Unleashing the Potential of Consistency Learning for Detecting and Grounding Multi-Modal Media Manipulation

Supplementary Material

A. Additional visualization

As shown in Fig. 1, we give the additional visualizations on DGM^4 dataset, which demonstrate the effectiveness of our proposed method in different scenarios.

B. Future work

The future work could be carried out based on the following aspects. First, the usage of large vision-language model (LVLM) should be explored on DGM^4 [2] task, as LVLM contains rich real-world knowledge which contributes to forgery reasoning. Second, more advanced constructing and supervising functions should be introduced on DGM^4 task, which enhances the ability to distinguish between genuine and forged information in extreme difficult cases. Third, due to the frequent issues of distortion and noise in image transmission on the internet, the robustness of detection is also worth exploring in the future.

C. Failure cases

As shown in Fig. 2, we give some failure cases. It could be observed that when the manipulated face is too small, the

Figure 1. **Visualization of detecting and grounding results on DGM^4 datasets.** Here, red box and text indicate the prediction of manipulated faces and words, while green box and text represent the corresponding ground truth. The first line is the prediction for text forgery, the second line is the prediction for image forgery, and the third line is the prediction for multi-modal forgery.

Figure 2. **Visualization of failure cases on DGM[4] datasets.** Here, red box and text indicate the prediction of manipulated faces and words, while green box and text represent the corresponding ground truth.

detecting or grounding tasks to image modality may fail. What's more, when the expression of the sentence is relatively vague and there is no well-aligned in semantics between image-text pair, the detecting or grounding tasks to text modality may fail.

D. Additional loss details

Here, we give additional details of losses used in different sub-tasks which are similar to [3]. V_{cls} and T_{cls} represent the class embeddings of the outputs of multi-modal interaction, while $\widetilde{V}_a$ and $\widetilde{T}_a$ represent the aggregated embeddings of the outputs of semantic consistency decoder. For binary classification, the loss L_{bcls} can be obtained by Eq. 1.

$$L_{bcls} = L_{ce}(C_b([V_{cls}, T_{cls}]), \hat{y}), \tag{1}$$

where L_{ce} is the cross entropy loss, C_b is the binary classifier and $\hat{y}$ is the corresponding ground truth. $[., .]$ is the concatenation operator. For fine-grained manipulated type prediction, the loss L_{fcls} can be calculated by Eq. 2.

$$L_{fcls} = L_{ce}(C_i(\widetilde{V}_a), \hat{y_i}) + L_{ce}(C_t(\widetilde{T}_a), \hat{y_t}), \tag{2}$$

where C_i is the FS/FA classifier, C_t is the TS/TA classifier. $\hat{y_i}$ and $\hat{y_t}$ are the corresponding ground truth. For grounding face manipulation, the loss is composed of L1 and GIoU loss [1] as shown in Eq. 3.

$$L_{img} = L_1(D_i(\widetilde{V}_a), \hat{b}) + L_{giou}(D_i(\widetilde{V}_a), \hat{b}), \tag{3}$$

where, D_i is the Bbox decoder, and $\hat{b}$ is the ground truth of Bbox. The overall loss is shown in Eq. 4.

$$L = L_{bcls} + \alpha L_{fcls} + \beta L_{img} + \gamma(Lc + Ls), \tag{4}$$

where Lc and Ls are the consistency loss of contextual and semantic consistency matrices, respectively. α, β and γ are the hyper-parameters to balance different loss functions. Based on the principle of achieving a uniform order of magnitude for all losses to prevent preference bias, we set them to 1, 0.1, and 10, respectively.

References

[1] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019. 2

[2] Rui Shao, Tianxing Wu, and Ziwei Liu. Detecting and grounding multi-modal media manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6904–6913, 2023. 1

[3] Jiazhen Wang, Bin Liu, Changtao Miao, Zhiwei Zhao, Wanyi Zhuang, Qi Chu, and Nenghai Yu. Exploiting modality-specific features for multi-modal manipulation detection and grounding. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4935–4939. IEEE, 2024. 2