Detecting and Grounding Multi-Modal Media Manipulation

Rui Shao[1,2]*, Tianxing Wu[2], Ziwei Liu[2]†

[1] School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen)
[2] S-Lab, Nanyang Technological University

shaorui@hit.edu.cn, {tianxing.wu, ziwei.liu}@ntu.edu.sg

https://github.com/rshaojimmy/MultiModal-DeepFake
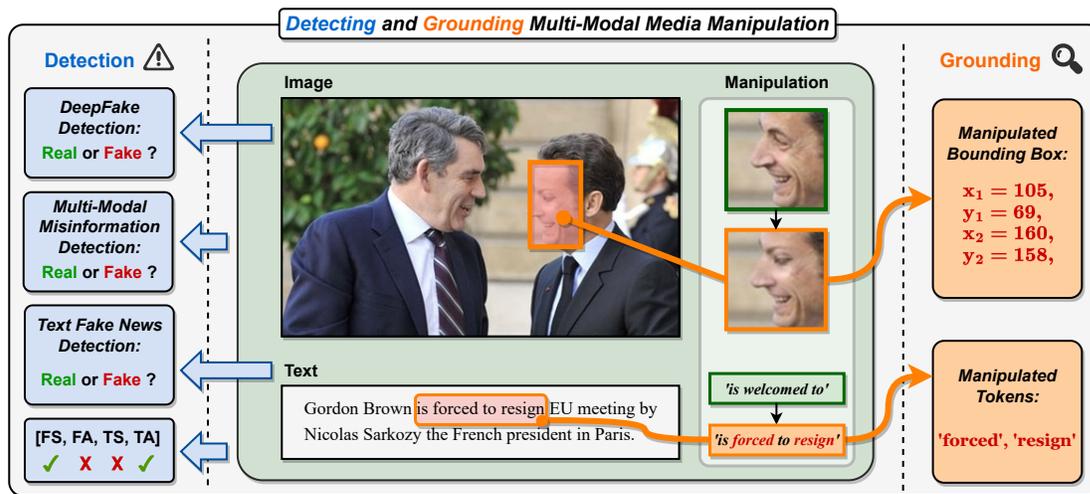
Figure 1. Different from existing single-modal forgery detection tasks, **DGM**[4] not only performs real/fake classification on the input image-text pair, but also attempts to detect more fine-grained manipulation types and ground manipulated image bboxes and text tokens. They provide more comprehensive interpretation and deeper understanding about manipulation detection besides the binary classification. (FS: Face Swap Manipulation, FA: Face Attribute Manipulation, TS: Text Swap Manipulation, TA: Text Attribute Manipulation)

Abstract

Misinformation has become a pressing issue. Fake media, in both visual and textual forms, is widespread on the web. While various deepfake detection and text fake news detection methods have been proposed, they are only designed for single-modality forgery based on binary classification, let alone analyzing and reasoning subtle forgery traces across different modalities. In this paper, we highlight a new research problem for multi-modal fake media, namely Detecting and Grounding Multi-Modal Media Manipulation (DGM[4]). DGM[4] aims to not only detect the authenticity of multi-modal media, but also ground the manipulated content (i.e., image bounding boxes and text tokens), which requires deeper reasoning of multi-modal media manipulation. To support a large-scale investigation, we construct the first DGM[4] dataset, where image-text pairs are manipulated by various approaches, with rich anno-

tation of diverse manipulations. Moreover, we propose a novel HierArchical Multi-modal Manipulation rEasoning tRansformer (HAMMER) to fully capture the fine-grained interaction between different modalities. HAMMER performs 1) manipulation-aware contrastive learning between two uni-modal encoders as shallow manipulation reasoning, and 2) modality-aware cross-attention by multi-modal aggregator as deep manipulation reasoning. Dedicated manipulation detection and grounding heads are integrated from shallow to deep levels based on the interacted multi-modal information. Finally, we build an extensive benchmark and set up rigorous evaluation metrics for this new research problem. Comprehensive experiments demonstrate the superiority of our model; several valuable observations are also revealed to facilitate future research in multi-modal media manipulation.

1. Introduction

With recent advances in deep generative models, increasing hyper-realistic face images or videos can be au-

*This work was done at S-Lab, Nanyang Technological University
†Corresponding author

Table 1. Comparison of the proposed $\mathbf{DGM}^4$ with existing tasks related to image and text forgery detection.

Problem Setting	Image Forgery		Text Forgery		Multi-Modal
	Detection	Grounding	Detection	Grounding	Forgery Detection
DeepFake Detection [26, 55]	✔	✘	✘	✘	✘
Text Fake News Detection [48, 53]	✘	✘	✔	✘	✘
Multi-Modal Misinformation Detection [1, 25]	✘	✘	✘	✘	✔
$\mathbf{DGM}^4$	✔	✔	✔	✔	✔

tomatically generated, which results in various security issues [35–41, 43, 52] such as serious *deepfake* problem [7, 12, 21, 34, 42] spreading massive fabrication on visual media. This threat draws great attention in computer vision community and various deepfake detection methods have been proposed. With the advent of Large Language Model, *e.g.*, BERT [6], GPT [31], enormous *text fake news* [48, 53] can be readily generated to maliciously broadcast misleading information on textual media. Natural Language Processing (NLP) field pays great attention to this issue and presents diverse text fake news detection methods.

Compared to a single modality, the multi-modal media (in form of image-text pairs) disseminates broader information with greater impact in our daily life. Thus, multi-modal forgery media tends to be more harmful. To cope with this new threat with a more explainable and interpretable solution, this paper proposes a novel research problem, namely **D**etecting and **G**rounding **M**ulti-**M**odal **M**edia **M**anipulation ($\mathbf{DGM}^4$). As shown in Table 1 and Fig. 1, two challenges are brought by $\mathbf{DGM}^4$: **1)** while current deepfake detection and text fake news detection methods are designed to detect forgeries of single modality, $\mathbf{DGM}^4$ demands simultaneously detecting the existence of forgery in both image and text modality and **2)** apart from binary classification like current single-modal forgery detection, $\mathbf{DGM}^4$ further takes grounding manipulated image bounding boxes (bboxes) and text tokens into account. This means existing single-modal methods are unavailable for this novel research problem. A more comprehensive and deeper reasoning of the manipulation characteristics between two modalities is of necessity. Note that some multi-modal misinformation works [1, 25] are developed. But they are only required to determine binary classes of multi-modal media, let alone manipulation grounding.

To facilitate the study of $\mathbf{DGM}^4$, this paper contributes the first large-scale $\mathbf{DGM}^4$ dataset. In this dataset, we study a representative multi-modal media form, *human-centric news*. It usually involves misinformation regarding politicians and celebrities, resulting in serious negative influence. We develop two different image manipulation (*i.e.,* face swap/attribute manipulation) and two text manipulation (*i.e.,* text swap/attribute manipulation) approaches to form the multi-modal media manipulation scenario. Rich annotations are provided for detection and grounding, including binary labels, fine-grained manipulation types, manipulated image bboxes and manipulated text tokens.

Compared to pristine image-text pairs, manipulated multi-modal media is bound to leave manipulation traces in manipulated image regions and text tokens. All of these traces together alter the cross-modal correlation and thus cause semantic inconsistency between two modalities. Therefore, reasoning semantic correlation between images and texts provides hints for the detection and grounding of multi-modal manipulation. To this end, inspired by existing vision-language representation learning works [16, 17, 30], we propose a novel **H**ier**A**rchical **M**ulti-modal **M**anipulation r**E**asoning t**R**ansformer (**HAMMER**) to tackle $\mathbf{DGM}^4$. To fully capture the interaction between images and texts, **HAMMER 1)** aligns image and text embeddings through manipulation-aware contrastive learning between two uni-modal encoders as shallow manipulation reasoning and **2)** aggregates multi-modal embeddings via modality-aware cross-attention of multi-modal aggregator as deep manipulation reasoning. Based on the interacted multi-modal embeddings in different levels, dedicated manipulation detection and grounding heads are integrated hierarchically to detect binary classes, fine-grained manipulation types, and ground manipulated image bboxes, manipulated text tokens. This hierarchical mechanism contributes to more fine-grained and comprehensive manipulation detection and grounding. Main contributions of our paper:

- We introduce a new research problem **D**etecting and **G**rounding **M**ulti-**M**odal **M**edia **M**anipulation ($\mathbf{DGM}^4$), with the objective of detecting and grounding manipulations in image-text pairs of human-centric news.
- We contribute a large-scale $\mathbf{DGM}^4$ dataset with samples generated by two image manipulation and two text manipulation approaches. Rich annotations are provided for detecting and grounding diverse manipulations.
- We propose a powerful **H**ier**A**rchical **M**ulti-modal **M**anipulation r**E**asoning t**R**ansformer (**HAMMER**). A comprehensive benchmark is built based on rigorous evaluation protocols and metrics. Extensive quantitative and qualitative experiments demonstrate its superiority.

2. Related Work

DeepFake Detection. To detect face forgery images, current deepfake detection methods are built based on spatial and frequency domains. Spatial-based deepfake detection methods exploit spatial visual cues, such as blending artifacts [19], textural features [4, 55, 57], 3D information [57], patch consistency [56] and noise character-

istics [10]. Frequency-based deepfake detection methods detect spectrum artifacts, like high-frequency components decomposed from Discrete Fourier Transform (DFT) [8], subtle frequency discrepancy derived from Discrete Cosine Transform (DCT) [29], up-sampling artifacts hidden in phase spectrum [23] and frequency-based metric learning [18]. Most of the above deepfake detection methods only perform binary classification in image media, not to mention manipulation grounding across multi-modalities.

Multi-Modal Misinformation Detection. Several existing works study the detection of multi-modal misinformation [1, 2, 13, 15, 25, 49]. Some of them deal with a small-scale human-generated multi-modal fake news [13, 15, 49], while others address out-of-context misinformation where a real image is paired with another swapped text without image and text manipulation [1, 2, 25]. All of these methods only perform binary classification based on simple image-text correlation. In contrast, DGM^4 studies large-scale machine-generated multi-media manipulation, which is closer to broad misinformation on the web in practice. Additionally, DGM^4 requires not only manipulation detection for binary classification, but also manipulation grounding with more interpretation for multi-modal manipulation.

3. Multi-Modal Media Manipulation Dataset

Most of existing misinformation datasets focus on single-modal image forgery [7, 12, 20, 34] or text forgery [44, 48, 53]. Some multi-modal datasets are built, but they usually contain a small amount of human-generated fake news [5, 13] or out-of-context pairs [2, 25] for binary forgery detection. To better facilitate the proposed novel research problem, we present DGM^4 dataset, studying large-scale machine-generated multi-modal media manipulation. DGM^4 dataset is constructed with diverse manipulation techniques on both **Image** and **Text** modality. All samples are annotated with rich, fine-grained labels that enable both **Detection** and **Grounding** of media manipulation.

3.1. Source Data Collection

Among all forms of multi-modal media, we specifically focus on *human-centric news*, in consideration of its great public influence. We thus develop our dataset based on the VisualNews dataset [22], which collects numerous image-text pairs from real-world news sources (The Guardian, BBC, USA TODAY, and The Washington Post). To formulate a human-centric scenario with meaningful context, we further conduct data filtering on both image and text modality, and only keep the appropriate pairs to form the source pool $O = \{p_o | p_o = (I_o, T_o)\}$ for manipulation.

3.2. Multi-Modal Media Manipulation

We employ two types of harmful manipulations on both image and text modality. 'Swap' type is designed to include relatively global manipulation traces, while 'Attribute' type introduces more fine-grained local manipulations. The manipulated images and texts are then randomly mixed with pristine samples to form a total of 8 fake plus one original manipulation classes. Distribution of manipulation classes and some samples are displayed in Fig. 2 (a).

Face Swap (FS) Manipulation. In this manipulation type, the *identity* of the main character is attacked by swapping his/her face with another person. We adopt two representative face swap approaches, SimSwap [3] and InfoSwap [9]. For each original image I_o, we choose one of the two approach to swap the largest face I_o^f with a random source face I_{celeb}^f from CelebA-HQ dataset [14], producing a face swap manipulation sample I_s. The MTCNN bbox of the swapped face $y_{\text{box}} = \{x_1, y_1, x_2, y_2\}$ is then saved as annotation for grounding.

Face Attribute (FA) Manipulation. As a more fine-grained image manipulation scenario, face attribute manipulation attempts to manipulate the *emotion* of the main character's face while preserving the identity. For example, if the original face is smiling, we deliberately edit it to the opposite emotion, *e.g.,* an angry face. To achieve this, we first predict the original facial expression of the aligned face I_o^f with a CNN-based network, then edit the face towards the opposite emotion using GAN-based methods, HFGI [47] and StyleCLIP [28]. After obtaining the manipulated face I_{emo}^f, we re-render it back to the original image I_o to obtain the manipulated sample I_a. Bbox y_{box} is also provided.

Text Swap (TS) Manipulation. In this scenario, the text is manipulated by altering its overall *semantic* while preserving words regarding main character. Given an original caption T_o, we use Named Entity Recognition (NER) model to extract the person's name as query 'PER'. Then we retrieve a different text sample T_o' containing same 'PER' entity from the source corpus O. T_o' is then selected as the manipulated text T_s. Note that we compute the semantic embedding of each text using Sentence-BERT [32] and only accept T_o' that has low cosine similarity with T_o. This ensures the retrieved text is not semantically aligned with T_o, so that the text semantic regarding the main character in the obtained pair $p_m = (I_o, T_s)$ is manipulated. After that, given M text tokens in T_s, we annotate them with a M-dimensional one-hot vector $y_{\text{tok}} = \{y_i\}_{i=1}^{M}$, where $y_i \in \{0, 1\}$ denotes whether the i-th token in T_s is manipulated or not.

Text Attribute (TA) Manipulation. Although news is a relatively objective media form, we observe that a considerable portion of news samples $p_o \in O$ still carry *sentiment* bias within the text T_o, as depicted in Fig. 2 (d). The malicious manipulation of text attributes, especially its sentiment tendency, could be more harmful and also harder to be detected as it causes less cross-modal inconsistency than text swap manipulation. To reflect this specific situation, we first use a RoBERTa [24] model to split the cap-
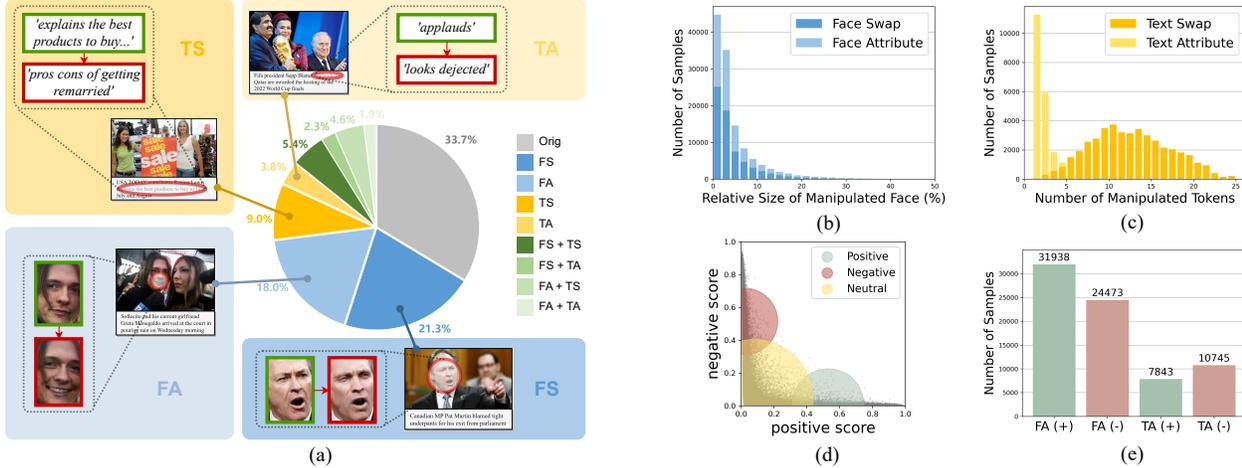
Figure 2. Statistics of **DGM**^{4} dataset. (a) Distribution of manipulation classes; (b) manipulated regions of most images are small-size, especially for face attribute manipulation; (c) manipulated tokens of text attribute manipulation are fewer than text swap manipulation; (d) distribution of text sentiment scores in the source pool. (e) number of manipulated samples towards each face/text attribute direction.

tions into positive, negative and neutral sentiment corpora: $\{O_+, O_-, O_{neu}\}$. Following [45], we replace all sentiment words of the original text T_o with the opposite sentiment text generated by a B-GST model trained on our own corpora $\{O_+, O_-\}$, obtaining T_a. Similar to text swap manipulation, all text tokens are also annotated with ground-truth vector y_{tok}.

Combination and Perturbation. Once all single-modal manipulations are finished, we combine the obtained manipulation samples I_s, I_a, T_s and T_a with the original (I_o, T_o) pairs. This forms a multi-modal manipulated media pool with full manipulation types: $P = \{p_m | p_m = (I_x, T_y), x, y \in \{o, s, a\}\}$. Each pair p_m in the pool is provided with a binary label y_{bin}, a fine-grained manipulation type annotation y_{mul}, aforementioned annotations y_{box} and y_{tok}. y_{bin} describes whether the image-text pair p_m is real or fake, and $y_{mul} = \{y_j\}_{j=1}^4$ is a 4-dimensional vector denoting whether the j-th manipulation type (*i.e.,* FS, FA, TS, TA) appears in p_m. To better reflect the real-world situation where manipulation traces may be covered up by noise, we employ random image perturbations on 50% of the media pool P, such as JPEG compression, Gaussian Blur, *etc*.

3.3. Dataset Statistics

The overall statistics of **DGM**^{4} dataset are illustrated in Fig. 2 (a). It consists a total of **230k** news samples, including 77,426 pristine image-text pairs and 152,574 manipulated pairs. The manipulated pairs contain 66,722 face swap manipulations, 56,411 face attribute manipulations, 43,546 text swap manipulations and 18,588 text attribute manipulations. $\sim$1/3 of the manipulated images and $\sim$1/2 of the manipulated text are combined together to form 32,693 mixed-manipulation pairs. Since both image and text attributes can be edited towards two opposite sentiment directions, we deliberately keep a balanced proportion to create

an emotionally-balanced dataset, as shown in Fig. 2 (e).

Furthermore, it can be observed from Fig. 2 (b)-(c) that the manipulated regions of most images and the number of manipulated text tokens are relatively small. This indicates **DGM**^{4} dataset provides a much more challenging scenario for forgery detection compared to existing deepfake and multi-modal misinformation datasets.

4. HAMMER

To address **DGM**^{4}, as illustrated in Fig. 3, we propose a **H**ier**A**rchical **M**ulti-modal **M**anipulation r**E**asoning t**R**ansformer (**HAMMER**), which is composed of two unimodal encoders (*i.e.,* Image Encoder E_v, Text Encoder E_t), Multi-Modal Aggregator F, and dedicated manipulation detection and grounding heads (*i.e.,* Binary Classifier C_b, Multi-Label Classifier C_m, BBox Detector D_v, and Token Detector D_t). All of these uni-modal encoders and multi-modal aggregator are built based on transformer-based architecture [46]. As mentioned above, modeling semantic correlation and capturing semantic inconsistency between two modalities can facilitate detection and grounding of multi-modal manipulation. However, there exist two challenges **1)** as discussed in Sec. 3.3 and shown in Fig. 2 (b)-(c), a large portion of multi-modal manipulations are minor and subtle, locating in some small-size faces and a few word tokens and **2)** much visual and textual noise [17] exists in multi-modal media on the web. As a result, some semantic inconsistencies caused by manipulation may be neglected or covered by noise. This demands more fine-grained reasoning of multi-modal correlation. To this end, we devise **HAMMER** to perform hierarchical manipulation reasoning which explores multi-modal interaction from shallow to deep levels, along with hierarchical manipulation detection and grounding. In the shallow manipulation reasoning, we carry out semantic alignment between image and
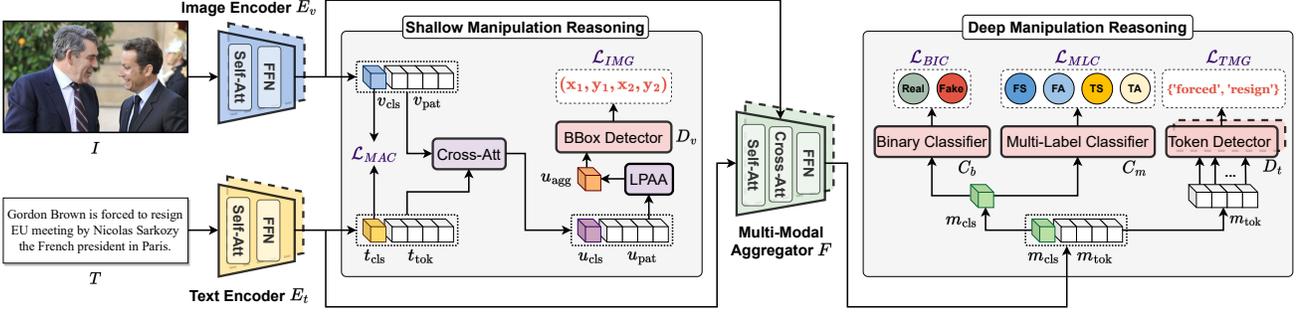
Figure 3. Overview of proposed **HAMMER**. It **1)** aligns image and text embeddings through manipulation-aware contrastive learning between Image Encoder E_v, Text Encoder E_t in shallow manipulation reasoning and **2)** further aggregates multi-modal embeddings via modality-aware cross-attention of Multi-Modal Aggregator F in deep manipulation reasoning. Based on the interacted multi-modal embeddings in different levels, various manipulation detection and grounding heads (Multi-Label Classifier C_m, Binary Classifier C_b, BBox Detector D_v, and Token Detector D_t are integrated to perform their tasks hierarchically. Modules with dashed lines mean they are the corresponding momentum versions of Image Encoder, Text Encoder, Multi-Modal Aggregator and Token Detector, respectively.

text embeddings through Manipulation-Aware Contrastive Loss $\mathcal{L}_{MAC}$, and conduct manipulated bbox grounding under Image Manipulation Grounding Loss $\mathcal{L}_{IMG}$. In the deep manipulation reasoning, based on deeper interacted multi-modal information generated by Multi-Modal Aggregator, we then detect binary classes with Binary Classification Loss $\mathcal{L}_{BIC}$, fine-grained manipulation types with Multi-Label Classification Loss $\mathcal{L}_{MLC}$, and ground manipulated text tokens via Text Manipulation Grounding Loss $\mathcal{L}_{TMG}$. By combing all the above losses, manipulation reasoning is performed hierarchically, contributing to a joint optimization framework as follows:

$$\mathcal{L} = \mathcal{L}_{MAC} + \mathcal{L}_{IMG} + \mathcal{L}_{MLC} + \mathcal{L}_{BIC} + \mathcal{L}_{TMG} \quad (1)$$

4.1. Shallow Manipulation Reasoning

Given an image-text pair $(I, T) \sim P$, we patchify and encode image I into a sequence of image embeddings via self-attention layers and feed-forward networks in Image Encoder as $E_v(I) = \{v_{cls}, v_{pat}\}$, where v_{cls} is the embedding of `[CLS]` token, and $v_{pat} = \{v_1, ..., v_N\}$ are embeddings of N corresponding image patches. Text Encoder extracts a sequence of text embeddings of T as $E_t(T) = \{t_{cls}, t_{tok}\}$, where t_{cls} is the embedding of `[CLS]` token, and $t_{tok} = \{t_1, ..., t_M\}$ are embeddings of M text tokens.
Manipulation-Aware Contrastive Learning. To help two uni-modal encoders better exploit the semantic correlation of images and texts, we align image and text embeddings through cross-modal contrastive learning. Nevertheless, some subtle multi-modal manipulations cause minor semantic inconsistency between two modalities, which are hardly unveiled by normal contrastive learning. To emphasize the semantic inconsistency caused by manipulations, **HAMMER** proposes manipulation-aware contrastive learning on image and text embeddings. Different from normal cross-modal contrastive learning pulling embeddings of original image-text pairs close while only pushing those

of unmatched pairs apart, manipulation-aware contrastive learning pushes away embeddings of manipulated pairs as well so that semantic inconsistency produced by them can be further emphasized. Following InfoNCE loss [27], we formulate image-to-text contrastive loss by:

$$\mathcal{L}_{v2t}(I, T^+, T^-) = -\mathbb{E}_{p(I,T)} \left[\log \frac{\exp(S(I, T^+)/\tau)}{\sum_{k=1}^{K} \exp(S(I, T_k^-)/\tau)} \right] \quad (2)$$

where τ is a temperature hyper-parameter, $T^- = \{T_1^-, ..., T_K^-\}$ is a set of negative text samples that are not matched to I *as well as that belong to manipulated image-text pairs*. Since `[CLS]` token serves as the semantic representation of the whole image and text, we use two projection heads h_v and $\hat{h}_t$ to map `[CLS]` tokens of two modalities to a lower-dimensional (256) embedding space for similarity calculation: $S(I, T) = h_v(v_{cls})^T \hat{h}_t(\hat{t}_{cls})$. Inspired by MoCo [11], we learn momentum uni-modal encoders $\hat{E}_v$, $\hat{E}_t$ (an exponential-moving-average version) and momentum projection heads for two modalities respectively. Two queues are used to store the most recent K image-text pair embeddings. Here $\hat{t}_{cls}$ are `[CLS]` tokens from text momentum encoders and $\hat{h}_t(\hat{t}_{cls})$ means projected text embeddings from text momentum projection head. Similarly, text-to-image contrastive loss is as follows:

$$\mathcal{L}_{t2v}(T, I^+, I^-) = -\mathbb{E}_{p(I,T)} \left[\log \frac{\exp(S(T, I^+)/\tau)}{\sum_{k=1}^{K} \exp(S(T, I_k^-)/\tau)} \right] \quad (3)$$

where $I^- = \{I_1^-, ..., I_K^-\}$ is a queue of K recent negative image samples that are not matched to T *as well as that belong to manipulated image-text pairs*. $S(T, I) = h_t(t_{cls})^T \hat{h}_v(\hat{v}_{cls})$. Inspired by [51], to maintain reasonable semantic relation within each single modality, we further carry out intra-modal contrastive learning in both modalities. We incorporate all the losses to form Manipulation-Aware Contrastive Loss as follows:

$$\mathcal{L}_{MAC} = \frac{1}{4}[\mathcal{L}_{v2t}(I, T^+, T^-) + \mathcal{L}_{t2v}(T, I^+, I^-) + \mathcal{L}_{v2v}(I, I^+, I^-) + \mathcal{L}_{t2t}(T, T^+, T^-)] \quad (4)$$

Manipulated Image Bounding Box Grounding. As mentioned above, FS or FA swaps identities or edits attributes of faces in images. This alters their correlation to corresponding texts in terms of persons' names or emotions. Given this, we argue that the manipulated image region could be located by finding local patches that have inconsistencies with the text embeddings. In this regard, we perform cross-attention between image and text embeddings to obtain patch embeddings that contain image-text correlation. Attention function [46] is performed on normalized query (Q), key (K), and value (V) features as:

$$\text{Attention}(Q, K, V) = \text{Softmax}(K^T Q / \sqrt{D})V \qquad (5)$$

Here we cross-attend the image embedding with text embedding, by treating Q as image embedding, K and V as text embedding as follows:

$$U_v(I) = \text{Attention}(E_v(I), E_t(T), E_t(T)) + E_v(I) \qquad (6)$$

where $U_v(I) = \{u_{\text{cls}}, u_{\text{pat}}\}$. $u_{\text{pat}} = \{u_1, ..., u_N\}$ are N image patch embeddings interacted with text information. Unlike [CLS] token u_{cls}, the patch tokens u_{pat} are generated with position encoding [46]. This means they possess richer local spatial information and thus are more suitable for manipulated image bbox grounding. Based on this analysis, we propose *Local Patch Attentional Aggregation (LPAA)* to aggregate the spatial information in u_{pat} via an attentional mechanism. This aggregation is performed by cross-attending a [AGG] token with u_{pat} as follows:

$$u_{\text{agg}} = \text{Attention}([\text{AGG}], u_{\text{pat}}, u_{\text{pat}}) \qquad (7)$$

Different from previous work [54] directly using [CLS] token for bbox prediction, we perform the manipulated bbox grounding based on the attentionally aggregated embedding u_{agg}. Specifically, we input u_{agg} into BBox Detector D_v and calculate Image Manipulation Grounding Loss by combing normal ℓ_1 loss and generalized Intersection over Union (IoU) loss [33] as follows:

$$\mathcal{L}_{IMG} = \mathbb{E}_{(I,T)\sim P}[\|\text{Sigmoid}(D_v(u_{\text{agg}})) - y_{box}\| \\ + \mathcal{L}_{\text{IoU}}(\text{Sigmoid}(D_v(u_{\text{agg}})) - y_{box})] \qquad (8)$$

4.2. Deep Manipulation Reasoning

Manipulated token grounding is a tougher task than manipulated bbox grounding as it requires deeper analysis and reasoning on the correlation between images and texts. For example, as illustrated in Fig. 3, we are able to detect the manipulated tokens in T, *i.e.,* 'force' and 'resign', only when we are aware of such negative words mismatching the positive emotion (*i.e.,* smiling faces) in I. Besides, we need to summarize multi-modal information to detect fine-grained manipulation types and binary classes. This demands a comprehensive information summary at this stage. To this end, we propose deep manipulation reasoning.

Manipulated Text Token Grounding. To model deeper multi-modal interaction, as depicted in Fig. 3, we propose modality-aware cross-attention to further lead text embedding $E_t(T)$ to interact with image embedding $E_v(I)$

through multiple cross-attention layers in Multi-Modal Aggregator F. This generates aggregated multi-modal embedding $F(E_v(I), E_t(T)) = \{m_{\text{cls}}, m_{\text{tok}}\}$. In particular, $m_{\text{tok}} = \{m_1, ..., m_M\}$ represent the deeper aggregated embeddings corresponding to each token in T. At this stage, each token in T has passed through multiple self-attention layers in E_t and cross-attention layers in F. In this way, each token embedding in m_{tok} not only entirely explores the context information of text, but also fully interacts with image features, which fits manipulated text tokens grounding. Moreover, grounding manipulated tokens is equal to labeling each token as real or fake. This is similar to sequence tagging task in NLP. Notably, unlike existing sequence tagging task mainly studied in text modality, manipulated text token grounding here can be regarded as a novel *multi-modal sequence tagging* since each token is interacted with two modality information. In this case, we use a Token Detector D_t to predict the label of each token in m_{tok} and calculate cross-entropy loss as follows:

$$\mathcal{L}_{tok} = \mathbb{E}_{(I,T)\sim P}\mathbf{H}(D_t(m_{\text{tok}}), y_{\text{tok}}) \qquad (9)$$

where $\mathbf{H}(\cdot)$ is the cross-entropy function. As mentioned, news on the web is usually noisy with texts unrelated to paired images [17]. To alleviate over-fitting to noisy texts, as shown in Fig. 3, we further learn momentum versions for Multi-Modal Aggregator and Token Detector, respectively, denoted as $\hat{F}$ and $\hat{D}_t$. We can obtain the multi-modal embedding from momentum modules as $\hat{F}(\hat{E}_v(I), \hat{E}_t(T)) = \{\hat{m}_{\text{cls}}, \hat{m}_{\text{tok}}\}$. Based on this, momentum Token Detector generates soft pseudo-labels to modulate the original token prediction, by calculating the KL-Divergence as follows:

$$\mathcal{L}_{tok}^{mom} = \mathbb{E}_{(I,T)\sim P}\text{KL}\left[D_t(m_{\text{tok}})\|\hat{D}_t(\hat{m}_{\text{tok}})\right] \qquad (10)$$

The final Text Manipulation Grounding Loss is a weighted combination as follows:

$$\mathcal{L}_{TMG} = (1 - \alpha)\mathcal{L}_{tok} + \alpha\mathcal{L}_{tok}^{mom} \qquad (11)$$

Fine-Grained Manipulation Type Detection and Binary Classification. Unlike current forgery detection works mainly performing real/fake binary classification, we expect our model to provide more interpretation for manipulation detection. As mentioned in Sec. 3.2, two image and two text manipulation approaches are introduced in **DGM**^{4} dataset. Given this, we aim to further detect four fine-grained manipulation types. As different manipulation types could appear in one image-text pair simultaneously, we treat this task as a specific *multi-modal multi-label classification*. Since [CLS] token m_{cls} aggregates multi-modal information after modality-aware cross-attention, it can be utilized as a comprehensive summary of manipulation characteristics. We thus concatenate a Multi-Label Classifier C_m on top of it to calculate Multi-Label Classification Loss:

$$\mathcal{L}_{MLC} = \mathbb{E}_{(I,T)\sim P}\mathbf{H}(C_m(m_{\text{cls}}), y_{\text{mul}}) \qquad (12)$$

Naturally, we also conduct a normal binary classification based on m_{cls} as follows:

$$\mathcal{L}_{BIC} = \mathbb{E}_{(I,T)\sim P}\mathbf{H}(C_b(m_{\text{cls}}), y_{\text{bin}}) \qquad (13)$$

Table 2. Comparison of multi-modal learning methods for DGM^4.

Categories	Binary Cls			Multi-Label Cls			Image Grounding			Text Grounding		
Methods	AUC	EER	ACC	mAP	CF1	OF1	IoUmean	IoU50	IoU75	Precision	Recall	F1
CLIP [30]	83.22	24.61	76.40	66.00	59.52	62.31	49.51	50.03	38.79	58.12	22.11	32.03
ViLT [16]	85.16	22.88	78.38	72.37	66.14	66.00	59.32	65.18	48.10	66.48	49.88	57.00
Ours	**93.19**	**14.10**	**86.39**	**86.22**	**79.37**	**80.37**	**76.45**	**83.75**	**76.06**	**75.01**	**68.02**	**71.35**

Table 3. Comparison of deepfake detection methods for DGM^4.

Categories	Binary Cls			Image Grounding		
Methods	AUC	EER	ACC	IoUmean	IoU50	IoU75
TS [26]	91.80	17.11	82.89	72.85	79.12	74.06
MAT [55]	91.31	17.65	82.36	72.88	78.98	74.70
Ours	**94.40**	**13.18**	**86.80**	**75.69**	**82.93**	**75.65**

Table 4. Comparison of sequence tagging methods for DGM^4.

Categories	Binary Cls			Text Grounding		
Methods	AUC	EER	ACC	Precision	Recall	F1
BERT [6]	80.82	28.02	68.98	41.39	63.85	50.23
LUKE [50]	81.39	27.88	76.18	50.52	37.93	43.33
Ours	**93.44**	**13.83**	**87.39**	**70.90**	**73.30**	**72.08**

Table 5. Ablation study of image modality.

Categories	Binary Cls			Image Grounding		
Methods	AUC	EER	ACC	IoUmean	IoU50	IoU75
Ours-Image	93.96	13.83	86.13	75.58	82.44	**75.80**
Ours	**94.40**	**13.18**	**86.80**	**75.69**	**82.93**	75.65

Table 6. Ablation study of text modality.

Categories	Binary Cls			Text Grounding		
Methods	AUC	EER	ACC	Precision	Recall	F1
Ours-Text	75.67	32.46	72.17	42.99	33.68	37.77
Ours	**93.44**	**13.83**	**87.39**	**70.90**	**73.30**	**72.08**

5. Experiments

Please refer to appendix for implementation details and rigorous setup of evaluation metrics.

5.1. Benchmark for DGM^4

Comparison with multi-modal learning methods. We adapt two SOTA multi-modal learning methods to **DGM**^{4} setting for comparison. Specifically, CLIP [30] is one of the most popular *dual-stream* approaches where two modalities are not concatenated at the input level. For adaptation, we make outputs of two streams interact with each other through cross-attention layers. Detection and grounding heads are further integrated on top of them. In addition, ViLT [16] is a representative *single-stream* approach where cross-modal interaction layers are operated on a concatenation of image and text inputs. We also adapt it by concatenating detection and grounding heads on corresponding outputs of the model. We tabulate comparison results in Table 2. The results show that the proposed method significantly outperforms both baselines in terms of all evaluation metrics. This demonstrates that hierarchical manipulation reasoning is more able to accurately and comprehensively model the correlation between images and texts and capture semantically inconsistency caused by manipulation, contributing to better manipulation detection and grounding.

Comparison with deepfake detection and sequence tag-

ging methods. We compare our method with competitive uni-modal methods in two single-modal forgery data splits, respectively. For a fair comparison, in addition to the original ground-truth regarding binary classification, we further integrate manipulation grounding heads into uni-modal models with corresponding annotations of grounding. For image modality, we tabulate the comparison with two SOTA deepfake detection methods in Table 3. For text modality, we compare two widely-used sequence tagging methods in NLP to ground manipulated tokens along with binary classification. We report the comparison results in Table 4. Tables 3 and 4 show that **HAMMER** performs better than uni-modal methods for single-modal forgery detection by a large margin. This indicates our method trained with multi-modal media also achieves promising manipulation detection and grounding performance in each single modality.

5.2. Experimental Analysis

Ablation study of two modalities. To validate the importance of multi-modal correlation for our model, we perform ablation study by only keeping corresponding input and network components with respect to image (Ours-Image) or text (Ours-Text) modality. We tabulate results in Tables 5 and 6, showing the performance of complete version of our model surpasses its ablated parts, especially in text modality. This suggests the performance degrades once one of the two modalities is missing without cross-modal interaction. This is to say, through exploiting correlation between two modalities via our model, more complementary information between them can be dug out to promote our task. Particularly, this correlation is more essential for manipulation detection and grounding in text modality.

Ablation study of losses. The considered losses and corresponding results obtained for each case are tabulated in Table 7. As evident from Table 7, removing the task-general loss, *i.e.,* $\mathcal{L}_{MAC}$, nearly all the performance degenerates. This implies manipulation-aware contrastive learning is indispensable for our task. After getting rid of any one of task-specific losses, *i.e.,* $\mathcal{L}_{MLC}$, $\mathcal{L}_{IMG}$ and $\mathcal{L}_{TMG}$, not only the performance of the corresponding task degrades dramatically, but also the overall binary classification performance probably becomes lower. Comparatively, our model with the complete loss function obtains the best performance in most of cases, indicating the effectiveness and complementarity of all losses. In particular, the first row of Table 7 represents the current multi-modal misinformation detection scenario where only $\mathcal{L}_{BIC}$ is used. Our method substantially outperforms this baseline on binary classification,

Table 7. Ablation study of losses in the proposed method.

Losses					Binary Cls			Multi-Label Cls			Image Grounding			Text Grounding		
BIC	MLC	MAC	IMG	TMG	AUC	EER	ACC	mAP	CF1	OF1	IoUmean	IoU50	IoU75	Precision	Recall	F1
✓					91.04	16.91	83.81	20.79	33.84	33.48	4.81	0.33	0.00	15.95	**78.70**	26.53
✓		✓	✓	✓	91.74	16.08	84.39	27.22	30.81	27.62	74.05	81.34	72.59	74.30	66.84	70.37
✓	✓		✓	✓	92.77	14.53	86.01	85.52	79.09	79.86	75.98	83.37	75.25	**77.82**	61.83	68.91
✓	✓	✓		✓	**93.21**	14.30	86.28	**86.29**	79.37	80.32	4.69	0.17	0.00	75.72	67.44	71.34
✓	✓	✓	✓		92.99	14.62	86.15	86.06	79.06	79.93	**76.51**	83.73	76.05	13.93	58.87	22.53
✓	✓	✓	✓	✓	93.19	**14.10**	**86.39**	86.22	**79.37**	**80.37**	76.45	**83.75**	**76.06**	75.01	68.02	**71.35**
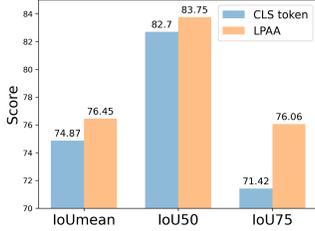
Figure 4. Efficacy of local patch attentional aggregation (LPAA).
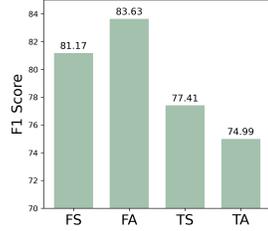
Figure 5. Performance of each manipulation type.

(a) GT: Fake-FS, Pred: Fake-FS   (b) GT: Fake-FA, Pred: Fake-FA

(c) GT: Fake-TS, Pred: Fake-TS   (d) GT: Fake-TA, Pred: Fake-TA

Figure 6. Visualization of detection and grounding results. Ground truth annotations are in red, and prediction results are in blue.

(a) Attention map in TS   (b) Attention map in TA

Figure 7. Grad-CAM visualizations on manipulated text tokens.

implying more manipulation grounding tasks in **DGM**^{4} facilitate binary classification as well.

Efficacy of LPAA. Regarding manipulated bbox grounding, we compare the usage of `[CLS]` token [54] with proposed LPAA in Fig. 4. Fig. 4 shows LPAA yields better performance under all metrics, verifying its efficacy.

Details of manipulation type detection. We plot the classification performance of each manipulation type based on the output of Multi-Label Classifier in Fig. 5. The results deliver more interpretation that text manipulation detection is harder than image modality and TA is the hardest case.

Visualization of manipulation detection and grounding. We provide some visualized results of manipulation detection and grounding in Fig. 6. Fig. 6 (a)-(b) show our method can accurately ground manipulated bboxes and detect correct manipulation types for both FA and FS. Furthermore, most of the manipulated text tokens in TS and all of those in TA are successfully grounded in Fig. 6 (c)-(d). All of them visually verify effective manipulation detection and grounding can be achieved by **HAMMER**.

Visualization of attention map. We provide Grad-CAM visualizations of our model regarding manipulated text tokens in Fig. 7. Fig. 7 (a) shows our model pays attention to surroundings of the character in image. These surroundings indicate the character is giving a speech, which is semantically distinct from text tokens manipulated by TS. As for TA, Fig. 7 (b) shows the per-word visualization with respect to the manipulated word ('mourn'). It implies our model focuses on the smiling face in image that is semantically inconsistent to the sad sentiment expressed from the manipulated word ('mourn'). These samples prove our model can indeed capture the semantic inconsistency between images and texts to tackle **DGM**^{4}.
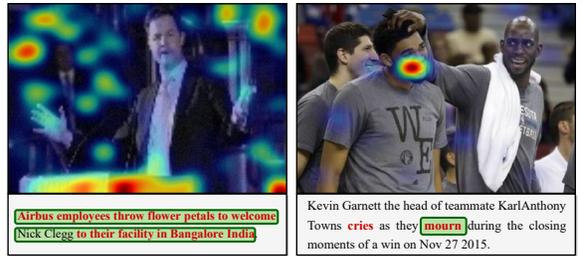
6. Conclusion

This paper studies a novel **DGM**^{4} problem, aiming to detect and ground multi-modal manipulations. We construct the first large-scale **DGM**^{4} dataset with rich annotations. A powerful model **HAMMER** is proposed and extensive experiments are performed to demonstrate its effectiveness.

Acknowledgements

References

[1] Sahar Abdelnabi, Rakibul Hasan, and Mario Fritz. Open-domain, content-based, multi-modal fact-checking of out-of-context images via online resources. In *CVPR*, 2022. 2, 3

[2] Shivangi Aneja, Chris Bregler, and Matthias Nießner. COSMOS: Catching Out-of-Context Misinformation with Self-Supervised Learning. In *ArXiv preprint arXiv:2101.06278*, 2021. 3

[3] Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. Simswap: An efficient framework for high fidelity face swapping. In *ACM MM*, 2020. 3

[4] Shen Chen, Taiping Yao, Yang Chen, Shouhong Ding, Jilin Li, and Rongrong Ji. Local relation learning for face forgery detection. In *AAAI*, 2021. 2

[5] Detection and visualization of misleading content on Twitter. Boididou, christina and papadopoulos, symeon and zampoglou, markos and apostolidis, lazaros and papadopoulou, olga and kompatsiaris, yiannis. *IJMIR*, 2018. 3

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. 2, 7

[7] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*, 2020. 2, 3

[8] Tarik Dzanic, Karan Shah, and Freddie Witherden. Fourier spectrum discrepancies in deep network generated images. *NeurIPS*, 2020. 3

[9] Gege Gao, Huaibo Huang, Chaoyou Fu, Zhaoyang Li, and Ran He. Information bottleneck disentanglement for identity swapping. In *CVPR*, 2021. 3

[10] Qiqi Gu, Shen Chen, Taiping Yao, Yang Chen, Shouhong Ding, and Ran Yi. Exploiting fine-grained face forgery clues via progressive enhancement learning. In *AAAI*, 2022. 3

[11] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 5

[12] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In *CVPR*, 2020. 2, 3

[13] Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *ACM MM*, 2017. 3

[14] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*, 2018. 3

[15] Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. Mvae: Multimodal variational autoencoder for fake news detection. In *WWW*, 2019. 3

[16] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, 2021. 2, 7

[17] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, 2021. 2, 4, 6

[18] Jiaming Li, Hongtao Xie, Jiahong Li, Zhongyuan Wang, and Yongdong Zhang. Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection. In *CVPR*, 2021. 3

[19] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *CVPR*, 2020. 2

[20] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-DF: A large-scale challenging dataset for deepfake forensics. In *CVPR*, 2020. 3

[21] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A new dataset for deepfake forensics. In *CVPR*, 2020. 2

[22] Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. Visual news: Benchmark and challenges in news image captioning. In *EMNLP*, 2021. 3

[23] Honggu Liu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu. Spatial-phase shallow learning: rethinking face forgery detection in frequency domain. In *CVPR*, 2021. 3

[24] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 3

[25] Grace Luo, Trevor Darrell, and Anna Rohrbach. NewsCLIPpings: Automatic Generation of Out-of-Context Multimodal Media. In *EMNLP*, 2021. 2, 3

[26] Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu. Generalizing face forgery detection with high-frequency features. In *CVPR*, 2021. 2, 7

[27] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 5

[28] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *ICCV*, 2021. 3

[29] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *ECCV*, 2020. 3

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2, 7

[31] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 2019. 2

[32] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP-IJCNLP*, 2019. 3

[33] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *CVPR*, 2019. 6

[34] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *ICCV*, 2019. 2, 3

[35] Rui Shao, Xiangyuan Lan, Jiawei Li, and Pong C. Yuen. Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In *CVPR*, 2019. 2

[36] Rui Shao, Xiangyuan Lan, and Pong C Yuen. Deep convolutional dynamic texture learning with adaptive channel-discriminability for 3d mask face anti-spoofing. In *IJCB*, 2017. 2

[37] Rui Shao, Xiangyuan Lan, and Pong C Yuen. Joint discriminative learning of deep dynamic textures for 3d mask face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 14(4):923–938, 2018. 2

[38] Rui Shao, Xiangyuan Lan, and Pong C. Yuen. Regularized fine-grained meta face anti-spoofing. In *AAAI*, 2020. 2

[39] Rui Shao, Pramuditha Perera, Pong C Yuen, and Vishal M Patel. Open-set adversarial defense. In *ECCV*, 2020. 2

[40] Rui Shao, Pramuditha Perera, Pong C Yuen, and Vishal M Patel. Federated generalized face presentation attack detection. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. 2

[41] Rui Shao, Pramuditha Perera, Pong C Yuen, and Vishal M Patel. Open-set adversarial defense with clean-adversarial mutual learning. *International Journal of Computer Vision*, 130(4):1070–1087, 2022. 2

[42] Rui Shao, Tianxing Wu, and Ziwei Liu. Detecting and recovering sequential deepfake manipulation. In *ECCV*, 2022. 2

[43] Rui Shao, Bochao Zhang, Pong C Yuen, and Vishal M Patel. Federated test-time adaptive face presentation attack detection with dual-phase privacy preservation. In *FG*, 2021. 2

[44] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 2017. 3

[45] Akhilesh Sudhakar, Bhargav Upadhyay, and Arjun Maheswaran. Transforming delete, retrieve, generate approach for controlled text style transfer. In *EMNLP-IJCNLP*, 2019. 4

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017. 4, 6

[47] Tengfei Wang, Yong Zhang, Yanbo Fan, Jue Wang, and Qifeng Chen. High-fidelity gan inversion for image attribute editing. In *CVPR*, 2022. 3

[48] William Yang Wang. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *ACL*, 2017. 2, 3

[49] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. Eann: Event adversarial neural networks for multi-modal fake news detection. In *KDD*, 2018. 3

[50] Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. LUKE: Deep contextualized entity representations with entity-aware self-attention. In *EMNLP*, 2020. 7

[51] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. In *CVPR*, 2022. 5

[52] Zitong Yu, Jun Wan, Yunxiao Qin, Xiaobai Li, Stan Z Li, and Guoying Zhao. Nas-fas: Static-dynamic central difference network search for face anti-spoofing. *IEEE transactions on pattern analysis and machine intelligence*, 43(9):3005–3023, 2020. 2

[53] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. Defending against neural fake news. In *NeurIPS*, 2019. 2, 3

[54] Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. In *ICML*, 2022. 6, 8

[55] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. In *CVPR*, 2021. 2, 7

[56] Tianchen Zhao, Xiang Xu, Mingze Xu, Hui Ding, Yuanjun Xiong, and Wei Xia. Learning self-consistency for deepfake detection. In *ICCV*, 2021. 2

[57] Xiangyu Zhu, Hao Wang, Hongyan Fei, Zhen Lei, and Stan Z Li. Face forgery detection by 3d decomposition. In *CVPR*, 2021. 2