The Fortieth AAAI Conference on Artificial Intelligence (AAAI-26)

Learning System Expansion with Efficient Heterogeneity-aware Knowledge Transfer

Gaole Dai[1], **Huatao Xu**[2], **Yifan Yang**[3], **Rui Tan**[*1], **Mo Li**[*2]
[1]Nanyang Technological University
[2]Hong Kong University of Science and Technology
[3]Microsoft Research Asia
{gaole001, tanrui}@ntu.edu.sg, {huatao, lim}@ust.hk, yifanyang@microsoft.com

Abstract

Modern AI services must continually adapt to newly joined domains, yet delivering high-quality customized models is hampered by label sparsity, domain shifts, and tight budgets. We formulate this challenge as the learning system expansion problem and introduce HaT, an efficient heterogeneity-aware knowledge-transfer framework. HaT first selects a small set of high-quality source models with minimal overhead, and then fuses their imperfect predictions through a sample-wise attention mixer. Later, it adaptively distills the fused knowledge into target models via a knowledge dictionary. Extensive experiments on different tasks and modalities show that HaT outperforms state-of-the-art baselines by up to 16.5% accuracy, and saves 31.1% training time and up to 93.0% traffic.

Code — https://github.com/MaginaDai/HaT-Public

Introduction

Deployment of customized learning models presents a critical dilemma. While building specialized models for each user, device, or environment (*domain*) on the edge can yield fine-grained performance and preserve privacy (Lu et al. 2024; Kong et al. 2023), producing these per-domain models is expensive. Each customized model demands substantial labeling and training efforts, yet in practice, many domains have only scarce labeled data due to prohibitive labeling costs (Gao et al. 2024; Dai et al. 2024). For example, a healthcare system might require customized models for different hospitals or even individuals (Ouyang et al. 2023), yet obtaining sufficient labeled data from each hospital/device remains costly and time-consuming. Moreover, modern learning systems typically serve numerous domains, in which cases the labeling and retraining overhead becomes unsustainable as the system grows.

While one might consider transferring existing models to new domains (Phan et al. 2024; Lu and Sun 2024), significant challenges arise from data and device heterogeneity. Models trained on a particular data distribution may fail to generalize to another, suffering from performance drops or inapplicability to unseen categories. Additionally, resource constraints of different devices, such as memory or computational power, further complicate direct model reuse (Li et al. 2024b). For instance, a smartphone and a smartwatch may both need to run the human activity recognition but have distinct available resources, making it difficult to transfer models directly between them. Thus, a more efficient and systematic strategy is needed to expand learning systems to more domains without incurring massive cost.

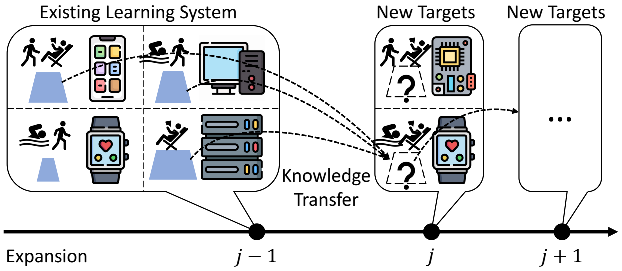


Figure 1: Expanding learning systems is challenging due to label scarcity and large heterogeneity.

We define this challenge as **learning system expansion**, illustrated in Figure 1. In this context, source domains, such as different users, devices, or datasets, maintain heterogeneous models to process local data. Target domains, on the other hand, have limited labeled data and abundant unlabeled data due to the high costs associated with labeling. The data in these target domains are non-IID with potential shifts in label space. For example, in in-home patient monitoring systems, customized models are deployed to accommodate the unique health conditions and sensor characteristics of each individual. As more users adopt such systems for proactive healthcare, the learning system must adapt to these new users without relying on extensive labeled data or imposing constraints on hardware. This leads to a critical question: *How can we effectively and efficiently expand learning systems to accommodate new target domains*?

Existing approaches struggle to handle this question. To handle data heterogeneity, domain adaptation is widely studied to enhance model robustness by aligning feature distributions across domains (Wilson, Doppa, and Cook 2021; He et al. 2023; Qu et al. 2024). Nevertheless, those works over-

---

[*] Corresponding Authors

look device heterogeneity, making it difficult to adopt existing models across diverse hardware environments. Knowledge distillation addresses device-side constraints by transferring knowledge from teacher models to student models (Hinton, Vinyals, and Dean 2015; Gou et al. 2021; Borup, Phoo, and Hariharan 2023; Peng et al. 2024). Yet existing methods assume that the teacher models are fully reliable and do not account for the impact of data heterogeneity. The non-IID data across domains makes source models less accurate on target domains. While personalized federated learning customizes models for each domain, its high training and communication overheads are unsuitable for the ever-growing learning system. Therefore, there is a significant gap in addressing the expansion problem.

To bridge this gap, we propose the Heterogeneity-aware Knowledge Transfer (HaT) framework with three key designs: 1) High-Quality Source Model Selection: HaT filters out low-quality source models using simple statistical features. The remaining models are further evaluated based on their performance on target domain data. This ensures that only reliable models contribute to the knowledge transfer. 2) Adaptive Knowledge Fusion and Injection: An attention-based mixer is trained to assign sample-wise weights to the predictions of each source model based on their representations similarity. A knowledge dictionary selectively stores the fused predictions, which are later injected into the target model. The transfer speed is dynamically adjusted based on the knowledge quality, ensuring that only useful knowledge is passed to the target model. 3) Efficient Communication and Joint Training. HaT encapsulates the model selection process within a communication protocol that only transmits models with high potential to the target domain, minimizing communication overhead. In addition, a low-cost joint training scheme is implemented to simultaneously update the target model and the mixer, ensuring minimal computational overhead while maintaining system performance.

Extensive experiments are conducted across multiple modalities and tasks to show the generalizability of HaT. HaT achieves up to 16.5% higher accuracy, which also reduces communication traffic by up to 93.0% and train time per epoch by 31.1%. The key contributions are as follows:

1. We address a practical learning system expansion problem characterized by label scarcity and both data and device heterogeneities.
2. We propose a general framework, HaT, for learning system expansion, which selects, fuses, and injects existing knowledge to deliver high-quality customized models with practical system overhead.
3. We evaluate the framework across various tasks, modalities, and architectures, demonstrating superior performance compared with baselines.

Related Works

Transfer Learning. Transfer learning explores methods to apply source models for new targets, addressing data or task heterogeneities (Pan and Yang 2009; Tan et al. 2018). In particular, domain adaptation has been extensively studied to align feature distributions between domains (Zhu et al. 2020; Wilson, Doppa, and Cook 2021; He et al. 2023; Qu et al. 2024). However, these approaches typically require access to both source and target domain data, which may not be feasible. In contrast, test-time adaptation techniques adapt models using only test data, enabling continual learning (Gong et al. 2024; Karmanov et al. 2024). Additionally, multi-source transfer learning methods aim to select source models with better generalizability (Tong et al. 2021; Agostinelli et al. 2022). Despite these advancements, most transfer learning approaches do not address device heterogeneity, which is a critical factor in learning system expansion. This limitation hinders the direct application of source models to target domains, where device-specific constraints must also be considered.

Knowledge Distillation. In knowledge distillation, a student model is trained using the knowledge from one or more teacher models, such as their predicted pseudo labels or intermediate features (Hinton, Vinyals, and Dean 2015; Liu, Zhang, and Wang 2020; Vemulapalli et al. 2024; Peng et al. 2024). Specifically, multi-teacher distillation approaches (Borup, Phoo, and Hariharan 2023; Liu, Zhang, and Wang 2020; Zhang, Chen, and Wang 2022) aggregate the knowledge of multiple teachers by assigning weights, aiming to provide the student model with more accurate and comprehensive knowledge. Most knowledge distillation studies assume high-quality teacher models are readily available (Hinton, Vinyals, and Dean 2015; Liu, Zhang, and Wang 2020; Zhang, Chen, and Wang 2022; Borup, Phoo, and Hariharan 2023). However, in the learning system expansion problem, the knowledge from source domain models may not directly transfer to the target domain due to data heterogeneity, leading to suboptimal performance.

Federated Learning. Federated learning (FL) focuses on collaboratively training a shared global model across decentralized clients (Zhang et al. 2021; Li et al. 2020; Yao et al. 2022; Criado et al. 2022). While personalized FL techniques (Tan et al. 2022; Collins et al. 2021) address learning under heterogeneity, they assume that all domains participate actively in training and focus on closed-world settings. In contrast, learning system expansion targets an open-world scenario, where new domains continually join. It focuses on the customized model construction for the new domains, rather than retraining among all domains.

Model Customization. Model customization has been extensively studied to meet specific computational and performance requirements (Wen et al. 2023; Li et al. 2024b). Some works explore pre-deployment or post-deployment model generation techniques (Cai et al. 2020; Wen et al. 2023) to search optimal architecture in terms of latency and accuracy. In contrast, HaT emphasizes the knowledge transfer process from the selected source models to any target models that satisfy the customized needs of target domains.

Learning Systems Expansion

Problem Formulation

We define the **Multi-Round System Expansion (MRSE)** problem to address the ever-growing nature of learning systems. At round j, there are $N^S(j)$ existing source domains, denoted as $\mathcal{D}^S(j) = \{\mathcal{D}_i^S(j)\}_{i=1}^{N^S(j)}$, and $N^T(j)$ target domains, $\mathcal{D}^T(j) = \{\mathcal{D}_i^T(j)\}_{i=1}^{N^T(j)}$. Each target domain in $\mathcal{D}^T(j)$ requires high-quality, customized models to meet its unique requirements. Once the models for target domains $\mathcal{D}^T(j)$ are curated, these domains become source domains in the subsequent round: $\mathcal{D}^S(j+1) = \mathcal{D}^S(j) \cup \mathcal{D}^T(j)$. The knowledge in $\mathcal{D}^S(j+1)$ is then leveraged to curate models for target domains in $\mathcal{D}^T(j+1)$. The primary objectives are: 1) maximize performance of the curated models on target domains; 2) minimize curation overhead, which includes communication and computation costs.

To better understand the process, the MRSE problem can be decomposed into individual **One-Time System Expansion (OTSE)** problems. For a specific target domain $\mathcal{D}_i^T(j) = \{X_i^T, Y_i^T, \zeta_i^T\}$, the goal is to curate a model based on the knowledge from source domains $\mathcal{D}^S(j) = \{X_i^S, Y_i^S, NN_i^S, \zeta_i^S\}$. However, source and target domains exhibit distributional differences between X_i^S and X_i^T, which hinder the direct applicability of the source model $NN_i^S = \{f_i^S, g_i^S\}$. The f_i^S and g_i^S are the encoder and the classifier. The label sets of source domains Y_i^S and target domain Y_i^T may not fully overlap, introducing additional complexity during expansion. Moreover, target domains impose constraints ζ_i^T, including memory usage and inference speed requirements, which must be considered during model curation. Besides, due to the high cost of labeling, only few ($\gamma\%$) data in $\mathcal{D}_i^T(j)$ is labeled, which is a general assumption to handle potential label space difference.

Connection with Real-life Scenarios

The learning system expansion problem is critical in real-world applications where the demand for customized models increases over time. For instance, to provide sleep monitoring or activity recognition service (Ouyang et al. 2023; Xu et al. 2021), models must adapt to individual health or motion conditions and sensor characteristics. As more users adopt these systems for proactive healthcare or interactions, the learning system must efficiently accommodate new users without relying on extensive labeled data or imposing significant hardware constraints.

A similar challenge arises in large-scale urban surveillance (Yuan et al. 2024), where new cameras are continuously deployed across diverse environments. Devices from different manufacturers may capture images with varying lighting conditions, viewing angles, and backgrounds, creating a need for model customization. This diversity in sensor characteristics and environmental conditions makes adapting models to new cameras both necessary and challenging.

HaT: Efficient Heterogeneity-aware Knowledge Transfer

Framework Overview

To address the learning system expansion problem, HaT, as presented in Figure 2, first select high-quality source models at a low cost. Later, the Sample-wise Knowledge Fusion is performed to aggregate the conflicting knowledge. Subsequently, the target model is trained with the Adaptive Knowledge Injection based on a low-cost training scheme. We further introduce the details of providing a customized model for one target $\mathcal{D}_t$ (instead of $\mathcal{D}_t^T$ for clarity), the processes of which are scalable for any number of targets.

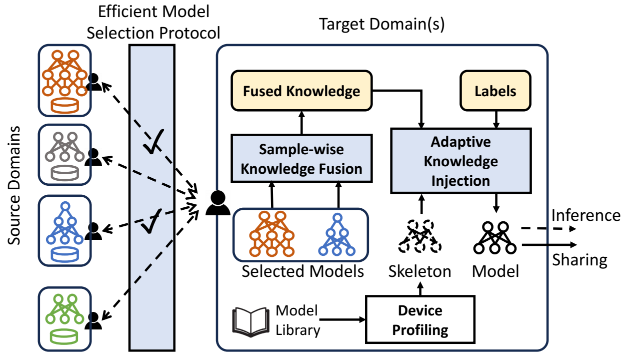


Figure 2: Framework overview of HaT.

Efficient Model Selection Protocol

Model selection is crucial for preventing negative transfer (Zhang et al. 2022), yet in an expanding learning system it quickly becomes prohibitively expensive as shown in Figure 3. The result is emulated with the statistics in (Warden 2018) (see Appendix A). Existing methods (Borup, Phoo, and Hariharan 2023; Li et al. 2019, 2024a) must transmit every source model to the target and benchmark it locally, incurring heavy communication and inference costs. The efficient model selection protocol in HaT sidesteps this bottleneck by performing a lightweight, feature-based pre-screening that filters out weak candidates before any model is transmitted, drastically reducing both traffic and computation.

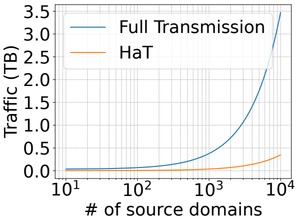


Figure 3: The traffic of the model selection for each target domain quickly becomes unaffordable as the learning system scales.

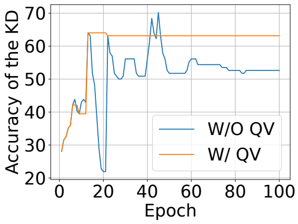


Figure 4: Quality verification (QV) stores only high-quality fused predictions in the KD, providing a stable training signal.

Feature-based Coarse Selection. A vector of lightweight statistical features l, e.g., [mean(X), var(X), skewness(X)], are first computed directly from raw data in each source and the target domain. These hand-crafted features capture coarse domain characteristics without requiring any learned model. The target device ranks all sources by feature-space similarity and requests model weights only from the top $\eta\%$ of candidates, thereby transmitting and evaluating a small, high-quality subset instead of the entire pool.

Centroids-Accuracy Joint Selection. To further select the high-quality models within the received model pool, source domain models are ranked by the product of each model's labeled set accuracy Acc_i and the distribution similarity on the unlabeled set. The top N_p models are then kept.

We approximate each domain's data distribution with class centroids, computed as the mean encoder output of all samples belonging to each class:

$$c_{t,m} = \frac{1}{K}\sum_{k=1}^{K} h_i(k), k \in \{k | max(g_i^S(h_i(k))) = m\} \quad (1)$$

where $h_i(k) = f_i^S(x(k))$ represents the features extracted by the encoder f_i^S of the i-th domain from the k-th data sample. To estimate pseudo-labels, we apply the classifier $g_i^S(h_i(k))$ followed by a $max(\cdot)$ operation. Then, the centroids of $\mathcal{D}_t$ and $\mathcal{D}_i^S$ are compared to estimate the data distribution similarity $s^{fine}(t,i)$: $s^{fine}(t,i) = \sum_{m=1}^{M} \text{sim}(c_{t,m}, c_{i,m}^S)/M$, where M is the number of overlapping classes between $\mathcal{D}_t$ and $\mathcal{D}_i^S$. A higher value of $s^{fine}(t,i)$ indicates a greater similarity in data distribution, implying the source model will likely perform better when applied to the target domain. To further enhance centroid accuracy, we introduce an entropy-based filtering step that excludes low-certainty samples. Specifically, the entropy $e(k)$ of the logits output $g_i^S(h_i(k))$ is calculated, and the $\omega = 75\%$ of features with the lowest entropy (i.e., the highest confidence) are selected for centroid extraction.

Sample-wise Knowledge Fusion from Multiple Imperfect Source Models

During learning system expansion, even the top-ranked source models can still mispredict on the target domain, unlike the near-oracle teachers assumed in standard knowledge distillation problem (Vemulapalli et al. 2024; Borup, Phoo, and Hariharan 2023). Adding to the difficulty, these source models differ in architecture, size, and output dimensionality, making standard fusion schemes inapplicable. We therefore merge their outputs using an attention-based mixer that is compatible with heterogeneous architectures.

Attention-based Mixer. The mixer first project the feature $h_t(k)$ from the target encoder through a linear layer L^{query} to the query vector $q(k)$. Similarly, the features $h_i(k)$, extracted by the selected models, are projected through the respective linear layers L_i^{key} to obtain the key vectors $key_i(k)$: $q(k) = L^{\text{query}}(h_t(k)), key_i(k) = L_i^{\text{key}}(h_i(k))$.

The mixer utilizes multiple linear layers L_i^{key} of tailored input dimension to accommodate the heterogeneous source architectures that extracts features of varying dimension. The output size of L_i^{key} is standardized to a common dimension. The similarities between the query and the keys are then calculated to obtain the attention score $w_i(k)$ for the i-th model on the data sample $x(k)$: $w_i(k) = \text{SoftMax}(q(k) \cdot key_i(k)), i = 1, 2, \ldots, N_p$.

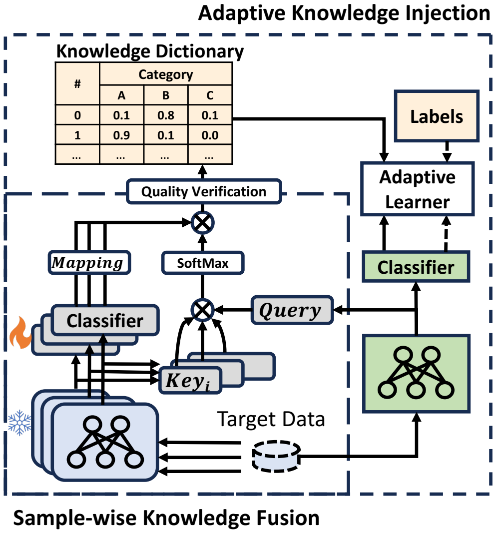


Figure 5: Details of the Sample-wise Knowledge Fusion and the Adaptive Knowledge Injection process.

The attention score $w_i(k)$ measures the feature similarity between the target and the selected models, which is used to aggregate the predictions from the selected classifiers $g_i(k)$:

$$p^{\text{mix}}(k) = \sum_i w_i(k) \cdot Map[g_i^S(h_i(k))], \quad (2)$$

where $Map[\cdot]$ projects the label spaces of different domains to a unified label space that includes all categories. If the features extracted by the target model and the i-th model are highly similar, a higher weight $w_i(k)$ is assigned to the prediction of the i-th model. This is because the i-th classifier is likely to be more accurate on data from a distribution that closely resembles the data it is trained with. Intuitively, the score reflects how well a source model processes a target sample. In cases of negative transfer, the mixer naturally assigns lower weights to such models to reduce their influence.

Algorithm 1: Low-cost Joint Training of HaT

```
**Input:** Target labeled and unlabeled data $X^{\{l,u\}}$, labels $Y^l$, selected models $\{f_i^S\}_{i=1}^{N_p}, \{g_i^S\}_{i=1}^{N_p}$,
**Output:** Target encoder and classifier $f_t, g_t$
Initialize $g_t, f_t, Mixer, KD$
# Encode target data with source encoders for later reuse
$\{h_i^{\{l,u\}}\}$ = ENCODING($X^{\{l,u\}}$, $\{f_i^S\}_{i=1}^{N_p}$)
for epoch in Epochs do
    # Obtain target features (sampled $X^u$ to reduce cost)
    $h_t^l, h_t^u = f_t(X^l), f_t(Sampling(X^u))$
    # Update Mixer (Eq. 2) to learn attention weights
    $Mixer$ = MIXER_UPDATE($\{h_i^l\}, h_t^l, Mixer, Y^l, \{g_i^S\}$)
    $f_t, g_t$ = MODEL_UPDATE($f_t, g_t, h_t^{\{l,u\}}, Y^l, KD$)
    if Quality_Verification($Mixer$) then
        # Update KD only if Mixer quality improves
        $KD$=DICT_UPDATE($KD, Mixer, \{h_i^u\}, h_t^u$)
    end if
end for
```

Cost-effective Adaptation To further improve the accuracy of the fused predictions, we might adapt all N_p selected source models, which, however, would be computationally expensive. Instead, the classifiers are trained jointly with the mixer, while their encoders remain frozen. This approach reduces the computational burden, as classifiers are typically lightweight (He et al. 2016). Additionally, freezing the encoders accelerates the knowledge aggregation process. By pre-computing and storing features for all target domain data using the frozen encoders, the mixer just fetch need features from memory and eliminates the need to repeatedly execute the forward pass of the selected encoders, which greatly reducing the overall computation time.

Adaptive Knowledge Injection with Verified Knowledge Dictionary

The target model selected given the training/inference-time constraints of the target domain can be trained with:

$$L_{\text{ada}} = L^{\text{label}} + \alpha\, L^{\text{distill}}(g_t \circ f_t(x(k)), p^{\text{mix}}(k)), \quad (3)$$

where L^{label} is the cross-entropy loss on the limited labeled data and L^{distill} is the distillation loss learning from the fused prediction. However, Figure 4shows that the accuracy of the mixer's fused predictions on the unlabeled set (blue curve) oscillates across epochs. Even the accuracy remains high in some cases, the aggregated results $p^{\text{mix}}(k)$ can change markedly during training, sometimes flipping a sample's pseudo-label from class A to class C. These fluctuations send conflicting gradient signals to the model, making supervision from $p^{\text{mix}}(k)$ unstable and slowing convergence.

Knowledge Dictionary with Quality Verification. To stabilize training, we introduce a knowledge dictionary (KD) guarded by a lightweight quality verification step. Specifically, we randomly select 20% of the labeled data from each target domain as a probing set. After each update of the mixer, the KD is refreshed with the latest fused predictions only if the mixer achieves improved accuracy on the probing set. As shown in Figure 4 using the HARBox dataset, the quality verification, which utilizes even a minimal probing set (as small as ten samples), reduces prediction fluctuations and maintains consistently high-quality pseudo-labels throughout the training process. In KD, each entry stores a soft label, allowing the target model to capture the confidence levels of the mixer. During subsequent epochs the target model is supervised on the unlabeled data by the KD. Accordingly, Equation 3 is replaced by:

$$L_{\text{ada}} = L^{\text{label}} + \alpha\, L^{\text{distill}}(g_t \circ f_t(x(k)), KD(k)), \quad (4)$$

Adaptive Learner. Given the varying quality of the fused predictions in the knowledge dictionary, an adaptive learner is employed to adjust the weight α of the distillation loss: $\alpha = m(Acc_{\text{train}} - b)$, where Acc_{train} represents the accuracy of the attention-based mixer on the training data. The m and b are predetermined hyperparameters. The m controls the scaling factor of the weight α, while b serves as a threshold to prevent the target model from learning from fused predictions of low quality. The weight α increases when the fused prediction accuracy is high, allowing the model to learn more effectively from reliable predictions.

Low-cost Joint Training To enable a cost-effective training process, a joint training scheme is developed as shown in Algorithm 1. The labeled and unlabeled data, $X^{\{l,u\}}$, are encoded by the frozen source encoders $\{f_i^S\}_{i=1}^{N_p}$ to high-level features, which are kept for later training. In each epoch, the f_t encodes $X^{\{l,u\}}$, generating representations h_t^l and h_t^u. To manage the computational overhead of processing a large volume of unlabeled data X^u, only a subset of X^u is randomly sampled in each epoch, with the sample size kept proportional to the size of the labeled data. This strategy ensures that the entire set of unlabeled data is progressively utilized over multiple iterations, thereby reducing training time and memory usage of each epoch without compromising model performance. The KD is updated only when the mixer's quality improves, minimizing the cost of generating pseudo-labels for all unlabeled data.

The cross-entropy loss is computed using p^{mix} and labels Y^l and minimized by one optimizer to train the mixer and the unfrozen classifiers (illustrated in gray in Figure 5). Equation (4) is minimized by a separate optimizer to train the target model (illustrated in green in Figure 5). After the model training, only g_t and f_t are stored for the inference.

Evaluations

Experiment Setting

Datasets. HaT is evaluated on five datasets, HARBox (Ouyang et al. 2021), ImageNet-R (Hendrycks et al. 2021), NinaPro (Pizzolato et al. 2017), Alzheimer's Disease (AD) (Ouyang et al. 2023), and Speech Command (Warden 2018), that span six modalities, four tasks, and different scales. For each dataset, different model architectures, e.g., convolutional neural networks and autoregressive models, are included as model libraries. (See Appendix A.)

Baselines. The five most relevant baselines from knowledge distillation, model aggregation, and domain adaptation are implemented for comparison, including DistillWeighted (Borup, Phoo, and Hariharan 2023), DistillNearest (Borup, Phoo, and Hariharan 2023), LEAD (Qu et al. 2024), and MEHLSoup (Li et al. 2024a). We also propose a baseline called AccDistill, which distills the knowledge from ensemble models selected from source domains. (See Appendix A.) Federated learning approaches are excluded from the comparison because they tackle a different problem setting than the learning-system expansion scenario (see Problem Formulation). All the baselines and HaT use the same amount of labeled data and information during training.

Real-world Testbed. We deploy our prototype on a two-tier setup: (i) a backend server that stores all source domain models and data, and (ii) an edge node, an NVIDIA Jetson Xavier, that hosts the target-domain data and executes training. The model training overhead, including storage, time and memory usage, is measured on the edge devices, which are closely correlated with energy consumption. Communication cost is recorded by capturing the cumulative network traffic exchanged between the server and the edge node.

Implementation Details. To demonstrate HaT's versatility, HaT is trained with full-parameter updates on all datasets except Speech Command, where we apply LoRA

Methods	**HARBox**	**ImageNet-R**	**NinaPro**	**AD**	**Speech Command**
LEAD	51.46	48.17	44.94	36.46	72.03
MEHLSoup	62.98	47.57	43.32	31.04	72.25
AccDistill	73.40	57.56	35.65	52.71	26.49
DistillNearest	74.95	57.64	40.75	58.12	20.30
DistillWeighted	75.42	57.66	41.02	56.46	22.18
HaT	**79.27**	**59.30**	**45.12**	**63.96**	**74.29**

Table 1: Accuracy comparison in the MRSE setting.

Method	**Multi-Rounds**	**One-Round**
DistillWeighted	75.42	67.51
HaT	**79.27**	**70.96**

Table 2: Accuracy under two expansion settings.

Method	**Traffic (MB)**	**Storage (MB)**	**Time (s)**
LEAD	508	158.0	8.46
MEHLSoup	508	451.7	10.91
AccDistill	1 786	474.0	31.30
DistillNearest	1 786	474.0	30.91
DistillWeighted	1 786	474.0	31.15
HaT	1 279	162.2	5.83

Table 3: System overhead comparison on ImageNet-R. For fairness, we standardized the batch size to 128 and used the same target model (ResNet-34).

fine-tuning. On the Speech Command dataset, training runs for 20 epochs, whereas on the other datasets training lasts 200 epochs. The learning rates of target models and mixer are searched among $\{$5e-4, 1e-3, 5e-3, 1e-2$\}$ for different datasets. The scaling ratio m and the bias b are determined using a grid search within the ranges $[1.0, 4.0]$ and $[0, 0.5]$, with step sizes of 0.5 and 0.1, respectively. The N_p is set to three. A sensitivity analysis is provided in Appendix B.

Performance in Multi-Round System Expansion

To assess HaT under MRSE, we partition each dataset's domains into successive groups of varying sizes, emulating different expansion speeds, and report average results for robustness. This staged release simulates an incremental learning-system expansion, where new domains arrive round by round (see Appendix A.)

HaT outperforms across every expansion scale and speed. Table 1 presents the performance across different rounds of expansion, with more detailed results provided in Appendix C. Compared to baselines that either leverage limited knowledge from source domains (LEAD and MEHLSoup) or transfer knowledge in a static manner (AccDistill, DistillNearest, and DistillWeighted), HaT delivers more effective customized models with higher accuracy for target domains under different system expansion speed. The reason is that HaT incrementally folds better models from each round into its source pool with better knowledge selection, fusion, and injection, it propagates higher-quality knowledge forward, yielding steady accuracy gains without error accumulation in subsequent rounds.

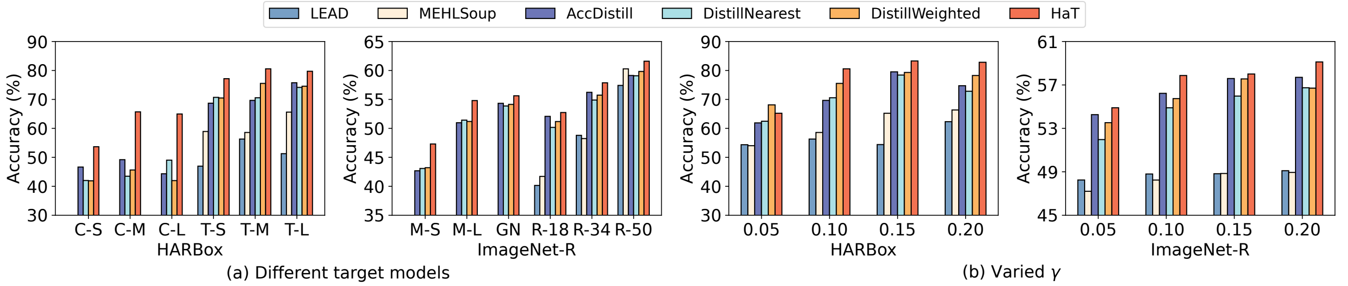
Continuous knowledge sharing boosts accuracy. We contrast MRSE with a *One-Round* variant on the HARBox dataset, where all new users are served in a single batch. Table 2 shows that accuracy is consistently higher under MRSE: earlier-round models act as additional knowledge source for later domains, boosting performance whenever successive domains share similar distributions.

HaT expands learning systems with superior efficiency. In Table 3, HaT cuts communication traffic despite the inevitable growth that comes with more source domains across all datasets. Note that the communication traffic of LEAP and MEHLSoup is not directly comparable to HaT because they can handle one or a few architecture-matched sources, a restriction that also limits their accuracy. Relative to the strongest multi-source baselines (AccDistill, DistillNearest, and DistillWeighted), HaT cuts selection-phase traffic by 28.4% on ImageNet-R, 41.1% on NinaPro, 31.6% on AD, 37.6% on HARBox, and 93.0% on Speech Command, with larger traffic savings observed on datasets that contain more source domains. For storage, LEAD is small because it only leverage a single model, yet HaT remains comparable even while leveraging multiple source models by storing only lightweight feature embeddings plus a classifier head per source. In addition, HaT records the shortest per-epoch runtime and converges at least 1.4× faster than the strongest distillation baselines. It converges in 79.4 epochs on average, versus 107.1, 115.8, 130.6 epochs for DistillNearest, DistillWeighted, and AccDistill. Collectively, these results confirm that HaT expands learning systems efficiently.

Robustness of HaT

We further evaluate HaT in the OTSE setting, focusing on its robustness to (i) different target-model architectures and (ii) severe label sparsity. We report average results across multiple randomly partitioned domains. Pre-processing details and comprehensive results across other datasets are provided in Appendix A and Appendix D.

HaT is robust across tasks, modalities, architectures, and label scarcity. Figure 6(a) shows that whereas domain adaptation and model merging methods can only leverage source models that share the same architecture, HaT fuses knowledge from heterogeneous sources and delivers the highest accuracy for most architectures on different tasks.

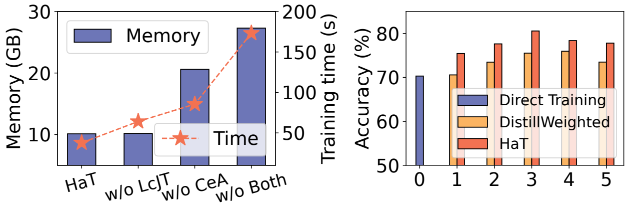


(a) Different target models (b) Varied γ

Figure 6: Performance comparisons with (a) varied architectures and (b) varied γ. In (a), C-(S, M, L), T-(S, M, L), M-(S, L), GN, R-(18, 34, 50) represents CPC-(s, M, L), TPN-(S, M, L), MobileNet-(S, L), GoogleNet, and ResNet-(18, 34, 50), respectively.

Design	HARBox	ImageNet-R	AD
HaT	**80.57**	**57.87**	**67.33**
w/o FbCS	73.86	56.86	63.83
w/o CAJS	79.92	56.45	67.17
w/o SwKF	69.05	56.14	60.17
w/o AKI	66.73	53.29	48.50
w/o QV	76.66	56.52	54.33

Table 4: Ablation study in OTSE. SwKF, AKI, QV refer to Sample-wise Knowledge Fusion, Adaptive Knowledge Injection, and Quality Verification.



(a) Design Choice (b) # of Selected Models

Figure 7: (a) The impact of the Cost-effective Adaptation (CeA) and Low-cost Joint Training (LcJT). The batch size is set to 16. (b) The impact of the number of selected models.

Figure 6(b) varies the portion of the labeled data γ. HaT either matches or exceeds the best baseline. In the few cases where a baseline edges ahead, we attribute the gap to the use of a fixed threshold b in the adaptive learner, which may yield a sub-optimal weight α when γ changes. Incorporating a dynamic threshold is left for future work.

Ablation Study

Design Effectiveness. As shown in Table 4, the Feature-based Coarse Selection and Centroids-Accuracy Joint Selection enhance performance by leveraging the statistical and high-level features that accurately reflect the domain similarity and the source models effectiveness. The combination of both selections demonstrates stronger generalizability across datasets. When labels are unavailable, HaT can leverages centroids to select, which slightly reduces the accuracy (by 1.3% on HARBox). Sample-wise Knowledge Fusion achieves an 11.5% accuracy improvement on HARBox, since the sample-wise weights learned by the mixer could more effectively combine predictions from source models. Adaptive Knowledge Injection boosts accuracy by dynamically scaling the distillation loss and selectively storing fused predictions. Specifically, quality verification contributes to an increase in accuracy of 6. 1% on average, filtering noisy pseudo-labels and stabilizing the training signal.

Optimizing the Training Overhead. Figure 7(a) presents the training overhead on the ImageNet-R dataset. The Cost-effective Adaptation, which partially tune the selected models during training, lead to a 2.0× reduction in memory usage and a 2.3× reduction in training time due to fewer parameters being optimized. Similarly, incorporating Low-cost Joint Training reduces the per-epoch training time from 64.1s to 37.8s. Overall, HaT achieves significant reductions in both training time (4.6×) and memory usage (2.7×), indicating a more energy-efficient training process.

More sources aren't always better. Figure 7(b) shows that as N_p increases, the accuracy first increases, which justifies the usage of multiple source models in HaT. When N_p continues to increase, the accuracy drops. It might be due to noisy or low-quality knowledge from additional sources or the limited capacity of the lightweight mixer when too many sources are combined. Across the full range of N_p, HaT remains superior to the baselines, confirming the value of its fusion strategy. Future work will explore hierarchical or sparsity-aware mixers to exploit larger pools.

Conclusions

Expanding existing learning systems to provide high-quality customized models for more domains is challenged by the limited labeled data and the data and device heterogeneities. To solve this problem, HaT first selects a small set of promising source models with small communication and inference overhead, and then fuses their knowledge by assigning sample-wise weights to their predictions. Later, HaT adaptively inject those fused knowledge into the customized models based on the knowledge quality. Experiments spanning multiple tasks, modalities, and models show that HaT consistently surpasses state-of-the-art baselines in accuracy while reducing system overhead, validating its practicality for real-world, large-scale learning-system expansion.

Acknowledgments

This research is supported by Singapore Ministry of Education under its AcRF Tier 1 grant RT14/22, the Global STEM Professorship Scheme of Hong Kong, the HKUST start up grant, and the Research Grants Council (RGC) General Research Fund (GRF) 16210425.

References

Agostinelli, A.; Uijlings, J.; Mensink, T.; and Ferrari, V. 2022. Transferability metrics for selecting source model ensembles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7936–7946.

Borup, K.; Phoo, C. P.; and Hariharan, B. 2023. Distilling from Similar Tasks for Transfer Learning on a Budget. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11431–11441.

Cai, H.; Gan, C.; Wang, T.; Zhang, Z.; and Han, S. 2020. Once for All: Train One Network and Specialize it for Efficient Deployment. In *International Conference on Learning Representations*.

Collins, L.; Hassani, H.; Mokhtari, A.; and Shakkottai, S. 2021. Exploiting shared representations for personalized federated learning. In *International conference on machine learning*, 2089–2099. PMLR.

Criado, M. F.; Casado, F. E.; Iglesias, R.; Regueiro, C. V.; and Barro, S. 2022. Non-iid data and continual learning processes in federated learning: A long road ahead. *Information Fusion*, 88: 263–280.

Dai, G.; Xu, H.; Yoon, H.; Li, M.; Tan, R.; and Lee, S.-J. 2024. ContrastSense: Domain-invariant Contrastive Learning for In-the-Wild Wearable Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(4): 1–32.

Gao, H.; Qian, K.; Ni, J.; Gan, C.; Hasegawa-Johnson, M. A.; Chang, S.; and Zhang, Y. 2024. Speech Self-Supervised Learning Using Diffusion Model Synthetic Data. In Salakhutdinov, R.; Kolter, Z.; Heller, K.; Weller, A.; Oliver, N.; Scarlett, J.; and Berkenkamp, F., eds., *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 14790–14810. PMLR.

Gong, T.; Kim, Y.; Lee, T.; Chottananurak, S.; and Lee, S.-J. 2024. SoTTA: Robust Test-Time Adaptation on Noisy Data Streams. *Advances in Neural Information Processing Systems*, 36.

Gou, J.; Yu, B.; Maybank, S. J.; and Tao, D. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6): 1789–1819.

He, H.; Queen, O.; Koker, T.; Cuevas, C.; Tsiligkaridis, T.; and Zitnik, M. 2023. Domain adaptation for time series under feature and label shifts. In *International Conference on Machine Learning*, 12746–12774. PMLR.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hendrycks, D.; Basart, S.; Mu, N.; Kadavath, S.; Wang, F.; Dorundo, E.; Desai, R.; Zhu, T.; Parajuli, S.; Guo, M.; et al. 2021. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8340–8349.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Karmanov, A.; Guan, D.; Lu, S.; El Saddik, A.; and Xing, E. 2024. Efficient Test-Time Adaptation of Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14162–14171.

Kong, R.; Li, Y.; Yuan, Y.; and Kong, L. 2023. Convrelu++: Reference-based lossless acceleration of conv-relu operations on mobile cpu. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services*, 503–515.

Li, J.; Qiu, S.; Shen, Y.-Y.; Liu, C.-L.; and He, H. 2019. Multisource transfer learning for cross-subject EEG emotion recognition. *IEEE transactions on cybernetics*, 50(7): 3281–3293.

Li, T.; Jiang, W.; Liu, F.; Huang, X.; and Kwok, J. T. 2024a. Learning Scalable Model Soup on a Single GPU: An Efficient Subspace Training Strategy. In *European conference on computer vision (ECCV)*.

Li, T.; Sahu, A. K.; Talwalkar, A.; and Smith, V. 2020. Federated learning: Challenges, methods, and future directions. *IEEE signal processing magazine*, 37(3): 50–60.

Li, X.; Li, Y.; Li, Y.; Cao, T.; and Liu, Y. 2024b. FlexNN: Efficient and Adaptive DNN Inference on Memory-Constrained Edge Devices. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*, 709–723.

Liu, Y.; Zhang, W.; and Wang, J. 2020. Adaptive multi-teacher multi-level knowledge distillation. *Neurocomputing*, 415: 106–113.

Lu, J.; and Sun, S. 2024. CauDiTS: Causal Disentangled Domain Adaptation of Multivariate Time Series. In Salakhutdinov, R.; Kolter, Z.; Heller, K.; Weller, A.; Oliver, N.; Scarlett, J.; and Berkenkamp, F., eds., *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 33113–33146. PMLR.

Lu, X.; Zhou, A.; Xu, Y.; Zhang, R.; Gao, P.; and Li, H. 2024. SPP: Sparsity-Preserved Parameter-Efficient Fine-Tuning for Large Language Models. In Salakhutdinov, R.; Kolter, Z.; Heller, K.; Weller, A.; Oliver, N.; Scarlett, J.; and Berkenkamp, F., eds., *Proceedings of the 41st International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR.

Ouyang, X.; Xie, Z.; Fu, H.; Cheng, S.; Pan, L.; Ling, N.; Xing, G.; Zhou, J.; and Huang, J. 2023. Harmony: Heterogeneous multi-modal federated learning through disentangled model training. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services*, 530–543.

Ouyang, X.; Xie, Z.; Zhou, J.; Huang, J.; and Xing, G. 2021. Clusterfl: a similarity-aware federated learning system for human activity recognition. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*, 54–66.

Pan, S. J.; and Yang, Q. 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10): 1345–1359.

Peng, B.; Fang, Z.; Zhang, G.; and Lu, J. 2024. Knowledge Distillation with Auxiliary Variable. In Salakhutdinov, R.; Kolter, Z.; Heller, K.; Weller, A.; Oliver, N.; Scarlett, J.; and Berkenkamp, F., eds., *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 40185–40199. PMLR.

Phan, M.; Brantley, K.; Milani, S.; Mehri, S.; Swamy, G.; and Gordon, G. J. 2024. When is Transfer Learning Possible? In Salakhutdinov, R.; Kolter, Z.; Heller, K.; Weller, A.; Oliver, N.; Scarlett, J.; and Berkenkamp, F., eds., *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 40642–40666. PMLR.

Pizzolato, S.; Tagliapietra, L.; Cognolato, M.; Reggiani, M.; Müller, H.; and Atzori, M. 2017. Comparison of six electromyography acquisition setups on hand movement classification tasks. *PloS one*, 12(10): e0186132.

Qu, S.; Zou, T.; He, L.; Röhrbein, F.; Knoll, A.; Chen, G.; and Jiang, C. 2024. LEAD: Learning Decomposition for Source-free Universal Domain Adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 23334–23343.

Tan, A. Z.; Yu, H.; Cui, L.; and Yang, Q. 2022. Towards personalized federated learning. *IEEE transactions on neural networks and learning systems*, 34(12): 9587–9603.

Tan, C.; Sun, F.; Kong, T.; Zhang, W.; Yang, C.; and Liu, C. 2018. A survey on deep transfer learning. In *Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part III 27*, 270–279. Springer.

Tong, X.; Xu, X.; Huang, S.-L.; and Zheng, L. 2021. A mathematical framework for quantifying transferability in multi-source transfer learning. *Advances in Neural Information Processing Systems*, 34: 26103–26116.

Vemulapalli, R.; Pouransari, H.; Faghri, F.; Mehta, S.; Farajtabar, M.; Rastegari, M.; and Tuzel, O. 2024. Knowledge Transfer from Vision Foundation Models for Efficient Training of Small Task-specific Models. In *International Conference on Machine Learning (ICML)*.

Warden, P. 2018. Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv preprint arXiv:1804.03209*.

Wen, H.; Li, Y.; Zhang, Z.; Jiang, S.; Ye, X.; Ouyang, Y.; Zhang, Y.; and Liu, Y. 2023. Adaptivenet: Post-deployment neural architecture adaptation for diverse edge environments. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*, 1–17.

Wilson, G.; Doppa, J. R.; and Cook, D. J. 2021. Calda: Improving multi-source time series domain adaptation with contrastive adversarial learning. *arXiv preprint arXiv:2109.14778*.

Xu, H.; Zhou, P.; Tan, R.; Li, M.; and Shen, G. 2021. Limubert: Unleashing the potential of unlabeled data for imu sensing applications. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*, 220–233.

Yao, C.-H.; Gong, B.; Qi, H.; Cui, Y.; Zhu, Y.; and Yang, M.-H. 2022. Federated multi-target domain adaptation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 1424–1433.

Yuan, T.; Zhang, X.; Liu, K.; Liu, B.; Chen, C.; Jin, J.; and Jiao, Z. 2024. Towards surveillance video-and-language understanding: New dataset baselines and challenges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22052–22061.

Zhang, C.; Xie, Y.; Bai, H.; Yu, B.; Li, W.; and Gao, Y. 2021. A survey on federated learning. *Knowledge-Based Systems*, 216: 106775.

Zhang, H.; Chen, D.; and Wang, C. 2022. Confidence-aware multi-teacher knowledge distillation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4498–4502. IEEE.

Zhang, W.; Deng, L.; Zhang, L.; and Wu, D. 2022. A survey on negative transfer. *IEEE/CAA Journal of Automatica Sinica*, 10(2): 305–329.

Zhu, Y.; Zhuang, F.; Wang, J.; Ke, G.; Chen, J.; Bian, J.; Xiong, H.; and He, Q. 2020. Deep subdomain adaptation network for image classification. *IEEE transactions on neural networks and learning systems*, 32(4): 1713–1722.