SustainML

Application Aware, Life-Cycle Oriented Model-Hardware Co-Design Framework for Sustainable, Energy Efficient ML Systems

Knowledge recycling scheme and functional knowledge core catalog - Initial

Deliverable D3.3

WP3 - Energy Consumption Optimized ML Toolkit and Methods

This project has received funding from the European Union's Horizon Europe research and innovation programme (HORIZON-CL4-2021-HUMAN-01) under grant agreement No 101070408

eProsima
The Middleware Experts

DFKI

Inria

upmem

RPTU Rheinland-Pfälzische Technische Universität Kaiserslautern Landau

IBM

Project

Title:	SustainML: Application Aware, Life-Cycle Oriented Model-Hardware Co-Design Framework for Sustainable, Energy Efficient ML Systems
Acronym:	SustainML
Coordinator:	eProsima
Grant agreement ID:	101070408
Call:	HORIZON-CL4-2021-HUMAN-01
Program:	Horizon Europe
Start:	01 October 2022
Duration:	36 months
Website:	https://sustainml.eu
E-mail:	sustainml@eprosima.com
Consortium:	**eProsima (EPROS)**, Spain **DFKI**, Germany **TU Kaiserslautern (TUK)**, Germany **University of Copenhagen (KU)**, Denmark **National Institute for Research in Digital Science and Technology (INRIA)**, France **IBM Research GmbH**, Switzerland **UPMEM**, France

Deliverable

Number:	**D3.3**
Title:	**Knowledge recycling scheme and functional knowledge core catalog - Initial**
Month:	24
Work Package:	WP3 - Energy Consumption Optimized ML Toolkit and Methods
Work Package leader:	KU
Deliverable leader:	IBM
Deliverable type:	R
Dissemination level:	Public (PU)
Date of submission:	30/09/2024
Version:	v1.0
Status:	Finished

Version history

Version	Date	Responsible	Author/Reviewer	Comments
v1.0	30/09/2024	IBM	IBM	

Executive summary

SustainML project aims to develop a design framework and an associated toolkit, so-called SustainML, that will foster energy efficiency throughout the whole life-cycle of Machine Learning (ML) applications: from the design and exploration phase that includes exploratory iterations of training, testing and optimizing different system versions through the final training of the production systems (which often involves huge amounts of data, computation and epochs) and (where appropriate) continuous online re-training during deployment for the inference process. The framework will optimize the ML solutions based on the application tasks, across levels from hardware to model architecture. It will also collect both previously scattered efficiency-oriented research, as well as novel Green-AI methods. Artificial Intelligence (AI) developers from all experience levels can make use of the framework through its emphasis on human-centric interactive transparent design and functional knowledge cores, instead of the common blackbox and fully automated optimization approaches.

This report corresponds to *Deliverable D3.3 - Knowledge recycling scheme and functional knowledge core catalog - Initial*3.3. The goal of this deliverable is to improve the traditional transfer learning paradigm by developing more efficient methods for reusing and fine-tuning pre-trained neural network components, known as functional knowledge cores, across different tasks and domains. These cores, drawn from the knowledge catalog established in Task 1.3, enable faster adaptation and deployment of models for new tasks, minimizing the need for retraining and reducing resource consumption. The contributions of this deliverable center on three main innovations:

- A semantic-aware knowledge distillation framework that enables efficient model compression while preserving critical functionality.
- A novel probabilistic approach to feature matching that enables scalable, training-free visual prompting.
- An advanced parameter-efficient fine-tuning methodology that allows for efficient model adaptation.

© 2024 SustainML | HORIZON-CL4-2021-HUMAN-01 | 101070408

Contents

Executive Summary **3**

Contents **4**

Acronyms **5**

1 Introduction **6**

2 Knowledge Recycling Techniques **6**
2.1 Semantic Aware Knowledge Distillation 6
2.2 Probabilistic feature matching for fast scalable visual prompting 8
2.3 Parameter Efficient Fine-Tuning of vision foundation models 10

3 Conclusion **12**

References **14**

Acronyms

FMs	Foundation Models.
IA3	Infused Adapter by Inhibiting and Amplifying Inner Activations.
IoU	Intersection over Union.
LLMs	Large Language Models.
LVMs	Large Vision Models.
NLP	Natural Language Processing.
PEFT	Parameter-Efficient Fine-tuning.
RFF	Random Fourier Features.
SAM	Segment Anything Model.
VPT	Visual Prompt Tuning.

1 Introduction

Machine learning systems have become increasingly complex and computationally intensive, raising concerns about their environmental impact and sustainability. Traditional approaches to ML development often require extensive retraining and adaptation when deploying models to new domains or applications, leading to significant energy consumption and resource utilization.

Knowledge recycling emerges as a promising paradigm to address these challenges by enabling efficient reuse and adaptation of existing model knowledge. Rather than training new models from scratch, knowledge recycling techniques allow practitioners to leverage pre-existing capabilities while minimizing additional computational overhead. This approach aligns with the growing need for sustainable AI solutions that can deliver high performance while reducing environmental impact.

The concept of functional knowledge cores represents a fundamental shift in how we approach model development and deployment. These cores serve as reusable building blocks that can be efficiently adapted to new tasks and domains, significantly reducing the resources required for model development and deployment. By establishing a systematic approach to knowledge reuse, we can dramatically improve the efficiency of ML systems throughout their lifecycle.

This deliverable explores novel methods for implementing these concepts in practice, focusing on techniques that enable efficient knowledge transfer and adaptation across different applications and domains. The work presented here builds upon recent advances in foundation models, knowledge distillation, and parameter-efficient fine-tuning, while introducing new approaches that specifically target the challenges of sustainable ML development.

2 Knowledge Recycling Techniques

This deliverable introduces three major contributions designed to enhance the adaptation and efficiency of Foundation Models (FMs). First, we present a distillation technique capitalize on the knowledge from large models while preserving key semantic structures. Second, we develop a probabilistic feature matching algorithm for visual prompting, allowing for training-free task adaptation. Lastly, we present a parameter-efficient fine-tuning method that is able to adapt segmentation models reducing computational costs by fine-tuning only a minimal subset of parameters.

2.1 Semantic Aware Knowledge Distillation

The paper "TSAK: Two-Stage Semantic-Aware Knowledge Distillation for Efficient Wearable Modality and Model Optimization in Manufacturing Lines" [1] introduces a novel approach for improving efficiency in wearable Human Activity Recognition (HAR) systems used in smart factories. HAR systems are relevant for tracking human body movements and interaction, but existing solutions often require complex models and multiple sensors to interpret the human complexity, which leads to high computational costs and reduced battery life when being deployed in wearable devices [2]. Traditionally, HAR models therefore prioritize performance over efficiency, using multiple sensor inputs (e.g., IMU, gyroscope) and large models, making them impractical for wearable applications [3]. Moreover, existing knowledge distillation (KD) techniques [4] do not effectively enhance smaller models when distilling knowledge from larger, multi-modal models [5].

As a solution, TSAK introduces a two-stage KD framework designed to compress complex HAR models while maintaining high classification performance:

- Teacher Model: The teacher model combines multi-modal sensor data, including accelerometer and capacitive data, from both hands. The architecture integrates self-attention and causal representa-

tion extraction modules, which capture temporal and spatial information crucial for accurate HAR in manufacturing environments.

- Semantic Classifier: The second stage applies five distinct knowledge distillation strategies, including Attention Representation (Attn-Rep), Causal Representation (Causal-Rep), Combined Representation (Combi-Rep), Logit-based distillation, and the Semantic Classifier (SC-Logit), to efficiently transfer knowledge from the teacher to the student model while preserving task-specific semantics and reducing model complexity.

TSAK was evaluated using two datasets. The first is a Smart Factory Dataset, a new dataset collected within this work in a smart factory testbed using wearable gloves equipped with IMU and capacitive sensors. The second dataset is based on an open available dataset, OpenPack [6], which contains packaging work activities, also recorded using wrist-worn sensors. Both datasets represent real-world manufacturing tasks, and the system was tested in scenarios such as walking, touching screens/buttons, opening/closing doors, and working inside factory modules.

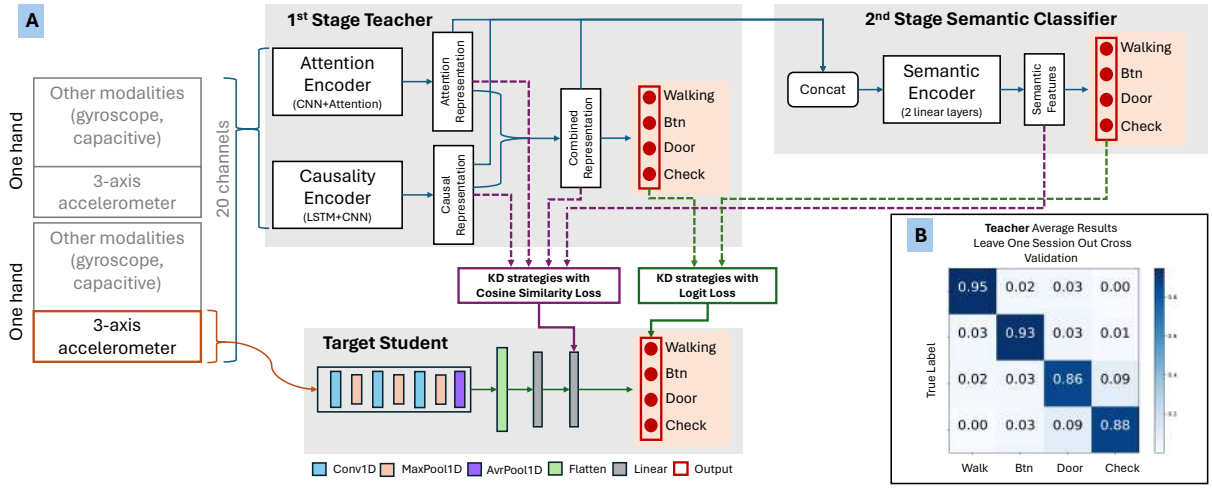

Figure 1: (A) In the TSAK knowledge distillation (KD) approach, five distillation methods were compared. The cosine similarity loss distills knowledge from one of the hidden vectors at a time; Attention Representation (Attn-Rep), Causal Representation (Causal-Rep), and Combined Representation (Combi-Rep). A shallow classifier is employed to merge and distill knowledge from all the hidden vectors simultaneously, preserving the semantics of the ground truth by logit-based KD (Semantic Classifier). Logit KD is also performed with the teacher's outputs. (B) Teacher average results with an F1 score of 85.91% across twelve users.

In more detail, as shown in Figure 1, the TSAK framework incorporates five distinct knowledge distillation strategies to optimize the transfer of information from the teacher model to the student model. These methods include Attention Representation (Attn-Rep), Causal Representation (Causal-Rep), and Combined Representation (Combi-Rep), which distill knowledge from various internal layers of the teacher. Additionally, Logit-based Knowledge Distillation transfers the teacher's soft outputs, while the Semantic Classifier (SC-Logit) method combines both feature-based and logit-based approaches. The SC-Logit method, in particular, yielded the best performance improvements, significantly enhancing the student model's accuracy while maintaining model efficiency.

The results of TSAK demonstrate significant improvements across both datasets, especially for the SC-Logit distillation strategy. For the Smart Factory Dataset, the student model trained using TSAK was highly efficient, achieving a model size that was 79% smaller than the teacher model. It ran 8.88 times

faster and consumed 96.6% less computational power. Importantly, these efficiency gains did not come at the cost of accuracy. The student model showed a notable increase in recognition performance, achieving up to 10% higher F1 scores compared to a model trained without TSAK. For key tasks of the dataset, such as door opening/closing and working inside factory modules, the student model reached an overall F1 score of 81.34%.

In the OpenPack Dataset, which focuses on industrial packaging work, TSAK also delivered improved results. Here, the student model saw an even more substantial performance improvement, with the F1 score increasing by 10.5% compared to the baseline model. The OpenPack student model, which was trained on one-handed acceleration data, demonstrated a strong recognition performance for packaging-related activities, showing that TSAK is versatile across different industrial tasks. The F1 score of the distilled student model reached 76.56%, which is 1.73% below the teacher model, demonstrating that TSAK successfully compresses the model while maintaining sufficient accuracy.

In summary, across both datasets, TSAK significantly improved both the efficiency and performance of the student model. It reduced the computational cost and size while preserving and, in many cases, enhancing the model's ability to recognize human activities in industrial environments. These results highlight TSAK's effectiveness in creating lightweight, high-performing models for wearable HAR systems.

2.2 Probabilistic feature matching for fast scalable visual prompting

This work presents a novel framework for visually guided image segmentation that exploits the capabilities of Vision FMs. Drawing on recent advancements in computer vision, the approach integrates multiple large-scale pre-trained models to address segmentation tasks with limited and sparsely annotated data provided interactively by the user. The framework uses a frozen feature extraction backbone to perform probabilistic feature matching between reference and target images. A pre-trained segmentation model translates user-provided scribbles into reference masks, which are then used to generate output target segmentation masks. The resulting framework, called Softmatcher, provides a versatile and efficient training-free architecture for image segmentation. It demonstrates efficiency and scalability for interactive, real-time segmentation in various visual domains, including use cases for technical visual inspection.

FMs have transformed the development and deployment of machine learning models by shifting from training task-specific models on tailored datasets to using generalist models pretrained on large-scale, diverse data. These models are then fine-tuned for a variety of tasks. In computer vision, models like Segment Anything Model (SAM) [7], CLIP [8], and self-supervised backbones such as DINO [9] and DINOv2 [10] have unlocked powerful functionalities such as object detection, semantic segmentation, and expressive embeddings, fueling a wide range of applications. Recent research has explored combining these models to tackle complex tasks in applied computer vision, such as visual inspection.

Building on these advancements, this work focuses on image segmentation, a fundamental task in computer vision with applications in fields such as medical imaging, autonomous driving, and visual inspection. The framework extends an earlier architecture, Matcher [11], which performs training-free few-shot segmentation using in-context examples and off-the-shelf FMs. Key improvements include integrating a pretrained segmentation model to convert user scribbles into reference masks and developing a scalable probabilistic feature soft-matching procedure, enabling real-time, interactive segmentation workflows with sparse user annotations.

The Softmatcher framework, highlighted in Figure 2 for interactive image segmentation guided by visual prompting operates in three main steps:

- **Prompting & reference segmentation:** The user provides scribbles on a reference image to indicate the object class for segmentation. These scribbles serve as sparse prompts for SAM, which generates a reference mask.

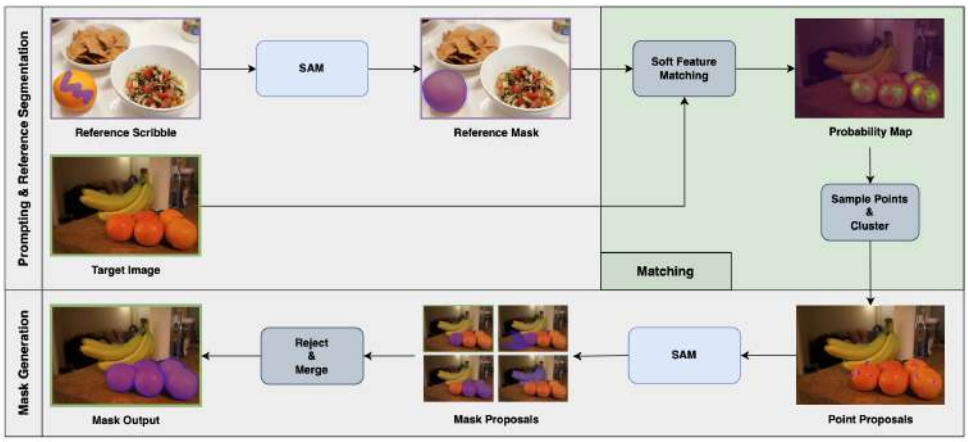

Figure 2: SoftMatcher Framework

- **Matching:** A probabilistic soft-matching process creates a probability map across target image pixels, quantifying how closely they match pixels in the reference mask. Points are sampled from this map, clustered, and used for the next step.
- **Mask generation:** The clustered points act as sparse prompts for SAM to generate mask proposals. These are filtered based on Intersection over Union (IoU) predictions and combined into the final mask output.

Softmatcher introduces key innovations over prior architectures like Matcher. The first is the intuitive prompting & reference segmentation step, where users provide coarse scribbles as prompts rather than detailed masks. This enhances interactivity and user engagement in the segmentation process. The second innovation is a computationally efficient matching step, optimized for low-latency segmentation. By improving the efficiency of the feature-matching process compared to previous methods like the Hungarian algorithm [12], the framework enables real-time interaction, allowing users to refine or augment the resulting segmentation masks with additional annotations.

Instead of using the computationally expensive Hungarian algorithm for feature matching, Softmatcher adopts an optimal transport approach based on a quadratic cosine similarity matrix as the cost function. This allows for more efficient approximations, including entropic regularization and a geometric mean of softmaxed cosine similarity maps between reference and target features. This approach results in a quadratic complexity in image patches, which significantly enhances performance.

The framework further improves efficiency by incorporating Random Fourier Features (RFF) to approximate the softmax computation of feature similarities. The Softmatcher RFF version offers even greater scalability, outperforming traditional methods in terms of speed, particularly as image sizes increase. Benchmark evaluations on the FSS-1000 dataset, which includes diverse and unique object classes, demonstrate the effectiveness of Softmatcher and its real-time interactivity for few-shot segmentation tasks.

An interactive Visual Prompting platform integrates this improved matching pipeline, enabling users to segment object classes by simply adding scribbles to reference images. Thanks to its computational efficiency, the system allows users to interact in real-time, refining segmentations iteratively with additional annotations. A seamless web interface, built with Angular and powered by a Python API back-end and Torch Serve inference, facilitates this interaction, enabling scribbles to be converted into segmentation masks and iteratively improved. Users can also work with multiple object classes by categorizing scribbles or converting masks from previous runs into new references.

2.3 Parameter Efficient Fine-Tuning of vision foundation models

In many visual inspection scenarios, prompting-based approaches provide an effective solution by enabling training-free and memory-efficient pipelines. These methods, leveraging the capabilities of vision FMs, excel at handling a variety of segmentation tasks, especially when user-provided prompts, such as scribbles or reference masks, guide the model towards precise outcomes. The computational efficiency and generalization of these models make them particularly well-suited for real-time, interactive workflows in environments with diverse but well-aligned data.

However, there are instances where prompting alone is insufficient. In more complex, out-of-distribution use cases, the pre-trained knowledge within foundation models may fail to generalize adequately. In such scenarios, the model's inherent limitations necessitate additional fine-tuning to better adapt to new data domains. Fine-tuning allows the model to be tailored to specific tasks while overcoming the gaps in the foundational knowledge.

One particularly effective approach to address these limitations is Parameter-Efficient Fine-tuning (PEFT). PEFT modifies only a small subset of model parameters, making it a computationally viable solution for improving the model's performance on challenging tasks without sacrificing its generalization capabilities or requiring excessive resources. This technique extends the capabilities of prompting-based workflows, ensuring robust performance even in demanding, out-of-distribution environments.

In particular, we studied an ensemble of parameter-efficient modalities for fine-tuning Large Vision Models (LVMs), which have become foundational in various computer vision tasks, similar to how Large Language Models (LLMs) are used in Natural Language Processing (NLP). While fine-tuning the entire model can yield task-specific improvements, it often requires significant computational resources and may compromise the model's generalization, especially in data-scarce environments. PEFT offers an alternative by modifying a smaller subset of model parameters, preserving generalization while reducing computational demands. The proposed approach leverages a combination of PEFT methods to optimize model performance and adaptation efficiency, maintaining the core knowledge of LVMs while facilitating efficient task-specific adaptations. Notably, this method applied specifically to the model head—such as in SAM's promptable decoder—demonstrates superior segmentation performance across diverse data availability scenarios, from large datasets to low-data settings.

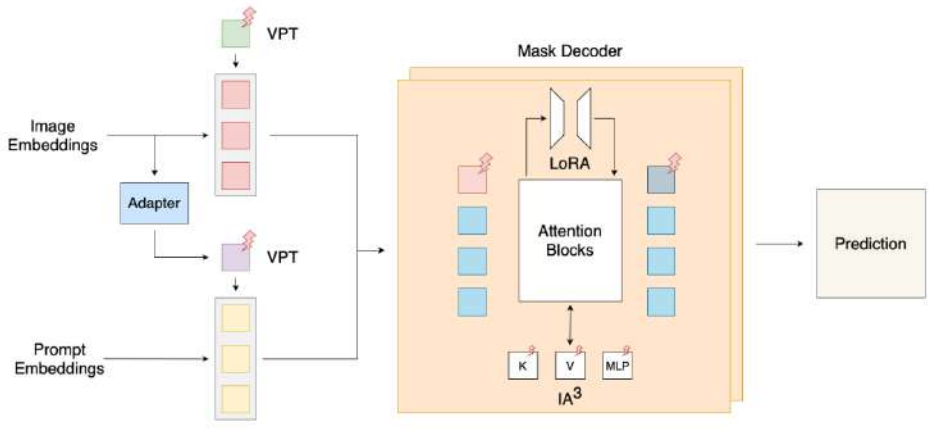

Figure 3: Ensembling of PEFT Techniques applied to Segment Anything Model.

This work provides a theoretical analysis and extensive experimentation, demonstrating the orthogonal nature of various PEFT methods. This inherent property enables the strategic combination of these

methods, resulting in a synergistic effect that amplifies their individual strengths. The proposed strategy consistently outperforms individual PEFT approaches and establishes a new state-of-the-art in the domain of parameter-efficient model adaptation. The orthogonality argument is supported by analyzing four different classes of PEFT methods: LoRA [13], IA3 [14], Prompt tuning, and adapter modules.

Our experiments center on the SAM for zero-shot semantic segmentation, with a particular focus on tuning its decoder head, which is also the core Foundation Model of the SoftMatcher Framework. The study aims to demonstrate that by carefully selecting and combining orthogonal PEFT techniques, it is possible to maximize the model's adaptation capabilities while tuning only a minimal set of parameters.

The method is grounded in a theoretical analysis that highlights the complementary nature of various PEFT approaches. By dissecting each technique, it becomes evident how these methods interact to enhance one another's strengths. Specifically:

- **LoRA** directly tunes model weights by identifying optimal low-rank directions that capture task-relevant knowledge. This method constrains optimization to a low-rank subspace, reducing degrees of freedom and facilitating faster convergence. LoRA leverages the redundancy in pre-trained weights, allowing for new capability acquisition without catastrophic forgetting of prior knowledge. However, its compute cost is higher compared to other PEFT methods due to the number of trained parameters, though it offers more stable convergence. There is a trade-off between LoRA's adaptation capacity and the degree of parameter efficiency.
- **Infused Adapter by Inhibiting and Amplifying Inner Activations (IA3)** modulates attention maps and MLPs by scaling network weights indirectly. While related to LoRA, IA3 operates at a different abstraction level, scaling the weights rather than tuning low-rank projections. This allows IA3 to softly modulate the network's behavior while preserving prior knowledge. However, IA3's scaling-based approach inherently limits its capacity to induce substantial changes in the model's behavior.
- **Prompt Tuning** has gained popularity for its extreme parameter efficiency, with performance approaching full fine-tuning. This method introduces a tunable prompt vector that acts as a learnable memory in the input space, allowing the model to store task-relevant information and utilize it for problem-solving. In addition to storing task-specific knowledge, prompt tuning has been shown to store high-level image information. Despite its compactness, prompt tuning's limited influence on the underlying model may result in lower performance ceilings, and challenges in prompt initialization can undermine its effectiveness.
- **Adapters** are lightweight, learnable modules that allow cross-flow of information between network branches. These modules inject additional capacity by enabling information transfer between internal layers, as seen in the early-last layer fusion modules of HQ-SAM [15]. Adapters enhance the model's expressive power, though they may not be sufficient alone to fully adapt the model to new tasks. They also exhibit more training instability compared to methods like LoRA.

This breakdown elucidates the interactions among various PEFT methods within the model architecture. LoRA achieves rapid convergence by directly adapting model weights, but it may overfit or underperform by focusing solely on weight adjustments. Here, prompt tuning serves as a valuable complement, introducing learnable memory registers that boost the model's capacity and support higher performance. Despite problematic initialization, prompt tuning can benefit from adapter outputs, which provide a better conditioning mechanism, as token representations are improved when initialized with patch information. IA3, by scaling network activations, harmonizes interactions between LoRA's weight modifications, the capacity added by prompts, and the information flow from adapters.

Inspired by this framework, a technique is proposed that integrates a specific implementation in SAM across all the categories outlined above. The resulting method is illustrated in Figure 3.

Table 2: Results of semantic segmentation across Kvasir, HQ-44k, Trans-10K, and CAMO datasets, highlighting the performance of various PEFT methods and their combinations. The analysis includes reporting outcomes at the 1st epoch and the 12th epoch for each dataset. The proposed method, D3AIL-VPT, outperforms both individual PEFT methods and their combined approaches.

	Kvasir-SEG		HQ-44k		Trans-10K		CAMO	
Method	ep. 1	ep. 12	ep. 1	ep. 12	ep. 1	ep. 12	ep. 1	ep. 12
D-VPT	81.32	86.65	87.16	88.49	81.10	83.88	75.37	78.64
D3-VPT	82.65	86.61	88.07	89.33	81.81	85.19	76.03	79.71
D3A-VPT	84.11	86.88	88.09	89.23	82.30	85.23	76.60	79.64
HQ-SAM	84.09	87.97	87.41	89.95	82.78	86.54	75.99	79.06
D3AL-VPT	**86.18**	**88.80**	**88.93**	**90.66**	**84.19**	**87.82**	**79.14**	**80.80**

The empirical findings support the analysis, as evidenced by performance on four out-of-distribution tasks: Fine Grained High Quality Masks (HQ-44k [15]), Camouflaged Objects Segmentation (CAMO [16]), Transparent Objects Segmentation (Trans-10K [17]), and Medical Image Segmentation (Kvasir-SEG [18]). The proposed ensemble of PEFTs consistently surpasses the performance of individual methods as well as the current state of the art.

The full ablation on the four datasets is highlighted in Table 2. The experimental protocol adopts a systematic approach, starting with Visual Prompt Tuning (Visual Prompt Tuning (VPT)) [19] as the baseline PEFT method. We progressively enhance the model by incorporating additional techniques: first, IA3 is added, followed by the integration of adapter modules, and finally, LoRA. This step-by-step process allows us to isolate and measure the specific impact of each method on overall performance.

More in detail, we begin with Deep VPT (D-VPT), which uses 12.8k parameters and demonstrates notable performance improvements. The addition of IA3 to D-VPT, resulting in D3-VPT, continues this trend of enhanced performance. We further introduce an MLP adapter layer with image aggregation enabling image features to condition the visual prompts. This method, termed D3A-VPT, achieves further gains. The final enhancement incorporates LoRA into the decoder, culminating in D3AL-VPT, which delivers the best results.

In this work, we proposed a novel fine-tuning strategy that combines multiple PEFT methods — LoRA, IA3, adapter modules, and VPT. Our experiments across datasets showed that this ensemble significantly outperforms individual methods and state-of-the-art approaches while remaining computationally efficient. The results demonstrate that combining PEFT techniques enhances model adaptability, especially in out-of-distribution scenarios, paving the way for more efficient and scalable fine-tuning strategies in vision models.

3 Conclusion

This deliverable has presented three complementary approaches to knowledge recycling and efficient model adaptation, each addressing different aspects of sustainable machine learning:

- The TSAK framework demonstrates how semantic-aware knowledge distillation can create efficient, deployable models for wearable devices while maintaining high performance. By achieving up to 96.6% reduction in computational power and 79% smaller model size while improving F1 scores by up to 10%, TSAK shows the potential for significant efficiency gains without compromising functionality.

- The Softmatcher framework introduces a novel approach to training-free visual prompting that enables real-time, interactive segmentation workflows. By combining probabilistic feature matching with efficient approximation techniques, it achieves scalable performance while maintaining the flexibility needed for diverse application scenarios.
- The parameter-efficient fine-tuning methodology showcases how combining multiple PEFT techniques can optimize model adaptation while minimizing resource requirements. The proposed D3AL-VPT approach consistently outperforms existing methods across multiple datasets, demonstrating the benefits of strategically combining orthogonal PEFT techniques.

Together, these contributions advance the state-of-the-art in sustainable machine learning by providing practical solutions for efficient knowledge transfer and model adaptation. The methods presented here enable practitioners to leverage the intrinsic capabilities foundation models while minimizing computational overhead, aligning with SustainML's goal of fostering energy-efficient ML applications throughout their lifecycle.

References

[1] Hymalai Bello et al. "TSAK: Two-Stage Semantic-Aware Knowledge Distillation for Efficient Wearable Modality and Model Optimization in Manufacturing Lines". In: *arXiv preprint arXiv:2408.14146* (2024).

[2] Oscar D Lara and Miguel A Labrador. "A survey on human activity recognition using wearable sensors". In: *IEEE communications surveys & tutorials* 15.3 (2012), pp. 1192–1209.

[3] Shibo Zhang et al. "Deep learning in human activity recognition with wearable sensors: A review on advances". In: *Sensors* 22.4 (2022), p. 1476.

[4] Jianping Gou et al. "Knowledge distillation: A survey". In: *International Journal of Computer Vision* 129.6 (2021), pp. 1789–1819.

[5] Yexu Zhou et al. "Tinyhar: A lightweight deep learning model designed for human activity recognition". In: *Proceedings of the 2022 ACM International Symposium on Wearable Computers*. 2022, pp. 89–93.

[6] Naoya Yoshimura et al. "Openpack: A large-scale dataset for recognizing packaging works in iot-enabled logistic environments". In: *2024 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE. 2024, pp. 90–97.

[7] Alexander Kirillov et al. *Segment Anything*. 2023. arXiv: 2304.02643 [cs.CV].

[8] Alec Radford et al. *Learning Transferable Visual Models From Natural Language Supervision*. 2021. arXiv: 2103.00020 [cs.CV].

[9] Mathilde Caron et al. *Emerging Properties in Self-Supervised Vision Transformers*. 2021. arXiv: 2104.14294 [cs.CV]. URL: https://arxiv.org/abs/2104.14294.

[10] Maxime Oquab et al. "Dinov2: Learning robust visual features without supervision". In: *arXiv preprint arXiv:2304.07193* (2023).

[11] Yang Liu et al. *Matcher: Segment Anything with One Shot Using All-Purpose Feature Matching*. 2024. arXiv: 2305.13310 [cs.CV]. URL: https://arxiv.org/abs/2305.13310.

[12] Harold W. Kuhn. "The Hungarian method for the assignment problem". In: *Naval Research Logistics (NRL)* 52 (1955).

[13] Edward J. Hu et al. "LoRA: Low-Rank Adaptation of Large Language Models". In: *CoRR* abs/2106.09685 (2021). arXiv: 2106.09685. URL: https://arxiv.org/abs/2106.09685.

[14] Haokun Liu et al. *Few-Shot Parameter-Efficient Fine-Tuning is Better and Cheaper than In-Context Learning*. 2022. arXiv: 2205.05638 [cs.LG].

[15] Lei Ke et al. *Segment Anything in High Quality*. 2023. arXiv: 2306.01567 [cs.CV].

[16] Trung-Nghia Le et al. "Camouflaged instance segmentation in-the-wild: Dataset, method, and benchmark suite". In: *IEEE Transactions on Image Processing* 31 (2021), pp. 287–300.

[17] Jun Hao Liew et al. "Deep Interactive Thin Object Selection". In: *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2021, pp. 305–314. DOI: 10.1109/WACV48630.2021.00035.

[18] Debesh Jha et al. *Kvasir-SEG: A Segmented Polyp Dataset*. 2019. arXiv: 1911.07069 [eess.IV].

[19] Menglin Jia et al. *Visual Prompt Tuning*. 2022. arXiv: 2203.12119 [cs.CV].