HiFE: Hierarchical Feature Ensemble Framework for Few-Shot Hypotheses Adaptation

Anonymous authors
Paper under double-blind review

Abstract

The process of transferring knowledge from a source domain to a target domain in the absence of source data constitutes a formidable obstacle within the field of source-free domain adaptation, often termed hypothesis adaptation. Conventional methodologies have been dependent on a robustly trained (strong) source hypothesis to encapsulate the knowledge pertinent to the source domain. However, this strong hypothesis is prone to overfitting the source domain, resulting in diminished generalization performance when applied to the target domain. To mitigate this issue, we advocate for the augmentation of transferable source knowledge via the integration of multiple (weak) source models that are underfitting. Furthermore, we propose a novel architectural framework, designated as the Hierarchical Feature Ensemble (HiFE) framework for Few-Shot Hypotheses Adaptation, which amalgamates features from both the strong and intentionally underfit source models. Empirical evidence from our experiments indicates that these weaker models, while not optimal within the source domain context, contribute to an enhanced generalization capacity of the resultant model for the target domain. Moreover, the HiFE framework we introduce demonstrates superior performance, surpassing other leading baselines across a spectrum of few-shot hypothesis adaptation scenarios.

1 Introduction

Domain adaptation (DA) (Ben-David et al., 2010) refers to the study of leveraging labeled data in a *source domain* (SD) to obtain a predicted model for a given *target domain* (TD) where labels are insufficient or unavailable. Conventional DA methods (Ahmed et al., 2021; Jiang et al., 2021; Kang et al., 2019; Sukhija et al., 2016; Wang et al., 2019) pose a potential risk of exposing private information caused by accessing the source data. To mitigate this concern, recent studies have introduced source-free DA, also referred to as *hypothesis adaptation* (HA) (Liang et al., 2020; Li et al., 2020; Yang et al., 2021; Yi et al., 2023), which leverages a source model to encode the knowledge from the source domain rather than the source data. Recently, *few-shot HA* (FHA) (Chi et al., 2021; Yazdanpanah & Moradi, 2022), which operates effectively in scenarios with limited labeled data from the target domain, has emerged as an appealing approach to address data scarcity. A typical application area of FHA is medical diagnostics. For instance, when employing Magnetic Resonance Imaging (MRI) as the SD and X-ray imaging as the TD, FHA can mitigate challenges such as the inaccessibility of source data owing to patient confidentiality and the scarcity of labeled target X-ray data. The existing FHA methods aim to adapt a strong model (with the best source accuracy) trained on the SD to generate a target hypothesis, which is supposed to outperform the one generated solely from limited target data.

Nevertheless, one single strong model may overfit the SD and perform worse on the TD after the adaptation, as indicated by the result of an experiment. In this experiment, we adapt models with different accuracies from the digit dataset SVHN to the target task MNIST. The findings depicted in Figure 1 demonstrate that even with a straightforward fine-tuning on the TD, under-fitted weak source models (e.g., Model-4 [source acc=76.0%]) exhibit superior performance compared to the strong one (e.g., Model-1 [source acc = 92.2%]). This indicates that **some weak hypotheses that convey source knowledge not encoded in the strong hypothesis could generalize better to the TD**. This research is focused on harnessing the

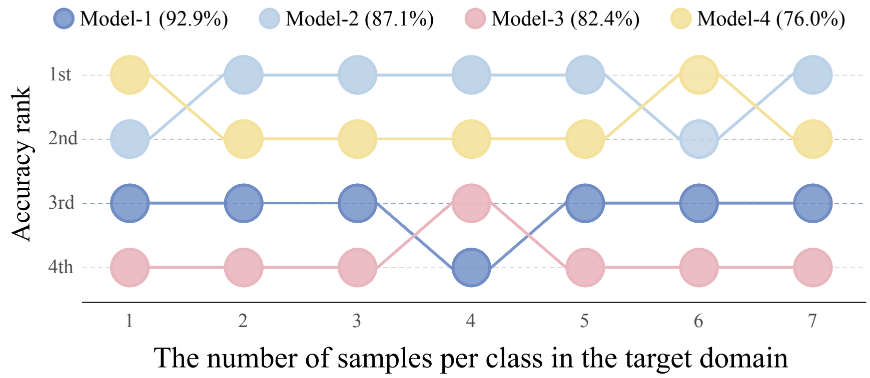


Figure 1: This figure depicts the FHA process from SVHN to MNIST. Four source models were generated using the training data from SVHN and fine-tuned with different quantities of samples from MNIST. The y-axis indicates the performance rank after adaptation, with the highest accuracy in the TD ranked first. Model-4, despite having the lowest source accuracy (76%), exhibits notably superior performance on the TD compared to Model-1, which has the highest source accuracy (92.9%).

readily available weak models for single-source FHA tasks. Notably, the routine training of source models naturally produces a series of weak intermediate models that are usually overlooked and discarded. However, these models are easily obtainable, as reaching out to the source provider to acquire additional weak models results in minimal additional cost. This approach is grounded in the practical advantage of utilizing resources that are often already at hand, yet underutilized.

A new problem: FHAW. Our motivation arises from the above observation that the weak models, although suboptimal for the source domain, may possess underlying knowledge that could be instrumental in the target domain. Thus, we propose to study a new problem: *few-shot hypotheses adaptation with weak models* (FHAW). Compared with previous single-source FHA approaches (Chi et al., 2021; Liang et al., 2020) that rely solely on a single strong hypothesis, FHAW aims to utilize several source hypotheses with varying degrees of accuracy to enhance the diversity of source models. Although some researchers have advocated utilizing multi-source-hypotheses for HA (Ahmed et al., 2021; Shu et al., 2022; Li et al., 2024), these approaches operate under the assumption that multiple *strong source* hypotheses from different domains are available, making them ineffective when presented with weak hypotheses. Our research investigates the significance of *weak source* hypotheses from the same SD in the HA process, bypassing the need for access to multiple source domains.

A viable approach to solve FHAW. To mitigate the potential negative transfer arising from weak models and extract valuable source knowledge for the target task, we have developed a new and efficient framework called the Hierarchical Feature Ensemble framework (HiFE) to address FHAW, as illustrated in Figure 2. The HiFE method employs hierarchical ensemble techniques to enhance the representativeness of intermediate features. It utilizes weighted residual units (WRU) to aggregate features induced by the source hypotheses, as illustrated in Figure 2 (b). WRU merges similar features with skip connections to reduce the risk of forgetting the source knowledge, alleviating the overfitting problem when fine-tuning with few-shot samples. Furthermore, we incorporate feature decorrelation learning (DeCL) by integrating a correlation penalty term into the standard classification loss, thereby enhancing the diversity of intermediate features (refer to Figure 2 (c)). Comprehensive results indicate that HiFE delivers state-of-the-art (SOTA) performance across various domain adaptation tasks.

Main contributions. Our contributions are three-fold:

- To the best of our knowledge, this is the initial investigation into the FHAW problem. FHAW holds practical relevance in numerous private data-based scenarios, as source providers are inclined to offer "redundant" weak models instead of disclosing sensitive datasets. Our work introduces a fresh perspective to encoder the source knowledge in the absence of the source data.

- We propose a new framework to aggregate all source hypotheses at the feature level to address the FHAW problem. We are the first to apply a hierarchical ensemble at the feature level in hypotheses adaptation. We effectively alleviate the over-fitting problem by the design of WRU and improve the generalization of the final hypothesis by incorporating feature DeCL loss under the few-shot setting.
- The comprehensive evaluation of the proposed HiFE methodology, conducted over an array of benchmark datasets—including MNIST, SVHN, USPS, CIFAR-10, STL-10, Amazon, DSLR, and Webcam—has established that our approach achieves performance on par with or exceeding current SOTA methods in various Feature Hashing Algorithm (FHA) tasks. Notably, as detailed in Table 2, the HiFE method surpasses the SOTA by an average accuracy of 4.3% in the digit dataset task USPS → MNIST. Similarly, in the task of adapting DSLR to Webcam datasets, as shown in Table 3, HiFE outperforms the SOTA by 3.6% in accuracy.

2 Related Work

This section presents a brief overview of the literature about traditional domain adaptation, hypothesis adaptation, multi-hypotheses adaptation, and ensemble methods for hypothesis adaptation.

Domain adaptation (DA). Traditional DA is a subfield of machine learning that focuses on learning a hypothesis for a TD when labeled data is insufficient or unavailable by leveraging labeled data from an SD. Numerous DA methods have been proposed for various tasks such as object classification (Liang et al., 2018), object detection (Hsu et al., 2020), and semantic segmentation (Zou et al., 2018). Existing approaches for DA can mainly be categorized into two classes: feature-based DA and instance-based DA. The former aims to learn a domain-invariant representation by minimizing the domain discrepancy in a shared space (Kang et al., 2019; Long et al., 2017). For example, Gradually Vanishing Bridge (Cui et al., 2020) uses bi-directional generation to learn domain-invariant representations. The latter minimizes the discrepancy by re-weighting the source samples for better training. Despite the success achieved by these methods, they require access to source data during the learning process, which incurs significant costs in terms of data transfer and storage as well as risks related to personal information leakage.

Hypothesis adaptation (HA). Researchers have started exploring source-free domain adaptation (SFDA), namely HA, to mitigate the issues arising from accessing source data. Early works addressed the problem by fine-tuning the source hypothesis on the target data (Girshick et al., 2014). However, recent studies have delved into unsupervised DA to investigate the limitations of this straightforward strategy (Ding et al., 2022; Liang et al., 2020; Yang et al., 2022; Yi et al., 2023). Among these methods, SHOT (Liang et al., 2020) proposes a representation learning framework to update the feature extractor through information maximization and self-supervised pseudo-labeling loss. In this framework, pseudo-labels of the target data are refined using the nearest centroids. Similarly, (Yi et al., 2023) views SFDA as the problem of learning with label noise and suggests exploiting the early-time training phenomenon to tackle the issue of pseudo-labels. Notably, these methods rely on large amounts of unlabeled data from the TD to purify the pseudo-labels. On the other hand, TOHAN (Chi et al., 2021) is the first study to explore the HA under a few-shot setting. It proposes generating an intermediate domain that is compatible with the TD to facilitate transfer learning. Many previous works rely on a strong source hypothesis for adaptation, which may not always be the most suitable one for adapting to a specific TD.

Multi-hypotheses adaptation (MHA). MHA extends the HA paradigm by integrating knowledge from source hypotheses from multiple domains. To tackle this, model selection methods (Nguyen et al., 2020; You et al., 2021) have been developed to estimate the transferability of each pre-trained hypothesis. However, the single selected hypothesis may not be able to carry the rich knowledge encapsulated in all of the source hypotheses. Thus, some researchers have turned to parameter ensemble methods (Ahmed et al., 2021; Rusu et al., 2016; Shu et al., 2022; Li et al., 2024). Yet, these approaches often require significant amounts of unlabeled target data to be effective, making them less tenable in the FHA setting. Moreover, these approaches operate under the assumption that each source hypothesis is a strong one from the corresponding SD, rendering them ineffective when presented with weak hypotheses. Besides, these approaches require accessing multiple source domains related to the TD, which is often not feasible. Our research focuses on a more practical and challenging scenario: only one SD is available.

Ensemble methods for HA. Ensemble methods are prominent research in machine learning (Dietterich, 2000; Dong et al., 2020; Sagi & Rokach, 2018; Eilers et al., 2022). These methods have demonstrated that combining multiple hypotheses is advantageous over a single hypothesis in classification and regression problems. However, traditional ensemble methods rely on weighted voting for the final decision and lack the capability of representation learning (Cao et al., 2012). Thus, some researchers have proposed feature-level ensembling. Studies have demonstrated the effectiveness of hierarchical feature representation in improving classification accuracy (Cai et al., 2018; Su et al., 2009). In FHA, the research on developing hierarchical feature-level ensemble methods to derive a comprehensive knowledge representation of all source hypotheses remains limited.

3 Few-Shot Hypotheses Adaptation with Weak Models

3.1 Problem Definition

We address the problem of few-shot hypotheses adaptation with weak models, where several pre-trained source hypotheses, including one strong and some weak hypotheses, are given. Let $\mathcal{X} \subset \mathbb{R}^d$ be an input space and $\mathcal{Y} := \{1, \dots, C\}$ be the label space, where C is the number of classes. To formalize the problem clearly, some definitions are presented as follows.

Definition 1. (Expected and empirical risk). Given a data distribution P over $\mathcal{X} \times \mathcal{Y}$, let $\mathcal{H} = \{h : \mathcal{X} \to \mathcal{Y}\}$ be the hypothesis space and $h_\theta \in \mathcal{H}$ with the parameter $\theta \in \Theta$, then the expected and empirical risks are defined as

$$L(\theta_h) = \mathbb{E}_{(x,y)\sim P}[\ell(\theta_h, x, y)],$$
$$\hat{L}(\theta_h, D) = \frac{1}{n}\sum_{i=1}^{n}(\ell(\theta_h, x_i, y_i)),$$

where ℓ is a proper loss function and $D = \{(x_i, y_i)\}_{i=1}^n \sim P^n$ denotes the *i.i.d.* n observations.

Definition 2. (Strong Hypothesis). Given a set of hypotheses $\hat{\mathcal{H}} = \{h^m\}_{m=1}^M$ from domain S with data D, where M is the hypothesis number, a hypothesis $h_s \in \hat{\mathcal{H}}$ is called a strong hypothesis if $\forall h \in \hat{\mathcal{H}}$, $\hat{L}(\theta_{h_s}, D) \leq \hat{L}(\theta_h, D)$.

Definition 3. (Weak Hypothesis). Given a set of hypotheses $\hat{\mathcal{H}} = \{h^m\}_{m=1}^M$ from domain S with data D, where M is the hypothesis number, a hypothesis $h_w \in \hat{\mathcal{H}}$ is called a weak hypothesis if $\exists h \in \hat{\mathcal{H}}$, $\hat{L}(\theta_h, D) < \hat{L}(\theta_{h_w}, D)$.

Problem 1. (Few-Shot Hypotheses Adaptation with Weak Models (**FHAW**)). Given a set of hypotheses $\hat{\mathcal{H}}$ with a strong hypothesis h_s and M weak hypotheses $\{h_w^m\}_{m=1}^M$ trained on the SD $P_S(X, Y)$, n_t target labeled data $D_t = \{(x_t^i, y_t^i)\}_{i=1}^{n_t}$ that *i.i.d.* drawn from $P_T(X, Y)$ with $n_t \ll n_s$ and $P_S(X, Y) \neq P_T(X, Y)$, FHAW is to learn a target hypothesis $h_t : \mathcal{X} \to \mathcal{Y}$ with h_s, $\{h_w^m\}_{m=1}^M$ and D_t to minimize the expected risk on the TD.

Comparison with FHA. FHA involves the use of a strong hypothesis derived from the SD. However, such a hypothesis is prone to over-fitting on the SD, and their generalizability towards the TD can be limited. To address this challenge, we introduce FHAW by leveraging multiple weak hypotheses to facilitate more effective adaptation. These weak hypotheses can be easily obtained by saving model snapshots during the training of the strong source model with minimal additional cost.

3.2 Addressing FHAW in Principle

We will present a theoretical view based on the PAC-Bayesian framework (Germain et al., 2009; McAllester, 1999; Masegosa, 2020) to demonstrate why we propose to incorporate multiple weak hypotheses for FHA and why our HiFE framework works. In the PAC-Bayesian framework, each hypothesis h_θ has prior knowledge of the hypothesis space Θ, and this prior distribution π is updated to a posterior distribution ρ after feeding samples D to h_θ. In FHAW, multiple models $\{h_{\theta_i}\}_{i=1}^M$ are given with $\theta_i \in \Theta_i$, $\theta = \{\theta_i\}_{i=1}^M$ and $\rho(\theta) = \prod_{i=1}^M \rho_i(\theta_i)$. For a given sample (x, y), we apply the cross-entropy loss $\ell(\theta, x, y) = -\log p(y|x, \theta)$. A bound theorem from (Deng et al., 2023) for the model ensemble is restated below, and the related theorems are shown in B.

Theorem 1. *(Model ensemble error bound (Deng et al., 2023)). Given a data distribution P over $\mathcal{X} \times \mathcal{Y}$, a set of model parameters $\{\Theta_i\}_{i=1}^M$ with associated prior $\{\pi_i\}_{i=1}^M$, where π_i is defined over Θ_i with $\pi_i(\theta_i) \sim \mathcal{N}(0, \sigma^2 I)$, a $\delta \in (0, 1]$, a real number $c > 0$, and $\rho_i(\theta_i)$ is a Dirac-delta distribution centered around θ_i' with $\rho_i(\theta_i) = \delta_{\theta_i'}(\theta_i)$, then we have that the $\mathbb{E}_{\rho(\theta)}(L(\theta))$ is upper bounded by*

$$\frac{1}{M}\sum_{i=1}^{M}\left(\hat{L}(\theta_i', D) + \frac{1}{2cn\sigma^2} \parallel \theta_i \parallel^2 + \frac{d_i}{2cn}\log(2\pi\sigma^2)\right) - \hat{\mathbb{V}}(\rho(\theta), D) + \frac{\epsilon}{cnL},$$

where $\hat{\mathbb{V}}(\rho(\theta), D)$ is the empirical version of a variance term $\mathbb{V}(\rho(\theta))$, which is defined as

$$\mathbb{E}_{\rho(\theta)}\mathbb{E}_{(x,y)\sim P}\left[\frac{1}{2M\max_\theta p(y|x,\theta)^2}\sum_{i=1}^{M}\left(p(y|x,\theta_i) - \frac{1}{M}\sum_{k=1}^{M}p(y|x,\theta_k)\right)^2\right],$$

and ϵ is defined as

$$\log\frac{\mathbb{E}_{\pi(\theta)}\mathbb{E}_{D\sim P^n}\left[e^{cn\left(\sum_{i=1}^{M}\left(L(\theta_i)-\hat{L}(\theta_i,D)\right)-M\left(\mathbb{V}(\theta)-\hat{\mathbb{V}}(\theta,D)\right)\right)}\right]}{\delta}.$$

In Theorem 1, the variance term $\hat{\mathbb{V}}(\rho(\theta), D)$ measures the diversity of all models (Masegosa, 2020). If there exists an input sample x such that $h_{\theta_i}(x) \neq h_{\theta_j}(x)$, then we have $\hat{\mathbb{V}}(\rho(\theta), D) > 0$. Therefore, in the setting of FHAW, adding weak hypotheses increases the diversity of the source models and provides opportunities to decrease this error bound. Minimizing the first term of the error bound in Theorem 1 is equivalent to finding $\theta = \{\theta_i\}_{i=1}^M$ by $\min_\theta \sum_{i=1}^M \left(\hat{L}(\theta_i, D) + \lambda_1 \parallel \theta_i \parallel^2 + \lambda_2 d_i\right)/M$, where d_i is the dimension of θ_i and λ_1, $\lambda_2 > 0$ are hyper-parameters. Based on this formula, we propose a hierarchical feature ensemble module to reduce the dimensionality of features.

4 Few-shot Hypotheses Adaptation via Hierarchical Feature Ensemble

To aggregate knowledge from both strong and weak source hypotheses, we introduce the HiFE framework, depicted in Figure 2. HiFE hierarchically merges features induced by all the source hypotheses. We assume each source hypothesis has been embedded with its specific discriminative knowledge about the SD. Hence, during the aggregation, we use a feature de-correction learning module, making the features as mutually independent as possible at each level to increase the representative power of the intermediate features. We describe the design insights of HiFE in Section 4.1 and Section 4.2.

4.1 Hierarchical Feature Ensemble

Ensemble learning is widely recognized as an effective approach for combining multiple learning methods and improving overall performance (Beven & Binley, 1992; Kuczera & Parent, 1998). While ensemble methods have been utilized for HA in past research (Ahmed et al., 2021), ensemble learning at the feature representation level has received relatively less attention. However, prior research has shown that hierarchical feature representations can significantly enhance classification accuracy. We propose a hierarchical feature ensemble-based approach for FHAW to leverage such benefits. Specifically, our method involves merging source features that contain knowledge of the SD using a *hierarchical feature ensemble module* before feeding them to the final classifier.

To simplify the feature extraction process with the source hypotheses, we follow (Motiian et al., 2017) and (Ahmed et al., 2021) to decompose each hypothesis h into two modules: a feature encoder $g : \mathcal{X} \to \mathbb{R}^d$ and a classifier $c : \mathbb{R}^d \to \mathbb{R}^C$, where d denotes the dimension of the output feature. Thus, we have $h(x) = c(g(x))$,

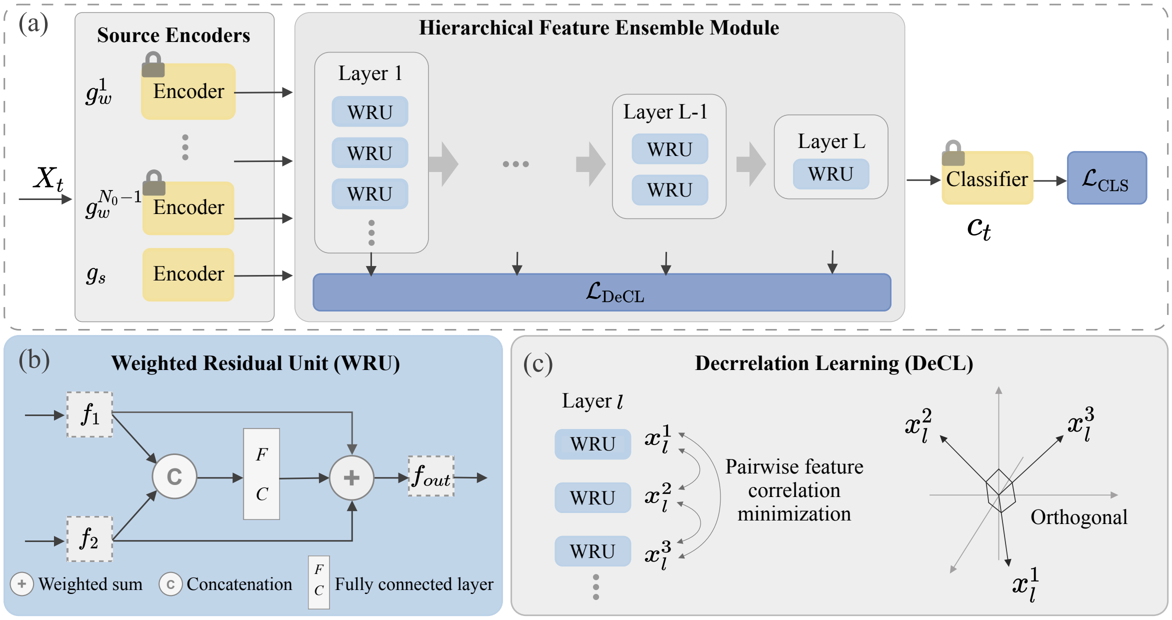


Figure 2: The architecture of HiFE framework. Each source hypothesis consists of a feature encoder and a classifier. We train a model with a hierarchical feature ensemble module to merge features from all source encoders. In this module, features are grouped according to the cosine similarity, and WRU merges the grouped features to generate new features for the next layer. Each WRU leverages skip connections to avoid over-fitting. Besides, we apply the decorrelation learning (DeCL) strategy by adding a correction penalty term to the loss function to encourage feature diversity. The target classifier c_t, initializing with the average of all the source classifiers, is fixed during the training. Only the parameters of the strong encoder g_s and the hierarchical feature ensemble module are updated.

i.e., $h = c \circ g$. In our problem setting, source hypotheses are decomposed to $c_s \circ g_s$ and $\{c_w^m \circ g_w^m\}_{m=1}^M$. Features induced by g_s, $\{g_w^m\}_{m=1}^M$ are fed into the feature ensemble module. Let x_l^i denote the i-th feature at layer l ($l \in \{0, 1, ..., L\}$) and $\{x_0^i\}_{i=1}^{N_0}$ be the N_0 input features. The hierarchical feature ensemble module aims to aggregate all these features into one single feature x_L^1 through a hierarchical method so that x_L^1 contains as much source knowledge as possible. To this end, we must tackle two questions: 1) which features to merge and 2) how to merge the chosen features.

1) *Which features to merge?* According to the Gestalt principles of psychology (Koffka, 2013), humans tend to group similar information during cognitive processing. Taking inspiration from this, we utilize feature similarity as a metric to group similar input features together from the previous layer for the purpose of merging. Given a set of N_l features, we first create a similarity matrix $S \in \mathbb{R}^{N_l \times N_l}$, where $S_{i,j} = \cos(x_l^i, x_l^j)$ is the cosine similarity of features x_l^i and x_l^j. Next, we repeatedly choose two or more features with the highest similarity and merge them into a new feature for the next layer. Such a hierarchical merge process repeats until only one output feature is left.

2) *How to merge features?* In the FHA setting, the small sample size problem limits the feasibility of maintaining adequate validation sets to assess performance before testing unknown samples. Without such validation sets, optimizing the model could cause over-fitting to the limited target data, leading to a local optimum and performance degradation (Goodfellow et al., 2014; Kirkpatrick et al., 2017). To address this issue, we propose the *Weighted Residual Unit* (WRU), adding the "shortcut connections" of the input features to the block output after feature merging (see Figure 2 (b)). The shortcut connections allow the upper layer's features to be directly sent to the next layer, maintaining the source knowledge during adaptation and alleviating the over-fitting problem. Within each WRU, the input features $\{f_i\}_{i=1}^K$ are concatenated and fed into a fully connected (FC) layer, where K is the number of input features. The output of the FC layer

is denoted as $\mathrm{FC}(\{f_i\}_{i=1}^K, W_{\mathrm{FC}})$ with W_{FC} be the learnable parameters of this layer. Unlike the shortcut connections that perform identity mapping in ResNet (He et al., 2016), we perform a weighted element-wise addition of $\{f_i\}_{i=1}^K$ and $\mathrm{FC}(\{f_i\}_{i=1}^K, W_{\mathrm{FC}})$ to balance the influence of different input features induced by the source hypotheses. The function of WRU can be formalized as follows.

$$\mathrm{WRU}(\{f_i\}_{i=1}^K) = \alpha_0 \cdot \mathrm{FC}(\{f_i\}_{i=1}^K, W_{\mathrm{FC}}) + \sum_{i=1}^{K} \alpha_i \cdot f_i, \tag{1}$$

where $\{\alpha_i\}_{i=0}^K$ are the learnable weights. We add the batch normalization and ReLu layers after the weighted sum for better performance. If the dimension of f_i is not equal to that of the output of FC, we can make a linear projection of f_i by extending α_i to a square matrice W_i to match the dimension. The application of WRU allows us to preserve some source knowledge and learn new information from the target samples simultaneously.

4.2 Decorrelation Learning

It has been commonly agreed that diversity is a success factor of ensemble algorithms. Different opinions from multiple classifiers are expected to reduce the generalization error. Traditional decorrelation learning (DeCL) methods encourage diversity explicitly by adding a correlation penalty term to the final error function (Liu & Yao, 1999; Shi et al., 2018; Wang et al., 2010). When it comes to feature ensemble, learning the features with good discriminative power is also essential for various high-level vision tasks (Wen et al., 2016; Cheng et al., 2018). To promote the learning of features widely distributed across the feature space and embed various forms of source knowledge, we apply DeCL in the feature space to encourage independence between features in each layer. In this regard, we introduce a cosine similarity penalty to decrease feature correlation and encourage feature diversity (see Figure 2 (c)). Specifically, we calculate the pairwise square values of cosine similarities for all features in the same layer and sum them up from all layers. The corresponding feature DeCL loss is defined as

$$\mathcal{L}_{\mathrm{DeCL}} = \sum_{l=1}^{L} \sum_{i=1}^{N_l-1} \sum_{j=i+1}^{N_l} \cos(x_l^i, x_l^j)^2, \tag{2}$$

where N_l is the number of features at layer l and $\cos(x_l^i, x_l^j) = (x_l^i \cdot x_l^j)/(\|x_l^i\| \cdot \|x_l^j\|)$. Furthermore, to enable the adaptation of the ensemble network to the TD, we incorporate the knowledge of TD by fitting the network to the labeled target data. To accomplish this, we adopt the standard cross-entropy loss, which is defined as follows,

$$\mathcal{L}_{\mathrm{CLS}} = \mathbb{E}_{(x_t, y_t)\sim P_T}[\mathrm{CE}(c_t(A(x_t)), y_t)], \tag{3}$$

where $\mathrm{CE}(\cdot)$ denotes the cross-entropy loss and $A(x_t)$ refers to the output of the feature ensemble module when fed x_t to the source encoders. To summarize, we train the ensemble network using joint supervision that combines the target supervised loss (Equation (3)) and a feature DeCL penalty term (Equation (2)) with a hyper-parameter β to trade off the two aspects (Equation (4)). The target supervised loss guides the network in learning the knowledge from the target samples, while the feature DeCL loss promotes mutual independence amongst features in each layer, thereby increasing the feature diversity and preserving the distinct discriminative knowledge of each source hypothesis.

$$\mathcal{L}(\beta) = (1-\beta) \cdot \mathcal{L}_{\mathrm{CLS}} + \beta \cdot \mathcal{L}_{\mathrm{DeCL}}. \tag{4}$$

5 Experiments

5.1 Experimental Setup

Datasets. We conduct experiments on various standard DA benchmarks to evaluate our approach.[1]

[1] The full code is available at `https://anonymous.4open.science/r/HFE-DCL-7761`.

Digits. We choose three-digit datasets, i.e., MNIST (**M**), USPS (**U**), and SVHN (**S**) for our experiments. Following (Motiian et al., 2017; Chi et al., 2021), we experiment with different numbers of target samples from 1 to 7 per class.

Office. We use three domains of the office datasets (Saenko et al., 2010): Amazon (**A**), DSLR (**D**), and Webcam (**W**). Each domain contains 31 object classes in the office environment. We conduct several experiments with different numbers of target samples per class ranging from 1 to 5.

Image classification. We use two image classification benchmarks CIFAR-10 (**CF**) (Krizhevsky, 2009) and STL-10 (**ST**) (Coates et al., 2011). Each benchmark consists of 10 classes of objects, and nine classes are overlapped. We remove the non-overlapped classes ("monkey" and "frog") and reduce the tasks to a 9-class classification problem following the procedure in (Shu et al., 2018). As the two domains are more complex than digits, we increase the number of target samples to 15 and 20 for each class.

Baseline methods. In the context of the novel FHAW problem setting, we establish our baseline comparisons by adapting and refining several established approaches in the field. We conducted a comprehensive evaluation against four existing methods for HA and their respective variations. Initially, SHOT (Liang et al., 2020), a hypothesis transfer learning framework tailored for unsupervised HA, served as a foundation. In our study, we preserved its model adaptation module, tweaking it to leverage labeled target data to support supervised HA, aligning it seamlessly with our experimental setup. The performance outcomes from employing solely the strong hypothesis with SHOT are denoted as *SHOT-strong*. The subsequent contender, TOHAN (Chi et al., 2021), specifically tackles the FHA challenge. Both SHOT and TOHAN are engineered to adapt a singular hypothesis to the TD independently. To evaluate against our multi-hypotheses adaptation approach, we extended these methods to *SHOT-ens* and *TOHAN-ens* by employing a straightforward ensemble technique, following the methodology outlined in (Ahmed et al., 2021). Furthermore, we included the models DECISION (Ahmed et al., 2021) and Bi-ATEN (Li et al., 2024), specifically designed for multi-source-free unsupervised HA, to contrast with our single-source multi-model HA strategy.

Network architecture. For digit recognition tasks, we employ the same architectures utilized in SHOT (Liang et al., 2020), namely using the LeNet-5 (LeCun et al., 1998) for MNIST, USPS, and a modified version of LeNet for the slightly more complex SVHN dataset. For the image classification tasks, we adopt ResNet-18 and ResNet-50 ((He et al., 2016)) as the backbones for the CIFAR-10/STL-10 and office datasets, respectively.

Source hypotheses preparation. We train a single optimal hypothesis as a strong source hypothesis for each SD and save seven intermediate snapshots as weak source hypotheses with varying accuracy levels. To acquire the hypotheses with different source accuracies, we first set an accuracy range $[\text{acc}_{min}, \text{acc}_s]$, where acc_{min} is a preset value around at 40-60% and acc_s is an estimation of the accuracy of the strong hypothesis. Then, we split this range into several uniform intervals and save one model snapshot at each interval to get weak hypotheses for each SD. The source data can be discarded after getting all the required source hypotheses. Table 1 shows the source models generated with the source dataset Mnist and their corresponding accuracy ranges. We generate 12 source models $\{h_i|_{i=1}^{12}\}$. According to our definition in Section 3, the first 11 models $\{h_i|_{i=1}^{11}\}$ are weak source hypotheses, while the last one h_{12} is the strong hypothesis. Among these models, we used $\{h_i|_{i=5}^{12}\}$ (8 models) in the experiments shown in Table 2 and Table 7, while all models are prepared for ablation studies. The process of training the target ensemble hypothesis with HiFE is detailed in Appendix A.

MI	AC	MI	AC	MI	AC	MI	AC
h_1	[40, 45)	h_4	[55, 60)	h_7	[70, 75)	h_{10}	[85, 90)
h_2	[45, 50)	h_5	[60, 65)	h_8	[75, 80)	h_{11}	[90, 95)
h_3	[50, 55)	h_6	[65, 70)	h_9	[80, 85)	h_{12}	[95, 100)

Table 1: The model indexes (MI) and their corresponding accuracy range (AC) of source models generated from the Mnist dataset.

Tasks	Hypothesis Number	Method	Number of Target Data per Class							Avg
			1	2	3	4	5	6	7	
$U \rightarrow M$	Single	SHOT-worst	42.1±1.2	44.3±0.8	49.9±1.0	48.4±1.6	50.7±0.6	50.9±1.1	50.9±0.8	48.2
		SHOT-best	92.1±1.5	93.4±1.2	93.7±0.9	93.6±1.0	93.7±1.5	93.5±0.8	94.0±0.6	93.4
		SHOT-strong	89.8±1.1	90.3±1.3	92.0±1.5	91.3±1.6	92.0±0.7	92.0±0.8	91.9±0.5	91.3
	Multiple	SHOT-ens	86.8±1.6	88.5±1.8	90.0±2.1	89.5±2.3	90.5±1.9	90.6±1.0	90.8±1.3	89.5
		TOHAN-ens	87.3±1.8	89.7±1.6	90.1±1.6	90.5±1.4	91.2±1.5	92.5±0.9	93.5±0.7	90.7
		DECISION	88.7±2.3	88.8±1.8	89.6±2.4	89.8±2.1	90.3±1.7	90.2±1.3	90.5±1.1	89.7
		Bi-ATEN	89.5±1.1	90.9±1.2	91.5±0.8	91.1±0.4	92.1±1.2	92.5±2.3	90.5±2.6	91.2
		HiFE (ours)	**92.7±0.8**	**94.9±0.2**	**95.0±0.4**	**95.2±0.6**	**95.4±0.5**	**95.4±0.7**	**96.1±0.3**	**95.0**
$S \rightarrow M$	Single	SHOT-worst	40.9±1.0	45.1±1.2	50.9±1.1	51.6±0.9	51.7±1.1	51.8±0.8	51.9±0.8	49.1
		SHOT-best	74.8±1.4	75.1±1.2	79.8±1.3	79.1±0.9	80.6±1.1	79.8±0.5	79.1±0.6	78.3
		SHOT-strong	74.5±2.0	73.5±1.1	78.7±1.8	78.2±1.5	78.8±1.3	78.6±0.9	78.7±0.8	77.3
	Multiple	SHOT-ens	75.6±2.2	74.9±1.2	81.2±2.6	81.5±1.4	82.0±1.3	81.6±1.0	81.7±1.5	79.8
		TOHAN-ens	79.0±1.9	**85.9±2.1**	87.5±1.6	89.5±1.1	90.1±1.4	90.6±1.2	91.1±0.9	87.7
		DECISION	71.9±1.3	72.1±2.1	72.5±2.0	73.4±1.5	75.0±1.2	76.7±1.5	79.2±1.0	74.4
		Bi-ATEN	75.1±1.7	77.2±1.1	77.1±2.3	79.7±2.5	80.1±2.9	82.5±1.9	83.1±1.6	79.3
		HiFE (ours)	**79.2±2.1**	85.7±2.0	**88.1±1.0**	**90.3±0.9**	**92.2±0.7**	**92.5±0.9**	**92.8±1.0**	**88.7**
$U \rightarrow S$	Single	SHOT-worst	15.8±2.1	14.8±1.9	14.7±0.9	14.3±1.5	14.3±1.7	14.0±0.9	14.4±0.5	14.6
		SHOT-best	32.6±1.1	32.4±1.6	34.5±1.2	37.3±2.0	38.4±0.8	40.6±0.6	40.5±0.7	36.6
		SHOT-strong	32.6±1.1	32.3±1.7	34.3±1.6	37.0±1.3	38.2±0.9	40.2±0.8	40.4±0.9	36.4
	Multiple	SHOT-ens	**33.3±2.1**	32.1±1.8	34.1±1.9	36.5±1.2	38.1±1.4	39.9±0.9	40.4±0.9	36.3
		TOHAN-ens	31.7±1.8	31.0±1.4	35.8±1.3	36.9±0.9	**40.5±0.6**	42.6±0.8	43.1±0.7	37.4
		DECISION	30.3±2.1	30.5±2.0	31.2±1.8	31.5±1.9	32.0±1.2	32.1±1.3	32.4±0.9	31.4
		Bi-ATEN	31.5±1.5	30.9±1.9	33.5±1.7	33.1±1.8	35.3±1.6	35.3±0.8	37.1±1.4	33.8
		HiFE (ours)	33.0±2.0	**32.9±1.2**	**37.5±0.8**	**39.8±0.8**	40.1±1.0	**42.7±1.1**	**43.3±0.9**	**39.5**

Table 2: Classification accuracy±standard deviation (%) on three adaptation tasks of **digit** datasets. M, U, and S refer to MNIST, USPS, and SVHN, respectively. The suffixes -best and -worst refer to the best and worst results after adapting each single source hypothesis. The suffixes -strong and -ens refer to the result of adapting the strong hypothesis and the ensemble of all hypotheses, respectively. Results of SHOT (Liang et al., 2020), TOHAN (Chi et al., 2021), DECISION (Ahmed et al., 2021), Bi-ATEN (Li et al., 2024), and our HiFE are presented. The highest accuracy is marked in bold.

5.2 Result Analysis

Results of digit classification tasks.

We evaluate the effectiveness of our approach on six closed-set adaptation tasks for digit classification. These tasks are by pairwise combinations of the three domains S, M, and U. We report the results of three tasks in Table 2 (more results can be found in Appendix C). Firstly, as shown in Table 2, there exist some weak hypotheses that can perform better than the strong hypothesis after adaptation (see the comparison of SHOT-best and SHOT-strong), supporting our motivation of adopting the weak hypotheses. Moreover, incorporating the weak hypotheses allows our proposed HiFE to outperform SHOT-strong. For instance, compared with the average accuracy of SHOT-strong (77.3%), HiFE leads to higher average accuracy (88.7%) in the task $S \rightarrow M$. Additionally, despite some weak hypotheses with bad adaptation performance (see SHOT-worst), HiFE can largely avoid the severe negative transfer and achieve the best performance than previous ensemble approaches. For example, HiFE outperforms the SOTA (TOHAN-ens) by 4.3% in the average accuracy of $U \rightarrow M$ task.

Results of office object classification tasks. We show the results of three closet-set adaptation tasks with office datasets in Table 3 (more results can be found in Appendix D). The proposed HiFE consistently improves the adaptation performance, boosting the average accuracy from 60.1% to 64.2% in task $W \rightarrow A$. While HiFE is designed to adapt source models from a single SD to a TD, our approach also works effectively when the source models come from multiple domains (see Appendix E). In addition, HiFE also outperforms the SOTA in the partial FHA scenario (as shown in Appendix F).

Tasks	Hypothesis Number	Methods	Number of Target Data per Class					Average
			1	2	3	4	5	
$A \to D$	Multiple	SHOT-ens	76.5±1.9	78.6±1.5	78.7±1.7	80.1±0.9	81.2±1.2	79.0
		TOHAN-ens	78.5±1.6	79.5±1.3	83.2±0.9	85.1±1.1	87.1±1.1	82.7
		DECISION	**79.2±1.5**	80.2±2.2	80.8±1.2	81.5±2.0	83.5±1.3	81.0
		Bi-ATEN	79.1±1.2	81.6±1.4	81.5±0.9	82.1±1.5	84.1±2.1	81.7
		HiFE (ours)	**79.2±1.0**	**84.3±0.4**	**85.7±1.0**	**86.2±0.9**	**89.2±0.8**	**85.0**
$D \to A$	Multiple	SHOT-ens	56.8±2.0	58.0±1.9	59.2±1.7	61.8±0.5	62.5±0.9	59.7
		TOHAN-ens	58.1±1.3	60.8±1.2	63.1±1.9	63.8±0.8	64.1±0.9	62.0
		DECISION	54.1±1.6	54.2±2.5	56.1±2.1	57.4±0.9	58.5±0.7	56.1
		Bi-ATEN	55.2±1.3	57.1±2.1	60.3±0.9	61.5±1.1	62.5±1.8	59.3
		HiFE (ours)	**61.8±1.0**	**64.7±0.7**	**67.2±0.6**	**66.8±1.0**	**67.5±0.9**	**65.6**
$W \to A$	Multiple	SHOT-ens	55.1±1.2	58.2±1.6	59.9±1.4	60.8±1.1	61.2±1.1	59.0
		TOHAN-ens	56.5±1.0	60.1±0.9	60.4±1.2	61.2±0.8	62.5±0.7	60.1
		DECISION	54.1±2.1	54.9±1.8	55.6±1.6	56.5±1.2	58.1±1.2	55.8
		Bi-ATEN	56.2±1.9	58.4±1.6	58.9±2.2	61.5±1.3	61.7±1.1	59.3
		HiFE (ours)	**62.5±2.5**	**65.1±1.5**	**64.4±1.2**	**64.3±0.9**	**64.8±0.8**	**64.2**

Table 3: Classification accuracy±standard deviation (%) on three adaptation tasks of office datasets. A, D, and W are abbreviations of Amazon, DSLR, and Webcam. The suffix -ens refers to the result of the ensemble of all adapted hypotheses. Results of SHOT (Liang et al., 2020), TOHAN (Chi et al., 2021), DECISION (Ahmed et al., 2021), Bi-ATEN (Li et al., 2024), and our HiFE are presented. The bold value represents the highest accuracy.

Results of image classification tasks. For image classification, we evaluate our approach on two adaptation tasks, $CF \to ST$ and $ST \to CF$. As shown in Table 4, we achieve 1.3% average improvement over the SOTA ensemble approaches in the $CF \to ST$ task.

Tasks	Methods	Data Number		Average
		15	20	
$CF \to ST$	SHOT-ens	70.3±0.4	70.5±0.6	70.4
	TOHAN-ens	67.5±0.6	69.8±0.5	68.7
	DECISION	70.4±0.4	70.6±0.6	70.5
	Bi-ATEN	70.7±0.3	70.9±0.5	70.8
	HiFE (ours)	**71.6±0.4**	**71.9±0.3**	**71.8**
$ST \to CF$	SHOT-ens	53.1±0.6	53.5±0.5	53.3
	TOHAN-ens	52.5±0.6	52.6±0.8	52.6
	DECISION	54.2±0.5	54.5±0.6	54.4
	Bi-ATEN	53.7±0.4	54.6±0.4	54.1
	HAFA	**55.0±0.3**	**55.3±0.3**	**55.2**

Table 4: Classification Accuracy±standard deviation (%) on two tasks between CIFAR-10 (CF) and STL-10 (ST).

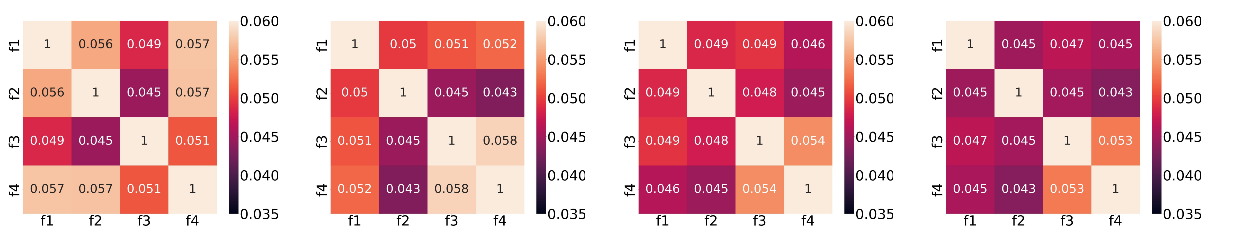
5.3 Ablation Studies

Ablation study on the feature DeCL loss. We study the advantage of our training loss by incorporating feature DeCL loss $\mathcal{L}_{\text{DeCL}}$ in Equation (4) with different β values ranging from 0 to 1.0 with the digit datasets. In this context, $\beta = 0$ corresponds to training the network using only the supervised loss $\mathcal{L}_{\text{CLS}}$, while $\beta = 1$ corresponds to training the network using only the feature DeCL loss $\mathcal{L}_{\text{DeCL}}$. As shown in Table 5, our optimal results generally occur at $\beta = 0.1$, which yields an average improvement of 1.4% compared to the result obtained when no feature DeCL loss is used ($\beta = 0$). Notably, even when the model is trained solely using the feature DeCL loss $\mathcal{L}_{\text{DeCL}}$ ($\beta = 1.0$), it still achieves an average improvement of 10.9% compared to the accuracy before the adaptation (WA), demonstrating the effectiveness of the feature DeCL loss. We also visualize the correlation matrixes of the features after the merge at the first layer of the task $U \to M$ when $\beta = 0.1$. As depicted in Figure 3, as the training progresses, the feature DeCL loss guides the decrease of most of the correlation values between the four features, thereby increasing feature diversity.

Ablation study on the number of weak hypotheses. We conduct an ablation study to analyze the impact of the number of weak hypotheses on the final performance, providing insights into choosing a proper

Tasks	The Value of β								WA
	0	0.1	0.2	0.3	0.5	0.7	0.9	1.0	
$S \rightarrow M$	93.2	**94.0**	93.9	93.2	92.3	90.0	86.4	77.0	67.1
$S \rightarrow U$	94.5	**95.6**	**95.6**	94.1	92.8	91.2	88.7	82.3	78.2
$M \rightarrow S$	52.2	54.2	**54.3**	53.2	52.8	52.7	52.1	46.0	23.2
$M \rightarrow U$	96.1	**97.1**	96.8	96.3	96.1	95.1	94.6	91.2	70.5
$U \rightarrow S$	44.5	**46.2**	44.5	42.5	40.1	39.6	39.1	33.2	26.2
$U \rightarrow M$	94.9	**95.9**	95.3	95.1	94.6	92.9	91.3	89.9	88.5
Average	79.2	**80.6**	80.1	79.1	78.1	76.9	75.4	69.9	59

Table 5: Ablation study on the feature decrrelation learning loss balance parameter β in Equation (4). M, U, and S are abbreviations of MNIST, USPS, and SVHN. WA indicates the accuracy of the model without the adaptation. The bold value represents the highest accuracy (%).



(a) $e = 0$ (b) $e = \frac{1}{3}E_{\max}$ (c) $e = \frac{2}{3}E_{\max}$ (d) $e = E_{\max}$

Figure 3: The correlation matrixes of the features after the merge at the first layer on digit task $U \rightarrow M$ when the β value is set to 0.1. (a) $\sim$ (d) shows the results over different training stages with e, $E_{\max}$ being the current and maximum number of epochs. As the training continues, we observe that most of the correlation values between the four features in this layer decrease (the darker the color, the lower the corresponding correlation value).

number of weak hypotheses to balance the cost and performance. We do this experiment in an adaptation task from domain Mnist to USPS. We select varying numbers (from 2 to 6) of source hypotheses from the models provided in Table 1 and make sure the accuracy range of each group of source models is the same. For each experiment, the selected weak hypotheses started from h_5 and ended with h_{11}, ensuring the selected accuracy range was [60, 95). The results are presented in $\text{Exp}_{1-1} \sim \text{Exp}_{1-5}$ of Table 6. As shown in Table 6, when the number of weak hypotheses is less than 3, the average accuracy (94.2% in Exp_{1-1}) is lower than that when the number of weak hypotheses is greater than 3. Moreover, the average accuracy remains consistent when the number of weak hypotheses exceeds 3 (see the comparison of the result of $\text{Exp}_{1-3} \sim \text{Exp}_{1-5}$).

Ablation study on the accuracy range of weak hypotheses. To investigate the impact of the accuracy range of weak hypotheses on the final performance, we leverage the source models presented in Table 1 and do the digit adaptation task from MNIST to USPS using source models with varying accuracy ranges. We conducted experiments from Exp_{2-1} to Exp_{2-4} as outlined in Table 6. Our results indicate that as the accuracy range increases, the performance after adaptation improves. It is important to note that weak source models with an accuracy lower than 55% may harm the final performance. Our comparison of Exp_{2-1} and Exp_{2-5} revealed that using weak hypotheses with such low accuracy resulted in worse performance than adaptation without weak hypotheses.

Ablation Study on the hierarchical layer number. To investigate the impact of the layer number in the ensemble module on the final performance, we have experimented with varying number of input features fed into WRU, which subsequently alters the number of merge layers within the hierarchical feature ensemble module. we conducted experiments from Exp_{3-1} to Exp_{3-4}, modifying the number of input features for each WRU. The results are presented in Table 6. Notably, in Exp_{3-4}, we utilized a simple weighted feature sum to merge all source features rather than using HEFM with WRU for feature aggregation. When we set the number of input features for each WRU to 2 or 4, with the corresponding layer number in the ensemble module to be 4 and 3, respectively, the adaptation average accuracy is similar. However, when the number

Exp ID	HN	Weak Hypothesis Accuracy Range	Weak Hypothesis Indices	FN	LN	Number of Target Data per Class				Average
						1	3	5	7	
Exp_{1-1}	2	[60, 95)	h_5, h_{11}	2	3	92.0±1.2	94.6±0.7	94.5±0.5	95.6±0.5	94.2
Exp_{1-2}	3	[60, 95)	h_5, h_8, h_{11}	2	3	92.3±1.0	94.8±0.6	95.7±0.7	96.1±0.7	94.7
Exp_{1-3}	4	[60, 95)	h_5, h_7, h_9, h_{11}	2	3	92.7±1.1	95.2±0.6	96.1±0.5	96.7±0.2	95.2
Exp_{1-4}	5	[60, 95)	$h_5, h_7, h_8, h_9, h_{11}$	2	3	92.6±1.0	95.3±0.7	96.0±0.5	96.6±0.4	95.1
Exp_{1-5}	6	[60, 95)	$h_5, h_6, h_7, h_8, h_9, h_{11}$	2	3	92.8±0.9	95.1±0.5	96.2±0.7	96.8±0.5	95.2
Exp_{2-1}	3	[40, 55)	h_1, h_2, h_3	2	3	91.1±0.7	93.8±0.6	94.5±0.8	95.2±0.8	93.7
Exp_{2-2}	3	[55, 70)	h_4, h_5, h_6	2	3	91.5±0.7	94.1±0.6	95.3±0.6	96.0±0.8	94.2
Exp_{2-3}	3	[70, 85)	h_7, h_8, h_9	2	3	92.2±0.6	94.7±0.8	95.7±0.8	96.7±0.6	94.8
Exp_{2-4}	3	[80, 95)	h_9, h_{10}, h_{11}	2	3	92.4±0.8	94.8±0.4	96.1±0.8	96.7±0.6	95.0
Exp_{2-5}	0	-	-	1	1	91.5±0.6	94.2±0.6	94.2±0.9	95.5±0.7	93.9
Exp_{3-1}	7	[60, 95)	$h_{5\sim11}$	2	4	93.0±1.4	95.3±0.5	96.0±0.3	96.7±0.3	95.3
Exp_{3-2}	7	[60, 95)	$h_{5\sim11}$	4	3	93.2±1.1	94.9±0.6	96.2±0.5	96.6±0.4	95.2
Exp_{3-3}	7	[60, 95)	$h_{5\sim11}$	8	1	92.3±1.2	93.1±1.0	95.4±0.8	95.6±0.5	94.1
Exp_{3-4}	7	[60, 95)	$h_{5\sim11}$	/	1	92.1±1.0	92.5±0.9	94.8±0.6	95.1±0.7	93.6

Table 6: Classification accuracy±standard deviation (%) on digit adaptation task $Mnist \rightarrow USPS$ with varying parameters including the number of weak hypotheses (HN), the accuracy ranges of the weak hypotheses, the number of input features to each WRU (FN), and the layer number (LN) in the hierarchical feature ensemble module.

of input features for each WRU increases to 8, we apply one WRU to merge the eight source encoders at once, and the adaptation average accuracy decreases to 94.1%.

6 Limitations and Future Work

HIFE leverages multiple source hypotheses with varying accuracy levels from the source domain to improve the performance of models in the target domain. By exploiting the diversity of source models, HIFE has the potential to enhance the generalization capabilities of the adapted models. However, the additional hypotheses result in increased model transfer and storage costs. Moreover, the increase in the number of parameters of the target model leads to higher computational costs. Nonetheless, we argue that the benefits of leveraging multiple source models with different strengths outweigh the costs associated with processing additional hypotheses, particularly when source data is absent for transfer. With the growing need to address privacy concerns and mitigate data-sharing challenges in real-world applications, opting for weak models simplifies collaboration between source providers and users.

For future research, it would be beneficial to investigate methods for generating weak hypotheses with higher diversity. Although the current experiments obtained weak hypotheses in the same run as generating the final strong hypotheses for simplicity, there is potential for improvement by obtaining weak hypotheses through different random seeds, hyperparameter choices, or training on different subsets of the source data. By increasing the diversity of weak hypotheses, we could obtain better performance after adaptation and further improve the effectiveness of the proposed approach.

7 Conclusion

In this paper, we investigate the potential of utilizing weak source hypotheses for domain adaptation and introduce a new problem setting termed "few-shot hypotheses adaptation with weak models". To tackle this problem, we design a new framework called HiFE, which leverages an array of readily available weak hypotheses to improve the adaptation performance of a strong source hypothesis. As a result, HiFE significantly mitigates the occurrence of over-fitting under the few-shot setting and achieves the SOTA performance across various adaptation tasks. This research introduces an innovative perspective for addressing the FHA problem in scenarios where the source data is inaccessible and the target data is limited. Additionally, this research shed light on the use of a weak source model to boost the practical application of transfer learning in scenarios where data privacy concerns are on the rise.

References

Sk Miraj Ahmed, Dripta S Raychaudhuri, Sujoy Paul, Samet Oymak, and Amit K Roy-Chowdhury. Unsupervised multi-source domain adaptation without access to source data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10103–10112, 2021.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010.

Keith Beven and Andrew Binley. The future of distributed models: model calibration and uncertainty prediction. *Hydrological processes*, 6(3):279–298, 1992.

Yaoming Cai, Xiaobo Liu, Yongshan Zhang, and Zhihua Cai. Hierarchical ensemble of extreme learning machine. *Pattern Recognition Letters*, 116:101–106, 2018.

Jiuwen Cao, Zhiping Lin, Guang-Bin Huang, and Nan Liu. Voting based extreme learning machine. *Information Sciences*, 185(1):66–77, 2012.

Gong Cheng, Ceyuan Yang, Xiwen Yao, Lei Guo, and Junwei Han. When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative cnns. *IEEE Transactions on Geoscience and Remote Sensing*, 56(5):2811–2821, 2018.

Haoang Chi, Feng Liu, Wenjing Yang, Long Lan, Tongliang Liu, Bo Han, William Cheung, and James Kwok. Tohan: A one-step approach towards few-shot hypothesis adaptation. In *Advances in Neural Information Processing Systems*, volume 34, pp. 20970–20982, 2021.

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 215–223. JMLR Workshop and Conference Proceedings, 2011.

Shuhao Cui, Shuhui Wang, Junbao Zhuo, Chi Su, Qingming Huang, and Qi Tian. Gradually vanishing bridge for adversarial domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12455–12464, 2020.

Danruo Deng, Guangyong Chen, Furui Liu, Jinpeng Li, Jianye Hao, and Heng Pheng-Ann. A simple and provable method to adapt pre-trained model across domains with few samples, 2023. URL `https://openreview.net/forum?id=eW2zCT1gm3`.

Thomas G Dietterich. Ensemble methods in machine learning. In *Multiple Classifier Systems: First International Workshop, MCS 2000 Cagliari, Italy, June 21–23, 2000 Proceedings 1*, pp. 1–15. Springer, 2000.

Ning Ding, Yixing Xu, Yehui Tang, Chao Xu, Yunhe Wang, and Dacheng Tao. Source-free domain adaptation via distribution estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7212–7222, 2022.

Xibin Dong, Zhiwen Yu, Wenming Cao, Yifan Shi, and Qianli Ma. A survey on ensemble learning. *Frontiers of Computer Science*, 14:241–258, 2020.

Dirk Eilers, Felippe Schmoeller Roza, and Karsten Roscher. Ensemble-based uncertainty estimation with overlapping alternative predictions. In *Deep Reinforcement Learning Workshop NeurIPS 2022*, 2022.

Pascal Germain, Alexandre Lacasse, François Laviolette, and Mario Marchand. Pac-bayesian learning of linear classifiers. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 353–360, 2009.

Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587, 2014.

Ian J Goodfellow, Mehdi Mirza, Aaron Courville, and Yoshua Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *Stat*, 1050:6, 2014.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Han-Kai Hsu, Chun-Han Yao, Yi-Hsuan Tsai, Wei-Chih Hung, Hung-Yu Tseng, Maneesh Singh, and Ming-Hsuan Yang. Progressive domain adaptation for object detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 749–757, 2020.

Pin Jiang, Aming Wu, Yahong Han, Yunfeng Shao, Meiyu Qi, and Bingshuai Li. Bidirectional adversarial training for semi-supervised domain adaptation. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pp. 934–940, 2021.

Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4893–4902, 2019.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

Kurt Koffka. *Principles of Gestalt psychology.* routledge, 2013.

A Krizhevsky. Learning multiple layers of features from tiny images. *Master's thesis, University of Tront*, 2009.

George Kuczera and Eric Parent. Monte carlo assessment of parameter uncertainty in conceptual catchment models: the metropolis algorithm. *Journal of Hydrology*, 211(1-4):69–85, 1998.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, and Si Wu. Model adaptation: Unsupervised domain adaptation without source data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9641–9650, 2020.

Xinyao Li, Jingjing Li, Fengling Li, Lei Zhu, and Ke Lu. Agile multi-source-free domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 13673–13681, 2024.

Jian Liang, Ran He, Zhenan Sun, and Tieniu Tan. Aggregating randomized clustering-promoting invariant projections for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41 (5):1027–1042, 2018.

Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International Conference on Machine Learning*, pp. 6028–6039. PMLR, 2020.

Yong Liu and Xin Yao. Ensemble learning via negative correlation. *Neural Networks*, 12(10):1399–1404, 1999.

Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *International Conference on Machine Learning*, pp. 2208–2217. PMLR, 2017.

Andres Masegosa. Learning under model misspecification: Applications to variational and ensemble methods. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5479–5491, 2020.

David A McAllester. Pac-bayesian model averaging. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, pp. 164–170, 1999.

Saeid Motiian, Quinn Jones, Seyed Iranmanesh, and Gianfranco Doretto. Few-shot adversarial domain adaptation. *Advances in Neural Information Processing Systems*, 30, 2017.

Tristan Needham. A visual explanation of jensen's inequality. *The American Mathematical Monthly*, 100(8): 768–771, 1993.

Cuong Nguyen, Tal Hassner, Matthias Seeger, and Cedric Archambeau. Leep: A new measure to evaluate transferability of learned representations. In *International Conference on Machine Learning*, pp. 7294–7305. PMLR, 2020.

Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016.

Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 213–226. Springer, 2010.

Omer Sagi and Lior Rokach. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1249, 2018.

Zenglin Shi, Le Zhang, Yun Liu, Xiaofeng Cao, Yangdong Ye, Ming-Ming Cheng, and Guoyan Zheng. Crowd counting with deep negative correlation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5382–5390, 2018.

Rui Shu, Hung H Bui, Hirokazu Narui, and Stefano Ermon. A dirt-t approach to unsupervised domain adaptation. *arXiv preprint arXiv:1802.08735*, 2018.

Yang Shu, Zhangjie Cao, Ziyang Zhang, Jianmin Wang, and Mingsheng Long. Hub-pathway: transfer learning from a hub of pre-trained models. *Advances in Neural Information Processing Systems*, 35: 32913–32927, 2022.

Yu Su, Shiguang Shan, Xilin Chen, and Wen Gao. Hierarchical ensemble of global and local classifiers for face recognition. *IEEE Transactions on Image Processing*, 18(8):1885–1896, 2009.

Sanatan Sukhija, Narayanan C Krishnan, and Gurkanwal Singh. Supervised heterogeneous domain adaptation via random forests. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pp. 2039–2045, 2016.

Boyu Wang, Jorge Mendez, Mingbo Cai, and Eric Eaton. Transfer learning via minimizing the performance gap between domains. *Advances in neural information processing systems*, 32, 2019.

Shuo Wang, Huanhuan Chen, and Xin Yao. Negative correlation learning for classification ensembles. In *The 2010 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2010.

Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 499–515. Springer, 2016.

Guanglei Yang, Hao Tang, Zhun Zhong, Mingli Ding, Ling Shao, Nicu Sebe, and Elisa Ricci. Transformer-based source-free domain adaptation. *arXiv preprint arXiv:2105.14138*, 2021.

Shiqi Yang, Shangling Jui, Joost van de Weijer, et al. Attracting and dispersing: A simple approach for source-free domain adaptation. *Advances in Neural Information Processing Systems*, 35:5802–5815, 2022.

Moslem Yazdanpanah and Parham Moradi. Visual domain bridge: A source-free domain adaptation for cross-domain few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2868–2877, 2022.

Li Yi, Gezheng Xu, Pengcheng Xu, Jiaqi Li, Ruizhi Pu, Charles Ling, A Ian McLeod, and Boyu Wang. When source-free domain adaptation meets learning with noisy labels. *arXiv preprint arXiv:2301.13381*, 2023.

Kaichao You, Yong Liu, Jianmin Wang, and Mingsheng Long. Logme: Practical assessment of pre-trained models for transfer learning. In *International Conference on Machine Learning*, pp. 12133–12143. PMLR, 2021.

Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 289–305, 2018.