Ingénierie des Systèmes d'Information
Vol. 31, No. 2, February, 2026, pp. 389-399
Journal homepage: http://iieta.org/journals/isi

A Hybrid Adversarial Transfer Learning Framework with Feature Transformation for Effective Cross-Domain Adaptation

Brahim Belgroun, Abdelaali Bekhouche*

ICOSI Laboratory, Abbes Laghrour University, Khenchela 40001, Algeria

Corresponding Author Email: bekhouche.abdelali@univ-khenchela.dz

https://doi.org/10.18280/isi.310207

Received: 23 September 2025
Revised: 10 December 2025
Accepted: 18 February 2026
Available online: 28 February 2026

Keywords:
cross-domain transfer learning, adversarial learning, feature transformation, domain adaptation, deep learning, Maximum Mean Discrepancy

ABSTRACT

This study introduces Hybrid Adversarial Transfer with Feature Transformation (HATFT), a novel hybrid framework that combines adversarial training with feature transformation for effective cross-domain transfer learning in visual recognition tasks. Unlike traditional methods that only focus on aligning domain distributions, HATFT integrates a feature transformation layer that explicitly reduces discrepancies between source and target domains. This enhances feature discriminability through adversarial learning. Evaluated on the widely used Office-31 benchmark (Amazon, WebCam, DSLR), HATFT achieves state-of-the-art performance, with an average accuracy of 88.4%, surpassing previous methods like lapCNN (84.73%) and DAN (87.83%). Specifically, HATFT achieves 85.8% on A→W and 97.2% on A→D, demonstrating robust domain adaptation capabilities. The sharp convergence of Maximum Mean Discrepancy (MMD) loss, along with the stabilization of adversarial losses, further highlights the effectiveness of the integrated training mechanism. This demonstrates significant improvements in both stability and generalization, confirming that HATFT is highly effective for cross-domain transfer learning tasks. The results indicate that HATFT not only enhances feature transferability but also provides a promising solution for addressing domain shifts in real-world applications such as medical imaging, e-commerce, and remote sensing. The findings suggest that combining adversarial learning with feature transformation is an effective strategy for improving cross-domain adaptation.

1. INTRODUCTION

A key challenge in machine learning is the dependence on large volumes of labeled data to develop accurate and reliable models. However, in many real-world applications, acquiring large annotated datasets is often expensive, time-consuming, or impractical. As a result, working with small datasets becomes particularly difficult, since limited training samples may hinder the model's ability to capture meaningful patterns and properly represent the underlying structure of the problem.

To mitigate the limitations associated with small datasets, several strategies have been proposed, including transfer learning [1], active learning [2], and data augmentation [3]. Among these techniques, transfer learning has proven to be particularly powerful, as it allows models to exploit knowledge acquired from a related source domain to enhance performance in a target domain [1]. By relying on pre-trained models that have already leaned rich feature representations from large datasets, transfer learning allows models to adapt to new tasks with limited data through fine-tuning. This capability is especially important in scenarios where domain shifts occur, as discrepancies between training and testing distributions can significantly degrade model performance [4].

However, cross-domain transfer learning brings additional challenges because of inherent discrepancies between the source and target domains [5, 6]. These discrepancies, commonly referred to as the *domain gap*, may arise from distribution shifts, feature space misalignment, label space inconsistencies, or task heterogeneity. Addressing these issues has become a central focus in domain adaptation research. Recent advances have explored various strategies to bridge domain gaps, including adversarial domain adaptation, meta-learning, and domain generalization, each targeting different aspects of cross-domain knowledge transfer.

Adversarial domain adaptation has gained significant attention for its ability to learn domain-invariant feature representations through adversarial training. In this framework, a feature extractor is trained to generate representations that cannot be distinguished by a domain discriminator, thereby reducing the statistical distance between source and target domains. While this strategy has shown promising results, recent studies highlight important limitations. In particular, excessive alignment of feature distributions may inadvertently reduce discriminative information, leading to a trade-off between domain invariance and task-specific predictiveness [7]. Several extensions have been proposed to address these challenges. For example, centroIDA introduces class-level alignment through accumulative class-centroids, while GLA-DA employs global-local alignment strategies designed for multivariate time series data, addressing semantic sparsity issues when adapting labeled source samples to unlabeled target data [8, 9].

Beyond adversarial learning, recent research has explored alternative paradigms such as meta-learning and domain generalization. Meta-learning approaches train models to rapidly adapt to new domains using limited data, often through task-based training strategies that encourage models to learn transferable initialization parameters [10, 11]. Domain generalization methods aim to train models that generalize to unseen domains without direct access to target-domain data during training, making them particularly useful in safety-critical applications or environments where target-domain data are unavailable [12, 13]. These approaches highlight the growing interest in developing models capable of handling increasingly complex domain shifts.

Understanding the nature of domain gaps is essential for designing effective transfer learning systems. Different types of distribution shifts require different adaptation strategies. For example, covariate shift occurs when input distributions differ between domains while the conditional relationship between features and labels remains stable, whereas concept drift involves changes in the underlying decision boundaries [14]. More complex scenarios include heterogeneous domain adaptation, where feature spaces differ across domains [15], and multi-source domain adaptation, which leverages multiple source domains but may introduce risks of negative transfer if feature alignment is not carefully handled [16, 17].

Recent applications demonstrate the practical impact of these advanced transfer learning strategies across diverse domains. In building energy prediction, adversarial domain adaptation combined with spatiotemporal graph convolutional networks enables accurate cross-district energy forecasting despite limited data availability [18]. In medical imaging, test-time adaptation techniques allow models to dynamically adjust during deployment using limited target-domain information [19, 20]. Similarly, materials science applications have integrated small-data modeling strategies with large language model screening to accelerate the discovery of new materials despite severe data constraints [21]. These developments illustrate the growing importance of robust domain adaptation techniques for real-world machine learning applications.

Despite these advances, several limitations remain in existing domain adaptation methods. Distribution alignment approaches based on discrepancy measures such as Maximum Mean Discrepancy (MMD) focus primarily on reducing statistical divergence between domains but may struggle to capture complex nonlinear transformations when domain shifts are substantial. Adversarial domain adaptation methods, while powerful, can suffer from training instability and may not always guarantee effective alignment of high-level semantic features. Furthermore, many existing approaches treat adversarial learning and feature transformation as independent components, rather than integrating them into a unified framework that explicitly models how feature representations should be adapted across domains.

To address these challenges, this paper proposes a Hybrid Adversarial Transfer with Feature Transformation (HATFT) framework for cross-domain transfer learning. The proposed approach integrates adversarial learning with an explicit feature transformation module and discrepancy-based alignment, enabling the model to simultaneously reduce distribution divergence and adapt feature representations across domains. By combining these complementary strategies within a unified training framework, HATFT aims to improve both the stability and effectiveness of cross-domain knowledge transfer.

The main contributions of this paper are summarized as follows:

- A novel HATFT framework is proposed, integrating adversarial learning with a feature transformation module to reduce domain discrepancies and enhance cross-domain generalization.
- A dual-objective optimization strategy is introduced, combining adversarial loss with MMD to simultaneously achieve domain invariance and feature alignment.
- An analytical study is presented to illustrate how the proposed architecture reduces both feature-space discrepancies and distribution divergence between source and target domains.
- Extensive experiments conducted on the Office-31 dataset (Amazon, DSLR, and Webcam domains) demonstrate the effectiveness of HATFT, achieving an average accuracy of 88.4% and outperforming several state-of-the-art methods including lapCNN and DAN.

2. TRANSFER LEARNING

Transfer learning is a method that exploits knowledge learned from a source domain $D_S = \{X_S, P_S\}$ to enhance learning performance in a target domain $D_T = (X_T, P_T)$ [22]. In this formulation, X denotes the feature space, while P denotes the marginal probability distribution. A specific task T is defined by a label space Y, a prior distribution P(Y), and a conditional probability distribution P(Y | X). This conditional probability is typically learned from training data, consisting of input-label pairs $x_i \in X$ and $y_i \in Y$ [22, 23].

The transfer learning process can be formally represented as:

$$M_T = \text{transfer}(M_S, T_S, D_S, T_T, D_T) \quad (1)$$

where, M_S refers to the model initially trained on the source task T_S within the source domain D_S. The transformed model M_T is then adapted to perform a new target task T_T in the target domain D_T.

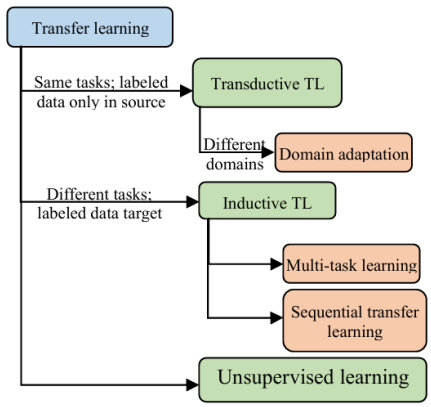

Figure 1. Taxonomy of transfer learning approaches, including transductive, inductive, and unsupervised categories

The primary goal of transfer learning is to estimate the target conditional probability distribution $P(Y_T \mid X_T)$ in the target domain D_T by utilizing the information acquired from

the source domain D_S and task T_S. This approach is particularly relevant when either the feature spaces differ ($D_S \neq D_T$) or the tasks are distinct ($T_S \neq T_T$). In practice, this often assumes the presence of a limited set of labeled samples from the target domain or a larger dataset of unlabeled target examples.

Based on how labeled data is used, transfer learning can be broadly classified into three main categories: transductive, inductive, and unsupervised transfer learning [24].

Figure 1 illustrates this classification, providing a clear taxonomy of transfer learning approaches.

2.1 Transductive transfer learning

In transductive transfer learning, the model addresses the same fundamental task across both the source and target domains; however, the data distributions between these domains vary [25]. The primary objective is to manage the domain shift, ensuring that the model can generalize effectively to the target domain despite discrepancies in distribution [26].

This form of transfer learning encompasses two main scenarios:

- Different Marginal Distributions: When $P_S(X_S) \neq P_T(X_T)$, indicating that the marginal probability distributions of the source and target domains are distinct. This scenario is typically referred to as domain adaptation [23, 25].
- Different Feature Spaces: When $X_S \neq X_T$, meaning that the feature spaces of the source and target domains differ.

2.2 Inductive transfer learning

In inductive transfer learning, the model has access to label information for instances within the target domain [22, 25]. This approach can be categorized into three distinct scenarios [23]:

- Different Prior Distributions: When $P_S(Y_S) \neq P_T(Y_T)$, meaning that the prior distributions of the source and target tasks differ, resulting in variations in label distributions across datasets [27, 28].
- Different Conditional Distributions: When $P_S(Y_S \mid X_S) \neq P_T(Y_T \mid X_T)$, indicating that the conditional probability distributions between the source and target tasks are not aligned. This situation arises when the class distributions in the source and target datasets are imbalanced [23]. In practice, this issue is often addressed using techniques such as over-sampling, under-sampling, or Synthetic Minority Over-sampling Technique (SMOTE) [29].
- Different Label Spaces: When $Y_S \neq Y_T$, meaning the source and target tasks have distinct label spaces, requiring different label assignments for the target domain. A key distinction here is whether tasks are learned simultaneously or sequentially. Learning tasks concurrently falls under multi-task learning, while addressing tasks one after another is referred to as sequential transfer learning [23].

2.3 Unsupervised transfer learning

Unsupervised transfer learning focuses on utilizing knowledge from a source domain D_S that lacks labeled data to enhance learning in a target domain D_T, where labeled examples are limited or entirely absent. The objective is to facilitate knowledge transfer from D_S to D_T without relying on labeled data from the target domain [30].

3. RELATED WORK

Cross-domain transfer learning has become an essential approach for enabling machine learning models to generalize across heterogeneous data distributions. Early research in this area primarily relied on shallow transfer methods, where handcrafted features and statistical measures were used to reduce the discrepancy between source and target domains. Classical techniques such as Transfer Component Analysis (TCA) and Geodesic Flow Kernel (GFK) attempted to align domain distributions through subspace projection and statistical matching [31]. Although these methods demonstrated promising results, their reliance on handcrafted representations limited their ability to capture complex semantic structures in high-dimensional data.

With the rapid development of deep learning, research gradually shifted toward deep domain adaptation methods that integrate representation learning with transfer mechanisms. These approaches leverage neural networks to learn high-level feature representations that are more transferable across domains. A major breakthrough was the introduction of adversarial domain adaptation frameworks. For example, Tzeng et al. [32] and Ganin et al. [33] proposed architectures in which a domain discriminator is trained to distinguish between source and target feature distributions, while a feature generator learns domain-invariant representations by attempting to fool the discriminator. This adversarial learning paradigm significantly improved cross-domain feature alignment and inspired many subsequent studies. To further enhance distribution alignment, several works incorporated statistical discrepancy measures such as MMD and correlation alignment to enforce closer matching between domain distributions [33].

More recent studies have explored hybrid and multi-level adaptation strategies that combine adversarial learning with additional alignment objectives. Long et al. [34] introduced a deep residual learning framework for domain adaptation that improves feature transferability across deep network layers. Zhang et al. [35] proposed adaptive adversarial normalization to dynamically align feature statistics during training, enabling better adaptation under distribution shifts. Similarly, Dayal et al. [36] presented the MDAMA framework, an adversarial multi-source domain adaptation method that aligns the target domain with a mixture distribution derived from multiple source domains, achieving improved performance in heterogeneous environments.

Beyond adversarial approaches, alternative learning paradigms such as meta-learning and contrastive learning have been investigated to improve cross-domain generalization. Finn et al. [37] introduced Model-Agnostic Meta-Learning (MAML), which enables models to rapidly adapt to new tasks or domains using only a few training samples. Building on this idea, Wang et al. [38] proposed a domain-generalization meta-learning framework that enhances robustness under extreme domain shifts by optimizing models for adaptability across diverse training domains.

Domain adaptation techniques have also demonstrated strong potential in real-world applications. In computer vision,

Yosinski et al. [39] and Oquab et al. [40] showed that features learned by deep convolutional neural networks on large-scale datasets can be effectively transferred to different visual recognition tasks. In industrial and Internet-of-Things (IoT) applications, Qin and Zhao [41] applied domain adaptation for battery health estimation, while Chen et al. [42] utilized cross-domain transfer learning to improve predictive performance in smart manufacturing systems.

Despite these advances, several challenges remain. Many adversarial domain adaptation methods suffer from training instability and imbalance between the generator and discriminator objectives, which can hinder effective feature alignment. In addition, discrepancy-based approaches may struggle to capture complex nonlinear relationships when the domain gap becomes large. Furthermore, existing methods often focus on either distribution alignment or feature transformation independently, limiting their ability to achieve robust cross-domain generalization.

To address these limitations, this study proposes a Hybrid Adversarial Transfer Learning with Feature Transformation (HATFT) framework. The proposed approach integrates adversarial domain adaptation with explicit feature transformation mechanisms, enabling simultaneous distribution alignment and feature consistency across domains. By combining these complementary strategies within a unified architecture, the HATFT model improves training stability and enhances the transferability of learned representations, leading to more reliable performance under significant domain shifts.

4. PROPOSED METHOD

In this research, we propose a novel hybrid approach that integrates adversarial training with a feature transformation layer to improve the alignment of feature spaces between domains. This method, referred to as HATFT, combines the advantages of adversarial learning with an innovative feature transformation layer to establish a robust and efficient framework for cross-domain transfer learning.

The core concept of HATFT involves leveraging adversarial training to reduce domain discrepancies by encouraging the feature generator to produce features that are invariant across domains. Concurrently, the feature transformation layer explicitly aligns the feature distributions of the source and target domains, effectively narrowing the domain gap. This dual approach ensures that the extracted features are both domain-invariant and well-aligned, enhancing the model's capacity to generalize across different domains.

4.1 Overview of Hybrid Adversarial Transfer with Feature Transformation

The HATFT architecture consists of four main components: a feature generator *G*, a feature transformation layer *T*, a domain discriminator *D*, and a task-specific classifier *H*, as illustrated in Figure 2. Each component plays a specific role in the domain adaptation process and is associated with a corresponding objective during training.

The feature generator *G* extracts high-level feature representations from both source and target domain inputs. These features are then passed through the feature transformation layer *T*, which learns to adapt the representations in order to better align the source and target feature distributions. The transformed features are subsequently provided to the domain discriminator *D*, which attempts to distinguish whether the features originate from the source or the target domain. Through this adversarial interaction, the generator is encouraged to produce domain-invariant representations.

To train the model, multiple loss functions are jointly optimized. The adversarial loss between the generator G and the discriminator D encourages the extraction of domain-invariant features, while the transformation loss L_T is specifically applied to the feature transformation layer T to regulate the alignment of feature representations.

As defined in Eq. (9), the overall training objective integrates these components through a weighted combination of losses:

$$L_{total} = L_G + L_D + \lambda L_T \quad (2)$$

where, L_G and L_D correspond to the adversarial objectives of the generator and discriminator, respectively, and L_T represents the transformation loss associated with the feature transformation layer T. The weighting parameter λ controls the contribution of the transformation loss to the overall optimization objective.

During training, the parameters of *G*, *T*, and *D* are updated jointly through backpropagation, ensuring that the generator learns domain-invariant features while the transformation layer explicitly aligns feature distributions between domains. As illustrated in Figure 2, this joint optimization allows the architecture to integrate adversarial learning with feature transformation in a unified framework.

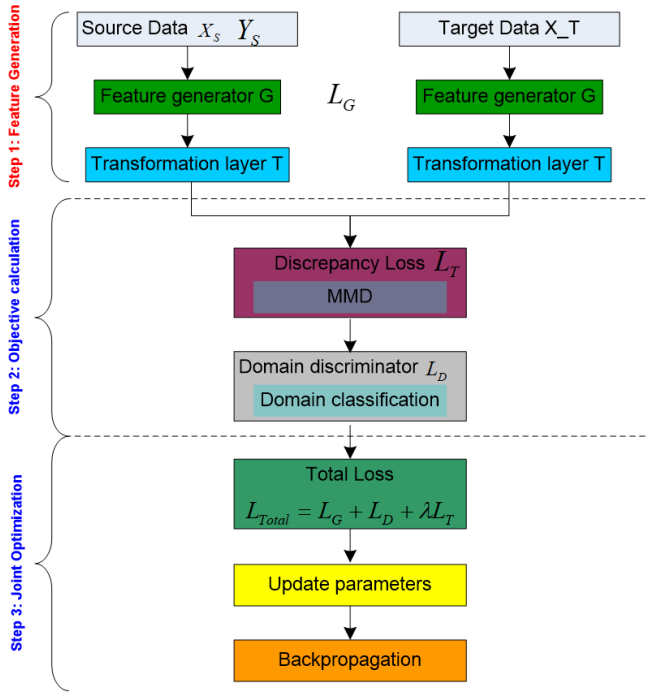

Figure 2. Architecture of the Hybrid Adversarial Transfer with Feature Transformation (HATFT) framework, showing the feature generator (G), feature transformation layer (T), domain discriminator (D), and task-specific head (H)

4.1.1 Feature generator (G)

The feature generator layer (G) is driven by the source domain data X_S, which comprises labeled pairs (x_s, y_s), and

the target domain data X_T, consisting of unlabeled examples (X_t).

By encoding the input data X_S and X_T into feature vectors $f_s = G(x_s)$ and $f_t = G(x_t)$ respectively, the feature generator seeks to develop domain-invariant features that form the basis for subsequent processing. Achieving domain invariance is essential for effective transfer learning, enabling the model to utilize knowledge from the source domain and apply it to the target domain despite potential differences in data distribution.

The ability of the feature generator to produce robust and adaptable features is key to the success of the HATFT method, as it helps minimize domain discrepancies and promotes better cross-domain generalization. The feature generator is structured with multiple layers and incorporates non-linear activation functions. Initially, the first layer h_1 processes the raw input data X, with each subsequent layer h_1 building on the output of the previous one. The final output layer h_L produces the result f= h_L. The feature generation process can be mathematically defined as follows:

$$\begin{cases} h_1 = \sigma(W_1 x + b_1) \\ h_l = \sigma(W_l h_{l-1} + b_l)\ for\ l = 2,3,\dots,L \end{cases} \quad (3)$$

where, W_1 and b_1 represent the weights and biases of the l^{th} layer, respectively. And σ denotes a non-linear activation function applied at each layer.

The feature generation process follows the steps outlined in the algorithm below.

Inputs:

Source domain data $X_S = \{x_{s_i}, y_{s_i}\}_{i=1}^{N_S}$

Target domain data $X_T = \{x_{T_j}\}_{j=1}^{N_T}$

Learning rates: $\eta G, \eta D, \eta T$

Balancing parameter λ

Initialize:

Neural network parameters for feature generator G, feature transformation layer T, and discriminator D

Training Loop:

For each training iteration **do**:

Feature Generation:

Source features: $f_{s_i} = G(x_{s_i}, \forall\, x_{s_i} \in X_S$

Target features: $f_{t_j} = G(x_{t_j}, \forall\, x_{t_j} \in X_T$

Feature Transformation:

Transform source features: $\hat{f}_{s_i} = T(f_{s_i})$

Transform target features: $\hat{f}_{t_j} = T(f_{t_j})$

Compute Losses:

Qescrimminator loss:

$$L_D = BCE\left(D\left(\hat{f}_{s_i}\right), 1\right) + BCE\left(D\left(\hat{f}_{t_j}\right), 0\right)$$

MMD/Transformation loss: $L_T = MMD\left(\hat{f}_{s_i}, \hat{f}_{t_j}\right)$

Total loss for G and T: $L_{total} = L_G + L_D + \lambda L_T$

Update Networks:

Update discriminator D parameters by minimizing L_{adv}

Update feature generator G parameters by minimizing L_{total}

Update feature transformation layer T parameters by minimizing L_{total}

End For

Outputs:

Transformed source features $\hat{F}_S = \{\hat{f}_{s_i}\}$

Transformed target features $\hat{F}_T = \{\hat{f}_{t_j}\}$

4.1.2 Feature transformation (T)

The feature transformation layer is a crucial component of the HATFT architecture, specifically designed to align the feature distributions of the source and target domains. This layer operates on the feature vectors f_s and f_t generated by the feature generator G, applying a series of transformations to make their distributions more similar. By addressing domain discrepancies that typically hinder transfer learning effectiveness, this layer helps minimize the domain gap.

This transformation not only reduces domain differences but also improves the model's robustness, enhancing its ability to generalize when exposed to unseen data from the target domain. The aligned features significantly boost the overall performance of the HATFT method by making transferred knowledge from the source domain more applicable and effective within the target domain.

The transformation function T involves multiple layers, optimized to minimize discrepancies between the source and target feature distributions. These layers may incorporate fully connected layers, normalization layers, and non-linear activation functions.

The transformation process can be defined as:

$$\begin{cases} t_1 = \sigma(W_1 f + b_1) \\ t_k = \sigma(W_k t_{k-1} + b_k)\ \text{for}\ k = 2,3,\dots,K \end{cases} \quad (4)$$

where, W_k and b_k represent the weights and biases of the k^{th} layer. And σ denotes a non-linear activation function.

The final transformed feature $f' = t_K$ corresponds to the output from the last layer.

4.1.3 Domain discriminator (D)

The domain discriminator D is responsible for distinguishing between features derived from the source domain and those from the target domain. Its main objective is to encourage the learning of domain-invariant features through adversarial training. The domain discriminator processes the transformed feature representations $F(f_s)$ from the source domain and $F(f_t)$ from the target domain, attempting to classify them accurately as either source or target domain features.

During training, the feature generator G is simultaneously trained to confuse the domain discriminator, making it harder to correctly differentiate between the two domains. This adversarial interaction minimizes domain-specific characteristics in the feature representations, pushing the generator to produce more domain-invariant features.

As a result, the model becomes better at generalizing across domains since the learned features become less dependent on specific traits from either the source or target domain. The domain discriminator's ability to drive this adversarial process is crucial for minimizing domain discrepancies and enhancing the overall effectiveness of the transfer learning framework.

The output of the domain discriminator D(f) represents the probability that the feature f originates from the source domain. Its objective function is to maximize the accuracy of domain classification for the input features.

The feature generator (G), on the other hand, aims to minimize the discriminator's ability to distinguish between the source and target features.

For the domain discriminator loss function L_D, we implement a hinge loss to encourage margin-based separation. This approach pushes the discriminator to output values closer to +1 for source domain features and -1 for target domain features, enhancing its ability to effectively distinguish between the two domains. The loss function for the domain

discriminator is defined as:

$$L_D = -E_{x_s \sim X_S}\left[\min\left(0, -1 + D\left(G(x_s)\right)\right)\right] - E_{x_t \sim X_T}\left[\min\left(0, -1 - D\left(G(x_t)\right)\right)\right] \quad (5)$$

This formulation penalizes incorrect classifications by enforcing a margin, encouraging stronger domain separation in the learned feature space.

For the feature generator loss function L_G, instead of the conventional adversarial loss, we use a feature matching loss. This loss minimizes the L_2 distance between the mean feature representations of the source and target domains. The objective is to directly align the statistical properties of the two domains, encouraging the generator to produce domain-invariant features. This method fosters better feature generalization and can significantly improve domain adaptation performance. The loss function is defined as:

$$L_G = \left\| E_{x_s \sim X_S}[G(x_s)] - E_{x_t \sim X_T}[G(x_t)] \right\|_2^2 \quad (6)$$

4.1.4 Task specific head

The role of the task-specific head (H) in processing transformed features to generate predictions relevant to a given task. Typically, this head is made up of fully connected layers followed by an output layer, with its architecture designed according to the specific nature of the task. For classification tasks, the output layer usually applies a softmax activation function to produce probabilities for each class, while for regression tasks, a linear activation function is used to generate continuous values. In the HATFT model, the task-specific head works in conjunction with the feature generator and feature transformation layer to ensure that the final predictions are based on domain-invariant features that are well-aligned between the source and target domains. This alignment effectively transfers knowledge from the source domain to the target domain, improving performance on the target task.

4.2 Theoretical analysis of Hybrid Adversarial Transfer with Feature Transformation

This section introduces the theoretical foundation of HATFT in the context of cross-domain transfer learning. The main objective is to transfer knowledge from a source domain (X_S, Y_S) to a target domain (X_T, Y_T), where the data distributions $P(X_S)$ and $P(X_T)$ are different. The key challenge lies in aligning the feature distributions $P(F_S)$ and $P(F_T)$ to ensure that a model trained on the source domain X_S can generalize effectively to the target domain X_T. This section presents a theoretical analysis of how HATFT addresses this challenge and enhances cross-domain transfer learning, supported by relevant mathematical formulations.

4.2.1 Adversarial training

The objective of adversarial training is to encourage the feature generator G to learn domain-invariant features through the use of a domain discriminator D. Here, D aims to differentiate between features extracted from the source and target domains, while G seeks to deceive D by producing features that are indistinguishable across domains.

In this updated framework, we introduce modified loss functions for both the discriminator and the generator to better enhance domain alignment.

The goal is now to minimize L_G while maximizing L_D, forming a revised min-max optimization problem:

$$\min_G \max_D (G, D) = L_G + L_D \quad (7)$$

At equilibrium, the discriminator D becomes unable to distinguish between $G(x_t)$ and $G(x_s)$, meaning $P(G(x_s)) = P(G(x_t))$. This equilibrium ensures that the learned features are domain-invariant, thereby improving the model's ability to generalize across domains.

4.2.2 Feature transformation layer

The feature transformation layer (T) further aligns the feature distributions by transforming the feature vectors $f_s = G(x_s)$ and $f_t = G(x_t)$ to minimize their discrepancy. It uses the following discrepancy loss function:

$$L_T = MMD\left(T(F_S), T(F_T)\right) \quad (8)$$

where, *MMD* is the Maximum Mean Discrepancy, a measure of the distance between two distributions. The overall objective of HATFT is to minimize the adversarial loss and the discrepancy loss:

$$\min_G \max_D (G, D) + \lambda L_T \quad (9)$$

Here, λ is a hyperparameter balancing the adversarial and discrepancy losses. Minimizing MMD ensures that the feature distributions are aligned, reducing the domain gap. This combined objective balances adversarial training and feature transformation, leading to improved cross-domain generalization.

In summary, HATFT leverages adversarial training to create domain-invariant features and a feature transformation layer to align feature distributions, thereby improving the effectiveness of cross-domain transfer learning.

5. EXPERIMENTS

This section presents the experimental setup designed to rigorously evaluate the performance of the HATFT method. We aim to demonstrate the efficacy of HATFT in enhancing cross-domain transfer learning by simultaneously minimizing domain discrepancy and optimizing adversarial alignment. Specifically, we assess the impact of the updated adversarial loss L_D and generator loss L_G in reducing the domain gap and improving model generalization.

This thorough and systematic approach ensures the validity and reliability of our findings, highlighting the practical advantages of the HATFT method in achieving robust and domain-invariant feature representations.

5.1 Datasets description

In this study, we utilized the Office-31 dataset, a widely recognized benchmark in cross-domain transfer learning. The dataset consists of three distinct subsets: (1) The Amazon dataset, which contains 2,817 product images sourced from the Amazon website. These images exhibit significant variability

in background and lighting conditions across 31 categories. (2) The DSLR dataset, comprising 498 high-quality images captured with a DSLR camera. These images maintain consistent lighting and background conditions, also spanning 31 categories. (3) The Webcam dataset, which includes 795 images taken with a low-resolution webcam. These images often display lower quality and higher noise levels, yet still cover the same 31 categories.

In our experiments, we used the Amazon dataset as the source domain for both the DSLR and Webcam datasets. Additionally, we used the DSLR dataset as the source domain for the Webcam dataset.

5.2 Implementation details

To ensure reproducibility of the proposed HATFT framework, we provide detailed implementation settings, including network architecture, hyperparameters, optimization strategy, and computational environment.

The feature generator G was implemented using a neural network consisting of three fully connected layers with [256, 128, 64] neurons respectively. The Rectified Linear Unit (ReLU) activation function was used in all hidden layers to introduce non-linearity, while a softmax activation was applied in the output classification layer.

The feature transformation layer T was constructed using two fully connected layers with batch normalization to stabilize training and improve distribution alignment between source and target domains.

The domain discriminator D consisted of two fully connected layers with 128 and 64 neurons, followed by a sigmoid activation function to output domain classification probability.

The trade-off parameter λ in Eq. (9) was selected empirically through grid search over the set $\{0.1, 0.3, 0.5, 0.7\}$. The best performance was obtained when $\lambda = 0.5$, which provided a good balance between adversarial learning and feature transformation loss.

The model was trained using the Adam optimizer with an initial learning rate of 1e-4. The learning rate was reduced by a factor of 0.1 if validation loss did not improve for 5 consecutive epochs. Training was conducted for 50 epochs with a batch size of 32.

Average training time per experiment was approximately 4–5 hours depending on domain transfer task.

5.3 Results and discussion

The results are presented in two sections: (1) The first section examines the training losses of both the discriminator and generator in our transfer learning model. (2) The second section compares the classification performance of the retrained model on the target domain dataset with previous works from the literature.

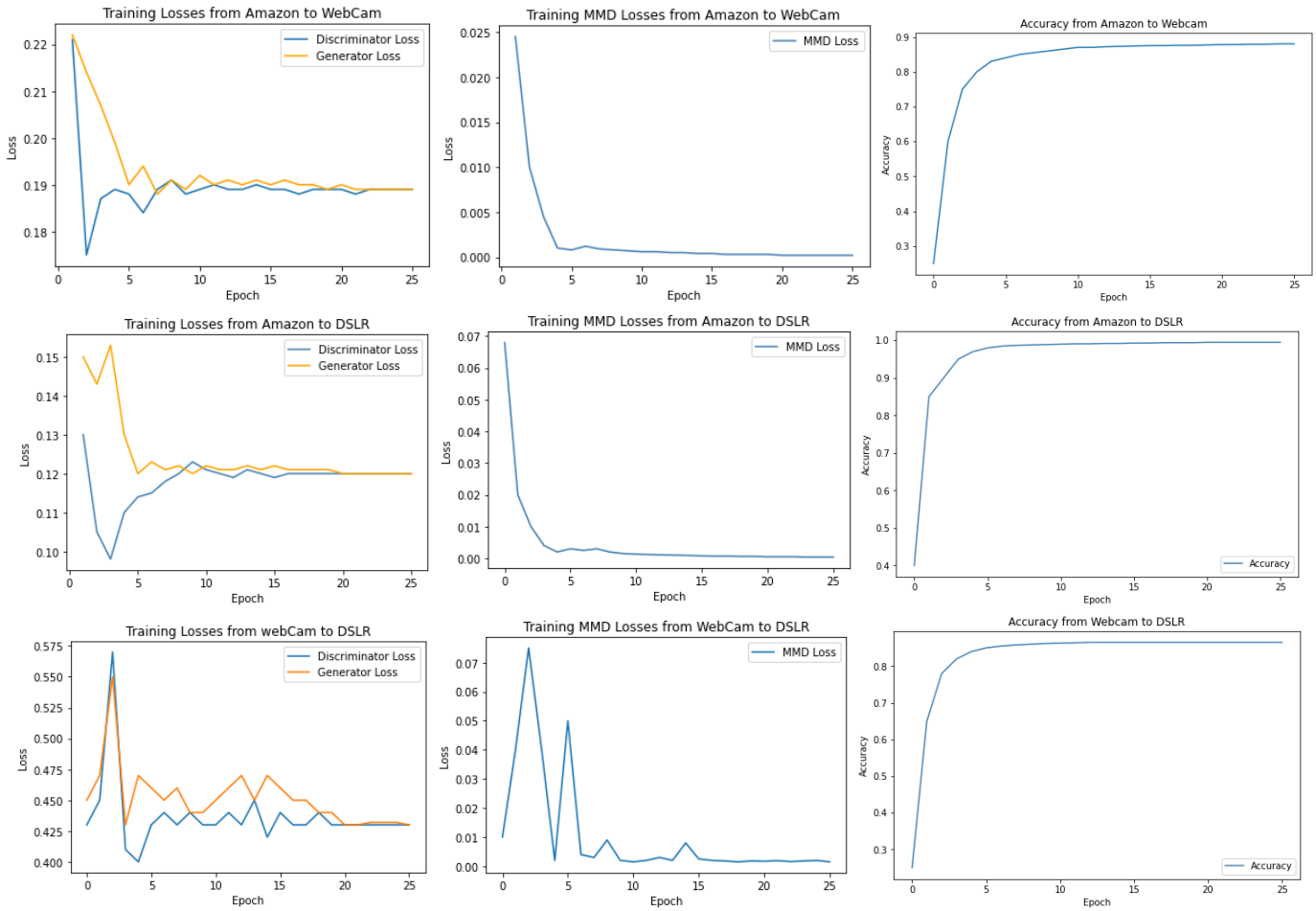

Figure 3. Performance analysis of Hybrid Adversarial Transfer with Feature Transformation (HATFT) across Office-31 domain shifts: Comparative evaluation of adversarial losses, Maximum Mean Discrepancy (MMD) discrepancy, and classification accuracy

5.3.1 Adversarial learning process results

The adversarial learning process of the proposed transfer learning model is illustrated in Figure 3. The generator loss initially starts at a relatively high value but decreases rapidly during the early training epochs, indicating that the feature generator quickly improves its ability to extract domain-invariant representations. Meanwhile, the discriminator loss starts lower but gradually increases before stabilizing, reflecting the adversarial competition between the generator and discriminator during training.

The stabilization of both generator and discriminator losses after approximately ten epochs suggests that the adversarial optimization process has reached a balanced state. This equilibrium indicates that the generator is producing features that are difficult for the discriminator to distinguish, while the discriminator maintains sufficient discriminative capability without dominating the training process.

The MMD loss exhibits a sharp decline during the initial training phase and approaches near-zero values as training progresses. This behavior reflects the effectiveness of the feature transformation layer T, which explicitly minimizes statistical divergence between source and target feature distributions. The rapid reduction of MMD loss suggests that the transformation mechanism successfully captures shared structural characteristics across domains, facilitating efficient domain alignment.

The convergence behavior observed in the loss curves is consistent with the design objective of the HATFT framework, where adversarial learning promotes representation robustness and the transformation layer enforces distribution-level alignment. Minor fluctuations observed in certain transfer directions, such as Amazon→DSLR and Webcam→DSLR, may be attributed to variations in domain complexity and noise characteristics across datasets.

The convergence behavior observed in the loss curves is consistent with the design objective of the HATFT framework, where adversarial learning promotes representation robustness and the transformation layer enforces distribution-level alignment. Minor fluctuations observed in certain transfer directions, such as Amazon→DSLR and Webcam→DSLR, may be attributed to variations in domain complexity and noise characteristics across datasets.

The convergence behavior observed in the loss curves is consistent with the design objective of the HATFT framework, where adversarial learning promotes representation robustness and the transformation layer enforces distribution-level alignment. Minor fluctuations observed in certain transfer directions, such as Amazon→DSLR and Webcam→DSLR, may be attributed to variations in domain complexity and noise characteristics across datasets.

It is also worth noting that the Office-31 dataset exhibits moderate domain shift complexity, which may contribute to the rapid convergence of discrepancy loss. The combination of adversarial training and explicit feature distribution alignment allows the HATFT model to efficiently reduce domain gap even when training data are limited.

Overall, the training dynamics demonstrate that the hybrid integration of adversarial learning and feature transformation contributes to stable optimization and effective cross-domain feature adaptation.

5.3.2 Classification performances of Hybrid Adversarial Transfer with Feature Transformation

The results presented in Table 1 show that HATFT achieves outstanding performance compared to existing methods on the Office-31 dataset. It obtains the highest accuracies on the A→D and A→W tasks (97.2% and 85.8%, respectively), surpassing all other approaches, including state-of-the-art methods such as DAN and lapCNN.

For the W→D task, HATFT achieves an accuracy of 82.2%, which is slightly lower than its performance on other transfer directions, indicating potential for further improvement in this domain shift. Nevertheless, this result remains competitive and superior to traditional methods like TCA and GFK.

Overall, HATFT reaches an average accuracy of 88.4%, demonstrating strong and balanced performance across all transfer tasks. These findings confirm that the proposed method offers a robust and effective solution for domain adaptation, benefiting from its innovative feature transformation and transfer learning strategies.

Table 1. Accuracy on Office-31 dataset with some methods of literature [43]

Method	A→W	A→D	W→D	Average
TCA	21.5±0.0	50.1±0.0	58.4±0.0	43.33
GFK	19.7±0.0	49.7±0.0	63.1±0.0	44.17
CNN	61.6 ± 0.5	95.4±0.3	99.0±0.2	85.33
lapCNN	60.4±0.3	94.7±0.5	99.1±0.2	84.73
DDC	61.8 ± 0.4	95.0±0.5	98.5±0.4	85.1
DAN	68.5 ± 0.4	96.0±0.3	99.0±0.2	87.83
DAN_7	63.2±0.2	94.8±0.4	98.9±0.3	85.63
DAN_8	63.8 ± 0.4	94.6±0.5	98.8±0.6	85.73
DAN_{SK}	63.3±0.3	95.6±0.2	99.0±0.4	85.96
HATFT	85.8	97.2	82.2	88.4

5.3.3 Ablation study

To evaluate the contribution of each core component of the proposed HATFT framework, an ablation study was conducted by removing key modules from the architecture. Specifically, we compared the full HATFT model with two variant configurations: (1) HATFT without the feature transformation layer (HATFT w/o T), and (2) HATFT without adversarial training (HATFT w/o Adv).

The results, summarized in Table 2, show that removing either component leads to a significant reduction in transfer performance across all domain adaptation tasks. The full HATFT achieves an average accuracy of 88.4%, outperforming HATFT w/o T (73.57%) and HATFT w/o Adv (65.67%). These results clearly demonstrate that both feature transformation and adversarial learning contribute substantially to the effectiveness of the model.

Table 2. Accuracy comparison between Hybrid Adversarial Transfer with Feature Transformation (HATFT), HATFT w:o T, and HATFT w/o Adv

Method	A→W	A→D	W→D	Average
HATFT	85.8	97.2	82.2	88.4
HATFT w/o T	71.6	82.4	66.7	73.57
HATFT w/o Adv	63.2	70.1	63.7	65.67

Figure 4 illustrates the training dynamics of the ablation variants. The full HATFT demonstrates faster convergence and lower final loss compared with the two simplified variants. Removing the feature transformation layer leads to slower convergence, while removing adversarial learning substantially reduces feature alignment and overall performance. Overall, the hybrid combination of adversarial

learning and feature transformation proves essential for effective domain adaptation, delivering superior performance across all transfer tasks.

Figure 4 illustrates the training dynamics of the ablation variants. The full HATFT model demonstrates faster convergence and lower final loss compared with the two simplified variants. Removing the feature transformation layer leads to slower convergence, while removing adversarial learning significantly reduces feature alignment effectiveness. These results confirm that the hybrid combination of adversarial learning and feature transformation contributes to improved domain adaptation.

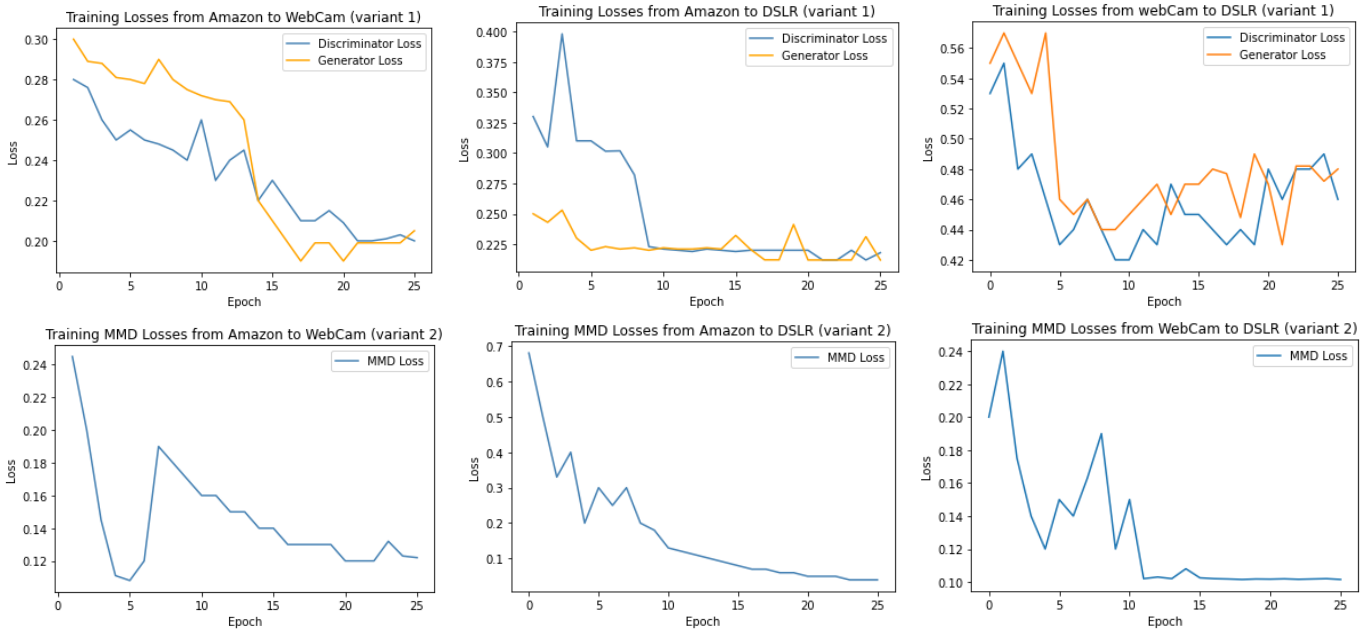

Figure 4. Ablation study of Hybrid Adversarial Transfer with Feature Transformation (HATFT): Impact of feature transformation and adversarial learning on training dynamics and final losses

6. CONCLUSION

In this paper, we proposed HATFT, a hybrid adversarial and feature transformation framework for cross-domain transfer learning. By combining a generator-discriminator architecture with MMD loss, HATFT effectively aligns source and target domain features, enabling robust feature transfer. Experiments on the Amazon, WebCam, and DSLR datasets demonstrated strong performance, with HATFT achieving 85.8% accuracy for Amazon→WebCam, 97.2% for Amazon→DSLR, and 82.2% for WebCam→DSLR, outperforming several state-of-the-art methods and achieving an average accuracy of 88.4%.

Theoretically, this work highlights that integrating adversarial alignment with feature transformation provides a stable and generalizable strategy for hybrid domain adaptation, offering a design principle for future model development. Practically, HATFT's architecture can be extended to other domain-shift scenarios, such as medical imaging, industrial defect detection, remote sensing, or real-time applications where labeled data are limited.

For future research, we recommend exploring multi-source adaptation, incorporating Vision Transformer-based feature generators, or leveraging self-supervised learning to further enhance cross-domain generalization. Evaluating HATFT on larger and more diverse datasets could also provide deeper insights into its robustness and scalability.

Overall, HATFT demonstrates a practical and theoretically grounded approach to addressing domain shift in visual recognition, offering guidance for model design and inspiration for future advances in domain adaptation.

REFERENCES

[1] Su, J., Yu, X., Wang, X., Wang, Z., Chao, G. (2024). Enhanced transfer learning with data augmentation. Engineering Applications of Artificial Intelligence, 129: 07602. https://doi.org/10.1016/j.engappai.2023.107602

[2] Lanjewar, M.G., Parab, J.S. (2024). CNN and transfer learning methods with augmentation for citrus leaf diseases detection using PaaS cloud on mobile. Multimedia Tools and Applications, 83(11): 31733-31758. https://doi.org/10.1007/s11042-023-16886-6

[3] Xu, Y., Liu, T., Yang, Y., Kang, J., Ren, L., Ding, H., Zhang, Y. (2024). ACVPred: enhanced prediction of anti-coronavirus peptides by transfer learning combined with data augmentation. Future Generation Computer Systems, 160: 305-315. https://doi.org/10.1016/j.future.2024.06.008

[4] Alhussan, A.A., Abdelhamid, A.A., Towfek, S.K., Ibrahim, A., Abualigah, L., Khodadadi, N., Khafaga, D.S., Al-Otaibi S., Ahmed, A.E. (2023). Classification of breast cancer using transfer learning and advanced al-biruni earth radius optimization. Biomimetics, 8(3): 270. https://doi.org/10.3390/biomimetics8030270

[5] Gardner, J., Popovic, Z., Schmidt, L. (2023). Benchmarking distribution shift in tabular data with tableshift. Advances in Neural Information Processing Systems, 36: 53385-53432.

[6] Arjovsky, M., Chintala, S., Bottou, L. (2017). Wasserstein GAN. arXiv:1701.07875. https://doi.org/10.48550/arXiv.1701.07875

[7] Dan, J., Liu, M., Xie, C., Yu, J., Xie, H., Li, R., Dong, S.

(2024). Similar norm more transferable: Rethinking feature norms discrepancy in adversarial domain adaptation. Knowledge-Based Systems, 296: 111908. https://doi.org/10.1016/j.knosys.2024.111908

[8] Sun, X., Wu, Z., Ji, Y., Zhan, Z. (2024). centroIDA: Cross-domain class discrepancy minimization based on accumulative class-centroids for Imbalanced Domain Adaptation. Expert Systems with Applications, 255: 124718. https://doi.org/10.1016/j.eswa.2024.124718

[9] Tu, G., Li, D., Lin, B., Zheng, Z., Ng, S.K. (2024). GLA-DA: Global-local alignment domain adaptation for multivariate time series. arXiv:2410.06671. https://doi.org/10.48550/ARXIV.2410.06671

[10] Wu, J., Zhang, T., Zhang, Y. (2024). Hybridprompt: Domain-aware prompting for cross-domain few-shot learning. International Journal of Computer Vision, 132(12): 5681-5697. https://doi.org/10.1007/s11263-024-02086-8

[11] Wang, J., Qiang, W., Su, X., Zheng, C., Sun, F., Xiong, H. (2024). Towards task sampler learning for meta-learning. International Journal of Computer Vision, 132(12): 5534-5564. https://doi.org/10.1007/s11263-024-02145-0

[12] Atghaei, A., Rahmati, M. (2025). Domain generalization via geometric adaptation over augmented data. Knowledge-Based Systems, 309: 112765. https://doi.org/10.1016/j.knosys.2024.112765

[13] Zhao, C., Zio, E., Shen, W. (2024). Domain generalization for cross-domain fault diagnosis: An application-oriented perspective and a benchmark study. Reliability Engineering & System Safety, 245: 109964. https://doi.org/10.1016/j.ress.2024.109964

[14] Yi, C.A., Chen, H., Xu, Y., Zhou, Y., Du, J., Cui, L., Tan, H. (2024). TADA: Temporal-aware Adversarial Domain Adaptation for patient outcomes forecasting. Expert Systems with Applications, 238: 122184. https://doi.org/10.1016/j.eswa.2023.122184

[15] Chen, Y., Zhou, H., Wang, Z., Zhong, P. (2024). Heterogeneous domain adaptation by class centroid matching and local discriminative structure preservation. Neural Computing and Applications, 36(21): 12865-12881. https://doi.org/10.1007/s00521-024-09786-9

[16] Yu, Y., Karimi, H.R., Shi, P., Peng, R., Zhao, S. (2024). A new multi-source information domain adaption network based on domain attributes and features transfer for cross-domain fault diagnosis. Mechanical Systems and Signal Processing, 211: 111194. https://doi.org/10.1016/j.ymssp.2024.111194

[17] Dong, C., Sun, D. (2024). Multi-source domain transfer learning with small sample learning for thermal runaway diagnosis of lithium-ion battery. Applied Energy, 365: 123248. https://doi.org/10.1016/j.apenergy.2024.123248

[18] Ruan, Y., Ma, Y., Meng, H., Xu, T., Yao, Y., Qian, F., Wang, C., Liu, W. (2026). A transfer learning framework using spatiotemporal graph convolutional network and adversarial domain adaptation for cross-district building group energy prediction. In Building Simulation, Beijing: Tsinghua University Press, pp. 1-21. https://doi.org/10.1007/s12273-025-1370-3

[19] Li, W., Chen, Y., Li, J., Wen, J., Chen, J. (2024). Learn then adapt: A novel test-time adaptation method for cross-domain fault diagnosis of rolling bearings. Electronics, 13(19): 3898. https://doi.org/10.3390/electronics13193898

[20] Qian, X., Lu, W., Zhang, Y. (2024). Adaptive wavelet-VNet for single-sample test time adaptation in medical image segmentation. Medical Physics, 51(12): 8865-8881. https://doi.org/10.1002/mp.17423

[21] Yu, Y., Xiong, J., Wu, X., Qian, Q. (2024). From small data modeling to large language model screening: A dual-strategy framework for materials intelligent design. Advanced Science, 11(45): 2403548. https://doi.org/10.1002/advs.202403548

[22] Pan, S.J., Yang, Q. (2010). A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering, 22(10): 1345-1359. https://doi.org/10.1109/TKDE.2009.191

[23] Ruder, S. (2019). Neural transfer learning for natural language processing. Doctoral dissertation, NUI Galway.

[24] Arnold, A., Nallapati, R., Cohen, W.W. (2007). A comparative study of methods for transductive transfer learning. In Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007), Omaha, NE, USA, pp. 77-82. https://doi.org/10.1109/ICDMW.2007.109

[25] Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., He, Q. (2020). A comprehensive survey on transfer learning. Proceedings of the IEEE, 109(1): 43-76. https://doi.org/10.1109/JPROC.2020.3004555

[26] Fu, W., Xue, B., Gao, X., Zhang, M. (2023). Genetic programming for document classification: A transductive transfer learning system. IEEE Transactions on Cybernetics, 54(2): 1119-1132. https://doi.org/10.1109/TCYB.2023.3338266

[27] Finkel, J.R., Manning, C.D. (2009). Hierarchical Bayesian domain adaptation. In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL '09), Association for Computational Linguistics, USA, pp. 602-610. https://doi.org/10.3115/1620754.1620842

[28] Chelba, C., Acero, A. (2006). Adaptation of maximum entropy capitalizer: Little data can help a lot. Computer Speech & Language, 20(4): 382-399. https://doi.org/10.1016/j.csl.2005.05.005

[29] Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16: 321-357. https://doi.org/10.1613/jair.953

[30] Michau, G., Fink, O. (2021). Unsupervised transfer learning for anomaly detection: Application to complementary operating condition transfer. Knowledge-Based Systems, 216: 106816. https://doi.org/10.1016/j.knosys.2021.106816

[31] Pan, S.J., Tsang, I.W., Kwok, J.T., Yang, Q. (2010). Domain adaptation via transfer component analysis. IEEE Transactions on Neural Networks, 22(2): 199-210. https://doi.org/10.1109/CVPR.2012.6247911

[32] Tzeng, E., Hoffman, J., Saenko, K., Darrell, T. (2017). Adversarial discriminative domain adaptation. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, pp. 2962-2971. https://doi.org/10.1109/CVPR.2017.316

[33] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., March, M., Lempitsky, V. (2016). Domain-adversarial training of neural networks. Journal of Machine Learning Research, 17(59): 1-35.

[34] Long, M., Zhu, H., Wang, J., Jordan, M.I. (2017). Deep

transfer learning with joint adaptation networks. In Proceedings of the 34th International Conference on Machine Learning, 70: 2208-2217.
[35] Zhang, D., Zhao, J., Zhou, L., Nunamaker, J.F. (2004). Can e-learning replace classroom learning? Communications of the ACM, 47(5): 75-79. https://doi.org/10.1145/986213.986216
[36] Dayal, A., Shrusti, S., Cenkeramaddi, L.R., Mohan, C.K., Kumar, A. (2025). Leveraging mixture alignment for multi-source domain adaptation. IEEE Transactions on Image Processing, 34: 885-898. https://doi.org/10.1109/TIP.2025.3532094
[37] Finn, C., Abbeel, P., Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In International Conference on Machine Learning, pp. 1126-1135.
[38] Wang, P., Yu, H., Jin, N., Davies, D., Woo, W.L. (2024). QuadCDD: A quadruple-based approach for understanding concept drift in data streams. Expert Systems with Applications, 238: 122114. https://doi.org/10.1016/j.eswa.2023.122114
[39] Yosinski, J., Clune, J., Bengio, Y., Lipson, H. (2014). How transferable are features in deep neural networks? NIPS'14: Proceedings of the 28th International Conference on Neural Information Processing Systems, 2: 3320-3328.
[40] Oquab, M., Bottou, L., Laptev, I., Sivic, J. (2014). Learning and transferring mid-level image representations using convolutional neural networks. In 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, pp. 1717-1724. https://doi.org/10.1109/CVPR.2014.222
[41] Qin, P., Zhao, L. (2023). A novel transfer learning-based cell SOC online estimation method for a battery pack in complex application conditions. IEEE Transactions on Industrial Electronics, 71(2): 1606-1615. https://doi.org/10.1109/TIE.2023.3250768
[42] Chen, Z., Kommrusch, S., Monperrus, M. (2022). Neural transfer learning for repairing security vulnerabilities in c code. IEEE Transactions on Software Engineering, 49(1): 147-165. https://doi.org/10.1109/TSE.2022.3147265
[43] Long, M.S., Zhu, H., Wang, J.M., Jordan, M.I. (2016). Unsupervised domain adaptation with residual transfer networks. Advances in Neural Information Processing Systems, 29.