Robust Domain Misinformation Detection via Multi-modal Feature Alignment

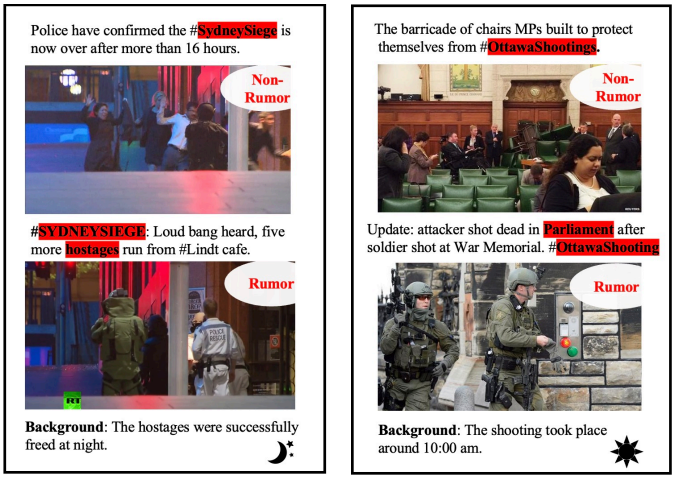
Hui Liu, Wenya Wang, Hao Sun, Anderson Rocha, and Haoliang Li

***Abstract*—Social media misinformation harms individuals and societies and is potentialized by fast-growing multi-modal content (i.e., texts and images), which accounts for higher "credibility" than text-only news pieces. Although existing supervised misinformation detection methods have obtained acceptable performances in key setups, they may require large amounts of labeled data from various events, which can be time-consuming and tedious. In turn, directly training a model by leveraging a publicly available dataset may fail to generalize due to domain shifts between the training data (a.k.a. source domains) and the data from target domains. Most prior work on domain shift focuses on a single modality (e.g., text modality) and ignores the scenario where sufficient unlabeled target domain data may not be readily available in an early stage. The lack of data often happens due to the dynamic propagation trend (i.e., the number of posts related to fake news increases slowly before catching the public attention). We propose a novel robust domain and cross-modal approach (RDCM) for multi-modal misinformation detection. It reduces the domain shift by aligning the joint distribution of textual and visual modalities through an inter-domain alignment module and bridges the semantic gap between both modalities through a cross-modality alignment module. We also propose a framework that simultaneously considers application scenarios of domain generalization (in which the target domain data is unavailable) and domain adaptation (in which unlabeled target domain data is available). Evaluation results on two public multi-modal misinformation detection datasets (Pheme and Twitter Datasets) evince the superiority of the proposed model.**

***Index Terms*—misinformation detection, domain generalization, domain adaptation, modality alignment, social media, and multimedia forensics.**

I. Introduction

MISINFORMATION has become a significant concern in contemporary society, threading all aspects of individuals and society [1, 2], because online social media lack serious verification processes and netizens usually cannot discriminate between fake and real news [3]. For example, during the 2016 presidential election cycle in the United States, false news stories claiming that Hillary Clinton ordered the murder of an FBI agent and participated in a satanic child abuse ring in a Washington pizza parlor were shared ostensibly through social media [4, 5]. While expert-based (e.g., PolitiFact[1], GossipCop[2]) and crowd-based efforts (such as Amazon Mechanical Turk[3]) for manual fact-checking tools have carried precious insights for misinformation detection, they cannot scale with the volume of news on social media [1].



(a) Sydney Siege (b) Ottawa Shooting

Fig. 1: Examples of Sydney Siege and Ottawa Shooting domains from Pheme Dataset. Sydney Siege was a terrorist attack in which a gunman held hostage ten customers and eight employees in Sydney on December 15-16, 2014. Ottawa Shooting took place on Ottawa's Parliament Hill, leading to the death of a Canadian soldier on October 22, 2014.

Various methods have been proposed to perform misinformation detection based on textual features [6]–[8] and propagation patterns [9]–[11]. As the increasing misinformation with images disseminates more quickly and is more believable, another line of exploration [12]–[15] exploits multi-modal features to verify misinformation. Despite the success of these algorithms, they typically require considerably large labeled datasets, which may not be feasible for real-world applications as data collection and annotation can be cumbersome and time-consuming.

Moreover, directly training with large-scale datasets may not generalize well to unseen events on account of the domain shift [16]–[19], as there exist discrepancies between data distributions across different domains, such as word frequency and image style as Fig. 1 depicts. For example, "Sydney Siege" and "hostages" frequently occur in the Sydney Siege event[4], while "Parliament" and "Ottawa" for Ottawa Shooting[5]. Additionally, the illumination conditions are dark and bright for these two events, caused by the different times of occurrence.

Recent studies resort to transfer learning to learn domain-

[1]https://www.politifact.com/.
[2]https://www.gossipcop.com/.
[3]https://www.mturk.com/.
[4]https://en.wikipedia.org/wiki/Lindt_Cafe_siege.
[5]https://en.wikipedia.org/wiki/2014_shootings_at_Parliament_Hill,_Ottaw.

robust misinformation detector through mitigating the distribution discrepancy between the source (a.k.a., training data) and the target domain (a.k.a., testing data) [18, 20]. However, there still exist two main limitations. First, intuitively, during the dissemination of a specific news event, the number of relevant posts increases slowly at first and rapidly when catching significant public attention [21, 22]. This indicates we cannot obtain sufficient data for the target domain early on. Hence, the methods above cannot be swiftly applied in this case as they require labeled [19, 23, 24] and unlabeled target domain data [16]–[20, 23, 24] to be available during training. Secondly, existing methods for cross-domain misinformation detection ignore the issue of discrepancy between visual and textual modalities. We argue that directly performing distribution alignment across domains without considering the gap between different modalities may not be optimal for capturing robust domain information for multi-modal misinformation detection.

We propose a unified, robust domain and cross-modality framework named **RDCM** for multi-modal misinformation detection that seeks to address the limitations above. The unified framework can be applied to two application scenarios: 1) real-time misinformation detection (i.e., when target domain data are not accessible during training, corresponding to domain generalization); and 2) offline misinformation detection (i.e., when unlabeled target domain data are available during training, which corresponds to domain adaptation).

To align multi-modal distributions and mitigate the modality gap between source and target domains, we propose to leverage an inter-domain alignment module based on the joint distribution of textual and visual features and a cross-modality alignment module based on contrastive learning for the multi-modal misinformation detection task. The inter-domain alignment module measures the joint distribution of modalities (i.e., image and text) based on the kernel mean embedding, reproducing the kernel Hilbert space (RKHS) [25] and then aligns the joint distribution of different domains by minimizing the corresponding Maximum Mean Discrepancy (MMD) [26].

We align distributions among multiple source domains for the scenario which requires real-time applications (a.k.a. domain generalization) and align distributions between each source and the target domain for the scenario where misinformation detection can be performed offline (a.k.a. domain adaptation).

Inspired by contrastive learning in self-supervised tasks [27]–[29], the cross-modality alignment module exploits contrastive learning to bridge the modality gap with a novel sampling strategy tailored for multi-modal misinformation detection. After inter-domain and cross-modal (i.e., feature alignment across different modalities in a single domain) alignment, we expect to extract domain-invariant textual and visual features of multi-modal posts and concatenate them for misinformation detection. The empirical study shows that our model yields state-of-the-art results on two public datasets.

The key contributions of this work are:

- A unified framework that tackles the domain generalization (target domain data is unavailable) and domain adaptation tasks (target domain data is available). This is necessary as obtaining sufficient unlabeled data in the target domain at an early stage of misinformation dissemination is difficult;
- Inter-domain and cross-modality alignment modules that reduce the domain shift and the modality gap. These modules aim at learning rich features that allow misinformation detection. Both modules are plug-and-play and have the potential to be applied to other multi-modal tasks.

II. Related Work

This section reviews domain generalization (DG), domain adaptation (DA), and robust domain misinformation detection.

A. Domain Generalization and Domain Adaptation

Supervised machine learning algorithms assume similar training and testing distributions, but practical deployment requires models to generalize well on unseen, out-of-distribution data. Domain generalization (DG) and domain adaptation (DA) address this challenge. DG learns from one or multiple source domains, while DA requires access to target domain data during training, making DG more difficult.

Domain generalization is widely used in computer vision and natural language processing. A recent survey [30] classified DG methods into three categories: data manipulation, representation learning, and learning strategy.

Data manipulation involves generating samples through data augmentation [31, 32] or data generation methods [33] to increase the diversity and quantity of source domain data.

Representation learning works are inspired by the theory that domain invariant representations are transferable to unseen domains [34]. These works aim to learn robust domain representation extraction functions by either aligning feature distributions among source domains [35]–[37] or disentangling features into different sub-spaces (domain-specific and domain-sharing space) [38, 39]. For instance, Li et al. [35] used adversarial autoencoders with Maximum Mean Discrepancy (MMD) distance to align distributions across different domains and learn a generalized latent feature representation. Ding and Fu [38] designed domain-specific and domain-sharing networks for the disentanglement in individual domains and across all domains, respectively.

Finally, the learning strategy-based DG methods focus on machine learning paradigms to enhance the generalization performance, such as meta-learning [40], ensemble learning [41], gradient-based DG [42], among others.

Domain adaptation methods differ from domain generalization in that they require access to target domain data during the training process [43]–[45]. These methods are categorized into two groups for single source domain visible during adaptation (SDA). One group uses explicit discrepancy measures, like H-divergence [46], MMD [25, 26], Wasserstein Distance [47, 48], and second-order statistics [43], to reduce the shift between source and target distributions. The other group employs adversarial learning, where a domain discriminator is confused in a min-max manner [49], to implicitly align the source and target distributions. Additionally, early theoretical analysis [50, 51] demonstrated that minimizing a weighted combination of source risks can achieve lower target error.

The above methods can also be applied when data from multiple source domains are available during training (MDA). Peng et al. [52] dynamically aligned moments of feature distributions of multiple source domains and the target domain with theoretical insights. Zhu et al. [53] proposed a two-stage alignment framework that aligned distributions of each pair of source and target domains and the outputs of classifiers. Despite the progress, effectively applying DG and DA methods to multi-modal settings with large semantic gaps among different modalities remains unsolved.

B. Robust Domain Misinformation Detection

The widespread presence of misinformation on social media has escalated the issue of social distrust, drawing significant attention from both society and the research community. However, many existing misinformation detection methods [6, 7, 9]–[14, 54] are domain-specific and may not perform effectively on unseen domains due to the domain shift. Moreover, these methods require extensive and diverse training data, which is impractical given the rapid accumulation of events and news.

Robust domain methods were developed aiming at the domain shift. Some work [19, 20, 23, 24] fall into domain adaptation, assuming access to target domain data during training. For instance, Mosallanezhad et al. [24] proposed a domain adaptive detection framework using reinforcement learning and incorporating auxiliary information. Silva et al. [20] introduced an unsupervised technique for selecting unlabeled news records to maximize domain coverage and preserve domain-specific and cross-domain knowledge through disentangle learning.

However, these methods may not accommodate the dynamic nature of misinformation generation and propagation, where target domain data might be unavailable during training. Limited access to timely target domain data hinders their real-time application. Another group of works explores using powerful search engines (e.g., Google) to retrieve background knowledge for fact-checking [55, 56]. Yet, unverified online information introduces noise that can negatively impact performance.

III. Proposed method

The multi-modal multi-domain misinformation detection framework comprises four components: Multi-modal Representation Extraction (Text and Image Encoders), Inter-domain Alignment, Cross-modality Alignment, and Classification. Textual and image features are extracted from a post using the corresponding encoders. The Inter-domain Alignment module removes domain-specific information while preserving domain-agnostic information. The Cross-modality Alignment combines textual and visual representations. The combined domain-robust and modality-aligned features are then used for misinformation detection. While designed for domain generalization (DG), the framework can be extended to unsupervised domain adaptation (DA) by adapting the inter-domain module to align distributions between source and target domains.

A. Task Definition

The goal of multi-modal misinformation detection is to determine the authenticity of a text and an associated image, classifying the pair as fake (rumor) or real (non-rumor). To address challenges posed by fast-emerging events and costly annotations, researchers have explored various domain adaptation methods [19, 20, 23, 23, 24] to learn robust domain features and mitigate domain shifts. However, these methods overlook the difficulty of collecting sufficient data in the target domain during the early stages of fake news dissemination and fail to consider the presence of multiple modalities in real-world news pieces. To address these issues, we propose a unified framework to handle the multi-modal misinformation detection task, making it suitable for both domain generalization (DG) and domain adaptation (DA) scenarios.

Formally, given $\mathcal{D}_S = \left\{\mathcal{D}_S^1, \mathcal{D}_S^2, \ldots, \mathcal{D}_S^M\right\}$ the collection of M labeled source domains and $\mathcal{D}_T$ the unlabeled target domain where all domains are defined based on different news events, our method aims to find a hypothesis in the given hypothesis space, which minimizes the classification error on $\mathcal{D}_T$. Each source domain can be represented as $\mathcal{D}_S^m = \{(t_n^m, v_n^m), y_n^m\}_{n=1}^{N_m}$ and the target domain can be denoted as $\mathcal{D}_T = \{(t_n, v_n)\}_{n=1}^{N_T}$, where N_m $(1 \leq m \leq M)$ is the number of samples in the m-th source domain, N_T is the number of samples in the target domain, and $y \in \{0, 1\}$ is the gold label (1 indicates fake information for the Twitter Dataset or the rumor for the Pheme Dataset and 0 otherwise). Additionally, (t, v) is a text-image pair, where t is a text sentence, and v is the corresponding image. We assume **no availability** of target domain data $\mathcal{D}_T$ in the scenario of DG.

B. Multi-modal Representation Extraction

Given an input text-image pair (t, v)[6] in each domain, following previous work [16, 18], we leverage a convolutional neural network (i.e., TextCNN [57]) with an additional two-layer perceptron (MLP) as the textual encoder to obtain the representation of t as $\mathbf{x}_t$:

$$\mathbf{x}_t = \mathbf{E}_t(t; \boldsymbol{\theta}_t), \tag{1}$$

where $\mathbf{x}_t \in \mathbb{R}^d$ is the final representation of t, $\mathbf{E}_t$ represents the textual encoder, and $\boldsymbol{\theta}_t$ represents the parameter of TextCNN and corresponding MLP. As large-scale pre-trained models have excelled in natural language processing tasks, we adopt the word embedding extracted by RoBERTa [58] as initializing word vectors of TextCNN, following existing work [16, 59]. The reason why we do not fine-tune RoBERTa is to avoid over-parameterization, which may harm the generalization ability of the model.

For image representation, given an image v, following existing methods [8, 19, 23], we use ResNet50 as the visual backbone neural network and choose the feature of the final pooling layer as the initial visual embedding. Then, similar to

[6]We omit the subscript for simplicity unless specifically stated.

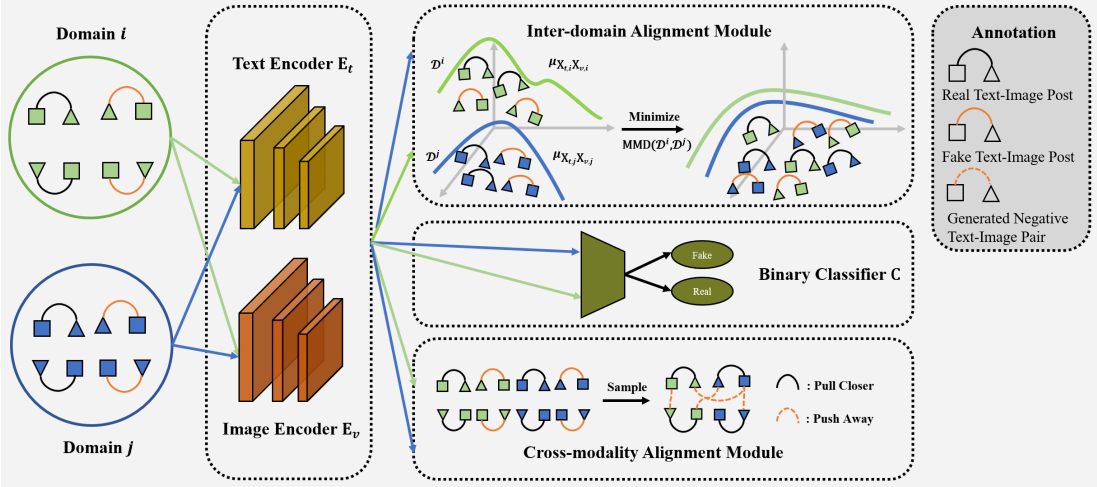


Fig. 2: Proposed robust domain and cross-modal framework. In the DG setup, we take multiple source domains as input and extract textual and visual features through the Text Encoder and the Image Encoder. Then we align the joint distributions of textual and visual modalities between each source domain pair by Inter-domain Alignment Module, reduce the modality gap by Cross-modality Alignment Module, and detect misinformation of source domains through Binary Classification. The DA setup takes multiple sources and the target domain as input. Compared with DG, it further aligns joint distributions between each source domain and the target domain but only performs cross-modal alignment and trains the classifier on source domains.

the text modality, we also use a MLP to reduce its dimension to d given as

$$\mathbf{x}_v = \mathbf{E}_v(v; \boldsymbol{\theta}_v), \tag{2}$$

where $\mathbf{x}_v \in \mathbb{R}^d$ is the final representation of the image v, $\mathbf{E}_v$ represents the visual encoder, and $\boldsymbol{\theta}_v$ represents the parameter of ResNet50 and the visual MLP.

We use $\mathbf{X}_t$ and $\mathbf{X}_v$ to denote random variables instantiated by $\mathbf{x}_t$ and by $\mathbf{x}_v$ in one domain. After extracting the textual and visual features of each text-image pair for multiple source domains $\{\mathcal{D}_S^m\}_{1\leq m\leq M}$ and target domain $\mathcal{D}_T$, we can empirically estimate the probability distribution of textual features $\mathbb{P}(\mathbf{X}_t)$ and the probability distribution of visual features $\mathbb{P}(\mathbf{X}_v)$ by drawing samples i.i.d. from variables $\mathbf{X}_t$ and $\mathbf{X}_v$ from each domain.

C. Multi-modal Feature Alignment

Multi-modal feature alignment aims to extract robust domain information for misinformation detection; as such, the trained model can be better generalized to unseen events. However, existing cross-domain-based methods for misinformation detection can be limited as most of them only focus on a single modality for misinformation detection. While one can perform marginal distribution alignment on textual features $\mathbf{X}_t$ and visual features $\mathbf{X}_v$, separately, or perform distribution alignment through feature concatenation or element-wise production [19, 24, 60], the correlation property across multiple modalities has been ignored, which may hinder robust domain misinformation detection when having textual and visual information as input. To tackle this limitation, we propose to explore domain covariance information on both the event level (i.e., domain) and sample level, corresponding to Inter-domain Alignment and Cross-modality Alignment, respectively.

1) Inter-domain Alignment: Among various inter-domain alignment methods based on distribution measurement, Maximum Mean Discrepancy (MMD) [26] has been proven to be effective where the distribution of samples can be formulated through kernel mean embedding [25] in a non-parametric manner. One intuitive way is to align the marginal distribution of textual and visual modality across domains through MMD, which can be defined as

$$\text{MMD}(\mathcal{D}_S^i, \mathcal{D}_S^j) = \|\mu_{\mathbf{X}_{t,i}} - \mu_{\mathbf{X}_{t,j}}\|_{\mathcal{H}}^2 + \|\mu_{\mathbf{X}_{v,i}} - \mu_{\mathbf{X}_{v,j}}\|_{\mathcal{H}}^2. \tag{3}$$

We use samples from the i-th and j-th source domains as example. μ denotes the kernel mean embedding operation in reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ [25], which is to compute the mean of latent features in the RKHS as $\mu_{\mathbf{X}}(\mathbb{P}) := \mathrm{E}_{\mathbf{X}}[\phi(\mathbf{X})] = \int_X \phi(x)\, \mathrm{d}\mathbb{P}(x)$ and ϕ denotes a kernel function. Here $\mu_{\mathbf{X}_{t,i}}$ and $\mu_{\mathbf{X}_{v,j}}$ indicate the textual mean embedding for the i-th source domain and the visual mean embedding for the j-th source domain, respectively. However, directly performing marginal distribution alignment may not capture the correlation information between textual and visual modalities. We propose to align the joint feature distribution upon textual and visual modalities where the kernel mean embedding can be formulated through the covariance operator $\otimes$ on RKHS [61] as

$$\mu_{\mathbf{x}_t, \mathbf{x}_v} = \mathrm{E}[\phi_t(\mathbf{X}_t) \otimes \phi_v(\mathbf{X}_v)]. \tag{4}$$

We can better capture the cross-covariance dependency between textual and visual modalities, contributing to robust domain multi-modal misinformation detection, and the new inter-domain alignment MMD can be formulated as

$$\text{MMD}(\mathcal{D}_S^i, \mathcal{D}_S^j) = \|\mu_{\mathbf{X}_{t,i}, \mathbf{X}_{v,i}} - \mu_{\mathbf{X}_{t,j}, \mathbf{X}_{v,j}}\|_{\mathcal{H}}^2. \tag{5}$$

We seek for the empirical estimate of $\text{MMD}(\mathcal{D}_S^i, \mathcal{D}_S^j)$ [61] which can be computed as

$$\begin{aligned}\text{MMD}(\mathcal{D}_S^i, \mathcal{D}_S^j) = &\frac{1}{N_i^2}\sum_{p=1}^{N_i}\sum_{q=1}^{N_i} k_v(\mathbf{x}_{v,i,p}, \mathbf{x}_{v,i,q})k_t(\mathbf{x}_{t,i,p}, \mathbf{x}_{t,i,q}) \\ &+\frac{1}{N_j^2}\sum_{p=1}^{N_j}\sum_{q=1}^{N_j} k_v(\mathbf{x}_{v,j,p}, \mathbf{x}_{v,j,q})k_t(\mathbf{x}_{t,j,p}, \mathbf{x}_{t,j,q}) \\ &-\frac{2}{N_i N_j}\sum_{p=1}^{N_i}\sum_{q=1}^{N_j} k_v(\mathbf{x}_{v,i,p}, \mathbf{x}_{v,j,q})k_t(\mathbf{x}_{t,i,p}, \mathbf{x}_{t,j,q}),\end{aligned} \tag{6}$$

where $\mathbf{x}_{v,i,p}$ denotes the latent feature of the p-th sample from the modality v of the domain i, k_t and k_v are Gaussian kernel functions to map extracted features $\mathbf{x}_t$ and $\mathbf{x}_v$ into RKHS, corresponding to the textual modality and visual modality, respectively.

Assume we have training samples from M different domains as $\mathcal{D}_S = \{\mathcal{D}_S^1, \mathcal{D}_S^2, \ldots, \mathcal{D}_S^M\}$, the inter-domain alignment loss based on data collected from textual and visual modalities can be formulated as

$$\mathcal{L}_{\text{inter}} = \binom{2}{M}\sum_{i=1}^{M-1}\sum_{j=i+1}^{M}\text{MMD}(\mathcal{D}_S^i, \mathcal{D}_S^j). \tag{7}$$

When some data for testing are available during training, we can extend the inter-domain alignment loss above by incorporating it with target domain data D^T as

$$\begin{aligned}\mathcal{L}_{\text{inter}} = &\binom{2}{M}\sum_{i=1}^{M-1}\sum_{j=i+1}^{M}\text{MMD}(\mathcal{D}_S^i, \mathcal{D}_S^j) \\ &+\binom{1}{M}\sum_{i=1}^{M}\text{MMD}(\mathcal{D}_S^i, \mathcal{D}_T).\end{aligned} \tag{8}$$

2) Cross-modality Alignment: Besides exploring domain-wise correlations between the textual and visual modalities, we are also interested in mining sample-wise correlations (i.e., aligning the textual and visual modalities based on a single sample). Hence, we propose a novel contrastive loss for the cross-modality alignment module to model pairwise relations between texts and images by pulling semantically similar pairs closer while pushing dissimilar ones away. Though recent vision-language contrastive learning methods have shown promising results in learning meaningful representations [29, 62], their sampling strategies for drawing positive and negative pairs may not be suitable for misinformation detection. Specifically, existing sampling methods derive positive pairs from the original input and negative pairs via random sampling in one minibatch. However, in our setting, cross-modal correspondence or similarity is more likely to only exist in real news rather than in misinformation scenarios. Besides, texts for different misinformation examples may use the same image in a specific event, which results in the image and text of many negative samples being close to each other in the semantic space. The observations above motivate us to design a metric for text-image similarity measurement, which can be further utilized for negative sample selection, contributing to cross-modality alignment through contrastive learning.

To tackle the above problems, we propose a novel sampling strategy by only taking real posts as positive samples and filtering out negative samples with high semantic similarity on the visual modality with a weighting function as follows:

$$\mathbb{I}((\mathbf{x}_{t,p}, \mathbf{x}_{v,p}), (\mathbf{x}_{t,q}, \mathbf{x}_{v,q})) = \begin{cases}0, & \text{if } \text{sim}(\mathbf{h}_p, \mathbf{h}_q) \geq \beta \\ \beta - \text{sim}(\mathbf{h}_p, \mathbf{h}_q) & \text{else},\end{cases} \tag{9}$$

Here p and q denote indices corresponding to the p-th and q-th samples in a minibatch, $\mathbf{h}_p$ and $\mathbf{h}_q$ denote the output of feature processing on $\mathbf{x}_{v,p}$ and $\mathbf{x}_{v,q}$ respectively[7]. $\text{sim}(\mathbf{h}_p, \mathbf{h}_q) = (\frac{\mathbf{h}_p \mathbf{h}_q^\top}{\|\mathbf{h}_p\|\|\mathbf{h}_q\|} + 1)/2$ represents the similarity between $(\mathbf{x}_{t,p}, \mathbf{x}_{v,p})$ and $(\mathbf{x}_{t,q}, \mathbf{x}_{v,q})$, and β is a threshold to remain semantic dissimilar pairs as negative samples. Regarding the feature processing function, one can choose an identity mapping for feature processing on visual modality. However, for the problem of misinformation detection, we are more interested in the instance-level information (i.e., object) instead of semantic information contained in the latent features. As a result, we take for $\mathbf{h}$ the output of the softmax layer of the backbone for the visual modality (e.g., ResNet50 in our model) which can measure the instance-level similarity between images well [63].

Especially, it is a good surrogate for similarity between $\mathbf{x}_{t,p}$ and $\mathbf{x}_{v,q}$ when we assume $\mathbf{x}_{t,p}$ and $\mathbf{x}_{v,p}$ of real posts are semantically relevant.

After performing a sample section to get the positive and negative text-image pairs, we leverage the contrastive loss objective in [28] and enhance it by our weighting function to learn cross-modal semantic alignment on source domains $\mathcal{D}_S$, which can be formulated as follows:

$$\mathcal{L}_{intra} = -\log\frac{e^{\frac{\tilde{\mathbf{x}}_{t,p}\tilde{\mathbf{x}}_{v,p}^\top}{\tau}}}{e^{\frac{\tilde{\mathbf{x}}_{t,p}\tilde{\mathbf{x}}_{v,p}^\top}{\tau}} + \sum_{q\neq p}^{B} e^{\frac{\tilde{\mathbf{x}}_{t,p}\tilde{\mathbf{x}}_{v,q}^\top}{\tau}}\mathbb{I}((\mathbf{x}_{t,p}, \mathbf{x}_{v,p}), (\mathbf{x}_{t,q}, \mathbf{x}_{v,q}))}, \tag{10}$$

where p represents the indices of real posts in a minibatch, q represents the indices of the other samples in this minibatch except the p-th sample, B is the minibatch size, and τ is a temperature hyperparameter. Additionally, we normalize $\mathbf{x}_t$ and $\mathbf{x}_v$ to $\tilde{\mathbf{x}}_t$ and $\tilde{\mathbf{x}}_v$ based on L_2 normalization to restrict the range of similarity scores, which have been widely adopted in [27, 28, 64]. Compared with the original loss in [28], our proposed $\mathcal{L}_{intra}$ can further push $\mathbf{x}_{v,q}$ of the hard negative samples far away from corresponding $\mathbf{x}_{t,p}$ in the shared feature space and mitigate the influence of inappropriate random sampling for multi-modal tasks [13, 64] to perform better modality alignment.

D. Classification

Given the textual feature $\mathbf{x}_t$ and visual feature $\mathbf{x}_v$ of one post (t, v) in source domains $\mathcal{D}_S$, we concatenate them for the final prediction:

$$\hat{\mathbf{y}} = \mathbf{C}(\mathbf{x}_t, \mathbf{x}_v; \boldsymbol{\theta}_c). \tag{11}$$

Here $\mathbf{C}$ is a classifier consisting of a MLP followed by a softmax activation function, $\boldsymbol{\theta}_c$ is its parameters, and $\hat{\mathbf{y}}$ is the predicted label. Then, the classifier is trained with cross-entropy loss against the ground-truth label $\mathbf{y}$ on source domains $\mathcal{D}_S$

[7]We omit the domain index here for simplicity.

as $\mathcal{L}_{cls} = -\mathbf{y}\log(\hat{\mathbf{y}})$, in which label 1 represents fake posts (rumors) while 0 means real posts (non-rumors) in our task.

In this work, we are especially concerned with the robust domain multi-modal misinformation detection that requires a model to simultaneously map textual and visual features into domain-invariant and modality-aligned semantic space to improve classification performance. As such, we combine $\mathcal{L}_{inter}$ in Eq. 7, $\mathcal{L}_{intra}$ in Eq. 10 and $\mathcal{L}_{cls}$ as the final form of our training objective in DG situation:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{inter} + \lambda_2 \mathcal{L}_{intra} + \mathcal{L}_{cls}, \tag{12}$$

where λ_1 and λ_2 are weighting parameters to balance the importance of $\mathcal{L}_{inter}$, $\mathcal{L}_{intra}$ and $\mathcal{L}_{cls}$. Moreover, we can easily extend our method to the DA situation by replacing $\mathcal{L}_{inter}$ in Eq. 7 as $\mathcal{L}_{inter}$ in Eq. 8 without changing our framework.

IV. Experiments

We devise experiments to answer the following research questions. For conciseness, **RQ4** and **RQ5**, the elaborate analysis of the inter-domain alignment module and cross-modality alignment module, are explained in the Appendix.

- **RQ1:** Do unlabeled target domain data and multiple modalities boost domain misinformation detection?
- **RQ2**: How effective is the proposed robust domain and cross-modal detection method (**RDCM**) compared with existing methods for misinformation detection?
- **RQ3:** How do the components of **RDCM** affect results?
- **RQ4:** How effective is the method to mitigate the domain shift by aligning the joint distribution of text and visual features represented by kernel mean embedding?
- **RQ5:** How effective is the sampling strategy for the cross-modality alignment module?

A. Data Preparation

We adopt two benchmark datasets, Pheme [65] and Twitter [66], to validate the effectiveness of the proposed misinformation detection approach **RDCM**.

Pheme dataset is constructed by collecting tweets related to five breaking news events: *Charlie Hebdo*, *Sydney Siege*, *Ferguson Unrest*, and *Ottawa Shooting* and *Germanwings Crash*. As the original Pheme dataset does not include images, we obtain relevant images through the Twitter API using the tweet ID contained in each sample if the sample has attached images, following [19]. In this work, we detect misinformation by incorporating text and image information. Thus, we remove the tweets without any text or image and finally get four event domains. If multiple images are attached to one post, we randomly retain one image and discard the others. The detailed statistics are listed in Table I.

The Twitter dataset collects text content, attached images/videos, and social context information related to 11 events. However, several events are removed from the experiments because of only having real or fake posts. Following the data cleaning method for the Pheme dataset, we only preserve samples containing texts and images and obtain four event domains, including *Hurricane Sandy*, the *Boston Marathon bombing*, *Malaysia*, and *Sochi Olympics*. It is worth noting that many samples have the same image in this dataset, which challenges the generation of negative multi-modal pairs for contrastive learning. The detailed statistics are listed in Table II.

Regarding the criterion of labels, in the Pheme dataset, the sample is labeled as a rumor when it is unverified[8] at the time of posting, it is labeled as non-rumor when it belongs to the other circulating information [67, 68]. Moreover, in the Twitter dataset, the sample is identified as fake when it shares an image that does not represent the event it refers to (e.g., maliciously tampering with images and reposting previously captured images in a different event). At the same time, they are considered real when it shares an image that legitimately represents the event it refers to [66]. As a result, a huge discrepancy exists between domains from different datasets.

To further verify the generalization of the proposed approach, we conduct Cross-dataset experiments between these two datasets. Especially we select three source domains from either the Pheme or Twitter dataset to train the model and evaluate its performance on the target domain from the other dataset. Finally, the results of four cases $\mathcal{COF} \rightarrow \mathcal{M}$, $\mathcal{CSF} \rightarrow \mathcal{A}$, $\mathcal{ABI} \rightarrow \mathcal{S}$ and $\mathcal{ABI} \rightarrow \mathcal{O}$ are reported in our experiments.

TABLE I: Statistics of Pheme Dataset

Event	Rumor	Non-Rumor	All
Charlie Hebdo ($\mathcal{C}$)	181	742	923
Sydney Siege ($\mathcal{S}$)	191	228	419
Ferguson Unrest ($\mathcal{F}$)	42	309	351
Ottawa Shooting ($\mathcal{O}$)	146	110	256

TABLE II: Statistics of Twitter Dataset

Event	Fake	Real	All
Hurricane Sandy ($\mathcal{A}$)	5461	6841	12302
Boston Marathon bombing ($\mathcal{B}$)	81	325	406
Malaysia ($\mathcal{M}$)	310	191	501
Sochi Olympics ($\mathcal{I}$)	274	127	398

B. Experimental Setup

1) Baselines: For comparison purposes, we adopt baselines from four categories: uni-modality, multi-modality, domain generalization, and domain adaptation baselines.

Uni-modality baselines comprise **TextCNN-rand**, **TextCNN-roberta**, **Bert** [69], and **ResNet** [70]. **TextCNN-rand**, **TextCNN-roberta**, and **Bert** are text modality-based models which only exploit textual information for classification. Both **TextCNN-rand** and **TextCNN-roberta** are based on TextCNN framework [57]. Their difference is that the workpiece embedding of **TextCNN-rand** uses random initialization, and **TextCNN-roberta** is initialized from the RoBERTa-base[9], which is frozen during training, following [18, 23]. **Bert** is a transformer-based pre-trained model, and we utilize one of its variants[10] to generate the embedding of [CLS] token for detection. We compare the model with the visual modality

[8]One post is defined as unverified when there is no evidence supporting it (e.g., logically self-consistent between the text and image) or there is no official confirmation from authoritative sources.

[9]https://huggingface.co/roberta-base

[10]https://huggingface.co/bert-base-uncased

method **ResNet** [70], which replaces the final classification layer as a binary classification layer for misinformation detection.

Multi-modality baselines include **Vanilla** [71] and **ModalityGat** [72], which take TextCNN and ResNet as textual and visual encoders, respectively. **Vanilla** concatenates textual and visual features to perform classification, similar to our proposed method without Inter-domain Alignment and Cross-modality Alignment components. On the other hand, **ModalityGat** introduces a gate mechanism to fuse the information from different modalities based on their corresponding importance.

Domain generalization baselines consist of **EANN** [18], **IRM** [73], **MLDG** [40] and **Fish** [42], among which the first two belong to representation learning based DG, and the last two belong to learning strategy based DG. In detail, **EANN** confuses an event domain discriminator in an adversarial manner to learn shared features among multiple events. **IRM** aims to estimate invariant and causal predictors from multiple source domains to improve the generalization performance on the target domain. **MLDG** is a meta-learning framework that simulates domain shift by synthesizing virtual meta-train and meta-test sets in each mini-batch. Finally, **Fish** matches the distribution of many source domains by maximizing the inner product between gradients of these domains. While there exists some work using data augmentation to improve the robustness of misinformation detection based on social networks [74, 75], these works are not designed for multi-modal based misinformation detection, and how to perform suitable data augmentation for multi-modal data is still an open question in the research community. We thus do not consider data augmentation in the baseline and will leave it in our future work.

Finally, domain adaptation baselines comprise **DAN** [44], **DANN** [45], **Coral** [43] and $\mathbf{M^3DA}$ [52]. **DAN** and **DANN** reduce domain discrepancy between the source and target domains by minimizing MMD metric and adversarial learning correspondingly. **Coral** aligns the second-order statistics of the source and target distributions using a nonlinear transformation. $\mathbf{M^3DA}$ employs moment matching to align each pair of source domains and each source domain with the target domain. Moreover, it further aligns the conditional probability distribution of output given input. **DAN**, **DANN**, **Coral** are single-source DA (SDA) methods, while the other belongs to multi-source DA (MDA) methods.

2) Implementation Details:

1) **Model Setting**. We adopt TextCNN and ResNet50 as the backbone framework to extract text and image features and map the features into d dimensions, using corresponding two-layer MLPs, for all models except **Bert**. Moreover, d is set to 256. TextCNN has three 1D convolutional layers with kernel sizes 3, 4, and 5, and the filter size of each layer is 100. While we finetune ResNet50 for the baseline **ResNet**, we freeze the weights of this visual encoder for the other models. We initialize TextCNN word embedding in the same way as **TextCNN-roberta**. As existing domain generalization and domain adaptation methods are devised for only one input modality, we apply these algorithms to the combined features. We concatenate text and image features and then use an external MLP to map them to d dimension.
2) **Domain Setting**. We select three events as source domains and the remaining one as the target domain. We combine three source domains as a source domain for SDA baselines (i.e., **DAN**, **DANN**, and **Coral**) while keeping these source domains individual for MDA approaches (i.e., $\mathbf{M^3DA}$ and our proposed **RDCM**).
3) **Training Setting**. The sample size of each domain is set to 32 for each minibatch. For data preprocessing, we first resize the image to 224×224 and then normalize pixel values to have a mean of [0.485, 0.456, 0.406] and a standard deviation of [0.229, 0.224, 0.225] to ensure compatibility with our visual backbone, ResNet50 [70]. For hyperparameters, we fix the sigma of Gaussian kernels as [2, 4, 8, 16] for both modalities (We adopt multi-kernel MMD in our experiments). If not otherwise stated, we set the threshold β in Eq. 9 to 0.5 and the temperature τ in Eq. 10 to 0.5. Moreover, we only finetune the weights of different losses λ_1 and λ_2 for our model by searching from [0.005, 0.1, 0.5, 1, 5, 10]. At last, We find that $\lambda_1 = 0.1$ and $\lambda_2 = 0.5$ achieve the best performance on the Pheme dataset, while $\lambda_1 = 1$ and $\lambda_2 = 1$ are optimal for the Twitter dataset and Cross dataset. We mainly finetune the loss weights for baselines by searching from [0.01, 0.1, 1, 10, 100, 1000] to find the best hyperparameter. We adopt Adam as the optimizer with a learning rate 0.001 and weight decay of 0.0005. All models are trained for 20 epochs on the Pheme and Cross datasets and 30 epochs on the Twitter dataset.
4) **Evaluation Protocol**. We utilize accuracy as the evaluation metric. In our work, we follow existing work in the community of domain generalization and domain adaptation [35, 76] and use the standard evaluation protocol. Especially, for each dataset, we divide each domain into a training set (70%) and a test set (30%) via random selection from the overall dataset and conduct a leave-one-domain-out evaluation. In domain generalization, we use the training split of source domains to train and select the optimal model based on the validation results of the testing split of source domains, while we employ the training split of source samples and the unlabelled target domain examples to train and also validate the model on the testing split in domain adaptation. For testing, we evaluate the model on the entire target domain for DG and DA. To avoid randomness, all experiments are repeated three times with different random seeds, and the average result and standard deviation are reported.

C. RQ1: Effectiveness of data collected from unlabeled target domain and multiple modalities

There are two motivations for our work. First, existing robust domain misinformation detection methods do not consider the dynamic propagation trend of online information. In other words, it is necessary to cover DG and DA for our method based on the availability of the target domain. Accordingly, an indispensable premise is that the target domain data could

TABLE III: Pheme dataset results of four groups of approaches.

Model		$\mathcal{COF} \rightarrow \mathcal{S}$ (%)	$\mathcal{CSF} \rightarrow \mathcal{O}$(%)	$\mathcal{CSO} \rightarrow \mathcal{F}$ (%)	$\mathcal{OFS} \rightarrow \mathcal{C}$(%)	Avg(%)
Uni-modality	TextCNN-rand	56.41±1.9	52.43±4.2	86.74 ±2.4	79.33±1.5	68.72±2.2
	TextCNN-roberta	**62.38**±1.4	**64.24**±1.0	**87.95**±0.2	**81.95**±0.3	**74.13**±0.3
	Bert [69]	60.53±2.4	57.29±1.0	79.72±2.6	78.72±0.4	69.07±1.5
	ResNet [70]	56.22±0.7	47.18±2.0	86.45±2.9	70.90±3.2	65.19±1.7
Multi-modality	Vanilla [71]	**65.79**±1.7	**64.67**±2.2	87.45±0.2	**81.02**±0.5	**74.73**±0.8
	ModalityGat [72]	56.09±2.3	47.48±5.0	**88.03**±0.0	80.32±0.2	67.98±1.3
Domain Generalization	EANN [18]	65.97±1.3	65.62±2.9	88.07±0.5	80.42±0.0	75.02±0.9
	IRM [73]	65.02±0.6	64.71±1.7	87.50±1.0	81.23±0.2	74.64±0.2
	MLDG [40]	64.41±2.4	64.84±0.4	88.35±0.2	81.56±0.1	74.79±0.5
	Fish [42]	55.87±2.1	43.58±0.5	88.03±0.0	75.88±4.7	65.84±0.6
	RDCM(DG)	**67.36**±1.8	**66.49**±2.7	**88.41**±0.6	**81.89**±0.0	**76.04**±0.9
Domain Adaptation	DAN [44]	67.09±0.3	62.46±1.4	86.04±2.4	80.56±0.2	74.04±0.9
	DANN [45]	69.24±1.2	64.67±2.8	87.66±0.6	81.29±0.2	75.72±1.1
	Coral [43]	**69.66**±0.7	64.19±2.4	85.60±2.5	80.70±0.0	75.04±1.0
	M^3DA [52]	66.75±1.7	66.63±0.6	88.20±0.4	81.06±0.3	75.66±0.6
	RDCM(DA)	67.49±1.7	**68.75**±1.0	**88.48**±0.0	**82.16**±0.3	**76.72**±0.3

TABLE IV: Twitter dataset results of four groups of approaches.

Model		$\mathcal{ABI} \rightarrow \mathcal{M}$ (%)	$\mathcal{BMI} \rightarrow \mathcal{A}$(%)	$\mathcal{AMI} \rightarrow \mathcal{B}$ (%)	$\mathcal{ABM} \rightarrow \mathcal{I}$(%)	Avg(%)
Uni-modality	TextCNN-rand	45.95±3.0	53.12±1.2	59.82±4.6	46.80±2.5	51.42±0.9
	TextCNN-roberta	46.31±0.7	56.12±0.4	69.76±1.1	40.01±1.0	53.05±0.5
	Bert [69]	58.44±2.9	54.55±0.7	75.27±2.1	**55.51**±3.6	60.94±1.3
	ResNet [70]	**76.89**±4.6	**54.73**±3.1	**83.40**±0.3	36.71±2.6	**62.93**±2.1
Multi-modality	Vanilla [71]	81.44±1.0	**61.11**±4.8	79.31±1.5	**40.12**±2.3	**65.50**±1.1
	ModalityGat [72]	**86.32**±0.3	59.55±0.3	**80.62**±0.2	34.61±3.1	65.28±0.6
Domain Generalization	EANN [18]	88.42±3.5	56.61±0.2	71.57±4.3	57.25±2.9	68.46±1.9
	IRM [73]	71.88±2.7	53.13±0.2	80.24±0.3	**58.36**±0.4	65.90±1.0
	MLDG [40]	86.25±6.5	56.23±0.7	78.94±0.2	51.20±8.7	68.16±3.6
	Fish [42]	71.86±5.3	55.61±0.0	79.56±0.5	45.11±6.6	63.03±3.8
	RDCM(DG)	**88.49**±0.7	**58.15**±1.9	**81.32**±1.8	52.48±2.3	**70.11**±0.6
Domain Adaptation	DAN [44]	89.37±1.0	58.29±0.7	77.80±1.6	44.21±4.7	67.42±2.5
	DANN [45]	89.49±1.0	60.01±0.2	78.27±1.8	49.62±3.3	69.35±2.1
	Coral [43]	89.91±0.3	60.38±1.7	78.41±1.5	47.52±5.8	69.05±2.8
	M^3DA [52]	89.99±3.2	55.94±0.7	79.35±0.8	**55.53**±2.0	70.20±1.3
	RDCM(DA)	**90.11**±0.6	**60.78**±1.4	**79.47**±1.9	55.50±3.1	**71.47**±0.7

TABLE V: Cross-dataset results of four groups of approaches.

Model		$\mathcal{COF} \rightarrow \mathcal{M}$ (%)	$\mathcal{CSF} \rightarrow \mathcal{A}$(%)	$\mathcal{ABI} \rightarrow \mathcal{S}$ (%)	$\mathcal{ABI} \rightarrow \mathcal{O}$(%)	Avg(%)
Uni-modality	TextCNN-roberta	49.74±0.4	56.11±0.1	53.84±1.3	**52.28**±1.2	52.99±1.6
	ResNet [70]	**53.67**±0.4	**58.32**±0.9	**58.02**±1.1	49.87±0.3	**54.97**±0.3
Multi-modality	Vanilla [71]	**48.66**±2.8	**57.28**±0.3	**59.40**±1.0	48.52±1.5	**53.47**±1.3
	ModalityGat [72]	38.46±0.5	55.86±0.2	56.14±0.4	**52.08**±1.8	50.64±0.6
Domain Generalization	EANN [18]	52.23±4.4	57.01±0.2	58.34±1.6	52.98±2.8	55.14±1.9
	IRM [73]	52.93±3.7	56.11±0.6	57.16±0.9	53.03±0.0	54.81±1.9
	MLDG [40]	53.30±0.6	55.28±0.1	56.64±1.0	52.82±0.6	54.51±0.5
	Fish [42]	47.78±1.5	51.23±4.6	53.49±2.3	48.00±3.4	50.12±2.6
	RDCM(DG)	**53.41**±1.7	**57.40**±0.2	**59.90**±2.0	**53.17**±0.5	**55.97**±1.2
Domain Adaptation	DAN [44]	53.29±0.8	57.26±0.2	59.12±2.0	51.57±1.8	55.31±0.5
	DANN [45]	54.66±6.7	57.03±0.8	55.10±0.5	51.08±1.3	54.47±2.2
	Coral [43]	54.20±3.1	58.01±1.0	56.36±1.5	51.48±2.4	55.01±1.2
	M^3DA [52]	53.61±1.8	58.36±0.3	58.84±1.0	51.34±1.2	55.54±1.0
	RDCM(DA)	**55.27**±2.6	**58.49**±0.5	**60.33**±0.6	**52.00**±2.5	**56.52**±1.0

TABLE VI: $\mathcal{A}$-distance of four cases for Pheme and Twitter datasets in DG and DA settings.

Pheme Dataset					
Model	Metric	$\mathcal{S}$	$\mathcal{O}$	$\mathcal{F}$	$\mathcal{C}$
RDCM(DG)	Acc(%)	67.36	66.49	88.41	81.89
	$\mathcal{A}$-distance	1.79	1.78	1.73	1.76
RDCM(DA)	Acc(%)	67.49	68.75	88.48	82.16
	$\mathcal{A}$-distance	1.75	1.76	1.64	1.73
Twitter Dataset					
Model	Metric	$\mathcal{M}$	$\mathcal{A}$	$\mathcal{B}$	$\mathcal{I}$
RDCM(DG)	Acc(%)	88.49	58.15	81.32	52.48
	$\mathcal{A}$-distance	1.69	1.62	1.64	1.90
RDCM(DA)	Acc(%)	90.11	60.78	79.47	55.50
	$\mathcal{A}$-distance	1.64	1.68	1.61	1.89

further boost the detection performance compared with DG. On the other hand, fewer recent approaches concentrate on the importance of the semantic gap between textual and visual modalities. However, a foundation of this motivation is that multi-modal methods could have advantages over uni-modal ones. As a result, we conduct comprehensive experiments and report the accuracy and standard error in Table III and Table IV, aiming to prove the validity of both motivations.

1) Importance of the Target Domain: We show the impact of unlabeled target domain data for improving the performance of misinformation detection. Some theoretical analyses [46, 50, 51] bound the target error in terms of the source error,

the divergence between the distributions of the source domain and the target domain, and other components. In other words, when reducing the discrepancy among source domains, we could improve the classification accuracy in the target domain by concurrently reducing the discrepancy between the target domain and source domains. In turn, we conduct two-sided Wilcoxon rank-sum statistic[11] for the average accuracy of DG and DA baselines on two datasets. The p-values of our tests (0.25 for the Pheme dataset and 0.12 for the Twitter Dataset) are more than 0.05.

2) Effectiveness of Multi-modal Methods: We illustrate the superiority of exploiting both modalities by analyzing the experimental results of unimodal and multi-modal methods. On the Pheme dataset, **Vanilla**, combining textual and visual features surpasses **TextCNN-roberta** with 0.60% improvement and **Resnet** with 9.54%. When on the Twitter dataset, this multi-modal method also brings 2.57% improvement compared with **ResNet**.

ResNet shows the opposite trend. It is possibly due to differences between two datasets, such as data collection ways and label protocols, which is a common case for practical applications. Especially, advisable multi-modal models could have the potential to combine complementary information from multiple modalities by filtering noise and resolving conflicts based on comprehending correlations between these modalities, which justifies the advantage of exploiting both texts and images for our task. We adopt **Vanilla** as the backbone for subsequent experiments.

Answer to RQ1: Target domain and multi-modal inputs effectively aid robust domain misinformation detection.

D. RQ2: Effectiveness of Our Method

Given the news propagation dynamics, it would be beneficial for robust domain approaches to cover domain adaptation and domain generalization simultaneously. To verify the effectiveness and versatility of our method for both settings, we compare **RDCM** with **Vanilla**, DG baselines, and DA baselines. Table III, Table IV and Table V show such results.

We first discuss the comparisons with **Vanilla**. On the Pheme dataset, the DG and DA versions of **RDCM** outperform **Vanilla** by 1%. And the superiority is more significant for the Twitter and Cross datasets. It evinces that inter-domain alignment and cross-modality alignment modules positively influence discriminating the misinformation.

Regarding DG baselines, **RDCM** consistently outperforms most of them by a clear margin and simultaneously achieves over 1% improvement compared with SOTA **EANN** on Pheme and Twitter datasets. Similarly, our proposed method also outperforms all DA baselines on three datasets. We suggest two possible reasons. First, we employ the kernel mean embedding to represent the joint distribution of textual and visual variables to perform domain alignment, which can capture the correlation between variables [25, 35] to reduce incorrect classification. Second, we further mitigate the semantic gap between text and image modalities based on contrastive learning to enable cross-modal misinformation detection compared to other baselines. We also observe that multi-source DA methods (e.g., $\mathbf{M^3DA}$ and **RDCM**) perform better than single-source DA methods (e.g., **DAN**, **DANN**, and **Coral**). Hence, we devise our inter-domain alignment component in the multi-source DA version.

Additionally, it is worth noting that the performance of the proposed method and the baselines significantly differ among the four target domains. For instance, we observe that our model performs better on cases $\mathcal{CSO} \rightarrow \mathcal{F}$ and $\mathcal{OFS} \rightarrow \mathcal{C}$ than it does on $\mathcal{COF} \rightarrow \mathcal{S}$ and $\mathcal{CSF} \rightarrow \mathcal{O}$. We suggest two possible causes for this phenomenon: 1) The domain gap between source and target domains for cases with poor generalization performance may be larger than those cases where models can generalize well. That is because the generalization performance largely depends on domain discrepancies between source domains and the target domain [46, 77]. To validate this conjecture, we exploit $\mathcal{A}$-distance[12], presented by Ben-David et al. [46], to measure domain discrepancies for different cases of Pheme and Twitter datasets in Tabel VI. The results show that models can learn more domain-invariant features in component cases than problematic ones to prove our conjectures. 2) In bad cases, the target domains may be more challenging and complex. For instance, the tweets labeled as rumors in $\mathcal{S}$ and $\mathcal{O}$ have more diversified styles and patterns [78]. As a result, it is difficult for models trained on source domains to learn beneficial invariance capable of covering the distribution of the intractable target domain.

Answer to RQ2: The proposed methods generally outperform different backbone networks, as well as all DG and DA baseline models based on two different settings, which evinces the effectiveness of our proposed **RDCM**.

E. RQ3: Analysis of Different Components

In this subsection, we conduct an ablation study to understand the impact of Inter-domain and Cross-modality Alignment modules of our proposed method. For brevity, we only report detection accuracy in DG.

We consider three variants, including removing the inter-domain alignment component (denoted as w/o inter), removing the cross-modality alignment component (denoted as w/o cross), and removing both components (denoted as w/o both). The results in Table VII are telling. Despite the performance drop in certain cases (e.g., $\mathcal{M}$ and $\mathcal{A}$ in the Twitter dataset) compared to other baselines, our model generally performs best when leveraging all these components. It suggests that our model benefits from both alignment modules. Moreover, **Ours** may overfit to cross-modality alignment loss for $\mathcal{M}$ and overfit to both inter-domain alignment and cross-modality alignment

[11] https://data.library.virginia.edu/the-wilcoxon-rank-sum-test/. The Wilcoxon Rank Sum Test is the non-parametric version of the two-sample t-test, which works when our samples are small.

[12] $\mathcal{A}$-distance is defined as $\hat{d}_{\mathcal{A}} = 2(1-\epsilon)$ where ϵ is the generalization error of a two-sample classifier (kernel SVM in our case, following [44]) trained on the binary problem to distinguish input samples between the source and target domains.

TABLE VII: Experimental results of ablation study in domain generalization.

Pheme Dataset					
	$\mathcal{S}$(%)	$\mathcal{O}$(%)	$\mathcal{F}$(%)	$\mathcal{C}$(%)	Avg(%)
Ours(DG)	**67.36**	**66.49**	**88.41**	81.89	**76.04**
w/o inter	66.61	66.17	88.30	81.47	75.64
w/o cross	67.33	65.41	88.19	**81.98**	75.73
w/o both	65.78	64.67	87.45	81.02	74.73
Twitter Dataset					
	$\mathcal{M}$(%)	$\mathcal{A}$(%)	$\mathcal{B}$(%)	$\mathcal{I}$(%)	Avg(%)
Ours(DG)	88.49	58.15	**81.32**	**52.48**	**70.11**
w/o inter	82.63	56.00	75.23	51.48	66.34
w/o cross	**91.08**	57.81	80.85	45.10	68.71
w/o both	81.44	**61.11**	79.31	40.12	65.50

(a) $\mathcal{COF} \rightarrow \mathcal{S}$ (b) $\mathcal{BMI} \rightarrow \mathcal{A}$

Fig. 3: The examples of Sydney Siege from $\mathcal{COF} \rightarrow \mathcal{S}$ in the Pheme Dataset and Hurricane Sandy from $\mathcal{BMI} \rightarrow \mathcal{A}$ in the Twitter Dataset. These samples are wrongly classified by **Vanilla** but can be identified correctly by our proposed **RDCM** (DG). Sydney Siege was a terrorist attack that a gunman held hostage ten customers and eight employees in Sydney on December 15-16, 2014. Moreover, Hurricane Sandy was extremely destructive and strong, affecting 24 states in the United States.

losses for $\mathcal{A}$, which can be mitigated by adjusting weights of different loss (λ_1 and λ_2). In turn, removing inter-domain alignment leads to a greater performance drop than cross-domain alignment, especially in the Twitter dataset. However, it is difficult to determine which is more important because of the comparable performance on the Pheme dataset.

Answer to RQ3: Each component of **RDCM** contributes positively to multi-modal misinformation detection task. Both components are important and could be assisted by each other.

F. Case Study

To further justify the effectiveness of our proposed model RDCM (DG), we provide case studies on samples that are misclassified by **Vanilla** [71] but are detected accurately by our proposed model, which incorporates domain alignment and cross-modal alignment modules.

As depicted in Fig. 3, **RDCM** excels at understanding semantic correspondences and contradictions between texts and images and learns more transferable implicit patterns for multimodal misinformation detection compared to **Vanilla**. For instance, identifying non-rumor and real samples may imply that our model can comprehend that "Australian PM Tony Abbott" and "turn off the lights" align with the person and the dark background in the attached images, respectively. Additionally, identifying the rumor sample in Fig. 3a depends on spotting the cross-modal irrelevance. We suggest the success of these samples may stem from our cross-modal alignment component that effectively reduces the modality gap through contrastive learning. Moreover, our model may learn contributive domain-invariant features better, such as the races of artificial synthesis as shown in the fake sample in Fig. 3b, owing to the inter-domain alignment module aligning the joint distribution of both modalities conditioned on their correlation information.

V. Conclusions & Future Work

In this paper, we tackled the problem of robust domain misinformation detection. We presented a robust domain and modality-alignment framework based on inter-domain and cross-modality alignment modules.

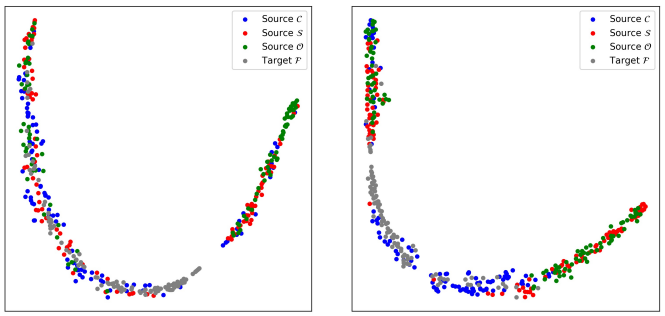
The kernel mean embedding underpins inter-domain alignment to represent the joint distribution of textual and visual modalities. It reduces the domain shift by minimizing the Maximum Mean Discrepancy between the joint distributions.

The cross-modality alignment module leverages a specific sample strategy to construct positive and negative samples and mitigate the modality gap based on contrastive learning. Experimental results show the effectiveness of the proposed method for robust domain misinformation detection.

For future work, extending the framework to handle multiple images and long-paragraph texts represents a key step forward. We also suggest exploring various multi-modality scenarios containing video and audio information to enrich the current text- and image-based representations.

VI. Limitations

While the proposed approach (**RDCM**) demonstrates versatility and effectiveness for the multimodal misinformation detection task in both domain generalization and domain adaptation scenarios, it is important to acknowledge two possible limitations. Firstly, **RDCM** employs Maximum Mean Discrepancy (MMD) as a metric to measure the domain discrepancy upon the joint distribution of textual and visual modalities. Although MMD offers theoretical merits, it does have certain deficiencies such as the sensitivity to kernel choices and computationally expensive calculations for large high-dimensional datasets (i.e., the computational complexity is O(n^2) where n represents the sample size) [26, 46]. Despite these drawbacks, our proposed method outperforms existing approaches in two publicly available datasets when the sigma of Gaussian kernels is fixed for both modalities and each domain contains a limited number of samples, because of the synergy



(a) **Joint** (b) **Text**

Fig. 4: t-SNE visualization of combined features belonging to three source domains ($\mathcal{C}$, $\mathcal{S}$ and $\mathcal{O}$) and one target domain ($\mathcal{F}$) for Pheme dataset. The features of domains $\mathcal{C}$ and $\mathcal{F}$ are mainly distributed in the two clusters on the left bottom of Fig. 4b while the features of these two domains scatter more evenly in Fig. 4a.

TABLE VIII: Comparison results of inter-domain alignment on different modalities in domain generalization.

Pheme Dataset					
Model	$\mathcal{S}$(%)	$\mathcal{O}$(%)	$\mathcal{F}$(%)	$\mathcal{C}$(%)	Avg(%)
Fusion	66.72	65.28	88.07	82.10	75.54
Vision	66.64	65.36	88.00	**82.18**	75.55
Text	64.89	64.84	88.00	81.94	74.92
Joint	**67.33**	**65.41**	**88.19**	81.98	**75.73**
Twitter Dataset					
Model	$\mathcal{M}$(%)	$\mathcal{A}$(%)	$\mathcal{B}$(%)	$\mathcal{I}$(%)	Avg(%)
Fusion	89.35	56.62	77.23	45.03	67.06
Vision	78.78	56.19	73.18	**49.85**	64.50
Text	85.98	**60.07**	80.41	49.15	**68.90**
Joint	**91.08**	57.81	**81.25**	44.70	68.71

of inter-domain alignment and intra-domain alignment modules. Secondly, our method specifically focuses on debunking fake image-text pairs. Nevertheless, the intricate nature of multimodal inputs permitted by social media platforms, such as short videos and emojis, further harms the deployment of our method in the real world. Therefore, we intend to address these two limitations in our future endeavors.

VII. Appendix

A. RQ4: Analysis of Inter-domain Alignment

In Inter-domain Alignment, we assume the domain shift exists in the joint distribution of multiple modalities instead of the marginal distribution of any individual modality. Furthermore, unlike simple fusion (e.g., concatenation), we employ the kernel mean embedding to represent the joint distribution. To show the superiority of this module, we conduct experiments on four models. The first model, **Fusion**, involves aligning the joint distribution of both modalities obtained by concatenation, described as $\text{MMD}(\mathcal{D}_S^i, \mathcal{D}_S^j) = \|\mu_{\mathbf{X}_{t,v,i}} - \mu_{\mathbf{X}_{t,v,j}}\|_{\mathcal{H}}^2$ where $\mathbf{X}_{t,v}$ represents the random variable of the concatenation of textual and visual features. The second, **Vision**, aligns the marginal distribution upon visual features, described as $\text{MMD}(\mathcal{D}_S^i, \mathcal{D}_S^j) = \|\mu_{\mathbf{X}_{v,i}} - \mu_{\mathbf{X}_{v,j}}\|_{\mathcal{H}}^2$. The third one, **Text**, aligns marginal distribution upon textual features, described as $\text{MMD}(\mathcal{D}_S^i, \mathcal{D}_S^j) = \|\mu_{\mathbf{X}_{t,i}} - \mu_{\mathbf{X}_{t,j}}\|_{\mathcal{H}}^2$. Finally, the fourth one, **Joint**, aligns the joint distribution of both modalities obtained by our proposed kernel mean embedding in Eq. 4, described as $\text{MMD}(\mathcal{D}_S^i, \mathcal{D}_S^j) = \|\mu_{\mathbf{X}_{t,i},\mathbf{X}_{v,i}} - \mu_{\mathbf{X}_{t,j},\mathbf{X}_{v,j}}\|_{\mathcal{H}}^2$.

TABLE IX: Comparison results of different contrastive learning methods in domain generalization.

Pheme Dataset					
Model	$\mathcal{S}$(%)	$\mathcal{O}$(%)	$\mathcal{F}$(%)	$\mathcal{C}$(%)	Avg(%)
Regular [64]	58.74	45.23	88.03	80.37	68.09
TextCon	65.66	65.36	88.06	81.47	75.21
ThresCon	64.76	65.41	88.00	80.63	74.70
Ours	**66.61**	**66.17**	**88.30**	**81.47**	**75.64**
Twitter Dataset					
Model	$\mathcal{M}$(%)	$\mathcal{A}$(%)	$\mathcal{B}$(%)	$\mathcal{I}$(%)	Avg(%)
Regular [64]	56.78	**60.42**	70.74	44.53	58.12
TextCon	77.18	56.21	73.48	50.90	64.44
ThresCon	73.76	56.03	70.85	49.57	62.55
Ours	**82.63**	56.00	**75.23**	**51.48**	**66.34**

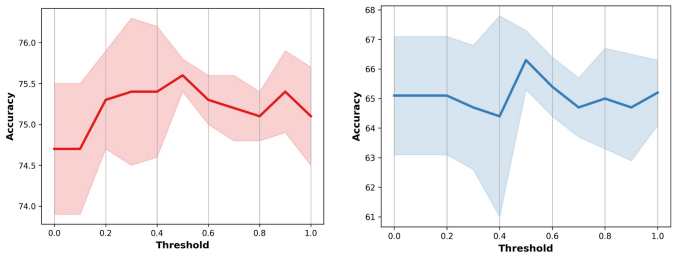
From Table VIII, we observe that **Joint** and **Fusion** usually have higher accuracy than **Text** and **Image**, which illustrates the effectiveness of aligning the joint distribution. It may be because deciding which modality mainly accommodates the domain shift is impractical. We further visualize the combined features of different domains extracted by **Joint** and **Text** using t-SNE embeddings in Fig. 4a and Fig. 4b, respectively. The figures show that the features are less discriminative when generated by **Joint**, especially for features of the target domain. It also suggests that the adaptation of joint distributions is more powerful than marginal distributions for our task. Besides, the boost of **Joint** is more significant than **Fusion**. Such empirical results and theoretical guarantees in Eq. 4 imply that the kernel mean embedding is more effective in modeling the joint distribution for our task.

Answer to RQ4: Aligning the joint distribution of textual and visual modalities achieves better performance than aligning their marginal distributions. Moreover, the mean kernel embedding is more advantageous for modeling the joint distribution compared with fusion through feature concatenation .

B. RQ5: Analysis of Cross-modality Alignment

In Cross-modality Alignment, we exclude positive and negative samples of low quality by only taking real posts as positive samples and the negative samples selected by our weighting function in Eq. 9 based on image similarity, respectively. To show the usefulness of this strategy (denoted as **Ours**), we compare it with three other kinds of contrastive learning methods. The first one, **Regular**, uses a common contrastive loss [64] based on random sampling. The second, **TextCon**, includes the weighting function but employs text modality-based similar scores instead. Finally, **ThresCon** removes the weighting function term and only considers real posts as positive samples.

As Table IX shows, **Regular** is dominated by the other three methods by a large margin, highlighting the importance of filtering out non-relevant samples. Moreover, our method outperforms **TextCon** and **ThresCon**, which demonstrates the



(a) Pheme Dataset (b) Twitter Dataset

Fig. 5: Performance of our cross-modality alignment module with different thresholds in domain generalization.

effectiveness of our proposed indicator function term in Eq. 9 that excludes low-quality artificial negative samples based on semantic similarity on the visual modality. In addition, we conduct experiments with different thresholds (i.e., β in Eq. 9) as Fig. 5 depicts. The increase in the threshold brings more noise. This figure shows that the performance first increases and then drops along the threshold increase. Thus, we advocate a tradeoff between sample number and sample noise.

Answer to RQ5: Our model benefits from the proposed sample strategy that can filter non-relevant samples.

REFERENCES

[1] X. Zhou and R. Zafarani, "A survey of fake news: Fundamental theories, detection methods, and opportunities," ACM Comput. Surv., vol. 53, no. 5, pp. 109:1–109:40, 2020.

[2] K. M. Caramancion, "The role of information organization and knowledge structuring in combatting misinformation: A literary analysis," in CSoNet, ser. Lecture Notes in Computer Science, vol. 13116. Springer, 2021, pp. 319–329.

[3] S. Wineburg and S. McGrew, "Evaluating information: The cornerstone of civic online reasoning," 2016.

[4] M. Hindman and V. Barash, "Disinformation, and influence campaigns on twitter," Knight Foundation: George Washington University, 2018.

[5] A. Willmore, "This analysis shows how viral fake election news stories outperformed real news on facebook," 2016.

[6] L. Hu, T. Yang, L. Zhang, W. Zhong, D. Tang, C. Shi, N. Duan, and M. Zhou, "Compare to the knowledge: Graph neural fake news detection with external knowledge," in ACL/IJCNLP (1). Association for Computational Linguistics, 2021, pp. 754–763.

[7] X. Zhou, A. Jain, V. V. Phoha, and R. Zafarani, "Fake news early detection: A theory-driven model," CoRR, vol. abs/1904.11679, 2019.

[8] Y. Wang, W. Yang, F. Ma, J. Xu, B. Zhong, Q. Deng, and J. Gao, "Weak supervision for fake news detection via reinforcement learning," in AAAI. AAAI Press, 2020, pp. 516–523.

[9] X. Yang, Y. Lyu, T. Tian, Y. Liu, Y. Liu, and X. Zhang, "Rumor detection on social media with graph structured adversarial learning," in IJCAI. ijcai.org, 2020, pp. 1417–1423.

[10] N. Ruchansky, S. Seo, and Y. Liu, "CSI: A hybrid deep model for fake news detection," in CIKM. ACM, 2017, pp. 797–806.

[11] L. Cheng, R. Guo, K. Shu, and H. Liu, "Causal understanding of fake news dissemination on social media," in KDD. ACM, 2021, pp. 148–157.

[12] Z. Jin, J. Cao, H. Guo, Y. Zhang, and J. Luo, "Multimodal fusion with recurrent neural networks for rumor detection on microblogs," in ACM Multimedia. ACM, 2017, pp. 795–816.

[13] Y. Chen, D. Li, P. Zhang, J. Sui, Q. Lv, L. Tun, and L. Shang, "Cross-modal ambiguity learning for multimodal fake news detection," in WWW. ACM, 2022, pp. 2897–2905.

[14] R. Tan, B. A. Plummer, and K. Saenko, "Detecting cross-modal inconsistency to defend against neural fake news," in EMNLP (1). Association for Computational Linguistics, 2020, pp. 2081–2106.

[15] P. Li, X. Sun, H. Yu, Y. Tian, F. Yao, and G. Xu, "Entity-oriented multi-modal alignment and fusion network for fake news detection," IEEE Trans. Multim., vol. 24, pp. 3455–3468, 2022.

[16] Y. Wang, F. Ma, H. Wang, K. Jha, and J. Gao, "Multimodal emergent fake news detection via meta neural process networks," in KDD. ACM, 2021, pp. 3708–3716.

[17] Y. Zhu, Q. Sheng, J. Cao, Q. Nan, K. Shu, M. Wu, J. Wang, and F. Zhuang, "Memory-guided multi-view multi-domain fake news detection," CoRR, vol. abs/2206.12808, 2022.

[18] Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, L. Su, and J. Gao, "EANN: event adversarial neural networks for multi-modal fake news detection," in KDD. ACM, 2018, pp. 849–857.

[19] H. Zhang, S. Qian, Q. Fang, and C. Xu, "Multimodal disentangled domain adaption for social media event rumor detection," IEEE Trans. Multim., vol. 23, pp. 4441–4454, 2021.

[20] A. Silva, L. Luo, S. Karunasekera, and C. Leckie, "Embracing domain differences in fake news: Cross-domain fake news detection using multi-modal data," in AAAI. AAAI Press, 2021, pp. 557–565.

[21] Y. Papanastasiou, "Fake news propagation and detection: A sequential model," Manag. Sci., vol. 66, no. 5, pp. 1826–1846, 2020.

[22] S. van der Linden, "Misinformation: susceptibility, spread, and interventions to immunize the public," Nature Medicine, vol. 28, no. 3, pp. 460–467, 2022.

[23] Y. Li, K. Lee, N. Kordzadeh, B. D. Faber, C. Fiddes, E. Chen, and K. Shu, "Multi-source domain adaptation with weak supervision for early fake news detection," in IEEE BigData. IEEE, 2021, pp. 668–676.

[24] A. Mosallanezhad, M. Karami, K. Shu, M. V. Mancenido, and H. Liu, "Domain adaptive fake news detection via reinforcement learning," in WWW. ACM, 2022, pp. 3632–3640.

[25] K. Muandet, K. Fukumizu, B. K. Sriperumbudur, and B. Schölkopf, "Kernel mean embedding of distributions: A review and beyond," Found. Trends Mach. Learn., vol. 10, no. 1-2, pp. 1–141, 2017.

[26] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. J. Smola, "A kernel two-sample test," J. Mach. Learn. Res., vol. 13, pp. 723–773, 2012.

[27] K. He, H. Fan, Y. Wu, S. Xie, and R. B. Girshick, "Momentum contrast for unsupervised visual representation learning," in CVPR. Computer Vision Foundation / IEEE, 2020, pp. 9726–9735.

[28] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, "A simple framework for contrastive learning of visual representations," in ICML, ser. Proceedings of Machine Learning Research, vol. 119. PMLR, 2020, pp. 1597–1607.

[29] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in ICML, ser. Proceedings of Machine Learning Research, vol. 139. PMLR, 2021, pp. 8748–8763.

[30] J. Wang, C. Lan, C. Liu, Y. Ouyang, and T. Qin, "Generalizing to unseen domains: A survey on domain generalization," in IJCAI. ijcai.org, 2021, pp. 4627–4635.

[31] J. Huang, D. Guan, A. Xiao, and S. Lu, "FSDR: frequency space domain randomization for domain generalization," in CVPR. Computer Vision Foundation / IEEE, 2021, pp. 6891–6902.

[32] K. Zhou, C. C. Loy, and Z. Liu, "Semi-supervised domain generalization with stochastic stylematch," CoRR, vol. abs/2106.00592, 2021.

[33] Z. Wang, Y. Luo, R. Qiu, Z. Huang, and M. Baktashmotlagh, "Learning to diversify for single domain generalization," in ICCV. IEEE, 2021, pp. 814–823.

[34] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira, "Analysis of representations for domain adaptation," in NIPS. MIT Press, 2006, pp. 137–144.

[35] H. Li, S. J. Pan, S. Wang, and A. C. Kot, "Domain generalization with adversarial feature learning," in CVPR. Computer Vision Foundation / IEEE Computer Society, 2018, pp. 5400–5409.

[36] J. Wang, W. Feng, Y. Chen, H. Yu, M. Huang, and P. S. Yu, "Visual domain adaptation with manifold embedded distribution alignment," in ACM Multimedia. ACM, 2018, pp. 402–410.

[37] X. Jin, C. Lan, W. Zeng, and Z. Chen, "Style normalization and restitution for domain generalization and adaptation," IEEE Trans. Multim., vol. 24, pp. 3636–3651, 2022.

[38] Z. Ding and Y. Fu, "Deep domain generalization with structured low-rank constraint," IEEE Trans. Image Process., vol. 27, no. 1, pp. 304–313, 2018.

[39] C. Liu, L. Wang, K. Li, and Y. Fu, "Domain generalization via feature variation decorrelation," in ACM Multimedia. ACM, 2021, pp. 1683–1691.

[40] D. Li, Y. Yang, Y. Song, and T. M. Hospedales, "Learning to generalize: Meta-learning for domain generalization," in AAAI. AAAI Press, 2018, pp. 3490–3497.

[41] K. Zhou, Y. Yang, Y. Qiao, and T. Xiang, "Domain adaptive ensemble learning," IEEE Trans. Image Process., vol. 30, pp. 8008–8018, 2021.

[42] Y. Shi, J. Seely, P. H. S. Torr, S. Narayanaswamy, A. Y. Hannun, N. Usunier, and G. Synnaeve, "Gradient matching for domain generalization," in ICLR. OpenReview.net, 2022.

[43] B. Sun and K. Saenko, "Deep CORAL: correlation alignment for deep domain adaptation," in ECCV Workshops (3), ser. Lecture Notes in Computer Science, vol. 9915, 2016, pp. 443–450.

[44] M. Long, Y. Cao, J. Wang, and M. I. Jordan, "Learning transferable features with deep adaptation networks," in ICML, ser. JMLR Workshop and Conference Proceedings, vol. 37. JMLR.org, 2015, pp. 97–105.

[45] Y. Ganin and V. S. Lempitsky, "Unsupervised domain adaptation by backpropagation," in ICML, ser. JMLR Workshop and Conference Proceedings, vol. 37. JMLR.org, 2015, pp. 1180–1189.

[46] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," Mach. Learn., vol. 79, no. 1-2, pp. 151–175, 2010.

[47] J. Shen, Y. Qu, W. Zhang, and Y. Yu, "Wasserstein distance guided representation learning for domain adaptation," in AAAI. AAAI Press, 2018, pp. 4058–4065.

[48] C. Lee, T. Batra, M. H. Baig, and D. Ulbricht, "Sliced wasserstein discrepancy for unsupervised domain adaptation," in CVPR. Computer Vision Foundation / IEEE, 2019, pp. 10 285–10 295.

[49] H. Li, W. Li, H. Cao, S. Wang, F. Huang, and A. C. Kot, "Unsupervised domain adaptation for face anti-spoofing," IEEE Trans. Inf. Forensics Secur., vol. 13, no. 7, pp. 1794–1809, 2018.

[50] Y. Mansour, M. Mohri, and A. Rostamizadeh, "Domain adaptation with multiple sources," in NIPS. Curran Associates, Inc., 2008, pp. 1041–1048.

[51] J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. Wortman, "Learning bounds for domain adaptation," in NIPS. Curran Associates, Inc., 2007, pp. 129–136.

[52] X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang, "Moment matching for multi-source domain adaptation," in ICCV. IEEE, 2019, pp. 1406–1415.

[53] Y. Zhu, F. Zhuang, and D. Wang, "Aligning domain-specific distribution and classifier for cross-domain classification from multiple sources," in AAAI. AAAI Press, 2019, pp. 5989–5996.

[54] G. Shan, B. Zhao, J. R. Clavin, H. Zhang, and S. Duan, "Poligraph: Intrusion-tolerant and distributed fake news detection system," IEEE Trans. Inf. Forensics Secur., vol. 17, pp. 28–41, 2022.

[55] S. Abdelnabi, R. Hasan, and M. Fritz, "Open-domain, content-based, multi-modal fact-checking of out-of-context images via online resources," in CVPR. IEEE, 2022, pp. 14 920–14 929.

[56] Y. Wang, S. Qian, J. Hu, Q. Fang, and C. Xu, "Fake news detection via knowledge-driven multimodal graph convolutional networks," in ICMR. ACM, 2020, pp. 540–547.

[57] Y. Kim, "Convolutional neural networks for sentence classification," in EMNLP. ACL, 2014, pp. 1746–1751.

[58] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized BERT pretraining approach," CoRR, vol. abs/1907.11692, 2019.

[59] Q. Nan, D. Wang, Y. Zhu, Q. Sheng, Y. Shi, J. Cao, and J. Li, "Improving fake news detection of influential domain via domain- and instance-level transfer," in COLING. International Committee on Computational Linguistics, 2022, pp. 2834–2848.

[65] A. Zubiaga, M. Liakata, and R. Procter, "Exploiting context for rumour detection in social media," in SocInfo (1), ser. Lecture Notes in Computer Science, vol. 10539. Springer, 2017, pp. 109–123.

[60] C. Yang, F. Zhu, G. Liu, J. Han, and S. Hu, "Multimodal hate speech detection via cross-domain knowledge transfer," in ACM Multimedia. ACM, 2022, pp. 4505–4514.

[61] L. Song and B. Dai, "Robust low rank kernel embeddings of multivariate distributions," in NIPS, 2013, pp. 3228–3236.

[62] T. Srinivasan, X. Ren, and J. Thomason, "Curriculum learning for data-efficient vision-language alignment," CoRR, vol. abs/2207.14525, 2022.

[63] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in CVPR. Computer Vision Foundation / IEEE Computer Society, 2018, pp. 3733–3742.

[64] S. Ging, M. Zolfaghari, H. Pirsiavash, and T. Brox, "COOT: cooperative hierarchical transformer for video-text representation learning," in NeurIPS, 2020.

[66] C. Boididou, K. Andreadou, S. Papadopoulos, D. Dang-Nguyen, G. Boato, M. Riegler, and Y. Kompatsiaris, "Verifying multimedia use at mediaeval 2015," in MediaEval, ser. CEUR Workshop Proceedings, vol. 1436. CEUR-WS.org, 2015.

[67] E. Kochkina, M. Liakata, and A. Zubiaga, "All-in-one: Multi-task learning for rumour verification," in Proceedings of the 27th International Conference on Computational Linguistics, 2018, pp. 3402–3413.

[68] A. Zubiaga, M. Liakata, R. Procter, G. Wong Sak Hoi, and P. Tolmie, "Analysing how people orient to and spread rumours in social media by looking at conversational threads," PloS one, vol. 11, no. 3, p. e0150989, 2016.

[69] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in NAACL-HLT (1). Association for Computational Linguistics, 2019, pp. 4171–4186.

[70] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in CVPR. IEEE Computer Society, 2016, pp. 770–778.

[71] H. Pan, Z. Lin, P. Fu, Y. Qi, and W. Wang, "Modeling intra and inter-modality incongruity for multi-modal sarcasm detection," in EMNLP (Findings), ser. Findings of ACL, vol. EMNLP 2020. Association for Computational Linguistics, 2020, pp. 1383–1392.

[72] Y. Cai, H. Cai, and X. Wan, "Multi-modal sarcasm detection in twitter with hierarchical fusion model," in ACL (1). Association for Computational Linguistics, 2019, pp. 2506–2515.

[73] K. Ahuja, K. Shanmugam, K. R. Varshney, and A. Dhurandhar, "Invariant risk minimization games," in ICML, ser. Proceedings of Machine Learning Research, vol. 119. PMLR, 2020, pp. 145–155.

[74] J. Ma, W. Gao, and K. Wong, "Detect rumors on twitter by promoting information campaigns with generative adversarial learning," in WWW. ACM, 2019, pp. 3049–3055.

[75] G. Luo, T. Darrell, and A. Rohrbach, "Newsclippings: Automatic generation of out-of-context multimodal media," in EMNLP (1). Association for Computational Linguistics, 2021, pp. 6801–6817.

[76] S. Motiian, M. Piccirilli, D. A. Adjeroh, and G. Doretto, "Unified deep supervised domain adaptation and generalization," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 5715–5725.

[77] H. Ye, C. Xie, T. Cai, R. Li, Z. Li, and L. Wang, "Towards a theoretical framework of out-of-distribution generalization," in NeurIPS, 2021, pp. 23 519–23 531.

[78] A. Zubiaga, A. Aker, K. Bontcheva, M. Liakata, and R. Procter, "Detection and resolution of rumours in social media: A survey," ACM Comput. Surv., vol. 51, no. 2, pp. 32:1–32:36, 2018.