Supplementary Material of
Detecting and Grounding Multi-Modal Media Manipulation

Rui Shao[1,2]*, Tianxing Wu[2], Ziwei Liu[2]†
[1] School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen)
[2] S-Lab, Nanyang Technological University
shaorui@hit.edu.cn, {tianxing.wu, ziwei.liu}@ntu.edu.sg
https://github.com/rshaojimmy/MultiModal-DeepFake

1. Experiments

1.1. Implementation Details.

All of our experiments are performed on 8 NVIDIA V100 GPUs with PyTorch framework [7]. Image Encoder is implemented by ViT-B/16 [3] with 12 layers. Both Text Encoder and Multi-Modal Aggregator are built based on a 6-layer transformer initialized by the first 6 layers and the last 6 layers of BERT$_{base}$ [2], respectively. Binary Classifier, Multi-Label Classifier, BBox Detector, and Token Detector are set up to two Multi-Layer Perception (MLP) layers with output dimensions as 2, 4, 4, and 2. We set the queue size $K = 65,536$. AdamW [5] optimizer is adopted with a weight decay of 0.02. The learning rate is warmed-up to $1e^{-4}$ in the first 1000 steps, and decayed to $1e^{-5}$ following a cosine schedule.

1.2. Evaluation Metrics.

To evaluate the proposed new research problem DGM^4 comprehensively, we set up rigorous evaluation protocols and metrics for all the manipulation detection and grounding tasks.

- **Binary classification:** Following current deepfake methods [6,8], we adopt Accuracy (ACC), Area Under the Receiver Operating Characteristic Curve (AUC), and Equal Error Rate (EER) for evaluation of binary classification.

- **Multi-label classification:** Like existing multi-label classification methods [1,4], we use mean Average Precision (MAP), average per-class F1 (CF1), and average overall F1 (OF1) for evaluating the detection of fine-grained manipulation types.

- **Manipulated image bounding box grounding:** To examine the performance of predicted manipulated bbox, we calculate the mean of Intersection over Union (IoUmean) between ground-truth and predicted coordinates of all testing samples. Moreover, we set two thresholds (0.5, 0.75) of IoU and calculate the average accuracy (correct grounding if IoU is above the threshold and versa vice), which are denoted as IoU50 and IoU75.

- **Manipulated text token grounding:** Considering the class imbalance scenario that manipulated tokens are much fewer than original tokens, we adopt Precision, Recall, F1 Score as metrics. This contributes to a more fair and reasonable evaluation for manipulated text token grounding.

References

[1] Emanuel Ben-Baruch, Tal Ridnik, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In *ICCV*, 2021. 1

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. 1

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020. 1

[4] Thibaut Durand, Nazanin Mehrasa, and Greg Mori. Learning a deep convnet for multi-label classification with partial labels. In *CVPR*, 2019. 1

[5] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1

[6] Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu. Generalizing face forgery detection with high-frequency features. In *CVPR*, 2021. 1

[7] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 1

[*] This work was done at S-Lab, Nanyang Technological University
[†] Corresponding author

[8] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. In *CVPR*, 2021. 1