Cognitive Computation
https://doi.org/10.1007/s12559-024-10316-x

RESEARCH

Multi-Modal Generative DeepFake Detection via Visual-Language Pretraining with Gate Fusion for Cognitive Computation

Guisheng Zhang[1] · Mingliang Gao[1] · Qilei Li[2] · Wenzhe Zhai[1] · Gwanggil Jeon[3]

Received: 24 October 2023 / Accepted: 28 May 2024

Abstract

With the widespread adoption of deep learning, there has been a notable increase in the prevalence of multimodal deepfake content. These deepfakes pose a substantial risk to both individual privacy and the security of their assets. In response to this pressing issue, researchers have undertaken substantial endeavors in utilizing generative AI and cognitive computation to leverage multimodal data to detect deepfakes. However, the efforts thus far have fallen short of fully exploiting the extensive reservoir of multimodal feature information, which leads to a deficiency in leveraging spatial information across multiple dimensions. In this study, we introduce a framework called Visual-Language Pretraining with Gate Fusion (VLP-GF), designed to identify multimodal deceptive content and enhance the accurate localization of manipulated regions within both images and textual annotations. Specifically, we introduce an adaptive fusion module tailored to integrate local and global information simultaneously. This module captures global context and local details concurrently, thereby improving the performance of image bounding-box grounding within the system. Additionally, to maximize the utilization of semantic information from diverse modalities, we incorporate a gating mechanism to strengthen the interaction of multimodal information further. Through a series of ablation experiments and comprehensive comparisons with state-of-the-art approaches on extensive benchmark datasets, we empirically demonstrate the superior efficacy of VLP-GF.

Keywords Multimodal deepfake · Deepfake detection · Generative AI · Cognitive computation · Manipulation grounding

Introduction

As deep learning technologies continue to advance at an extraordinary rate, deepfakes have attracted significant attention from both the academic and technological domains. Deepfake systems utilize sophisticated Generative Adversarial Networks (GANs) [1] to generate highly convincing yet completely synthetic multimedia content, which often involves manipulated text or images. The initial foray into the realm of deepfakes was driven by benign intentions. Across a multitude of domains [2–6], this technology has substantially augmented user convenience. This fusion of deepfake detection and cognitive computation represents a cutting-edge approach to safeguarding individual privacy and asset security in the face of the deepfake challenge.

However, with the advancement of technology, there arises a worrisome outlook. The negative ramifications of deepfake technology encompass the dissemination of misinformation, covert digital maneuvers, and the creation of deceptive simulations [7]. The deceptive nature of these manipulated media fragments can undermine confidence in both visual and auditory evidence. The dissemination of falsified information not only encroaches upon individual privacy but also undermines societal stability. Consequently, given the challenges presented by deepfakes, a multitude of scholars in the field are actively developing detection methods [8–13]. They apply these techniques to mitigate the widespread threat of deepfakes. In real-world scenarios, deepfakes extend beyond isolated instances involving images or text. They often encompass deepfakes involving two or more modalities. While acknowledging the commend-

✉ Mingliang Gao
mlgao@sdut.edu.cn

✉ Gwanggil Jeon
gjeon@gmail.com

[1] School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo 255000, China

[2] School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, UK

[3] Department of Embedded Systems Engineering, Incheon National University, Incheon 22012, South Korea

Published online: 25 June 2024

able efforts in unimodal deepfake detection, it is crucial to recognize that the challenges posed by multimodal versions are more formidable.

In contrast to solely employing fabricated images or text, multimodal content deepfakes are more difficult to distinguish, increasing the potential risks in our daily routines. In response to this urgent concern, numerous scholars have extensively investigated the realm of multimodal deepfake detection [14–19]. Some studies [16, 17, 19] concentrate on scrutinizing human-generated multimodal fabrications, whereas others [14, 15, 18] aim at identifying out-of-context manipulation by amalgamating genuine images with modified text. The recent study, HAMMER [20], endeavors to identify and interpret multimodal deepfakes. The model undergoes training using extensive deepfake datasets that simulate prevalent instances of real-world disinformation. Beyond mere binary deepfake identification, HAMMER provides a nuanced interpretation of manipulative elements. Although these frameworks occupy a crucial role in the realm of multimodal deepfake detection, their proficiency in cross-domain embeddings and the utilization of diverse feature information is deficient.

In this work, the presented network, Vision-Language Pre-training with Gate Fusion module (VLP-GF), endeavors to enhance the accuracy of task detection by facilitating the interaction of information from diverse features. Consequently, this enables the system to conduct accurate multi-modal detection and verify fake information. Specifically, the VLP-GF integrates the corresponding embeddings of images and text into a shared feature space. Similar feature embeddings tend to cluster closely in the feature space during cross-modal interactions, whereas dissimilar feature embeddings exhibit more significant spatial separation. Similarly, this principle applies to intra-modal interactions as well. Additionally, VLP-GF promotes an adaptive interaction between global and local features, thereby enhancing the contextual depth of local features in contrast to their original representations. This augmentation provides significant benefits in enhancing the task of grounding image-bounding boxes. The proposed framework incorporates a gating mechanism for consolidating features extracted from diverse modalities. This method improves the viability of performing binary and multi-class classification tasks by amalgamating data from various modalities. In sum, the contributions of the work are three-fold:

- We propose a comprehensive framework for the detection of multimodal deepfakes that utilizes a wide range of features to enhance information sharing. Our framework exhibits exceptional performance when compared to the most recent state-of-the-art (SOTA) methods, which suggests significant potential for real-world applications.
- We introduce a module for an adaptive fusion of local and global features to simultaneously capture global context and local details. This enhancement improves the model's capacity to understand and represent intricate data, which enhances performance in localizing image-bounding boxes.
- We employ a gating mechanism to combine image and text features and facilitate a thorough integration of visual-linguistic semantics. This approach aims to improve the accuracy of binary and multiclass classifications.

The subsequent sections of this paper are structured as follows: In "Related Work" section, we provide a comprehensive review of the pertinent literature. In "Method" section, we elucidate the intricacies of the proposed VLP-GF model. "Experiments" section encompasses comprehensive comparative analyses and ablation studies aimed at evaluating the proposed model. The conclusion is presented in "Conclusion" section.

Related Work

Deefake Detection

Deepfake detection constitutes a forefront research domain dedicated to the identification and mitigation of deepfake content. Detecting deepfakes often leans towards a binary classification approach. Within the deepfake detection framework, various classifiers are employed to differentiate between authentic and manipulated images. In the context of deepfake detection, a clear dichotomy arises: unimodal techniques concentrate on a single sensory modality, whereas multimodal approaches incorporate multiple modalities.

Unimodal Deepfake Detection In the early stages, unimodal deepfake detection methods [21–25] have shown promising results. In pursuit of improving the detection of deepfake content within resource-constrained environments, Chen et al. [24] employed DefakeHop++, a lightweight yet powerful framework that extends the principles established by the earlier DefakeHop method. This approach has been specifically tailored and fine-tuned for deployment on devices characterized by limited computational resources, which include smartphones and edge computing platforms. Patel et al. [25] reported a deep-CNN (D-CNN) architecture aimed at improving the detection of deepfake phenomena. The method utilizes images from various sources, thereby enhancing its overall applicability.

Multimodal Deepfake Detection The formulation of multimodal deepfakes necessitates the artful fusion of information originating from various modalities, such as images, text, and sound. This amalgamation process is executed through advanced deep-learning techniques. Multimodal deepfake detection methods aim to meticulously examine the authenticity of these manipulated elements. There are plenty of works [14–17, 19, 20] that study the detection of multimodal deepfakes. Abdelnabi et al. [14] put forward the Consistency-Checking Network (CCN), which emulates the hierarchical cognitive processes employed by humans across different modalities. The architecture can accurately detect deepfakes by employing consistent cues extracted from diverse online sources and the provided image-caption correlation. Recently, a Hierarchical Multi-modal Manipulation Reasoning Transformer (HAMMER) [20] was proposed to examine the complex interactions among different modalities. This approach employs a pair of uni-modal encoders to conduct surface-level analysis through the application of manipulation-aware contrastive learning. A multimodal aggregator enhances these features even further. Therefore, this model enables a more profound comprehension via modality-aware cross-attention mechanisms.

In this work, we put forward a comprehensive defense strategy against multimodal deepfake manipulations. The framework enables the detection of modal image distortions, such as facial alterations, as well as textual misinformation. The comparison between conventional multi-modal deepfake detection and our proposed approach is illustrated in Fig. 1. In contrast to conventional approaches, our proposed framework can concurrently execute multiple tasks.

Gate Fusion

Initially influenced by the gating mechanisms inherent in LSTMs and GRUs, gate fusion has developed into a versatile methodology for selectively integrating features from various modalities or network layers. Through strategic feature amalgamation, gate fusion empowers the model to precisely concentrate on the essential data elements pertinent to the current task. This enhancement significantly contributes to improved generalization and overall efficacy. There exists a considerable body of research [26–30] focused on gated fusion methodologies. Arevalo et al. [26] proposed a concept called the Gated Multimodal Unit (GMU) with the aim of adaptively learning how to integrate information from various input streams. The GMU dynamically allocates weights to different input modalities to regulate its activation. To incorporate visual information into both large-scale text-only Neural Machine Translation (NMT) and multimodal NMT, Zhang et al. [28] employed an attention layer equipped with gated weighting. This mechanism functions to integrate visual and textual information, which subsequently serves as input for the decoder in the generation of target translations. Zhu et al. [27] introduced the Dynamic Memory Generative Adversarial Network (DM-GAN). Within this framework, a memory-writing gate is employed to emphasize crucial textual information, while a response gate combines images with memory attributes. Zhang et al. [30] put forward a Multimodal-CoT model, a two-stage architecture meticulously crafted to integrate textual and visual components. The Multimodal-CoT employs an adaptive gated fusion mechanism to achieve optimal fusion of text and image features.

In this work, drawing from the aforementioned methods, we employ a gated fusion mechanism to enhance the synergy between textual and visual elements. By doing so, we utilize cross-modal information to improve feature discrimination by aggregating features from various modalities.

Fusion of Global and Local Information

In the interdisciplinary domain of deep learning and computer vision, the integration of global and local information emerges as a fundamental subject of investigation across various research pursuits. The global features offer a holistic perspective of the image, while local features concentrate on intricacies within designated regions. The amalgamation of these attributes consistently enhances the robustness and precision of models. Numerous studies [31–36] have seamlessly integrated global and local features, which leads to improved outcomes. Fang et al. [31] build a method for consolidating distinctive facial features for face detection. The approach involved the utilization of the Adaptive Neuro-Fuzzy Inference System (ANFIS) and Support Vector Machine (SVM) to create personalized feature profiles for each subject and integrate both global and local features. Zou et al. [33] put forward a fusion paradigm termed multiscale completed local binary patterns. By using BoVM [37] and spatial pyramid matching (SPM) [38] with global ones via MS-CLBP, the method effectively explores the symbiotic relationship between local and global domains. Yang et al. [35] proposed

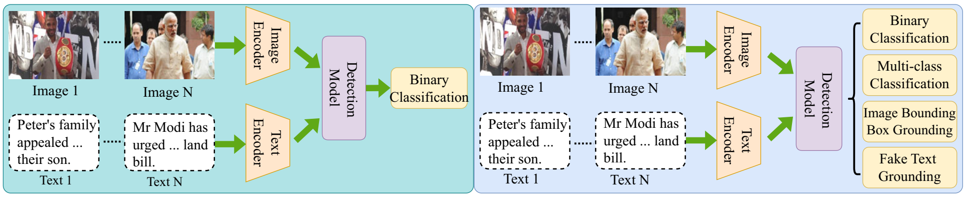


Image 1
Image N
Image Encoder
Peter's family appealed ... their son.
Text 1
Mr Modi has urged ... land bill.
Text N
Text Encoder
Detection Model
Binary Classification
Image 1
Image N
Image Encoder
Peter's family appealed ... their son.
Text 1
Mr Modi has urged ... land bill.
Text N
Text Encoder
Detection Model
Binary Classification
Multi-class Classification
Image Bounding Box Grounding
Fake Text Grounding

(a). Traditional multi-modal deepfake detection

(b). The proposed multi-modal deepfake detection

Fig. 1 Comparison between the conventional method and our method. **a** Traditional multi-modal deepfake detection. **b** The proposed multi-modal deepfake detection

a Deep Orthogonal Local and Global (DOLG) paradigm as a method for achieving seamless image retrieval. In DDLG model, an orthogonal fusion module effectively integrates both local and global features. Through objective-driven training, these components bolster each other to generate a concise and representative descriptor.

In this work, we introduce an Adaptive Global–Local Feature Fusion Module. By integrating global features into local features, we can grasp the comprehensive environmental and contextual context of an object. Through the synthesis of these characteristics, we can achieve a more nuanced interpretation of objects and scenes depicted in images. This can improve the accuracy of bounding box placement.

Contrastive Learning

Contrastive learning is a prominent and widely explored research domain within the fields of natural language processing, computer vision, and deep learning. It involves acquiring representations by contrasting positive and negative examples, and it emphasizes the similarity between data points while pushing dissimilar ones apart. In recent years, contrastive learning has exhibited outstanding performance across a wide range of tasks, particularly in the domains of unsupervised and semi-supervised learning. Gutmann and Hyvärinen [39] introduced an estimation approach tailored for parameterized statistical models. Utilizing nonlinear logistic regression, this approach aims to differentiate observed data from noise generated by the model. A SimCLR model [40] was proposed by Chen et al. which stands as a testament to the power of contrastive learning. The objective of this model is to enhance the alignment between an original data image and its various augmented iterations. The method dissolves ties between disparate modified facets of images by drawing upon contrastive loss. Oord et al. [41] reported an all-encompassing unsupervised learning technique named Contrastive Predictive Coding. This technique efficiently enables the latent space to predict subsequent samples through the use of a probabilistic contrastive loss. Recently, a method grounded in Triple Contrastive Learning (TCL) [41] was introduced by Yang et al. The TCL intensively amplifies the mean mutual information to capture the localized and structural nuances present in both image and textual inputs, and it connects local domains within the image or text to provide a comprehensive overview.

In this work, to extract the semantic connections between images and text, we employ cross-modal contrastive learning to align image and text embeddings produced by two unimodal encoders.

Vision-Language Pre-Train Methods

In the evolving landscape of deep learning, Vision-Language Pre-training (VLP) Methods have emerged as a seminal research direction. It investigates the convergence of various visual information, including images and videos, alongside linguistic expressions such as text-based descriptions and sentences. The utility of VLP extends across diverse visual endeavors, notably in visual question responding [42], textual discernment in visual deduction [43], visual referential idioms [44], and phrase-centric positioning [45]. Recently, numerous researchers have delved into Vision-Language Pre-training Methods [46–51]. Tan and Bansal [48] introduced a Learning Cross-Modality Encoder Representations from Transformers (LXMERT) approach for comprehending the correlations between visual and linguistic elements. This method enables the acquisition of intra-modal as well as cross-modal associations. Bhargava [46] proposed an adaptive technique to enhance model clarity and computational efficacy. They also investigated attention durations by employing both sparse and structured dropout strategies. This sheds light on how their attention mechanisms operate across both visual and linguistic challenges. Li et al. [50] reported a loss function called ALBEF, designed to synchronize visual and textual data. This synchronization is accomplished prior to their integration through cross-modal attention, and it bolsters the foundation of vision-language learning.

In this work, we employ the extensive vision-language pre-training framework, ALBEF, to establish alignment between unimodal image and text representations, thereby revealing their shared semantic content Subsequently, these aligned text-image pairs are employed for multimodel deepfake detection.

Method

Overview

This work aims to tackle the challenge of countering multimodal deepfakes by developing a comprehensive framework that integrates data from both images and text. The main goals of the proposed model encompass detecting potential deepfake alterations in images, identifying altered regions within images, recognizing textual manipulation in descriptions, and pinpointing changes in specific words. Constructing a comprehensive framework to simultaneously detect deepfakes and establish connections between visual and linguistic modalities represents a formidable undertaking. Detecting multimodal misinformation is increasingly challenging due to the wide array of data sources and the disruptive influence of deepfake-related distortions Therefore, within the domain of multimodal detection, the enhancement of detection performance by maximizing the utilization of diverse features presents a significant and essential challenge.

In pursuit of obtaining a multimodal representation and harnessing their collaborative capabilities for precise misinformation identification, we introduce a framework known as Vision-Language Pre-training with Gate Fusion (VLP-GF) framework. The primary objective of this model is to effectively leverage a wide range of features. These features encompass a variety of global and local characteristics, as well as attributes related to both images and text. The overall structure of the VLP-GF is illustrated in Fig. 2.

The presented framework consists of three main components: (a) The multi-modal feature extraction and alignment module. This module employs a Vision Transformer (ViT) [52] for image feature extraction, BERT [53] for text feature extraction, and utilizes contrastive learning for semantic alignment. (b) Feature fusion module: This transcends mere amalgamation and actively focuses on the acquisition of distinguishing insights from the concurrent manifestation of the two modalities. Additionally, to enhance contextual information on the local features, our approach employs an adaptive technique to merge global and local features. (c) Multimodal detection modules: The proposed model utilizes multi-task learning modules to gain fine-grained detection and grounding.

Consider a paired input denoted as $M = [I, T]$, where I signifies an image and T encompasses a textual description. The primary objective of the proposed framework is to extract latent semantic information inherent in the given input. Subsequently, the framework employs this information to detect and identify multimodal deepfakes.

To achieve this, we use a ViT and BERT to extract features from images and texts. These models play a pivotal role in transforming input data into unimodal representations, and they offer both conciseness and expressiveness in depicting the underlying semantic content. To achieve semantic alignment across multimodal data, we recommend employing intra-modal and cross-modal contrastive learning techniques, which can facilitate their convergence in the feature space. In pursuit of maximizing the collaborative influence among these representations, we have developed a gated fusion strategy proficient in harnessing multi-modal features to their maximum potential. Meanwhile, we employ a Local–Global Feature model (LGF) to capture both global context and local details. The architecture of the proposed framework has been meticulously designed to enable multimodal deepfake detection and leverage various distinctive traits to their maximum potential.

Vision-Language Feature Gated Fusion

For multimodal detection tasks, it is essential to facilitate robust interactions between different modalities. Therefore, following the extraction of visual embeddings denoted as $P_I(i)$ using the image encoder, and linguistic features represented as $P_T(t)$ via the text encoder, they undergo cross-attention mechanisms to facilitate inter-modal information exchange. Cross-modal attention is a variant of self-attention. Cross-modal attention facilitates the interaction between textual and visual components, which can enhance comprehension of semantic relationships across different data modalities. The structural representation of cross-attention is depicted in Fig. 3.

Specifically, by utilizing the cross-attention framework, one modality is assigned as the query (Q), and a different modality operates as both the key (K) and value (V). Subsequently, these constituents are input into the cross-attention layer to enable the blending of modalities. The cross-attention is represented as

$$\text{Cross-Attention}(Q, K, V) = \omega\left(\frac{Q \cdot K^T}{\sqrt{d_k}}\right) \cdot V, \tag{1}$$

where $\omega(\cdot)$ is an activation function, typically embodied as $Softmax(\cdot/\sqrt{d_k})$. In tackling the challenge of multimodal deepfake detection, the proposed method incorpo-

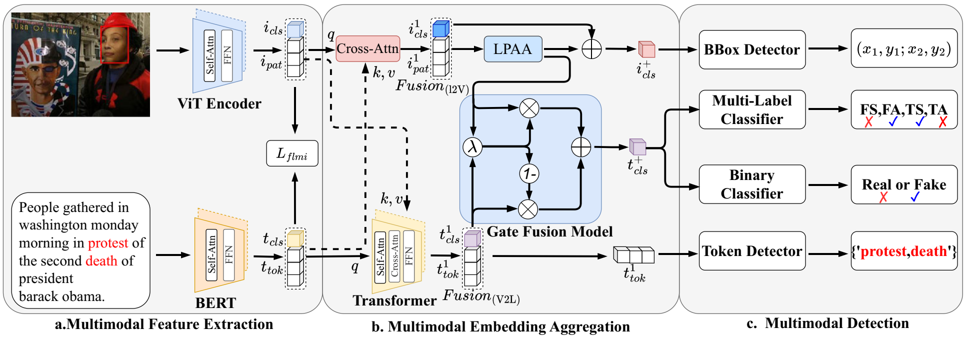


Fig. 2 The overarching structure of VLP-GF comprises three essential components: **a** A multi-modal feature extraction and alignment module. **b** A multimodal feature fusion module that integrates features from varied modalities. **c** Multimodal deepfake detection modules for detailed detection and bounding box grounding

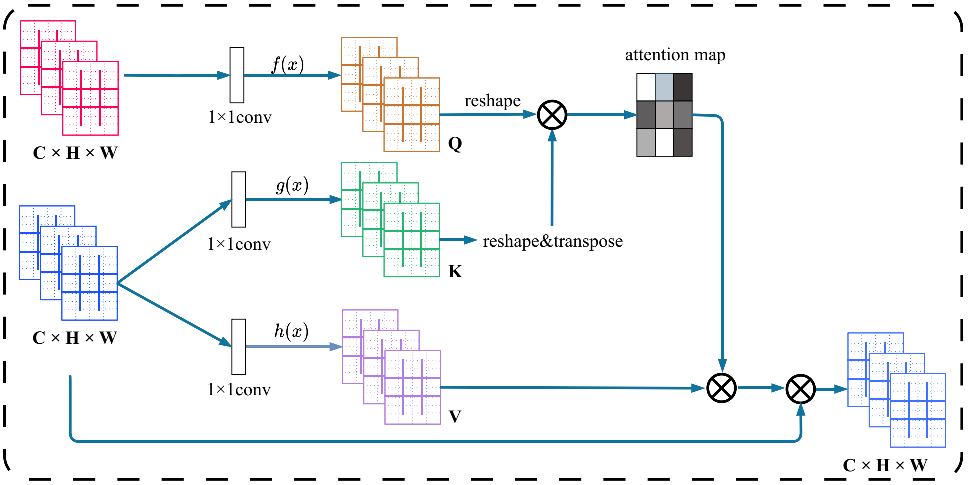


Fig. 3 The structural representation of cross-attention

rates both image and text modalities through two distinct fusion techniques: Language-to-Vision (L2V) and Vision-to-Language (V2L).

In the Vision-to-Language fusion model, we consider image embedding as the primary foundation for information preservation, while using text embedding to provide additional and supporting context. Therefore, the fusion feature of Vision-to-Language is defined as

$$Fusion_{(\mathrm{L2V})} = \text{Cross-Attn}(P_I(i), P_T(t), P_T(t)), \tag{2}$$

where the $Fusion_{(\mathrm{L2V})} = [i^1_{\mathrm{cls}}, i^1_{\mathrm{pat}}]$. In the context of the Language-to-Vision (L2V) fusion model, we utilize visual data to augment the textual modality and enhance the textual modality, similar to the concept of Vision-to-Language fusion. We formulate the fusion feature of Language-to-Vision (L2V) as

$$Fusion_{(\mathrm{V2L})} = \text{Cross-Attn}(P_T(t), P_I(i), P_I(i)), \tag{3}$$

where the $Fusion_{(\mathrm{V2L})} = [t^1_{\mathrm{cls}}, t^1_{\mathrm{tok}}]$.

The HAMMER [20] employs $Fusion_{(\mathrm{V2L})}$ to authenticate both images and textual content, while also applying it in multi-classification tasks. However, both types of tasks require information from two different modalities. Consequently, to enhance the quality of the vision class token in the V2L model, we utilize gated fusion to merge features generated by the Local Patch Attentional Aggregation (LPAA) module with the original V2L features. The LPAA module adheres to the HAMMER framework. The fused output $Fusion^+_{(\mathrm{V2L})}$ is obtained by

$$\alpha = \text{Sigmoid}(W_1 Fusion_{(\mathrm{V2L})} + W_2 \text{LPAA}(i^1_{\mathrm{pat}})), \tag{4}$$

$$\begin{aligned} Fusion^+_{(\mathrm{V2L})} &= \alpha Fusion_{(\mathrm{V2L})} \\ &\quad + (1-\alpha)\text{LPAA}(i^1_{\mathrm{pat}}), \end{aligned} \tag{5}$$

where W_1 and W_2 are learnable parameters, and they are generated by a fully connected network. Following the augmentation of embeddings, the resulting vector $Fusion^+_{(\mathrm{V2L})}$ is deployed for ensuing multitask learning applications.

Local–Global Feature Aggregation

Within the provided framework, the primary aim of the manipulated image bounding box grounding task is to pinpoint areas of image alteration through the identification of local patches that deviate from text embeddings. Consequently, a refined feature, denoted as $Fusion_{(\mathrm{L2V})}$, is achieved by effectively integrating image embeddings with text embeddings using the cross-attention model. The patch tokens i^1_{pat} contained within $Fusion_{(\mathrm{L2V})}$ are constructed by incorporating positional encoding. This indicates an enhanced level of localized spatial data richness within their possession. Therefore, Shao et al. [20] introduced the Local Patch Attentional Aggregation (LPAA) model to enhance the task of grounding bounding boxes in manipulated images. The LPAA model utilizes an attention-based mechanism for synthesizing spatial data from i^1_{pat}. Synthesis is achieved through cross-attention involving the $[AGG]$ token and i^1_{pat}. This aggregation is formulated as

$$\text{LPAA}(i^1_{\mathrm{pat}}) = \text{Cross-Attn}(Agg, i^1_{\mathrm{pat}}, i^1_{\mathrm{pat}}). \tag{6}$$

To optimize the task of grounding bounding boxes in manipulated images, the VLP-GF framework uses a Local–Global Feature model (LGF) to concurrently capture global context and local details. Specifically, the proposed framework significantly improves task accuracy by integrating the global feature i^1_{cls} and harmoniously combining it with the local feature i^1_{pat} to provide a more comprehensive understanding of the context. The GLF model represents an adaptive synthesis that combines both localized and holistic features to generate the resulting feature denoted as i^+_{cls}.

The feature i^{+}_{cls} is obtained as

$$\beta = Sigmoid(Fc(Cat(\text{LPAA}(i^{1}_{\text{pat}}), i_{\text{cls}}))), \tag{7}$$

$$i^{+}_{\text{cls}} = \beta\text{LPAA}(i^{1}_{\text{pat}}) + (1-\beta)i_{\text{cls}}, \tag{8}$$

where Cat, Fc, and $Sigmoid$, respectively, represent the data concatenation, the fully connected operation, and sigmoid operation. The weight β is capable of adaptive adjustment which is generated by a fully connected network. In contrast with feature LPAA(i^{1}_{pat}), the feature i^{+}_{cls} embodies an expanded scope of context-driven information.

Manipulation-Aware Contrastive Learning

The objective of contrastive learning is to align the embeddings of paired image-text instances, while simultaneously inducing divergence in the embeddings of non-paired instances. To enhance the utilization of semantic correlations between images and text, two unimodal encoders employ cross-modal and intra-modal contrastive learning approaches to align their respective embeddings.

Specifically, in the context of a multimodal image-text pair (I, T), this proposed method performs image segmentation on I and then creates a series of image embeddings using a ViT. For the text component, this framework utilizes BERT to extract embeddings from the textual data. Together, the image embeddings and text embeddings form the feature extractor pair (P_I, P_T). In the realm of Cross-Modal Alignment (CMA), the principal goal is to maximize the Mutual Information (MI) shared between paired images and textual content, based on the premise that they convey identical semantic meanings. However, considering the computational challenges associated with directly maximizing Mutual Information (MI) for continuous and high-dimensional variables [54]. Therefore, an alternative approach is employed, which minimizes the InfoNCE loss [41] to derive a lower-bound approximation of MI.

In the proposed framework, cross-modal contrastive learning is implemented using the InfoNCE loss in both the image-to-text and text-to-image directions. Specifically, for the image-to-text context, the InfoNCE loss can be formally defined as follows:

$$\begin{aligned}&\mathcal{L}_{i2t}(I, T^{+}, T^{-})\\&= \mathbb{E}_{p(I,T)}\left[-\log \frac{e^{\left(\text{Sim}\left(P_I(I), P_T(T^{+})\right)/\tau\right)}}{\sum_{k=1}^{K} e^{\left(\text{Sim}\left(P_I(I), P_T(T_k^{-})\right)/\tau\right)}}\right],\end{aligned} \tag{9}$$

where the parameter τ a temperature hyper-parameter. T^{+} is a set of positive text examples that are matched to I. Conversely, T^{-} is a set of negative text examples that exhibit no correspondence with I. Similarly, the InfoNCE loss of text-to-image is mathematically denoted as

$$\begin{aligned}&\mathcal{L}_{t2i}(T, I^{+}, I^{-})\\&= \mathbb{E}_{p(T,I)}\left[-\log \frac{e^{\left(\text{Sim}\left(P_T(T), P_I(I^{+})\right)/\tau\right)}}{\sum_{k=1}^{K} e^{\left(\text{Sim}\left(P_T(T), P_I(I_k^{-})\right)/\tau\right)}}\right].\end{aligned} \tag{10}$$

Therefore, the cross-modal contrastive loss can be formulated as

$$\mathcal{L}_{cross} = \frac{1}{2}[\mathcal{L}_{t2i}(T, I^{+}, I^{-}) + \mathcal{L}_{i2t}(I, T^{+}, T^{-})]. \tag{11}$$

Different from the cross-modal contrastive learning, intra-modal contrastive learning aims to comprehend semantic disparities that distinguish positive and negative instances within a single modality. Concerning the image modality, the contrast loss within its modality can be formally defined as

$$\begin{aligned}&\mathcal{L}_{i2i}(I, I^{+}, I^{-})\\&= \mathbb{E}_{p(I,I)}\left[-\log \frac{e^{\left(\text{Sim}\left(P_I(I), P_I(I^{+})\right)/\tau\right)}}{\sum_{k=1}^{K} e^{\left(\text{Sim}\left(P_I(I), P_I(I_k^{-})\right)/\tau\right)}}\right].\end{aligned} \tag{12}$$

Symmetrically, the contrast loss of text modality can be denoted as

$$\begin{aligned}&\mathcal{L}_{t2t}(T, T^{+}, T^{-})\\&= \mathbb{E}_{p(T,T)}\left[-\log \frac{e^{\left(\text{Sim}\left(P_T(T), P_T(T^{+})\right)/\tau\right)}}{\sum_{k=1}^{K} e^{\left(\text{Sim}\left(P_T(T), P_T(T_k^{-})\right)/\tau\right)}}\right].\end{aligned} \tag{13}$$

The intra-modal contrastive loss is expressed as

$$\mathcal{L}_{intra} = \frac{1}{2}[\mathcal{L}_{i2i}(I, I^{+}, I^{-}) + \mathcal{L}_{t2t}(T, T^{+}, T^{-})]. \tag{14}$$

Both cross-modal and intra-modal contrastive losses play a crucial role in influencing the semantic alignment between images and text. The proposed framework adheres to the conceptual framework of HAMMER while addressing this issue. We contend that both cross-modal and intra-modal contrastive losses are equally significant in addressing this problem. Therefore, the overall contrast loss is formulated as

$$\mathcal{L}_{flmi} = \gamma\mathcal{L}_{cross} + (1-\gamma)\mathcal{L}_{intra}. \tag{15}$$

Multi-Task Learning

The proposed framework addresses not only the task of discerning the authenticity of both images and text but also extends its scope to encompass three additional tasks: multi-classification, bounding box grounding for manipulated images, and token grounding for manipulated text. The presented model demonstrates the ability to authenticate images and detect the specific forgery operations applied to them. Additionally, it is capable of identifying fake facial regions. Similarly, the proposed model is not only employed for text authentication but also for categorizing different types of text forgery and identifying manipulated words within the text. These loss functions of four tasks are obtained through various training supervisory techniques as described below.

Image Bounding Box Grounding To localize the manipulated region within the image, our proposed framework utilizes a three-layer Multilayer Perceptron (MLP) as the Bounding Box (BBox) Detector. In contrast to HAMMER, we employ feature i^{+}_{cls} for the purpose of bounding box detection. This choice is motivated by the fact that i^{+}_{cls} not only captures rich local information but also encompasses specific global characteristics.

Concretely, we input the feature i^{+}_{cls} into the Bounding Box Detector (D_{bbox}) to calculate the Loss for Image Manipulation Grounding. This computation combines the normal $L1$ loss with the generalized Intersection over Union (IoU) loss [55], and it is expressed as

$$\begin{aligned}\mathcal{L}_{IMG} = \mathbb{E}_{(I,T)\sim P}\Big[&\left\|\text{Sigmoid}\left(D_{bbox}\left(i^{+}_{\text{cls}}\right)\right) - y_{\text{box}}\right\| \\ &+ \mathcal{L}_{\text{IoU}}\left(\text{Sigmoid}\left(D_{bbox}\left(i^{+}_{\text{cls}}\right)\right) - y_{\text{box}}\right)\Big],\end{aligned} \tag{16}$$

where the y_{box} represents the ground truth information concerning bounding box detection.

Binary Classification and Multi-Classification Detection The proposed network simultaneously tackles both binary classification for real or fake detection and multiclass classification for counterfeit operation type detection tasks.

The fusion feature $Fusion^{+}_{(\text{V2L})}$ is derived from the gated fusion of the feature $Fusion_{(\text{V2L})}$ and LPAA(i^{1}_{cls}) amalgamating a richer set of textual and image characteristics. Therefore, these classification and detection tasks entail the utilization of the feature denoted as t^{+}_{cls} extracted from the head region of the $Fusion^{+}_{(\text{V2L})}$ model. For the binary classification task, the loss is formulated as

$$\mathcal{L}_{BIC} = \mathbb{E}_{(I,T)\sim P}\mathbb{H}\left(C_b\left(t^{+}_{\text{cls}}\right), y_{\text{b}}\right), \tag{17}$$

where $H(\cdot)$ is defined as a Cross-Entropy function. The C_{b} serves as a binary classifier. For the multi-classification task, the proposed framework exhibits the ability to perform nuanced deepfake analysis. Specifically, it endeavors to ascertain the precise nature of manipulations that include face swap or text swap (FS or TS) and face attribute or text attribute (FA or TA) manipulations. The feature t^{+}_{cls} is fed into a multi-label classifier C_m for the computation of the multi-label classification loss. The multi-classification loss is formulated as

$$\mathcal{L}_{MLC} = \mathbb{E}_{(I,T)\sim P}\mathbb{H}\left(C_m\left(t^{+}_{\text{cls}}\right), y_{\text{m}}\right), \tag{18}$$

where a multi-label classifier C_m is a three-layer MLP. y_{m} is the ground truth label of the multi-classification detection.

Manipulated Text Token Grounding The L2V fusion feature, denoted as $Fusion_{(\text{L2V})}$, not only captures textual context comprehensively but also interacts effectively with image features. The component t^{1}_{tok} within $Fusion_{(\text{L2V})}$ represents comprehensive embeddings for individual text tokens. It aligns with identifying manipulated text tokens. In the task of grounding manipulated text tokens, our framework aims to highlight altered words in the text. This task resembles sequence tagging tasks in the field of Natural Language Processing (NLP). We use a momentum-infused adaptation of the detector module in alignment with prior research. The proposed method uses a Bert-based Token Detector, denoted as D_t, to ground the token t^{1}_{tok} within the D_t to identify manipulated text tokens. Therefore, the overall objective function of this task is expressed as

$$\begin{aligned}&\mathcal{L}_{\text{tok}} = \mathbb{E}_{(I,T)\sim P}[-y_{\text{tok}}\log(D_t(t^{1}_{\text{tok}}))],\\ &\mathcal{L}^{\text{m}}_{\text{tok}} = \mathbb{E}_{(I,T)\sim P}\text{KL}\left[h_t\left(t^{1}_{\text{tok}}\right)\|D^{m}_{t}\left(t^{m}_{\text{tok}}\right)\right],\\ &\mathcal{L}_{\text{tmg}} = (1-\mu)\mathcal{L}_{tok} + \mu\mathcal{L}^{m}_{\text{tok}},\end{aligned} \tag{19}$$

where the t^{m}_{tok} is the momentum version of t^{1}_{tok}. The parameter γ is a balancing factor. And the KL denotes the KL-Divergence algorithm. Hence, the overall loss function for the proposed method is formulated as follows:

$$\begin{aligned}\mathcal{L}_{\text{all}} = &\mathcal{L}_{\text{flmi}} + \lambda_1\mathcal{L}_{\text{img}} + \lambda_2\mathcal{L}_{\text{tmg}}\\ &+ \lambda_3\mathcal{L}_{\text{mdb}} + \lambda_4\mathcal{L}_{\text{mdm}}.\end{aligned} \tag{20}$$

Experiments

Implementation Details

We used the vision transformer [52] for generating image embeddings and BERT [53] for text embeddings, respectively. We implemented a learning rate warm-up phase, during which it gradually increased to 1e-4 over the initial 1000 steps. Subsequently, a decay phase, following a cosine-based schedule, lowered the learning rate to 1e-6. The batch size was set to 64. The architectural framework used for the Binary Classifier, Multi-Label Classifier, Bounding Box Detector, and Token Detector comprises two Multi-Layer Perceptron (MLP) layers, each with output dimensions of 2, 4, 4, and 2. The W_1 and W_2 in Eq. (5) were experimentally set as $W_1 = W_2 = 1$. The weight β in Eq. (8) was set as $\beta = 1$. The weight γ in Eq. (15) was set to 0.5. The hyperparameter in Eq. (20) was set as $\lambda_1 = 0.1$, $\lambda_2 = \lambda_3 = \lambda_4 = 1$. The framework is implemented in PyTorch [56] framework with 2 NVIDIA A100 GPUs.

Datasets

The DGM^4 dataset [20], proposed by Shao et al., functions as a publicly accessible repository meticulously crafted to foster extensive research on machine-engineered media distortions. This dataset employs a range of manipulation techniques in both visual and textual domains. Each sample is enriched with detailed labels that enable the identification and localization of manipulated media. With a compilation of 230,000 news samples, the DGM^4 dataset includes 77,426 unaltered image-text pairs and 152,574 manipulated pairings. The manipulated instances are classified into four categories: 66,722 facial swappings (FS), 56,411 alterations of facial attributes (FA), 43,546 text swap manipulations (TS), and 18,588 text attribute manipulations (TA). In order to generate 32,693 mixed-distortion pairings, researchers combined approximately one-third of the manipulated images with half of the manipulated text.

Evaluation Metrics

To assess the effectiveness of the VLP-GF model, we conduct a comprehensive analysis that encompasses both objective and subjective metrics. In the domain of objective appraisal, this work examines four distinct tasks: binary classification, multi-class discrimination, allocation of bounding boxes for manipulated images, and token grounding for altered text. These tasks are evaluated using a range of metrics. In the context of binary classification, we utilize Area Under the Curve (AUC), Equal Error Rate (EER), and Accuracy (ACC) as performance metrics. The AUC quantifies the overall performance of the classifier through the Receiver Operating Characteristic (ROC) curve, while the EER determines the point of equilibrium between the False Acceptance Rate (FAR) and the False Rejection Rate (FRR). ACC calculates the quotient of precise judgments to total instances. We employ an ensemble of metrics, which encompass the mean Average Precision (mAP), class-specific F1 score (CF1), and the overall F1 score (OF1). The mAP quantifies the average precision across different classes, and it effectively balances precision and recall. The CF1 assesses the algorithm's ability to accurately classify individual instances, whereas the OF1 provides a comprehensive metric across all classes. Image manipulation tasks are assessed by employing metrics including the mean Intersection over Union (IoUmean) as well as IoU at thresholds of 50% and 75%. IoU is a widely used criterion in visual computational tasks, used to quantify the degree of overlap between annotations. For manipulated text grounding, Precision, Recall, and F1-score are the chosen metrics. Notably, higher scores across all metrics except for EER signify improved system efficiency. Conversely, a lower EER signifies superior performance.

Comparison with State-of-the-Art Methods

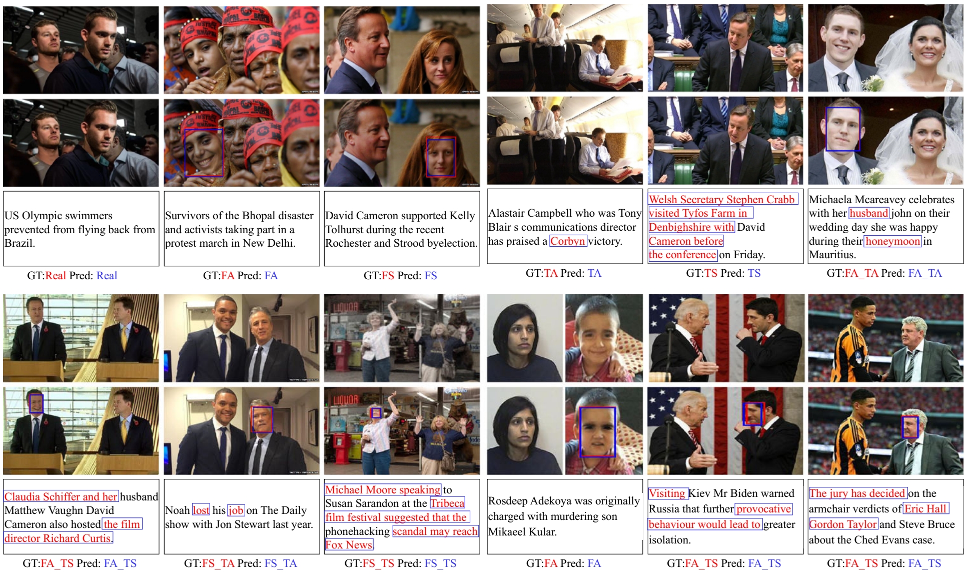
To validate the effectiveness of our proposed framework, we conducted assessments that involved comparisons with multimodal detection models as well as methodologies for deepfake detection and sequence tagging.

Comparison with Multimodal Deepfake Detection Models In order to assess the effectiveness of the proposed methodology in the field of multi-modal deepfake detection, we performed a comparative analysis by comparing it with three state-of-the-art (SOTA) multi-modal learning approaches. These three methods under scrutiny are Contrastive Language-Image Pre-training (CLIP) model [57], Vision-and-Language Transformer (ViLT) [58], and HAMMER [20]. Specifically, the CLIP model epitomizes an efficient and scalable approach to acquiring knowledge guided by natural language. This approach facilitates effortless zero-shot knowledge transfer to a broad range of existing datasets, which demonstrates the remarkable versatility and adaptability of the CLIP. The ViLT offers a remarkably straightforward architecture for vision-and-language models. It employs the transformer module to extract and process visual features, thus eliminating the necessity for a distinct deep visual embedder. The results of the evaluation concerning multimodal deepfake detection models are shown in Table 1. In comparison to the second-best model HAMMER [20], the proposed VLP-GF framework demonstrates

Table 1 Comparison with the SOTA methods. The best results are shown in bold

Categories	Binary Cls			Multi-label Cls			Image grounding			Text grounding		
Methods	**AUC↑**	**ACC↑**	**EER↓**	**mAP↑**	**CF1↑**	**OF1↑**	**IoUmean↑**	**IoU50↑**	**IoU75↑**	**Precision↑**	**Recall↑**	**F1↑**
CLIP [57]	83.22	76.40	24.61	66.00	59.52	62.31	49.51	50.03	38.79	58.12	22.11	32.03
ViLT [58]	85.16	78.38	22.88	72.37	66.14	66.00	59.32	65.18	48.10	66.48	49.88	57.00
HAMMER[a] [20]	92.29	85.48	15.49	85.36	79.20	78.63	76.15	83.24	76.09	75.48	66.49	70.70
VLP-GF (Ours)	**92.84**	**86.13**	**14.45**	**85.65**	**80.02**	**79.07**	**76.73**	**83.89**	**76.24**	**76.42**	**66.80**	**71.29**

[a] Performance evaluation was conducted using the identical platform as ours with the officially released code



US Olympic swimmers prevented from flying back from Brazil.

GT:Real Pred: Real

Survivors of the Bhopal disaster and activists taking part in a protest march in New Delhi.

GT:FA Pred: FA

David Cameron supported Kelly Tolhurst during the recent Rochester and Strood byelection.

GT:FS Pred: FS

Alastair Campbell who was Tony Blair s communications director has praised a Corbyn victory.

GT:TA Pred: TA

Welsh Secretary Stephen Crabb visited Tyfos Farm in Denbighshire with David Cameron before the conference on Friday.

GT:TS Pred: TS

Michaela Mcareavey celebrates with her husband john on their wedding day she was happy during their honeymoon in Mauritius.

GT:FA_TA Pred: FA_TA

Claudia Schiffer and her husband Matthew Vaughn David Cameron also hosted the film director Richard Curtis.

GT:FA_TS Pred: FA_TS

Noah lost his job on The Daily show with Jon Stewart last year.

GT:FS_TA Pred: FS_TA

Michael Moore speaking to Susan Sarandon at the Tribeca film festival suggested that the phonehacking scandal may reach Fox News.

GT:FS_TS Pred: FS_TS

Rosdeep Adekoya was originally charged with murdering son Mikaeel Kular.

GT:FA Pred: FA

Visiting Kiev Mr Biden warned Russia that further provocative behaviour would lead to greater isolation.

GT:FA_TS Pred: FA_TS

The jury has decided on the armchair verdicts of Eric Hall Gordon Taylor and Steve Bruce about the Ched Evans case.

GT:FA_TS Pred: FA_TS

Fig. 4 The visualization pertains to specific instances within the DGM^4 dataset. Elements in red signify the ground truth (GT), and the content encapsulated in the blue box corresponds to the prediction (Pred)

outstanding performance across diverse domains. It demonstrates notable proficiency in tasks related to binary and multi-classification, as well as in aligning manipulated entities within both visual and textual contexts. The proposed framework improves the ACC, CF1, IoUmean, IoU50, Precision, and F1 by 0.76%, 1.03%, 0.76%, 0.78%, 1.25%, and 0.83%, respectively. Moreover, the VLP-GF framework reduces the EER by 6.71%.

In Fig. 4, we illustrate the outcomes of the proposed method within the context of manipulation detection and grounding using visualizations. The red-highlighted text represents the ground truth (GT), while the content enclosed in the blue delineated area represents the prediction (Pred). Similar visualization patterns are observed in image detection. The VLP-GF framework effectively anchors manipulated bounding boxes while precisely identifying the manipulation types relevant to both FA and FS.

Details of Multi-classification Detection The classification performance for each manipulation type is graphically represented in Fig. 5, which utilizes the output of the Multi-Label Classifier. The results offer nuanced insights and reveal that discerning text manipulation proves to be more intricate than that of image modality, with TA manipulation posing the utmost challenge.

Comparison with Unimodal Deepfake Detection Models We conducted a comparative analysis between our proposed method and competing unimodal techniques in two distinct unimodal synthetic data partitions. To ensure a fair evaluation, we incorporated a grounding module into the unimodal model along with the initial binary ground truth and supplied corresponding grounding annotations. In the context of single-modal image forgery detection, we conducted a comparative analysis between our proposed approach and two prominent methodologies, namely TS [59] and MAT [60]. The TS framework comprises three functional modules: a multi-scale high-frequency feature extraction module, a residual-guided spatial attention module, and a cross-modal attention module. These modules work together to effectively utilize high-frequency features. The MAT is a multi-attention deepfake detection architecture. With the assistance of attention maps, this module combines low-level textural and high-level semantic features extracted from images, which improves the detection performance of the system. The comparative outcomes are presented in Table 2. In the context of evaluating unimodal text detection, we aim to compare the proposed methodology with two well-established sequence tagging approaches in Natural Language Processing (NLP), namely BERT [53] and LUKE [61]. The BERT was developed with the aim of pre-training comprehensive bidirectional representations from the unlabeled text by simultaneously conditioning on preceding and succeeding

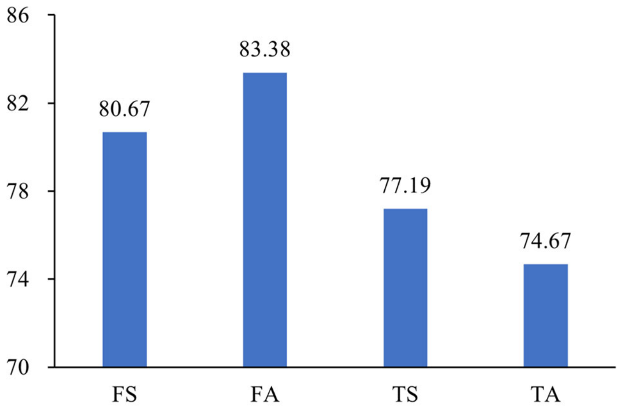


Fig. 5 Performance of multi-classification detection

Table 2 Comparison of image deepfake detection methods. The best results are highlighted in bold

Categories		Binary Cls			Image grounding	
Methods	**AUC↑**	**EER↓**	**ACC↑**	**IoUmean↑**	**IoU50↑**	**IoU75↑**
TS [59]	**91.80**	17.11	82.89	72.85	79.12	74.06
MAT [60]	91.31	17.65	82.36	**72.88**	78.98	**74.70**
VLP-GF (Ours)	91.61	**15.95**	**84.64**	72.85	**80.81**	67.88

Table 3 Comparison of text deepfake detection methods. The best results are highlighted in bold

Categories		Binary Cls			Text grounding	
Methods	**AUC↑**	**EER↓**	**ACC↑**	**Precision↑**	**Recall↑**	**F1↑**
BERT [53]	80.82	28.02	68.98	41.39	63.85	50.23
LUKE [61]	81.39	27.88	76.18	50.52	37.93	43.33
VLP-GF (Ours)	**91.66**	**16.20**	**84.47**	**72.91**	**64.50**	**68.45**

contexts across all layers. The LUKE framework, another pre-trained model, generates contextualized representations for words and entities through the utilization of the transformer architecture. It incorporates an enhanced transformer framework that integrates an innovative self-attention mechanism designed to improve entity sensitivity. The comparative findings are presented in Table 3. According to Tables 2 and 3, the VLP-GF demonstrates a significant performance superiority compared to unimodal methods in the realm of detecting single-modal forgeries. This marked enhancement distinctly signals that our approach, trained on multimodal data, excels in identifying and pinpointing manipulations across various modalities and shows promising effectiveness in detecting and establishing manipulations within each specific modality.

Ablation Studies

Ablation Study of the Key Components In order to assess the effectiveness of the essential elements, ablation experiments were performed using the DGM^4 dataset, as shown in Table 4. In these ablation studies, HAMMER [20] serves as the baseline. The GF denotes the gated fusion model. The utilization of the GF model within the proposed framework facilitates the retrieval of enhanced cross-modal information. The LGF represents the local–global feature aggregation model. Within this module, the envisaged framework adeptly captures both overarching context and nuanced particulars. This enhances the capacity of models for understanding and representing intricate datasets. Drawing upon the information presented in Table 4, the experimental results indicate that the proposed framework significantly enhances the performance of the detection task by effectively utilizing diverse features.

Efficacy of the Local–Global Feature Aggregation Regarding the placement of the manipulated bounding box, we conducted a comparative analysis that utilized both LPAA [20] and the proposed LGF, as shown in Fig. 6. The results presented in Fig. 6 unequivocally demonstrate that LGF outperforms the alternatives across all metrics and it substantiates the efficacy of the LGF model.

Table 4 Ablation study of the key components. The best results are highlighted in bold

Methods			Binary Cls			Multi-label Cls			Image grounding			Text grounding		
Baseline	**GF**	**LGF**	**AUC↑**	**ACC↑**	**EER↓**	**mAP↑**	**CF1↑**	**OF1↑**	**IoUmean↑**	**IoU50↑**	**IoU75↑**	**Precision↑**	**Recall↑**	**F1↑**
✓			92.29	85.48	15.49	85.36	79.20	78.63	76.15	83.24	76.09	75.48	66.49	70.70
✓	✓		**93.05**	86.10	14.59	86.04	**80.29**	**79.32**	75.45	82.53	75.32	74.75	67.56	71.08
✓		✓	92.33	85.66	15.18	84.96	79.33	78.64	76.46	83.57	76.38	75.23	66.38	70.53
✓	✓	✓	92.84	**86.13**	**14.45**	**85.65**	80.02	79.07	**76.73**	**83.89**	**76.24**	**76.42**	**66.80**	**71.29**

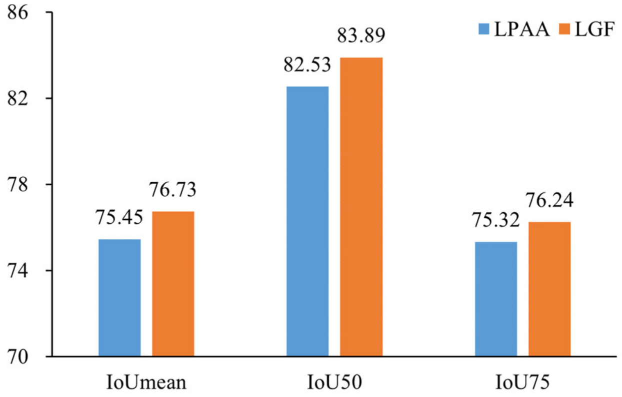


Fig. 6 Efficacy of local–global feature aggregation model (LGF)

Conclusion

In this work, we presented the VLP-GF framework, which is meticulously designed for the detection of intricate multimodal forgeries. In contrast to traditional methods, this framework excels in harnessing semantic information among features. To extract complementary knowledge from cross-modal representations, we employed a gated fusion mechanism that facilitates the seamless integration of cross-modal information by VLP-GF sub-models. Furthermore, we developed an adaptive module for the fusion of local and global features, which allows for the concurrent capture of both global context and local details. This enhancement bolsters the model's comprehension and feature representation capabilities when dealing with intricate data. These innovations significantly enhance the performance of multitask detection. The experimental findings substantiate the superior performance of VLP-GF when compared to SOTA methodologies in both multimodal and unimodal contexts. However, this framework heavily relies on the Transformer architecture and multi-head attention mechanisms, which leads to a significant surge in computational resource requirements, escalation in parameter count, and data redundancy. Therefore, in future work, our primary objective will be to attain network efficiency without compromising system detection performance.

Author Contributions Gao and Zhang wrote the main manuscript text. Li, Zhai, and Jeon conducted simulation. All authors reviewed the manuscript.

Data Availability Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Declarations

Competing Interests The authors declare no competing interests.

References

1. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. Adv Neural Inf Process Syst. 2014;27.
2. Prezja F, Paloneva J, Pölönen I, Niinimäki E, Äyrämö S. Deepfake knee osteoarthritis x-rays from generative adversarial neural networks deceive medical experts and offer augmentation potential to automatic classification. Sci Rep. 2022;12(1):18573.
3. Kim YS, Song HJ, Han JH. A study on the development of deepfake-based deep learning algorithm for the detection of medical data manipulation. Webology. 2022;19(1):4396–409.
4. Kietzmann J, Mills AJ, Plangger K. Deepfakes: perspectives on the future reality of advertising and branding. Int J Advert. 2021;40(3):473–85.
5. Lu H, Chu H. Let the dead talk: how deepfake resurrection narratives influence audience response in prosocial contexts. Comput Hum Behav. 2023;145:107761.
6. Waqas N, Safie SI, Kadir KA, Khan S, Khel MHK. Deepfake image synthesis for data augmentation. IEEE Access. 2022;10:80847–57.
7. Kumar S, Shah N. False information on web and social media: a survey. arXiv:1804.08559 [Preprint]. 2018. Available from: http://arxiv.org/abs/1804.08559.
8. Li Q, Gao M, Zhang G, Zhai W. Defending deepfakes by saliency-aware attack. IEEE Trans Comput Soc Syst. 2023;1–8. https://doi.org/10.1109/TCSS.2023.3271121.
9. Chang X, Wu J, Yang T, Feng G. Deepfake face image detection based on improved VGG convolutional neural network. In: 2020 39th Chinese Control Conference (CCC). IEEE; 2020. pp. 7252–6.
10. Hsu CC, Zhuang YX, Lee CY. Deep fake image detection based on pairwise learning. Appl Sci. 2020;10(1):370.
11. Raza A, Munir K, Almutairi M. A novel deep learning approach for deepfake image detection. Appl Sci. 2022;12(19):9820.
12. Li Q, Gao M, Zhang G, Zhai W, Chen J, Jeon G. Towards multimodal disinformation detection by vision-language knowledge interaction. Inf Fusion. 2023;102037.
13. Guarnera L, Giudice O, Battiato S. Deepfake detection by analyzing convolutional traces. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020. pp. 666–7.
14. Abdelnabi S, Hasan R, Fritz M. Open-domain, content-based, multi-modal fact-checking of out-of-context images via online resources. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022; pp. 14940–9.
15. Aneja S, Bregler C, Nießner M. Cosmos: catching out-of-context misinformation with self-supervised learning. arXiv:2101.06278 [Preprint]. 2021. Available from: http://arxiv.org/abs/2101.06278.
16. Jin Z, Cao J, Guo H, Zhang Y, Luo J. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In: Proceedings of the 25th ACM international conference on Multimedia. 2017. pp. 795–816.
17. Khattar D, Goud JS, Gupta M, Varma V. Mvae: multimodal variational autoencoder for fake news detection. In: The World Wide Web Conference. 2019. pp. 2915–21.
18. Luo G, Darrell T, Rohrbach A. Newsclippings: Automatic generation of out-of-context multimodal media. arXiv:2104.05893 [Preprint]. 2021. Available from: http://arxiv.org/abs/2104.05893.
19. Wang Y, Ma F, Jin Z, Yuan Y, Xun G, Jha K, Su L, Gao J. Eann: event adversarial neural networks for multi-modal fake news detection. In: Proceedings of the 24th ACM Sigkdd International Conference on Knowledge Discovery & Data Mining. 2018. pp. 849–57.

20. Shao R, Wu T, Liu Z. Detecting and grounding multi-modal media manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023. pp. 6904–13.
21. Xuan X, Peng B, Wang W, Dong J. On the generalization of GAN image forensics. In: Chinese Conference on Biometric Recognition. Springer; 2019. pp. 134–41.
22. Zhang Y, Zheng L, Thing VL. Automated face swapping and its detection. In: 2017 IEEE 2nd International Conference on Signal and Image Processing (ICSIP). IEEE; 2017. pp. 15–9.
23. Coccomini DA, Caldelli R, Falchi F, Gennaro C, Amato G. Cross-forgery analysis of vision transformers and CNNs for deepfake image detection. In: Proceedings of the 1st International Workshop on Multimedia AI against Disinformation. 2022. pp. 52–8.
24. Chen HS, Hu S, You S, Kuo CCJ, et al. Defakehop++: an enhanced lightweight deepfake detector. APSIPA Trans Signal Inf Process. 2022;11(2).
25. Patel Y, Tanwar S, Bhattacharya P, Gupta R, Alsuwian T, Davidson IE, Mazibuko TF. An improved dense CNN architecture for deepfake image detection. IEEE Access. 2023;11:22081–95.
26. Arevalo J, Solorio T, Montes-y Gómez M, González FA. Gated multimodal units for information fusion. arXiv:1702.01992 [Preprint]. 2017. Available from: http://arxiv.org/abs/1702.01992.
27. Zhu M, Pan P, Chen W, Yang Y. Dm-gan: dynamic memory generative adversarial networks for text-to-image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019. pp. 5802–10.
28. Zhang Z, Chen K, Wang R, Utiyama M, Sumita E, Li Z, Zhao H. Neural machine translation with universal visual representation. In: International Conference on Learning Representations. 2019.
29. Li B, Lv C, Zhou Z, Zhou T, Xiao T, Ma A, Zhu J. On vision features in multimodal machine translation. arXiv:2203.09173 [Preprint]. 2022. Available from: http://arxiv.org/abs/2203.09173.
30. Zhang Z, Zhang A, Li M, Zhao H, Karypis G, Smola A. Multimodal chain-of-thought reasoning in language models. arXiv:2302.00923 [Preprint]. 2023. Available from: http://arxiv.org/abs/2302.00923.
31. Fang Y, Tan T, Wang Y. Fusion of global and local features for face verification. In: 2002 International Conference on Pattern Recognition, vol. 2. IEEE; 2002. pp. 382–5
32. Eskandari M, Toygar Ö. Fusion of face and iris biometrics using local and global feature extraction methods. SIViP. 2014;8: 995–1006.
33. Zou J, Li W, Chen C, Du Q. Scene classification using local and global features with collaborative representation fusion. Inf Sci. 2016;348:209–26.
34. Zhu Y, Jiang Y. Optimization of face recognition algorithm based on deep learning multi feature fusion driven by big data. Image Vis Comput. 2020;104:104023.
35. Yang M, He D, Fan M, Shi B, Xue X, Li F, Ding E, Huang J. Dolg: single-stage image retrieval with deep orthogonal fusion of local and global features. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021. pp. 11772–81.
36. Zhao X, Yu Y, Ni R, Zhao Y. Exploring complementarity of global and local spatiotemporal information for fake face video detection. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE; 2022. pp. 2884–8.
37. Yang Y, Newsam S. Bag-of-visual-words and spatial extensions for land-use classification. In: Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems. 2010. pp. 270–9.
38. Chen S, Tian Y. Pyramid of spatial relations for scene-level land use classification. IEEE Trans Geosci Remote Sens. 2014;53(4): 1947–57.
39. Gutmann M, Hyvärinen A. Noise-contrastive estimation: a new estimation principle for unnormalized statistical models. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. JMLR Workshop and Conference Proceedings; 2010. pp. 297–304.
40. Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. In: International conference on machine learning. PMLR; 2020. pp. 1597–607.
41. Oord AVD, Li Y, Vinyals O. Representation learning with contrastive predictive coding. arXiv:1807.03748 [Preprint]. 2018. Available from: http://arxiv.org/abs/1807.03748.
42. Antol S, Agrawal A, Lu J, Mitchell M, Batra D, Zitnick CL, Parikh D. Vqa: visual question answering. In: Proceedings of the IEEE international conference on computer vision. 2015. pp. 2425–33.
43. Suhr A, Zhou S, Zhang A, Zhang I, Bai H, Artzi Y. A corpus for reasoning about natural language grounded in photographs. arXiv:1811.00491 [Preprint]. 2018. Available from: http://arxiv.org/abs/1811.00491.
44. Cirik V, Morency LP, Berg-Kirkpatrick T. Visual referring expression recognition: what do systems actually learn? In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 2 (Short Papers). 2018. pp. 781–7.
45. Karpathy A, Joulin A, Fei-Fei LF. Deep fragment embeddings for bidirectional image sentence mapping. Adv Neural Inf Process Syst. 2014;27.
46. Bhargava P. Adaptive transformers for learning multimodal representations. arXiv:2005.07486 [Preprint]. 2020. Available from: http://arxiv.org/abs/2005.07486.
47. Alberti C, Ling J, Collins M, Reitter D. Fusion of detected objects in text for visual question answering. In: 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019. Association for Computational Linguistics; 2019. pp. 2131–40.
48. Tan H, Bansal M. Lxmert: learning cross-modality encoder representations from transformers. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019. pp. 5100–11.
49. Li W, Gao C, Niu G, Xiao X, Liu H, Liu J, Wu H, Wang H. Unimo: towards unified-modal understanding and generation via cross-modal contrastive learning. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (vol. 1: Long Papers). 2021. pp. 2592–607.
50. Li J, Selvaraju RR, Gotmare AD, Joty S, Xiong C, Hoi S. Align before fuse: vision and language representation learning with momentum distillation. In: Advances in Neural Information Processing Systems. 2021.
51. Bugliarello E, Cotterell R, Okazaki N, Elliott D. Multimodal pretraining unmasked: a meta-analysis and a unified framework of vision-and-language BERTs. Trans Assoc Comput Linguist. 2021;9:978–94.
52. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, et al. An image is worth 16x16 words: transformers for image recognition at scale. arXiv:2010.11929 [Preprint]. 2020. Available from: http://arxiv.org/abs/2010.11929.
53. Devlin J, Chang MW, Lee K, Toutanova K. Bert: pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805 [Preprint]. 2018. Available from: http://arxiv.org/abs/1810.04805.
54. Belghazi MI, Baratin A, Rajeswar S, Ozair S, Bengio Y, Courville A, Hjelm RD. Mine: mutual information neural estimation. arXiv:1801.04062 [Preprint]. 2018. Available from: http://arxiv.org/abs/1801.04062.
55. Rezatofighi H, Tsoi N, Gwak J, Sadeghian A, Reid I, Savarese S. Generalized intersection over union: a metric and a loss for bound-

ing box regression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019. pp. 658–66.

56. Paszke A, Gross S, Chintala S, Chanan G, Yang E, DeVito Z, Lin Z, Desmaison A, Antiga L, Lerer A. Automatic differentiation in pytorch. In: NIPS-W. 2017.
57. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, et al. Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. PMLR; 2021. pp. 8748–63.
58. Kim W, Son B, Kim I. Vilt: vision-and-language transformer without convolution or region supervision. In: International Conference on Machine Learning. PMLR; 2021. pp. 5583–94.
59. Luo Y, Zhang Y, Yan J, Liu W. Generalizing face forgery detection with high-frequency features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021. pp. 16317–26.
60. Zhao H, Zhou W, Chen D, Wei T, Zhang W, Yu N. Multi-attentional deepfake detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021. pp. 2185–94.
61. Yamada I, Asai A, Shindo H, Takeda H, Matsumoto Y. Luke: deep contextualized entity representations with entity-aware self-attention. arXiv:2010.01057 [Preprint]. 2020. Available from: http://arxiv.org/abs/2010.01057.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.