Supplementary Material of "DRAW: Defending Camera-shooted RAW against Image Manipulation"

Xiaoxiao Hu[1,2*], Qichao Ying[1,2*], Zhenxing Qian[1,2†], Sheng Li[1,2], Xinpeng Zhang[1,2]
[1]School of Computer Science, Fudan University
[2]Key Laboratory of Culture & Tourism Intelligent Computing, Fudan University

In this supplementary material, we provide the details of the hybrid attack layer, the baseline designs and the experimental settings. Also, more experimental results on the imperceptibility of RAW protection and the performance of robust image manipulation detection are presented.

1. Details of the Hybrid Attack Layer.

Manipulation Mask Generation. Real-world image tampering may be oriented on regions of interest within a targeted image. However, code-driven realistic image manipulation can be expensive and time-consuming. We study the natural distribution of tampered areas by observing the binary masks in CASIA dataset [3]. The location of forgery within an image roughly follows a uniform distribution except for corners, and for most manipulated images, the total area of forged contents is within the range of 5%-30%. For simplification, we assume that the location of forgery within an image during training roughly follows a uniform distribution and the accumulated manipulated squared area is within the range of $[0, 0.3]$.

We apply free-form mask generation [20] to arbitrarily select areas within $\hat{\mathbf{I}}$ according to a binary mask $\mathbf{M}$.

$$\hat{\mathbf{I}}_t = \hat{\mathbf{I}} \cdot (1 - \mathbf{M}) + \mathbf{R} \cdot \mathbf{M}, \tag{1}$$

where $\mathbf{R}$ is the source of manipulation.

Image Manipulation Simulation. For image manipulation, we simulate the most common types of tampering, which include *copy-moving*, *splicing* and *inpainting*. The simulation of different kinds of attacks can be reflected by the composition of $\mathbf{R}$ in Eq. (1). For *copy-moving*, we let $\mathbf{R}$ in Eq. (1) as a spatially-shifted version of $\hat{\mathbf{I}}$. For *splicing*, we use another random RGB image as $\mathbf{R}$. However, we find that this setting of attack will encourage the network to widen the distribution gap between $\hat{\mathbf{I}}$ and natural RGB images to better distinguish each other, thus greatly decreasing the overall image quality. To address this, we also apply an enhanced *splicing* attack named *coincident-splicing* that "coincidentally" use another protected RGB $\hat{\mathbf{I}}'$ as $\mathbf{R}$. For *inpainting*, we use the open-source model from LAMA [16] and ZITS [4] to generate the inpainted result as $\mathbf{R}$. We iteratively and evenly perform the above *three* types of attacks for balanced training.

Image Distortion Simulation. Similar to HiDDeN [22], we simulate typical image lossy post-processing operations to enhance the robustness of the proposed method. The involved attacks include the following: (1) *rescaling*, which resizes the image by an arbitrarily resizing rate $r \in [50\%, 150\%]$, (2) *median blurring*, which blurs the image using median filter whose kernel size k is arbitrarily selected from $[3, 5]$, (3) *Additive White Gaussian Noise* (AWGN), which adds Gaussian noise evenly on the image, where the standard value s ranges from zero to one, (4) *Gaussian blurring*, which is similar to the median blurring but the kernel is different, (5) *JPEG compression*, which compresses the image using the popular Diff-JPEG [14] with tunable JPEG quality factors.

Color Adjustment Simulation. Most users prefer manually adjusting the brightness or contrast after RAW files are automatically rendered into RGBs. Therefore, we also simulate typical color adjustment operations to mitigate their impacts on the performance of our method. The involved attacks include the following: (1) *Hue adjustment*: the image hue is adjusted by converting the image to HSV and cyclically shifting the intensities in the hue channel. The image is then converted back to the original image mode. The hue factor is set within the range of $[-0.05, 0.05]$. (2) *contrast enhancement*: we adjust the contrast of an image, where the contrast factor is set within the range of $[0.7, 1.5]$. (3) *saturation adjustment*: we adjust the color saturation of the image, where the factor is set within the range of $[0.7, 1.5]$. (4) *brightness adjustment*: we adjust the brightness of the image, where the factor is set within the range of $[0.7, 1.5]$. The differentiable data augmentation functions applied during training are implemented by the APIs from the "torchvision" package.

Real-world Attack Involvement. The real-world attacks

*Xiaoxiao Hu and Qichao Ying contribute equally to this work.
†Corresponding author: Zhenxing Qian (zxqian@fudan.edu.cn)

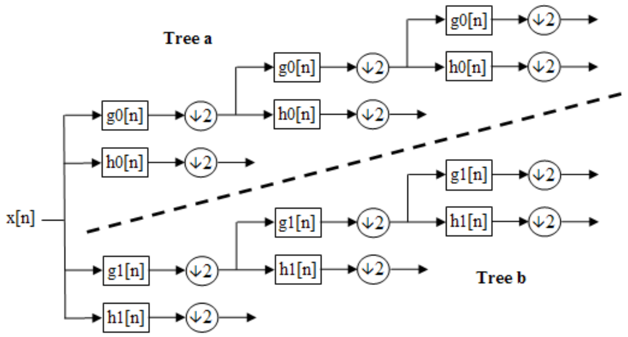

Figure 1. **Illustration of DT-CWT,** which is a two-dimensional transform that decomposes an image into six frequency subbands at each level of the transform.

are implemented by the APIs from the "cv2" package, e.g., *cv2.GaussianBlur* for Gaussian blurring and *cv2.imencode* for JPEG compression. These functions are performed on PIL images in "ndarray" format, which requires that we transform the 32-bit float-typed tensors into 8-bit integer-based arrays. Therefore, quantization attack is also automatically considered by the introduction of real-world attacks. In each iteration of the training stage, we perform the corresponding real-world attacks using the same setting from the simulated methods.

2. Frequency Learning in Deep Networks

Existing Methodologies. Frequency-learning is an efficient way to reduce computation resource costs. For example, Xu et al. [19] proposes a learning-based frequency selection method to identify trivial frequency components in the input images, which can be removed without performance loss. According to [8], self-attention layers can be replaced with simple Fourier transformations to speed up Transformer encoder architectures under limited accurary sacriface. FEDformer [21] exploits the sparse representation in Fourier transform to capture the global view of time series. In addition, frequency-domain information has shown great potential in revealing subtle differences between real and fake images, such as in face forgery detection tasks, where it can help detect generated faces [5, 1, 17] or synthesized images [11, 9, 10] based on face-swapping techniques.

However, the above-listed work only replaces interpolation with DWT or DCT, which still requires heavy computation. In order to design a new lightweight network with frequency learning, we must effectively combine the advantages of wavelet transform and CNN architecture.

DT-CWT Transformation. The Dual-Tree Complex Wavelet Transform (DT-CWT) is a type of wavelet transform used in signal and image processing. It was introduced by Kingsbury [12] and is an extension of the discrete wavelet transform (DWT) that uses complex wavelets. The DT-CWT is a two-dimensional transform that decomposes an image into six frequency subbands at each level of the transform. These subbands are formed by filtering the image with two sets of filters, one for the real part of the wavelet and the other for the imaginary part. The filters are designed to have good directional selectivity and to be approximately shift-invariant.

The DT-CWT has been successfully applied to various computer vision tasks, including image denoising [6], image super-resolution [7], and object detection [15, 13]. For example, in object detection, the DT-CWT can be used to extract features that are both scale and orientation invariant, which can improve the accuracy of the detector. In image super-resolution, the DT-CWT can be used to extract high-frequency information that is lost during image downsampling, which can then be used to reconstruct a higher-resolution image. With the development of modern CNN networks, researchers prefer learning end-to-end feature extractors in favor of pre-designed filters, which possibly results in a downgraded role of wavelet transform played in computer vision tasks. However, compared to cascaded learnable convolutional layers, DT-CWT transform still contain several advantages as follows.

Bringing DT-CWT into CNNs. Introducing DT-CWT into CNNs can have several advantages for our task and beyond. First, the DT-CWT is robust to noise in image data, as it can extract features at multiple scales and orientations. This can help improve the performance of CNNs on modifying the higher-frequency details. Second, the DT-CWT can extract rich features that are both scale and orientation invariant, which can improve the discriminative power of CNNs. This can be particularly useful in content-aware protective signal embedding. Thirdly, the DT-CWT can reduce the complexity of CNNs by reducing the number of filters required in the initial layers. It also provides both magnitude and phase information, which can be used to visualize and interpret the learned features.

In MPF-Net, we combine the benefit of DT-CWT with Fourier frequency learning, where FFT can mitigate the issue of focusing too much on local patterns. Besides, global information aggregation and lower computational complexity is achieved by the proposed HFC and PFF mechanisms.

3. More Experimental Results

Fig. 2 shows more experimental results on the imperceptibility of RAW protection. From the results, we see that the injected signal is weak and the generated protected RGB images are not affected in their overall visual quality.

To justify the generalizability to lossy transmission, we randomly handcraft 150 manipulated images, upload them onto several famous OSNs and download them for detection. Also, we test the performance against dual JPEG and

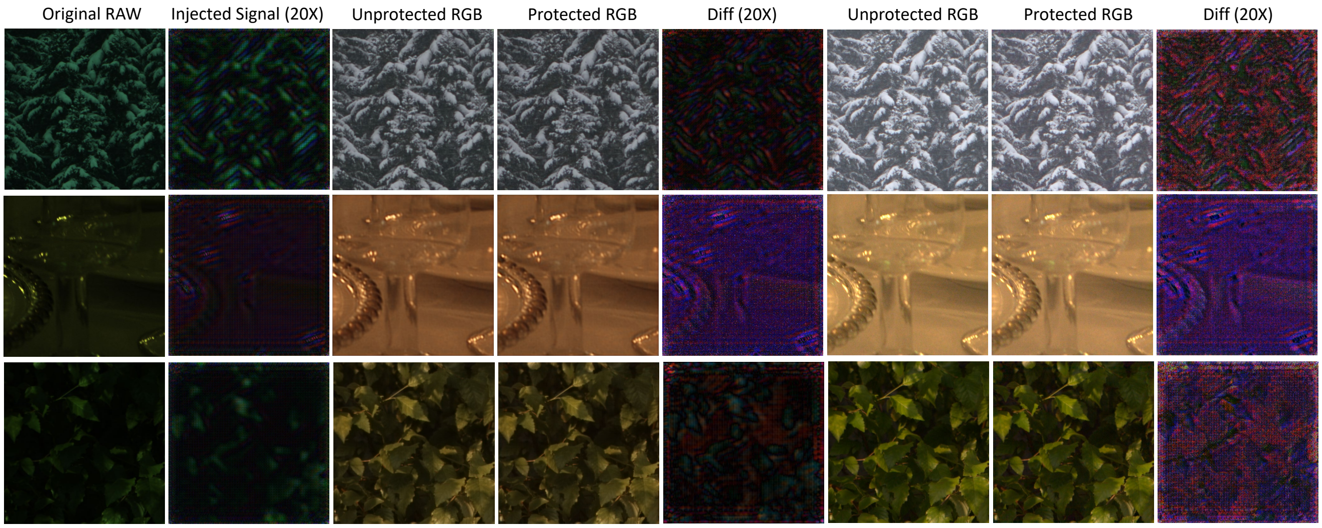

Figure 2. **Examples of protected RAWs and the corresponding protected RGBs.** In each test, we apply two ISPs for rendering (upper: InvISP / TradISP, middle: LibRAW / TradISP, lower: CycleISP / LibRAW). The RAW images are visualized through bilinear demosaicing.

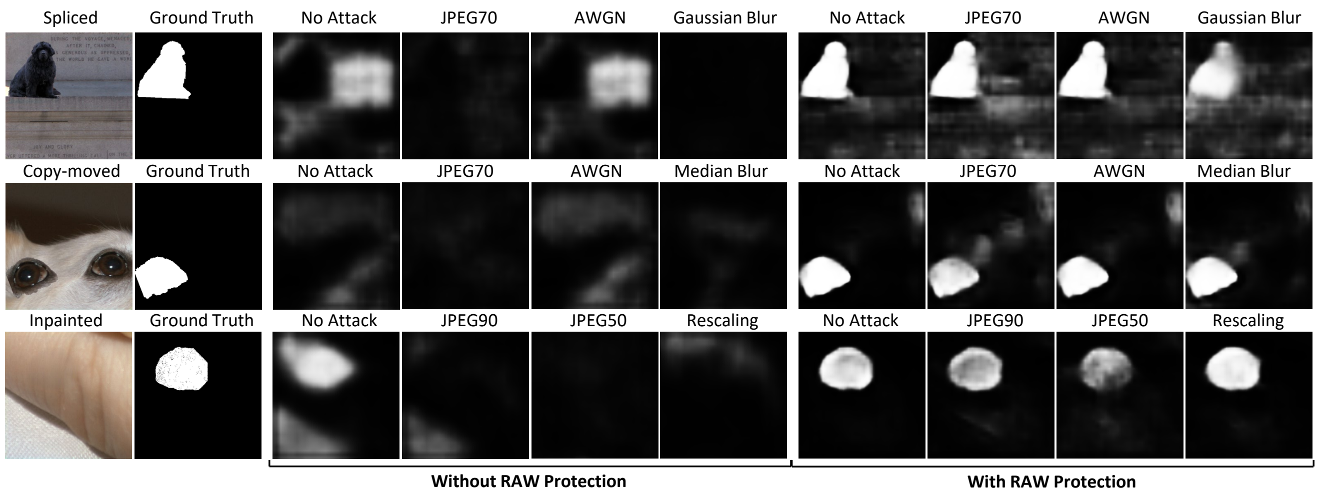

Figure 3. **Example of performance gain on MVSS with DRAW.** The protection helps the detector locate the forged area despite the presence of lossy image operations.

Table 1. **Generalizability to lossy transmission and untrained perturbations.** Dataset: RAISE.

Forgery	S&P		Dual JPEG		Facebook		Weibo		WeChat	
	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU
splicing	.839	.855	.657	.683	.917	.920	.902	.897	.763	.728
copymove	.854	.850	.692	.729	.905	.910	.859	.870	.637	.688
inpainting	.687	.711	.377	.423	.665	.598	.623	.577	.410	.355

salt & pepper attack ($p = 5\%$) which are untrained types for DRAW. As shown in Table 1, DRAW can effectively resist lossy OSN transmission, and its protection remains valuable against unknown lossy operations.

Fig. 3 and Fig. 4 respectively show some examples of performance gain on MVSS [2] and RIML [18] with DRAW. The protection helps the two detectors locate the forged area despite the presence of lossy image operations.

4. Details of the Baseline Methods.

Fig. 5 illustrates the pipeline overview of the two baseline methods, namely, image forgery detection with pure robust training and image forgery detection using RGB protection. Detailed settings are specified as follows.

RAW Protection vs Pure Robust Training. We validate the impact of RAW protection on the performance of DRAW by first removing the RAW protection stage. The corresponding fidelity terms are also removed. In this case, no camera imaging pipeline is considered and the training technique of hybrid attacking layer involvement is solely responsible for improving the robustness of image manipulation localization, which is close to RIML. According to the experiments, the baseline can indeed noticeably boost

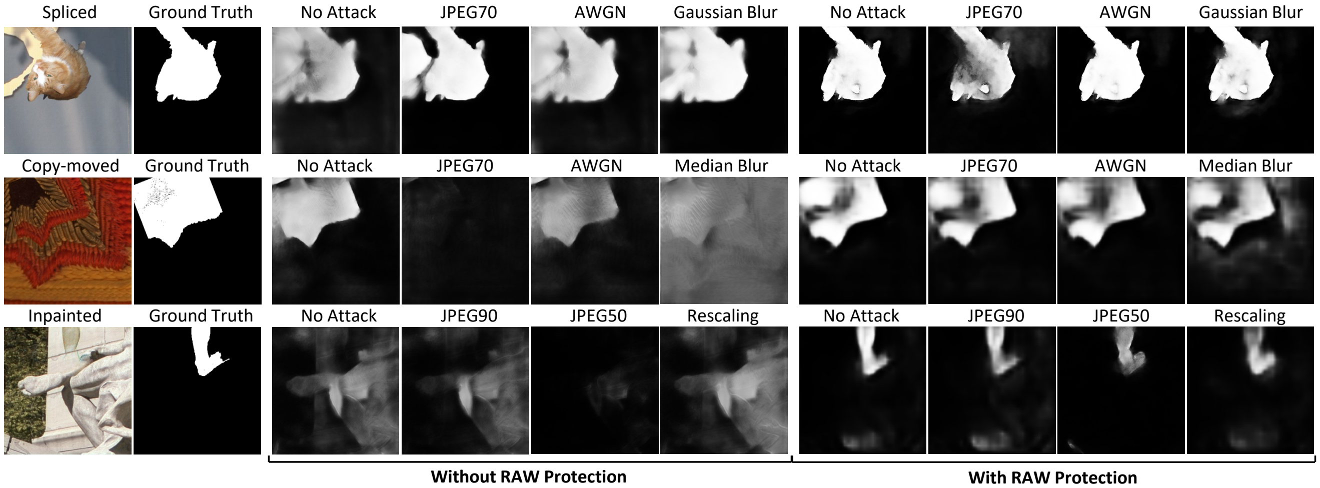

Figure 4. **Example of performance gain on RIML with DRAW.** The protection helps the detector locate the forged area despite the presence of lossy image operations.

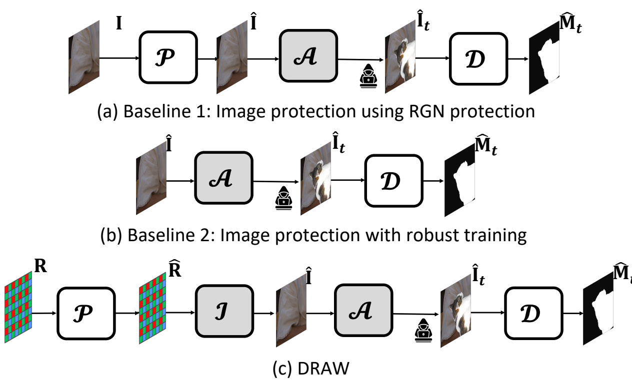

Figure 5. **Pipeline comparison** between DRAW and baselines.

the performance of manipulation detection under lossy image operations, but the overall accuracy is worse than that of RAW protection. The reason is that finding a universally-existing trace to unveil manipulation is difficult in the real world, unlike injecting an outer-sourced signal for later retrieval.

RAW protection vs RGB protection. We compare RAW protection with RGB protection, in which we modify the original image for anti-manipulation protection. The ISP process is also ruled out in the pipeline. The RAW protection term is therefore removed and the hyper-parameters are changed as $\beta = 1, \gamma = 0.01, \epsilon = 0.005$. Though the two schemes ideally can come up with the same solution where after image rendering, the protective signal embedded within RAW could be the same or close compared with that embedded directly within RGB, the experimental results show that successful RGB protection is more difficult compared to RAW protection. The reason is that RAW protection can adaptively introduce protection with the help of content-related procedures, e.g., demosaicing and noise reduction, within the subsequent ISP algorithms that suppress unwanted artifacts and biases. Besides, RAW data modification enjoys a much larger search space that allows transformations from the original image into another image with high density upon sampling.

5. Other Implementation Details

We train all network-based ISP pipelines using RGB images rendered by the libraw library as supervision and these pre-trained ISPs will be frozen when training the RAW protection network. We find that for different RAW datasets, the performances of cross-dataset RGB image rendering of ISP networks are not satisfactory. Therefore, for each RAW dataset, we separately train their exclusive ISP networks. In contrast, our protection network is transferable, and we train the network based on a single benchmark dataset, e.g., RAISE, and conduct experiments on other datasets on this model without further fine-tuning.

References

[1] Keshigeyan Chandrasegaran, Ngoc-Trung Tran, and Ngai-Man Cheung. A closer look at fourier spectrum discrepancies for cnn-generated images detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7200–7209, 2021. 2

[2] Xinru Chen, Chengbo Dong, Jiaqi Ji, Juan Cao, and Xirong Li. Image manipulation detection by multi-view multi-scale supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14185–14193, 2021. 3

[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 Information Processing*, pages 422–426. IEEE, 2013. 1

[4] Qiaole Dong, Chenjie Cao, and Yanwei Fu. Incremental transformer structure enhanced image inpainting with masking positional encoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11358–11368, 2022. 1

[5] Tarik Dzanic, Karan Shah, and Freddie Witherden. Fourier spectrum discrepancies in deep network generated images. *Advances in neural information processing systems*, 33:3022–3032, 2020. 2

[6] Paul Hill, Alin Achim, and David Bull. The undecimated dual tree complex wavelet transform and its application to bivariate image denoising using a cauchy model. In *2012 19th IEEE international conference on image processing*, pages 1205–1208. IEEE, 2012. 2

[7] Sara Izadpanahi and Hasan Demirel. Motion based video super resolution using edge directed interpolation and complex wavelet transform. *Signal Processing*, 93(7):2076–2086, 2013. 2

[8] James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontanon. Fnet: Mixing tokens with fourier transforms. *arXiv preprint arXiv:2105.03824*, 2021. 2

[9] Jiaming Li, Hongtao Xie, Jiahong Li, Zhongyuan Wang, and Yongdong Zhang. Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6458–6467, 2021. 2

[10] Honggu Liu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu. Spatial-phase shallow learning: rethinking face forgery detection in frequency domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 772–781, 2021. 2

[11] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII*, pages 86–103. Springer, 2020. 2

[12] Ivan W Selesnick, Richard G Baraniuk, and Nick C Kingsbury. The dual-tree complex wavelet transform. *IEEE signal processing magazine*, 22(6):123–151, 2005. 2

[13] Sandeep Singh Sengar and Susanta Mukhopadhyay. Moving object detection using statistical background subtraction in wavelet compressed domain. *Multimedia Tools and Applications*, 79(9-10):5919–5940, 2020. 2

[14] Richard Shin and Dawn Song. Jpeg-resistant adversarial images. In *NIPS 2017 Workshop on Machine Learning and Computer Security*, volume 1, 2017. 1

[15] Yue-Hui Sun and Ming-Hui Du. Face detection using dt-cwt on spectral histogram. In *2006 International Conference on Machine Learning and Cybernetics*, pages 3637–3642. IEEE, 2006. 2

[16] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2149–2159, 2022. 1

[17] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8695–8704, 2020. 2

[18] Haiwei Wu, Jiantao Zhou, Jinyu Tian, and Jun Liu. Robust image forgery detection over online social network shared images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13440–13449, 2022. 3

[19] Kai Xu, Minghai Qin, Fei Sun, Yuhao Wang, Yen-Kuang Chen, and Fengbo Ren. Learning in the frequency domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1740–1749, 2020. 2

[20] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4471–4480, 2019. 1

[21] Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International Conference on Machine Learning*, pages 27268–27286. PMLR, 2022. 2

[22] Xinshan Zhu, Yongjun Qian, Xianfeng Zhao, Biao Sun, and Ya Sun. A deep learning approach to patch-based image inpainting forensics. *Signal Processing: Image Communication*, 67:90–99, 2018. 1