SCAM: Systematic Classification and Analysis of Misinformation

Iyyakutti Iyappan Ganapathi
iyyakutti.ganapathi@ku.ac.ae
Syed Sadaf Ali
syed.ali@ku.ac.ae
Maregu Assefa
maregu.habtie@ku.ac.ae
Sajid Javed
sajid.javed@ku.ac.ae
Naoufel Werghi
naoufel.werghi@ku.ac.ae

C2PS & KUCARS, Khalifa University, Abu Dhabi, UAE

Abstract

In recent years, image manipulation has become a significant threat due to the widespread availability of techniques that can be easily accessed by everyday users, including those with just a mobile phone. Such manipulations can have a profound impact on society, as altered visuals can easily influence public perception. Common forms of image forgery, such as copy-move and splicing, are now joined by more advanced methods involving generative networks and other sophisticated techniques, making detection increasingly challenging. Traditional approaches are often trained on specific types of forgeries, raising concerns about their ability to generalize to new and unseen manipulation techniques. In response to these challenges, this paper introduces a training-free approach to detecting and localizing manipulated images. Our framework leverages the capabilities of large language models (LLMs) to perform visual analysis through natural language descriptions. For each input image, a detailed text description is generated using a visual language model (VLM), which is then further analyzed by an LLM to assess the context and identify any manipulated objects. This structured analysis generates key information, including the detection of manipulation, the identification of the manipulated object, and the confidence level of the evaluation. Our approach has been tested on six different datasets and demonstrates performance that is comparable to existing state-of-the-art techniques that rely on extensive training with diverse datasets. By eliminating the need for training, our method offers a scalable and flexible solution to the increasingly sophisticated challenge of image manipulation detection.

1 Introduction

The widespread use of digital image manipulation has increasingly challenged the trustworthiness of visual content across various domains, including media, marketing, and forensic

investigations. Content manipulation poses a significant threat to society, especially as it spreads rapidly through social networks. Among different types of content, image, text, and audio, images are particularly vulnerable due to the ease with which visual information can be understood [2, 14, 30]. Real-time applications, such as Fake Image Detector [8], FotoForensics [12], and Ghiro [15], have been developed to detect image manipulations. However, most of these techniques rely on error level analysis (ELA) [11, 35], which is effective primarily against low-level forgeries but struggles with high-resolution, retouched, and GAN-generated images. Moreover, ELA operates by analyzing the differences between two compression levels and is mostly applicable to JPEG-compressed images.

Copy-move and splicing are among the most common types of image forgeries [13, 18, 33]. Additionally, other widely used editing techniques, such as cropping, resizing, rotating, cloning, blurring, sharpening, and color adjustment, should also be considered when detecting manipulation [38]. Many existing techniques employ dual-branch networks [1, 17, 26, 36] to detect and localize these manipulations. These dual-branch architectures combine spatial and frequency domain features and have been shown to be effective in detecting and localizing image manipulations. However, they come with certain limitations, often requiring extensive and diverse datasets to achieve good generalization. The process of collecting and labeling datasets that encompass a broad spectrum of forgery types and scenarios is time-consuming and costly. Moreover, models trained on specific types of manipulations may become overly specialized, making them less effective when confronted with new or evolved forgery methods [3]. In real-world scenarios, forgeries can vary significantly, necessitating models that are more generalizable to unseen data. This implies that models should either be unsupervised or trained on a vast array of manipulations to develop a comprehensive understanding of existing techniques. However, achieving this level of generalization is challenging and may not be feasible.

Recent advances in natural language processing (NLP) and the development of large language models (LLMs) have opened new avenues for cross-modal analysis, enabling the effective integration of textual descriptions and visual data. These models have demonstrated remarkable capabilities in generating detailed textual descriptions from images and in understanding and analyzing complex visual scenes through natural language [5, 34, 39].

Building on these advancements, we propose a framework for detecting and localizing manipulated images without the need for additional training. Our approach leverages the descriptive and analytical power of foundation models to perform image analysis. The framework begins by passing an input image through a visual language model (VLM) to generate a comprehensive description. These descriptions are then processed by an LLM to produce structured JSON data, which includes key indicators such as *detected manipulation* (yes or no), the identity of the *manipulated object*, and an *explainable decision* associated with the manipulation assessment. This approach offers several advantages over traditional approaches, including scalability, flexibility, and transparency. Our framework is independent of specific forgery types and can be easily adapted to various forms of image manipulation, making it a versatile tool for image authentication and forensic analysis. Additionally, the use of LLMs enables the generation of interpretable outputs that can be seamlessly integrated into existing workflows, thereby enhancing the overall robustness of image content verification processes. We do not train our proposed framework on any datasets. The pretrained knowledge gained by the LLMs and the vision foundation models is leveraged in the proposed framework to detect the forgery. The proposed approach has demonstrated performance comparable to that of state-of-the-art techniques.

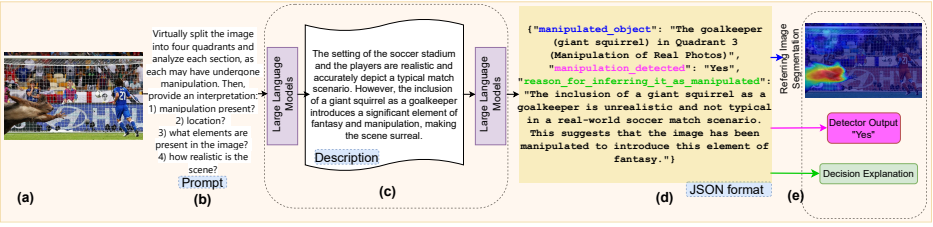

Figure 1: Overview of the proposed framework for detecting and localizing image manipulations: (a) Input: Forged image, (b) Initial Setup: Creation of a prompt script, (c) Analysis Phase: The script is processed by a visual language model (VLM) to produce a detailed description of potential manipulations. This description is then fed to a LLM, which generates a JSON format response detailing key aspects such as detection status, the presence and type of manipulation, and the reasons for such inferences, (d) Localization: The extracted key values are used to localize the areas of forgery in the original image.

2 Related works

Foundational models, particularly those that combine vision and language, have become cornerstones in modern AI research and applications [20, 22, 25, 34]. These models, built on large-scale datasets and advanced architectures, have demonstrated remarkable robustness and adaptability across a wide range of tasks, from image classification to complex segmentation. In the field of image manipulation detection and localization, these methods have significantly enhanced the ability to identify and precisely localize manipulated content in images, effectively countering the growing sophistication of forgery techniques [21].

Existing methods for detecting manipulated images have typically relied on large labeled datasets. These techniques have been widely used to identify image anomalies, such as copy-move, splicing, and inpainting forgeries. For example, in [40], a method that focuses on learning rich features for manipulation detection, however, these methods often struggle to generalize to novel manipulations, highlighting the need for more flexible and adaptable models. Recently, visual-language models have leveraged their ability to understand both visual content and textual descriptions, enabling a more comprehensive analysis of images for manipulation detection and localization [24, 29]. In [21], zero-shot image classification and retrieval techniques have been adapted for manipulation detection by identifying inconsistencies between image content and textual descriptions. By utilizing the robust contextual understanding provided by these models, the localization of manipulated areas has been improved without the need for extensive retraining. The Segment Anything Model (SAM) [23] has also been applied in the context of manipulation localization as a versatile tool capable of segmenting objects in images with minimal prompts. This approach has proven particularly effective in scenarios where the manipulation is subtle or context-dependent, highlighting the model's ability to generalize across different types of images and tasks.

In the domain of image manipulation detection and localization, the architecture of many advanced systems is typically designed with two distinct but interconnected branches [1, 17, 26, 36]. One branch usually processes the RGB input image in the spatial domain, focusing on the standard pixel-level information. The other branch is usually designed to capture and analyze the frequency domain information of the image. This is based on the understanding that all manipulations performed in the spatial domain, such as copy-move

or splicing, inevitably leave traces in the frequency domain. Techniques like discrete cosine transform (DCT), discrete fourier transform (DFT) or discrete wavelet transform (DWT) are commonly used to transform the image data into the frequency domain. The features extracted from this branch provide additional insights into the image's underlying structure, often revealing inconsistencies that are not easily detectable in the spatial domain alone. By combining spatial and frequency domain information, the dual-branch architecture improves the overall detection accuracy, particularly for subtle or sophisticated manipulations.

3 Proposed Methodology

In this section, we present a training-free approach for detecting and localizing manipulation in an image. The proposed methodology leverages the capabilities of large language models (LLMs) to extract meaningful descriptions from images, which are then processed to generate structured data. This data guides a segmentation model to localize the forgery and to provide a contextual explanation of the manipulation. Given an input image $I \in R^{H\times W\times 3}$ where H and W denote the height and width of the image, respectively, the first step involves generating a detailed textual description D of the image using a vision language model (VLM) [25].

$$D = \text{VLM}(I) \tag{1}$$

Here, D represents the description of the visual content in I, including objects, scenes, and potential signs of manipulation. The VLM is pre-trained on large-scale image-text pairs, allowing it to understand and describe complex visual elements.

3.1 Structured Data Generation

The generated description D is then fed into a LLM to produce structured data, which includes the following components:

- Manipulation Detected: $M \in \{0,1\}$, where $M = 1$ indicates that manipulation is detected, and $M = 0$ indicates no manipulation.
- Manipulated Object: O, a text label identifying the object or region believed to be manipulated.
- Explanation: E, a textual explanation providing the reasoning behind the manipulation detection.

This process represented as:

$$\{M, O, E\} = \text{LLM}(D) \tag{2}$$

In the proposed methodology, the process of generating detailed descriptions from images is critical to the accuracy and effectiveness of manipulation detection and localization. To enhance the granularity of analysis and improve detection accuracy, we employ a specific prompt strategy that instructs the VLM to virtually split the image into four quadrants and analyze each section independently. This approach allows the model to focus on smaller, potentially manipulated regions that might be overlooked when analyzing the image as a whole. The prompt used to guide the VLM in analyzing each image is structured as follows:

```
Virtually split the image into four quadrants and analyze
each section, as each may have undergone manipulation.
Then, provide an interpretation:
Manipulation present?
Location?
What elements are present in the image?
How realistic is the scene?
What inferences can be made from the image?
summarize the overall interpretation as a small paragraph
with manipulated object, location, and the reason.
```

This prompt is designed to systematically guide the VLM through a detailed inspection of each quadrant of the image, ensuring that no subtle manipulations are missed. The VLM independently evaluates each section and then synthesizes the findings into a comprehensive interpretation that includes the identification of manipulated objects, their locations, and the reasoning behind the detection.

3.2 Segmentation and Localization

Using the manipulated object O identified in the previous step, the segmentation model S localizes the manipulated region within the image I. The segmentation process can be described as:

$$M_{\text{seg}} = S(I, O) \tag{3}$$

where $M_{\text{seg}} \in R^{H \times W}$ is a binary mask that highlights the regions of the image that have been identified as manipulated. The segmentation model S [27] uses the structured data to focus on specific areas within the image, improving the precision of the localization. The segmentation model S used is a weakly supervised framework that uses referring texts to accurately localize target objects without requiring expensive pixel- or box-level annotations. Referring texts naturally provide descriptive details that aid in localizing target objects within images. Positive expressions highlight the target objects, whereas negative expressions refer to objects in other images. This differentiation enables the model to distinguish between relevant and irrelevant information, improving its focus on the correct target during the localization process.

The final step involves interpreting the structured data and the segmentation results. The binary mask M_{seg}, along with the manipulation detection result M, the manipulated object O, and the explanation E, provide a comprehensive understanding of the manipulation.

$$\text{Output} = \{M, O, M_{\text{seg}}, E\} \tag{4}$$

This output not only highlights whether manipulation is present but also localizes the manipulated regions and provides a reasoned explanation for the detection. The complete workflow of the proposed methodology is summarized in Algorithm 1.

3.3 Implementation Details

The proposed methodology is implemented using PyTorch and leverages the Vision-Language Model (VLM) [25] for generating detailed image descriptions, while GPT-4 serves as the

Algorithm 1 Manipulation Detection and Localization

Image I Manipulation detection result M, Manipulated object O, Segmentation mask M_{seg}, Explanation E Generate description D from image I using VLM:

$$D \leftarrow \text{VLM}(I)$$

Generate structured data $\{M, O, E\}$ using LLM:

$$\{M, O, E\} \leftarrow \text{LLM}(D)$$

Localize manipulated region using segmentation model S:

$$M_{\text{seg}} \leftarrow S(I, O)$$

Return M, O, M_{seg}, and E

large language model (LLM) for extracting structured data. Specifically, GPT-4 is utilized to produce structured outputs for *Manipulation Detection*, *Manipulated Object*, and *Explanation*. For the segmentation task, the approach integrates textual information from the structured data with the input image using a referring image segmentation model that localizes the manipulated region. Notably, the segmentation model operates based on preexisting capabilities without additional task-specific training, making the proposed method highly adaptable and scalable. It is capable of handling a wide range of manipulation types without the need for retraining. The implementation is executed on two NVIDIA RTX 3090 GPUs

4 Experimental Analysis

In this section, we evaluate the performance of the proposed training-free framework for image manipulation detection and localization. The evaluation is conducted on multiple datasets that include various types of forgeries, predominantly copy-move and splicing. We also compare our approach with five state-of-the-art techniques to demonstrate its effectiveness. We computed the F1 score to measure the performance of our framework.

4.1 Datasets

The datasets used in our experiments are CASIA v1.0 [10], Columbia [31], NIST16 [16], In-the-Wild [19], Coverage [37], IMD 2020 [32]. Columbia and In-the-Wild datasets are designed explicitly for splicing forgery, providing a range of images with well-labeled spliced regions; Coverage is explicitly for copy move forgery and the other datasets (CASIA v1.0, NIST16, and IMD 2020) contain images with splicing, copy-move, and removal forgeries, providing a comprehensive testbed for manipulation detection methods. The data sets were chosen to provide a diverse set of challenges, ensuring that the proposed framework is tested across different types of manipulation and scenarios.

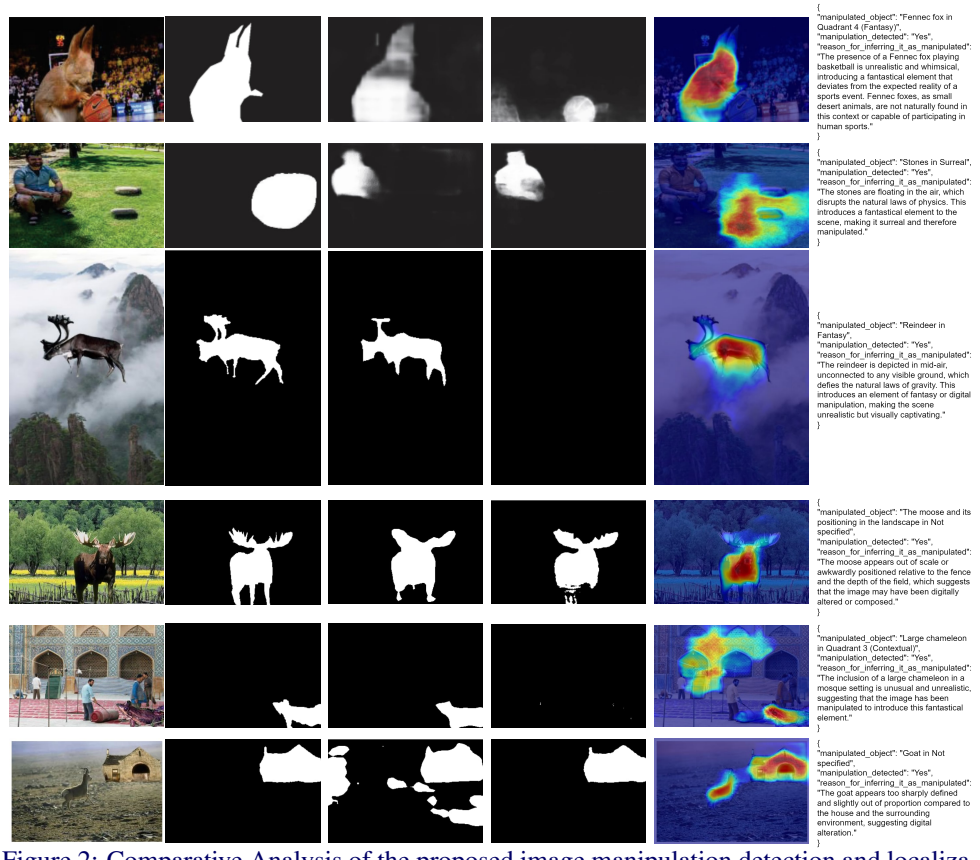

Figure 2: Comparative Analysis of the proposed image manipulation detection and localization. The analysis includes six samples, each row presenting the original image alongside the ground truth and results from MVSS-Net [9], PSCC-Net [28], and our technique. Additionally, each row includes the generated prompt fed to the segmentation network. This setup facilitates a direct qualitative comparison, highlighting the effectiveness of each method in handling complex image contexts.

4.2 Qualitative Analysis

The comparative analysis involved evaluating the results of the segmentation in multiple samples, where each row featured the original image, its corresponding ground truth, the results of MVSS-Net and PSCC-Net, our results, and the generated prompts that served as input to the segmentation network [27]. Our comprehensive qualitative evaluation reveals that our proposed training-free framework exhibits a performance that aligns well with current state-of-the-art techniques such as MVSS-Net [9] and PSCC-Net [28].

In all evaluated cases, our technique demonstrated improved localization and consistently outperformed MVSS-Net. Although our results were comparable to those of PSCC-Net in several instances, they remained competitive in others, highlighting the robustness of our approach. The generated prompts effectively identified the manipulated objects with high accuracy, which significantly contributed to the precision of the segmentation output. The localization of manipulated objects within specific image quadrants was generally precise.

Table 1: Comparison of F1 Scores for Manipulation Localization. This table presents a comparative analysis of F1 scores across four datasets, showcasing the performance of our technique against three other methods in the field of image manipulation localization.

Technique	Dataset					
	CASIAv1.0 [10]	Columbia [31]	Coverage [37]	NIST16 [14]	IMD2020 [33]	In-the-Wild [19]
DeepLabv3 [6]	0.430	0.432	0.148	0.228	0.208	0.201
RRU-Net [9]	0.295	0.270	0.100	0.201	0.163	0.169
MantraNet [38]	0.155	0.360	0.280	0.050	0.168	0.305
PSCC [28]	0.330	0.512	0.230	0.176	0.200	0.298
MVSS-Net [8]	0.435	0.353	0.350	0.300	0.276	0.205
Ours	0.210	0.380	0.116	0.150	0.170	0.180

However, challenges arose in correctly specifying the location, leading to instances where the location was ambiguously noted as 'not specified.' Some cases also revealed limitations in the ability of the segmentation network to accurately map descriptions to manipulated areas. This issue was particularly pronounced in complex scenarios, where the network struggled to align the detailed prompt descriptions with the visual data.

4.3 Quantitative Analysis

To validate the performance of our proposed framework, we compared it against five state-of-the-art manipulation detection techniques. The comparison is based on the standard F1-Score metric for localization accuracy. Table 1 presents the performance of DeepLabv3, RRU-Net, MantraNet, PSCC-Net, MVSS-Net, and our proposed method. Our framework demonstrates competitive performance across all datasets, particularly excelling in splicing forgery detection. For example, MVSS-Net consistently achieves the highest performance across many datasets, followed by PSCC-Net, which secures the second-highest scores. However, PSCC-Net performs the lowest on the Columbia dataset, with a score of 0.36. Our method shows on-par performance with MantraNet [38], and in some instances, it even surpasses RRU-Net [9]. This indicates that our method is effective across various types of manipulations, demonstrating its versatility and practical applicability. Additionally, our framework exhibits relatively balanced performance across the datasets, with F1-Scores ranging from 0.51 to 0.72. In many cases, these results are close to state-of-the-art performances, achieved without any additional training. The method's ability to generalize to different types of forgeries without the need for retraining underscores its robustness and adaptability.

4.4 Failure Cases

The proposed framework demonstrates strong results on splicing forgeries, which involve copying and pasting regions from one image into another. This type of manipulation is easier for the framework to detect because it can identify the contextual inconsistency and clearly infer the presence of manipulation. However, the performance is less optimal when handling copy-move forgeries, where a region is copied and pasted within the same image. In these cases, it becomes challenging to understand the contextual meaning due to the subtlety of the manipulation. As a result, forgeries that would typically be identifiable may go undetected when using foundation models. The lack of significant visual differences between the duplicated regions makes it difficult for vision language models (VLMs) to generate descriptive insights, which, in turn, complicates the task of segmentation networks to accurately local-

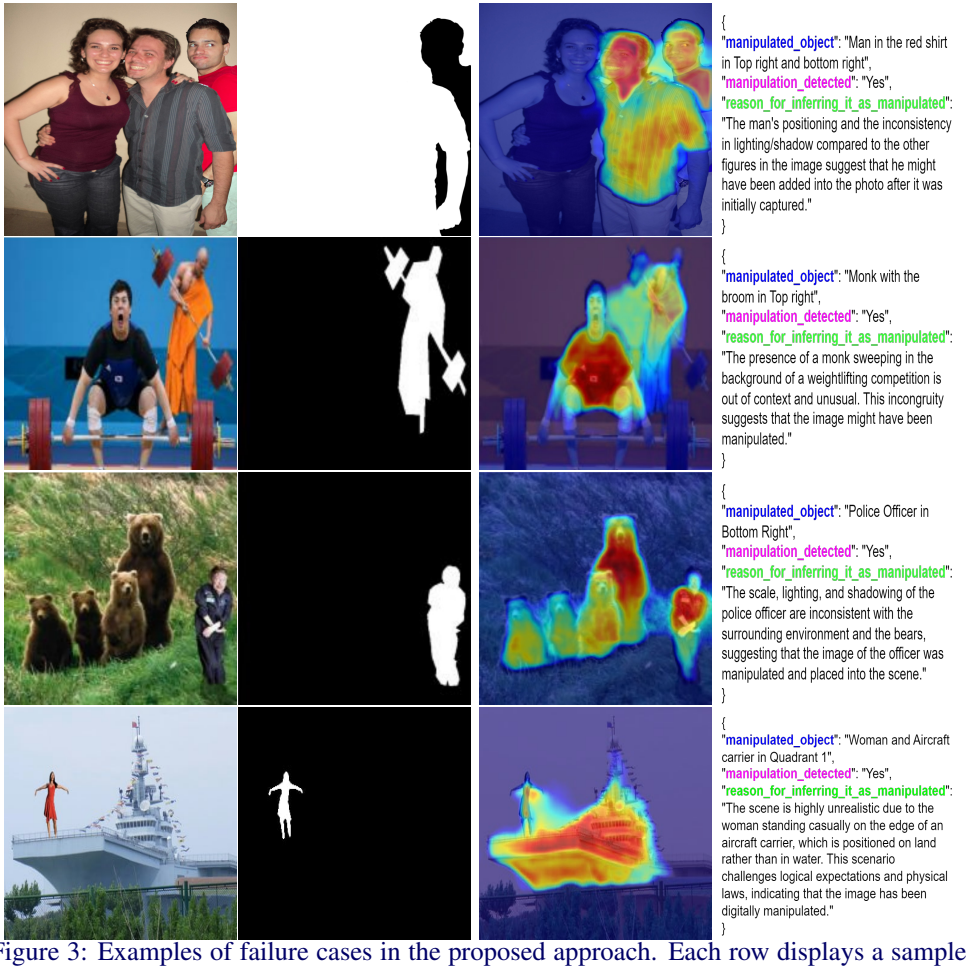

Figure 3: Examples of failure cases in the proposed approach. Each row displays a sample from a different dataset, showcasing the original image, the ground truth, and the predicted mask. The last column in each row contains the generated prompt used as input to the segmentation network.

ize the manipulated areas. Furthermore, the framework faces challenges when dealing with highly post-processed images and those generated by generative networks, such as GANs. These images are often subjected to extensive filtering, noise addition, and other alterations that obscure the manipulative artifacts. Since the proposed network has not been fine-tuned on specific image manipulation datasets, the LLMs and segmentation models used have not encountered such forgeries during their pre-training phase. As a result, the model's ability to detect and localize these sophisticated manipulations is limited. Figure 3 illustrates a few failure cases. In the first row, the derived text for the input image is *man in the red shirt in top right and bottom right*. However, the localization result targets both men in the image, indicating that the segmentation model focused more on "top right, bottom, man" and gave less emphasis to the word "red." Consequently, both men were segmented as manipulated regions. Similarly, in the second row, the model mistakenly identifies the weightlifting rod as a broom, producing the text *Monk with the broom in top right*. In the third row, the de-

rived text is *police officer in bottom right*, which the segmentation model also struggles to interpret accurately. The issue becomes more pronounced when proper nouns like "monk" or "police officer" are used, as the model tends to get confused and has difficulty distinguishing them, as opposed to more general terms like "human" or "animal." Finally, in the last row, the model mistakenly concludes that the ship is on land due to the presence of nearby plants, leading it to assume the ship should be in water. Consequently, both the ship and the woman on board are considered manipulated regions in the output.

4.5 Conclusions and Future Directions

The proposed framework has shown promise due to its reliance on detailed text descriptions generated by Vision-Language Models (VLMs). These descriptions allow the model to grasp the contextual meaning of the image content and produce structured data that can be effectively utilized for classification and localization, even without explicit training on forgery detection tasks. To further enhance the framework's performance, particularly in challenging scenarios like copy-move forgeries and highly post-processed or generative images, future work will focus on prompt enhancement and fine-tuning segmentation networks on a diverse set of manipulated images. This fine-tuning process will enable the models to better recognize and adapt to a wider range of manipulation types, thereby improving their generalization capabilities. Additionally, a future direction involves the development of specialized models, referred to as *forgeryGPTs*, which would be explicitly designed for the task of forgery detection and localization. These models would be trained on extensive datasets of manipulated images and would incorporate the latest advances in both language modeling and visual analysis. By leveraging the strengths of these two fields, forgeryGPTs could provide significant performance and reliability in detecting image forgeries across various domains.

References

[1] Susmit Agrawal, Prabhat Kumar, Siddharth Seth, Toufiq Parag, Maneesh Singh, and R Venkatesh Babu. Sisl: self-supervised image signature learning for splicing detection & localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22–32, 2022.

[2] Syed Sadaf Ali, Iyyakutti Iyappan Ganapathi, Ngoc-Son Vu, Syed Danish Ali, Neetesh Saxena, and Naoufel Werghi. Image forgery detection using deep learning by recompressing images. *Electronics*, 11(3):403, 2022.

[3] Belhassen Bayar and Matthew C Stamm. 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] Xiuli Bi, Yang Wei, Bin Xiao, and Weisheng Li. Rru-net: The ringed residual u-net for image splicing forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al.

Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[6] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(4): 834–848, 2018.

[7] Xinru Chen, Chengbo Dong, Jiaqi Ji, Juan Cao, and Xirong Li. Image manipulation detection by multi-view multi-scale supervision. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14185–14193, 2021.

[8] Fake Image Detector. Fake image detector. https://www.fakeimagedetector.com/. Accessed: 2024-08-15.

[9] Chengbo Dong, Xinru Chen, Ruohan Hu, Juan Cao, and Xirong Li. Mvss-net: Multi-view multi-scale supervised networks for image manipulation detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(3):3539–3553, 2023.

[10] Jing Dong, Wei Wang, and Tieniu Tan. CASIA image tampering detection evaluation database. In *2013 IEEE China Summit and International Conference on Signal and Information Processing, ChinaSIP 2013, Beijing, China, July 6-10, 2013*, pages 422–426. IEEE, 2013.

[11] Hany Farid. Exposing digital forgeries from jpeg ghosts. In *IEEE Transactions on Signal Processing*, volume 16, pages 1706–1710. IEEE, 1999.

[12] Foto Forensics. Foto forensics. https://fotoforensics.com/. Accessed: 2024-08-15.

[13] Jessica Fridrich, David Soukal, Jan Lukas, et al. Detection of copy-move forgery in digital images. In *Proceedings of digital forensic research workshop*, volume 3, pages 652–63. Cleveland, OH, 2003.

[14] Iyyakutti Iyappan Ganapathi, Sajid Javed, Syed Sadaf Ali, Arif Mahmood, Ngoc-Son Vu, and Naoufel Werghi. Learning to localize image forgery using end-to-end attention network. *Neurocomputing*, 512:25–39, 2022.

[15] Ghiro. Ghiro. https://getghiro.org/. Accessed: 2024-08-15.

[16] Haiying Guan, Mark Kozak, Eric Robertson, Yooyoung Lee, Amy N Yates, Andrew Delgado, Daniel Zhou, Timothee Kheyrkhah, Jeff Smith, and Jonathan Fiscus. Mfc datasets: Large-scale benchmark datasets for media forensic challenge evaluation. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pages 63–72. IEEE, 2019.

[17] Fabrizio Guillaro, Davide Cozzolino, Avneesh Sud, Nicholas Dufour, and Luisa Verdoliva. 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*, pages 20606–20615, 2023.

[18] Qixian Hao, Ruyong Ren, Shaozhang Niu, Kai Wang, Maosen Wang, and Jiwei Zhang. Ugee-net: Uncertainty-guided and edge-enhanced network for image splicing localization. *Neural Networks*, page 106430, 2024.

[19] Minyoung Huh, Andrew Liu, Andrew Owens, and Alexei A Efros. Fighting fake news: Image splice detection via learned self-consistency. In *Proceedings of the European conference on computer vision (ECCV)*, pages 101–117, 2018.

[20] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

[21] Sohail Ahmed Khan and Duc-Tien Dang-Nguyen. Clipping the deception: Adapting vision-language models for universal deepfake detection. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pages 1006–1015, 2024.

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

[23] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[24] Joshua Krinsky, Alan Bettis, Qiuyu Tang, Daniel Moreira, and Aparna Bharati. Exploring saliency bias in manipulation detection. *arXiv preprint arXiv:2402.07338*, 2024.

[25] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[26] Shuaibo Li, Wei Ma, Jianwei Guo, Shibiao Xu, Benchong Li, and Xiaopeng Zhang. Unionformer: Unified-learning transformer with multi-view representation for image manipulation detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12523–12533, 2024.

[27] Fang Liu, Yuhao Liu, Yuqiu Kong, Ke Xu, Lihe Zhang, Baocai Yin, Gerhard Hancke, and Rynson Lau. Referring image segmentation using text supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22124–22134, 2023.

[28] Xiaohong Liu, Yaojie Liu, Jun Chen, and Xiaoming Liu. Pscc-net: Progressive spatio-channel correlation network for image manipulation detection and localization. *IEEE Trans. Circuits Syst. Video Technol.*, 32(11):7505–7517, 2022.

[29] Xuntao Liu, Yuzhou Yang, Qichao Ying, Zhenxing Qian, Xinpeng Zhang, and Sheng Li. Prompt-iml: Image manipulation localization with pre-trained foundation models through prompt tuning. *arXiv preprint arXiv:2401.00653*, 2024.

[30] Soumyaroop Nandi, Prem Natarajan, and Wael Abd-Almageed. Trainfors: A large benchmark training dataset for image manipulation detection and localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 403–414, 2023.

[31] Tian-Tsong Ng, Jessie Hsu, and Shih-Fu Chang. Columbia image splicing detection evaluation dataset. *DVMM lab. Columbia Univ CalPhotos Digit Libr*, 2009.

[32] Adam Novozamsky, Babak Mahdian, and Stanislav Saic. 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*, pages 71–80, 2020.

[33] Chenfan Qu, Yiwu Zhong, Chongyu Liu, Guitao Xu, Dezhi Peng, Fengjun Guo, and Lianwen Jin. Towards modern image manipulation localization: A large-scale dataset and novel methods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10781–10790, 2024.

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

[35] Matthew C Stamm and KJ Liu. Forensic detection of image tampering using intrinsic statistical fingerprints in histograms. In *Proceedings: APSIPA ASC 2009: Asia-Pacific Signal and Information Processing Association, 2009 Annual Summit and Conference*, pages 563–572. Asia-Pacific Signal and Information Processing Association, 2009 Annual . . . , 2009.

[36] Junke Wang, Zuxuan Wu, Jingjing Chen, Xintong Han, Abhinav Shrivastava, Ser-Nam Lim, and Yu-Gang Jiang. Objectformer for image manipulation detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2364–2373, 2022.

[37] Bihan Wen, Ye Zhu, Ramanathan Subramanian, Tian-Tsong Ng, Xuanjing Shen, and Stefan Winkler. Coverage—a novel database for copy-move forgery detection. In *2016 IEEE international conference on image processing (ICIP)*, pages 161–165. IEEE, 2016.

[38] Yue Wu, Wael AbdAlmageed, and Premkumar Natarajan. Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 9543–9552. Computer Vision Foundation / IEEE, 2019.

[39] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13041–13049, 2020.

[40] Peng Zhou, Xintong Han, Vlad I Morariu, and Larry S Davis. Learning rich features for image manipulation detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1053–1061, 2018.