Enhancing Tampered Text Detection through Frequency Feature Fusion and Decomposition (Supplementary Material)

Zhongxi Chen[1,2*†], Shen Chen[2†], Taiping Yao[2✉], Ke Sun[1], Shouhong Ding[2], Xianming Lin[1✉], Liujuan Cao[1], and Rongrong Ji[1]

[1] Key Laboratory of Multimedia Trusted Perception and Effcient Computing, Ministry of Education of China, Xiamen University, 361005, P.R. China.
[2] Youtu Lab, Tencent, China.
chenzhongxi@stu.xmu.edu.cn

In light of the main paper's page constraints, we have included comprehensive details and extended results in this supplementary document. The content is structured as follows: Section 1 elucidates the evaluation metrics employed in our study. Section 2 details the reproduction process of the competing models. In Section 3, we explore our model's performance across a spectrum of frequency domain inputs. Visual representations, including Grad-CAM, features after Wavelet-like Frequency Enhancement and model predictions, are provided in Section 4. Lastly, Section 5 addresses the limitations of our research.

1 Evaluation Metrics

To rigorously assess the performance of the proposed model, we employ a suite of evaluation metrics that are standard in the field of computer vision and object detection. These metrics provide a comprehensive understanding of the model's effectiveness from various perspectives:

- **Precision (P)**: This metric quantifies the accuracy of the positive predictions made by the model. It is defined as the ratio of true positive detections to the total number of positive detections (both true positives and false positives). Mathematically, it is expressed as $P = \frac{TP}{TP+FP}$, where TP represents true positives and FP represents false positives.
- **Recall (R)**: Also known as sensitivity, this metric measures the model's ability to correctly identify all relevant instances. It is calculated as the ratio of true positive detections to the actual number of positive samples, which includes both true positives and false negatives. The formula for recall is $R = \frac{TP}{TP+FN}$, with FN denoting false negatives.
- **F1-score**: The F1-score is the harmonic mean of precision and recall, providing a single score that balances both the false positives and false negatives. It is particularly useful when the class distribution is imbalanced. The F1-score is computed using the relation $F1 = \frac{2 \cdot P \cdot R}{P+R}$.

* Work done during an internship at YouTu Lab, Tencent.
† Equal Contribution. ✉ Corresponding authors.

Table 1: Performance comparison of our model with the integration of different frequency domain features.

	DocTamper-T				DocTamper-FCD				DocTamper-SCD			
	IoU	P	R	F	IoU	P	R	F	IoU	P	R	F
+ RGB	0.842	0.804	0.762	0.782	0.602	0.783	0.652	0.712	0.628	0.731	0.716	0.723
+ High Pass	0.839	0.803	0.762	0.782	0.589	0.767	0.654	0.706	0.625	0.730	0.720	0.725
+ NoisePrint	0.844	0.807	0.768	0.787	0.617	0.750	0.675	0.711	0.649	0.740	0.732	0.736
+ SRM	0.849	0.814	0.770	0.791	0.636	0.777	0.682	0.726	0.655	0.740	0.720	0.730
+ FFT	0.856	0.818	0.794	0.806	0.700	0.809	0.750	0.779	0.664	0.729	0.770	0.748
+ Bayar	0.867	0.833	0.795	0.813	0.714	0.816	0.768	0.791	0.681	0.750	0.769	0.760
+ DCT	0.895	0.873	0.840	0.857	0.878	0.927	0.905	0.916	0.748	0.806	0.818	0.812

– **Intersection over Union (IoU)**: This metric, also known as the Jaccard index, evaluates the overlap between the predicted and ground truth bounding boxes. It is defined as the size of the intersection divided by the size of the union of the predicted and true bounding box. The formula for IoU is $IoU = \frac{TP}{TP+FP+FN}$.

These metrics collectively provide a robust framework for evaluating the proposed model's detection capabilities, ensuring a well-rounded analysis of its performance.

2 Detailed Reproduction of Competing Models

In our study, we replicated the performance of ten different competing models. These include UperNet [14], SegFormer [15], Swin-UPer [8], Mask2Former [1], ConvNext [9], ConvNextV2 [13], InterImage [12] and DTD [10], all of which were implemented using the mmsegmentation framework [2]. PCSS-Net [7], CAT-Net [6] were reproduced using publicly available code. For consistency, we employed the default configurations provided by mmsegmentation [2], which encompassed settings for the optimizer, learning rate scheduler, and other parameters. Notably, during our reproduction efforts, we identified and corrected data loading errors in the publicly available test code for DTD [10]. These errors had previously caused a marked decline in the performance of the pretrained models. After making the necessary corrections, our reproduced model's performance aligned closely with the results reported in the original publications.

3 Evaluation with Varied Frequency Domain Inputs

Our investigation into the influence of different frequency domain inputs on our model's performance is documented in Table 1. We systematically integrated a series of frequency-specific features—High Pass, SRM [4], NoisePrint [3], Fast Fourier Transform (FFT), Bayar convolution, and Discrete Cosine Transform (DCT)—into the model. To mitigate the effects of potential complexity increases,

we included a configuration that incorporated RGB data as an additional input for comparison.

The experimental outcomes reveal that the inclusion of Bayar convolution and FFT features notably improves the model's performance. The most substantial enhancement, however, is achieved with the addition of DCT features, underscoring their pivotal role in bolstering the model's ability to detect tampered content.

4 Additional Visualizations

4.1 Visualizations of Grad-CAM

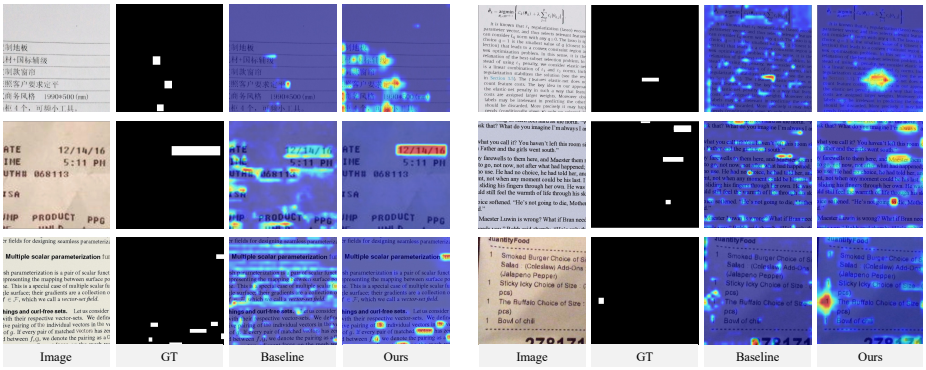

Fig. 1: Comparative visualizations of Class Activation Mapping (CAM) highlighting the detection areas in both the baseline model and our proposed Feature Fusion and Decomposition Network (FFDN). The baseline model activations are concentrated at the tampered edges, while the FFDN model shows expanded activations that cover both the edges and the internal textures of the tampered regions, demonstrating the FFDN's superior capability to identify and analyze tampering traces.

To demonstrate the effectiveness of our Feature Fusion and Decomposition Network (FFDN), we compared Class Activation Mapping (CAM) [11] visualizations of our model against a baseline model, both trained with identical hyper-parameters on RGB and DCT inputs. Our FFDN's Wavelet-like Frequency Enhancement (WFE) module is designed to explicitly decompose and separate features into high and low-frequency bands, enhancing the model's ability to detect and retain high-frequency tampering indicators.

As discussed in the main text, potential high-frequency tampering traces include artifacts primarily found at the tampering edges, such as blending traces left behind, and inconsistencies within the tampered area's texture, which may arise from transformations like affine changes.

As depicted in Figure 1, the activation areas of the baseline model are predominantly located at the edges of the tampered regions, suggesting that the baseline model primarily detects tampering by identifying blending traces at the edges. In contrast, the activation areas of our FFDN model are notably more extensive, encompassing not only the edges but also the internal texture information of the tampered regions, indicating a more holistic analysis of the tampered regions.

This comprehensive detection is made possible by the WFE's proficiency in preserving high-frequency textures, which are essential for identifying subtle tampering traces. Consequently, our FFDN model not only outperforms the baseline in detecting tampering but also provides a more detailed representation of the tampered areas, leading to a significant advancement in tampering detection accuracy.

4.2 Visualization of Features After WFE

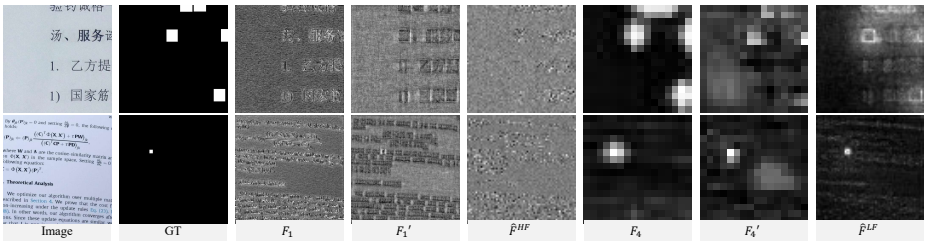

Fig. 2: Visualization of high- and low-frequency components before and after applying our Wavelet-like Frequency Enhancement (WFE) module

Figure 2 illustrates the high- and low-frequency components before and after the application of our Wavelet-like Frequency Enhancement (WFE) module. Among them, F_1 and F_4 are the original features, F_1' and F_4' denote the features enhanced by this module. The $\hat{F}^{HF}$ and $\hat{F}^{LF}$ symbolize the aggregated high-frequency and low-frequency features, respectively. The enhanced high-frequency component exhibits a significant improvement in delineating subtle boundary traces. Correspondingly, the aggregated low-frequency features more precisely demarcate the tampered regions.

4.3 Visualizations of Model Ablation

To demonstrate the distinct impact of our model's components, we conducted a visual ablation study, the results of which are depicted in Figure 3. Performance comparisons were drawn among four configurations: the baseline ConvNextV2 [13] model enhanced with DCT information, the model sans Visual

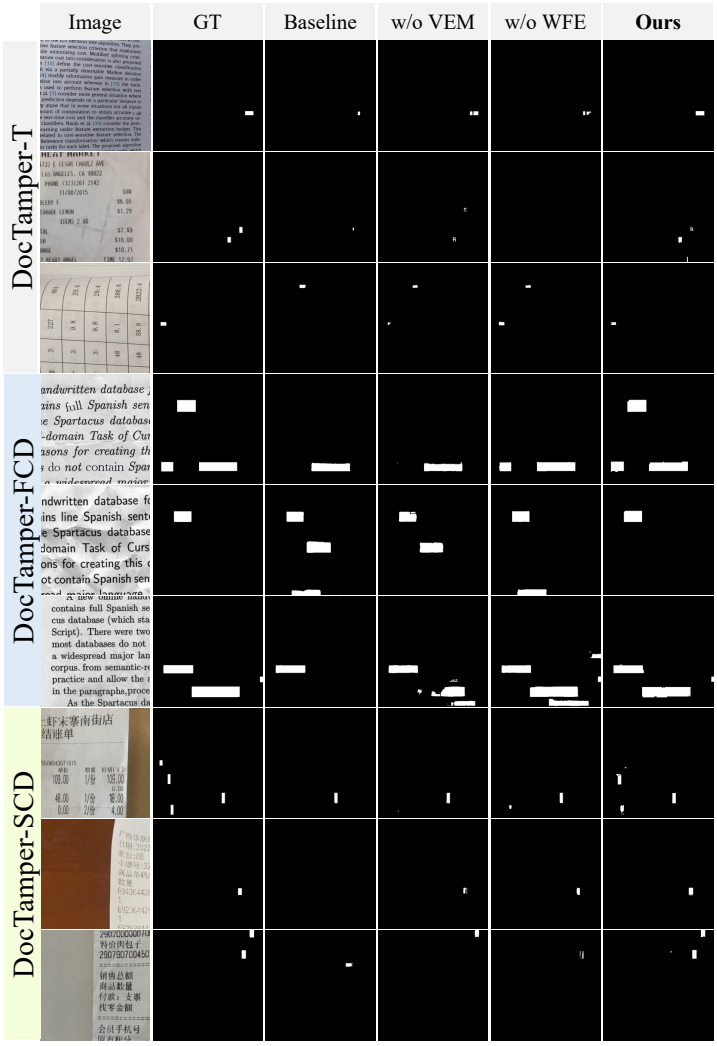

Fig. 3: A visual ablation study showcasing the performance impact of individual components in our model. The comparison includes the baseline ConvNextV2 model with DCT information, the model without the Visual Enhancement Module (w/o VEM), the model without the Wavelet-like Frequency Enhancement (w/o WFE), and our fully integrated model.

Enhancement Module (w/o VEM), the model lacking Wavelet-like Frequency Enhancement (w/o WFE), and our fully equipped model.

The comparative analysis revealed that both VEM and WFE modules contribute to performance gains, notably in the precise detection and localization of tampered areas. Furthermore, when combined, these modules operate in harmony, yielding a synergistic effect that significantly elevates the overall efficacy of our model.

4.4 More Visualizations of Model Predictions

We provide additional visualizations of model predictions on the DocTamper-T, DocTamper-FCD, and DocTamper-SCD datasets [10] in Figure 4, Figure 5, and Figure 6, respectively. These visualizations further demonstrate the robustness and effectiveness of our model in detecting tampered text regions across diverse tampering scenarios. Also the visualizations of model predictions on the T-SROIE dataset [5] are shown in Figure 7.

5 Limitations

Our research marks progress in the field of document image tampering detection, yet it is imperative to recognize its limitations. One primary constraint is the potential vulnerability of our model to sophisticated tampering methods. For instance, large-scale generative models like stable-diffusion can closely replicate genuine image attributes. These methods produce highly realistic tampered images that may evade detection by our system, highlighting the need for continuous improvement and adaptation to new tampering techniques.

Moreover, our research has focused on the visual domain, and integrating multimodal data sources could be explored to further enhance tampering detection capabilities. Future work could address these limitations by developing more efficient models, expanding dataset diversity, and exploring multimodal approaches.

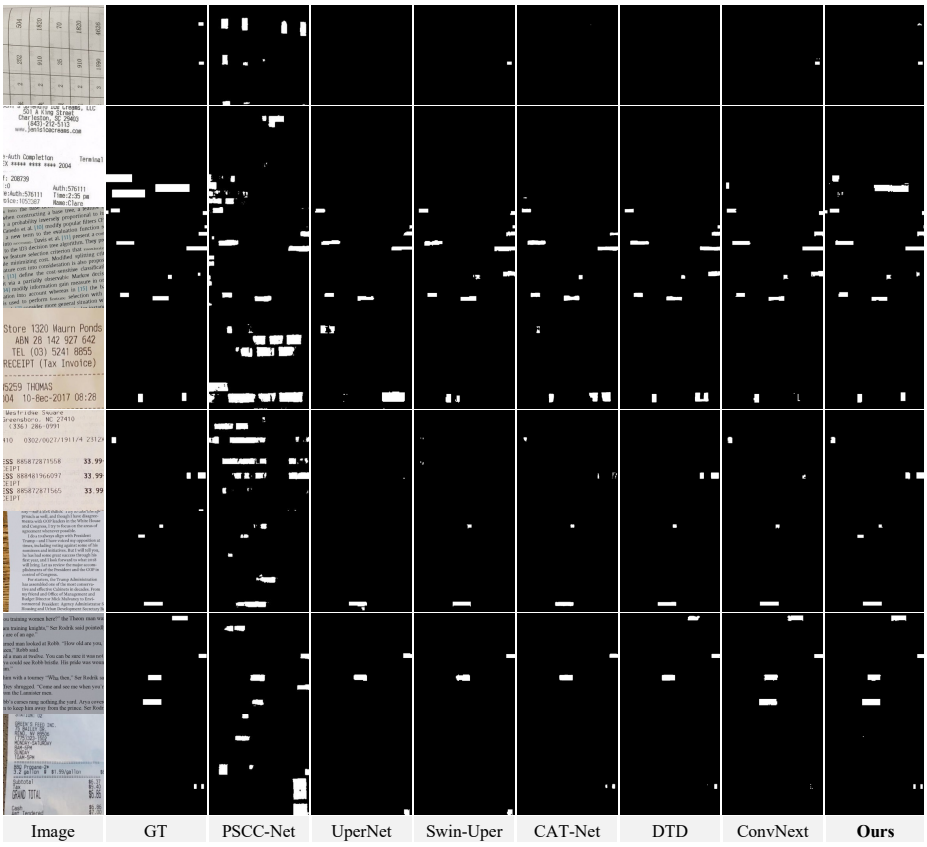

Fig. 4: Additional visualizations of model predictions on DocTamper-T dataset.

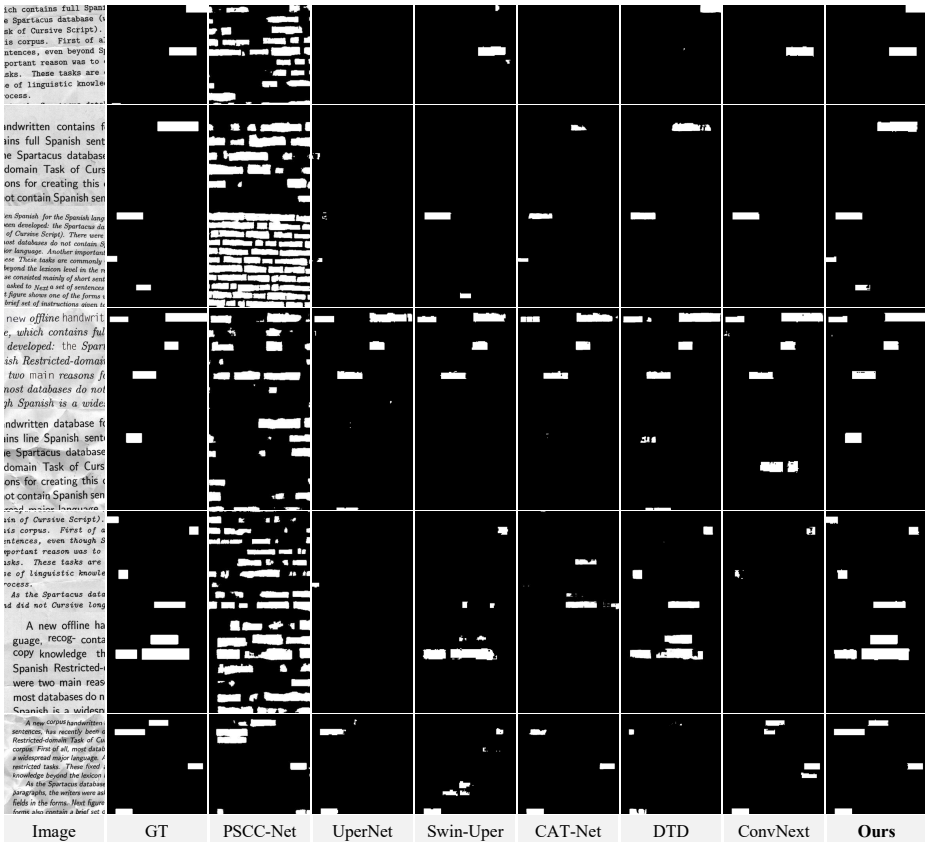

Fig. 5: Additional visualizations of model predictions on DocTamper-FCD dataset.

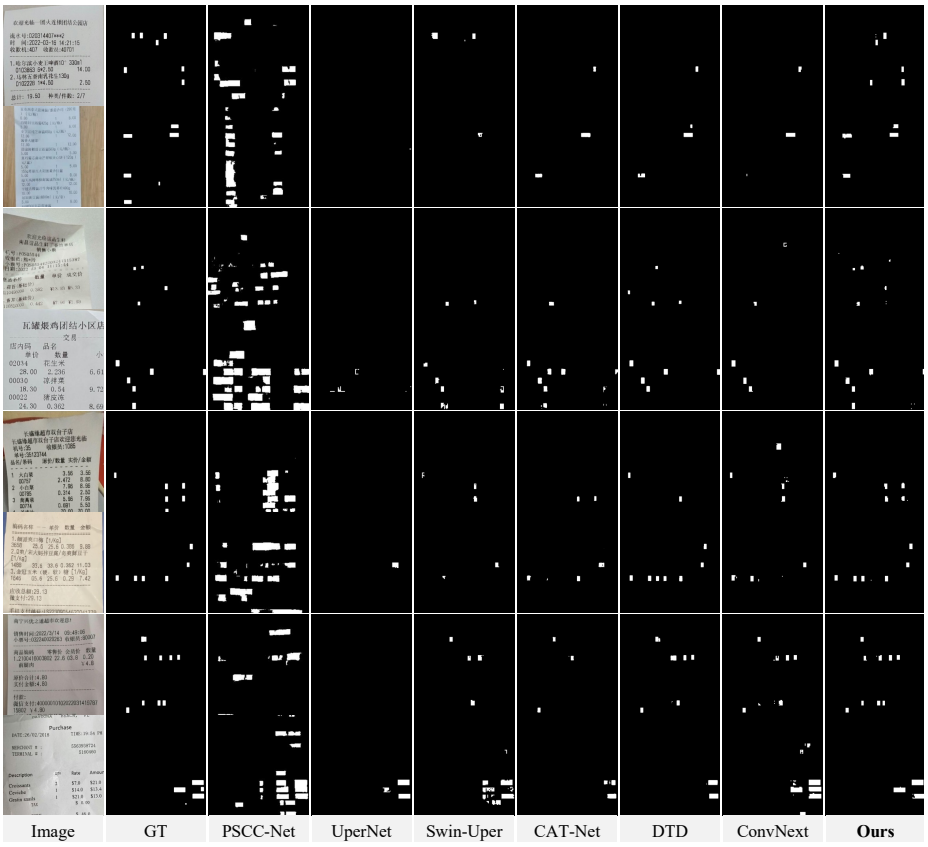

Fig. 6: Additional visualizations of model predictions on DocTamper-SCD dataset.

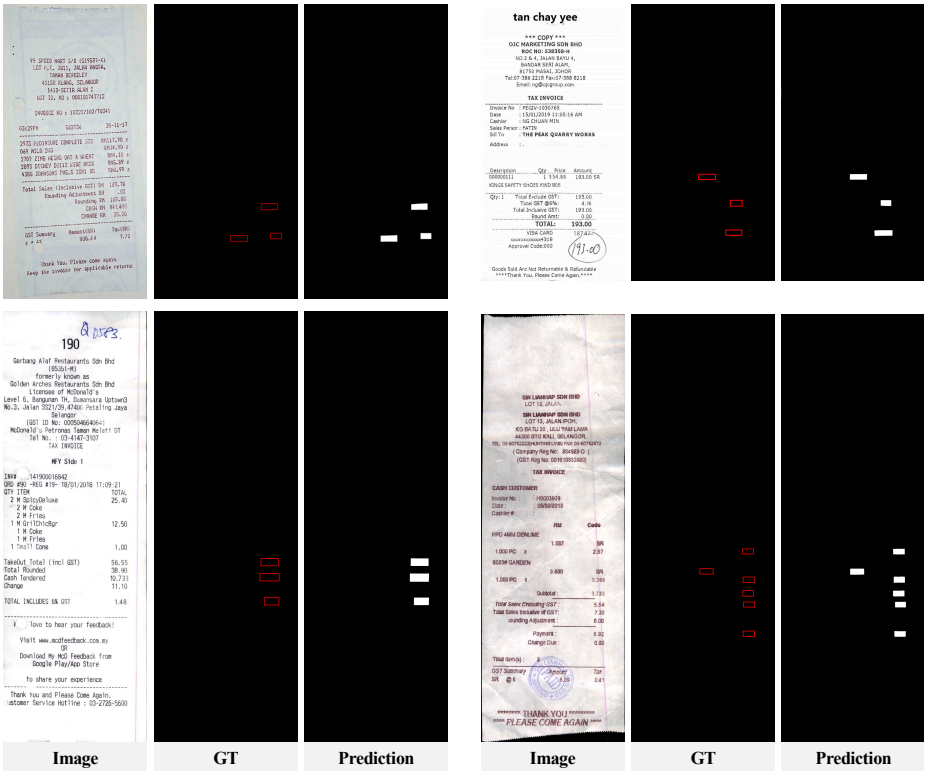

Fig. 7: Additional visualizations of model predictions on the T-SROIE dataset.

References

1. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1290–1299 (2022)
2. Contributors, M.: MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. `https://github.com/open-mmlab/mmsegmentation` (2020)
3. Cozzolino, D., Verdoliva, L.: Noiseprint: A cnn-based camera model fingerprint. IEEE Transactions on Information Forensics and Security **15**, 144–159 (2019)
4. Fridrich, J., Kodovsky, J.: Rich models for steganalysis of digital images. IEEE Transactions on information Forensics and Security **7**(3), 868–882 (2012)
5. Huang, Z., Chen, K., He, J., Bai, X., Karatzas, D., Lu, S., Jawahar, C.: Icdar2019 competition on scanned receipt ocr and information extraction. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 1516–1520. IEEE (2019)
6. Kwon, M.J., Nam, S.H., Yu, I.J., Lee, H.K., Kim, C.: Learning jpeg compression artifacts for image manipulation detection and localization. International Journal of Computer Vision **130**(8), 1875–1895 (2022)
7. Liu, X., Liu, Y., Chen, J., Liu, X.: 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)
8. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
9. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11976–11986 (2022)
10. Qu, C., Liu, C., Liu, Y., Chen, X., Peng, D., Guo, F., Jin, L.: Towards robust tampered text detection in document image: New dataset and new solution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5937–5946 (2023)
11. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
12. Wang, W., Dai, J., Chen, Z., Huang, Z., Li, Z., Zhu, X., Hu, X., Lu, T., Lu, L., Li, H., et al.: Internimage: Exploring large-scale vision foundation models with deformable convolutions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14408–14419 (2023)
13. Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I.S., Xie, S.: Convnext v2: Co-designing and scaling convnets with masked autoencoders. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16133–16142 (2023)
14. Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J.: Unified perceptual parsing for scene understanding. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 418–434 (2018)
15. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in Neural Information Processing Systems **34**, 12077–12090 (2021)