Supplementary Materials of Multi-view Feature Extraction via Tunable Prompts is Enough for Image Manipulation Localization

Table 1: Detailed information about the training and test datasets we used. We use cpmv, spl, and imp to represent the Copy-Move, Splicing, and Inpainting manipulation techniques, respectively. '-' indicates that the data is not available.

Dataset	Fake	Authentic	cpmv	spl	inp
Training					
CASIA2[1]	7,491	5,063	3,235	1,828	0
Testing					
CASIA1[1]	800	920	459	461	0
COVER[10]	100	100	100	0	0
Columbia[5]	183	180	0	180	0
NIST16[2]	0	564	68	288	208
IMD20[7]	414	2,010	-	-	-
DEF-12K[4]	6,000	6,000	2,000	2,000	2,000

1 Details of Training and Test Datasets

We solely utilize CASIA2 [1] to train Prompt-IML. 6 public test datasets are utilized for evaluation, including CASIA1 [1], NIST16 [2], COVERAGE [10], Columbia [5], IMD2020 [7], and DEFACTO [4]. The detailed information for each dataset can be found in Tab.1.

2 Evaluation through AUC Metric

AUC is another commonly used evaluation metric for IML task. To comprehensively assess our model's performance, we report the AUC scores of our model on 5 test datasets in Tab.2. The missing information in the table is due to differences in experimental protocols. Our approach exhibits superior performance across 5 datasets, with an average improvement of 2.8% over IML-ViT.

Table 2: Image Manipulation Localization Performance (AUC score). We highlight the best results in each column in bold. '-' indicates that the data is not available due to the different experimental protocols.

Method	CASIA1	Columbia	NIST16	COVER	IMD20	Average
ObjectFormer[9], CVPR22	0.882	-	-	-	-	-
CFL-Net[6], WACV23	0.863	-	0.799	-	-	-
SAFL-Net[8], ICCV23	0.908	-	-	-	-	-
IML-ViT[3], AAAI24	0.931	0.962	0.818	**0.918**	0.892	0.904
Prompt-IML	**0.954**	**0.978**	**0.891**	0.913	**0.923**	**0.932**

3 Ablation Study of Prompts

We first explore the impact of tunable prompts quantity on model performance. Specifically, we set the number of prompts to 5, 10, 20, and 30, and report the F1 scores of each setting on 6 test datasets in the top part of Tab.3. Due to the minor performance differences, we use 5 as the default number of prompts, as this setting is more resource-efficient in computing self-attention. Then, we explore the impact of shallow prompt and deep prompt. In the shallow prompt experiment, we use fully connected layers to adjust the dimensions C_i of the prompts between the backbone's layers to meet the size requirements. We report the F1 scores in the bottom part of Tab.3 and choose the deep prompt strategy due to the experiments' results.

Table 3: Ablation study of prompts. The upper part shows the impact of the number of prompts, while the lower part demonstrates the differences between shallow prompt and deep prompt. We highlight the best results in each column in bold.

	CASIA1	Columbia	NIST16	COVER	DEF-12K	IMD20
5	0.686	0.882	0.415	**0.429**	**0.237**	**0.471**
10	**0.713**	**0.906**	0.410	0.404	0.233	0.456
20	0.691	0.880	0.399	0.392	**0.237**	0.435
30	0.701	0.903	**0.416**	0.412	0.229	0.459
shallow prompt	0.684	**0.896**	0.373	0.350	0.219	0.402
deep prompt	**0.686**	0.882	**0.415**	**0.429**	**0.237**	**0.471**

4 Additional Localization Results

To fully showcase the outstanding performance of our method, we present additional results in Fig. 1. The tampered images are sourced from 5 different test datasets, with significant variations in the size of the manipulated areas. Experimental results demonstrate the superior generalization capability of our model.

References

[1] Jing Dong, Wei Wang, and Tieniu Tan. 2013. Casia image tampering detection evaluation database. In *2013 IEEE China summit and international conference on signal and information processing*. IEEE, 422–426.

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

[3] Xiaochen Ma, Bo Du, Xianggen Liu, Ahmed Y Al Hammadi, and Jizhe Zhou. 2023. Iml-vit: Image manipulation localization by vision transformer. *arXiv preprint arXiv:2307.14863* (2023).

[4] Gaël Mahfoudi, Badr Tajini, Florent Retraint, Frederic Morain-Nicolier, Jean Luc Dugelay, and PIC Marc. 2019. DEFACTO: Image and face manipulation dataset. In *2019 27Th european signal processing conference (EUSIPCO)*. IEEE, 1–5.

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

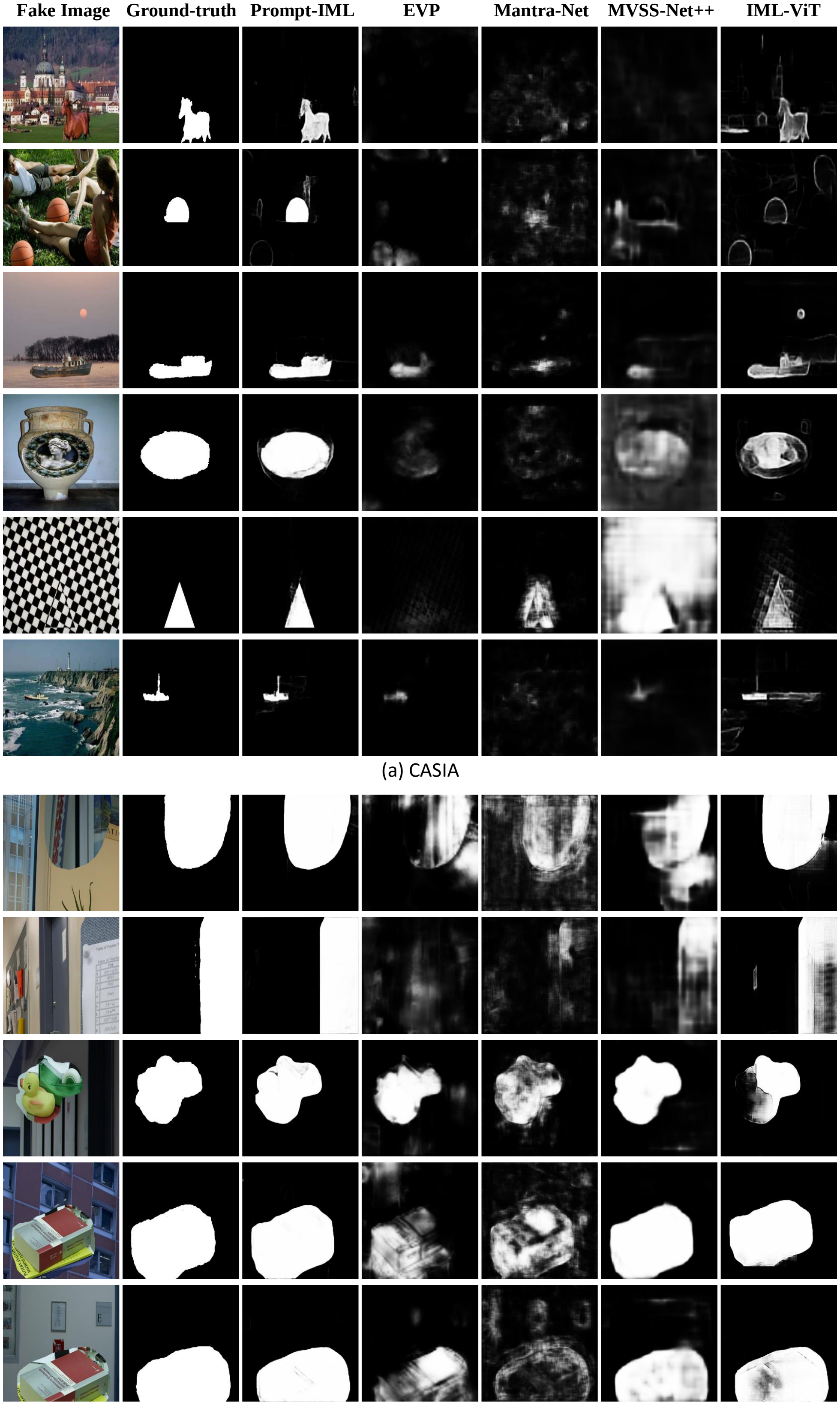
[6] Fahim Faisal Niloy, Kishor Kumar Bhaumik, and Simon S Woo. 2023. Cfl-net: Image forgery localization using contrastive learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 4642–4651.

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

[8] Zhihao Sun, Haoran Jiang, Danding Wang, Xirong Li, and Juan Cao. 2023. 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*. 22424–22433.

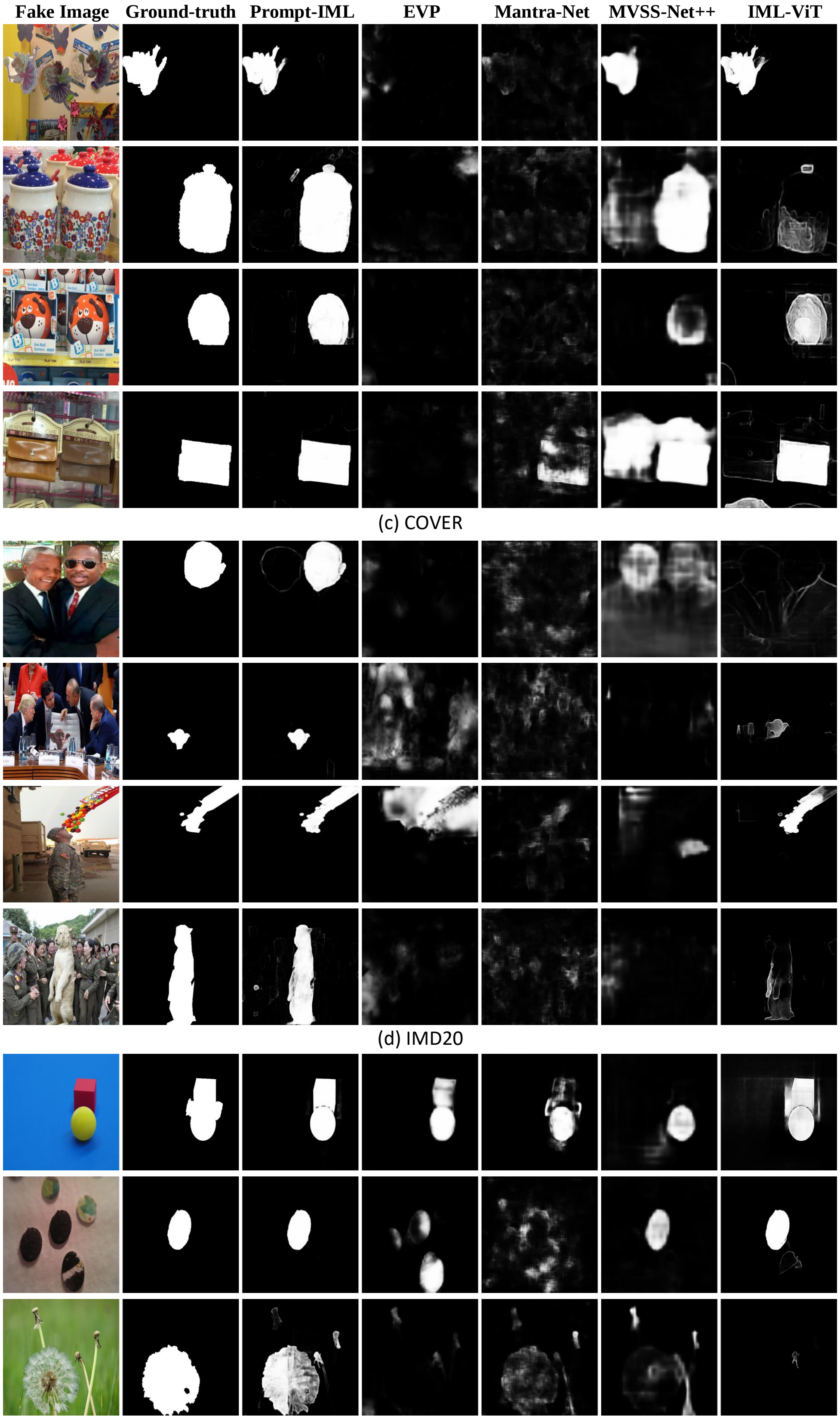
[9] Junke Wang, Zuxuan Wu, Jingjing Chen, Xintong Han, Abhinav Shrivastava, Ser-Nam Lim, and Yu-Gang Jiang. 2022. Objectformer for image manipulation detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2364–2373.

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



(a) CASIA

(b) Columbia



(c) COVER

(d) IMD20

(e) NIST16

Figure 1: Additional manipulation localization results on images originating from 5 datasets. Columns from left to right are: fake image, ground-truth, Prompt-IML, EVP, Mantra-Net, MVSS-Net++ and IML-ViT.