Noise-assisted Prompt Learning for Image Forgery Detection and Localization (Supplementary Material)

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

University of Science and Technology of China, Hefei, 230026, China

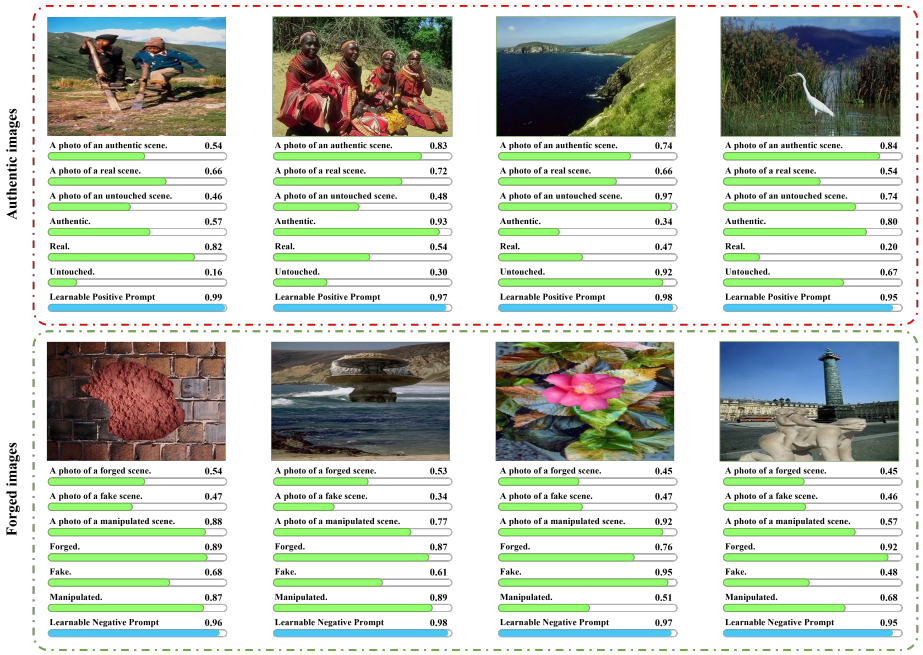

Fig. 1: Perception capability of CLIP for the authenticity-forgery attributes of images.

1 Perception Capability of CLIP for Forgery

In Figure 1, we illustrate the perception capability of the frozen CLIP regarding the authenticity-forgery attributes of images and the performance using different prompts. It can be observed that for each image, a semantically relevant prompt related to the concepts of authentic or forged can achieve high scores in the CLIP space. This indicates the potential of CLIP in distinguishing between forged and authentic images. However, achieving this with fixed discrete prompts is

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

not straightforward. For instance, in the case of a real image (the first image in the first row of Figure 1), replacing a similar concept like "authentic" with "real" results in a significant increase in CLIP scores. Conversely, in the opposite case (the second image in the first row of Figure 1), "authentic" becomes the most suitable prompt. This suggests that the optimal prompt may vary for each image due to the diversity of forged images. Therefore, finding precise words or sentences to uniformly describe the authenticity-forgery attributes of images is challenging, and prompt engineering is labor-intensive and time-consuming to annotate each image. Given these considerations, we employ learnable prompts to describe the authenticity-forgery attributes of images and achieve optimal performance.

2 Baseline Models of the Manuscript

In the manuscript, our method is compared with various baseline models. They are described as follows.

RGB-N [19]: This model adopts a two-stream Faster R-CNN network, including an RGB stream and a noise stream, to discover tampering features and noise inconsistency within an image separately.

ManTraNet [18]: This model 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 [6]: This model constructs a pyramid of local self-attention blocks to model the relationship between image patches at multiple scales.

PSCCNet [9]: This model processes the image in a two-path procedure: a top-down path that extracts local and global features and a bottom-up path that detects whether the input image is manipulated.

ObjectFormer [15]: This model extracts high-frequency features of the images and combines them with RGB features as multimodal patch embeddings to capture subtle manipulation traces in the RGB domain.

HiFi-Net [4]: This model introduces a hierarchical fine-grained approach to IFDL representation learning. Specifically, it conducts fine-grained classification at various levels, leveraging their hierarchical dependencies for enhanced performance.

SAFL-Net [14]: This model constrains a feature extractor to learn semantic-agnostic features by designing specific modules with corresponding auxiliary tasks. Meanwhile, it leverages boundary supervision to identify inconsistencies in the features around the tampered boundary and design a feature conversion structure to ensure the coherence of the auxiliary task and the primary task.

3 More Ablation Studies

3.1 Prompt initialization

To mitigate the difficulty of finding an accurate vector in a continuous semantic space to measure the concept of authenticity-forgery, we use a pair of discrete

prompts as initializations for learnable prompts. We try three sets of prompt words: "Untouched/Manipulated", "Authentic/Forged", and "Real/Fake". The results on the CASIA dataset using these three sets of prompt words as initializations and random initialization are shown in Table 1. It can be observed that using discrete prompts as initialization is better than random initialization, and employing "Real/Fake" is an appropriate choice.

Table 1: Results on CASIA dataset using different prompt initialization. The image-level F1 score (%) for the detection task and the pixel-level F1 score (%) for the localization task are reported.

	Random initialization	Untouched/Manipulated	Authentic/Forged	Real/Fake
Detection	92.1	95.2	94.9	**98.75**
Localization	74.2	75.1	75.3	**77.90**

3.2 Comparison with the Linear Probe

To further validate the effectiveness of our method, we replace prompt learning and the text encoder with the linear probe. Specifically, following [13], we feed the features outputted by the image encoder into a linear classifier for training. This approach referred to as the linear probe, replaces the text-image similarity calculation in our method for forgery detection. As shown in Table 2, our method not only outperforms the linear probe in image-level detection performance but also excels in pixel-level localization performance. This indicates that the prompt learning in our method not only facilitates forgery detection by leveraging CLIP priors but also refines localization through the guidance of learnable prompts.

Table 2: Comparison with the linear probe on the CASIA and COVER datasets. 'Image' refers to image-level AUC (%), and 'Pixel' refers to pixel-level AUC (%).

	Linear Probe	Ours
Image (CASIA)	98.7	**99.8**
Pixel (CASIA)	88.9	**91.3**
Image (COVER)	61.5	**73.2**
Pixel (COVER)	95.1	**98.2**

3.3 Ablation on localization decoder

Since the localization decoder is not the primary contribution of this paper, we use a simple U-shaped decoder. To further explore the impact of this localization decoder on performance, we conduct ablation experiments. In Table 3, we design the following variants: replacing FiLM of the decoder with concatenation (w/o FiLM), using convolution blocks instead of transformer blocks (w/o transformer), and omitting skip connections (w/o U-Net). It is evident that our

approach is the optimal choice. Furthermore, compared to the variants in Table 4 of the manuscript, modifications to the decoder design do not result in significant performance degradation. This indicates that the impact of decoder design on performance is relatively small compared to the core components of our method, IDPL and FENA.

Table 3: Ablation results of our scheme using different variants of localization decoders. Pixel-level AUC (%)and F1 scores (%) are reported.

Variants	CASIA		NIST16	
	AUC	F1	AUC	F1
w/o FiLM	89.1	71.5	95.6	85.1
w/o transformer	90.4	75.1	97.9	88.2
w/o U-Net	89.6	74.3	97.2	87.6
Ours	**91.3**	**77.9**	**99.8**	**89.3**

4 More Experimental Setups

Pre-training Data For splicing, we employ the MS COCO [8] to generate spliced images, where one annotated region is randomly selected per image and pasted into a different image after several transformations. We adopt the same transformation as [1], including the scale, rotation, shift, and luminance changes. Since the spliced region is not always an object, we create random outlines using the Bezier curve [10] and fill them to create splicing masks. For copy-move, the datasets from MS COCO and [17] are adopted. For removal, we adopt the SOTA inpainting method [7] to fill one annotated region that is randomly removed from each chosen MS COCO image. We randomly add Gaussian noise or apply the JPEG compression algorithm to the generated data to resemble the visual quality of images in realistic scenarios. It is worth noting that our method of synthesizing datasets follows the previous seminal works [9, 15].

Testing Datasets Our test dataset includes CASIA [2], Coverage [16], Columbia [5], Nist Nimble 2016 (NIST16) [3] and IMD20 [11]. Specifically, CASIA, which contains two types of tampered images (splicing and copy-move), is widely used in the image forgery domain. Coverage provides 100 images, and all of them are generated by copy-move tampering technique. Columbia consists of 180 splicing images, whose size ranges from 757×568 to 1152×568. NIST16 is a high-quality dataset. IMD20 collects real-life manipulated images from Internet, and involves all three manipulations as well.

Evaluation Metrics To quantify the localization performance, following previous works [6, 15], we use pixel-level Area Under Curve (AUC) and F1 score on manipulation masks. To evaluate detection performance, we use image-level AUC and F1 score. To further measure the miss detection rate and false alarm rate, we report specificity and sensitivity on some challenging datasets. Since binary masks are required to compute F1 scores, we adopt the Equal Error Rate (EER) threshold to binarize them.

Implementation details The input images are resized to 512×512. In this work, the backbone is CLIP ViT-B/16 [13]. We enable CLIP to accept different image sizes by interpolating the positional embeddings. Implemented by PyTorch, our model is trained with NVIDIA V100. We use Adam as the optimizer, and the learning rate decays from 10^{-4} to 10^{-7}. We train 100 epochs with a batch size of 8, and the learning rate decays by 10 times every 30 epochs.

5 Limitations and Discussions

The proposed model achieves excellent results on several datasets and performs well in generalizability as well as in resolving false alarms. However, challenges remain when facing complex scenarios, such as the hybrid lossy operations employed by online social networks and the forged image containing multiple tampered regions. The former affects the robustness of the network, while the latter impacts the accuracy of forgery localization. Therefore, our next step will be to handle these real-world complex scenarios better.

In addition, some visual artifacts introduced by image editing are less perceptible in the RGB domain but become noticeable in the frequency domain [12]. However, directly incorporating frequency domain information faces challenges in our approach. This is because frequency domain information, such as the Fourier domain, differs significantly from spatial domain information, presenting a larger disparity than that between noise and RGB domains. This substantial difference results in a mismatch between the CLIP prior and frequency domain information. Hence, introducing frequency domain information in CLIP-IFDL is a direction to be explored in future work, for example, through strategies involving signal modulation.

References

1. 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)
2. 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)
3. 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)
4. 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)
5. 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)
6. 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)
7. Li, J., Wang, N., Zhang, L., Du, B., Tao, D.: Recurrent feature reasoning for image inpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7760–7768 (2020)
8. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
9. 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)
10. Mortenson, M.E.: Mathematics for computer graphics applications. Industrial Press Inc. (1999)
11. 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)
12. Qian, Y., Yin, G., Sheng, L., Chen, Z., Shao, J.: Thinking in frequency: Face forgery detection by mining frequency-aware clues. In: European conference on computer vision. pp. 86–103. Springer (2020)
13. 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)
14. 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)
15. 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)

16. 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)
17. 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)
18. 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)
19. 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)