Towards Modern Image Manipulation Localization: A Large-Scale Dataset and Novel Methods

Supplementary Material

Abstract

In this supplementary material, we first present additional details pertaining to the proposed Category-Aware Auto-Annotation (CAAA), MIML dataset and APSC-Net. Subsequently, we present the comparison experiments for constrained image splicing localization on the widely-used synthetic benchmarks. Additionally, we present the performance of the classification model within the CAAA. Furthermore, we present extensive experiments regarding the APSC-Net and the proposed MIML dataset. Finally, we present additional qualitative results for visual comparison.

1. More Details of the CAAA

In this section, we present additional details about the $Corr$ function, the model's structure and training configuration of the proposed Category-Aware Auto-Annotation.

1.1. More Details of the Corr Function

As described in Section 3.3 of the paper, the correlation function is the one widely used in previous works [10, 11, 21]. To be specific, given two feature map $F_a, F_b \in \mathbb{R}^{h\times w\times d}$, and $f_a(i_a, j_a) \in F_a$, $f_b(i_b, j_b) \in F_b$ denote the d-dimension vector at specific positions. The cross-correlation maps $c_{a,b} \in \mathbb{R}^{h\times w\times(h\times w)}$ contain the scalar product of a pair of individual vectors $f_a(i_a, j_a)$, $f_b(i_b, j_b)$ at each position $(i_{a,b}, j_{a,b}, k_{a,b})$, as equation (1).

$$c_{a,b}(i_{a,b}, j_{a,b}, k_{a,b}) = f_a(i_a, j_a)^T f_b(i_b, j_b) \tag{1}$$

in which

$$\begin{aligned} i_b = mod(i_a + i_t, h), \quad & j_b = mod(j_a + j_t, w) \\ i_{a,b} = i_a, \quad j_{a,b} = j_b \quad & and \quad k_{a,b} = w \cdot i_t + j_t \end{aligned} \tag{2}$$

The constraints in equation (2) mean that the correlation maps in the corresponding channel $k_{a,b}$ must satisfy the strong spatial restriction. To reduce the negative impact of uncorrelated signals, the average, maximum and sorted correlation maps are generated as:

$$c^{avg}_{a,b}(i_{a,b}, j_{a,b}) = \frac{1}{h\times w}\sum_{k_{a,b}} c_{a,b}(i_{a,b}, j_{a,b}, k_{a,b}) \tag{3}$$

$$\begin{aligned} c^{max}_{a,b}(i_{a,b}, j_{a,b}) = argmax_{k_{a,b}}(c_{a,b}(i_{a,b}, j_{a,b}, k_{a,b})) \\ where \quad 0 \le k_{ab} \le (h\times w) \end{aligned} \tag{4}$$

$$\begin{aligned} c^{srt}_{a,b}(i_{a,b}, i_{a,b}, k) = c_{a,b}(i_{a,b}, i_{a,b}, k_t) \\ k_t \in Top\text{-}K(sort_{k_{a,b}}(sum(c_{a,b}[:,:,k_{a,b}]))) \end{aligned} \tag{5}$$

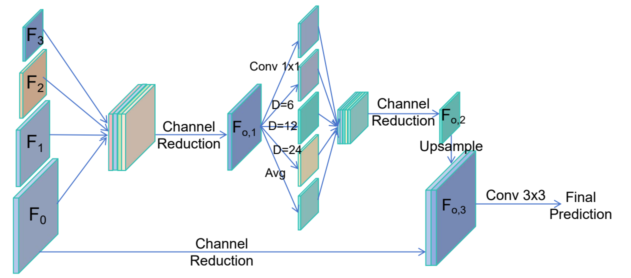

Figure 1. The detailed structure of the decoder in the proposed Difference-Aware Semantic Segmentation and the Semantic Aligned Correlation Matching. 'D' denotes dilated conv-layer with dilation size D. 'Avg' denotes global average-pooling.

where $Top\text{-}K$ denotes the function that selects the indexes of the top-K values (K is empirically set to 14). The resulting output feature maps are denoted by $\hat{c}_{a,b}$ = $[c^{avg}_{a,b}, c^{max}_{a,b}, c^{srt}_{a,b}]$, and $\hat{c}_{a,b} \in \mathbb{R}^{h\times w\times(K+2)}$, in which 2 dimensions correspond to the average and maximum correlation maps, whereas the remaining K dimensions are the sorted correlation maps. Similarly, by replacing F_b with F_a, we can obtain self-correlation maps $\hat{c}_{a,a}$ = $[c^{avg}_{a,a}, c^{max}_{a,a}, c^{srt}_{a,a}]$. For the sake of clarity, we denote the correlation function $Corr$ employed in our model as:

$$Corr(F_a, F_b) = [\hat{c}_{a,b}, \hat{c}_{a,a}] \tag{6}$$

1.2. More Details of the Structure

Both the proposed Difference-Aware Semantic Segmentation (DASS) and the proposed Semantic Aligned Correlation Matching (SACM) employ the encoder-decoder structure. We adopt VAN [5] and ConvNeXt [13] as the encoder backbone model for them respectively. Inspired by DeepLabV3+ [1], we utilize the decoder with dilated conv-layers for both of them, as shown in Fig 1. Given four input features maps F_0, F_1, F_2, F_3, we first resize them to the same resolution as F_1 and concatenate them at the channel dimension. Then, channel reduction is performed to obtain $F_{o,1}$. Consequentially, we extract features from $F_{o,1}$ using conv-layers with dilation $(1, 6, 12, 24)$ and global average-pooling, concatenate the results and reduce the channels to get $F_{o,2}$. Afterwards, $F_{o,2}$ is concatenated with a channel-reduced version of F_0 to get $F_{o,3}$, and $F_{o,3}$ is utilized for the final prediction. For DASS, the input features F_0, F_1, F_2, F_3 are the output of the encoder. For SACM, the input features F_0, F_1, F_2 are the output of the encoder, the F_3 is the correlation features F_{corr}.

1.3. More Details of the Training Configuration

In the experiments in Section 6.1 of the paper, we adopt Cross-Entropy loss and AdamW optimizer [14] with the learning rate linearly decaying from 1e-4 to 1e-6. A sampling ratio of approximately 5:1:1 is utilized for the synthetic, CASIAv2 and IMD20 datasets respectively. The models are trained for 160k iterations with a batch-size of 8. The IMD20 dataset is split into SPG and SDG using the classifier described in Section 4.1 of the paper.

2. More Details of the MIML Dataset

The proposed MIML dataset comprises a total of 123,150 manually forged images, with 76,978 images belonging to the Shared Probe Group and 46,172 images belonging to the Shared Donor Group. The statistics about the image resolution and the proportion of forged area within the MIML dataset are presented in Fig 2.

3. More Details of the APSC-Net

In this section, we present the detailed structure of the Calibration Kernel Mapping Network and the Classification Head of the Self-Calibration module in the APSC-Net. The APSC-Net has a total of 143M parameters.

3.1. More Details of the CKMN

The detailed structure of the Calibration Kernel Mapping Network (CKMN) is presented in Table 1. Firstly, the input mask prediction with a shape of (B, 1, H, W) is downsampled to (B, 1, 64, 64) utilizing bi-linear interpolation. Here ‘B‘ denotes the batch-size. Subsequently, the mask is fed into the CKMN, which produces an output vector of (B, 961). Afterwards, the vector is reshaped into (B, 1, 31, 31) to obtain the calibration kernel.

LayerName	IN_C	OUT_C	K	S	IN_S	OUT_S
Conv-BN-ReLU	1	32	5	2	64	32
Conv-BN-ReLU	32	64	5	2	32	16
Conv-BN-ReLU	64	128	5	2	16	8
Conv-BN-ReLU	128	256	5	2	8	4
Conv-BN-ReLU	256	512	3	1	4	4
Avg-Pooling	512	512	4	1	4	1
Linear	512	961	-	-	1	1

Table 1. Detailed structure of the Calibration Kernel Mapping Network. ‘IN_C‘ denotes the input channels, ‘OUT_C‘ denotes the output channels, ‘K‘ denotes the kernel size of the conv-layer, ‘S‘ denotes the stride of the conv-layer, ‘IN_S‘ denotes the input shape, ‘OUT_S‘ denotes the output shape.

3.2. More Details of the Classification Head

The detailed structure of the Classification Head is presented in Table 2. The classification head takes the concatenation of F_o and F_{ref2} as input, and determines whether the input image is manipulated or not at image-level.

LayerName	IN_C	OUT_C	K	S	IN_S	OUT_S
Conv-BN-ReLU	3072	512	1	1	64	64
Max-Pooling	512	512	2	2	64	32
Conv-BN-ReLU	512	256	3	1	32	32
Max-Pooling	256	256	2	2	32	16
Conv-BN-ReLU	256	256	3	1	16	16
Max-Pooling	256	256	16	1	16	1
Linear	256	2	-	-	1	1

Table 2. Detailed structure of the Classification Head. ‘IN_C‘ denotes the input channels, ‘OUT_C‘ denotes the output channels, ‘K‘ denotes the kernel size of the conv-layer, ‘S‘ denotes the stride of the conv-layer, ‘IN_S‘ denotes the input shape, ‘OUT_S‘ denotes the output shape.

3.3. Comparison between previous methods

There have been many designs for image manipulation localization, however, our APSC-Net differs from previous methods in the following aspects:

For multi-view perception, previous methods simply concatenate different feature maps in the channel dimension (*e.g.* CAT-Net [8], MVSS-Net [2]). While ours fuses different feature maps with adaptive weights.

For prediction refinement, previous works (*e.g.* PSCC-Net [9]) initialize with the coarsest prediction derived from the highest level feature map. While ours initializes with the finest prediction calibrated with a learnable kernel.

4. Extensive Experiments

In this section, we conduct extensive experiments to further evaluate the effectiveness of the proposed MIML dataset, Category-Aware Auto-Annotation and APSC-Net.

4.1. Comparison Experiments for MIML

To further evaluate the effectiveness of the proposed MIML dataset, we replace it with DEFACTO [15], a dataset for image manipulation localization that synthesized with elaborately designed pipelines. We re-train the PSCC-Net [9] and CAT-Net [8] utilizing this dataset with the same training configuration and sampling ratio as that of MIML. As shown in Table 3, the incorporation of DEFACTO does not lead to an discernible improvement in the models' performance. In contrast, the inclusion of MIML significantly enhances the models. It is the high-quality of our MIML dataset that brings the improvement, rather than the mere increase in size and diversity of the training data.

Dataset	PSCC-Net [9]									
	IoU					F1				
	Ori	+DEFACTO	+Ours	gain(DEFACTO)	gain(Ours)	Ori	+DEFACTO	+Ours	gain(DEFACTO)	gain(Ours)
CASIAv1 [3]	.401	.394	.609	-2%	**+52%**	.430	.429	.649	-0%	**+51%**
NIST16 [4]	.247	.223	.402	-10%	**+62%**	.295	.270	.476	-8%	**+61%**
Coverage [20]	.197	.231	.395	+17%	**+100%**	.218	.256	.477	+17%	**+118%**
IMD20 [16]	.125	.137	.470	+10%	**+277%**	.156	.171	.541	+10%	**+247%**
Average	.243	.246	.469	+2%	**+93%**	.275	.282	.536	+2%	**+95%**

Dataset	CAT-Netv2 [8]									
	IoU					F1				
	Ori	+DEFACTO	+Ours	gain(DEFACTO)	gain(Ours)	Ori	+DEFACTO	+Ours	gain(DEFACTO)	gain(Ours)
CASIAv1 [3]	.660	.673	.691	+2%	**+5%**	.703	.715	.728	+2%	**+4%**
NIST16 [4]	.239	.220	.353	-8%	**+48%**	.287	.261	.422	-9%	**+47%**
Coverage [20]	.245	.200	.302	-18%	**+23%**	.286	.230	.389	-19%	**+36%**
IMD20 [16]	.157	.164	.547	+4%	**+248%**	.192	.200	.629	+4%	**+228%**
Avgrage	.325	.314	.473	-3%	**+46%**	.367	.352	.542	-4%	**+48%**

Table 3. Comparison study on the proposed MIML dataset. '+DEFACTO' denotes the inclusion of DEFACTO dataset during training. '+Ours' denotes the inclusion of our MIML dataset during training. 'gain' denotes the ratio of improvement in performance.

4.2. Extensive CIML Experiments

For a comprehensive comparison with the previous constrained image splicing localization methods, we retrain the proposed Semantic Aligned Correlation Matching model with a million synthetic data for 6 epochs, fixing the input size to 256 $\times$ 256 following the previous works [18, 23]. The model is evaluated on the widely-used synthetic datasets, Combination Sets [11]. As shown in Table 4, our model achieves state-of-the-art performance.

Method	Difficult			Normal		
	IoU	MCC	NMM	IoU	MCC	NMM
DMVN [21]	.2772	.3533	-.4382	.6818	.7570	.4042
DMAC [11]	.5433	.6584	.1026	.8317	.8833	.6877
AttentionDM [10]	.7228	.8108	.4793	.8980	.9320	.8253
SADM [23]	.7759	.8128	.5129	.9040	.8288	.8265
MSTAF [18]	.8394	.8918	.7064	.9510	.9700	.9151
Ours	**.8507**	**.9132**	**.7371**	**.9548**	**.9725**	**.9291**

Table 4. Comparison study for the proposed Semantic Aligned Correlation Matching model on the Combination Sets [11].

4.3. Classification Performance Evaluation

To evaluate the performance of the classification model within the proposed Category-Aware Auto-Annotation, we randomly picked 500 image pairs from the IMD20 dataset, and manually divided them into SPG and SDG, resulting in a final tally of 258 SDG and 242 SPG pairs. Considering that a very small proportion of SPG image pairs are not spatially aligned, which could negatively impact the prediction's quality, we also include a linear classification layer to filter them out. To construct the misaligned SPG image pair for training, we randomly crop a rectangular region from an image in an SPG pair and resize the region to its source image's resolution. Totally, 14 pairs from the annotated SPG are misaligned. The classification results are presented in Table 5. It is evident that the voting ensemble of the classification models produces accurate enough outcomes.

4.4. Extensive Experiments for APSC-Net

Comparison Study for APSC-Net. We further fine-tune our pre-trained APSC-Net following the widely-used training splits [17, 24] of the specific datasets, and compare it with SOTA methods on the remaining testing splits. As shown in Table 6, our APSC-Net still outperforms SOTA methods, showing its strong generalization ability.

Robustness Evaluation for APSC-Net We evaluate the robustness of the pre-trained APSC-Net on NIST16 with the AUC metric following the standard setting in previous works [17, 24]. As shown in Table 7, our APSC-Net shows satisfactory robustness against the common distortions.

Ablation Study for APSC-Net. The Adaptive Perception (AP) module is designed to enable the model to adaptively select an optimal combination of observations. The Segmentation-based Self Calibration (SSC) and Classification-based Self Calibration (CSC) are designed to assist the model in getting more accurate predictions by in-depth analyses with its initial predictions. We conduct ablation study to verify the effectiveness of these components. As shown in Table 8, all of the proposed modules contribute towards a higher performance of the APSC-Net.

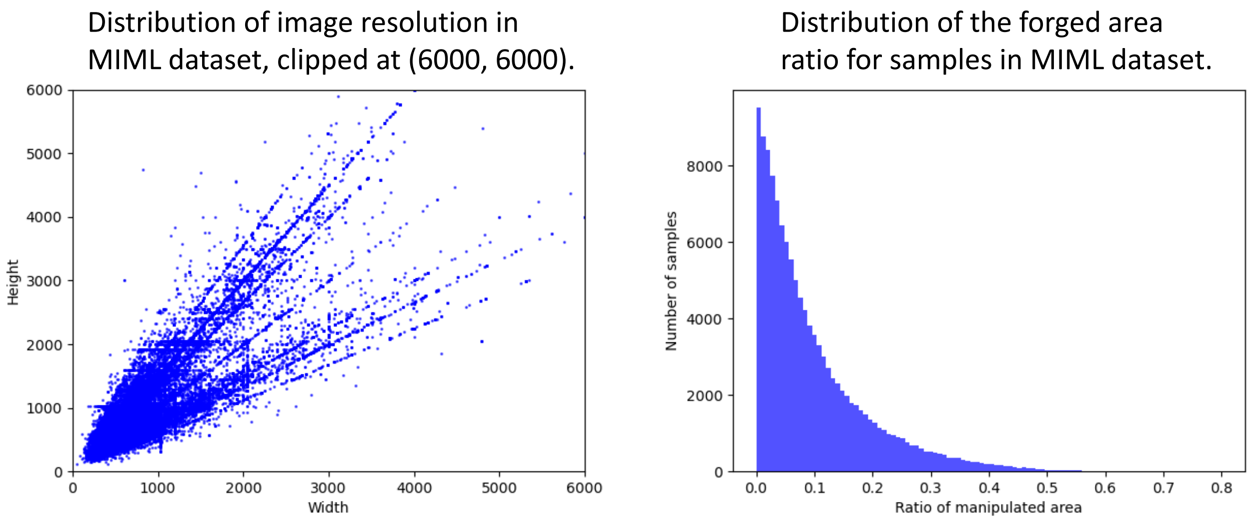

Figure 2. Some statistics of our MIML dataset.

method	SDG			SPG			Misaligned		
	P	R	F	P	R	F	P	R	F
DiNAT [6]	.992	.992	.992	1	.996	.998	.867	.929	.897
SwinTrans [12]	.996	.981	.988	1	.996	.998	.737	1	.849
ConvNeXt [13]	.996	.992	.994	1	.996	.998	.875	1	.933
Ensemble	.996	1	.998	1	.996	.998	1	1	1

Table 5. Classification experiments for SDG, SPG and misaligned SPG. 'P' denotes precision, 'R' denotes recall, 'F1' denotes F1-score. 'Ensemble' denotes the ensemble of the three models.

Method	CASIAv1		NIST16		Coverage	
	AUC	F1	AUC	F1	AUC	F1
RGB-N [25]	.795	.408	.937	.722	.817	.437
SPAN [7]	.838	.382	.961	.582	.937	.558
MVSS-Net [2]	.877	.522	.942	.814	.849	.504
CL-Net [26]	.895	.584	.985	.823	.857	.512
PSCC-Net [9]	.875	.554	.996	.819	.941	.723
ObjectFormer [19]	.882	.579	.996	.824	.957	.758
NCL [24]	.864	.598	.912	.831	.928	.801
SAFL-Net [17]	.908	.740	.997	.879	.970	.803
Ours (w/ MIML)	**.983**	**.860**	**.998**	**.914**	**.976**	**.878**

Table 6. IML comparison study for the fine-tuned models.

5. Visualization

In this section, we present qualitative results for our MIML dataset and APSC-Net. The qualitative results for ablation study on our MIML dataset are shown in Fig 3, the qualitative results for comparison study on the pre-trained APSC-Net are shown in Fig 4, and the qualitative results for ablation study on our APSC-Net are shown in Fig 5.

Method	Ori	Resize		Blur		JPEG	
		.78×	.25×	k=3	k=15	q=100	q=50
ManTra-Net [22]	.795	.774	.755	.774	.746	.779	.744
MVSS-Net [2]	.788	.783	.775	.786	.758	.788	.788
SPAN [7]	.840	.832	.802	.831	.792	.836	.807
PSCC-Net [9]	.855	.853	.850	.854	.800	.854	.854
ObjectFormer [19]	.872	.872	.863	.860	.803	.864	.862
SAFL-Net [17]	.888	.884	.869	.881	.877	.886	.881
NCL [24]	.912	.856	.831	.840	.806	.843	.819
Ours (w/ MIML)	**.928**	**.917**	**.888**	**.907**	**.901**	**.922**	**.907**

Table 7. IML robustness evaluation on the NIST16 dataset. 'Ori' denotes no distortion, 'k' denotes the kernel size of Gaussian Blur and 'q' denotes the quality of JPEG compression.

Setting	AP	SSC	CSC	MIML	CASIAv1		NIST16		IMD20	
					IoU	F1	IoU	F1	IoU	F1
(1)					.711	.779	.346	.410	.273	.342
(2)	✓				.763	.805	.350	.393	.315	.368
(3)	✓	✓			.798	.833	.387	.424	.331	.381
(4)	✓	✓	✓		.799	.837	.398	.436	.339	.391
(5)	✓	✓	✓	✓	**.810**	**.848**	**.525**	**.590**	**.679**	**.760**

Table 8. IML ablation study for the APSC-Net. 'AP' denotes the proposed Adaptive Perception module. 'SSC' denotes the proposed Segmentation-based Self Calibration. 'CSC' denotes the proposed Classification-based Self Calibration.

References

[1] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 1

[2] Chengbo Dong, Xinru Chen, Ruohan Hu, Juan Cao, and Xirong Li. Mvss-net: Multi-view multi-scale supervised networks for image manipulation detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3539–3553, 2022. 2, 4

[3] Jing Dong, Wei Wang, and Tieniu Tan. Casia image tampering detection evaluation database. In *2013 IEEE China Summit and International Conference on Signal and Infor-*

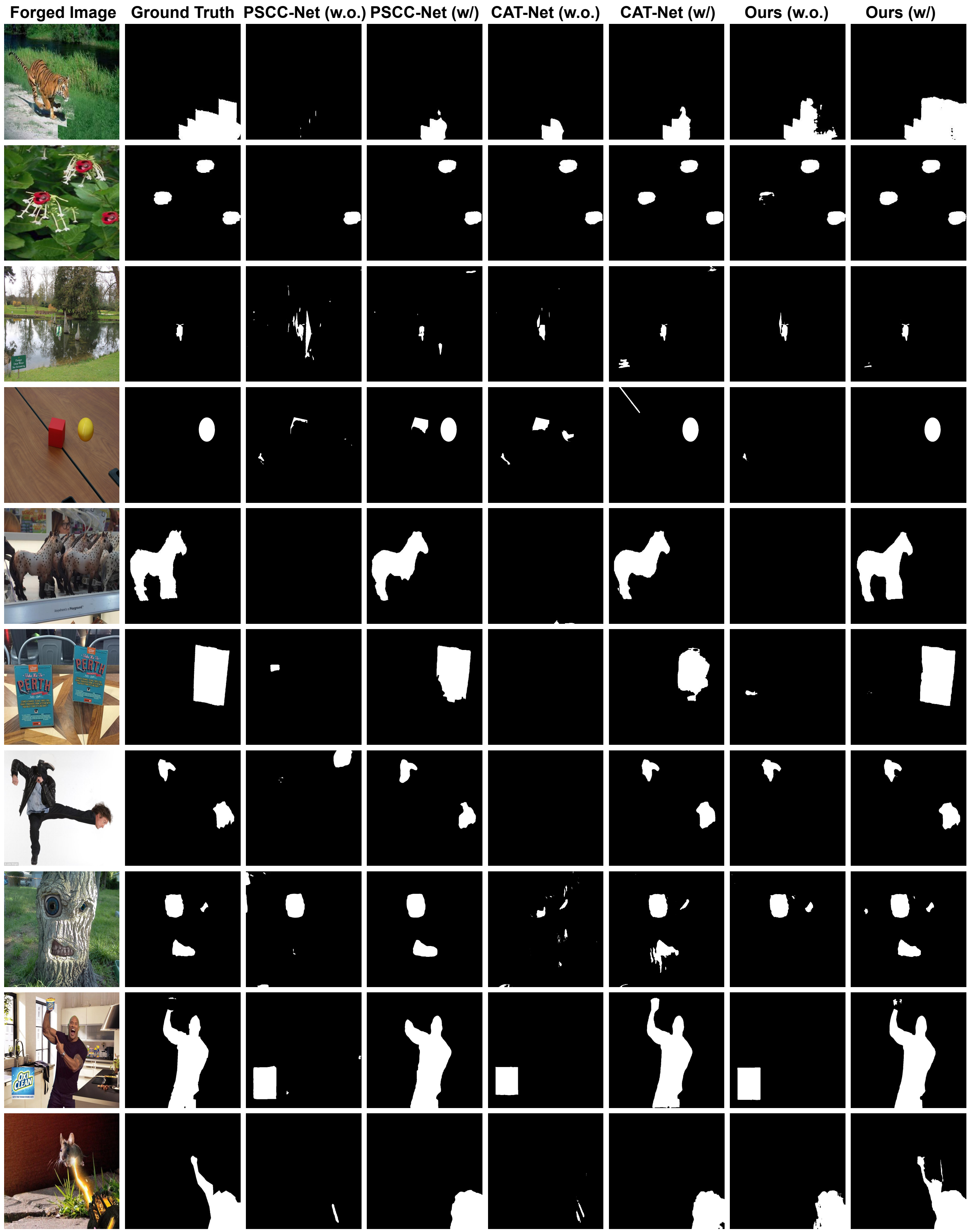

Figure 3. Qualitative results for ablation study on our MIML dataset across the CASIAv1, NIST16, Coverage and IMD20 datasets.

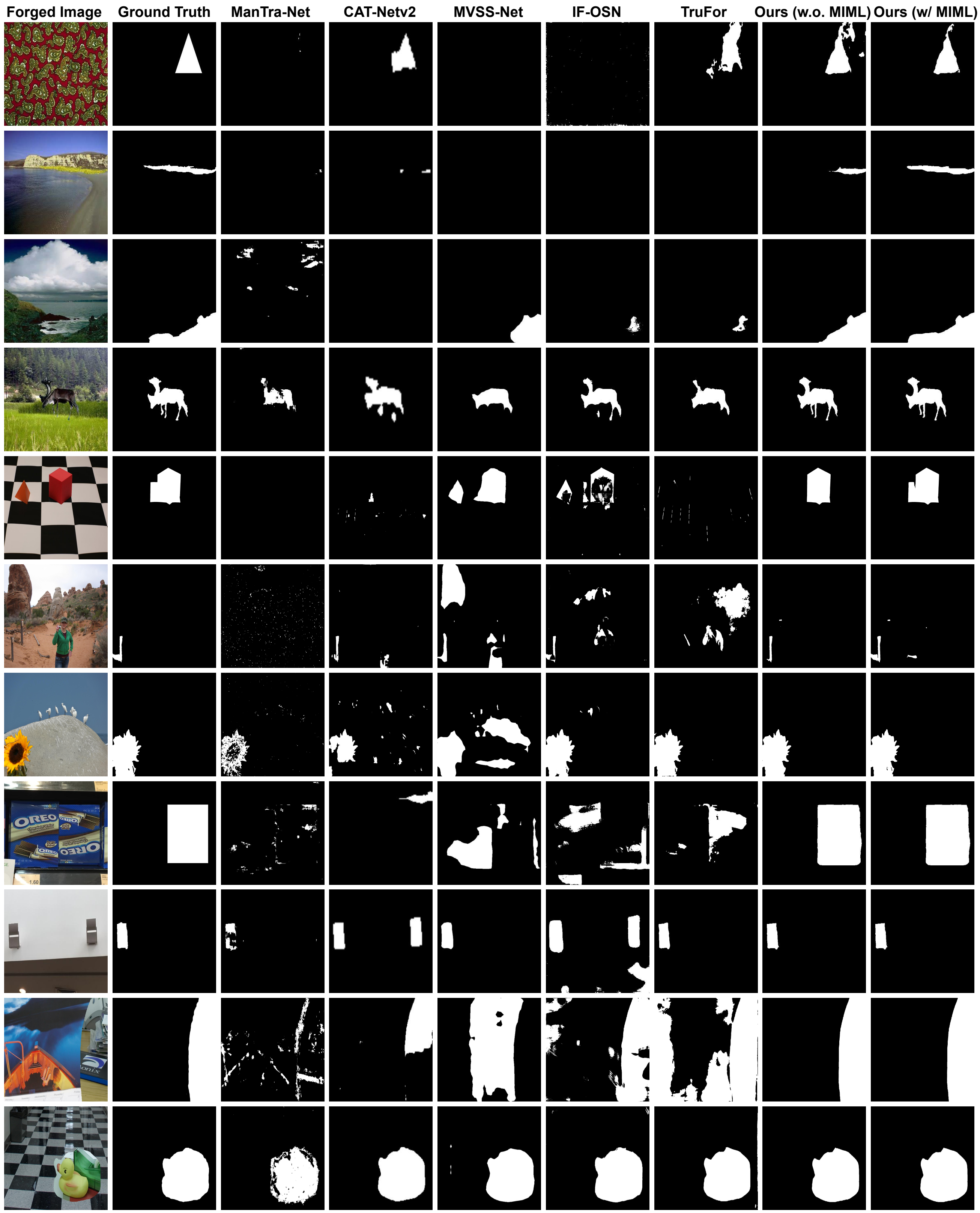

Figure 4. Qualitative results for comparison study on our APSC-Net across the CASIAv1, NIST16, Coverage and Columbia datasets.

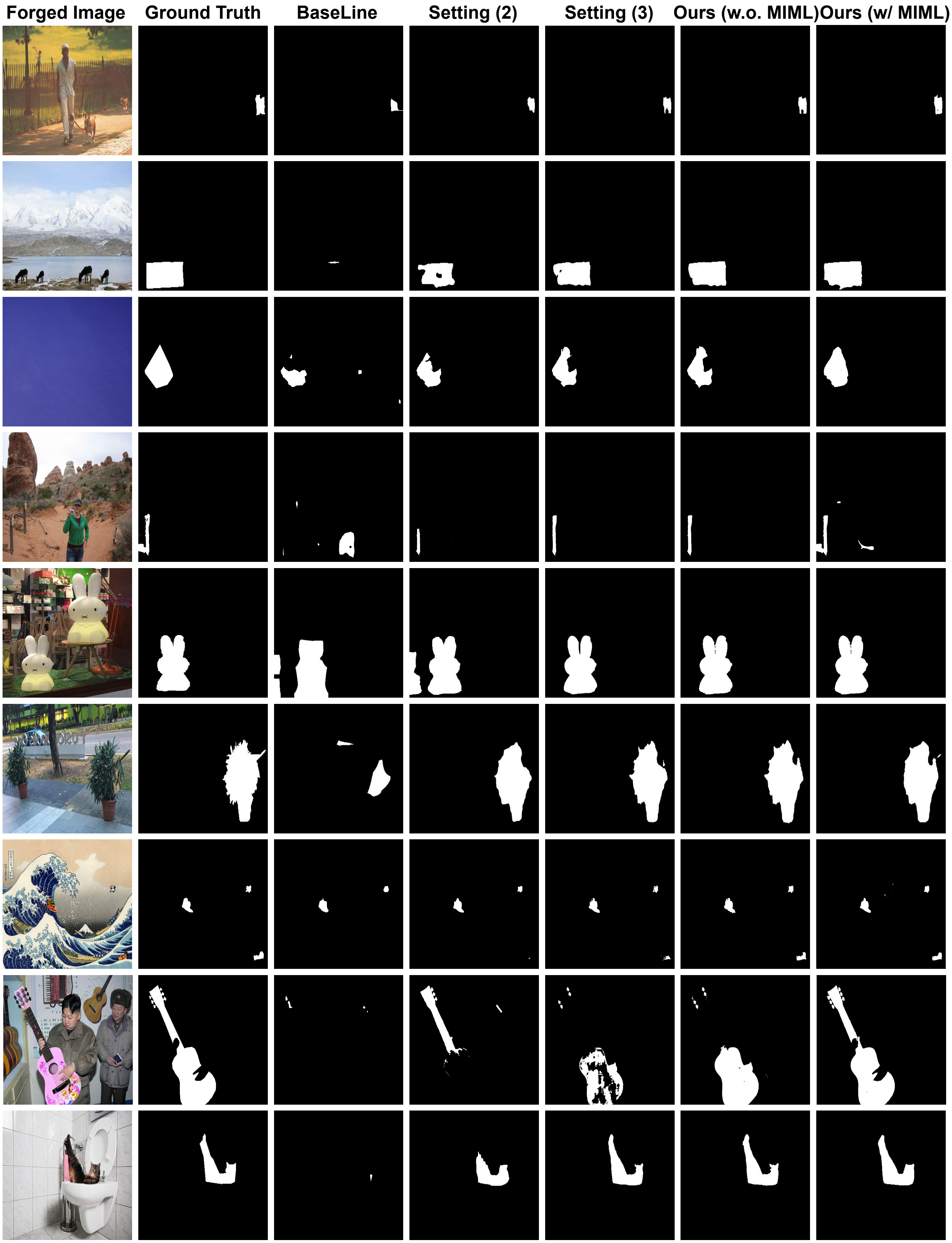

Figure 5. Qualitative results for ablation study conducted on our APSC-Net across the CASIAv1, NIST16, Coverage and IMD20 datasets. ‘Setting (2)‘ denotes the inclusion of the Adaptive Perception module. ‘Setting (3)‘ denotes the inclusion of both the Adaptive Perception module and the Segmentation-based Self-Calibration.

mation Processing, pages 422–426, 2013. 3

[4] 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. 3

[5] Meng-Hao Guo, Cheng-Ze Lu, Zheng-Ning Liu, Ming-Ming Cheng, and Shi-Min Hu. Visual attention network. *Computational Visual Media*, 9(4):733–752, 2023. 1

[6] Ali Hassani, Steven Walton, Jiachen Li, Shen Li, and Humphrey Shi. Neighborhood attention transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6185–6194, 2023. 4

[7] Xuefeng Hu, Zhihan Zhang, Zhenye Jiang, Syomantak Chaudhuri, Zhenheng Yang, and Ram Nevatia. Span: Spatial pyramid attention network for image manipulation localization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 312–328. Springer, 2020. 4

[8] Myung-Joon Kwon, Seung-Hun Nam, In-Jae Yu, Heung-Kyu Lee, and Changick Kim. Learning jpeg compression artifacts for image manipulation detection and localization. *International Journal of Computer Vision*, 130(8):1875–1895, 2022. 2, 3

[9] Xiaohong Liu, Yaojie Liu, Jun Chen, and Xiaoming Liu. Pscc-net: Progressive spatio-channel correlation network for image manipulation detection and localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 32 (11):7505–7517, 2022. 2, 3, 4

[10] Yaqi Liu and Xianfeng Zhao. Constrained image splicing detection and localization with attention-aware encoder-decoder and atrous convolution. *IEEE Access*, 8:6729–6741, 2020. 1, 3

[11] Yaqi Liu, Xiaobin Zhu, Xianfeng Zhao, and Yun Cao. Adversarial learning for constrained image splicing detection and localization based on atrous convolution. *IEEE Transactions on Information Forensics and Security*, 14(10):2551–2566, 2019. 1, 3

[12] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 4

[13] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 1, 4

[14] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proceedings of the International Conference on Learning Representations*, pages 10012–10022, 2019. 2

[15] 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. 2

[16] 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 (WACV) Workshops*, 2020. 3

[17] Zhihao Sun, Haoran Jiang, Danding Wang, Xirong Li, and Juan Cao. 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*, pages 22424–22433, 2023. 3, 4

[18] Yuxuan Tan, Yuanman Li, Limin Zeng, Jiaxiong Ye, Xia Li, et al. Multi-scale target-aware framework for constrained image splicing detection and localization. *arXiv preprint arXiv:2308.09357*, 2023. 3

[19] Junke Wang, Zuxuan Wu, Jingjing Chen, Xintong Han, Abhinav Shrivastava, Ser-Nam Lim, and Yu-Gang Jiang. Objectformer for image manipulation detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2364–2373, 2022. 4

[20] 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, 2016. 3

[21] Yue Wu, Wael Abd-Almageed, and Prem Natarajan. Deep matching and validation network: An end-to-end solution to constrained image splicing localization and detection. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1480–1502, 2017. 1, 3

[22] Yue Wu, Wael AbdAlmageed, and Premkumar Natarajan. 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*, pages 9543–9552, 2019. 4

[23] Shengwei Xu, Shanlin Lv, Yaqi Liu, Chao Xia, and Nan Gan. Scale-adaptive deep matching network for constrained image splicing detection and localization. *Applied Sciences*, 12 (13):6480, 2022. 3

[24] Jizhe Zhou, Xiaochen Ma, Xia Du, Ahmed Y Alhammadi, and Wentao Feng. Pre-training-free image manipulation localization through non-mutually exclusive contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22346–22356, 2023. 3, 4

[25] Peng Zhou, Xintong Han, Vlad I Morariu, and Larry S Davis. Learning rich features for image manipulation detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1053–1061, 2018. 4

[26] Peng Zhou, Bor-Chun Chen, Xintong Han, Mahyar Najibi, Abhinav Shrivastava, Ser-Nam Lim, and Larry Davis. Generate, segment, and refine: Towards generic manipulation segmentation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 13058–13065, 2020. 4