# Multimodal Fake News Detection: MFND Dataset and Shallow-Deep Multitask Learning

**Ye Zhu**[1] , **Yunan Wang**[1] and **Zitong Yu**[2,3,4*]
[1]School of Artificial Intelligence, Hebei University of Technology
[2]School of Computing and Information Technology, Great Bay University
[3]Guangdong Provincial Key Laboratory of Intelligent Information Processing & Shenzhen Key Laboratory of Media Security, Shenzhen University
[4]Dongguan Key Laboratory for Intelligence and Information Technology

## Abstract

Multimodal news contains a wealth of information and is easily affected by deepfake modeling attacks. We present a new Multimodal Fake News Detection dataset (MFND) containing 11 manipulated types, designed to detect and localize highly authentic fake news. Furthermore, we propose a Shallow-Deep Multitask Learning (SDML) model for fake news, which fully uses unimodal and mutual modal features to mine the intrinsic semantics of news. Under shallow inference, we propose the momentum distillation-based light punishment contrastive learning for fine-grained uniform spatial image and text semantic alignment, and an adaptive cross-modal fusion module to enhance mutual modal features. Under deep inference, we design a two-branch framework to augment the image and text unimodal features, respectively merging with mutual modalities features. Experiments on both mainstream and our proposed datasets demonstrate the superiority of the model.

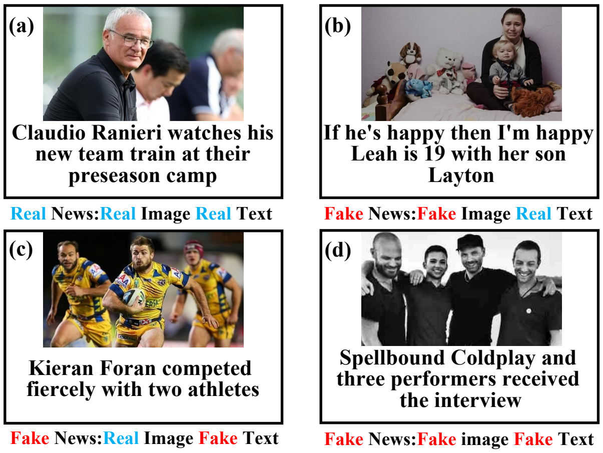
(a)
Claudio Ranieri watches his new team train at their preseason camp
Real News:Real Image Real Text
(b)
If he's happy then I'm happy Leah is 19 with her son Layton
Fake News:Fake Image Real Text
(c)
Kieran Foran competed fiercely with two athletes
Fake News:Real Image Fake Text
(d)
Spellbound Coldplay and three performers received the interview
Fake News:Fake image Fake Text

Figure 1: Illustrates of the news from the MFND dataset. (a) Real News of Real Image Real Text, (b) Fake News of Fake Image Real Text, (c) Fake News of Real Image Fake Text, (d) Fake News of Fake Image Fake Text.

## 1 Introduction

Multimodal news composed of visual and textual components has become the mainstream information dissemination method. However, the malicious abuse of Large Language Models (LLMs) [Kenton and Toutanova, 2019; Radford *et al.*, 2019] and Deep Generative Models [Dolhansky *et al.*, 2020; Zhang *et al.*, 2024b] are challenging media credibility, brought lots of negative impacts to society. Computer Vision (CV) and Natural Language Processing (NLP) fields have respectively proposed many approaches [Xie *et al.*, 2024; Gao *et al.*, 2024], but most of these do not consider news as a whole, or provide inaccurate analysis of semantics.

Multimodal fake news detection research can be outlined into two categories: traditional learning methods [Liu *et al.*, 2015a; Ma *et al.*, 2015] and deep learning methods. Recent studies use deep learning methods [Wang *et al.*, 2024; Luvembe *et al.*, 2024] which are based on neural networks to capture features, but most of them are limited to binary detection. From a multimodal perspective, fake news requires multitasking for detection and localization to mine deep inference information, Fig. 1 illustrates four types of news from the MFND dataset, where the semantic similarity values of true and fake news are alike, and it is difficult to reason effectively only through the shallow binary categorization task with mutual features.

To solve the above problems, we propose a Shallow-Deep Multitask Learning (SDML) model for fake news detection and localization. We propose a light punishment contrastive learning based on momentum distillation for over-penalization of hard negative image-text pair in feature alignment. Moreover, we propose adaptive cross-modal fusion to blend the aligned image and text unimodal features and spontaneously adjust the weight between the two modalities. Finally, we leverage the enhanced unimodal and multimodal features to predict images and text detection and localization results in different branches. To counter the more truthful AI fake news, we establish the MFND dataset which is about media news with humans as the principal part. The proposed MFND dataset contains four multi-modality types and employs post-processing to simulate the real scene, providing media news true-false binary labels, manipulated image la-

[*]Corresponding Author

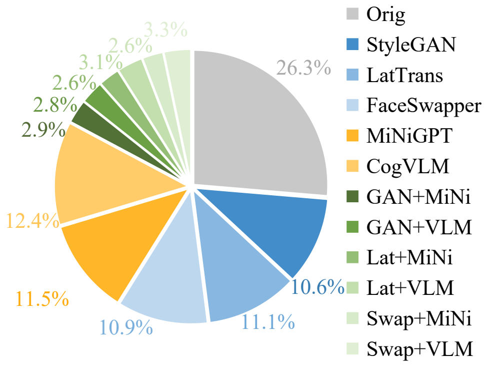

Figure 2: Statistics of our proposed MFND dataset.

bels, manipulated text labels, and manipulated image localization labels.

Overall, the main contributions of this paper are as:

- We contribute the MFND dataset which uses 11 state-of-the-art image and text manipulation methods and provides rich detection and localization labels that fit a wide range of realistic scenarios.
- We propose a Shallow-Deep Multitask Learning (SDML) model for fake news detection and localization, which fuses image and text features after alignment, combining mutual modality with augmented unimodality for fine-grained semantic inference.
- The proposed SDML model achieves state-of-the-art detection and localization performance on four benchmark datasets under both multi-modal multi-task and multi-modal single-task settings.

## 2 Related Work

### 2.1 Deepfake Detection

Current image deepfake researches are mostly based on spatial and frequency domains, such as texture features, noise distribution, blending artifacts, and so on. Zhao [Zhao *et al.*, 2021] applied multi-attention to obtain image local details and aggregate low-level and high-level texture features. Shi [Shi *et al.*, 2023] proposed stacked multi-scale transformers to mine image structural anomalies on blocks of different sizes. Some recent methods incorporate the frequency domain information as data enhancement, Gao [Gao *et al.*, 2023] using the frequency and spatial domains as dual streams to locate results in a hierarchical fusion manner. Li [Li *et al.*, 2023] fused graph convolutional processing features to predict image forgery binary masks guided by dual attention. However, none of them combine with textual modalities whose form have more widespread and harmful.

### 2.2 Multimodal Fake News Detection

Several works have investigated multimodal fake news detection. Qian [Qian *et al.*, 2021] encodes features then hierarchically fuse and sends them to decode getting news results. Zhang [Zhang *et al.*, 2021] introduces cross-contrastive learning and an attention mechanism guides reasoning. Ma [Ma *et al.*, 2024] integrates inconsistencies at the event level and utilizes graph capture plausibility for robust predictions. These works are limited to binary detection, and datasets are miniature while fake news structures are mostly image-text exchanged without deepfake technology. Shao [Shao *et al.*, 2023] defined detection and localization tasks for the first time and built a dataset containing deepfake forgery. However, the types of manipulation are not rich enough. In addition, due to the complexity of the similarity relationship between different texts, the strict text localization labels hinder capturing the accurate image-text depth semantics, and localization annotation also requires extra computational cost. In this paper, we redefine the task of multimodal fake news detection and localization and build a dataset containing more deepfake news with a small labeling cost.

| Datasets | Image | Text | Deepfake | Forgery | Size |
|---|---|---|---|---|---|
| Twitter | ✗ | ✗ | ✗ | ✗ | 13924 |
| Pheme | ✗ | ✗ | ✗ | ✗ | 5790 |
| Weibo | ✗ | ✗ | ✗ | ✗ | 9128 |
| DGM$^4$ | ✓ | ✓ | GAN | 8 | 230000 |
| MFND | ✓ | ✓ | LLM | 11 | 125000 |

Table 1: Comparison with the other four multimodal fake news datasets.

## 3 MFND Dataset

Most established fake multimodal datasets usually adopt coarse-grained annotations for binary detection and collect hand-generated or contextual semantic outer pairs. To facilitate the study of multimodal fake news detection, the large-scale and diverse Multimodal Fake News Detection (MFND) dataset is introduced, which covers keyword and sentiment reversal, summary induction, and keyword substitution manipulation techniques. Specifically, MFND is built on the original VisualNews dataset [Liu *et al.*, 2020], which contains numerous real social news. We select image-text pairs centered on humans as the source pool $O = \{p_o \mid p_o = (I_o, T_0)\}$ through data filtering.

### 3.1 Data Collection

Deepfake generates fake faces based on deep learning methods mainly divided into three categories, i.e., Entire Face Synthesis (EFS) [Karras *et al.*, 2021; Ho *et al.*, 2020], Attribute Manipulation (AM) [Yao *et al.*, 2021; Pernuš *et al.*, 2023], and Face Swap (FS) [Ma *et al.*, 2019; Xu *et al.*, 2022]. EFS uses generative techniques to obtain non-existent fake faces from random noise, AM edits the facial attributes of the original image to produce new forged faces through deep learning models, and FS utilizes neural networks to replace the face of the source image with the face of the target image. We create the MFND dataset through the above forgery techniques.

**Real Image Real Text**

The source pool $O$ after filtering the VisualNews dataset contains 200k valid image-text pairs, we randomly draw some sample pairs from it as the initial part of the MFND dataset,

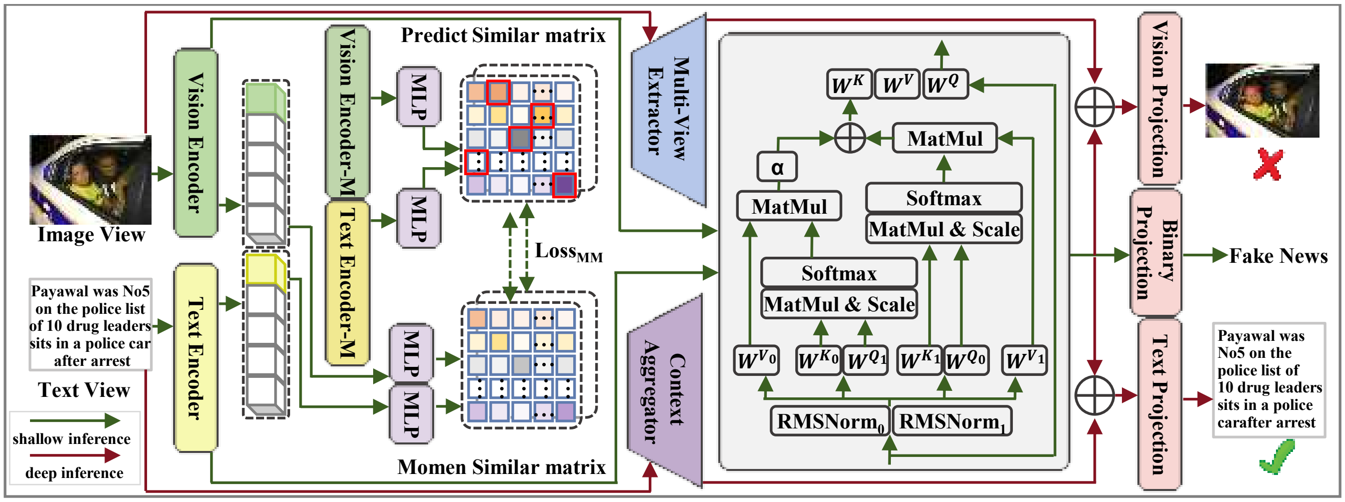

Figure 3: Illustration of the proposed Shallow-Deep Multitask Learning (SDML) method. As for the shallow inference with green lines, we encode two single modalities using different pre-trained Encoders, align the embeddings by contrastive learning, and obtain mutual modality after adaptive fusion for media news binary classification. As for the deep inference with red lines, we enhance features from image and text modalities in a two-branch framework, combined with the fusion feature for detection and localization.

given an example as Fig. 1(a). We record binary labels of the media news, images, and texts respectively, constantly updating the source pool to remove the drawn sample pairs.

### Fake Image Real text

We subject the sampled pairs with an equal probability to EFS, AM, and FS manipulation. Specifically utilizing CelebAHQ [Karras, 2017] pre-trained StyleGAN3 changes the main faces in the image to noise-generated forgery faces, selecting some common categories in forty types of LatTrans to edit the image's facial attribute, setting the pre-trained FaceSwapper by CelebA [Liu *et al.*, 2015b] as the source image to replace stochastic face in the target image. Given an example as Fig. 1(b), and record the forged image and mark the grounding bound box $y_{box} = \{x_1, y_1, x_2, y_2\}$ for detection and localization.

### Real Image Fake text

Text data exists multiple representations for the same semantic, which leads to insufficient compatibility pairs generated by text attribute editing or swap corresponding to traditional emotion reverse theory, and also unsuitable to combine the EFS images with simple textual person name replacement samples. Therefore, we through the state-of-the-art technology MiNiGPT-v2 [Chen *et al.*, 2023] and CogVLM [Wang *et al.*, 2023] in Multimodal Large Language Models (MLLMs) [Tu *et al.*, 2024] fundamental both image and text content to produce forged text with equal probability, the similarity between image-text sample pairs is controlled to be 50% to 75% considering the noise in the real media news. Given an example as Fig. 1(c), and recorded text labels for the detection.

### Fake Image Fake Text

To simulate the correlation between image and text in fake news, we randomly draw some sample pairs from the source pool manipulate the image part according to the above-mentioned means, and then combine the forged image as the proposed real image to generate forged text with MLLMs. Given an example as Fig. 1(d), and record media news real or fake binary labels, forgery or origin image labels, authentic or falsehood text labels, and image manipulation localization labels.

## 3.2 Data Statistics

The overall statistics of the MFND dataset are illustrated in Fig. 2. It contains 125k multimodal fake news samples, including 32869 real image real text news pairs, 40821 fake image real text news pairs, 29826 real image fake text news pairs, and 21484 fake image fake text news pairs. The dataset is divided into three parts, the 95k pairs are part of model training, another 15k pairs are part of testing and the remaining 15k are for testing.

We summarize the information of the major existing multimodal fake news datasets and ours in Table. 1. Earlier datasets such as Twitter [Khattar *et al.*, 2019], Pheme [Zubiaga *et al.*, 2017], Weibo [Jin *et al.*, 2017] without deepfake techniques and fake samples are generated by mismatched image and text pairs and contain only binary labels of the media news, recent DGM[4] [Shao *et al.*, 2023] sets up image and text manipulation methods based on reversed sentiment transformations, containing both image and text localization annotations. MFND updates the technique grounded in real scenes and text semantic representations, utilizes more sophisticated deepfake methods in forged images exchanges sentiment-recognition-based text for generated text, provides rich fine-grained annotations, and removes redundant text localization.

# 4 Method

As shown in Fig. 3, the overall framework of the proposed method Shallow-Deep Multitask Learning (SDML) contains shallow and deep inference stages. As shallow, we encode the image and text modality into a sequence of embeddings by two dedicated uni-modal Encoders, update the alignment results in uniform space under the momentum distillation-based Light Punishment Contrastive Learning module (LPCL), and learn the mutual modal features through Adaptive Cross-Modal Fusion module (ACMF) for predicting the Fake News Binary Detection outputs. Under deep, we design the two-branch framework to augment the uni-modal features for image and text separately, and eventually predict Image Forgery Detection and Localization and Text Forgery Detection with dedicated projections.

## 4.1 Shallow Inference

Given an input image-text pair $(V, T) \in D$, where $V$, $T$, and $D$ represent image, text, and dataset, we obtain the encoded image and text single modality features $V^e$ and $T^e$ by two pre-trained Encoders. The vision encoder chooses the 12-layer ViT-B/16 which has been initialized with ImageNet and adds a fully connected layer to dimension transform the encoded results. The text encoder chooses the first 8 layers of BERT-16 to extract the text semantics and also using a fully connected layer obtain the finally transformed text embedding.

### Light Punishment Contrastive Learning

In multimodal studies, contrastive learning is generally used to embed two single modalities into a unified space thereby eliminating semantic differences. Contrastive learning adopts a unique one-hot encoding method thus all negative pairs will be penalized resulting in an inaccurate multimodal embedding. We propose utilizing knowledge distillation to obtain more views' modality representation, combined with contrastive learning to soften the high punishment caused by weak correlations, such as mismatched image and text content in positive pairs, and effective matching of misaligned text in negative pairs.

Specifically, we put the encoded single modal embeddings $V^e$ and $T^e$ in two dedicated Multi-Layer Perceptron (MLP) then the features are transformed into the low-dimensional space obtain $e^v$ and $e^t$, utilize the similarity function $(sim)$ measures similarity scores via dot product. Within a batch size $N$ of training, the predicted image-text similarity $p_{\mathrm{ij}}^{\mathrm{v}\to\mathrm{t}}$ also as the text-image similarity $p_{\mathrm{ij}}^{\mathrm{t}\to\mathrm{v}}$ can be calculated as follows

$$p_{\mathrm{ij}}^{\mathrm{t}\to\mathrm{v}} = \frac{\exp\left(\mathrm{sim}\left(e_i^v, e_j^t\right)/\tau\right)}{\sum_{j=1}^{N}\exp\left(\mathrm{sim}\left(e_i^v, e_j^t\right)/\tau\right)}, \tag{1}$$

where $\tau$ is the learnable temperature parameter with an initial value of 0.07, the results compose the similarity matrix $P$. At this point, the image-text contrastive loss $\mathcal{L}_{ITC}$ can be calculated by the mean of image-text and text-image loss, in which the two losses are both calculated as follows

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} y_{ij}\log p_{ij}, \tag{2}$$

where the $y_{ij}$ is the corresponding one-hot vector to the real labels, in which the negative pair is denoted as 0 and the positive pair is denoted as 1.

To balance the presence of weak correlations, we choose the unimodal exponential moving average version of the base model as the momentum model, the momentum features are denoted as $V^s$ and $T^s$, obtaining $s^v$ and $s^t$ and unifying the dimension with the above $e^v$ and $e^t$. We maintain two queues of size $M$ $(M \ll N)$ to store the most recent image-text pairs. The momentum image-text similarity $s_{\mathrm{ij}}^{v\to\mathrm{t}}$ and momentum text-image similarity $s_{\mathrm{ij}}^{t\to\mathrm{v}}$ are computed as follows

$$s_{\mathrm{m}}^{\mathrm{t}\to\mathrm{v}} = \frac{\exp\left(\mathrm{sim}\left(s^v, s_m^t\right)/\tau\right)}{\sum_{m=1}^{M}\exp\left(\mathrm{sim}\left(s^v, s_m^t\right)/\tau\right)}. \tag{3}$$

Results compose the momentum similarity matrix $S$. Momentum Multimodal loss $\mathcal{L}_{MM}$ is calculated by cross-entropy using the similarity matrix $P$ and momentum similarity matrix $S$. The final momentum distillation-based light punishment contrastive loss is calculated as follows, with $\lambda$ being the learnable parameter with an initial value of 0.02:

$$\mathcal{L}_{LC} = \mathcal{L}_{ITC} + \lambda\mathcal{L}_{MM}. \tag{4}$$

### Adaptive Cross-Modal Fusion

Image and text as different modalities contain rich independent semantics and shared semantics. Existing research attempts to map image-to-text space or text-to-image, which ignores the unique qualities of discrete and continuous information resulting in semantic hidden. The activity of image and text features present higher and lower, thus needing to measure the importance as a shared part in modal fusion. We propose an adaptive cross-modal fusion module that can preserve the private characteristics of a single modality and adaptively adjust the mutual modal feature weights.

Specifically, given a sequence of image-text feature pairs, the previous layer vectors are denoted as $H_{l-1}$ where $l \in [1, L]$. We normalize the modalities method to the same dimension by RMSNorm. Define the image-text attention operation using the key projection matrix and value projection matrix for images and the query projection matrix for text calculated as follows

$$H_l^{Q_0} = H_{l-1}W_l^{Q_0}, \tag{5}$$

$$H_l^{K_1} = H_{l-1}W_l^{K_1}, H_l^{V_1} = H_{l-1}W_l^{V_1}, \tag{6}$$

$$C_l^0 = \mathrm{Softmax}\left(\frac{H_l^{Q_0}H_l^{K_1T}}{\sqrt{d}}\right)H_l^{V_1}, \tag{7}$$

where $W_l^{Q_0}$ $W_l^{K_1}$ $W_l^{V_1}$ are the learnable projection matrix, $C_l^0$ is the image-text cross-context feature, as a local shared feature it contains the intersection part with the text while retains more image feature, in the same way, we obtain the text-image cross-context feature $C_l^1$. To adaptive handle mutual modality information, we use the learnable modal weight parameter $\alpha$, to sum up the local image-text cross-context feature and text-image cross-context feature, bring the result $C$ to pass through another attention mechanism again with the image-text sequence handle global fusion, add a linear aggregate layer final obtain mutual modality, which is calculated as follows

$$C = (1-\alpha)C_l^0 + \alpha C_l^1, \tag{8}$$

$$F = \text{ Linear } \left(\text{Softmax}\left(\frac{H_0 C^T}{\sqrt{D}}\right) C\right). \tag{9}$$

To ensure modal collaboration in the same semantic space also used a shared Feed Forward Network (FFN). Mutual modality features $F$ are fed to a dedicated fake news detection projection to predict real and false news.

## 4.2 Deep inference

Mutual modality fuses semantically aligned image and text-sharing information, existing studies generally detect and locate forgery image and text in a single branch by mutual features. The instability of noise effects in real media news leads to semantic gaps between single modality and fusion modality while deepfake techniques exacerbate such semantic inaccuracies, thus it is difficult to predict all the forgery categories only through mutual modality features. The dual branch combines unimodal independent information, and deep reasoning for image and text to provide different views making the semantic gap complemented.

### Image Forgery Detection and Localization

Image modality as visual perceptual information has an important impact on the media propagation degree, and manipulation techniques against image modality are sophisticated and complex and require fine-grained inference. To enhance the key visual to weaken other interferences, we utilize the advantages of two image encode methods namely combining multi-level and multi-scale visual perception to form a Multi-View Extractor (MVE).

Specifically, we adopt ViT-B/16 to model the long-range contextual information of the image by extracting different encoder layers, obtain multi-level visual intermediate feature mapping $\left\{I_a^k\right\}_{k=1}^n$, where $n$ denotes the number of layers, splice all the visual features along the channel dimensions to capture the subtle changes in the image. The multi-level visual perceptual features contain initial spatial information and later semantic indications, calculated as follows

$$I_a^1 = \text{ViTBlock}_1(V), \tag{10}$$

$$I_a^k = \text{ViTBlock}_k\left(I_a^{k-1}\right), k = 2, \ldots, n, \tag{11}$$

$$I_a = \text{Concat}\left[I_a^1, I_a^2, \ldots, I_a^n\right]. \tag{12}$$

To model the local information of the image, we choose CNN network which performs well in edge and texture processing, obtain multi-scale visual features $\left\{I_b^k\right\}_{k=1}^m$, where $m$ is the number of scales of the image including multiple resolutions such as $\frac{H}{4} * \frac{W}{4}$, $\frac{H}{8} * \frac{W}{8}$, $\frac{H}{16} * \frac{W}{16}$ where $H * W$ is the initial resolution, transform to a uniform channel dimension and connect all the extracted visual features. The multi-scale visual perceptual features contain spatial details at multiple grains, calculated as follows

$$I_b^1 = \text{ConvBlock}_1(V_1), \tag{13}$$

$$I_b^k = \text{ConvBlock}_k\left(I_b^{k-1}\right), k = 2, \ldots, m, \tag{14}$$

$$I_b = \text{Concat}\left[I_b^1, I_b^2, \ldots, I_b^m\right]. \tag{15}$$

Splice multi-level and multi-scale visual perceptual features by channel and align dimension with the mutual modality features by a linear layer to obtain independent image modality features, sum up it with mutual modality features $F$. The image deep inference features $I_p$ are fed to the dedicated detection and localization projection to predict the forgery or origin image and mark the forged bounding box, by cross-entropy loss for binary classification and smoothing L1 loss for bounding box regression, the total image loss is calculated as follows

$$\mathcal{L}_{IMG} = \mathcal{L}_{IB\text{ic}}(I_p) + \mathcal{L}_{\text{IBox}}\left(I_p\right). \tag{16}$$

### Text Forgery Detection

The same semantic content can be represented in many different textual representations, thus the detection of text modality requires deep contextual interaction information. Pre-trained LLMs are trained on masses of language libraries to help them adapt to downstream tasks. Research in several areas has made significant progress based on this technology.

In this paper, we choose the last 4 layers of BERT-16 and a linear layer as the Context Aggregator (CA) and align dimension with mutual modality features $F$ to obtain a composite feature $T_p$ that fuses text uni-modal features and shared mutual modal features, feed into the dedicated text detection projection to predict authentic or falsehood text binary classification. The total loss $\mathcal{L}_{TOTAL}$ consists of four components is calculated as follows:

$$\mathcal{L}_{TOTAL} = \mathcal{L}_{LC} + \mathcal{L}_{BIC} + \mathcal{L}_{IMG} + \mathcal{L}_{TEX}. \tag{17}$$

# 5 Experiments

## 5.1 Datasets, Metrics, and Settings

The Twitter dataset [Khattar *et al.*, 2019] consists of tweets containing textual information, visual information, and social contextual information related to them. The Weibo dataset [Jin *et al.*, 2017] is from Xinhua News Agency and Weibo where each tweet contains three elements i.e. tweet id, text, and image. To comprehensively evaluate the model, we use the accuracy (ACC), area under the receiver operating characteristic curve (AUC), and the mean of F1 scores (mF1) to evaluate fake news detection, AUC, and ACC values for image and text detection. To verify the bounding box prediction, image localization calculates the intersection ratio between the true and predicted coordinates(IoU), sets thresholds 0.5, 0.75, and 0.9 to calculate the average accuracy, and selects IoUmean and IoU50 as the evaluation scales.

For the multi-modal multi-task approaches, CLIP [Radford *et al.*, 2021] co-trains an image encoder and a text encoder to predict the correct pairing of a batch of image-text samples. ViLT [Kim *et al.*, 2021]greatly simplifies the processing of visual inputs in the same convolution-free manner as text inputs. DGM$^4$ [Shao *et al.*, 2023] fuses the methods of contrast learning, multimodal feature fusion Attentional Mechanisms. RPPG-Fake [Zhang *et al.*, 2024a] explores the problem through the generation of propagation paths. For multi-modal single-task approaches, HMCAN [Qian *et al.*, 2021] uses a hierarchical contextual attention network, MEAN [Wei *et al.*, 2022] uses a multi-modal generator

| Categories | Multimodal | | | Image Binary | | Image Grounding | | Text Binary | |
|---|---|---|---|---|---|---|---|---|---|
| Methods | AUC | ACC | mF1 | AUC | ACC | IoUmean | IoU50 | AUC | ACC |
| CLIP | 79.53 | 73.48 | 73.09 | 80.54 | 72.16 | 46.04 | 46.04 | 69.61 | 65.46 |
| ViLP | 82.52 | 77.43 | 77.58 | 82.88 | 75.51 | 50.39 | 54.99 | 71.13 | 68.61 |
| DGM$^4$ | 91.62 | 84.92 | 84.73 | 92.59 | 85.47 | 74.32 | 81.25 | 94.02 | 92.71 |
| RPPG-Fake | 91.42 | 84.22 | 84.67 | 90.11 | 83.45 | 70.26 | 79.59 | 90.47 | 88.44 |
| **SDML (Ours)** | **92.43** | **85.54** | **85.83** | **95.65** | **88.41** | **77.83** | **84.39** | **95.76** | **93.10** |

Table 2: Comparison of multi-modal multi-task models on MFND dataset.

| Categories | Multimodal | | | Image Binary | | Image Grounding | | Text Binary | |
|---|---|---|---|---|---|---|---|---|---|
| Methods | AUC | ACC | mF1 | AUC | ACC | IoUmean | IoU50 | AUC | ACC |
| CLIP | 83.10 | 76.21 | 76.77 | 83.75 | 75.38 | 49.26 | 50.08 | 72.97 | 68.68 |
| ViLP | 85.30 | 78.26 | 78. 44 | 85.11 | 78.56 | 60.36 | 68.84 | 75.00 | 71.37 |
| DGM$^4$ | 93.23 | **86.39** | 86.11 | 93.22 | 87.23 | 76.70 | 83.49 | 95.16 | 94.27 |
| RPPG-Fake | 92.02 | 85.69 | 85.97 | 90.64 | 84.03 | 71.19 | 81.44 | 93.48 | 92.95 |
| **SDML (Ours)** | **93.67** | 86.14 | **86.33** | **96.12** | **89.31** | **78.36** | **84.60** | **96.14** | **94.94** |

Table 3: Comparison of multi-modal multi-task models on DGM$^4$ dataset.

| Datasets | Weibo | | Twitter | |
|---|---|---|---|---|
| Methods | ACC | mF1 | ACC | mF1 |
| HMCAN | 88.52 | 88.55 | 89.71 | 89.54 |
| MEAN | 89.49 | 89.15 | 78.42 | 78.49 |
| COOLANT | 92.30 | **92.63** | 90.04 | 90.81 |
| Event-Randar | 91.94 | 91.90 | 92.84 | 92.31 |
| **SDML (Ours)** | **92.61** | 92.44 | **93.07** | **93.61** |

Table 4: Comparison with single-task models on three datasets.

and a dual discriminator for adversarial training, COOLANT [Zhang *et al.*, 2021] uses an attention-guided contrast and cross-fertilization framework, and Event-Radar [Ma *et al.*, 2024] uses a graph-structured event consistent encoder and multi-view fusion.

All experiments are implemented on the Pytorch deep learning framework. The momentum queue size is set to 65535, the media detection, image detection, image localization, and text detection projections are set to three different multi-layer perceptual with output dimensions of 2, 2, 4, and 2. The model trained for 100 epochs with a batch size of 64, AdamW optimizer, with a weight decay of 0.005, in the first 1000 steps the learning rate is warmed up to $5e^{-6}$, decaying to $5e^{-7}$ after the cosine scheduling.

## 5.2 Comparison Results

**Results on MFND dataset.** We compare three multi-modal multi-task baseline methods with SDML on our newly proposed MFND dataset, and the results are shown in Table 2. SDML significantly outperforms the three baseline methods on all evaluation metrics, and the image localization task shows the most pronounced superiority, with IoUmean and IoU50 values exceeding the highest value by 3.51% and 3.14%, respectively. This suggests that shallow-deep inference under dual branching can comprehensively and accurately capture the interactions between images and text, and track the semantic changes caused by the operations.

**Results on DGM$^4$ dataset.** We compare our methods with the same multi-modal multi-task methods on the DGM$^4$ dataset, and the comparison results are listed in Table 3. The results show that our method outperforms DGM$^4$ by 2% in both AUC and ACC values on the image detection task, which indicates that unimodal data enhancement guides feature representation learning. SDML also performs superiorly on multimodal and text detection tasks, e.g., the AUC value of multimodal is improved by 0.44% and the AUC value of text is enhanced by 0.98%.

**Comparison with single-task methods.** We also evaluate our method as well as four multi-modal single-task methods on two other datasets, all of which perform binary detection of multimodal fake news. The comparison results are shown in Table 4, where the ACC and F1 values of our models are significantly better than the other baseline methods on all datasets. In addition, comparing the different performances of SDML on the same dataset, the effect of binary classification under single-task decreases significantly, which suggests that multi-task guides the model to learn stable features to improve the prediction accuracy.

## 5.3 Ablation Study

**Ablation of modalities.** To validate the multimodal relevance of images and text, we perform an ablation experiment, and the results from rows 1-2 of Table 5 show that the full multimodal version outperforms the eliminated portion. The experiments demonstrate that multimodal information interacts with each other and capturing common semantics is particularly important for feature learning.

**Ablation of different modules.** To verify the importance of different modules to our SDML model, we set up several ab-

| Categories | Multimodal | | | Image Binary | | Image Grounding | | Text Binary | |
|---|---|---|---|---|---|---|---|---|---|
| Methods | AUC | ACC | mF1 | AUC | ACC | IoUmean | IoU50 | AUC | ACC |
| w/o Image | | | | | | | | 77.62 | 73.77 |
| w/o Text | | | | 94.89 | 88.08 | 77.21 | 83.92 | | |
| w/o LPCL | 90.43 | 83.35 | 83.65 | 93.84 | 86.55 | 65.46 | 77.37 | 91.62 | 89.58 |
| w/o ACMF | 89.06 | 82.19 | 81.94 | 92.70 | 85.37 | 64.08 | 74.09 | 91.73 | 89.43 |
| w/o MVE | 91.25 | 84.15 | 84.13 | 93.38 | 86.25 | 74.16 | 82.75 | 93.86 | 91.27 |
| w/o CA | 91.44 | 84.92 | 84.49 | 94.77 | 87.71 | 76.49 | **84.47** | 93.29 | 91.09 |
| **SDML (Ours)** | **92.43** | **85.54** | **85.83** | **95.65** | **88.41** | **77.83** | 84.39 | **95.76** | **93.10** |

Table 5: Ablation results of modalities and modules. 'LPCL', 'ACMF', 'MVE', and 'CA' are short for 'Light Punishment Contrastive Learning', 'Adaptive Cross-Modal Fusion', 'Multi-View Extractor', and 'Context Aggregator', respectively.

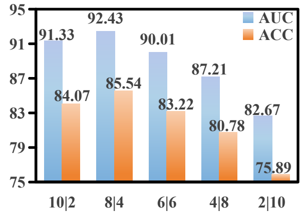

Figure 4: Ablation on layer numbers of text encoder and Contextual Aggregator.

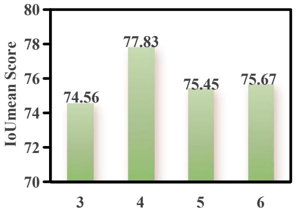

Figure 5: Ablation on numbers of Multi-view Extractor.

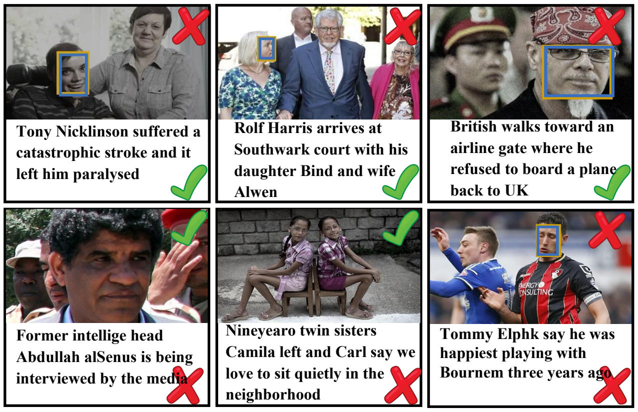
Tony Nicklinson suffered a catastrophic stroke and it left him paralysed
Rolf Harris arrives at Southwark court with his daughter Bind and wife Alwen
British walks toward an airline gate where he refused to board a plane back to UK
Former intellige head Abdullah alSenus is being interviewed by the media
Nineyearo twin sisters Camila left and Carl say we love to sit quietly in the neighborhood
Tommy Elphk say he was happiest playing with Bournem three years ago

Figure 6: Visualization of our predictions (real in green Checkmark while fake in red XSolid). The ground truth and prediction bounding boxes in the images are in yellow and blue, respectively.

lation experiments, and the results are listed in rows 3-6 of Table 5. We can find that removing light punishment contrastive learning (row 3) or removing adaptive cross-modal fusion (row 4) results in a sharp decrease in accuracy for each task, suggesting that the model needs to complete modal alignment and fusion at an early stage to facilitate deeper reasoning. In addition, removing the multi-view extractor (row 5) and context aggregator (row 6) is correspondingly less effective, suggesting that the unimodal semantics contain separate information from the mutual modality and can be used as complementary augmentation data.

**Impact of layer numbers of text encoder and Contextual Aggregator.** We consider five scenarios regarding the layers for the text encoder and contextual aggregator. The settings and results are shown in Fig. 4. In terms of the evaluation metrics AUC and ACC, models with ratio setting 8|4 (i.e., 8 layers text encoder and 4 layers contextual aggregator) obtain the best results while performance degrades faster when the ratio becomes larger.

**Impact of numbers of Multi-view Extractor.** The multi-view extractor is composed of multi-layer and multi-scale visual perception modules. Here we compare layers number 3, 4, 5, and 6 in combination with the corresponding scale number, and select the IoUmean as the evaluation score. The results in Fig. 5, illustrate that the best performance can be achieved when the values of n and m are set to be 4.

### 5.4 Visualization and Analysis

We provide some visualization results for fake news detection and localization in Figure 6. The first row is all fake news in the Fake Images Real Text category, generated by FS, EFS, and AM from left to right, and the second row is all fake news in the Fake Text category, in which the faked images of the rightmost news samples are combined and the rest of the samples are real images. The visualization results use pairwise error symbols to mark the detection results and use bounding boxes to locate the forgery images, which demonstrates that our method can accurately detect manipulated images under different types of deep forgery techniques, as well as manipulated text generated by a large language model under the influence of real media news noise.

## 6 Conclusion

In this paper, we propose a large-scale complex MFND dataset with richer annotations to forge multimodal news through the latest generation techniques. To facilitate the detection and localization tasks in fake news detection, we propose a Shallow-Deep Multitask Learning (SDML) model to accomplish a deeper understanding of the interactions between unimodal and mixed modalities, improving its interpretability and robustness. We evaluate its effectiveness and superiority by conducting baseline comparison synthesis experiments on mainstream and MFND datasets.

## Acknowledgments

This work was supported by Natural Science Foundation of Hebei Province under the Grant F2024202017, Basic Research Project of Hebei Universities in Shijiazhuang under the Grant 241790817A, the Central Guidance Fund for Local Science and Technology Development Projects under the Grant 246Z0106G, Guangdong Basic and Applied Basic Research Foundation under the Grant 2023A1515140037, Guangdong Provincial Key Laboratory under the Grant 2023B1212060076, and Guangdong Key Laboratory of Information Security Technology, Sun Yat-sen University. The computational resources are supported by SongShan Lake HPC Center (SSL-HPC) in Great Bay University.

## References

[Chen *et al.*, 2023] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023.

[Dolhansky *et al.*, 2020] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*, 2020.

[Gao *et al.*, 2023] Zan Gao, Chao Sun, Zhiyong Cheng, Weili Guan, Anan Liu, and Meng Wang. Tbnet: A two-stream boundary-aware network for generic image manipulation localization. *IEEE Transactions on Knowledge and Data Engineering*, 35(7):7541–7556, 2023.

[Gao *et al.*, 2024] Zhangwei Gao, Zhe Chen, Erfei Cui, Yiming Ren, Weiyun Wang, Jinguo Zhu, Hao Tian, Shenglong Ye, Junjun He, Xizhou Zhu, et al. Mini-internvl: a flexible-transfer pocket multi-modal model with 5% parameters and 90% performance. *Visual Intelligence*, 2(1):1–17, 2024.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[Jin *et al.*, 2017] Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 795–816, 2017.

[Karras *et al.*, 2021] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in neural information processing systems*, 34:852–863, 2021.

[Karras, 2017] Tero Karras. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[Kenton and Toutanova, 2019] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2. Minneapolis, Minnesota, 2019.

[Khattar *et al.*, 2019] Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. Mvae: Multimodal variational autoencoder for fake news detection. In *The world wide web conference*, pages 2915–2921, 2019.

[Kim *et al.*, 2021] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR, 2021.

[Li *et al.*, 2023] Dong Li, Jiaying Zhu, Menglu Wang, Jiawei Liu, Xueyang Fu, and Zheng-Jun Zha. Edge-aware regional message passing controller for image forgery localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8222–8232, 2023.

[Liu *et al.*, 2015a] Xiaomo Liu, Armineh Nourbakhsh, Quanzhi Li, Rui Fang, and Sameena Shah. Real-time rumor debunking on twitter. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, page 1867–1870, New York, NY, USA, 2015. Association for Computing Machinery.

[Liu *et al.*, 2015b] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

[Liu *et al.*, 2020] Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. Visual news: Benchmark and challenges in news image captioning. *arXiv preprint arXiv:2010.03743*, 2020.

[Luvembe *et al.*, 2024] Alex Munyole Luvembe, Weimin Li, Shaohau Li, Fangfang Liu, and Xing Wu. Caf-odnn: Complementary attention fusion with optimized deep neural network for multimodal fake news detection. *Information Processing & Management*, 61(3):103653, 2024.

[Ma *et al.*, 2015] Jing Ma, Wei Gao, Zhongyu Wei, Yueming Lu, and Kam-Fai Wong. Detect rumors using time series of social context information on microblogging websites. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, page 1751–1754, New York, NY, USA, 2015. Association for Computing Machinery.

[Ma *et al.*, 2019] Jianxin Ma, Chang Zhou, Peng Cui, Hongxia Yang, and Wenwu Zhu. Learning disentangled representations for recommendation. *Advances in neural information processing systems*, 32, 2019.

[Ma *et al.*, 2024] Zihan Ma, Minnan Luo, Hao Guo, Zhi Zeng, Yiran Hao, and Xiang Zhao. 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)*, pages 5809–5821, 2024.

[Pernuš *et al.*, 2023] Martin Pernuš, Vitomir Štruc, and Simon Dobrišek. Maskfacegan: High resolution face editing with masked gan latent code optimization. *IEEE Transactions on Image Processing*, 2023.

[Qian *et al.*, 2021] Shengsheng Qian, Jinguang Wang, Jun Hu, Quan Fang, and Changsheng Xu. Hierarchical multi-modal contextual attention network for fake news detection. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 153–162, 2021.

[Radford *et al.*, 2019] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[Shao *et al.*, 2023] Rui Shao, Tianxing Wu, and Ziwei Liu. Detecting and grounding multi-modal media manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6904–6913, 2023.

[Shi *et al.*, 2023] Zenan Shi, Haipeng Chen, and Dong Zhang. Transformer-auxiliary neural networks for image manipulation localization by operator inductions. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(9):4907–4920, 2023.

[Tu *et al.*, 2024] Xiaoguang Tu, Zhi He, Yi Huang, Zhi-Hao Zhang, Ming Yang, and Jian Zhao. An overview of large ai models and their applications. *Visual Intelligence*, 2(1):1–22, 2024.

[Wang *et al.*, 2023] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023.

[Wang *et al.*, 2024] Wei-Yao Wang, Yu-Chieh Chang, and Wen-Chih Peng. Style-news: Incorporating stylized news generation and adversarial verification for neural fake news detection. *arXiv preprint arXiv:2401.15509*, 2024.

[Wei *et al.*, 2022] Pengfei Wei, Fei Wu, Ying Sun, Hong Zhou, and Xiao-Yuan Jing. Modality and event adversarial networks for multi-modal fake news detection. *IEEE Signal Processing Letters*, 29:1382–1386, 2022.

[Xie *et al.*, 2024] Xinyu Xie, Yawen Cui, Tao Tan, Xubin Zheng, and Zitong Yu. Fusionmamba: Dynamic feature enhancement for multimodal image fusion with mamba. *Visual Intelligence*, 2(1):37, 2024.

[Xu *et al.*, 2022] Yangyang Xu, Bailin Deng, Junle Wang, Yanqing Jing, Jia Pan, and Shengfeng He. High-resolution face swapping via latent semantics disentanglement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7642–7651, 2022.

[Yao *et al.*, 2021] Xu Yao, Alasdair Newson, Yann Gousseau, and Pierre Hellier. A latent transformer for disentangled face editing in images and videos. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13769–13778, 2021.

[Zhang *et al.*, 2021] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 833–842, 2021.

[Zhang *et al.*, 2024a] Litian Zhang, Xiaoming Zhang, Ziyi Zhou, Xi Zhang, Senzhang Wang, Philip S Yu, and Chaozhuo Li. Early detection of multimodal fake news via reinforced propagation path generation. *IEEE TKDE*, 2024.

[Zhang *et al.*, 2024b] Yaning Zhang, Zitong Yu, Tianyi Wang, Xiaobin Huang, Linlin Shen, Zan Gao, and Jianfeng Ren. Genface: A large-scale fine-grained face forgery benchmark and cross appearance-edge learning. *IEEE Transactions on Information Forensics and Security*, 2024.

[Zhao *et al.*, 2021] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2185–2194, 2021.

[Zubiaga *et al.*, 2017] Arkaitz Zubiaga, Maria Liakata, and Rob Procter. 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*, pages 109–123. Springer, 2017.