Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 11, No. 3, 2025

Systematic Analysis of Source-Free Domain Adaptation Methods

Zhiyi Miao

Department of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing, Jiangsu Province, 210094, China

Abstract: Source-Free Domain Adaptation (SFDA) aims to address the challenge of effectively transferring a source domain model to a target domain when the target domain data is unlabeled and the source domain data is unavailable. Traditional Unsupervised Domain Adaptation (UDA) methods rely on simultaneous access to both source and target domain data. However, in many practical scenarios, such as medical data privacy protection or resource-constrained devices, direct access to source domain data is not feasible. SFDA leverages only a pre-trained source domain model and unlabeled target domain data to update the model, avoiding the direct use of source domain data and thereby meeting privacy and security requirements. This paper provides a systematic classification and review of SFDA research methods, categorizing them into three main types: data-related methods, model-related methods, and loss-related methods. Data-related methods replace missing source data by extracting data or feature augmentation information from pre-trained models; model-related methods reduce domain discrepancies by optimizing feature representations or utilizing information in the feature space; and loss-related methods enhance the model's generalization ability through specific loss functions. This paper aims to offer a clear research roadmap for researchers in the field by systematically classifying and analyzing existing SFDA methods, facilitating the selection of appropriate methods or the development of new strategies to address specific problems.

Keywords: Source-Free Domain Adaptation; Transfer Learning; Domain Adaptation; Deep Learning; Pseudo-Labeling.

1. Introduction

In the field of machine learning, deep neural networks have achieved remarkable results in many practical applications due to their strong fitting capabilities. Traditional supervised learning methods rely on large amounts of labeled data, but the cost of labeling is high, and in some cases, such as privacy protection and data sensitivity issues, obtaining sufficient labeled data becomes impractical.

To address this issue, UDA improves model performance on the target domain by reducing the distribution discrepancy between the source and target domains. UDA methods primarily achieve domain alignment by minimizing discrepancy loss or through adversarial learning, allowing the model to learn from the source domain and adapt to the target domain. However, traditional UDA typically assumes that both source and target domain data are accessible during the adaptation process, which is unrealistic in many practical applications.

To reduce reliance on source domain data, SFDA [1][2] has emerged and quickly become a focus of research. The core idea of SFDA is to adapt the model to the target domain without directly using source domain data, but rather by initializing the target domain model with a pre-trained source domain model and updating the model using unlabeled target domain data. In the SFDA field, researchers have proposed various methods to address the unavailability of source domain data, aiming to improve the model's generalization ability on the target domain while reducing dependence on source domain data. This paper categorizes these research methods into three main types: data-related, model-related, and loss-related, as shown in Figure 1.

Data-related methods replace missing source data by extracting data information from pre-trained models, thereby reducing the difficulty of knowledge transfer. These methods include source domain estimation, which generates features or data similar to the source domain style by analyzing the statistical characteristics of the source domain, and image style transfer, which simulates the source domain image style by altering the style of target domain images.

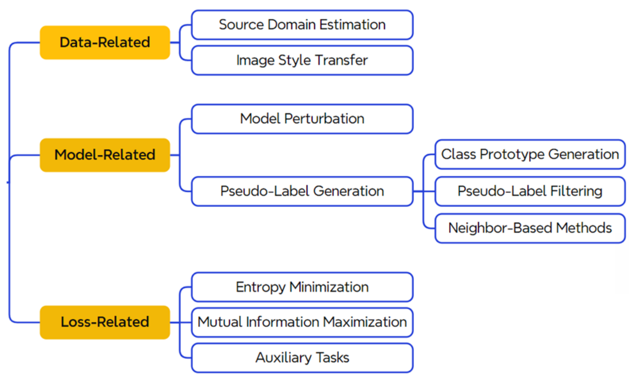

Figure 1. Classification of SFDA Methods in This Paper

Loss-related methods enhance the model's generalization ability by designing specific loss functions, including entropy minimization, which aims to reduce the uncertainty of model predictions; mutual information maximization, which maximizes information sharing between source and target domain features to reduce distribution discrepancies; and other loss methods, such as auxiliary tasks, which introduce additional prediction tasks to provide learning objectives for the model.

This paper provides a systematic summary and in-depth analysis of existing SFDA methods, classifying them based on their characteristics. Through a detailed review and analysis of a large number of cutting-edge literature, this paper aims to provide a clear and comprehensive research roadmap for researchers in the SFDA field, facilitating more efficient development and broader application in the future.

2. Problem Setting

A domain D is defined as a joint distribution $P(x \times y)$ on the space $X \times Y$, where $x \in X$, $y \in Y$, X represents the dataset, and Y represents the corresponding label set. In the SFDA problem, only the source domain model f_S trained on the source domain $Ds = \{X_s, Y_s\}$ and the target domain dataset X_T from $D_T = \{X_T\}$ are available, as shown in Figure 2. The goal is to transfer the trained source domain model f_S to the target domain D_T, so that the predictions obtained from processing the target domain data X_T are as consistent as possible with the true values.

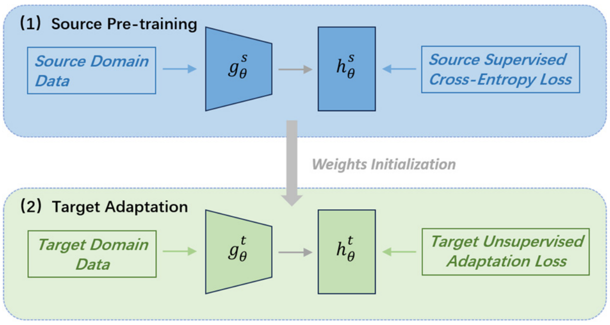

Figure 2. Flowchart of SFDA

3. Data-Related Methods

Data-related methods primarily replace missing source data by extracting data information from pre-trained models, and can be divided into source domain estimation and image style transfer, both of which rely on pre-trained source models to leverage the stored knowledge of source domain data for target domain adaptation.

3.1. Source Domain Estimation

Source domain estimation methods typically rely on pre-trained source models, which are trained on source domain data and store the statistical characteristics of the source domain. By analyzing these statistical characteristics, source domain estimation methods can generate features or data similar to the source domain style for subsequent domain adaptation tasks.

3C-GAN [3] proposes a framework based on conditional generative adversarial networks (GAN), introducing a data generator G(·;β) and a binary discriminator D(·;φ), optimizing an objective function similar to conditional GAN:

$$\min_{\beta} \max_{\varphi} E_{x_t \in X_t}[logD(x_t)] + E_{y_t,z}[\log(1 - D(G(y_t, z)))] \\ -\lambda_s E_{y_t,z} \Sigma_c (y_t = c) \log p(y_c | G(y_t, z), \alpha) \quad (1)$$

where z is a random noise vector, y_t is a predefined label, $\lambda_s > 0$ is a balancing parameter, and α represents the parameters of the pre-trained prediction model. This method alternately optimizes the parameters of the generator and discriminator, enabling the generator to produce multiple substitute labeled source instances for subsequent domain alignment steps. On the other hand, VDM-DA [4] uses the weight vectors of the pre-trained source model as initial class centers and constructs a virtual domain by randomly sampling features, ensuring that the virtual domain has a similar distribution to the source domain in the feature space.

3.2. Image Style Transfer

The goal of image style transfer is to render images with the same content in different styles, typically involving two key loss functions. The content loss ensures that the semantic information of the image remains consistent during style transfer. The style transfer loss achieves effective style transfer based on the statistical characteristics of multiple intermediate layers. In previous methods, most statistical characteristics were presented in the form of Gram matrices or instance normalization, requiring access to source domain style data. However, in the SFDA setting, obtaining source domain style data is not feasible. Therefore, Hou et al. proposed an innovative source-free image style transfer method [5], using the mean μ_n^{stored} and variance μ_n^{stored} of the nth layer image batch stored in the batch normalization (BN) layer as representatives of the source domain style, and constructed the following loss functions:

$$L_{content} = \|F_N(\widetilde{x_t}) - F_N(x_t)\|_2$$

$$L_{style} = \frac{1}{N}\sum_{n=1}^{N} \left\|\mu_n^{\text{current}} - \mu_n^{\text{stored}}\right\|_2 \\ + \left\|\sigma_n^{\text{current}} - \sigma_n^{\text{stored}}\right\|_2 \quad (2)$$

where the feature maps in the source classifier contain N layers, and $\widetilde{x_t}$ represents the image with the source domain style obtained through the generator.

Meanwhile, some methods focus on enhancing sample diversity during training using AdaIN to improve the model's robustness to style variations, thereby obtaining models with strong generalization capabilities.

4. Model-Related Methods

Model-related methods reduce the discrepancy between the source and target domains by optimizing feature representations or utilizing information in the feature space, thereby improving the model's adaptability and performance on the target domain. Based on the processing of the feature space, model-related methods can be divided into model perturbation and pseudo-label generation.

4.1. Model Perturbation

Model perturbation in SFDA can be divided into various techniques, each optimizing feature representations in different ways. MC Dropout [6] estimates epistemic uncertainty by activating dropout during the model's forward propagation.

Additionally, the Mean Teacher framework can be used to form a strong teacher model and a learnable student model, ensuring that the student model's predictions align with the teacher model. The teacher model's weights are gradually updated using the formula:

$$\theta_t = (1 - k)\theta_t + k\theta \quad (3)$$

where θ_t represents the teacher model's weights, θ represents the student model's weights, and k is the momentum coefficient. The teacher model is considered a temporal ensemble of the student model, providing more accurate predictions.

4.2. Pseudo-Label Generation

Pseudo-labels are temporary labels generated by predicting unlabeled data using a pre-trained model. Based on the generation and usage strategies of pseudo-labels, pseudo-label generation methods can be divided into class prototype generation, pseudo-label filtering, and neighbor-based methods.

4.2.1. Class Prototype Generation

The core idea of class prototype generation is to simplify the classification task by calculating representative samples (i.e., prototypes) for each class. Prototypes are typically obtained through clustering algorithms or neural training, effectively capturing class characteristics in the dataset.

A straightforward method is to select class prototypes based on self-entropy, choosing samples with self-entropy above a certain threshold as class prototypes. Although simple, this method may not select representative prototypes, as high self-entropy samples could be outliers or noise. Ding et al. defined the source classifier's weights as class prototypes [7], finding that using source classifier weights as class prototypes yielded higher average accuracy than using entropy criteria.

Another effective method calculates the center of each class based on clustering:

$$c_k = \frac{\sum_{x_t \in X_t} \delta_k(F \circ G(x_t)) F(x_t)}{\sum_{x_t \in X_t} \delta_k(G(x_t))} \quad (4)$$

where c_k represents the center of the kth class, δ_k represents the kth element in the softmax operation, F is the feature extractor, and G is the prediction head. This method generates class centers through clustering algorithms, better capturing the distribution of each class in the target domain. SFDA uses these class prototypes to obtain more robust pseudo-labels for training the network.

4.2.2. Pseudo-Label Filtering

Since pseudo-labels may contain errors, filtering mechanisms are needed to remove noisy labels and retain high-quality pseudo-labels. An end-to-end filtering mechanism is an effective method, identifying reliable samples only when the Hausdorff distance between a sample and its most similar class prototype is less than its distance to the second most similar class prototype.

On the other hand, noisy label learning not only filters incorrect pseudo-labels but also learns useful information from noisy labels for self-refinement. SHOT++ [8] attempts to propagate information between high-confidence and low-confidence samples using MixMatch to improve the reliability of low-confidence samples. Some methods introduce ensemble learning into the SFDA setting, first performing data augmentation on target domain samples from multiple perspectives based on input and feedback, then refining pseudo-labels across multiple versions using negative ensemble learning.

4.2.3. Neighbor-Based Methods

Since similar data points often have similar attributes or labels, the labels of target points can be inferred by analyzing their neighboring points. Neighbor-based methods calculate the distance between target points and other points in the dataset to find "neighbors" and use their information for prediction, as shown in Figure 3.

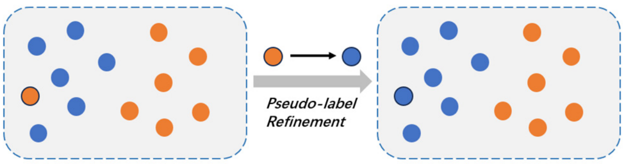

Figure 3. Neighbor-Based Methods

A common method is to perform K-means clustering in the target domain and use the prediction results of neighbors within the same cluster to generate more accurate pseudo-labels. CaiDA [9] proposes a greedy chain search strategy to find the nearest neighbor in the anchor set, interpolate the nearest anchor into the target feature, and use the prediction results of the synthesized feature. Some methods design different weighting schemes for neighbors during aggregation. Memory banks and online refinement maintain a memory bank storing target data features and predictions, allowing online refinement of pseudo-labels. Specifically, refined pseudo-labels are obtained using the formula:

$$\hat{p} = \frac{1}{m} \sum_{j \in N_i} q_j \quad (5)$$

where N_i represents the indices of the m nearest neighbors of g(xi) in the memory bank. Through memory banks and online refinement, more robust pseudo-labels are obtained, improving the model's adaptability.

5. Loss-Related Methods

Loss-related methods primarily include entropy minimization, mutual information, and auxiliary tasks. These methods rely on the intrinsic structure and statistical characteristics of the data rather than external label information, enhancing the model's generalization ability through specific loss functions. Entropy minimization leverages the distribution characteristics of the data, reducing the model's prediction uncertainty on unlabeled data to help the model make more accurate predictions in SFDA. Mutual information maximizes information sharing between source and target domain features, reducing distribution discrepancies, and is suitable for SFDA. Auxiliary tasks introduce additional prediction tasks to provide learning signals, enhancing the model's understanding of images. Although their goals differ, they all aim to optimize model performance, making it perform better in the SFDA field.

5.1. Entropy Minimization

Traditional UDA methods primarily rely on feature distribution alignment to enhance the model's adaptability on the target domain, reducing the discrepancy between the source and target domains to improve generalization. However, when source domain data is unavailable, these methods are no longer applicable. Researchers have proposed starting from the model's output, assuming that an adaptive model should produce highly certain predictions for each sample, i.e., achieving entropy minimization. This result-oriented constraint can guide the model's optimization process.

Specifically, the core of the entropy minimization strategy is to quantify the uncertainty of the model's output and minimize it. In the context of UDA, this means the model is encouraged to produce high-confidence predictions, reducing uncertainty on target domain data. Based on robustness measures in information theory, the entropy minimization loss can be expressed as:

$$H(\hat{Y}_t|X_t) = \frac{1}{n_t} \sum_{i=1}^{n_t} \sum_{c=1}^{C} p_\theta(y_c|x_i) \log p_\theta(y_c|x_i) \quad (6)$$

where $p_\theta(y_c|x_i)$ is the model's predicted probability that the ith sample belongs to the cth class, and nt is the number of samples.

5.2. Mutual Information Maximization

Mutual information maximization is a clustering-based regularization that enhances the model's feature representation ability by maximizing the mutual information between input data and model output. Mutual information $I(X_t, \hat{Y}_t)$ measures the dependency between two random variables X_t and $\hat{Y}_t$, defined as:

$$I(X_t, \hat{Y}_t) = H(\hat{Y}_t) - H(\hat{Y}_t | X_t) \quad (7)$$

where $H(\hat{Y}_t)$ is the entropy of $\hat{Y}_t$, representing its uncertainty, and $H(\hat{Y}_t | X_t)$ is the conditional entropy of $\hat{Y}_t$ given X_t, representing the remaining uncertainty of $\hat{Y}_t$ when X_t is known. Mutual information can be further expressed as:

$$I(X_t, \hat{Y}_t) = -\sum_{c=1}^{C} \bar{p}_\theta(y_c) \log \bar{p}_\theta(y_c) + \frac{1}{n_t}\sum_{i=1}^{n_t}\sum_{c=1}^{C} p_\theta(y_c|x_i) \log p_\theta(y_c|x_i) \quad (8)$$

where $\bar{p}_\theta(y_c) = \frac{1}{n_t}\sum_i p_\theta(y_c|x_i)$ represents the estimated class label distribution's cth element. By increasing $H(\hat{Y}_t)$, i.e., increasing the entropy of $\hat{Y}_t$, the model is encouraged to generate a more uniform label distribution, avoiding the tendency to assign all samples to the same class.

Mutual information maximization was first introduced in SHOT [10] and SHOT++ for image classification tasks. By maximizing mutual information, SHOT and SHOT++ effectively learn target domain feature representations, making them as consistent as possible with the source domain classifier's output, thereby improving model performance on the target domain. Subsequently, many SFDA methods have adopted mutual information maximization. In multimodal sentiment analysis, hierarchical mutual information maximization is used to improve multimodal fusion and eliminate modality discrepancies.

5.3. Auxiliary Tasks

Auxiliary tasks allow the model to learn useful feature representations from data without explicit labels, guiding the model to learn the intrinsic structure and patterns of the data.

Rotation prediction is a common auxiliary task, requiring the model to predict the angle by which an input image has been rotated. This forces the model to learn the spatial structure and orientation information of objects in the image, which is valuable for subsequent recognition tasks. The SHOT++ algorithm introduces a relative rotation prediction task based on self-supervised learning, requiring the model to predict both the semantic classification of the image and the relative rotation angle between images. This multi-task learning approach enables the model to simultaneously learn classification information and spatial relationships, providing richer feature representations.

6. Conclusion

To meet the privacy and security requirements of practical scenarios, various SFDA methods have emerged. To help researchers better understand and apply SFDA methods, this paper provides a comprehensive and in-depth summary and classification of existing SFDA methods. Through a detailed review and analysis of a large number of cutting-edge literatures, this paper categorizes existing SFDA methods into three main types: data-related methods, model-related methods, and loss-related methods. This systematic summary and classification aim to provide a clear research roadmap for researchers in the SFDA field, promoting future development and application.

References

[1] Yu You, Feng Lin, Wang Gege. A Pseudo-Label-Based Semi-Supervised Few-Shot Learning Model [J]. Acta Electronica Sinica, 2019, 47(11): 2284-2291.

[2] Wang Shipeng, Sun Jian, Xu Zongben. A Source-Free Domain Adaptation Method Based on Feature Confidence [J]. Chinese Journal of Biomedical Engineering, 2024, 43(2): 129-143.

[3] R. Li, Q. Jiao, W. Cao, H.-S. Wong, and S. Wu, “Model adaptation: Unsupervised domain adaptation without source data,” in Proc. CVPR, 2020, pp. 9641–9650.

[4] J. Tian, J. Zhang, W. Li, and D. Xu, “Vdm-da: Virtual domain modeling for source data-free domain adaptation,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 32, no. 6, pp. 3749–3760, 2022.

[5] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in International conference on machine learning. PMLR, 2015, pp. 448–456.

[6] Y. Gal and Z. Ghahramani, “Dropout as a bayesian approximation: Representing model uncertainty in deep learning,” in Proc. ICML, 2016, pp. 1050–1059.

[7] Y. Ding, L. Sheng, J. Liang, A. Zheng, and R. He, “Proxymix: Proxy-based mixup training with label refinery for source-free domain adaptation,” arXiv preprint arXiv:2205.14566, 2022.

[8] J. Liang, D. Hu, Y. Wang, R. He, and J. Feng, “Source data-absent unsupervised domain adaptation through hypothesis transfer and labeling transfer,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021.

[9] J. Dong, Z. Fang, A. Liu, G. Sun, and T. Liu, “Confident anchorinduced multi-source free domain adaptation,” in Proc. NeurIPS, 2021, pp. 2848–2860.

[10] J. Liang, D. Hu, and J. Feng, “Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation,” in Proc. ICML, 2020, pp. 6028–6039.