SAFL-Net: Semantic-Agnostic Feature Learning Network with Auxiliary Plugins for Image Manipulation Detection

Zhihao Sun[1,2], Haoran Jiang[3], Danding Wang[1,2]*, Xirong Li[4], Juan Cao[1,2]
[1]Institute of Computing Technology, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
[3]School of Mathematics Science, University of Chinese Academy of Sciences
[4]MoE Key Lab of Data Engineering and Knowledge Engineering, Renmin University of China
{sunzhihao21s,wangdanding,caojuan}@ict.ac.cn,
jianghaoran21@mails.ucas.ac.cn, xirong@ruc.edu.cn

A. Correlation Between Manipulation Regions and Semantic Information

To quantify the correlation between tampered regions and semantic information in a dataset and to reflect the difference in this correlation between various datasets, we perform statistics on the semantic distribution of tampered regions in representative datasets.

In our core experiments, we select six publicly available datasets, including Columbia [8], Coverage [9], CASIAv1 [2], CASIAv2 [2], IMD20 [7], NIST16 [3]. Images in Columbia dataset are tampered with by splicing two images together in a random shape, so the tampering is too simple and there is no significant semantic information in tampered areas. And Coverage dataset contains only 100 tampered images, we therefore neglect to count the semantic distribution of these two datasets. The annotations provided in CASIAv1 and CASIAv2 datasets include the categories of the tampered regions, so we can directly analyze the labels to obtain statistics. However, there are no similar annotations in IMD20 dataset and NIST16 dataset, and we manually count the different semantic information contained in the manipulation regions.

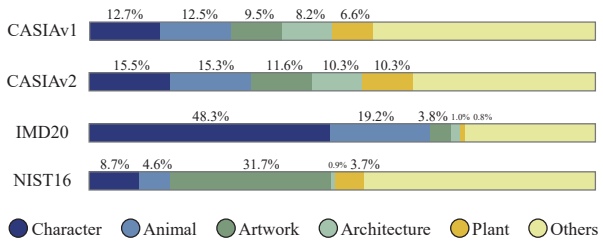

Figure 1. The statistics of correlation between manipulation regions and semantic information.

*Corresponding author.

As shown in Figure 1, we count the number of tampered regions with common semantic information including *character*, *animal*, *artwork*, *architecture*, *plant* and *others*, and transform them into percentages for easy comparison between various datasets. The difference in the correlation between different datasets is obvious, and this phenomenon is sufficient to show that such semantically related distribution in limited data is unreliable for the pictures from an unseen scene to be detected.

B. Synthesized Dataset

B.1. ProDEFACTO

DEFACTO [6] is a recent publicly available dataset containing a large amount of manipulated images which are automatically generated based on MS COCO [5]. The dataset offers three types of manipulated images and corresponding tampering strategies to enhance the authenticity of tampered images. By leveraging the annotations provided by MS COCO, it ensures that the forgeries produced are meaningful and take into account the semantic context.

The tampering traces in this dataset are relatively simple and easy to learn and fit, lacking the post-processing operations that simulate reality. Therefore, a model trained directly on this dataset may not generalize well. To address this issue, we use Albumentations [1] to design a post-processing operation pipeline, as shown in the Figure 2, to enhance the original data and create prodefacto. While this operation increases the difficulty of fitting tampering traces, it significantly improves the generalization of the model.

Unlike the data augmentation operations used in the training phase, which aimed to increase data richness, we apply several post-processing operations commonly seen in real media scenes to decrease image quality. For each image, we randomly select two operations and applied them

Setup	Pixel-level localization (F1)					Image-level detection				Com-F1
	spli.	*cpmv.*	*inpa.*	*ps.*	MEAN	AUC	Sen.	Spe.	F1	
0: Seg+Cls#0	66.5	39.2	25.6	51.7	45.8	84.1	55.4	**97.6**	70.7	55.6
1: Seg+Cls#1 (Baseline)	68.7	42.8	29.9	54.7	49.0	84.4	75.3	84.2	79.5	60.0
2: Seg+Cls#1+bdSup#0	64.4	44.9	30.1	49.6	47.3	80.7	79.4	63.0	70.3	56.6
3: Seg+Cls#1+BGM#1	70.4	48.0	35.5	55.6	52.4	87.3	80.7	72.6	76.4	62.2
4: Seg+Cls#1+BGM#2	71.8	49.9	40.6	58.0	55.1	87.1	84.7	92.5	88.4	67.8
5: Seg+Cls#1+BGM#2+SSM#1	70.1	**53.2**	38.0	64.1	56.4	90.8	**90.9**	86.1	88.4	68.9
6: Seg+Cls#1+BGM#2+SSM#2	**72.4**	51.7	**41.4**	**68.6**	**58.5**	**91.1**	89.5	88.8	**89.2**	**70.7**

Table 1. Ablation study of different modules in SAFL-Net on different manipulation types.

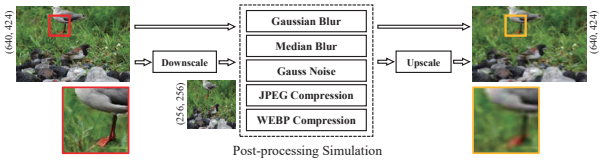

Figure 2. Post-processing operations of ProDEFACTO.

in a random order. Before the processing, we downscale the image with a probability of 0.5, and then upscale it back to the original resolution afterwards, with the goal of further obscuring the tampering traces and simulating more challenging scenarios.

B.2. PSBattles

PSBattles [4] provides a more sophisticated form of manipulation that is highly relevant to real-world scenarios and more challenging to detect. The tampered images in the dataset were collected from a large community of image manipulation enthusiasts, but they do not come with pixel-level annotations of the manipulated regions.

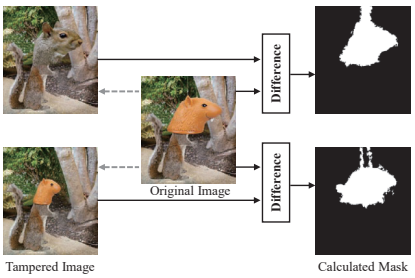

Figure 3. Calculating pixel-level masks for PSBattles.

Figure 3 demonstrates that we calculate the difference between the tampered image and the authentic image provided in the dataset. Based on this correspondence, we binarize the difference using a threshold of 0.2 and generate the tampered region mask, with the goal of achieving the most accurate results possible.

C. Ablation Study

Through the utilization of our well-structured dataset that encompasses four distinct types of tampering, our study enables a more comprehensive analysis within the ablation experiment section. Specifically, we investigate the effectiveness of each of our proposed strategies in relation to the detection of each type of tampering as shown in Table 1.

Our proposed Boundary Guidance Module (BGM) aims to guide the model in detecting subtle feature differences between the interior and exterior regions of the boundary by leveraging boundary guidance. Comparing Setup#4 and Setup#1, we find that BGM is highly effective in detecting *inpa.* manipulation (from 42.8 to 49.9). This is due to the fine feature differences between the tampered and original regions generated by the inpainting algorithm. Furthermore, BGM is also highly effective in detecting *cpmv.* tampering (from 29.9 to 40.6), indicating that our module not only serves a differential role, but the residual connection involved can effectively extract traces on the boundaries. However, the improvement of the BGM in detecting *ps.* tampering is not significant (from 54.7 to 58.0). This may be due to the fact that the boundary annotations in this type of data are not precise.

By comparing Setup#6 and Setup#4, we find that the Semantic Suppression Module (SSM) is particularly helpful in detecting *ps.* tampering (from 58.0 to 68.6). Considering that the *ps.* tampering type involves data with rich semantic information about manipulation, this improvement effectively demonstrates the ability of SSM to learn semantic-agnostic feature. Unfortunately, we also find that the introduction of the contrastive learning strategy in the SSM leads to a decline in ability to detect *cpmv.* tampering (from 53.2 to 51.7). This may be due to the fact that there are no significant feature differences between the tampered and authentic regions in copy-move operation, and the repetition of semantic information in the image may aid the detection. However, the SSM is still helpful in detecting copy-move tampering overall (from 49.9 to 51.7).

References

[1] Alexander Buslaev, Vladimir I Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A Kalinin. Albumentations: fast and flexible image augmentations. *Information*, 11(2):125, 2020. 1

[2] 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*, pages 422–426. IEEE, 2013. 1

[3] 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. 1

[4] Silvan Heller, Luca Rossetto, and Heiko Schuldt. The ps-battles dataset-an image collection for image manipulation detection. *arXiv preprint arXiv:1804.04866*, 2018. 2

[5] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1

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

[7] 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. 1

[8] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000. 1

[9] 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. 1