Active Multi-source Domain Adaptation for Multimodal Fake News Detection

Yanping Chen[1, 2], Weijie Shi[3], Mengze Li[3], Yue Cui[3, 4], Jiaming Li[4], Ruiyuan Zhang[5], Hao Chen[3], Hanghui Guo[6], Shimin Di[7], Ziyi Liu[3], Jia Zhu[6], Jiajie Xu[1, 2*]

[1]School of Computer Science and Technology, Soochow University
[2]Key Laboratory of Data Intelligence and Advanced Computing, Soochow University
[3]The Hong Kong University of Science and Technology
[4]Alibaba Group
[5]Hong Kong Generative AI Research and Development Center
[6]Zhejiang Normal University
[7]Southeast University

Abstract

Multimodal fake news detection plays a crucial role in combating online misinformation. The inherent domain diversity of news in the real world has driven the development of cross-domain detection methods. However, these detection methods either suffer from significant performance degradation due to semantic and deception pattern shifts between the training (source) and test (target) domains or heavily rely on annotated labels. To address the problems, we propose ADOSE, an active multi-source domain adaptation framework for multimodal fake news detection which actively annotates a small subset of target samples to improve detection performance. Specifically, for domain shifts, we design a multi-expert classifier network based on refined features to comprehensively capture and adapt to the semantic space and deception patterns of news across different domains. To maximize adaptation performance with limited annotation cost, we propose a least-disagree uncertainty selector equipped with a diversity calculator for selecting the most informative samples. The selector leverages the uncertainty of inconsistent predictions before and after perturbations by multiple classifiers as an indicator of unfamiliar samples. It further incorporates diversity scores derived from multi-view features to ensure the chosen samples achieve maximal coverage of target domain features. The extensive experiments on multiple datasets show that ADOSE outperforms existing domain adaptation methods by 2.45% $\sim$ 9.1%, indicating the superiority of our model.

1 Introduction

Multimodal fake news detection aims to identify the authenticity of online news composed of multiple modalities such as text and images. With the proliferation of various social networks like Twitter and Weibo, this task plays a critical role in combating misleading information and has garnered significant attention (Liu et al. 2024a; Zhang et al. 2024b).

In the real world, the disparity in descriptive content and linguistic style often leads to the categorization of news into different domains such as society, education, and healthcare. Leveraging news from multiple source domains to detect news in a target domain enables the adaptation of extensive news knowledge and reduces reliance on labeled target domain data (Yang et al. 2022; Cui and Jia 2025). Conventional domain adaptation methods transfer knowledge from labeled source domains to unlabeled target domains, but struggle with performance degradation when confronting significant domain shifts in data distribution (Ran and Jia 2023; Liu et al. 2023b). Active Domain Adaptation (ADA) methods strike a balance between labeling cost and adaptation performance (Ma, Gao, and Xu 2021). ADA assists

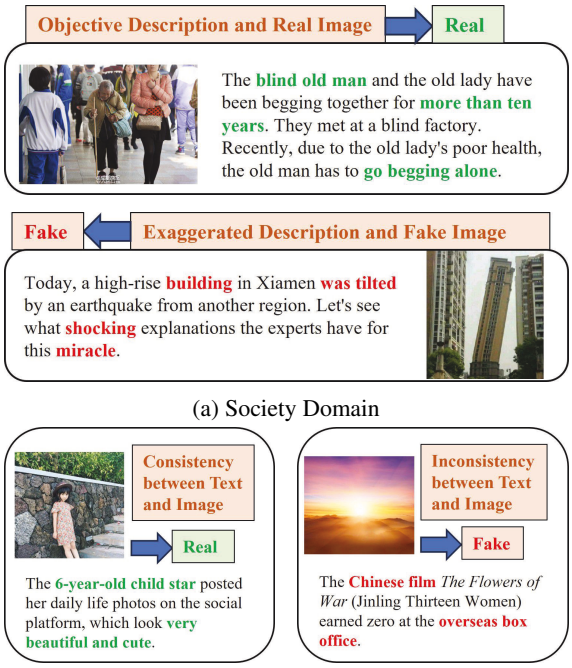


Figure 1: The comparison of deception patterns between the social and entertainment domains in the Weibo dataset, with "Real" labeled as real news and "Fake" labeled as false news.

*Corresponding author: xujj@suda.edu.cn.

knowledge transfer for the target domain by strategically annotating informative target samples within a limited budget. Therefore, we introduce ADA for fake news detection and define the new task scenario as multi-source active domain adaptation for multimodal fake news detection.

Despite the success of ADA in various multimedia tasks such as object detection and semantic segmentation (Nakamura, Ishii, and Yamashita 2024; Liu et al. 2024b), it faces two key challenges in our task setting. **First, how to address varying semantic spaces and deception patterns across domains.** For example, in the social domain (as shown in Figure 1a), the semantic space centers around social events, and the deception pattern manifests as exaggerated descriptions and fabricated images, which are considered intra-modal knowledge errors (Dong et al. 2024). In contrast, in the entertainment domain, fake celebrity news is mainly characterized by the inconsistency between text and image, which relies on cross-modal interaction (as seen in Figure 1b) (Ma et al. 2024; Madaan et al. 2024). **Second, how to design an active sample selection strategy for target domain adaptation.** An effective strategy requires identifying samples that best reflect the domain shifts. However, the unknown veracity and multimodal nature of news in the target domain make it challenging to design appropriate metrics for evaluating sample representativeness (Kim et al. 2023; Lu et al. 2024).

In this paper, we propose ADOSE, an Active Multi-source Domain Adaptation Framework for Multimodal Fake News Detection. (1) For semantic space shifts, we employ adversarial networks and contrastive learning to refine domain-invariant and semantically aligned multimodal features. These features are used to construct classifiers that capture distinct deception patterns, including two classifiers based on intra-modal dependencies and one based on inter-modal dependencies. The above processes are integrated into a Modal-dependency Expertise Fusion Network (MEFN) for adapting to varying deception patterns. (2) To select informative samples in the target domain, we design a dual selection strategy targeting domain shift and feature distribution. Specifically, we utilize prediction uncertainty as an indicator of domain shifts and propose a Least-disagree Uncertainty Selector (LUS) to identify the samples with maximal uncertainty. LUS applies Gaussian perturbations to estimate each sample's proximity to decision boundaries. We further select the most discriminative samples from the multimodal latent feature space using a Multi-view Diversity Calculator (MDC). This strategy allows us to identify the most valuable target samples for annotation. The main contributions of this paper are as follows:

- We introduce ADOSE, an active domain adaptation framework for multimodal fake news detection, aiming to address the adaptation problem from multiple source domains to a single target domain.
- We develop a modal-dependency expertise fusion network that captures both intra-modal and inter-modal dependencies, enabling fake news pattern recognition across domains.
- We design an effective dual selection strategy to identify the most informative target samples for annotation under domain shift.

2 Related Work

2.1 Multimodal Fake News Detection

Multimodal fake news detection has attracted increasing attention, with key efforts focusing on effective text-image fusion. Recent works explore diverse fusion strategies: COOLANT (Wang et al. 2023) applies cross-modal contrastive learning with auxiliary loss softening and attention integration; MSACA (Wang et al. 2024) enhances semantic alignment via hierarchical multi-scale image features and selective attention; MMDFND (Tong et al. 2024) addresses cross-domain fusion using domain embeddings and stepwise transformers. Other advances include AKA-Fake (Zhang et al. 2024a), which builds knowledge subgraphs via reinforcement learning and learns with heterogeneous graphs; Event-Radar (Ma et al. 2024), modeling event-level inconsistencies through multi-view graph fusion; and NSLM (Dong et al. 2024), which incorporates neuro-symbolic reasoning with pseudo-Siamese networks. Additionally, FKA-Owl (Liu et al. 2024c) enhances large vision-language models using forgery-specific knowledge, and FoRM (Li et al. 2024) improves token-level reasoning with coarse-to-fine relation attention.

2.2 Active Domain Adaptation

Active domain adaptation enhances model generalization to a new target domain by selecting a small, informative subset of target samples for annotation. Previous studies include CLUE (Prabhu et al. 2021), which uses uncertainty-weighted clustering for diverse sample selection, and EADA (Xie et al. 2022), which employs energy-based sampling to bridge source-target gaps. DiaNA (Huang et al. 2023) proposes a divide-and-adapt framework based on uncertainty and domainness. For probabilistic uncertainty modeling, Diffusion-Based ADA (Du and Li 2023) and DUC (Xie et al. 2023) introduce variational inference and the Dirichlet distribution, respectively. Recently, Ada-iD (Han et al. 2024) optimizes intrusion detection via multi-task learning and Category-Aware ADA (Xiao, Gu, and Liu 2024) targets category imbalance with influence functions. RASC (Zhang et al. 2024d) employs a reconfigurability-aware sample selection strategy to mine semantic information. Detective (Zhang et al. 2024c) extends ADA to multi-source settings with evidential deep learning and LUET (Sun et al. 2025) integrates local uncertainty and energy transfer alignment constraints to achieve adaptation. In contrast, we design a multi-expert fusion network and a fine-grained multimodal uncertainty selector to adapt to domain shifts and select informative target samples for fake news detection.

3 Problem Statement

Formally, we have access to M fully labeled source domains and a target domain with actively labeled target data within a pre-defined budget $\mathcal{B}$. All domains are defined based on different news events or topics. The i-th source domain $\mathcal{S}_i = \{(t_k, v_k), y_k\}_{k=1}^{N_{s,i}}$ contains $N_{s,i}$ labeled samples,

where $i \in M$ and the target domain $\mathcal{T}_u = \{(t_k, v_k)\}_{k=1}^{N_{tu}}$ contains N_{tu} unlabeled samples. And, (t, v) is a text-image pair, where t is a text sentence, and v is the corresponding image. The size of budget $\mathcal{B}$ is set to N_{tl}, and the labeled target domain is denoted as $\mathcal{T}_l = \{(t_k, v_k), y_k\}_{k=1}^{N_{tl}}$, where $N_{tl} \ll N_{tu}$ and $N_{tl} \ll N_{s.i}$. The multiple source domains and target domain have the same label space $\mathcal{Y} = \{0, 1\}$ (0 indicates real news and 1 indicates fake news). We aim to learn a model $\mathcal{M}(\cdot)$ adapting from $\{\mathcal{S}_i\}_{i=1}^{M} \cup \mathcal{T}_l$ to $\mathcal{T}_u$, i.e., the model can generalize well on unseen samples from $\mathcal{T}_u$.

4 Methodology

In this section, we introduce our model, ADOSE, which comprises three components: Modal-dependency Expertise Fusion Network (MEFN), Least-disagree Uncertainty Selector (LUS) and Multi-view Diversity Calculator (MDC). The structure of the model is illustrated in Figure 2.

4.1 Modal-dependency Expertise Fusion Network

The design motivation of the Modal-dependency Expertise Fusion Network (MEFN) is twofold: on one hand, it leverages domain-invariant feature extraction and multimodal representation alignment to mitigate distribution shifts in multi-domain scenarios; on the other hand, it enhances adaptability to modal interactions in different news contexts by explicitly modeling various modality dependencies and performing late fusion operations.

Domain Invariant Feature Extraction. Given an input text-image pair (t, v), following previous work (Wang et al. 2021; Liu et al. 2023a), we leverage a convolutional neural network (i.e., TextCNN (Chen 2015)) with an additional perceptron (MLP) as the textual encoder E_t to obtain the representation of t as e_t. For image representation, following existing methods (Zhang et al. 2020; Li et al. 2021), we use ResNet50 as the visual backbone neural network and further generate visual representation with reduced dimensionality of v as e_v:

$$e_t = E_t(t; \theta_t), \quad e_v = E_v(v; \theta_v) \tag{1}$$

where $e_t, e_v \in \mathbb{R}^d$, E_v represents the visual encoder, and θ_t, θ_v represents the parameter of the textual encoder and the visual encoder respectively.

The feature distribution shift among source samples and target samples is a key obstacle for domain adaptation. In order to generate domain invariant features, we train the textual encoder and visual encoder through adversarial networks, which has been proven effective in recent domain adaptation studies (Du and Li 2023; Lu, Huang, and Hu 2024). We design a textual discriminator D_t (for invariant text features) and a visual discriminator D_v (for invariant image features), each with $M + 1$ classes (M source domains and one target domain). Fine-grained domain adversarial learning is then performed, where textual adversarial loss $\mathcal{L}_{abv\_t}$ and visual adversarial loss $\mathcal{L}_{abv\_v}$ are defined between any two different domains:

$$\mathcal{L}_{abv\_t} = -\mathbb{E}_{e_t \sim (S \cup \mathcal{T}_u)} \log D_t(d_l \mid GRL(e_t)) \tag{2}$$

$$\mathcal{L}_{abv\_v} = -\mathbb{E}_{e_v \sim (S \cup \mathcal{T}_u)} \log D_v(d_l \mid GRL(e_v)) \tag{3}$$

where $d_l \in \{0, 1, 2...M\}$ is the domain label of e_t and e_v, and GRL denotes the gradient reversal layer. Here, $D_t(d_l \mid GRL(e_t))$ and $D_v(d_l \mid GRL(e_v))$ represent the probabilities predicted by the discriminators for the ground-truth domain label. We minimize the feature generator and maximize the overall adversarial loss which can be defined as $\mathcal{L}_{adv} = \mathcal{L}_{abv\_t} + \mathcal{L}_{abv\_t}$ to play the min-max game.

Multimodal Representation Alignment. To learn expert knowledge of cross-modal interactions, it is necessary to align semantically consistent text and images. Based on the domain invariant text feature e_t and image feature e_v, we utilize MLP transformations to further obtain textual and visual alignment representation x_t and x_v:

$$x_t = \text{MLP}(e_t; \theta_{mt}), \quad x_v = \text{MLP}(e_v; \theta_{mv}) \tag{4}$$

In the task of fake news detection, cross-modal correspondence or similarity is more likely to only exist in real news rather than in misinformation scenarios. Besides, text for different misinformation examples may use the same image in a specific event domain, which results in the images and text of many negative samples being close to each other in the semantic space. Inspired by RDCM (Liu et al. 2023a), we adopt the following sampling strategy to overcome this limitation, which only takes real posts as positive samples and filters out negative samples with high semantic similarity on the visual modality:

$$\mathbb{I}(e_v^i, e_v^j) = \begin{cases} 0, & \text{if sim}(e_v^i, e_v^j) \geq \beta \\ \beta - \text{sim}(e_v^i, e_v^j), & \text{else} \end{cases} \tag{5}$$

where $\text{sim}(e_v^i, e_v^j) = \frac{1}{2}(\frac{e_v^i {e_v^j}^T}{\|e_v^i\| \|e_v^j\|} + 1)$ and β is a threshold to remain semantic dissimilar pairs as negative samples. Then, we leverage the contrastive objective to update the parameters of MLP i.e. θ_{mt} and θ_{mv} so that we obtain multimodal semantic alignment features (x_t and x_v):

$$\mathcal{L}_{ctr} = -\log \frac{\exp\left(\frac{x_t^i {x_v^i}^T}{\tau}\right)}{\exp\left(\frac{x_t^i {x_v^i}^T}{\tau}\right) + \sum_{i \neq j}^{b} \exp\left(\frac{x_t^i {x_v^j}^T}{\tau}\right) \mathbb{I}(e_v^i, e_v^j)} \tag{6}$$

where i represents the indices of real posts in a batch, j represents the indices of the other samples in this batch except the i-th sample, b is the batch size, and τ is a temperature hyperparameter.

Expertise Fusion Network. After the above feature refinement process, we obtain domain-invariant unimodal feature representations e_t, e_v and aligned multimodal representations x_t, x_v. We concatenate x_t and x_v to create the cross-modal feature representation. Using the unimodal features, we build a text classifier $cls\_t(e_t;)$ and an image classifier $cls\_v(e_v;)$ to capture intra-modal dependencies. Based on the cross-modal features, we construct a classifier $cls\_c([x_t, x_v];)$ to model inter-modal dependencies. Finally, we combine the outputs of these three classifiers to generate the predictions of the expertise fusion network. This process is formulated as follows:

$$score_y = \log P_{cls\_t}(y) + \log P_{cls\_v}(y) + \log P_{cls\_c}(y)$$

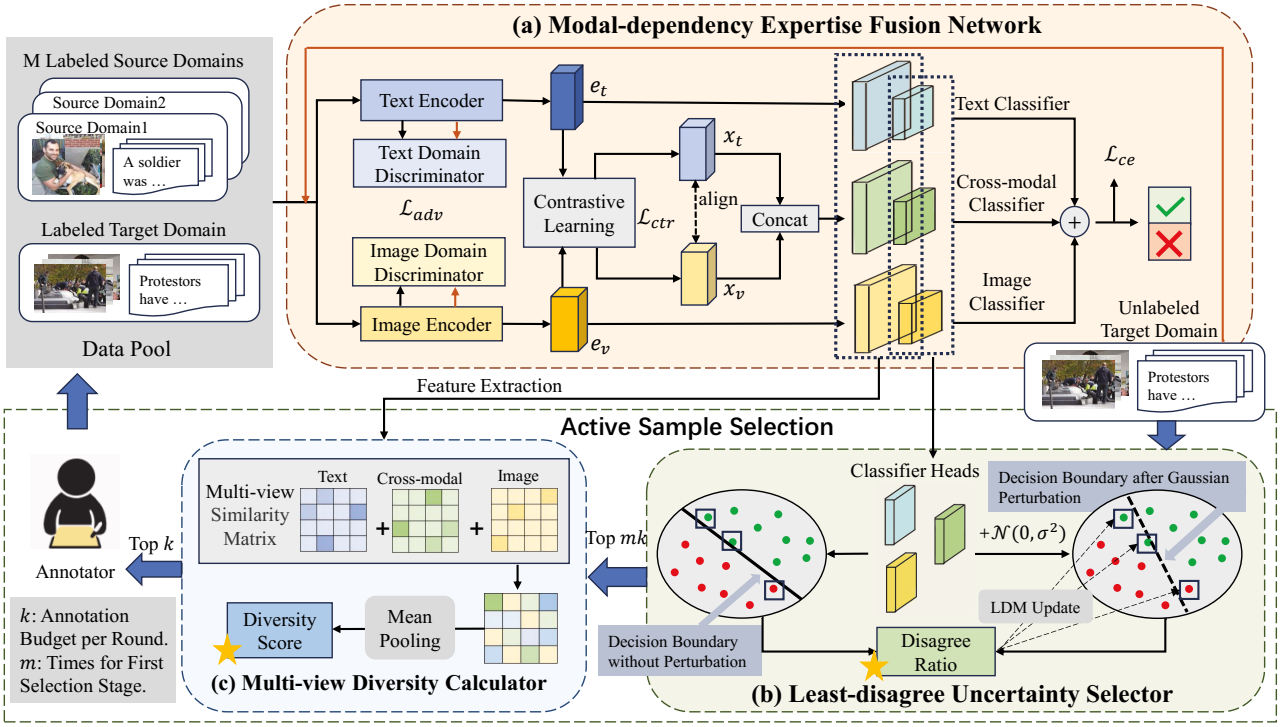


Figure 2: The network architecture of ADOSE. (a) Modal-dependency Expertise Fusion Network (MEFN) utilizes multiple classifiers to detect fake news based on adversarial training and contrastive learning. (b) and (c) are the Least-disagree Uncertainty Selector (LUS) and Multi-view Diversity Calculator (MDC) that respectively measure the uncertainty (first selection stage) and diversity (second selection stage) of target samples.

$$P_{mefn} = \text{softmax}\left(\frac{score_y}{\log \sum_y \exp(score_y)}\right) \tag{7}$$

where y denotes the news label (0 or 1), $P_{cls\_}$ represents the predicted probability distributions, and P_{mefn} is the final probability distribution from our MEFN module.

4.2 Least-disagree Uncertainty Selector

The erroneous or uncertain decisions of the model on the target domain highlight the distribution shift from the source domain. Selecting the most representative misclassified or low-confidence samples can help the model more effectively adapt to the unseen distribution. As misclassifications cannot be directly identified, we focus on highly uncertain samples, which are often close to the decision boundary and thus more likely to be misclassified. Inspired by recent active learning work (Cho et al. 2023; Hanneke et al. 2014), we develop a Least-disagree Uncertainty Selector (LUS) (Figure 2(b)) by introducing the Least Disagree Metric (LDM) to quantify the closeness of a sample to the decision boundary. Let $\mathcal{X} = \{(e_t, e_v, x_t, x_v)\}$ and $\mathcal{Y}$ denote the feature and label spaces, respectively, and $\mathcal{H}$ denote the hypothesis space, encompassing all possible functions $h : \mathcal{X} \to \mathcal{Y}$. The disagree probability measure between two hypotheses h_1 and h_2 is defined as follows:

$$\rho(h_1, h_2) := \mathbb{P}_{\mathcal{X} \sim \mathcal{T}_u}[h_1(\mathcal{X}) \neq h_2(\mathcal{X})] \tag{8}$$

For a given hypothesis $g \in \mathcal{H}$ and $\mathcal{X}_0 \in \mathcal{X}$, let $\mathcal{H}_0 := \{h \in \mathcal{H} | h(\mathcal{X}_0) \neq g(\mathcal{X}_0)\}$ be the set of hypotheses disagreeing with g in their prediction for $\mathcal{X}_0$. Further, the least disagree metric (LDM) is defined as

$$L(g, \mathcal{X}_0) := \inf_{h \in \mathcal{H}_0} \rho(h, g). \tag{9}$$

The true LDM in our task is not computable due to the complexity of the neural network architecture. We follow the approximation method for LDM proposed in the work (Cho et al. 2023) to estimate the LDM value for each target domain sample. Specifically, we use the statistical probability of prediction disagreement by the model on target domain samples to replace the probability $\mathbb{P}$ and construct the hypothesis set $\mathcal{H}_0$ with a set finite number of variances. The estimator denoted by $L_e(g, \mathcal{X}_0)$ is defined as follows:

$$L_e(g, \mathcal{X}_0) := \inf_{h \in \mathcal{H}_0^K} \left\{ \frac{1}{N_{tu}} \sum_{i=1}^{N_{tu}} \mathbb{I}[h(\mathcal{X}_i) \neq g(\mathcal{X}_i)] \right\} \tag{10}$$

where $\mathbb{I}$ is an indicator function, N_{tu} is the number of target domain unlabeled samples for approximating ρ, and K is the number of sampled hypotheses for approximating L. We define a set of linearly increasing variances $\{\sigma_k^2\}_{k=1}^K$ $(0 < \sigma_k \leq 1)$ to apply Gaussian perturbations to the model weights, thereby constructing the hypothesis set.

We apply perturbations to the expert fusion network to influence its predictions by introducing Gaussian perturbations $\mathcal{N}(\cdot, \sigma^2)$ to the weights of the last layer of each modal classifier, as the weights of the last layer critically affect the prediction results. Before perturbation, we denote the weights of the last layer of the three classifiers as $\mathbb{W}_{cls\_t}, \mathbb{W}_{cls\_v}, \mathbb{W}_{cls\_c}$ respectively, and the expert fusion network as F. After perturbation, the weights of the last layer of the three classifiers are denoted as $\widetilde{\mathbb{W}}_{cls\_t}, \widetilde{\mathbb{W}}_{cla\_v}, \widetilde{\mathbb{W}}_{cls\_c}$, and the expert fusion network as $\widetilde{F}$. The perturbation process and the definitions of F and $\widetilde{F}$ are as follows:

$$\widetilde{\mathbb{W}}_{cls} \sim \mathcal{N}(\mathbb{W}_{cls}, \sigma^2), cls \in \{cls\_t, cls\_v, cls\_c\} \quad (11)$$

$$F = F(; \mathbb{W}_{cls\_t}, \mathbb{W}_{cls\_v}, \mathbb{W}_{cls\_c}) \quad (12)$$

$$\widetilde{F} = \widetilde{F}(; \widetilde{\mathbb{W}}_{cls\_t}, \widetilde{\mathbb{W}}_{cls\_v}, \widetilde{\mathbb{W}}_{cls\_c}) \quad (13)$$

The perturbation with varying variances is conducted over K rounds. We use the fusion network with perturbed weights to re-predict target samples and compute the proportion of changed predictions, which serves as the candidate LDM for each update round. Samples with smaller L_e values indicate that, compared to other samples, they are more susceptible to changes in the decision boundary. The formula for updating the L_e of samples during the entire perturbation process is expressed as follows:

$$L_e(\mathcal{X}_i) = \begin{cases} \min(L_e(\mathcal{X}_i), \rho(\widetilde{F}, F)), & \text{if } F(\mathcal{X}_i) \neq \widetilde{F}_{k,j}(\mathcal{X}_i) \\ L_e(\mathcal{X}_i), & \text{else} \end{cases} \quad (14)$$

Here, $\widetilde{F}_{k,j}$ denotes the fusion network after the j-th weight sampling in the k-th round of perturbation using variance σ_k and $L_e(\mathcal{X}_i)$ is initialized to 1. After the entire perturbation process is completed, each target domain sample obtains an L_e value. We sort the L_e values of all samples in ascending order and select the top mk samples as candidate samples for the next stage of selection, where m is a hyperparameter and k is the budget for active selection in each round.

4.3 Multi-view Diversity Calculator

Further, we refine the selection process by increasing sample diversity (Figure 2(c)). Specifically, instead of directly using the input features of the classifiers, we leverage the shallow networks of each classifier (the first layer of MLP) in the fusion network to extract multi-view features. This is because the features obtained before the final classification have greater discriminative value for model training. We denote the feature representations of the three modalities obtained from these networks as f_t, f_v, f_c. Then, we use cosine similarity to compute the inter-sample similarity matrices for each modality and then obtain a global similarity matrix. Furthermore, we aggregate the similarity of a sample with all other samples to calculate the diversity score d_i of i-th sample, which is defined as follows:

$$d_i = \frac{1}{3N_{tu}} \sum_{j=1}^{N_{tu}} \left[cos(f_t^i, f_t^j) + cos(f_v^i, f_v^j) + cos(f_c^i, f_c^j) \right] \quad (15)$$

A larger d_i indicates that the sample is more dissimilar to other samples in the feature space. Finally, we select the top k samples with the largest d_i values from the mk samples for annotation. The final selected samples integrate uncertainty estimation and diversity calculation, and are considered the most informative for facilitating domain adaptation.

4.4 ADOSE Training

We define the objective loss of the MEFN module as $\mathcal{L}_{ce}$, which consists of the BCE loss $\mathcal{L}_{efn}$ from the expertise fusion network and the BCE loss $\mathcal{L}_{cls}$ from multiple classifiers, i.e., $\mathcal{L}_{ce} = \mathcal{L}_{efn} + \lambda_c \mathcal{L}_{cls}$ (λ_c is a hyperparameter). Let N_l be the total number of labeled samples from all domains, and $P_\cdot^i(y)$ be the prediction probability for the i-th sample. The expertise generated by multiple classifiers should exhibit as much consistency as possible on true news. In other words, we expect that the prediction of the classifiers on positive samples are consistent. Accordingly, the negotiation loss $\mathcal{L}_{nego}$ between classifiers is defined based on JS divergence as follows:

$$\mathcal{L}_{nego} = \frac{1}{2N_{l+}} \sum_{i=1}^{N_{l+}} [\text{JS}(P_{cls\_t} || P_{cls\_c}) + \text{JS}(P_{cls\_v} || P_{cls\_c})] \quad (16)$$

where N_{l+} is the total number of positive labeled samples (i.e., real news). We introduce hyperparameters λ_a, λ_t and λ_n to define the overall loss for ADOSE:

$$\mathcal{L} = \mathcal{L}_{ce} + \lambda_a \mathcal{L}_{adv} + \lambda_t \mathcal{L}_{ctr} + \lambda_n \mathcal{L}_{nego} \quad (17)$$

5 Experiments

5.1 Experiment Setup

Datasets. Our model is evaluated on two real-world datasets: Pheme (Zubiaga, Liakata, and Procter 2017) and Weibo (Wang et al. 2018). Pheme is collected from five news events related to terrorist attacks. We retain four event domains with enough available text and images: Charlie Hebdo (C), Sydney Siege (S), Ferguson Unrest (F) and Ottawa Shooting (O). Weibo is divided into nine topic domains by the work (Tong et al. 2024). Similar to the setup of Pheme, we remove some domains with a small number of samples and finally get four topic domains: society (S), entertainment (E), education (D) and health (H). We split the above datasets into 70% training sets and 30% testing sets according to each domain. To meet the domain adaptation setting, three news domains are set as source domains, and the remaining one is the target domain.

Implementation Details. We use TextCNN and ResNet50 as the backbone framework to extract initial text and image embeddings and map the embeddings into 256 dimensions. Adam (Kingma and Ba 2014) is adopted as the optimizer with a learning rate 1e-3, weight decay of 5e-4 and batch size of 16. Following (Han et al. 2024), we set the total labeling budget $\mathcal{B}$ as 10%, which is divided into 5 selection rounds, i.e., the labeling budget in each round is $k = \mathcal{B}/5$. The hyperparameters are determined through experimental search as $\lambda_c = \lambda_t = 0.5$, $\lambda_a = \lambda_n = 0.2$. And $m = 2$ is set

Dataset	Method	*SOF* → *C*	*COF* → *S*	*CSF* → *O*	*CSO* → *F*	Avg. Acc	Avg. F1(fake)	Avg. F1(real)
Pheme	RDCM	76.58	63.77	67.53	88.67	74.13	43.82	80.38
	ADOSE-UDA	75.60	67.71	74.02	86.79	76.03	54.54	80.92
	Detective	77.07	75.59	74.02	87.73	78.60	47.30	83.15
	EADA	75.12	70.86	81.81	87.73	78.88	51.95	81.51
	Entropy	77.56	73.22	80.51	88.67	79.99	63.05	82.87
	RASC	78.04	76.37	72.72	88.67	78.95	58.33	82.68
	LUET	75.60	74.80	74.02	88.67	78.27	59.70	82.24
	ADOSE(ours)	**80.00**	**77.95**	**84.41**	**90.56**	**83.23**	**64.17**	**86.29**
Dataset	Method	*EDH* → *S*	*SDH* → *E*	*SEH* → *D*	*SED* → *H*	Avg. Acc	Avg. F1(fake)	Avg. F1(real)
Weibo	RDCM	75.12	77.31	77.98	88.09	79.62	72.27	79.77
	ADOSE-UDA	75.72	78.41	83.48	84.52	80.53	75.30	80.95
	Detective	81.14	82.15	80.73	89.28	83.32	77.61	84.47
	EADA	81.84	81.71	78.89	89.88	83.08	76.32	84.07
	Entropy	81.44	81.93	80.73	**91.66**	83.94	78.89	85.15
	RASC	76.93	76.43	83.48	89.88	81.68	78.23	82.05
	LUET	80.84	81.71	86.23	88.09	84.21	81.11	83.92
	ADOSE(ours)	**84.15**	**83.70**	**87.15**	**91.66**	**86.66**	**82.67**	**87.23**

Table 1: Performance comparison of domain adaptation accuracy (%) and average F1 score (%) of different methods with ADOSE on two datasets: Pheme and Weibo. C, S, O, F represent the four event domains in Pheme, and S, E, D, H represent the four topic domains in Weibo.

for Pheme, while $m = 5$ for Weibo. All codes are developed using PyTorch and an NVIDIA A100 GPU.

Evaluation Metrics. We employ accuracy as the main evaluation metric. The average accuracy (Avg. Acc) is the mean of four accuracy results obtained from different experimental settings. We additionally report the average F1 scores for fake news i.e., Avg. F1 (fake) and for real news i.e., Avg. F1 (real).

Baselines. To conduct a comprehensive evaluation of our proposed model, we compare it with unsupervised domain adaptation (UDA) and ADA approaches. UDA methods include RDCM (Liu et al. 2023a) and ADOSE-UDA which is derived from our model by removing the process of actively selecting target samples (i.e., the LUS and MDC). For ADA methods, we choose EADA (Xie et al. 2022), Detective (Zhang et al. 2024c), RASC (Zhang et al. 2024d), LUET (Sun et al. 2025) as well as the most common method, Entropy (Wang and Shang 2014).

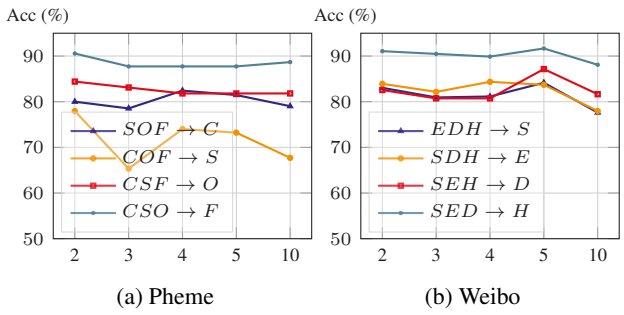
Dataset	Method	Acc	F1(fake)	F1(real)
Pheme	**ADOSE**	**83.23**	**64.17**	**86.29**
	w/o MEFN	79.23	52.33	82.67
	w/o LUS	77.67	59.43	81.72
	w/o MDC	81.18	52.38	84.54
Weibo	**ADOSE**	**86.66**	**82.67**	**87.23**
	w/o MEFN	83.65	77.41	85.19
	w/o LUS	80.45	76.29	81.02
	w/o MDC	84.06	79.30	85.09

Table 2: Ablation study results (Avg. %).

5.2 Overall Performance

We design four multi-source domain adaptation scenarios for each dataset and conduct evaluations from both domain-specific and average perspectives. Table 1 summarizes the quantitative experiment results of our framework and baselines on Pheme and Weibo, respectively. We make the following observations:

- In general, ADOSE achieves the best performance in almost all the adaptation scenarios compared to various baselines (tied with Entropy under the $SED \rightarrow H$ setting on Weibo). In particular, our method substantially outperforms others in terms of average accuracy (Avg. Acc), achieving an improvement of $3.24\% \sim 9.1\%$ on the Pheme dataset and $2.45\% \sim 7.04\%$ on the Weibo dataset. Meanwhile, our model is also optimal for the other two F1 metrics.
- Another easily observable trend is that all ADA methods outperform UDA methods, where UDA methods mainly leverage the sample features without utilizing their labels in the target domain. The trend is more pronounced in the $COF \rightarrow S$ setting of Pheme and the $EDH \rightarrow S$ of Weibo. It suggests the effectiveness of active annotation in addressing domain adaptation challenges.
- Evidence-based, energy-based, and entropy-based ADA methods (*e.g.*, Detective, EADA, Entropy), generally fail to exert their advantages in the binary news classification task, as evidenced by lower Avg. F1(fake). RASC and LUET lack consideration of the complexity introduced by multiple source domains and multimodal features. In contrast, our model adapts to domain discrepancies and multimodal characteristics, achieving optimal performance on both Avg. F1(fake) and Avg. F1(real).



(a) Pheme (b) Weibo

Figure 3: Accuracy trends with varying m (the x-axis represents the different values of m as 2, 3, 4, 5, 10).

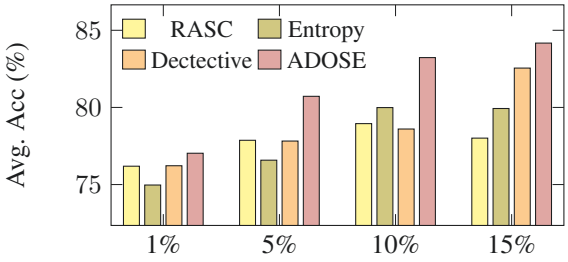


Figure 4: The results on Weibo dataset under various annotation budgets.

5.3 Ablation Study

To evaluate the effectiveness of each component in the ADOSE framework, we conduct an ablation study by omitting each component individually. The experimental settings are as follows: **(1) w/o MEFN**: The Modal-dependency Expertise Fusion Network module is omitted, and classification is performed solely by relying on the cross-modal classifier. **(2) w/o LUS**: The Least-disagree Uncertainty Selector, which quantifies the closeness of samples to the decision boundary to select uncertain target samples, is removed. **(3) w/o MDC**: The Multi-view Diversity Calculator for alleviating feature redundancy is omitted. Table 2 presents the results of ablation studies.

We observe that the original ADOSE outperforms all variants, demonstrating the effectiveness of each component. We find that the module with the greatest impact on detection accuracy is the LUS, followed by MEFN, with MDC having the least impact. This aligns with our module design intention, where LUS serves as the core selection strategy, determining the overall performance of domain adaptation, and MEFN is more important than MDC due to its role in modeling multimodal dependencies.

5.4 Additional Analysis

Hyperparameter sensitivity. We investigate the impact of the parameter m in the LUS module, which controls the relative influence of uncertainty and diversity in the selection strategy. A smaller m indicates that the LUS module plays a dominant role in the selection process, while a larger m reflects greater influence from the MDC module. Figure 3 illustrates the detection accuracy under varying m values on two datasets. The figure shows that the optimal choices of m for Pheme and Weibo are 2 and 5, respectively. This phenomenon can be explained by the fact that the Pheme dataset exhibits less inter-domain differences compared to Weibo (whose domains differ significantly in topic), making uncertainty estimation more crucial than diversity measurement.

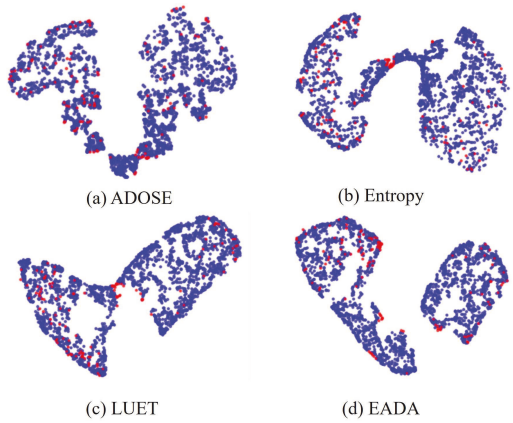


Figure 5: T-SNE visualization of target features in task $EDH \rightarrow S$ on Weibo.

Varying labeling budget. We show the performance on Pheme under various labeling budgets (1% to 15%) in Figure 4. We can observe that ADOSE outperform other ADA methods with varying labeling budgets, demonstrating that our method is suitable for various labeling budgets. Moreover, we can observe that as the labeling budget increases, our method can continuously improve the adaptation performance, demonstrating that our sample selection strategy is stable and effective, and can continuously select the most valuable samples for annotation.

Feature visualization. Figure 5 shows the t-SNE visualization results of multimodal features learned by various methods on the social domain of the Weibo dataset, where blue dots represent unlabeled target samples and red dots represent selected samples. Compared to other ADA methods, ADOSE's feature space is more concentrated, and the selected samples exhibit a uniform and diverse distribution, further confirming the effectiveness of our proposed active selection strategy.

6 Conclusion

In this paper, we introduce multi-source active domain adaptation into the multimodal fake news detection task to assist knowledge transfer in the target domain. We propose ADOSE, which addresses the semantic space and deception pattern discrepancies in news, while strategically selecting target samples for domain adaptation. The experimental results conducted on the two real-world datasets prove the effectiveness of our ADOSE model.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (Grant No.62272334, No.62572335, No.62572336 and No.62577050) and the Priority Academic Program Development of Jiangsu higher education institutions.

References

Chen, Y. 2015. *Convolutional neural network for sentence classification*. Master's thesis, University of Waterloo.

Cho, S. J.; Kim, G.; Lee, J.; Shin, J.; and Yoo, C. D. 2023. Querying Easily Flip-flopped Samples for Deep Active Learning. In *The Twelfth International Conference on Learning Representations*.

Cui, C.; and Jia, C. 2025. Towards Real-World Rumor Detection: Anomaly Detection Framework with Graph Supervised Contrastive Learning. In *Proceedings of the 31st International Conference on Computational Linguistics*, 7141–7155.

Dong, Y.; He, D.; Wang, X.; Jin, Y.; Ge, M.; Yang, C.; and Jin, D. 2024. Unveiling implicit deceptive patterns in multimodal fake news via neuro-symbolic reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 8354–8362.

Du, Z.; and Li, J. 2023. Diffusion-based probabilistic uncertainty estimation for active domain adaptation. *Advances in Neural Information Processing Systems*, 36: 17129–17155.

Han, F.; Ye, P.; Duan, S.; and Wang, L. 2024. Ada-iD: Active Domain Adaptation for Intrusion Detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 7404–7413.

Hanneke, S.; et al. 2014. Theory of disagreement-based active learning. *Foundations and Trends® in Machine Learning*, 7(2-3): 131–309.

Huang, D.; Li, J.; Chen, W.; Huang, J.; Chai, Z.; and Li, G. 2023. Divide and adapt: Active domain adaptation via customized learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7651–7660.

Kim, Y.-Y.; Cho, Y.; Jang, J.; Na, B.; Kim, Y.; Song, K.; Kang, W.; and Moon, I.-C. 2023. Saal: sharpness-aware active learning. In *International Conference on Machine Learning*, 16424–16440. PMLR.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Li, J.; Bin, Y.; Peng, L.; Yang, Y.; Li, Y.; Jin, H.; and Huang, Z. 2024. Focusing on relevant responses for multi-modal rumor detection. *IEEE Transactions on Knowledge and Data Engineering*.

Li, Y.; Lee, K.; Kordzadeh, N.; Faber, B.; Fiddes, C.; Chen, E.; and Shu, K. 2021. Multi-source domain adaptation with weak supervision for early fake news detection. In *2021 IEEE International Conference on Big Data (Big Data)*, 668–676. IEEE.

Liu, G.; Zhang, J.; Liu, Q.; Wu, J.; Wu, S.; and Wang, L. 2024a. Uni-Modal Event-Agnostic Knowledge Distillation for Multimodal Fake News Detection. *IEEE Transactions on Knowledge and Data Engineering*.

Liu, H.; Wang, W.; Sun, H.; Rocha, A.; and Li, H. 2023a. Robust domain misinformation detection via multi-modal feature alignment. *IEEE Transactions on Information Forensics and Security*, 19: 793–806.

Liu, S.; Jiang, Z.; Li, Y.; Peng, J.; Wang, Y.; and Lin, W. 2024b. Density matters: improved core-set for active domain adaptive segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 13999–14007.

Liu, X.; Li, P.; Huang, H.; Li, Z.; Cui, X.; Liang, J.; Qin, L.; Deng, W.; and He, Z. 2024c. Fka-owl: Advancing multimodal fake news detection through knowledge-augmented lvlms. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 10154–10163.

Liu, Y.; Qiao, L.; Lu, C.; Yin, D.; Lin, C.; Peng, H.; and Ren, B. 2023b. OSAN: A one-stage alignment network to unify multimodal alignment and unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3551–3560.

Lu, H.; Xie, Y.; Yang, X.; and Yan, J. 2024. Boundary Matters: A Bi-Level Active Finetuning Method. *Advances in Neural Information Processing Systems*, 37: 35945–35972.

Lu, Y.; Huang, H.; and Hu, X. 2024. Style Adaptation and Uncertainty Estimation for Multi-Source Blended-Target Domain Adaptation. *Advances in Neural Information Processing Systems*, 37: 87042–87060.

Ma, X.; Gao, J.; and Xu, C. 2021. Active universal domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8968–8977.

Ma, Z.; Luo, M.; Guo, H.; Zeng, Z.; Hao, Y.; and Zhao, X. 2024. Event-radar: Event-driven multi-view learning for multimodal fake news detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5809–5821.

Madaan, D.; Makino, T.; Chopra, S.; and Cho, K. 2024. Jointly Modeling Inter-& Intra-Modality Dependencies for Multi-modal Learning. *Advances in Neural Information Processing Systems*, 37: 116084–116105.

Nakamura, Y.; Ishii, Y.; and Yamashita, T. 2024. Active Domain Adaptation with False Negative Prediction for Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 28782–28792.

Prabhu, V.; Chandrasekaran, A.; Saenko, K.; and Hoffman, J. 2021. Active domain adaptation via clustering uncertainty-weighted embeddings. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8505–8514.

Ran, H.; and Jia, C. 2023. Unsupervised cross-domain rumor detection with contrastive learning and cross-attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 13510–13518.

Sun, Y.; Shi, G.; Dong, W.; Li, X.; Dong, L.; and Xie, X. 2025. Local Uncertainty Energy Transfer for Active Domain Adaptation. *IEEE Transactions on Image Processing*.

Tong, Y.; Lu, W.; Zhao, Z.; Lai, S.; and Shi, T. 2024. MMDFND: Multi-modal Multi-Domain Fake News Detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 1178–1186.

Wang, D.; and Shang, Y. 2014. A new active labeling method for deep learning. In *2014 International joint conference on neural networks (IJCNN)*, 112–119. IEEE.

Wang, J.; Zhang, H.; Liu, C.; and Yang, X. 2024. Fake news detection via multi-scale semantic alignment and cross-modal attention. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2406–2410.

Wang, L.; Zhang, C.; Xu, H.; Xu, Y.; Xu, X.; and Wang, S. 2023. Cross-modal contrastive learning for multimodal fake news detection. In *Proceedings of the 31st ACM international conference on multimedia*, 5696–5704.

Wang, Y.; Ma, F.; Jin, Z.; Yuan, Y.; Xun, G.; Jha, K.; Su, L.; and Gao, J. 2018. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*, 849–857.

Wang, Y.; Ma, F.; Wang, H.; Jha, K.; and Gao, J. 2021. Multimodal emergent fake news detection via meta neural process networks. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, 3708–3716.

Xiao, W.; Gu, J.; and Liu, H. 2024. Category-aware active domain adaptation. In *Forty-first International Conference on Machine Learning*.

Xie, B.; Yuan, L.; Li, S.; Liu, C. H.; Cheng, X.; and Wang, G. 2022. Active learning for domain adaptation: An energy-based approach. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 8708–8716.

Xie, M.; Li, S.; Zhang, R.; and Liu, C. H. 2023. Dirichlet-based uncertainty calibration for active domain adaptation. *arXiv preprint arXiv:2302.13824*.

Yang, R.; Ma, J.; Lin, H.; and Gao, W. 2022. A weakly supervised propagation model for rumor verification and stance detection with multiple instance learning. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, 1761–1772.

Zhang, H.; Qian, S.; Fang, Q.; and Xu, C. 2020. Multimodal disentangled domain adaption for social media event rumor detection. *IEEE Transactions on Multimedia*, 23: 4441–4454.

Zhang, L.; Zhang, X.; Zhou, Z.; Huang, F.; and Li, C. 2024a. Reinforced adaptive knowledge learning for multimodal fake news detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 16777–16785.

Zhang, Q.; Liu, J.; Zhang, F.; Xie, J.; and Zha, Z.-J. 2024b. Natural language-centered inference network for multi-modal fake news detection. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, 2542–2550.

Zhang, W.; Lv, Z.; Zhou, H.; Liu, J.-W.; Li, J.; Li, M.; Li, Y.; Zhang, D.; Zhuang, Y.; and Tang, S. 2024c. Revisiting the domain shift and sample uncertainty in multi-source active domain transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16751–16761.

Zhang, Z.; Shen, C.; Lü, S.; and Zhang, S. 2024d. Reconfigurability-aware selection for contrastive active domain adaptation. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 5545–5553.

Zubiaga, A.; Liakata, M.; and Procter, R. 2017. Exploiting context for rumour detection in social media. In *Social Informatics: 9th International Conference, SocInfo 2017, Oxford, UK, September 13-15, 2017, Proceedings, Part I 9*, 109–123. Springer.