Noise-assisted Prompt Learning for Image Forgery Detection and Localization

Dong Li†, Jiaying Zhu†, Xueyang Fu*, Xun Guo, Yidi Liu,
Gang Yang, Jiawei Liu, and Zheng-Jun Zha

School of Information Science and Technology and MoE Key Laboratory of
Brain-inspired Intelligent Perception and Cognition,
University of Science and Technology of China, Hefei, 230026, China
{dongli6, zhujy53, xunguo, liuyidi2023, yg1997}@mail.ustc.edu.cn
{xyfu, jwliu6, zhazj}@ustc.edu.cn

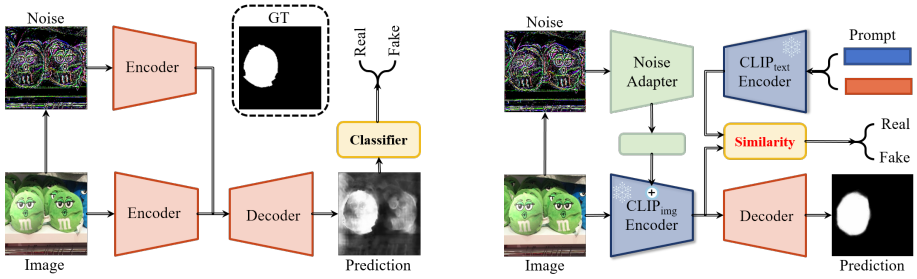
Abstract. We present CLIP-IFDL, a novel image forgery detection and localization (IFDL) model that harnesses the power of Contrastive Language Image Pre-Training (CLIP). However, directly incorporating CLIP in forgery detection poses challenges, given its lack of specific prompts and forgery consciousness. To overcome these challenges, we tailor the CLIP model for forgery detection and localization leveraging a noise-assisted prompt learning framework. This framework comprises instance-aware dual-stream prompt learning and a forgery-enhanced noise adapter. We initially create a pair of learnable prompts as negative-positive samples in place of discrete prompts, then fine-tune these prompts based on each image's features and categories. Additionally, we constrain the text-image similarity between the prompts and their corresponding images to update the prompts. Moreover, We design a forgery-enhanced noise adapter that augments the image encoder's forgery perceptual ability via multi-domain fusion and zero linear layers. By doing so, our method not only extracts pertinent features but also benefits from the generalizability of the open-world CLIP prior. Comprehensive tests indicate that our method outperforms existing ones in terms of accuracy and generalizability while effectively reducing false alarms.

Keywords: Image forgery detection and localization · CLIP · Prompt learning

1 Introduction

The rapid evolution of media technology and editing tools has made image manipulation increasingly commonplace. Risks associated with these manipulated images cut across various sectors, including copyright watermark removal, generation of fake news, and evidence falsification in court proceedings [35, 66, 67]. Consequently, the field of Image Forgery Detection and Localization (IFDL) has gained increased attention, aiming to determine if images have been altered and

† Co-first authors contributed equally, *Corresponding author.



(a) Previous methods. **(b)** Our proposed method.

Fig. 1: Comparison between previous methods and our proposed method. (a) Previous methods train all parameters and perform detection based on localization predictions. (b) Our method leverages priors brought by CLIP and utilizes prompt learning and noise adapters to perform detection and localization relatively independently.

identify the altered regions. However, with the fast-paced advancement of image forgery techniques, such as diffusion models [21,40,45,48,49], IFDL faces an ongoing challenge to keep pace with new forgery types. Meanwhile, false alarms on authentic images can also disrupt media distribution, leading to negative consequences. Thus, it is crucial to develop accurate and generalized IFDL methods.

Deep learning has significantly advanced Image Forgery Detection and Localization (IFDL). For example, RGB-N [73] uses noise features from a steganalysis-rich model filter layer to detect inconsistencies between genuine and tampered regions. MVSS-Net [7] leverages noise views and boundary artifacts to learn multi-view features, while ObjectFormer [55] detects subtle alterations from the high-frequency parts of images. However, these methods often underperform in real-world applications. While superior to traditional ones, learning-based methods may struggle with out-of-distribution detection, i.e., handling images tampered with in ways different from those in the training set. Furthermore, most advanced methods prioritize forgery localization, treating detection as a subsequent task [10, 19, 74] based on global integrity scores derived from forgery localization predictions [24, 47, 62], as shown in Figure 1. This approach often results in poor detection accuracy and high false alarm rates [19]. In a realistic setting, forged images are relatively rare [19], and the high rate of false alarms for real images can create more problems than the algorithms solve. Thus, there is a pressing need for methods that can accurately detect and localize forgeries while minimizing false alarms.

In this work, we investigate the potential of Contrastive Language-Image Pre-Training (CLIP) [43] to enhance IFDL. This is because CLIP has shown significant capabilities in zero-shot image recognition [43, 71] and the perception of abstract concepts [54]. Furthermore, we find that CLIP has the potential to discriminate between authentic and forged images (See supplementary material). We aim to leverage these extensive visual-language priors encapsulated in the CLIP model for forgery detection and localization, with the ultimate goal of

improving generalization and minimizing false alarms. However, directly applying CLIP to IFDL presents challenges. **First**, abstract semantic prompts like "fake" and "real" are difficult to accurately correlate with each image, making it tough to find a universal prompt capable of handling various types of forgeries. **Second**, the CLIP prior is derived from 400 million image-text pairs primarily composed of real images; meanwhile, CLIP appears to be insensitive to local object regions [27]. This results in the insufficient local perceptual capability of CLIP's image encoder for forgery.

To address the challenges of applying CLIP to IFDL, we propose a novel noise-assisted prompt learning framework named CLIP-IFDL. This framework addresses the challenges through instance-aware dual-stream prompt learning (IDPL) and forgery-enhanced noise adapter (FENA). For the prompt finding issue, IDPL first establishes a pair of learnable prompts as positive and negative sample pairs to replace discrete prompts, then adjusts the prompts in a doubly learnable manner based on the category and visual features of each image. On this basis, we update the prompts and conduct forgery detection by constraining the text-image similarity between the prompt pair and the corresponding images in the CLIP latent space. To tackle the issue of local forgery perception, we design the FENA, which enhances CLIP's perception of local forgeries through multi-domain fusion, zero-linear layers, and memory mechanisms. Through mutual enhancement, CLIP-IFDL achieves accurate and generalized image forgery detection and localization, effectively reducing false alarms on authentic images.

Our contributions are as follows:

- We introduce a novel method for image forgery detection and localization, CLIP-IFDL, leveraging the perceptual capability of CLIP.
- We propose instance-aware dual-stream prompt learning, finding accurate prompts to describe the authenticity-forgery attributes based on the category and visual features of each image.
- We develop a forgery-enhanced noise adapter to enhance the network's perception of local forgeries while avoiding catastrophic forgetting of CLIP priors caused by extensive fine-tuning.

Extensive experiments on several representative benchmarks demonstrate that our method surpasses state-of-the-art methods in terms of accuracy, generalization, and false alarm mitigation.

2 Related Works

2.1 Image Forgery Detection and Localization

Most early works propose to detect a specific type of forgery, including splicing [2,4,9,11,24,28,38,59,69], copy-move [8,15,25,57,60,61], and removal [1,58,64,75]. While these methods demonstrate satisfactory performance in detecting specific types of forgery, they exhibit limitations in practical applications due to the prevalence of unknown and diverse forgery types. Consequently, recent studies

have emphasized the need for an approach to tackle multiple forgery types with one model. RGB-N [73] introduces a dual-stream network, where one stream extracts RGB features to capture visual artifacts, and the other stream leverages noise features to model the inconsistencies between tampered and untouched regions for image forgery localization. ManTra-net [62] leverages an end-to-end network, which extracts image manipulation trace features and identifies anomalous regions by assessing how different a local feature is from its reference features. SPAN [23] attempts to model the spatial correlation via local self-attention blocks and pyramid propagation. MVSS-Net [7] has designed an edge-supervised branch that uses edge residual blocks to capture fine-grained boundary detail in a shallow to deep manner. ObjectFormer [55] captures forged traces from the high-frequency part of the image to attempt image forgery localization in the frequency domain. TruFor [19] outputs a reliability map to reduce false alarms and allow for a large scale analysis, which is important in forensic applications. ERMPC [31] proposes a two-step coarse-to-fine framework to explicitly model the inconsistency between the forged and authentic regions with edge information. In this work, we exploit the perceptual capabilities of CLIP and prompt learning to explore the potential of visual language priors for image forgery detection and localization, thereby improving performance.

2.2 CLIP Extensions and Prompting

CLIP [43] shows remarkable performance in zero-shot classification, thanks to the knowledge learned from 400 million carefully curated image-text pairs. Multiple derivative works across different sub-fields have emerged, such as object detection [29,65], image segmentation [37,46,70], image enhancement [34], image editing [42]. In addition to high-level semantic information, recent research [54] shows that the rich visual language priors encapsulated in CLIP can also be used to assess the quality and abstract perception of images in a zero-shot manner. These studies inspired us to exploit CLIP for image forgery detection and localization. Prompt engineering is popularized by the success of the GPT series [5, 44]. In the NLP domain, various prompt design approaches have been proposed recently, with one type focusing on prompt engineering by mining or generating proper discrete prompts [16, 26, 50]. Besides, continuous prompts circumvent the restriction from pre-trained language models and are adopted on NLP tasks [17, 30, 33]. For vision, CLIP finds the design of prompts matters for downstream tasks, so it improves the performance of visual classification by adding the prefix "a photo of" before the object name. Based on CLIP, CoOp [71] proposes Contextual Optimization, a method specifically designed to adapt CLIP-like visual language models for downstream image recognition. In contrast, our method accurately extracts abstract real-fake image representations through adaptive prompt learning for each image, instead of high-level semantic information in CLIP.

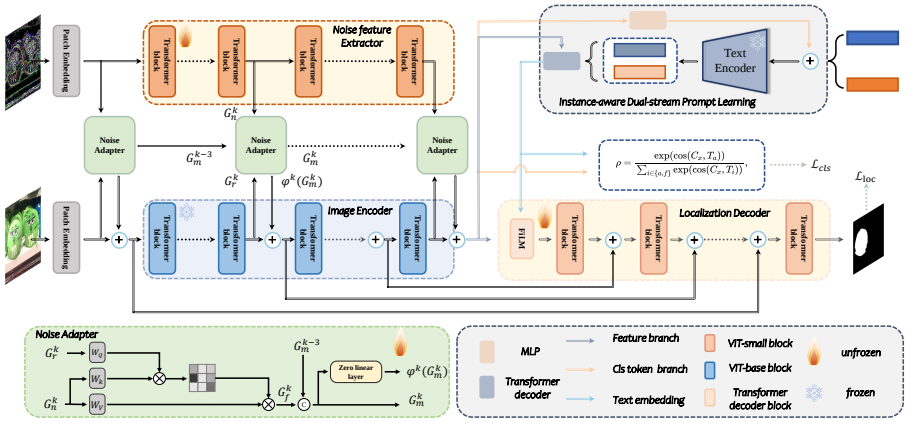


Fig. 2: An overview of the proposed framework CLIP-IFDL. The input is a suspicious image (H×W×3), and the output is a predicted mask (H×W×1), which localizes the forged regions. We perform forgery detection by calculating the text-image similarity between the prompt pair and the corresponding images in the CLIP latent space. The feature branch consists of the class token and the visual token. Instance-aware Dual-stream Prompt Learning is shown in Figure 3.

3 Methodology

3.1 Overview

Existing methods face the problems of poor generalization and high false alarm rate. CLIP demonstrates excellent zero-shot perception capabilities for distinguishing between authentic and forged image attributes, holding potential for addressing the aforementioned issues. Therefore, we propose CLIP-IFDL based on CLIP. Figure 2 is an overview of the framework. We freeze CLIP's image and text encoders to maintain priors. Building upon this, we devise Instance-aware Dual-stream Prompt Learning (IDPL) and Forgery-enhanced Noise Adapter (FENA) to leverage CLIP's potential in the field of IFDL. IDPL adaptively seeks suitable prompts for each image based on its category and visual features, addressing the challenge of accurately describing abstract forgery concepts with prompts. FENA aims to alleviate CLIP's insufficient perception of local forgeries. It explores the organic integration of adapters, cross-domain attention mechanisms, and memory mechanisms to incorporate noise containing forgery information into the frozen CLIP, thereby promoting forgery localization.

Formally, the input image is represented as $X \in \mathbb{R}^{H\times W\times 3}$, where H and W represent the height and width of the image. In practice, this CLIP image encoder uses the ViT-base [13] structure, which has 12 layers. The output features of each layer of the encoder are $\{G_r^1, G_r^2, \cdots, G_r^{12}\}$, and G_r^0 denotes the stem layer. The output of the final layer of the image encoder and the output of the text encoder are fed together into IDPL to perform forgery detection

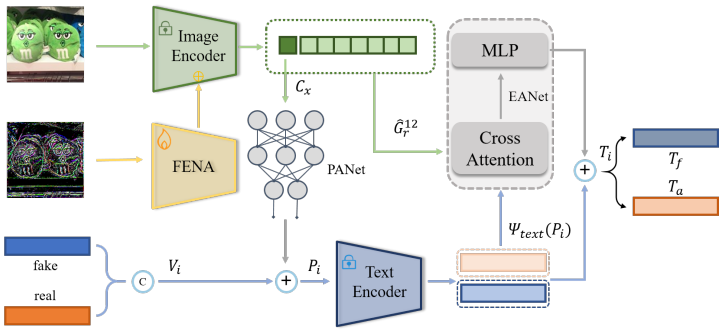


Fig. 3: Instance-aware Dual-stream Prompt Learning. FENA denotes the forgery-enhanced noise adapter.

while facilitating semantic representation of authenticity-forgery attributes. The proposed FENA introduces forgery information into the k-th layer of the image encoder to enhance the image feature, denoted as $\hat{G}_r^k$, through the noise feature extractor and the noise adapter. In our work, $k \in \{0, 3, 6, 9, 12\}$. Finally, the enhanced image features $\hat{G}_r^k$ and the text embedding output by IDPL are jointly fed into a classical decoder to output the predicted forgery localization map $G_{out} \in \mathbb{R}^{H \times W \times 1}$.

3.2 Instance-aware Dual-stream Prompt Learning

Unlike semantic segmentation and object detection, the authenticity-forgery attribute is not high-level semantic information but a relatively abstract concept, which makes using discrete prompts like CLIP [43] unable to obtain accurate results for IFDL. Furthermore, the forgery traces may be different for each image, therefore we propose Instance-aware Dual-stream Prompt Learning (IDPL), as shown in Figure 3.

First, we employ learnable vectors as prompts in a continuous space to represent authenticity-forgery attributes. Given an authentic image $X_a \in \mathbb{R}^{H \times W \times 3}$ and a forged image $X_f \in \mathbb{R}^{H \times W \times 3}$, we set two learnable vectors $V_a \in \mathbb{R}^{N \times 512}$ and $V_f \in \mathbb{R}^{N \times 512}$, respectively. N represents the number of embedded tokens in each prompt. It is worth noting that these two learnable vectors are not randomly initialized, but use words as the initialization vector, that is, V_a, and V_f are initialized by using the embedding of the two words "real" and "fake". Initiating the process with a thoughtfully curated set of discrete prompts simplifies the challenge of finding a precise vector within the continuous semantic space for the assessment of abstract concepts.

Then, we use dual-stream learning to adjust the prompts from coarse to fine. It contains a prompt adjustment network (PANet) and an embedding adjustment network (EANet). PANet adjusts the initial prompt based on the image category, which is written as:

$$P_i = V_i + \mathrm{PAN}\left(C_x\right), \tag{1}$$

where C_x is the CLS token of the feature $\hat{G}_r^{12}$ output by the last layer of the CLIP-based encoder for the input X, $P_i = \{P_a, P_f\}$ are the prompts after initial adjustment, and PAN denotes PANet. In our work, PANet is a two-layer bottleneck structure (Linear-ReLU-Linear), in which the hidden layer reduces the input dimension by 16 times. Then, P_i is fed into a frozen text encoder to obtain text embedding $\Psi_{text}(P_i)$, which contains CLIP prior. Next, EANet is used to adjust the text embedding. Specifically, we take the text embedding $\Psi_{text}(P_i)$ as the query and feed it together with the image features $\hat{G}_r^{12}$ into the EANet, composed of a transformer decoder, to search for visual clues related to authenticity-forgery attributes. Then, a residual connection is used to obtain T_i. The process is calculated as

$$T_i = \Psi_{text}(P_i) + \alpha \mathrm{EAN}\left(\Psi_{text}(P_i), \hat{G}_r^{12}\right), \tag{2}$$

where Ψ_{text} is the frozen text encoder, α is a learnable parameter and EAN is EANet. In practice, it is the Transformer decoder [53]. For α, we initialize with very small values (e.g., 10^{-3}) to prevent image features from overshadowing the prompt in the early training stages, avoiding the direct loss of prompt information. In short, IDPL not only improves prompts for each image in the category but also looks for relevant clues in image features to further adjust prompts. Meanwhile, this design also builds a bridge between the image encoder and the prompting, which is beneficial to the optimization of the entire framework.

After obtaining $T_i = \{T_a, T_f\}$, we calculate the image-text similarity ρ in the CLIP content space, as shown in Figure 2:

$$\rho = \frac{\exp(\cos(C_x, T_a))}{\sum_{i\in\{a,f\}} \exp(\cos(C_x, T_i))}, \tag{3}$$

where $\cos(\cdot,\cdot)$ denotes cosine similarity. Based on this, we then use the binary cross-entropy loss L_{cls} to distinguish authentic and forged images to optimize the learnable parameters of the prompts. The process can be formulated by:

$$\mathcal{L}_{cls} = -(y \cdot log(\rho)) + (1-y) \cdot log(1-\rho), \tag{4}$$

where y is the label of the current image. We assigned the label '1' to the authentic image and the label '0' to the forged image. This is to ensure that with the network's optimization, the distance between the authentic image X_a and the prompt T_a becomes closer.

4 Forgery-enhanced Noise Adapter

To maintain the open-world CLIP prior, we freeze the parameters of CLIP. However, the knowledge of CLIP comes from 400 million image-text pairs, primarily consisting of natural images. Additionally, CLIP seems to be insensitive to local object regions [27]. These factors result in the insufficient perception of local forgeries by the image encoder of CLIP, leading to inadequate pixel-level localization capability. Using noise information enables the discovery of tampering

traces that are nearly invisible in the RGB domain, thereby yielding strong forgery detection performance [3,7,31,56,62,73]. However, it is inappropriate to directly introduce noise information into the CLIP encoder due to the information gap between noise and RGB. Therefore we propose the Forgery-enhanced Noise Adapter (FENA), which consists of the noise feature extractor and the customized noise adapter, as shown in Figure 2.

First, following [7], we use BayarConv [62] to obtain the noise feature G_n. For the sake of parameter efficiency and alignment with the image features of CLIP, we use ViT-small [52] as the backbone of the noise feature extractor. It has twelve layers, and the output features of each layer are set to $\left\{G_n^1, G_n^2, \cdots, G_n^{12}\right\}$. In particular, we use G_n^0 to represent the feature before inputting the first layer, i.e., the feature of stem layer.

To reduce the computational cost, we do not feed noise information into each layer of the image encoder. Following [63], we select $\{0, 3, 6, 9, 12\}$ layers. We then incorporate the noise features into the image encoder using the noise adapter. It is worth noting that we do not directly add conditions to the frozen backbone network, which is different from the conventional adapter [39, 68]. Instead, we introduce a fusion strategy to bridge the gap between noise and RGB while exploring forgery information. In practice, we utilize the cross attention to perform the fusion of two information to obtain G_f^k:

$$G_f^k = \text{softmax}((W_q G_r^k)(W_k G_n^k)^T) W_v G_n^k, \tag{5}$$

where $k \in \{0, 3, 6, 9, 12\}$. W_q, W_k and W_v are all weight matrixes, and softmax is the softmax function. We explore the role of the fusion strategy for the adapter, and replacing it with a complex fusion module may have better performance. Furthermore, to facilitate the signal flow across iterative stages, we also introduce a simple persistent memory mechanism [6, 72] to augment information representation by leveraging memory in the fusion space. The process can be written as

$$G_m^k = \begin{cases} G_f^k & k = 0 \\ \psi(\text{Cat}(G_f^k, G_m^{k-3})) & k \in \{3, 6, 9, 12\} \end{cases}, \tag{6}$$

where ψ denotes the linear layer and Cat denotes the concatenation. Then, inspired by [68], we connect G_m^k and G_r^k with the zero linear layer φ^k, which is the linear layer with both weight and bias initialized to 0. It is computed as:

$$\hat{G}_r^k = G_r^k + \varphi^k\left(G_m^k\right), k \in \{0, 3, 6, 9, 12\} \tag{7}$$

where $\hat{G}_r^k$ is the K-layer feature of the image encoder that has been operated by the adapter. φ^k protects the backbone by eliminating the random noise used as a gradient in the initial training step. We explore the organic integration of adapters, cross-domain attention mechanisms, and memory mechanisms to design FENA. It efficiently enhances CLIP's sensitivity to local forgeries while avoiding the disruption of CLIP's priors. Therefore, FENA represents an effective approach for leveraging CLIP's potential in the field of image forgery localization, contributing significantly to the community.

4.1 Forgery Localization Decoder

Following [36, 37], we employ a U-Net shaped decoder for forgery localization. As shown in Figure 2, we feed the image features $\hat{G}_r^k, k \in \{0, 3, 6, 9, 12\}$ to the decoder before each transformer block. For network efficiency, we utilize a small decoder proposed by [37], comprising merely 1.12M parameters. In addition, to make full use of text information to enhance the network's forgery localization performance, we use Feature-wise Linear Modulation (FiLM) [14] to input the text features T_i to the decoder. FiLM applies feature-wise affine transformation to its input, allowing the modulation of image features by text features, which can be expressed as

$$\text{FiLM}(\hat{G}_r^{12}) = \gamma(T_i) \odot \hat{G}_r^{12} + \beta(T_i), \tag{8}$$

where γ and β are both linear layers, and $\odot$ is the Hadamard product. $\hat{G}_r^{12}$ is the image feature and T_i is the text feature after the prompt learning. Finally, with the help of multi-level image features and the text feature, the decoder can obtain a predicted forgery localization map $G_{out} \in \mathbb{R}^{H \times W \times 1}$.

4.2 Optimization

As shown in Figure 2, the loss function of our method consists of two components: detection loss and localization loss. We compute the forgery detection loss $\mathcal{L}_{cls}$ relying on the similarity between the image and the learned text embedding in the CLIP content space, as shown in Sec. 3.2. This is distinct from the majority of previous methods, which convert pixel-level localization predictions into binary detection results, introducing a higher risk of false alarms on authentic images [7]. The localization loss of our method is derived from the final prediction G_{out} and the ground-truth mask $Y \in \mathbb{R}^{H \times W \times 1}$. The overall loss function is written as:

$$\mathbf{L} = \lambda_1 \mathcal{L}_{loc}\left(Y, G_{out}\right) + \lambda_2 \mathcal{L}_{cls} \tag{9}$$

where $\mathcal{L}_{loc}$ denotes the Dice loss and λ_1, λ_2 are the parameters to balance the two terms in the loss function. It is worth noting that the detection loss and localization loss of our method are relatively independent, and there is no fixed sequential relationship between them, which alleviates false alarms on authentic images. However, they both contribute to optimizing the extraction of forgery features, ensuring their mutual promotion.

5 Experiments

5.1 Experimental Setup

Pre-training Data We create a sizable image tampering dataset and use it to pre-train our model. This dataset includes three categories: 1) splicing, 2) copy-move, and 3) removal. Details can be found in the supplementary material.
Testing Datasets Following [36, 55], we evaluate our model on CASIA [12], Coverage [57], Columbia [22], NIST16 [18] and IMD20 [41]. Specifically, IMD20 collects real-life manipulated images from Internet. We apply the same training/testing splits as [23, 55] to fine-tune our model for fair comparisons.

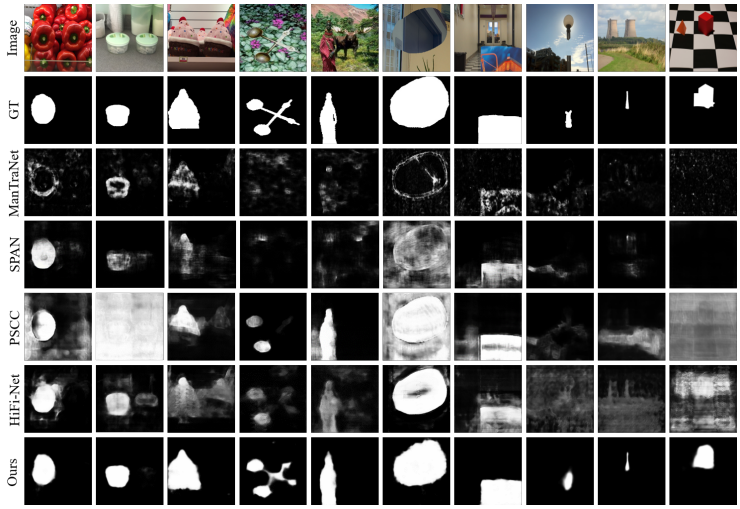


Fig. 4: Visualization of the predicted manipulation mask by different methods. From top to bottom, we show forged images, GT masks, and predictions of ManTraNet, SPAN, PSCC-Net, HiFi-Net and ours.

5.2 Image Forgery Localization

Following SPAN [23] and ObjectFormer [55], our model is compared with other state-of-the-art methods under two settings: 1) training on the synthetic dataset and evaluating on the full test datasets, and 2) fine-tuning the pre-trained model on the training split of test datasets and evaluating on their test split. It's worth noting that in both stages, the main CLIP model remains frozen.

Pre-trained Model Table 1a shows the localization performance of pre-trained models for different methods on five datasets under pixel-level AUC. We compare our CLIP-IFDL with MantraNet [62], SPAN [23], PSCCNet [36], ObjectFormer [55], HiFi-Net [20] and SAFL-Net [51] when evaluating pre-trained models.

The pre-trained CLIP-IFDL achieves the best localization performance on Coverage, CASIA, NIST16 and IMD20 and ranks third on Columbia. In partic-

Table 1: Image forgery detection and localization results. (a) Localization performance of the pre-train model. (b) Localization performance of the fine-tuned model. (c) Detection performance on *CASIA-D* dataset. [Key: **Best**; Second Best].

Localization	Data	Columbia	Coverage	CASIA	NIST16	IMD20
		Metric: AUC(%) – Pre-trained				
ManTraNet	64K	82.4	81.9	81.7	79.5	74.8
SPAN	96K	93.6	92.2	79.7	84.0	75.0
PSCCNet	100K	98.2	84.7	82.9	85.5	80.6
ObjectFormer	62K	95.5	92.8	84.3	87.2	82.1
HiFi-Net	100K	**98.3**	93.2	85.8	87.0	82.9
SAFL-Net	48K	96.9	93.5	90.9	88.8	96.5
Ours	60K	97.6	**94.3**	**92.5**	**89.7**	**97.8**

(a)

Localization	Coverage	CASIA	NIST16
	Metric: AUC(%) / F1(%) – Fine-tuned		
RGB-N	81.7/43.7	79.5/40.8	93.7/72.2
SPAN	93.7/55.8	83.8/38.2	96.1/58.2
PSCCNet	94.1/72.3	87.5/55.4	99.6/81.9
ObjectFormer	95.7/75.8	88.2/57.9	99.6/82.4
HiFi-Net	96.1/80.1	88.5/61.6	98.9/85.0
SAFL-Net	97.0/80.3	90.8/74.0	99.7/87.9
Ours	**98.2**/**81.3**	**91.3**/**77.9**	**99.8**/**89.3**

(b)

Detection	AUC(%)	F1(%)
ManTraNet	59.94	56.69
SPAN	67.33	63.48
PSCC	99.65	97.12
ObjectFormer	99.70	97.34
HiFi-Net	99.50	97.40
SAFL-Net	99.50	98.40
Ours	**99.83**	**98.75**

(c)

Table 2: Comparison of detection results on challenging datasets.

Method	COVER				IMD20			
	AUC	Spe.	Sen.	F1	AUC	Spe.	Sen.	F1
ManTraNet	0.500	0.0	100.0	0.0	0.500	0.0	100.0	0.0
SPAN	0.500	0.0	100.0	0.0	0.500	0.0	100.0	0.0
PSCC	0.658	91.0	19.0	31.4	0.631	92.0	20.9	34.1
ObjectFormer	0.534	33.0	57.0	41.8	0.510	30.2	56.2	39.3
HiFi-Net	0.513	93.0	7.0	13.0	0.469	97.8	1.4	2.8
SAFL-Net	0.653	53.0	47.0	49.8	0.612	54.3	50.0	52.1
Ours	0.732	70.0	61.0	65.2	0.678	73.9	68.7	71.2

ular, CLIP-IFDL achieves a 97.8 % performance on the IMD20, which consists of real-life images. It indicates that our method not only possesses the superior ability to capture tampering traces but also generalizes well to realistic scenarios.

We fail to achieve the best performance on Columbia, falling behind HiFi-Net 0.7 % and under AUC. We contend that the explanation may be that the distribution of their synthesized training data closely resembles that of the Columbia dataset. This is further supported by the results in Table 1b, which show that CLIP-IFDL performs better than HiFi-Net in terms of both AUC and F1 scores.

Fine-tuned Model The network weights of the pretrained model are used to initiate the fine-tuned models that will be trained on the training split of Coverage, CASIA, and NIST16 datasets, respectively. We evaluate the fine-tuned models of different methods in Table 1b. As for AUC and F1, our model achieves significant performance gains. This validates that our method can precisely capture various subtle manipulation traces through instance-aware dual-stream prompt learning and forgery-enhanced noise adapter.

5.3 Image Forgery Detection

To demonstrate the image-level discrimination capability of the network, we also consider the forgery detection task. Following ObjectFormer [55], we conduct experimental comparisons on the CASIA-D dataset introduced by [36]. As shown in Table 1c, our method has excellent detection performance. To further measure the miss detection rate and false alarm rate, we conduct additional comparisons on challenging datasets such as Coverage [57] and IMD20 [41] in Table 2. Spe denotes specificity, where higher values imply fewer false alarms for authentic images, while Sen stands for sensitivity, indicating fewer missed forgeries detection. Our method once again ranks among the top performers, obtaining high specificity while ensuring F1. This indicates that CLIP-IFDL effectively mitigates false alarms, leveraging a relatively independent detection paradigm along with CLIP priors. Our method also performs well in AUC, demonstrating its capability to accurately distinguish between forged and genuine images.

Table 3: The performance on NIST16 under various distortions. AUC scores are reported (in %), (Blur: GaussianBlur, Noise: GaussianNoise, Compress: JPEGCompress.)

Distortion	SPAN	ObjectFormer	HiFi-Net	SAFL-Net	Ours	
no distortion	83.95	87.18	87.0	88.79	**89.68**	
Resize(0.78×)	83.24	87.17	86.9	88.39	**89.31**	↓0.37
Resize(0.25×)	80.32	86.33	86.5	86.92	**87.93**	↓1.75
Blur($k = 3$)	83.10	85.97	86.1	88.13	**89.14**	↓0.54
Blur($k = 15$)	79.15	80.26	81.0	87.68	**88.56**	↓1.12
Noise($\sigma = 3$)	75.17	79.58	81.9	/	**89.13**	↓0.45
Noise($\sigma = 15$)	67.28	78.15	79.5	/	**86.75**	↓2.83
Compress($q = 100$)	83.59	86.37	86.5	88.56	**89.58**	↓0.10
Compress($q = 50$)	80.68	86.24	86.0	88.07	**88.93**	↓0.65

5.4 Robustness Evaluation

To analyze the robustness of our model for localization, we follow the distortion settings in [55] to degrade the raw forged images from NIST16. These distortion types include resizing images to different scales (Resize), applying Gaussian blur with a kernel size k (GaussianBlur), adding Gaussian noise with a standard deviation σ (GaussianNoise), and performing JPEG compression with a quality factor q (JPEGCompress). We compare the forgery localization performance (AUC scores) of our pre-trained models with SPAN and ObjectFormer on these corrupted data, and report the results in Table 3. Our model demonstrates better robustness against various distortion techniques. It is worth noting that JPEG compression is commonly performed when uploading images to social media. And our model performs significantly better on compressed images.

5.5 Ablation Study

In this section, we conduct experiments to demonstrate the effectiveness of our method. The instance-aware dual-stream prompt learning (IDPL) is designed to find accurate prompts describing the authenticity-forgery attributes of each image thus facilitating the utilization of CLIP priors. It consists of the prompt adjustment network (PANet) and the embedding adjustment network (EANet). PANet adaptively adjusts the prompts according to the category of each image, while EANet looks for relevant clues for authenticity-forgery attributes in image features to adjust the embedding of the prompt. The forgery-enhanced noise adapter (FENA) is designed to enhance CLIP's perception of forgery.

Table 4: Ablation results using different variants of our scheme.

Variants	CASIA		NIST16	
	AUC	F1	AUC	F1
baseline	76.2	45.7	81.3	70.5
w/o PANet	83.3	50.8	84.4	74.2
w/o EANet	84.2	51.1	88.9	77.6
w/o FENA	88.6	57.1	94.9	83.2
Ours	**91.3**	**77.9**	**99.8**	**89.3**

Table 5: The effect of different unfreezing strategies.

Unfreezing	CASIA	NIST16	IMD20
a	83.1	80.3	75.4
b	83.4	82.6	86.7
c	87.3	83.4	90.1
Ours	**92.5**	**89.7**	**97.8**

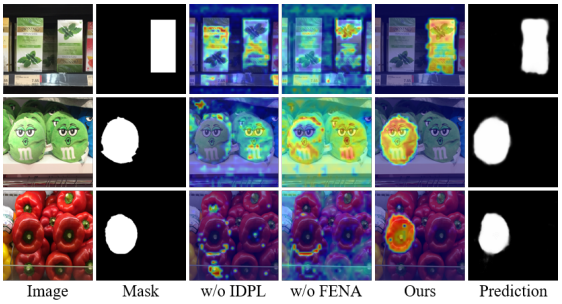


Fig. 5: Visualization of IDPL and FENA. From left to right, we display the forged images, masks, GradCAM of the feature map without (w/o) IDPL and without FENA and with both, and predictions.

To evaluate the effectiveness of PANet, EANet and FENA, we remove them separately from our method and evaluate the forgery localization performance as shown in Table 4. The baseline denotes that we solely employ CLIP and forgery localization decoder, with two learnable vectors as prompts. It can be seen that without PANet, the AUC scores decrease by 8.70 % on CASIA and 15.43 % on NIST16, while without EANet, the AUC scores decrease by 7.77 % on CASIA and 10.92 % on NIST16. Besides, when FENA is discarded, performance degradation is also observed in Table 4, i.e., 4.91 % in terms of AUC and 6.83 % in terms of F1 on NIST16. The performance drop by removing PANet or EANet is more pronounced compared to removing FENA, indicating the significance of instance-aware prompt learning in facilitating localization tasks.

In Table 5, we compare the results of various unfreezing strategies on three datasets to validate the contribution of the prior brought by the frozen CLIP to the generalization performance. "Ours" represents the pre-trained CLIP-IFDL model to reflect generalization. "a" denotes unfreezing all CLIP parameters, "b" denotes unfreezing CLIP's image encoder, and "c" denotes unfreezing CLIP's text encoder. It can be observed that unfreezing all parameters leads to forgetting the CLIP prior, resulting in poorer generalization. Moreover, unfreezing either the image or text encoder degrades the network's performance on the real-life dataset IMD20. This indicates that the prior from frozen CLIP contributes to the network's generalization to realistic scenarios.

5.6 Visualization Results

Qualitative results As shown in Figure 4, we provide predicted masks of various methods. Since the source codes of ObjectFormer [55] and SAFL-Net [51] are not available, their predictions are not available. The results demonstrate that our method could not only locate the tampering regions more accurately but also develop sharp boundaries. It benefits from the ability of our model to effectively distinguish between the two regions with the help of CLIP prior and noise adapter.

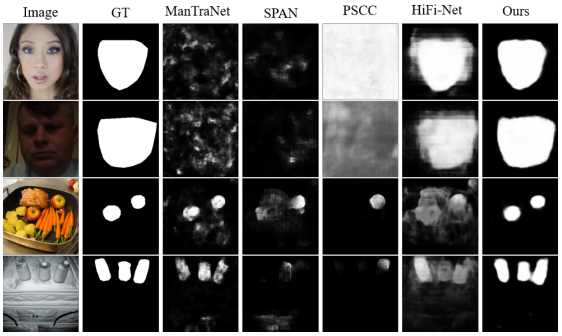


Fig. 6: More visualization results on Faceshifter and CocoGlide.

Visualization of IDPL To verify the effect of the instance-aware dual-stream prompt learning (IDPL), we show the change of features with and without the prompt learning in Figure 5. It can be seen that IDPL can improve the accuracy of forgery localization. The network without IDPL will make false judgments about objects that are similar to the forgery.
Visualization of FENA We show the change of features with and without the forgery-enhanced noise adapter (FENA) in Figure 5. It is clear that FENA facilitates the learning of forgery features and obtains more accurate contours of forged regions.
Visualization on other challenging datasets To further validate the robust generalization of our method, our models are also compared visually on two challenging out-of-distribution datasets. These two datasets are face tampering images Faceshifter [32] and diffusion-based tampering images CocoGlide [19], respectively, and both of them have very different types of forgery than our training set. As shown in Figure 6, our method still achieves the best visualization results. This fully illustrates the strong generalization of our method to detect unseen forgeries.

6 Conclusion

In this paper, we propose a novel paradigm for image forgery detection and localization CLIP-IFDL, by leveraging the potential of CLIP. This method not only leverages the open-world CLIP prior to distinguish between forged and authentic images but also identifies forged regions, enhancing generalization and reducing false alarms. Initially, we create a learnable prompt pair, updating it by aligning the text-image similarity in the CLIP latent space. Moreover, we design a forgery-enhanced noise adapter that enhances the perceptual ability of the image encoder for forgery. To our knowledge, this is the first attempt to utilize prompt learning and the CLIP prior for IFDL. Extensive experiments on several representative benchmarks demonstrate that our method outperforms existing methods in terms of accuracy, generalization, and false alarm mitigation.

Acknowledgements. This work was supported by National Natural Science Foundation of China (NSFC) under Grants 62225207, 62276243, and 62106245.

References

1. Aloraini, M., Sharifzadeh, M., Schonfeld, D.: Sequential and patch analyses for object removal video forgery detection and localization. IEEE Transactions on Circuits and Systems for Video Technology **31**(3), 917–930 (2020)
2. Amerini, I., Uricchio, T., Ballan, L., Caldelli, R.: Localization of jpeg double compression through multi-domain convolutional neural networks. In: 2017 IEEE Conference on computer vision and pattern recognition workshops (CVPRW). pp. 1865–1871. IEEE (2017)
3. Bayar, B., Stamm, M.C.: Constrained convolutional neural networks: A new approach towards general purpose image manipulation detection. IEEE Transactions on Information Forensics and Security **13**(11), 2691–2706 (2018)
4. Bondi, L., Lameri, S., Guera, D., Bestagini, P., Delp, E.J., Tubaro, S., et al.: Tampering detection and localization through clustering of camera-based cnn features. In: CVPR Workshops. vol. 2 (2017)
5. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
6. Chen, J., Sun, Y., Liu, Q., Huang, R.: Learning memory augmented cascading network for compressed sensing of images. In: European Conference on Computer Vision. pp. 513–529. Springer (2020)
7. Chen, X., Dong, C., Ji, J., Cao, J., Li, X.: Image manipulation detection by multi-view multi-scale supervision. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14185–14193 (2021)
8. Cozzolino, D., Poggi, G., Verdoliva, L.: Efficient dense-field copy–move forgery detection. IEEE Transactions on Information Forensics and Security **10**(11), 2284–2297 (2015)
9. Cozzolino, D., Poggi, G., Verdoliva, L.: Splicebuster: A new blind image splicing detector. In: 2015 IEEE International Workshop on Information Forensics and Security (WIFS). pp. 1–6. IEEE (2015)
10. Cozzolino, D., Poggi, G., Verdoliva, L.: Data-driven digital integrity verification. In: Multimedia Forensics, pp. 281–311. Springer Singapore Singapore (2022)
11. Cozzolino, D., Verdoliva, L.: Noiseprint: a cnn-based camera model fingerprint. IEEE Transactions on Information Forensics and Security **15**, 144–159 (2019)
12. Dong, J., Wang, W., Tan, T.: Casia image tampering detection evaluation database. In: 2013 IEEE China summit and international conference on signal and information processing. pp. 422–426. IEEE (2013)
13. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
14. Dumoulin, V., Perez, E., Schucher, N., Strub, F., Vries, H.d., Courville, A., Bengio, Y.: Feature-wise transformations. Distill **3**(7), e11 (2018)
15. D'Amiano, L., Cozzolino, D., Poggi, G., Verdoliva, L.: A patchmatch-based dense-field algorithm for video copy–move detection and localization. IEEE Transactions on Circuits and Systems for Video Technology **29**(3), 669–682 (2018)

16. Gao, T., Fisch, A., Chen, D.: Making pre-trained language models better few-shot learners. arXiv preprint arXiv:2012.15723 (2020)
17. Gu, Y., Han, X., Liu, Z., Huang, M.: Ppt: Pre-trained prompt tuning for few-shot learning. arXiv preprint arXiv:2109.04332 (2021)
18. Guan, H., Kozak, M., Robertson, E., Lee, Y., Yates, A.N., Delgado, A., Zhou, D., Kheyrkhah, T., Smith, J., Fiscus, J.: Mfc datasets: Large-scale benchmark datasets for media forensic challenge evaluation. In: 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW). pp. 63–72. IEEE (2019)
19. Guillaro, F., Cozzolino, D., Sud, A., Dufour, N., Verdoliva, L.: Trufor: Leveraging all-round clues for trustworthy image forgery detection and localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20606–20615 (2023)
20. Guo, X., Liu, X., Ren, Z., Grosz, S., Masi, I., Liu, X.: Hierarchical fine-grained image forgery detection and localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3155–3165 (June 2023)
21. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
22. Hsu, Y.F., Chang, S.F.: Detecting image splicing using geometry invariants and camera characteristics consistency. In: 2006 IEEE International Conference on Multimedia and Expo. pp. 549–552. IEEE (2006)
23. Hu, X., Zhang, Z., Jiang, Z., Chaudhuri, S., Yang, Z., Nevatia, R.: Span: Spatial pyramid attention network for image manipulation localization. In: European conference on computer vision. pp. 312–328. Springer (2020)
24. Huh, M., Liu, A., Owens, A., Efros, A.A.: Fighting fake news: Image splice detection via learned self-consistency. In: Proceedings of the European conference on computer vision (ECCV). pp. 101–117 (2018)
25. Islam, A., Long, C., Basharat, A., Hoogs, A.: Doa-gan: Dual-order attentive generative adversarial network for image copy-move forgery detection and localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4676–4685 (2020)
26. Jiang, Z., Xu, F.F., Araki, J., Neubig, G.: How can we know what language models know? Transactions of the Association for Computational Linguistics **8**, 423–438 (2020)
27. Jiao, S., Wei, Y., Wang, Y., Zhao, Y., Shi, H.: Learning mask-aware clip representations for zero-shot segmentation. Advances in Neural Information Processing Systems **36** (2024)
28. Kniaz, V.V., Knyaz, V., Remondino, F.: The point where reality meets fantasy: Mixed adversarial generators for image splice detection. Advances in Neural Information Processing Systems **32** (2019)
29. Kuo, W., Cui, Y., Gu, X., Piergiovanni, A., Angelova, A.: F-vlm: Open-vocabulary object detection upon frozen vision and language models. arXiv preprint arXiv:2209.15639 (2022)
30. Lester, B., Al-Rfou, R., Constant, N.: The power of scale for parameter-efficient prompt tuning. arXiv preprint arXiv:2104.08691 (2021)
31. Li, D., Zhu, J., Wang, M., Liu, J., Fu, X., Zha, Z.J.: Edge-aware regional message passing controller for image forgery localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8222–8232 (2023)
32. Li, L., Bao, J., Yang, H., Chen, D., Wen, F.: Faceshifter: Towards high fidelity and occlusion aware face swapping. arXiv preprint arXiv:1912.13457 (2019)

33. Li, X.L., Liang, P.: Prefix-tuning: Optimizing continuous prompts for generation. arXiv preprint arXiv:2101.00190 (2021)
34. Liang, Z., Li, C., Zhou, S., Feng, R., Loy, C.C.: Iterative prompt learning for unsupervised backlit image enhancement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8094–8103 (2023)
35. Lin, X., Wang, S., Deng, J., Fu, Y., Bai, X., Chen, X., Qu, X., Tang, W.: Image manipulation detection by multiple tampering traces and edge artifact enhancement. Pattern Recognition **133**, 109026 (2023)
36. Liu, X., Liu, Y., Chen, J., Liu, X.: Pscc-net: Progressive spatio-channel correlation network for image manipulation detection and localization. IEEE Transactions on Circuits and Systems for Video Technology (2022)
37. Lüddecke, T., Ecker, A.: Image segmentation using text and image prompts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7086–7096 (2022)
38. Lyu, S., Pan, X., Zhang, X.: Exposing region splicing forgeries with blind local noise estimation. International journal of computer vision **110**(2), 202–221 (2014)
39. Mou, C., Wang, X., Xie, L., Zhang, J., Qi, Z., Shan, Y., Qie, X.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. arXiv preprint arXiv:2302.08453 (2023)
40. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021)
41. Novozamsky, A., Mahdian, B., Saic, S.: Imd2020: A large-scale annotated dataset tailored for detecting manipulated images. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops. pp. 71–80 (2020)
42. Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., Lischinski, D.: Styleclip: Text-driven manipulation of stylegan imagery. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2085–2094 (2021)
43. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
44. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
45. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents, 2022. URL https://arxiv.org/abs/2204.06125 **7** (2022)
46. Rao, Y., Zhao, W., Chen, G., Tang, Y., Zhu, Z., Huang, G., Zhou, J., Lu, J.: Denseclip: Language-guided dense prediction with context-aware prompting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18082–18091 (2022)
47. Rao, Y., Ni, J., Xie, H.: Multi-semantic crf-based attention model for image forgery detection and localization. Signal Processing **183**, 108051 (2021)
48. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
49. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems **35**, 36479–36494 (2022)

50. Shin, T., Razeghi, Y., Logan IV, R.L., Wallace, E., Singh, S.: Autoprompt: Eliciting knowledge from language models with automatically generated prompts. arXiv preprint arXiv:2010.15980 (2020)
51. Sun, Z., Jiang, H., Wang, D., Li, X., Cao, J.: Safl-net: Semantic-agnostic feature learning network with auxiliary plugins for image manipulation detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22424–22433 (2023)
52. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International conference on machine learning. pp. 10347–10357. PMLR (2021)
53. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
54. Wang, J., Chan, K.C., Loy, C.C.: Exploring clip for assessing the look and feel of images. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 2555–2563 (2023)
55. Wang, J., Wu, Z., Chen, J., Han, X., Shrivastava, A., Lim, S.N., Jiang, Y.G.: Objectformer for image manipulation detection and localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2364–2373 (2022)
56. Wang, T., Chow, K.P.: Noise based deepfake detection via multi-head relative-interaction. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 14548–14556 (2023)
57. Wen, B., Zhu, Y., Subramanian, R., Ng, T.T., Shen, X., Winkler, S.: Coverage—a novel database for copy-move forgery detection. In: 2016 IEEE international conference on image processing (ICIP). pp. 161–165. IEEE (2016)
58. Wu, H., Zhou, J.: Iid-net: Image inpainting detection network via neural architecture search and attention. IEEE Transactions on Circuits and Systems for Video Technology **32**(3), 1172–1185 (2021)
59. Wu, Y., Abd-Almageed, W., Natarajan, P.: Deep matching and validation network: An end-to-end solution to constrained image splicing localization and detection. In: Proceedings of the 25th ACM international conference on Multimedia. pp. 1480–1502 (2017)
60. Wu, Y., Abd-Almageed, W., Natarajan, P.: Busternet: Detecting copy-move image forgery with source/target localization. In: Proceedings of the European conference on computer vision (ECCV). pp. 168–184 (2018)
61. Wu, Y., Abd-Almageed, W., Natarajan, P.: Image copy-move forgery detection via an end-to-end deep neural network. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1907–1915. IEEE (2018)
62. Wu, Y., AbdAlmageed, W., Natarajan, P.: Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9543–9552 (2019)
63. Xu, M., Zhang, Z., Wei, F., Hu, H., Bai, X.: Side adapter network for open-vocabulary semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2945–2954 (2023)
64. Yang, Q., Yu, D., Zhang, Z., Yao, Y., Chen, L.: Spatiotemporal trident networks: detection and localization of object removal tampering in video passive forensics. IEEE Transactions on Circuits and Systems for Video Technology **31**(10), 4131–4144 (2020)

65. Zang, Y., Li, W., Zhou, K., Huang, C., Loy, C.C.: Open-vocabulary detr with conditional matching. In: European Conference on Computer Vision. pp. 106–122. Springer (2022)
66. Zhang, F., Liu, J., Xie, J., Zhang, Q., Xu, Y., Zha, Z.J.: Escnet: Entity-enhanced and stance checking network for multi-modal fact-checking. In: Proceedings of the ACM on Web Conference 2024. pp. 2429–2440 (2024)
67. Zhang, F., Liu, J., Zhang, Q., Sun, E., Xie, J., Zha, Z.J.: Ecenet: Explainable and context-enhanced network for muti-modal fact verification. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 1231–1240 (2023)
68. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023)
69. Zhang, Y., Zhu, G., Wu, L., Kwong, S., Zhang, H., Zhou, Y.: Multi-task se-network for image splicing localization. IEEE Transactions on Circuits and Systems for Video Technology (2021)
70. Zhou, C., Loy, C.C., Dai, B.: Extract free dense labels from clip. In: European Conference on Computer Vision. pp. 696–712. Springer (2022)
71. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (2022)
72. Zhou, M., Yan, K., Pan, J., Ren, W., Xie, Q., Cao, X.: Memory-augmented deep unfolding network for guided image super-resolution. International Journal of Computer Vision **131**(1), 215–242 (2023)
73. Zhou, P., Han, X., Morariu, V.I., Davis, L.S.: Learning rich features for image manipulation detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1053–1061 (2018)
74. Zhu, J., Li, D., Fu, X., Yang, G., Huang, J., Liu, A., Zha, Z.J.: Learning discriminative noise guidance for image forgery detection and localization. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 7739–7747 (2024)
75. Zhu, X., Qian, Y., Zhao, X., Sun, B., Sun, Y.: A deep learning approach to patch-based image inpainting forensics. Signal Processing: Image Communication **67**, 90–99 (2018)