Information Fusion 102 (2024) 102037

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

# Towards Multimodal Disinformation Detection by Vision-language Knowledge Interaction

Qilei Li [a,b], Mingliang Gao [a,*], Guisheng Zhang [a], Wenzhe Zhai [a], Jinyong Chen [a], Gwanggil Jeon [c,**]

[a] School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo, 255000, China
[b] School of Electronic Engineering and Computer Science, Queen Mary University of London, London, E1 4NS, United Kingdom
[c] Department of Embedded Systems Engineering, Incheon National University, Incheon, 22012, South Korea

ARTICLE INFO

*Keywords:*
Multimodal disinformation
Deepfake detection
Manipulation grounding
Information aggregation
Cross-modality interaction

ABSTRACT

Disinformation created by artificial neural networks has been widespread along with the recent rapid progress in multimodal learning, and the arising of vision-language foundation models. This disinformation caused a substantial negative impact on society. To solve this pressing issue, numerous efforts have been made to detect either image deepfake or text manipulation. These methods generally focus on a single modality while ignoring the complementary knowledge provided by the counterpart in the other modalities. In this paper, we aim to detect multimodal disinformation and further identify manipulated image areas or text tokens. To this aim, a novel framework termed Vision-language Knowledge Interaction (ViKI) is designed to explore the semantic correlation of an object in different modalities. Specifically, we propose a vision-language embedding regulator to build a joint feature space in which the embeddings of the same semantic are well-aligned. Besides, we perform cross-modality knowledge interaction so as to aggregate uni-modality embedding by adaptively injecting cross-modality information. By exploring vision-language knowledge jointly, ViKI produces accurate predictions for detecting and grounding disinformation. We demonstrate the superiority of ViKI by ablation studies and comparisons with the state-of-the-art methods on large-scale benchmarks. Notably, ViKI outperforms the state-of-the-art works by a rise of 3.71% in precision and 2.14% in CF1 respectively.

## 1. Introduction

Convolutional Neural Networks (CNN) [1] have advanced rapidly over the last decade. Deepfake technology [2], which relies on CNN to generate realistic-looking images, has drawn significant attention. This technology can be found in widespread applications across various domains [3–8], which significantly enhances convenience for individuals. However, the misuse of deepfake is prevalent, and the deepfake model strives to create deceptive disinformation that appears genuine but does not exist in reality. The dissemination of such disinformation poses significant risks, including misguidance, confusion, and potential harm to individuals' reputations and privacy. Moreover, such disinformation can lead to various negative outcomes, such as increased fraud, manipulation of political processes, and the fragmentation of society [9–11].

To mitigate the negative effects of deepfake, researchers are actively engaged in the development of deepfake detection technologies [12–16], with the objective to identify effectively and address the pervasive threat posed by misleading information. Based on the manipulated content, the domain of deepfake detection can be categorized into two primary divisions: unimodal and multimodal detection, which might involve crafted visual representations, textual descriptions, or a combination of both. There are numerous works [17–20] focusing on unimodal deepfake detection, specifically targeting fabricated images or manipulated text. However, it is important to note that disinformation in real-world typically encompass both manipulated images and fabricated text. Consequently, while these investigations have demonstrated impressive performance in detecting unimodal deepfake, they fall short when it comes to effectively tackle multimodal deepfakes. Fig. 1 depicts the difference between unimodal and multimodal deepfake detection techniques. As shown in the figure, while the unimodal approach solely identifies the authenticity of a single modality, the multimodal method not only discerns the truthfulness of modalities but also engages in multi-classification, bounding box localization for manipulated images, and token grounding for manipulated text.

[*] Corresponding author at: School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo, 255000, China.
[**] Corresponding author at: Department of Embedded Systems Engineering, Incheon National University, Incheon, 22012, South Korea.
*E-mail addresses:* mlgao@sdut.edu.cn (M. Gao), ggjeon@gmail.com (G. Jeon).

https://doi.org/10.1016/j.inffus.2023.102037
Received 29 June 2023; Received in revised form 26 August 2023; Accepted 20 September 2023
Available online 23 September 2023

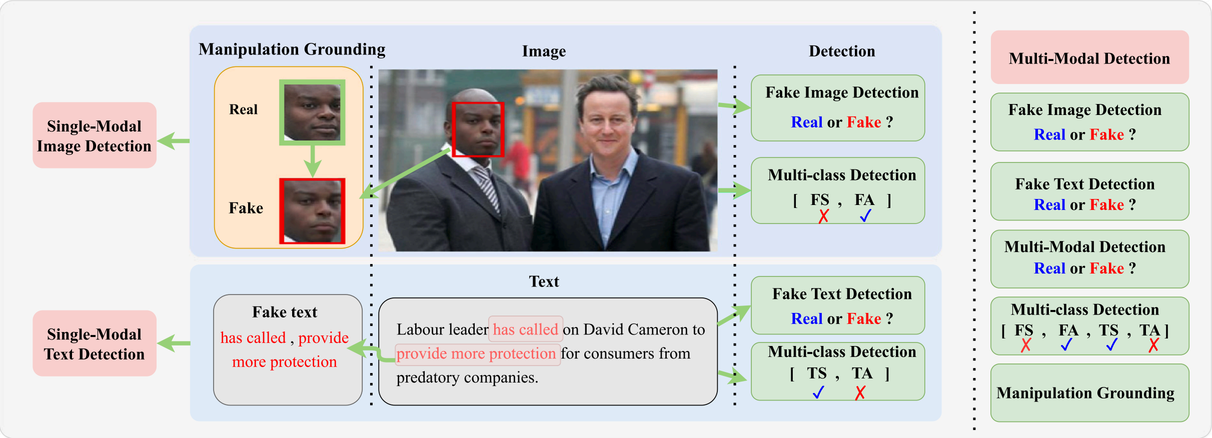


**Fig. 1.** Comparison of single-modal disinformation detection (left) and multi-modal disinformation detection. By jointly utilizing vision and language information, the multimodal-based method can simultaneously detect and grounding the forgery detail. This entitles the learned model to be more practical in solving real-world multimodal disinformation issues.

Compared to purely images or text, the detection of multimodal disinformation poses greater challenges since it typically involves fabrications in both visual and textual domains, where the information is represented in different formats. In order to address this formidable challenge, numerous works have embarked into multimodal deepfake detection [21–26]. Several investigations were dedicated to detecting small-scale instances of multi-modal fake news [23,24,26]. Alternative approaches [21,22,25] concentrate on combating out-of-context misinformation that involves the juxtaposition of authentic images with manipulated text, without any manipulation of the images or text. Recently, HAMMER [27] was proposed to detect and ground multimodal deepfake by extensive training on a large-scale deepfake dataset. Apart from detecting deepfake in a binary classification setting, HAMMER also entails in-depth interpretation in manipulation grounding to locate the region of the manipulated image, or the token of text sequence. Although these models are integrated for multimodal disinformation detection and grounding, their diagrams in exploring the complementary knowledge in different modalities can be further improved by considering the interaction with cross-domain embedding.

In this work, we aim to learn more robust and discriminative vision-language embeddings to perform accurate multimodal misinformation detection and grounding. We achieve this objective by the proposed Vision-language Knowledge Interaction (ViKI) model, which is a unified framework to perform vision-language knowledge interaction and regularize feature learning. Specifically, ViKI binds the paired image and text embeddings into a joint feature space with a novel hypothesis space regulation. The assumption is that abstract embeddings of the same object are tightly clustered in the feature space across modalities, whereas embeddings of different objects are more distant. Consequently, ViKI simultaneously optimizes a metric learning objective and a geometry distance minimization objective to ensure effective alignment of cross-modality features. Besides, considering that reasoning of multimodal disinformation requires discriminative representations from both vision and language modalities, while the balancing of these two components is varied depends on the specific task. To leverage the complementary knowledge in cross-modality embedding, we propose ViKI that adaptively aggregates unimodal embeddings by exploring the cross-modality prompt. With the aggregated embeddings, the multitask learning-heads can produce accurate predictions which can simultaneously detect the fine-grained type of multimodal deepfake, the area of deepfake region in an image, and the location of manipulated text tokens. We summarize the contribution of this paper from the following three aspects:

- We design a unified framework that can produce discriminative and informative representations for multimodal disinformation detection. It can achieve superior performance compared with the latest state-of-the-art (SOTA) methods and is promising in real-world applications.
- We propose to align the multimodal embedding by jointly optimizing a metric learning objective, as well as a geometry distance minimization objection, which can enable the embeddings for the same instance to be closer in the hypothesis space.
- We propose the cross-modality knowledge interaction mechanism to adaptively aggregate unimodal embedding with cross-modality knowledge. This enables the formulation of a discriminative fused representation which facilitates the learning of multitasking objectives.

The remainder of the paper is organized as follows. In Section 2, the related literature is reviewed. In Section 3, we detail the proposed ViKI model. In Section 4, comprehensive comparison and ablation studies are conducted to evaluate the proposed method. The conclusion is drawn in Section 5.

## 2. Related work

### 2.1. Deepfake detection

With the rise of a generative model and the potential misuses, the emergence of deepfake detection techniques has been imperative. Deepfake detection refers to the process of identifying and distinguishing between genuine and manipulated media content created by deepfake techniques. It is categorized into two primary divisions: unimodal deepfake detection and multimodal deepfake detection.

**Unimodal Deepfake Detection** Numerous research efforts [28–30] have been dedicated to developing robust and efficient techniques for detecting and mitigating the spread of manipulated media content. Nguyen et al. [28] proposed to utilize capsule networks for detecting different types of spoofing attacks. The study expanded the application of capsule networks beyond their initial purpose and applied them to solve inverse graphics problems. Xue et al. [30] introduced GLFNet, a network that fuses global and local facial features to detect forged images with perturbations. The method leverages physiological characteristics and deep learning to capture forgery traces and improve robustness.

**Multimodal Deepfake Detection** Multimodal deepfake detection [21–23,26] has become popular and important research topic due to the increasing concern over the deceptive use of multimodal deep learning techniques. Abdelnabi et al. [21] proposed a Consistency-Checking

Network (CCN) to enhance the accuracy and reliability of the analysis by using comprehensive consistency checking. Another multimodal deepfake detection model HAMMER [27] was employed by Shao et al. which can capture fine-grained interactions between different modalities. The HAMMER incorporates shallow manipulation reasoning through contrastive learning between uni-modal encoders and deep manipulation reasoning through modality-aware cross-attention using a multi-modal aggregator.

In this study, our attention is directed toward the realm of multimodal deepfake due to its substantial relevance in societal privacy concerns. Our approach involves the concurrent mitigation of face manipulation within the image modality and the dissemination of textual disinformation, all within a cohesive framework.

### 2.2. *Large-scale foundation models*

In the realm of deep learning and artificial intelligence, researchers have observed the ascendance of Large-Scale Foundation Models. Notably, the focus has shifted towards the exploration of vision-language pre-training models and the transformer paradigm.

**Vision-Language Pre-training Models** To alleviate the need for training models from scratch, vision-language pre-training methods have gained considerable attention and research focus in multimodal tasks [31–35]. An align the image and text representations Before Fusing (ALBEF) [31] method was proposed by Li et al.. The ALBEF model improves vision and language learning without requiring extra bounding box annotations. Tiong et al. [32] proposed a Plug-and-Play VQA (PNP-VQA) model for zero-shot VQA without training. To ensure accurate question answering and generate comprehensive image captions, it leverages network interpretation as an interface between pre-trained language models and vision-language models.

**Transformer** In the past years, with the emergence of the Transformer model capturing widespread attention, deep learning has witnessed remarkable achievements across various domains. Initially proposed by Vaswani et al. [36], the Transformer model has showcased an exceptional performance in the domain of machine translation tasks. Radford et al. [37] proposed Generative Pre-trained Transformer (GPT), a text generation model based on the Transformer architecture, aimed at generating natural language text. For instance, Vision Transformer (ViT) [38], a visual attention model based on the Transformer framework, has gained wide recognition. The ViT framework challenges the conventional image recognition tasks and replaces traditional CNNs.

In this work, we use pre-trained vision and language models as the feature extractor to obtain the representative embedding from different modalities. Inspired by the self-attention and cross-attention mechanism in Fig. 2, we perform cross-modality information aggregation to fully explore the complementary knowledge, which helps further detection and grounding on the disinformation.

### 2.3. *Feature space alignment*

Feature space alignment refers to the process of aligning or mapping the feature spaces of different modalities or domains. This alignment aims to facilitate cross-modal or cross-domain tasks by ensuring that the learned representations capture similar or compatible information across different modalities or domains. The advancement of feature space alignment has manifested itself in numerous avenues of research, spanning object recognition [39,40], face antispoofing [41,42], and medical imaging analysis [43,44]. Sun et al. [45] used a domain alignment network to align feature representations from the source and target domains. The method reduces the distribution discrepancy and improves the target classification accuracy.

Recently, a significant amount of scholarly attention has been directed towards ameliorating distribution disparities among distinct domains, and this is another research direction closely related to feature space alignment. Long et al. [46] employed Maximum Mean Discrepancy (MMD) to measure distribution disparity between different domains. Their approach aims to align feature spaces by minimizing the distance between feature distributions from distinct domains. Li et al. [39] sought to minimize the MMD distance between the distributions of the source domain based on the hidden-layer features. Additionally, they employed adversarial learning to ensure similarity between the feature distributions and a predetermined prior distribution. Inspired by the aforementioned methods, we further build a geometry distance minimization objective to supervise the learning of joint feature space, in which the embedding with the same semantics is aligned with a short distance.

### 2.4. *Contrastive learning*

Contrastive learning is a self-supervised representation learning framework. It refines feature representation by augmenting positive pair similarity and diminishing that of negative pairs in a concealed space. It is based on information theory and has been extensively studied in various domains. One of its initial formulations is the InfoMax principle [47] was put forward by Linsker et al.. The InfoMax principle seeks to maximize the mutual information between input and output representations. Building upon this notion, Hadsell et al. [48] extended the framework by introducing the Siamese network architecture. Siamese networks facilitate representation learning by minimizing the contrastive loss between pairs of samples. In the early stages of contrastive learning [48,49], although these techniques have the potential to improve the training performance of the system, they can also introduce the challenge of training instability. In order to tackle the challenge of slow training in contrastive learning, Sohn et al. [50] introduced the concept of multi-class N-pair loss. This approach fundamentally involves aggregating multiple data instances for simultaneous comparison, thereby achieving error averaging and enhancing training stability. A notable advancement in contrastive learning is the introduction of the Info Normalized Cross Entropy (InfoNCE) loss by Oord et al. [51]. The InfoNCE loss is designed to maximize agreement between positive pairs while minimizing agreement between negative pairs. This loss function has gained widespread adoption in various contrastive learning approaches, such as SimCLR [52] and MoCo [53]. The recent success of contrastive learning can be attributed to the effectiveness of large-scale self-supervised training. By leveraging the structural characteristics of unlabeled data, contrastive learning has demonstrated superior performance across diverse domains, including image recognition [53], natural language processing [54], and speech recognition [55]. In this work, besides employing contrastive learning for feature alignment, we propose a geometry regularizer using MMD distance to align the vision and language embedding into a joint feature space, so as to utilize the synergy for accurate disinformation detection.

## 3. Proposed method

### 3.1. *Problem formulation*

In this work, the primary objective is to confront multimodal misinformation through the lens of a unified framework. The proposed framework endeavors to extract and aggregate knowledge from the realms of vision and language. To be more specific, the proposed model grapples with the challenge of discerning truth from falsity within a complex landscape. Specifically, given an image depicting a person in a specific scenario and a corresponding textual description of that scenario, our model aims to achieve various objectives, including: (1) Uncovering the presence of deepfake alterations, should they be present within the input image. (2) Pinpointing the precise regions within the image that have undergone modification. (3) Detecting any form of textual manipulation or tampering within the description. (4) Identifying the manipulation of particular words, thereby illuminating the

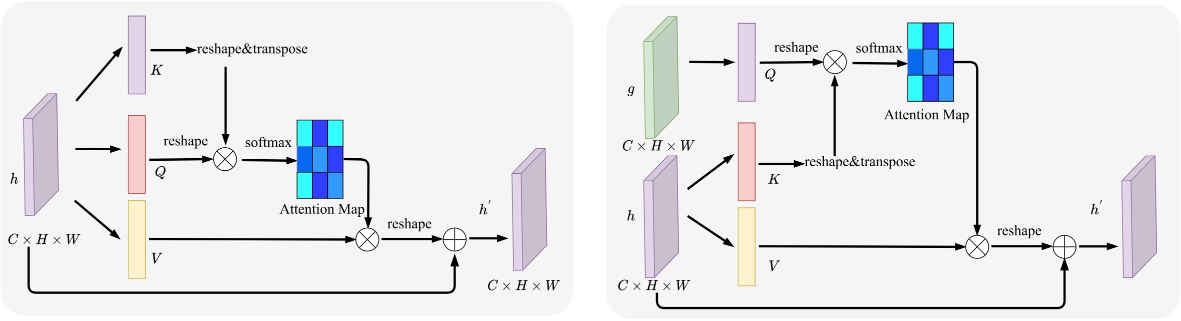
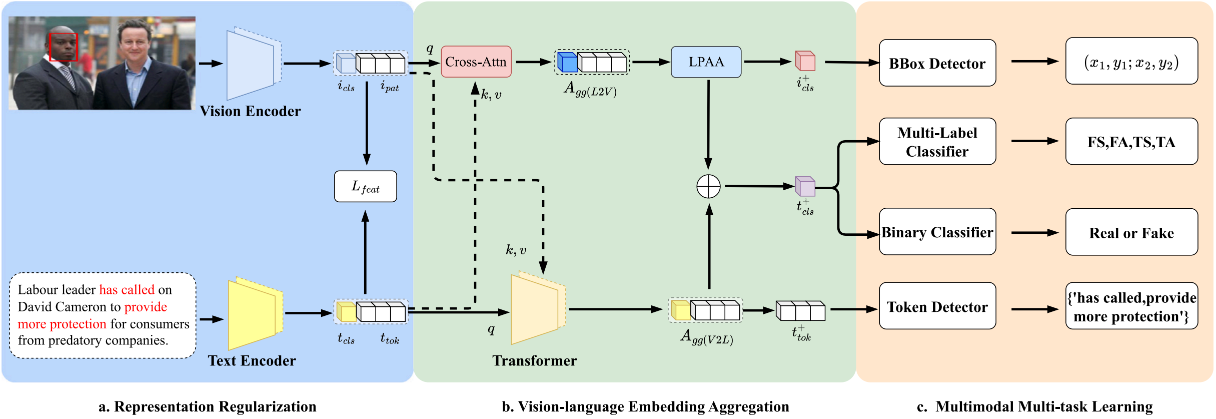


**Fig. 2.** Comparison on self-attention (left) and cross-attention (right). In self-attention, the query (Q), key (K), and value (V) all come from the same feature activation. However, in cross-attention, the query (Q) comes from a different source than the key (K) and value (V).



a. Representation Regularization

b. Vision-language Embedding Aggregation

c. Multimodal Multi-task Learning

**Fig. 3.** The overall framework of ViKI. It is composed of three key components. (a) A vision-language feature extractor utilizing representation regularization for a concise feature domain. (b) An integrator merging features from diverse modalities. (c) Multifaceted learning modules for detailed detection and grounding.

semantic alterations. It is non-trivial to derive a unified framework that is capable of learning discriminative representations from the realms of vision and language modalities to perform joint deepfake detection and grounding. Multimodal misinformation detection inherently presents a greater challenge, as it involves the representation of information across diverse modalities. Moreover, the integration of such multimodal information is further compounded by the noisy artifacts introduced by deepfake practices. It is non-trivial to design a unified framework that possesses the remarkable ability to discern discriminative representations from the realms of vision and language is no trivial feat. Additionally, the subsequent fusion of these representations becomes paramount, as it enables the holistic consideration of complementary information for the purposes of misinformation detection and grounding.

### 3.2. Methodology overview

To learn multimodal representation and leverage their synergy for fine-grained misinformation detection, we propose a unified model, Vision-language Knowledge Interaction (ViKI), which learns a compact feature space to facilitate multimodal representation fusion. The overall framework of ViKI is depicted in Fig. 3. It consists of three hierarchical modules. (1) A vision-language feature extractor which employs representation regularization to construct a compact feature space for vision and language knowledge. (2) The vision-language knowledge aggregator, designed to fuse features extracted from distinct modalities. It goes beyond mere combination and interactively concentrates on capturing discriminative knowledge from the joint representation of the two modalities. (3) Multitask learning heads that are corresponding to each learning objective to achieve fine-grained detection and grounding. The transformer module is constructed from 12 layers of attention blocks. The first six layers are dedicated to fine-grained text feature extraction by utilizing self-attention mechanisms, whereas the subsequent six layers focus on the amalgamation of text and image attributes by cross-attention mechanisms. We employ the lightweight Multi-Layer Perceptrons (MLPs) to design the bounding box detector, multi-label classifier, and binary classifier. All of them are designed as 3-layer, and the dimension of the final layers is determined by the specific task accordingly.

Let us consider a paired input denoted as $X = [I, T]$, where $I$ represents an image and $T$ represents its textual description. In order to extract the semantic information embedded within this input, we leverage the capabilities of the corresponding image encode $E_I$ and text encoders $E_T$. These encoders play a crucial role in transforming the input into uni-modal representations. These representations serve as compact yet expressive depictions of the underlying semantic content. Despite the fact that multimodal information exists in distinct modalities, we impose regularization to encourage their proximity in the feature space. This regularization is based on the premise that the multimodal representations share the same high-level semantic representation. To fully explore the synergy between these representations, we further design an objective-oriented multimodal fusion strategy. This strategy involves adaptively injecting knowledge from one modality into another, guided by the specific task at hand and the corresponding features. The architecture of this framework model is crafted to enable end-to-end training and optimization using conventional learning optimizers.

### 3.3. Vision-language representation regularization

Given a multimodal pair $[I, T]$, we extract the unimodality representation by the corresponding feature extractor pair $[E_I, E_T]$. To this end, benefiting from the latest development in self-attention and cross-attention mechanisms, we use a vision transformer (ViT) [38] as the feature extractor for the visual modality, and BERT [54] as the feature extractor for the text modality.

Specifically, the image is firstly cropped into non-overlapping patches in the size of 16 × 16, which are further flatted and projected into high-dimension feature embeddings, with a class token attached at the first dimension. The embeddings are denoted as $E_I(I) = [i_{\text{cls}}, i_{\text{pat}}]$, where $x_{\text{cls}}$ is the class token which has the same dimension as the patch projection. For the text modality, the input description is tokenized and encoded by BERT to get the embedding $[t_{\text{cls}}, t_{\text{tok}}]$. Representations from these two modalities are originally in distinct feature space although they carry the same semantic information when representing the same content. In order to mine the correlation between the cross-modality representation and align the feature distance in the hypnosis space, recent work [27] has proposed to use contrastive learning to pull the positive multimodal representation while pushing away the negative one. However, we argue that the intrinsic noisy caused by the disinformation has already perturbed the semantic information, causing contrastive learning to be sub-optimized. To solve this problem, in this work, we further propose to regulate the multimodal feature distance with a geometry alignment regularization for the positive pairs when there is no misinformation existing. Specifically, we first calculate the contrastive learning objective by considering the correlation among image feature $E_I(I)$ and text feature $E_T(T)$ as

$$\mathcal{L}_{v2t}(I,T^+,T^-) = \mathbb{E}_{p(I,T)}\left[-\log\frac{\exp\left(\text{Sim}\left(E_I(I), E_T(T^+)\right)/\tau\right)}{\sum_{k=1}^{K}\exp\left(\text{Sim}\left(E_I(I), E_I(T_k^-)\right)/\tau\right)}\right], \tag{1}$$

where $\text{Sim}(\cdot)$ calculates the similarity by inner product, and $\tau$ is the temperature factor. $T^+$ is the text sample with the same semantic meaning as $I$, while $T^-$ is the negative sample that carries different information. Similarly, the text-to-vision contrastive loss $\mathcal{L}_{t2i}(T, I^+, I^-)$, the uni-modality contrastive losses $\mathcal{L}_{t2t}(T, T^+, T^-)$ and $\mathcal{L}_{i2i}(I, I^+, I^-)$, are jointly considered to further regulate the formulation of the joint feature space. Therefore, the overall contrastive learning objective is formulated as

$$\begin{aligned}\mathcal{L}_{\text{con}}(I,T) &= \mathcal{L}_{v2t}(I,T^+,T^-) + \mathcal{L}_{t2i}(T,I^+,I^-)\\ &+ \mathcal{L}_{t2t}(T,T^+,T^-) + \mathcal{L}_{i2i}(I,I^+,I^-)\end{aligned} \tag{2}$$

In practices, we follow the existing works [27,31,53] to use the token representations $i_{\text{cls}}$ and $t_{\text{cls}}$ for calculating the similarity in Eq. (2). To alleviate the interface due to the noise in $t_{\text{cls}}$, we use its counterpart $t^m_{\text{cls}}$ produced by the momentum text encoder [53] to calculate the similarity by the inner product as $[i^{\text{T}}_{\text{cls}} \cdot t^m_{\text{cls}}]$.

To further ensure the coherent and consistency of multimodal features, a maximum mean discrepancy (MMD) regularization is further applied to minimize the distance between the image feature and the text feature. Specifically, given the feature pairs from an original image–text pair, which is not manipulated in either modality, we calculate the MMD distance of them and use as a learning objective as

$$\mathcal{L}_{\text{dist}}(I,T) = \text{dist}(i_{\text{cls}}, t_{\text{cls}}). \tag{3}$$

By considering jointly the contrastive objective and the distance regularization, the final loss function can be formulated as

$$\mathcal{L}_{\text{feat}} = \eta_1\mathcal{L}_{\text{con}} + \eta_2\mathcal{L}_{\text{dist}}. \tag{4}$$

### 3.4. Vision-language feature aggregation

Once the vision embedding $E_I(I)$ and language feature $E_T(T)$ are obtained, the complementary cross-modality information will be explored for feature discrimination augmentation by aggregating the knowledge from one modality to another. We achieve this objective by employing cross-attention layers to interactive embeddings in different modalities by modeling each of them as query (Q), key (K), and value (V). Specifically, by considering Q as the source information, the information from K-V can be aggregated into Q by

$$\text{Cross-Attn}(Q,K,V) = \text{Softmax}(K^TQ/\sqrt{D})V, \tag{5}$$

where $\text{Softmax}(K^TQ)$ will form an attention map to guide the knowledge flow from $V$ to $Q$. In our specific context, for image deepfake grounding, we consider image embedding as the primary source of knowledge, while the text embedding will be used as the auxiliary component to provide additional and complementary information. Therefore, we formulate the language-to-vision (L2V) aggregation as

$$Agg_{(\text{L2V})} = \text{Cross-Attn}(E_I(I), E_T(T), E_T(T)). \tag{6}$$

Similarly, for text manipulation grounding, we enhance the language embedding with the help of vision embedding by performing vision-to-language (V2L) aggregation as

$$Agg_{(\text{V2L})} = \text{Cross-Attn}(E_T(T), E_I(I), E_I(I)). \tag{7}$$

To further refine the vision class token $I_{\text{cls}}$, we follow HAMMER [27] to employ an LPAA module to attentively learn more important knowledge for image grounding. For the multimodal detection tasks, we argue that solely using the information in $Agg_{(\text{V2L})}$ is inadequate as the information within it is biased to language modality, while we experimentally demonstrated that vision modality contributes more to these tasks. To solve this problem, we designed an adaptive fusion mechanism to inject the vision information to $Agg_{(\text{V2L})}$ by

$$\begin{aligned}Agg^+_{(\text{L2V})} &= Agg_{(\text{L2V})} + \alpha E_I(I),\\ Agg^+_{(\text{V2L})} &= Agg_{(\text{V2L})} + \beta\text{LPAA}(i_{\text{cls}}),\end{aligned} \tag{8}$$

where $Agg^+_{(\cdot)}$ is the refined embedding. It should be noted that for $Agg^+_{(\text{V2L})}$, we update the class token in $E_I(I)$ with $\text{LPAA}(i_{\text{cls}})$. After the embedding enhancement, we have $Agg^+_{(\text{L2V})} = [i^+_{\text{cls}}, i^+_{\text{pat}}]$, and $Agg^+_{(\text{V2L})} = [t^+_{\text{cls}}, t^+_{\text{tok}}]$, which are used for subsequent multitask learning.

### 3.5. Multimodal multi-task learning

Our framework is versatile at detecting and grounding multimodal disinformation from the following four perspectives, namely image deepfake grounding, text manipulated grounding, multimodal deepfake detection, and deepfake fine-grained classification. These objectives are achieved by different training supervisions as following.

**Image Deepfake Grounding** To grounding the area of the image which has been manipulated, a bounding box detection head is designed which is composed of a three-layer MLP. We denote the bounding detection head as $h_i(\cdot)$ which takes the augmented image embedding $Agg^+_{(\text{V2L})}$ as the input and reproduce the corresponding bounding box prediction, which is supervised by the grounding objective as

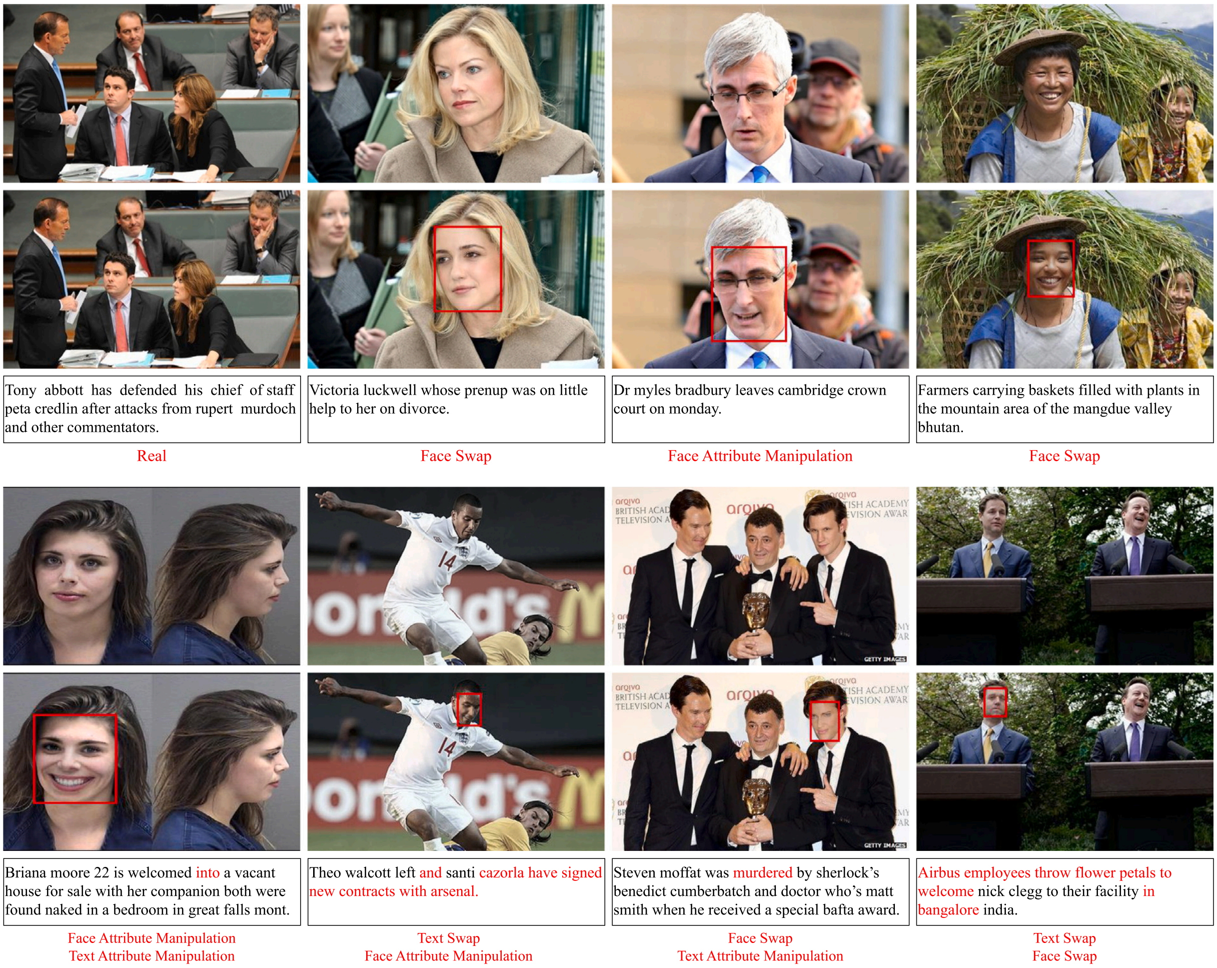
$$\begin{aligned}\hat{y_{\text{box}}} &= \text{Sigmoid}(h_i(i^+_{cls})),\\ \mathcal{L}_{\text{img}} &= \mathbb{E}_{(I,T)\sim P}\left[\|\hat{y_{\text{box}}} - \ y_{\text{box}}\| + \mathcal{L}_{\text{IoU}}(\hat{y_{\text{box}}} - y_{\text{box}})\right],\end{aligned} \tag{9}$$

where $y_{\text{box}}$ is the ground truth for bounding box detection.

**Text Manipulation Grounding** To grounding the manipulated text token, the token representation $t^+_{\text{tok}}$ is fed into a token manipulation detector to detect the manipulation location. Similar to existing works, we also employ a momentum version of the aggregator and detector, which is denoted as $h^m_t$. The overall objective function is formulated as

$$\begin{aligned}\mathcal{L}_{\text{tok}} &= \mathbb{E}_{(I,T)\sim P}[-y_{\text{tok}}\log(h_t(t^+_{\text{tok}}))],\\ \mathcal{L}^{\text{m}}_{\text{tok}} &= \mathbb{E}_{(I,T)\sim P}\text{KL}\left[h_t\left(t^+_{\text{tok}}\right) \parallel h^m_t\left(t^m_{\text{tok}}\right)\right],\\ \mathcal{L}_{\text{tmg}} &= (1-\gamma)\mathcal{L}_{tok} + \gamma\mathcal{L}^m_{\text{tok}},\end{aligned} \tag{10}$$

where $h_t$ is a token detector composed of a three-layer MLP, $h^m_t$ is the image embedding extracted by the momentum-updated feature extractor, and $\gamma$ is a balancing factor.


Tony abbott has defended his chief of staff peta credlin after attacks from rupert murdoch and other commentators.

Real

Victoria luckwell whose prenup was on little help to her on divorce.

Face Swap

Dr myles bradbury leaves cambridge crown court on monday.

Face Attribute Manipulation

Farmers carrying baskets filled with plants in the mountain area of the mangdue valley bhutan.

Face Swap

Briana moore 22 is welcomed into a vacant house for sale with her companion both were found naked in a bedroom in great falls mont.

Face Attribute Manipulation
Text Attribute Manipulation

Theo walcott left and santi cazorla have signed new contracts with arsenal.

Text Swap
Face Attribute Manipulation

Steven moffat was murdered by sherlock's benedict cumberbatch and doctor who's matt smith when he received a special bafta award.

Face Swap
Text Attribute Manipulation

Airbus employees throw flower petals to welcome nick clegg to their facility in bangalore india.

Text Swap
Face Swap

**Fig. 4.** Visualization on a few samples from the DGM$^4$ [27] dataset, which includes various types of manipulations on vision or language modality. The first row is the untouched raw images, the second row is the manipulated image, the third row is the corresponding description, and the fourth row is the label for the manipulation.

**Multimodal Deepfake Detection** Once the high-level representation is fused by the image and text embedded is obtained, it can be used to reason if any deepfake, regardless of the modality, has occurred on the given vision-language pair. To achieve this end, we take the $t^{+}_{\text{cls}}$ token in the $Agg^{+}_{\text{V2L}}$ as the input to the tailored detection network to perform binary deepfake reasoning. We design the detection network $h_{\text{bd}}$ as a light-weight MLP with three FC layers, which outputs the softmax prediction to indicate the probability for the deepfake. The detection network is supervised by the conventional binary classification loss as

$$\mathcal{L}_{\text{mdb}} = \mathbb{E}_{(I,T)\sim P} - [y_{\text{b}}\log(h_{\text{bd}}\left(t^{+}_{\text{tok}}\right)) + (1-y_{\text{b}})\log(1-h_{\text{bd}}\left(t^{+}_{\text{tok}}\right)))]. \tag{11}$$

**Fine-grained Deepfake Justification** Apart from the above binary deepfake detection, ViKI is capable of fine-grained deepfake justification which aims to determine the specific type of manipulations, such as face/text swap (FS/TS), or face/text attribute (FA/TA) manipulation. This is essentially a four-way classification task achieved by an MLP-based classifier with the following learning objective:

$$\mathcal{L}_{\text{mdm}} = \mathbb{E}_{(I,T)\sim P} - [\sum_{i=1}^{4} y^{(i)}_{\text{m}}\log(h^{(i)}_{\text{md}}\left(t^{+}_{\text{tok}}\right))], \tag{12}$$

where $y^{(i)}_{m}$ is the one-hot encoded label for the $i$th type of manipulation, while $h^{(i)}_{\text{md}}\left(t^{+}_{\text{tok}}\right)$ is the corresponding prediction by the multi-class deepfake detector $h_{\text{md}}$. By combining the proposed embedding regularization $\mathcal{L}_{\text{feat}}$ and the multitask learning objectives, the final loss function is formulated as:

$$\mathcal{L}_{\text{train}} = \mathcal{L}_{\text{feat}} + \alpha_1\mathcal{L}_{\text{img}} + \alpha_2\mathcal{L}_{\text{tmg}} + \alpha_3\mathcal{L}_{\text{mdb}} + \alpha_4\mathcal{L}_{\text{mdm}}. \tag{13}$$

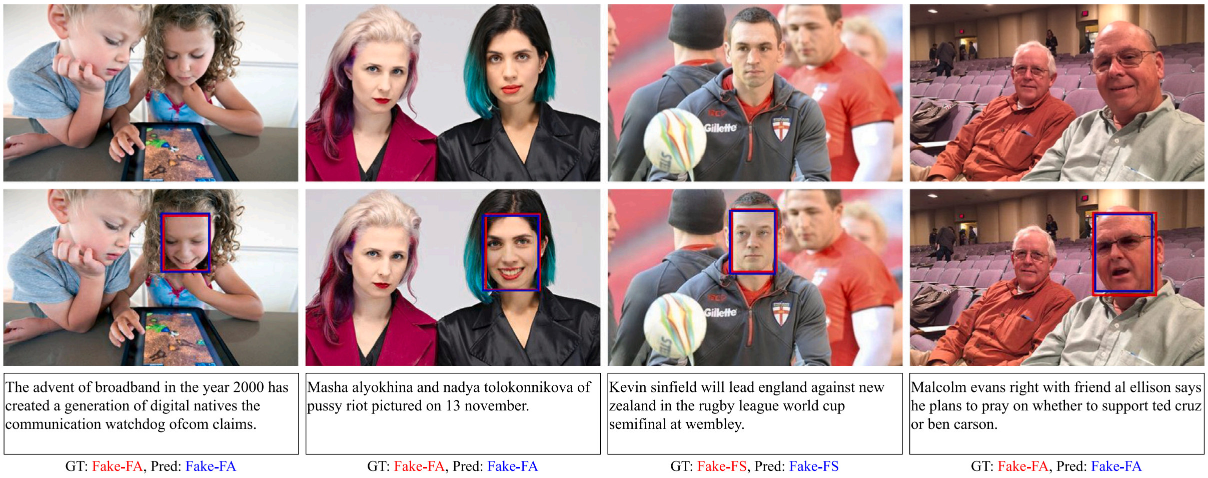
## 4. Experimental results and analysis

### *4.1. Implementation details*

We employed the vision transformer (ViT-B/16) [38] as the image encoder and BERT [54] as the text encoder. All the images were resized to 256 × 256. The batch size and training epoch were set to 32 and 50, respectively. We used a cosine annealing strategy to warm up the learning rate to $1\times10^{-4}$ in the first 1000 iterations, and then decayed to $1\times10^{-6}$ in the remaining steps. The AdamW optimizer [56] was adopted for updating the parameters and the weight decay ratio was set to 0.02. The balancing factors in Eq. (4) were experimentally set as: $\eta_1 = 0.1, \eta_2 = 10$. The hyperparameter in Eq. (13) were set as: $\alpha_1 = 0.1, \alpha_2 = \alpha_3 = \alpha_4 = 1$. For the aggregation objective in Eq. (8), the learnable factors $\alpha$ are initialized to 1.0 and $\beta$ was initialized to 10. All experiments were conducted on 4 NVIDIA A100 GPUs with the PyTorch [57] framework.

### *4.2. Benchmark details*

The DGM$^4$ dataset [27] comprises 230,000 news samples, including 77,426 pristine image–text pairs and 152,574 manipulated pairs. The manipulated pairs consist of 66,722 face swap manipulations, 56,411 face attribute manipulations, 43,546 text swap manipulations, and



The advent of broadband in the year 2000 has created a generation of digital natives the communication watchdog ofcom claims.

GT: Fake-FA, Pred: Fake-FA

Masha alyokhina and nadya tolokonnikova of pussy riot pictured on 13 november.

GT: Fake-FA, Pred: Fake-FA

Kevin sinfield will lead england against new zealand in the rugby league world cup semifinal at wembley.

GT: Fake-FS, Pred: Fake-FS

Malcolm evans right with friend al ellison says he plans to pray on whether to support ted cruz or ben carson.

GT: Fake-FA, Pred: Fake-FA

**Fig. 5.** Visualization for a partial of samples in DGM$^4$ datasets detected by the proposed method.

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

| Categories | Binary Cls | | | Multi-Label Cls | | | Image grounding | | | Text grounding | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Methods | AUC | ACC | EER↓ | mAP | CF1 | OF1 | IoUmean | IoU50 | IoU75 | Precision | Recall | F1 |
| CLIP [58] | 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 [59] | 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 [27] | 93.19 | 86.39 | 14.10 | 86.22 | 79.37 | **80.37** | 76.45 | 83.75 | **76.06** | 75.01 | **68.02** | 71.35 |
| ViKI (Ours) | **93.51** | **86.67** | **13.87** | **86.58** | **81.07** | 80.10 | **76.51** | **83.95** | 75.77 | **77.79** | 66.06 | **72.44** |

**Table 2**
Comparison of image Deepfake detection methods.

| Categories | Binary Cls | | | Image grounding | | |
|---|---|---|---|---|---|---|
| Methods | AUC | EER↓ | ACC | IoUmean | IoU50 | IoU75 |
| TS [60] | 91.80 | 17.11 | 82.89 | 72.85 | 79.12 | 74.06 |
| MAT [18] | 91.31 | 17.65 | 82.36 | 72.88 | 78.98 | **74.70** |
| ViKI (Ours) | **91.85** | **15.92** | **84.90** | **75.93** | **82.16** | 74.57 |

**Table 3**
Comparison of sequence manipulation tagging methods.

| Categories | Binary Cls | | | Text grounding | | |
|---|---|---|---|---|---|---|
| Methods | AUC | EER↓ | ACC | Precision | Recall | F1 |
| BERT [54] | 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 |
| ViKI (Ours) | **92.31** | **15.27** | **85.35** | **78.46** | **65.09** | **71.15** |

18,588 text attribute manipulations. About one-third of the manipulated images and half of the manipulated text are combined to form 32,693 mixed manipulation pairs. To maintain emotional balance, the dataset ensures an equal representation of positive and negative sentiment directions through modifications of the image and text attributes. Most manipulated images have small regions of manipulation, and the manipulated text tokens are relatively few. This makes the DGM$^4$ dataset more challenging for forgery detection compared to existing deepfake and multi-modal disinformation datasets. Illustrations from the DGM$^4$ dataset's particular instances are shown in Fig. 4.

### 4.3. Evaluation metrics

In order to ascertain the effectiveness of the proposed framework, we carry out a comprehensive evaluation, considering both objective and subjective criteria. For objective evaluation, evaluation of the four tasks, namely binary classification, multi-classification, manipulated image bounding box grounding, and manipulated text token grounding, involves the application of diverse assessment metrics. Three evaluation metrics, namely Area Under the Curve (AUC), Equal Error Rate (EER), and Accuracy (ACC), are employed to assess the binary classification task. The AUC metric captures the overall performance of a binary classifier through the ROC curve. The EER determines the balanced performance point where FAR and FRR are equal on the ROC curve. The ACC measures the accuracy of the classifier by calculating the ratio of correct predictions to total samples. To evaluate the performance of the multi-classification task, we adopt multiple evaluation metrics: mean Average Precision (mAP), class-wise F1 score (CF1), and overall F1 score (OF1). The mAP is a measure that calculates the average precision across diverse classes, incorporating both precision and recall. This metric offers a comprehensive assessment of the model's capability to accurately rank the classes. The CF1 evaluates the model's performance for individual classes by considering both precision and recall. It offers insights into the model's accuracy in classifying instances within each specific class. The OF1 computes the harmonic mean of precision and recall across all classes, providing a comprehensive measure of the model's overall performance in the multi-classification task. When assessing the performance of a task related to manipulated image bounding box grounding, several evaluation metrics are employed. These include the calculation of mean Intersection over Union (IoUmean), Intersection over Union at a threshold of 50% (IoU50), and Intersection over Union at a threshold of 75% (IoU75). Intersection over Union (IoU) is a commonly used evaluation metric in computer vision tasks, especially in object detection and image segmentation. It measures the overlap between predicted bounding boxes or segmentation masks and ground truth annotations. For the purpose of evaluating the performance of a manipulated text token grounding task, the metrics employed include Precision, Recall, and F1-score. Precision measures the accuracy of the positive predictions, Recall quantifies the ability to correctly identify positive instances, and F1-score combines both metrics to provide a balanced assessment of the model's performance. Among these twelve evaluation metrics, excluding the EER metric, all other metrics demonstrate that larger values correspond to superior system performance. And smaller values of the EER metric indicate improved system performance.

**Table 4**
Ablation studies on the critical modules in the proposed ViKI model. The best results are shown in **Red**.

| Categories | Binary Cls | | | Multi-Label Cls | | | Image grounding | | | Text grounding | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Methods | AUC | ACC | EER↓ | mAP | CF1 | OF1 | IoUmean | IoU50 | IoU75 | Precision | Recall | F1 |
| Baseline | 92.08 | 85.25 | 15.50 | 85.22 | 78.94 | 78.39 | 75.56 | 82.84 | 74.70 | 73.39 | 67.43 | 70.28 |
| Baseline+$\mathcal{L}_{\text{dist}}$ | 93.18 | 86.57 | 14.07 | 86.21 | 80.57 | 79.50 | 76.16 | 83.64 | 75.26 | 76.38 | 67.07 | 71.42 |
| Baseline+$Agg^{+}_{(\text{V2L})}$ | 93.28 | 86.40 | 14.10 | 86.68 | 80.63 | 79.66 | 76.38 | 83.81 | 75.48 | 75.33 | **68.48** | 71.74 |
| Baseline+$\mathcal{L}_{\text{dist}}$+$Agg^{+}_{(\text{V2L})}$ | **93.51** | **86.67** | **13.87** | **86.58** | **81.07** | **80.10** | **76.51** | **83.95** | **75.77** | **77.79** | 66.06 | **72.44** |

**Table 5**
Ablation study of image modality.

| Categories | Binary Cls | | | Image grounding | | |
|---|---|---|---|---|---|---|
| Methods | AUC | EER↓ | ACC | IoUmean | IoU50 | IoU75 |
| Image | 84.11 | 23.21 | 76.79 | 73.88 | 81.00 | 73.83 |
| **Joint** | **91.85** | **15.92** | **84.90** | **75.93** | **82.16** | **74.57** |

**Table 6**
Ablation study of text modality.

| Categories | Binary Cls | | | Text grounding | | |
|---|---|---|---|---|---|---|
| Methods | AUC | EER↓ | ACC | Precision | Recall | F1 |
| Text | 60.93 | 42.75 | 54.83 | 74.47 | 62.90 | 68.20 |
| **Joint** | **92.31** | **15.27** | **85.35** | **78.46** | **65.09** | **71.15** |

### *4.4. Comparison with state-of-the-art methods*

**Comparison with multimodal detection models** To establish the effectiveness of the proposed technique, comprehensive comparisons were made against diverse multi-modal learning methods and deepfake detection and sequence tagging methods. In order to assess the efficacy of the proposed methodology in the realm of multi-modal deepfake detection, a comparative study is conducted by comparing it with three SOTA multi-modal learning methods. To ensure a consistent dataset selection, the evaluation is performed on the DGM$^4$ dataset. Specifically, Vision-and-Language Transformer (ViLT) [59], a minimalist VLP model, is introduced in this research. It takes a monolithic approach, simplifying the visual processing to mirror the convolution-free methodology utilized for textual inputs. CLIP models [58] undergo comprehensive task acquisition during pre-training to optimize their training objective. This acquired task proficiency permits zero-shot transitions to many established datasets using natural language prompts effectively. Referring to Table 1, compared with the second-best HAMMER [27], ViKI exhibits superior performance in areas such as binary classification, multi-classification, and the grounding of manipulated entities in both images and texts. Moreover, the proposed method reduces EER by 1.63%. In Figs. 6 to 8, we present visual results of manipulation detection and grounding. The proposed method can effectively ground manipulated bounding boxes and identify the correct types of manipulation for both FA and FS. Moreover, most of the manipulated text tokens were successfully grounded as evidenced in Figs. 6 to 8. These visualizations confirm the efficacy of ViKI in achieving accurate manipulation detection and grounding (see Fig. 5).

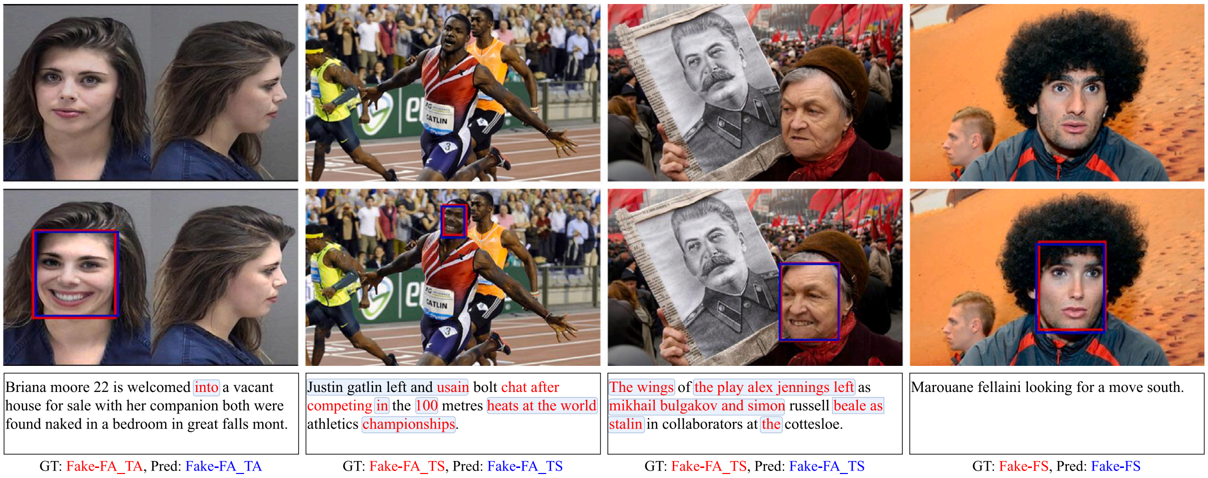
**Comparison with unimodal detection models** We conducted a comparative evaluation against competitive unimodal methods in two separate forgery data splits. Concerning the imaging modality, we compared ViKI with two established techniques, namely TS [60] and MAT [18], for comparative evaluation. TS model composes a module detecting high-frequency features for a new modality, a residual-driven module directing RGB discernment toward forgery signs, and a module harnessing dual modalities to boost mutual feature understanding. The MAT framework is designed to discern localized distinctive characteristics from diverse facial focal areas. In enhancing the network's efficacy, a unique regional loss formulation is embedded, complemented by an augmentation approach that influenced by attention, which can promote antagonistic training. For textual modality, a comparative evaluation is performed on two widely adopted sequence tagging approaches utilized in NLP for the purpose of grounding manipulated tokens, alongside binary classification. The examined techniques comprise BERT [54] and LUKE [61]. The BERT model is derived to pre-train profound bidirectional insights from non-labeled text. The model coalesces both left and right contextual cues in all strata. LUKE, grounded in a Wikipedia dataset, offers context-specific portrayals tailored for entity-related tasks. An entity-sensitive self-attention mechanism is incorporated, which discerns token types in its attention score computation. For this testing, ViKI extracted multimodal embedding from the paired samples while only unimodal-related objectives are used for supervising the model training. For image-modal deepfake detection, we excluded $\mathcal{L}_{\text{tmg}}$ and $\mathcal{L}_{\text{mdm}}$ in Eq. (13), while for text-modal manipulation detection, we excluded $\mathcal{L}_{\text{img}}$ and $\mathcal{L}_{\text{mdm}}$. The comparison results are shown in Tables 2 and 3, from which we observe that ViKI is robust performing detection and grounding for both unimodal deepfakes compared with other SOTA methods. This is attributed to the effective interactive Knowledge aggregation strategy in our model design (see Table 7).

### *4.5. Ablation study*

**Component Analysis** We first ablated the effectiveness of each component in our model design and report the results in Table 4. The baseline model is the reproduced HAMMER [27] by the official released code. We ablated the regularization $\mathcal{L}_{\text{dist}}$ and the adaptive fused representation $Agg^{+}_{\text{V2L}}$. We observed from Table 4 that both components improve the overall performance, and the improvement is more significant when they are jointly employed. The alignment of multimodal representations promotes the formulation of a compact feature space, and facilitates the subsequent feature aggregation in which complementary knowledge is utilized to enhance the learned feature. Therefore, we illustrated the efficiency of the two critical modules and the necessity to incorporate them simultaneously to learn more representative and information multimodal embeddings.

**Effectiveness of Cross-modality Knowledge** In this study, we highlight the necessity of exploring cross-modality knowledge by comparing the results from using solely unimodal (denoted as 'Image' or 'Text') or multimodal (denoted as 'Joint') knowledge. For unimodal design, unimodal information is employed as the input, without the feature regularization $\mathcal{L}_{\text{feat}}$ and knowledge aggregation. For the joint design, multimodal information is used as the input and only the model-specific training objectives are employed to supervise the training. As shown in Tables 5 and 6, using the knowledge from the different is benefiting for the formulation of a discriminative representation for either image deepfake detection or text manipulation tagging. It is highly necessary to exploit complementary information from different modalities to promote the learned embeddings.

**Influence of Aggregation Strategy** The proposed multimodal embedding aggregation module aims to enhance the representation of unimodal feature with the complementary information in the cross-modality embedding. We compared the proposed aggregation strategy, which uses the LPAA($E_I(I)$) as the residual component, with the other three counterparts. The design of these aggregators is shown in Fig. 9.



Briana moore 22 is welcomed into a vacant house for sale with her companion both were found naked in a bedroom in great falls mont.

GT: Fake-FA_TA, Pred: Fake-FA_TA

Justin gatlin left and usain bolt chat after competing in the 100 metres heats at the world athletics championships.

GT: Fake-FA_TS, Pred: Fake-FA_TS

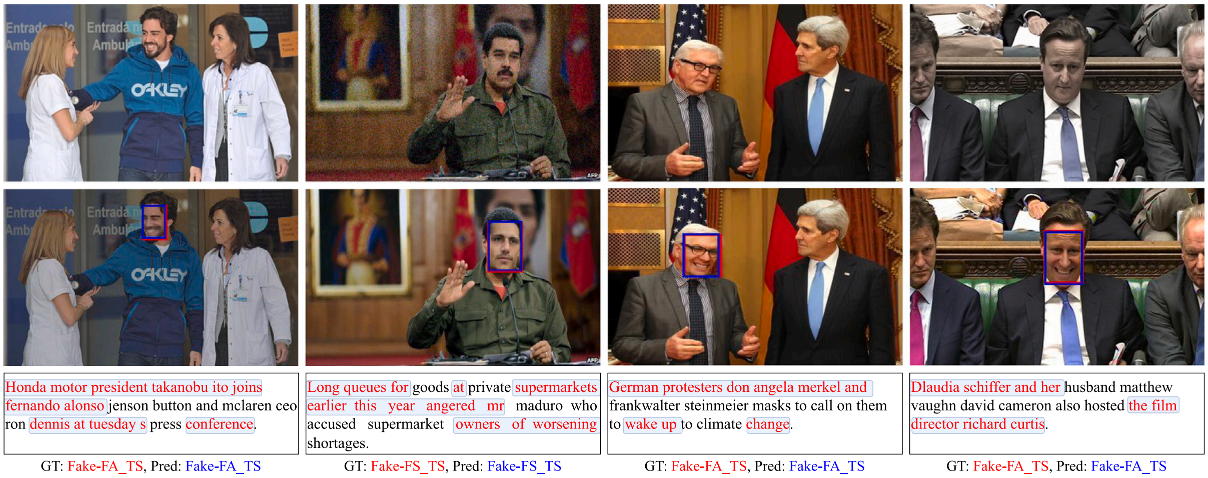
The wings of the play alex jennings left as mikhail bulgakov and simon russell beale as stalin in collaborators at the cottesloe.

GT: Fake-FA_TS, Pred: Fake-FA_TS

Marouane fellaini looking for a move south.

GT: Fake-FS, Pred: Fake-FS

**Fig. 6.** Visualization for a partial of samples in DGM$^4$ datasets. The text shown in red is the ground truth (GT), while the text inside the blue box is the prediction (Pred).



Honda motor president takanobu ito joins fernando alonso jenson button and mclaren ceo ron dennis at tuesday s press conference.

GT: Fake-FA_TS, Pred: Fake-FA_TS

Long queues for goods at private supermarkets earlier this year angered mr maduro who accused supermarket owners of worsening shortages.

GT: Fake-FS_TS, Pred: Fake-FS_TS

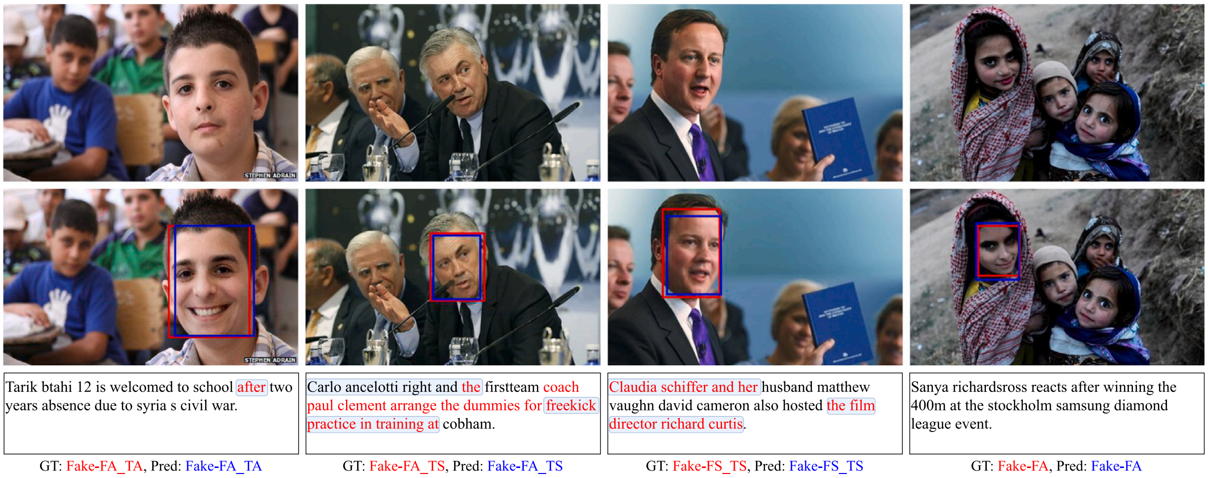
German protesters don angela merkel and frankwalter steinmeier masks to call on them to wake up to climate change.

GT: Fake-FA_TS, Pred: Fake-FA_TS

Dlaudia schiffer and her husband matthew vaughn david cameron also hosted the film director richard curtis.

GT: Fake-FA_TS, Pred: Fake-FA_TS

**Fig. 7.** Visualization for a partial of samples in DGM$^4$ datasets. The text shown in red is the ground truth (GT), while the text inside the blue box is the prediction (Pred).



Tarik btahi 12 is welcomed to school after two years absence due to syria s civil war.

GT: Fake-FA_TA, Pred: Fake-FA_TA

Carlo ancelotti right and the firstteam coach paul clement arrange the dummies for freekick practice in training at cobham.

GT: Fake-FA_TS, Pred: Fake-FA_TS

Claudia schiffer and her husband matthew vaughn david cameron also hosted the film director richard curtis.

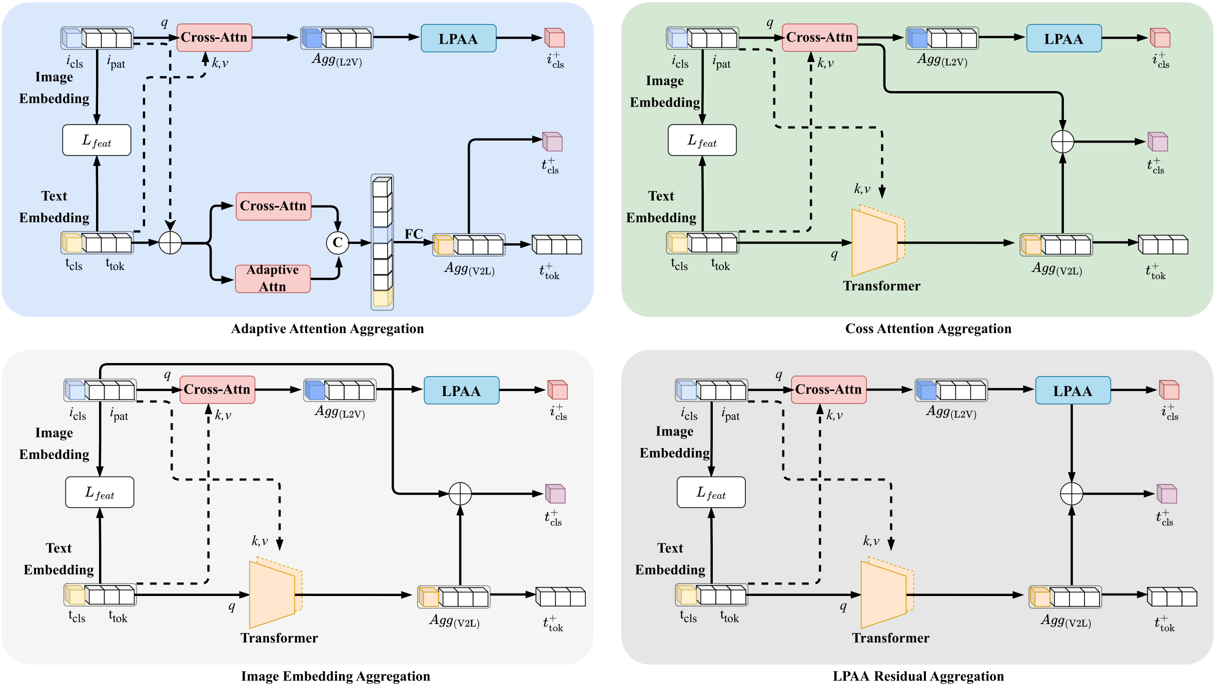
GT: Fake-FS_TS, Pred: Fake-FS_TS

Sanya richardsross reacts after winning the 400m at the stockholm samsung diamond league event.

GT: Fake-FA, Pred: Fake-FA

**Fig. 8.** Visualization for a partial of samples in DGM$^4$ datasets. The text shown in red is the ground truth (GT), while the text inside the blue box is the prediction (Pred).

For adaptive attention aggregation, a parallel attention group, include a cross-attention and adaptive attention, was employed to fuse the multimodal information. For cross-attention aggregation, the output of the cross-attention layer was used to enhance the aggregated feature $Agg_{(\mathrm{V2L})}$. For the image embedding aggregation, a skip connection was directly used to incorporate the image embedding into $Agg_{(\mathrm{V2L})}$. For the LPAA residual aggregation, the image information refined by the LPAA unit was used for the feature enhancement. The detection results for these four designs are shown in Fig. 2, which experimentally demonstrates the LPAA residual aggregation can result in a more



Adaptive Attention Aggregation

Coss Attention Aggregation

Image Embedding Aggregation

LPAA Residual Aggregation

**Fig. 9.** Ablation study on four types of cross-modal embedding aggregation strategies.

**Table 7**
Comparison results for the four types of cross-modal embedding aggregation strategies. The best results are shown in **Red**.

| Categories | Binary Cls | | | Multi-Label Cls | | | Image grounding | | | Text grounding | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Methods | AUC | ACC | EER↓ | mAP | CF1 | OF1 | IoUmean | IoU50 | IoU75 | Precision | Recall | F1 |
| Adaptive attention aggregation | 90.67 | 83.74 | 16.67 | 78.49 | 76.11 | 74.11 | 76.65 | 83.89 | 76.03 | 76.17 | 61.04 | 67.77 |
| Image embedding aggregation | 93.34 | 86.57 | 14.07 | 86.21 | 80.57 | 79.61 | 76.05 | 83.51 | 75.19 | 77.14 | 65.68 | 70.95 |
| Coss attention aggregation | 93.37 | 86.40 | 14.07 | 86.41 | 80.56 | 79.61 | 76.14 | 83.52 | 75.52 | 77.39 | 66.14 | 71.32 |
| LPAA residual aggregation | 93.51 | 86.67 | 13.87 | 86.58 | 81.07 | 80.10 | 76.51 | 83.95 | 75.77 | 77.79 | 66.06 | 72.44 |

representative multimodal embedding to better detect multimodal disinformation. Therefore, we use the LPAA residual aggregation in our ViKI design.

## 5. Conclusion

In this work, we proposed a novel model termed Vision-language Knowledge Interaction (ViKI) for detecting disinformation in multimodal data. ViKI is capable of jointly detecting and grounding disinformation in both vision and language modalities, which is bootstrapped by the aligned vision-language embeddings that are extracted and aggregated by ViKI. To enable the embedding to be discriminative in representing the multimodal objectives, ViKI binded the vision and language features by jointly optimizing a metric learning objective and a geometry distance minimization objective, which contributes to a well-aligned compact hypothesis space. To mine the complementary knowledge in cross-modality representation, a knowledge interaction mechanism was designed in ViKI to adaptively incorporate cross-modality information, which is discriminative for the subsequent multitask learning. We experimentally demonstrated the effectiveness of ViKI over the existing SOTA methods under both multimodal and unimodal settings. In contrast to SOAT methods, ViKI has demonstrated comprehensive advancements simultaneously on four distinct tasks, including binary classification, multi-classification, grounding of manipulated image bounding boxes, and grounding of manipulated text tokens. Notably, there is a rise of 3.71% in precision and 2.14% in CF1 relative to SOAT. The proposed method is proficient in detecting multimodal deepfakes. However, it requires an extensive computational commitment. Hence, forthcoming investigations are required to focus on enabling the network lightweight while maintaining the efficacy of the fusion mechanism.

## CRediT authorship contribution statement

**Qilei Li:** Conceptualization. **Mingliang Gao:** Methodology. **Guisheng Zhang:** Visualization, Software. **Wenzhe Zhai:** Investigation, Validation. **Jinyong Chen:** Software, Validation, Data curation, Investigation. **Gwanggil Jeon:** Reviewing and editing.

## Declaration of competing interest

None Declared.

## Data availability

No data was used for the research described in the article.

## References

[1] Z. Li, F. Liu, W. Yang, S. Peng, J. Zhou, A survey of convolutional neural networks: Analysis, applications, and prospects, IEEE Trans. Neural Netw. Learn. Syst. 33 (12) (2022) 6999–7019.

[2] Y. Taigman, M. Yang, M. Ranzato, L. Wolf, Deepface: Closing the gap to human-level performance in face verification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1701–1708.

[3] F. Prezja, J. Paloneva, I. Pölönen, E. Niinimäki, S. Äyrämö, DeepFake knee osteoarthritis X-rays from generative adversarial neural networks deceive medical experts and offer augmentation potential to automatic classification, Sci. Rep. 12 (1) (2022) 18573.

[4] Y.S. Kim, H.J. Song, J.H. Han, A study on the development of deepfake-based deep learning algorithm for the detection of medical data manipulation, Webology 19 (1) (2022) 4396–4409.

[5] J. Kietzmann, A.J. Mills, K. Plangger, Deepfakes: perspectives on the future "reality" of advertising and branding, Int. J. Advert. 40 (3) (2021) 473–485.

[6] B. Sivathanu, R. Pillai, B. Metri, Customers' online shopping intention by watching AI-based deepfake advertisements, Int. J. Retail Distrib. Manag. 51 (1) (2023) 124–145.

[7] H. Lu, H. Chu, Let the dead talk: How deepfake resurrection narratives influence audience response in prosocial contexts, Comput. Hum. Behav. 145 (2023) 107761.

[8] N. Waqas, S.I. Safie, K.A. Kadir, S. Khan, M.H.K. Khel, DEEPFAKE image synthesis for data augmentation, IEEE Access 10 (2022) 80847–80857.

[9] S. Greengard, Will deepfakes do deep damage? Commun. ACM 63 (1) (2019) 17–19.

[10] L. Verdoliva, Media forensics and deepfakes: an overview, IEEE J. Sel. Top. Sign. Proces. 14 (5) (2020) 910–932.

[11] J. Ternovski, J. Kalla, P. Aronow, The negative consequences of informing voters about deepfakes: Evidence from two survey experiments, J. Online Trust Saf. 1 (2) (2022).

[12] Y. Zhang, L. Zheng, V.L. Thing, Automated face swapping and its detection, in: 2017 IEEE 2nd International Conference on Signal and Image Processing (ICSIP), IEEE, 2017, pp. 15–19.

[13] X. Wang, N. Thome, M. Cord, Gaze latent support vector machine for image classification improved by weakly supervised region selection, Pattern Recognit. 72 (2017) 59–71.

[14] S. Bai, Growing random forest on deep convolutional neural networks for scene categorization, Expert Syst. Appl. 71 (2017) 279–287.

[15] A. Raza, K. Munir, M. Almutairi, A novel deep learning approach for deepfake image detection, Appl. Sci. 12 (19) (2022) 9820.

[16] D.A. Coccomini, R. Caldelli, F. Falchi, C. Gennaro, G. Amato, 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. 52–58.

[17] P. Bharadwaj, Z. Shao, Fake news detection with semantic features and text mining, Int. J. Nat. Lang. Comput. (IJNLC) 8 (2019).

[18] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, N. Yu, Multi-attentional deepfake detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 2185–2194.

[19] Q. Su, M. Wan, X. Liu, C.-R. Huang, et al., Motivations, methods and metrics of misinformation detection: an NLP perspective, Nat. Lang. Process. Res. 1 (1–2) (2020) 1–13.

[20] Y. Hou, Q. Guo, Y. Huang, X. Xie, L. Ma, J. Zhao, Evading DeepFake detectors via adversarial statistical consistency, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 12271–12280.

[21] S. Abdelnabi, R. Hasan, M. Fritz, 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. 14940–14949.

[22] S. Aneja, C. Bregler, M. Nieß ner, Cosmos: Catching out-of-context misinformation with self-supervised learning, 2021, arXiv preprint arXiv:2101.06278.

[23] Z. Jin, J. Cao, H. Guo, Y. Zhang, J. Luo, 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.

[24] D. Khattar, J.S. Goud, M. Gupta, V. Varma, Mvae: Multimodal variational autoencoder for fake news detection, in: The World Wide Web Conference, 2019, pp. 2915–2921.

[25] G. Luo, T. Darrell, A. Rohrbach, Newsclippings: Automatic generation of out-of-context multimodal media, 2021, arXiv preprint arXiv:2104.05893.

[26] Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, L. Su, J. Gao, 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. 849–857.

[27] R. Shao, T. Wu, Z. Liu, Detecting and grounding multi-modal media manipulation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 6904–6913.

[28] H.H. Nguyen, J. Yamagishi, I. Echizen, Capsule-forensics: Using capsule networks to detect forged images and videos, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 2307–2311.

[29] X. Xuan, B. Peng, W. Wang, J. Dong, On the generalization of GAN image forensics, in: Biometric Recognition: 14th Chinese Conference, CCBR 2019, Zhuzhou, China, October 12–13, 2019, Proceedings, Springer, 2019, pp. 134–141.

[30] Z. Xue, X. Jiang, Q. Liu, Z. Wei, Global–local facial fusion based GAN generated fake face detection, Sensors 23 (2) (2023) 616.

[31] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, S.C.H. Hoi, Align before fuse: Vision and language representation learning with momentum distillation, in: Advances in Neural Information Processing Systems, Vol. 34, 2021, pp. 9694–9705.

[32] A.M.H. Tiong, J. Li, B. Li, S. Savarese, S.C. Hoi, Plug-and-play vqa: Zero-shot vqa by conjoining large pretrained models with zero training, 2022, arXiv preprint arXiv:2210.08773.

[33] J. Li, D. Li, C. Xiong, S. Hoi, Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, in: International Conference on Machine Learning, PMLR, 2022, pp. 12888–12900.

[34] J. Guo, J. Li, D. Li, A.M.H. Tiong, B. Li, D. Tao, S. Hoi, From images to textual prompts: Zero-shot visual question answering with frozen large language models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 10867–10877.

[35] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K.V. Alwala, A. Joulin, I. Misra, ImageBind: One embedding space to bind them all, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 15180–15190.

[36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, Vol. 30, 2017.

[37] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al., Improving language understanding by generative pre-training, 2018.

[38] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, 2020, arXiv preprint arXiv:2010.11929.

[39] H. Li, S.J. Pan, S. Wang, A.C. Kot, Domain generalization with adversarial feature learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5400–5409.

[40] M. Ghifary, W.B. Kleijn, M. Zhang, D. Balduzzi, Domain generalization for object recognition with multi-task autoencoders, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2551–2559.

[41] R. Shao, X. Lan, J. Li, P.C. Yuen, Multi-adversarial discriminative deep domain generalization for face presentation attack detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 10023–10031.

[42] Y. Jia, J. Zhang, S. Shan, X. Chen, Single-side domain generalization for face anti-spoofing, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 8484–8493.

[43] H. Li, Y. Wang, R. Wan, S. Wang, T.-Q. Li, A. Kot, Domain generalization for medical imaging classification with linear-dependency regularization, Adv. Neural Inf. Process. Syst. 33 (2020) 3118–3129.

[44] S. Aslani, V. Murino, M. Dayan, R. Tam, D. Sona, G. Hamarneh, Scanner invariant multiple sclerosis lesion segmentation from MRI, in: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), IEEE, 2020, pp. 781–785.

[45] B. Sun, J. Feng, K. Saenko, Return of frustratingly easy domain adaptation, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2016.

[46] M. Long, Y. Cao, J. Wang, M. Jordan, Learning transferable features with deep adaptation networks, in: International Conference on Machine Learning, PMLR, 2015, pp. 97–105.

[47] R. Linsker, Self-organization in a perceptual network, Computer 21 (3) (1988) 105–117.

[48] R. Hadsell, S. Chopra, Y. LeCun, Dimensionality reduction by learning an invariant mapping, in: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), Vol. 2, IEEE, 2006, pp. 1735–1742.

[49] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: A unified embedding for face recognition and clustering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 815–823.

[50] K. Sohn, Improved deep metric learning with multi-class n-pair loss objective, in: Advances in Neural Information Processing Systems, Vol. 29, 2016.

[51] A.v.d. Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding, 2018, arXiv preprint arXiv:1807.03748.

[52] P.H. Chen, W. Wei, C.-j. Hsieh, B. Dai, Overcoming catastrophic forgetting by generative regularization, 2019, arXiv preprint arXiv:1912.01238.

[53] K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, Momentum contrast for unsupervised visual representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9729–9738.

[54] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018, arXiv preprint arXiv:1810.04805.

[55] S. Schneider, A. Baevski, R. Collobert, M. Auli, Wav2vec: Unsupervised pre-training for speech recognition, 2019, arXiv preprint arXiv:1904.05862.

[56] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: ICLR2019, 2019.

[57] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer, Automatic differentiation in pytorch, in: NIPS-W, 2017.

[58] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.

[59] W. Kim, B. Son, I. Kim, Vilt: Vision-and-language transformer without convolution or region supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 5583–5594.

[60] Y. Luo, Y. Zhang, J. Yan, W. Liu, Generalizing face forgery detection with high-frequency features, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 16317–16326.

[61] I. Yamada, A. Asai, H. Shindo, H. Takeda, Y. Matsumoto, LUKE: Deep contextualized entity representations with entity-aware self-attention, 2020, arXiv preprint arXiv:2010.01057.