ADCD-Net: Robust Document Image Forgery Localization via Adaptive DCT Feature and Hierarchical Content Disentanglement (Supplementary Material)

Kahim Wong[1], Jicheng Zhou[1], Haiwei Wu[2], Yain-Whar Si[1], Jiantao Zhou[1]✉
[1]State Key Laboratory of Internet of Things for Smart City, Department of Computer and Information Science, University of Macau
[2]School of Information and Software Engineering, University of Electronic Science and Technology of China
{yc37437, mc35093, fstasp, jtzhou}@um.edu.mo, haiweiwu@uestc.edu.cn

1. More Severe Degradations

Table. A presents the average F1 scores across three test sets against more severe degradations. ADCD-Net consistently outperforms existing methods, surpassing the second-best by an average margin of 20%.

Method	JPEG-60	Crop-0.7	Shift-50	Resize-0.7	AVG
TruFor	.599	.420	.590	.608	.554
DTD	.324	.082	.135	.605	.287
Ours	.655	.599	.666	.755	.669

Table A. Robustness test on more severe degradations.

2. Recompressing with Same QF

Theoretically, when recompress an images with the same QF multiply times, the only new damage is re-quantisation after tiny rounding drifts introduced by JPEG decoding which is negligible. We further conduct an experiment in Table B to verify the cases.

QF	95	90	85	80	75
Compress once	.762	.877	.783	.812	.738
Recompress two times	.762	.877	.782	.812	.738
Recompress three times	.762	.877	.782	.812	.738

Table B. F1 of recompress with same QF.

3. Analysis on PPE

We argue that PPE can be selectively employed when high confidence pristine area exists (*e.g.* the BG of deepfake portraits and forged documents, which is less informative and predominantly pristine). Using BG as pristine prior is justified by: 1) BG is less informative, making them less prone to forgery; 2) Minimal IoU between ground-truth forgery and BG regions, evidenced by a low average IoU (0.031) across four datasets in DocTamper (Table. C).

Dataset	`Train`	`Test`	`FCD`	`SCD`	AVG
IoU($\mathbf{X}_{\text{bg}}, \mathbf{Y}$)	.037	.030	.038	.021	.031

Table C. IoU between $\mathbf{X}_{\text{bg}}$ and $\mathbf{Y}$ cross datasets.

4. HCD Clarification

As Fig. 3 shows, shuffled forgery features are sent *only* to the reconstruction branch, while the localization branch uses the spatially ordered forgery features. $\mathbf{X}_{\text{dct}}$ is computed by applying an 8×8 block DCT to the Y-channel of $\mathbf{X}$ (standard JPEG), capturing local rather than global statistics. Thus, HCD alleviates text–background bias of the forgery feature yet still preserve local RGB/DCT cues for accurate forgery localization.

5. Unified vs. Multi-Scale Alignment Score

While separate $\hat{s}_{\text{aln}}$ heads could be trained for each scale, our goal is only to detect whether the DCT grid is aligned for an given image. A single classifier producing a unified $\hat{s}_{\text{aln}}$ reduces computation and limits scale-dependent variance. Fig. A shows the structure of these settings. Table D shows that the F1 performance of unified structure is slightly better than that of the multi-scale one in most types of distortion. It should also be noted that the multi-scale structure gives rather poor F1 score in J-85 (JPEG with QF 85), may be due to the high prediction variance introduced by multiple $\hat{s}_{\text{aln}}$ heads.

Setting	N-30	B-7	D-.7	J-85	C-.98	S-1	R-98
Unified $\hat{s}_{\text{aln}}$	**.697**	**.707**	<u>.755</u>	**.783**	<u>.689</u>	**.717**	**.692**
Multi-scale $\hat{s}_{\text{aln}}$	<u>.690</u>	<u>.700</u>	**.760**	<u>.650</u>	**.691**	<u>.667</u>	<u>.670</u>

Table D. F1 of unified and multi-scale $\hat{s}_{\text{aln}}$ across the distortions.

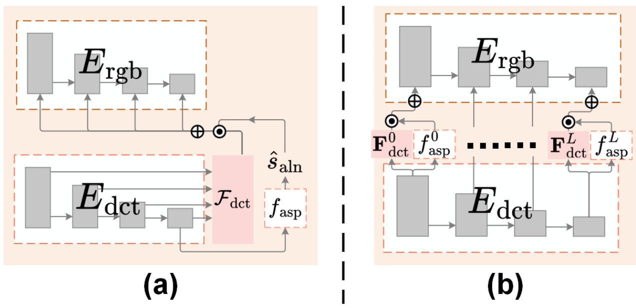

Figure A. Structure of (a) Unified $\hat{s}_{\text{aln}}$ and (b) Multi-scale $\hat{s}_{\text{aln}}$.

6. More Visualization

Fig. B illustrates the qualitative detection results obtained from representative test images. As shown, natural image forgery localization methods, such as TruFor and ConvNext, exhibit limitations in detecting numerous forged regions. Document-specific methods, exemplified by MA-Net, also remain insufficient in this regard. Although DTD demonstrates superior performance in some instances, it still fails to accurately identify forged regions accurately. In contrast, our proposed method consistently achieves high accuracy in detecting forged regions and exhibits stable performance across cross-domain testing scenarios.

Fig. C shows the reconstructed content and forgery features in the RGB and DCT domains. To enhance visualization, the overly smooth forgery RGB image is converted to grayscale and processed with histogram equalization. The 4-channel output of D_{rec} comprises the RGB image (channels 1-3) and the DCT coefficients (last channel). The HCD module accurately recovers content and preserves expected forgery patterns in both domains.

7. DCT Reconstruction Details

In practice, as shown in Fig. 6, $\hat{s}_{\text{aln}}$ is typically much greater than zero in most samples, reaching a non-zero minimum of 3×10^{-7}. We observe that the DCT reconstruction loss is reduced by 29.4% compared to the case when $\hat{s}_{\text{aln}} = 1$.

8. Implementation Details

ADCD-Net uses the Restormer encoder [4] as E_{rgb} and the Restormer decoder as D_{rec} and D_{frg}, initialized with the DocRes checkpoint [5]. For E_{dct}, we adopt the Frequency Perception Head from DTD [3], and we use CRAFT [1] as the OCR model. We follow the CLTD training strategy from DTD [3] and optimize with AdamW [2] at a learning rate of 3×10^{-4} over 100k steps (batch size 16, decayed to 1×10^{-5} with a cosine annealing schedule). The loss weights are set as $\lambda_{\text{aln}} = \lambda_{\text{rec}} = \lambda_{\text{frg}} = 1$, $\lambda_{\text{ce}} = 3$, and $\lambda_{\text{con}} = [0.001, 0.005, 0.02, 0.1]$ across different scale, such that all losses are in similar scale. Experiments are run on 4 NVIDIA GeForce RTX 4090 24G GPUs, the entire training takes about 33 hours. The predictions binarized at 0.5. Our model is trained on the DocTamper [3] training set and evaluated on three cross-domain test sets (`Test`, `FCD`, and `SCD`). For more details please refer to our code at https://github.com/KAHIMWONG/ACDC-Net.

9. Explanation of Datasets Used

We use the DocTamper `Training set` for model training and evaluate on the `Testing set`, `DocTamper-FCD`, and `DocTamper-SCD`. "DTD" names the forgery localization model, while "DocTamper" names the forgery document dataset. We summarizes the datasets used in different figures and tables in Table E.

Datasets	Fig./Table	Remark
`Training set`	N/A	Train our model
	Fig. 7	1,000 random images
`Testing set` + `DocTamper-FCD` + `DocTamper-SCD`	Fig. 4, Table 1, Fig. 5, Table 3, Table 4, Table A, Table B, Table D	Average F1 across three test sets
	Fig. 6, Table 5	1,000 random images per set (3,000 total)
`Tianchi 2023 DDT train set`	Table 2	1,000 random pristine images

Table E. Datasets used in different figures and tables.

References

References

[1] Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee. Character region awareness for text detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9365–9374, 2019. 2

[2] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 2

[3] Chenfan Qu, Chongyu Liu, Yuliang Liu, Xinhong Chen, Dezhi Peng, Fengjun Guo, and Lianwen Jin. Towards robust tampered text detection in document image: New dataset and new solution. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5937–5946, 2023. 2

[4] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5728–5739, 2022. 2

[5] Jiaxin Zhang, Dezhi Peng, Chongyu Liu, Peirong Zhang, and Lianwen Jin. Docres: A generalist model toward unifying document image restoration tasks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 15654–15664, 2024. 2

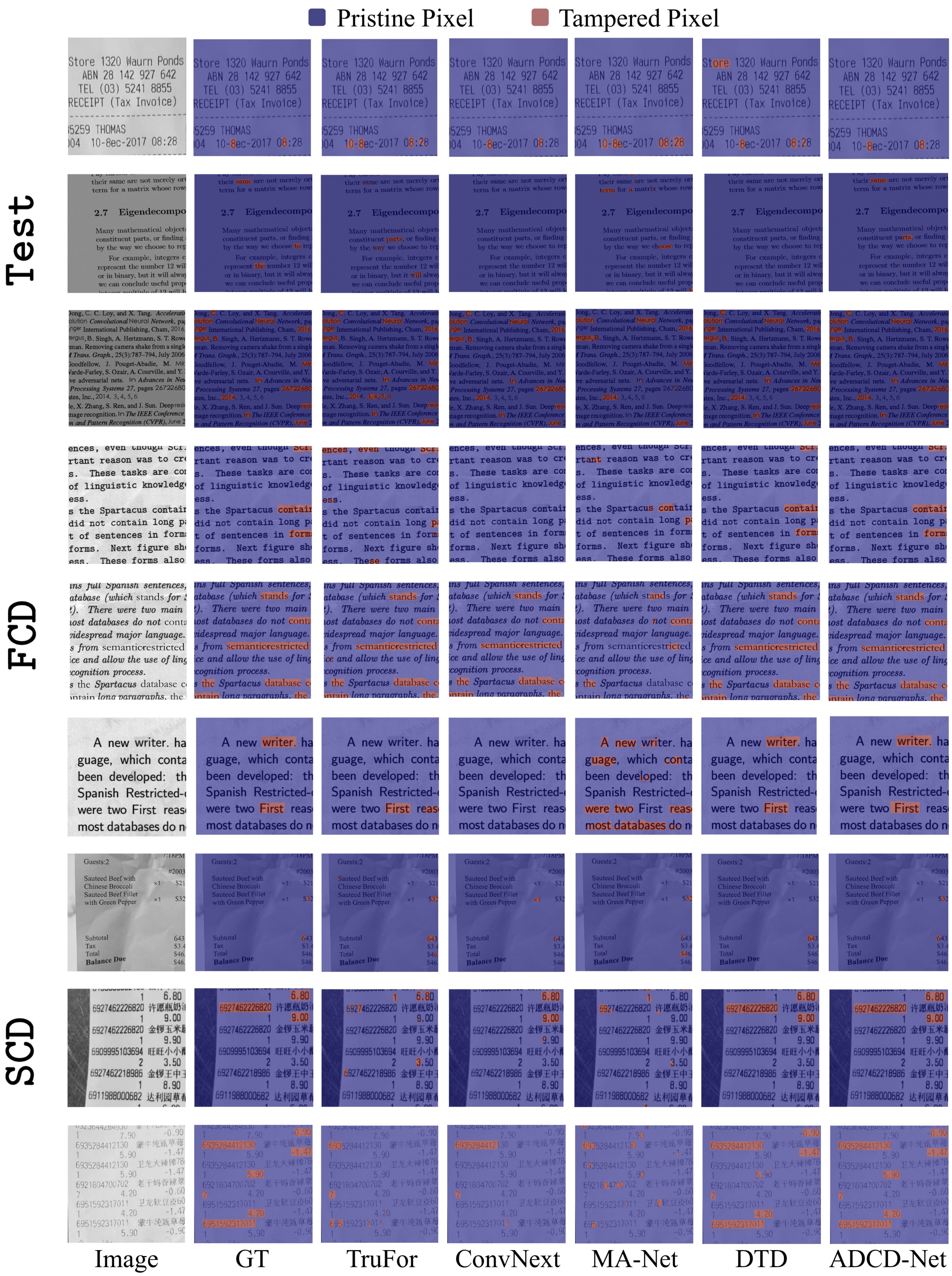

Figure B. Qualitative results on the three test sets comparing ADCD-Net with SOTA methods.

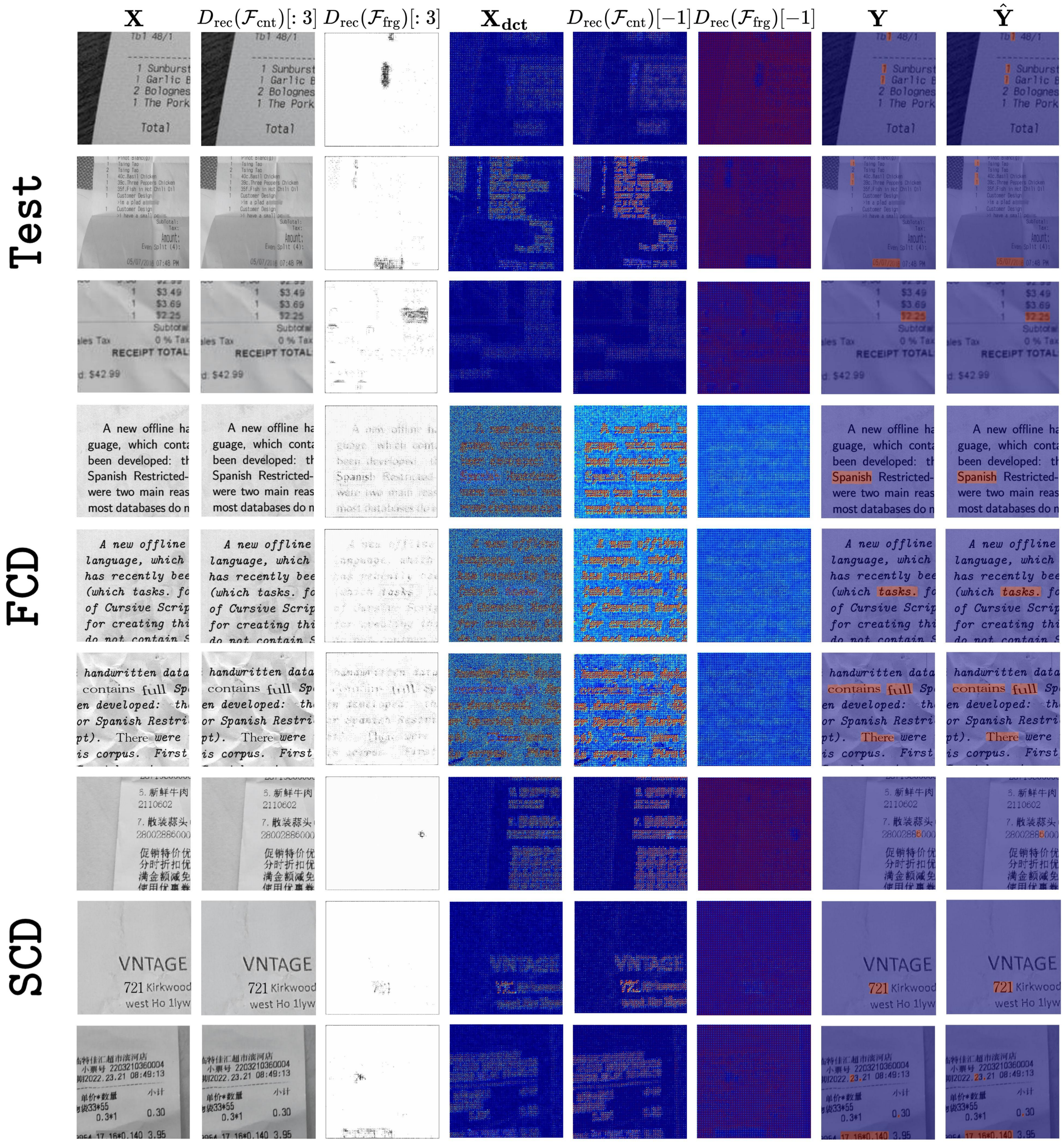

Figure C. Visualization of HCD feature reconstructions: content features from the RGB domain $D_{\text{rec}}(\mathcal{F}_{\text{cnt}})[:3]$ and DCT domain $D_{\text{rec}}(\mathcal{F}_{\text{cnt}})[-1]$, and forgery features from the RGB domain $D_{\text{rec}}(\mathcal{F}_{\text{frg}})[:3]$ and DCT domain $D_{\text{rec}}(\mathcal{F}_{\text{frg}})[-1]$. Also shown are the original RGB image $\mathbf{X}$, DCT coefficients $\mathbf{X}_{\text{dct}}$, model prediction $\hat{\mathbf{Y}}$, and ground truth $\mathbf{Y}$.