $\mathbb{D}(\mathcal{T}_{\text{IN}}^{ViT}, \mathcal{D}_{\text{CIFAR100}})$	1.4943	1.2862	**0.2377**

D Algorithms of Continual Distillation

An undetailed algorithm and a detailed algorithm of the training procedure in Continual Distillation are presented in Algorithm 1 and Algorithm 2.

Algorithm 1 Continual Distillation overview, more details in Algorithm 2

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

for $t = 1$ to N **do**
  $\mathbb{D}_{\mathcal{S}}(\mathcal{T}_{\mathcal{D}_t^{\mathcal{T}}}, \mathcal{D}^{\mathcal{S}})$
end for
return Trained student model $\mathcal{S}$

Algorithm 2 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 $p_t(x) = \mathcal{T}_t(x)$
    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
end for
return Trained student model $\mathcal{S}_\theta$