JPEG Compression-aware Image Forgery Localization

Menglu Wang
University of Science and Technology of China
Hefei, Anhui, China
vault@mail.ustc.edu.cn

Xueyang Fu*
University of Science and Technology of China
Hefei, Anhui, China
xyfu@ustc.edu.cn

Jiawei Liu
University of Science and Technology of China
Hefei, Anhui, China
jwliu6@ustc.edu.cn

Zheng-Jun Zha
University of Science and Technology of China
Hefei, Anhui, China
zhazj@ustc.edu.cn

ABSTRACT

Image forgery localization, which aims to find suspicious regions tampered with splicing, copy-move or removal manipulations, has attracted increasing attention. Existing image forgery localization methods have made great progress on public datasets. However, these methods suffer a severe performance drop when the forged images are JPEG compressed, which is widely applied in social media transmission. To tackle this issue, we propose a wavelet-based compression representation learning scheme for the specific JPEG-resistant image forgery localization. Specifically, to improve the performance against JPEG compression, we first learn the abstract representations to distinguish various compression levels through wavelet integrated contrastive learning strategy. Then, based on the learned representations, we introduce a JPEG compression-aware image forgery localization network to flexibly handle forged images compressed with various JPEG quality factors. Moreover, a boundary correction branch is designed to alleviate the edge artifacts caused by JPEG compression. Extensive experiments demonstrate the superiority of our method to existing state-of-the-art approaches, not only on standard datasets, but also on the JPEG forged images with multiple compression quality factors.

CCS CONCEPTS

• **Applied computing** → **Computer forensics**; • **Computing methodologies** → **Computer vision**.

KEYWORDS

image forgery localization, JPEG compression, contrastive learning, wavelet transform

*Xueyang Fu is the corresponding author (xyfu@ustc.edu.cn). This work was supported by the National Key R&D Program of China under Grant 2020AAA0105702, the National Natural Science Foundation of China (NSFC) under Grants U19B2038 and 61901433, the University Synergy Innovation Program of Anhui Province under Grants GXXT-2019-025, the Fundamental Research Funds for the Central Universities under Grant WK2100000024, and the USTC Research Funds of the Double First-Class Initiative under Grant YD2100002003.

MM '22, October 10–14, 2022, Lisboa, Portugal

ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3547749

ACM Reference Format:
Menglu Wang, Xueyang Fu, Jiawei Liu, and Zheng-Jun Zha. 2022. JPEG Compression-aware Image Forgery Localization. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, Lisboa, Portugal, 9 pages. https://doi.org/10.1145/3503161.3547749

1 INTRODUCTION

With the development of image editing techniques and the popularization of various editing software, image forgery becomes easier and has been widely used on social media, leading to a decline in the credibility of image information and negative effects on the daily life. The most common manipulations are splicing (copying regions from an authentic image and pasting them to other images), copy-move (copying and pasting regions within the same images) and removal (eliminating regions from an authentic image). Therefore, image forgery localization (IFL) attracts increasing attention, which aims to locate suspicious regions on the tampered images (splicing, copy-move, removal). While the tampered images are often compressed during the transmission and re-posting of real social media, which brings multiple degradation artifacts such as blocking artifacts, ringing effects, blurring and color distortion. These complex compression artifacts overlap with the image tampering traces, making it more difficult to locate the subtle tampering traces. Although IFL approaches have made great progress on high-quality datasets, these methods suffer a severe performance drop when the forged images are JPEG compressed. Therefore, the trained models are difficult to generalize to JPEG compressed images in real social media scenarios. In this paper, we consider the JPEG compressed image forgery localization (JPEG-IFL) on the common manipulations (splicing, copy-move, and removal), which aims to locate the tampered regions on JPEG compressed forged images.

Traditional IFL methods rely on the hand-crafted features to figure out the different statistical distribution between tampered regions and the authentic regions [9, 22, 44]. In recent years, due to the powerful representation ability, deep learning based methods have been explored to directly learn a nonlinear mapping from forged images to its corresponding masks. For example, ManTra-Net [38] detects forged pixels by identifying local anomalous features in an end-to-end way. MVSS-Net [4] learns a multi-view feature with multi-scale supervised networks to jointly exploit the noise view and the boundary artifacts. However, these learning-based methods suffer a severe performance drop when the forged images are JPEG compressed, which is widely applied in social media transmission.

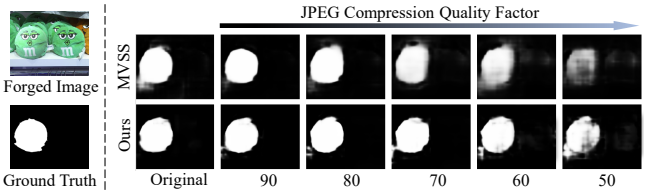

Figure 1: The JPEG image forgery localization results under different compression quality factors. As the quality factor (QF) decreases, the masks predicted by the latest approach MVSS [4] are getting blurry. While our network can predict accurate masks under multiple compression quality factors.

As shown in Figure 1, when the JPEG compression quality factor (QF) decreases, the masks predicted by MVSS [4] are getting blurry. Especially in QF=60 and QF=50, a lot of artifacts appear and the method is almost incapable of predicting the accurate tampered area masks. Recently, Rao *te al.* [27] adopts the domain adaptation strategy to alleviate JPEG compression problem without fully exploring the JPEG compression information, resulting in limited performance. Therefore, it is emergency to explore an efficient method to deal with JPEG compressed image forgery localization.

Inspired by the success of degradation representation learning in super resolution [33], we propose a new wavelet-based compression representation learning scheme for the specific JPEG image forgery localization (JPEG-IFL). The motivation of our method is that the compression representations fully characterize the information of different compression degrees, which can be utilized as an important cue for JPEG-IFL. Since JPEG compression is highly correlated with the frequency domain, a wavelet-based representation learner is firstly introduced to obtain compression representations. Then, a JPEG compression-aware image forgery localization network is designed to leverage the learned representations. Specifically, we decompose the forged image into different frequencies with wavelet transformation to extract distinguishable compression representations through contrastive learning strategy. Based on the learned representations, we introduce a JPEG compression-aware IFL network to flexibly handle forged images with various JPEG quality factors. Moreover, a boundary correction branch is further designed to alleviate the edge artifacts caused by JPEG compression. To our knowledge, this is the first attempt to explicitly utilize degradation information (low-level representations) on image forgery localization (high-level vision task). Extensive experiments shows the effectiveness of our method on IFL as well as JPEG-IFL.

Contributions of this work can be summarized as follows:

- We propose a general framework for JPEG image forgery localization (JPEG-IFL) by taking compression representations into consideration. The compression quality factors, which are ignored in existing methods, are fully utilized to help design algorithms for the specific JPEG-IFL problem.
- We design a wavelet-based encoder to extract distinguishable compression representations through contrastive learning strategy. Based on the learned representations, we introduce a JPEG compression-aware IFL network to flexibly handle forged images compressed with various JPEG quality factors.
- We incorporate the boundary correction module into our JPEG-IFL network, which can effectively alleviate the edge artifacts caused by JPEG compression.
- Extensive experimental results prove superior performance of our approach compared with the state-of-the-art methods, not only on standard IFL datasets, but also on the JPEG forged images with multiple compression quality factors.

2 RELATED WORK

Image forgery localization (IFL) has been an increasingly important research topic in computer vision community. Tremendous methods have been proposed in this area. These methods can be divided into two categories, i.e., traditional methods [5, 9, 22, 24, 32, 44, 45] and deep-learning based methods [4, 6, 14, 27–29, 37–40, 47].

2.1 Hand-crafted Feature Based IFL Approaches

Traditional methods rely on the hand-crafted feature to figure out the different statistical distribution between tampered regions and the authentic regions [9, 22, 44]. Relying on the statistics of pristine natural images, some works [24, 44] detect tempered region based on the deviation. For example, sparse descriptors [24] obtain a high accuracy for copy-move [5]. More complex dense descriptors [30] could achieve better performance. A feature-based procedure [44] is proposed to tell apart regions subject to median filtering from region treated by other forms of processing. Steganalysis rich model [9] (SRM) works in a similar manner by digging out the inconsistent local noise variances between different regions within an authentic image. While blind local noise estimation methods [23, 25] expose region splicing by revealing inconsistencies in local noise levels. On the other hand, tampering possibility maps [22], which are obtained by adjusting statistical feature-based detector and copy-move forgery detector, are integrated to improve the performance of forgery localization. These hand-crafted methods achieve effective forgery localization before the era of deep learning and provide various enlightening analyses of image forgery localization.

2.2 Deep Learning Based IFL Approaches

The introduction of neural-network-based solutions has brought significant improvement for image forgery localization. Tremendous architecture designs provide various effective solutions for image forgery localization. For example, a two-stream network [47] uses both an RGB stream and a noise stream to learn rich features for image manipulation detection, where Faster R-CNN [29] is integrated. The multi-task fully convolutional network (MFCN) [31] is proposed to automatically learn the relevant features for image forgery localization, which does not require explicit feature extraction. ManTra-Net [38] is an end-to-end network that performs both detection and localization without extra preprocessing and post-processing. MVSS-Net [4] learns a multi-view feature to jointly exploit the noise view and the boundary artifacts. The architecture of a two-branch convolutional neural network [28] is presented as an expressive local descriptor to automatically learn hierarchical representations. Recently, inspired by the success of transformer, TransForensics [14] introduces transformers into this area which captures discriminative representations and obtain high-quality

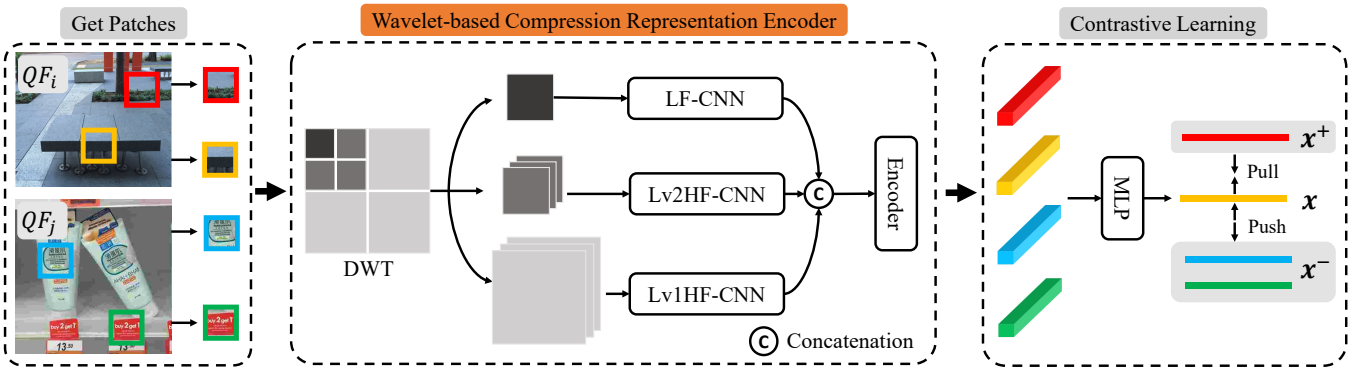

Figure 2: The architecture of the Compression Representation Learning Network (CRL-Net). It aims to obtain the compression-related representations. The CRL-Net consists of three parts: getting patches, wavelet-based compression representation encoder and contrastive learning strategy. The input patches include query patches (annotated with a yellow box), positive patch (annotated with a red box) and negative patches (annotated with blue and green boxes), which take compression degree as division basis. Then, we disentangle patches into multiple level frequency bands to extract compression related features. Finally, through contrastive learning strategy, the extracted features with the same compression degree are encouraged to be similar, while the extracted features with different compression degrees are encouraged to be dissimilar.

mask predictions. Aforementioned IFL methods suffer severe performance drops when the forged images are degraded, which will bring more complex artifacts [17–19, 26, 41]. Particularly, JPEG compression is widely used in social media. To solve this problem, the method [8] analyzes the impact of JPEG compression on forgery task. The domain adaptation network [27] is proposed to alleviate the domain shift between uncompressed images and JPEG-agent images, which consists of a Siamese backbone and a compression approximation network. The method [2] aims to extract compression-insensitive features from both uncompressed and compressed forgeries using an adversarial learning strategy. However, the JPEG compression information is not fully explored and this method suffers from boundary artifacts caused by compression.

3 METHODOLOGY

The proposed framework consists of the compression representation learning network and JPEG compression-aware image forgery localization network, which are illustrated in Figure 2 and Figure 3 respectively. In this section, we firstly present the problem definition and the overall architecture of the proposed approach. Then we introduce each component in the following subsections.

3.1 Problem Definition and Overview

As mentioned above, the regular image forgery localization methods will suffer a severe performance drop when the forged images are JPEG compressed, which is widely applied in social media transmission. For practical application requirements, we focus on JPEG-resistant image forgery localization problem. Specifically, we denote the forged images in the standard datasets as $D = \{d_i\}_{i=1}^{N}$ and their corresponding forgery localization masks as $M = \{m_i\}_{i=1}^{N}$, where N is the total number of forged images. The forgery process can be represented as $d_i = F(a_i)$, where a_i is the authentic image and F is a tampering operation (e.g., splicing, copy-move or removal). In our setting, we reformulate the forgery process with the consideration of compression as $y_i = JPEG(d_i) = JPEG(F(a_i))$. JPEG compression operation brings complex degradation artifacts such as blocking artifacts, ringing effects, blurring [11, 12, 34], which are overlapped with the image tampering traces. That makes it more difficult to locate the forged area accurately. Therefore, our goal is to perceive and alleviate the compression artifacts impact so as to achieve accurate localization. Given a JPEG compressed forged image y_i, we aim to design a JPEG-resistant forgery localization model G to predict the forged region mask $\hat{m}_i$, which is denoted as $\hat{m}_i = G(y_i)$.

To achieve this goal, we propose a novel framework which consists of two components: compression representation learning network (CRL-Net) and JPEG compression-aware image forgery localization network (CA-IFL), whose architectures are depicted in Figure 2 and Figure 3, respectively. The CRL-Net aims to extract representations that are highly correlated with JPEG compression, which can be utilized as an important cue for the following CA-IFL. Based on the compression representations, the CA-IFL aims to perceive and alleviate the compression artifacts impact so as to locate the forged area as accurately as possible. Our model is trained in two steps. **First**, to obtain adequate compression representations, we train a wavelet-based CRL-Net through contrastive learning strategy, as illustrated in Figure 2. The CRL-Net consists of three parts: getting patches, wavelet-based compression representation encoder and contrastive learning strategy. The input patches include query patches, positive patches and negative patches, which take compression degree as division basis. Then, since the JPEG compressed images are highly sensitive in frequency domain, we use wavelet-based network to extract compression related features from these patches. Finally, through contrastive learning strategy, the extracted features with the same compression degree are encouraged to be similar, while the different ones are encouraged to be away from each other. Therefore, the well-trained CRL-Net can extract corresponding compression representations from JPEG

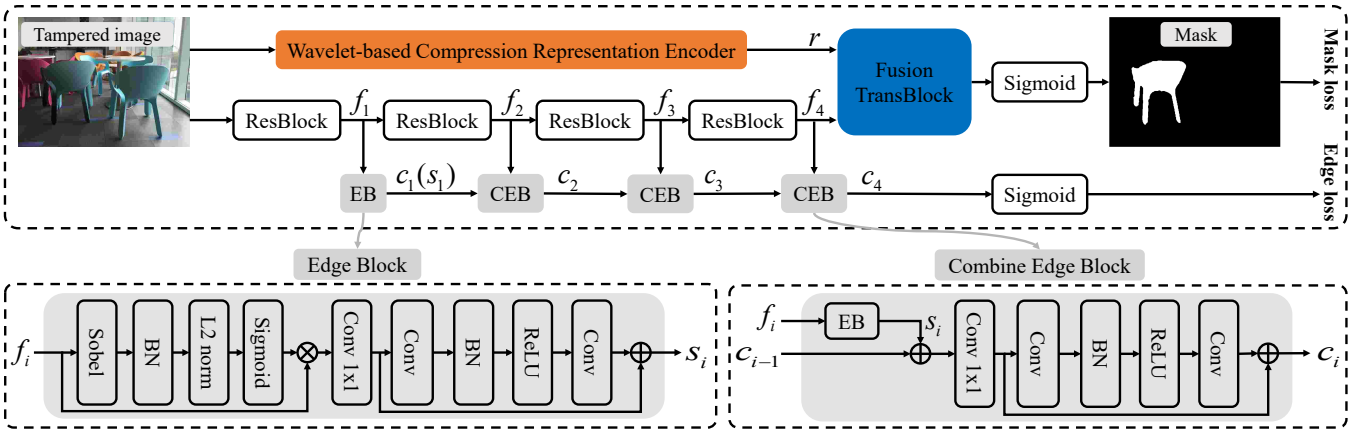

Figure 3: The overall architecture of the proposed JPEG Compression-aware Image Forgery Localization network (CA-IFL). Based on the learned compression representations, we train the CA-IFL network to predict the accurate localization masks for JPEG compressed forged images. Specifically, the CA-IFL adopts ResBlocks as backbone to extract basic feature. Furthermore, it introduces two important branches. The first one is the compression-aware branch, including wavelet-based compression representation encoder and fusion TransBlock. The second one is the boundary correction branch (shown as tandem gray blocks: EB and CEBs), which aims to alleviate the edge artifacts caused by JPEG compression.

compressed forged images. **Second**, based on the learned compression representations, we train the CA-IFL network to obtain accurate forgery localization for JPEG compressed images, as shown in Figure 3. The CA-IFL consists of two important branches: the wavelet-based compression representation encoder as mentioned above and the boundary correction branch, which is designed to alleviate the edge artifacts caused by JPEG compression. The details of the CRL-Net and CA-IFL will be discussed below.

3.2 Compression Representation Learning Network (CRL-Net)

As mentioned above, in the first step, we train the CRL-Net to obtain the compression representations, which is highly related with JPEG compression degrees. As illustrated in Figure 2, we utilize wavelet-based architecture and contrastive learning strategy for compression representation learning. The specific architecture and training strategy are described below.

Get patches. We utilize JPEG compressed forged images to get patches. Given an image patch as the query patch (annotated with a yellow box in Figure 2). The patches with the same compression quality factor (QF) are considered as positive patches (annotated with a red box). We simply take query patch and positive patches from the same image. Conversely, patches from different QFs are referred as negative patches (annotated with blue and green boxes). The CRL-Net takes the three types of patches as input.

Wavelet-based compression representation encoder. The architecture of the encoder is depicted in Figure 2. It takes the query patch, positive patches and negative patches as input and aims to predict their corresponding compression representations. Since the compression information is highly sensitive in frequency domain, we adopt 2D discrete wavelet transform (DWT) to convert image into frequency domain (as suggested in WaveFill [43]). Specifically, for the first iteration of the decomposition, we utilize DWT to apply low-pass and high-pass wavelet filters alternatively along image columns and rows (followed by downsampling), which produces 4 sub-bands including LL, LH, HL, and HH. Then, for the second iteration, the LL is used to produce LL^2, LH^2, HL^2, and HH^2. The LL^2 is used for the third iteration and so on. Hence, if there are T iterations, $3T+1$ wavelet sub-bands are produced, which include LL^T, $\{LH^i\}_{i=1}^T$, $\{HL^i\}_{i=1}^T$, $\{HH^i\}_{i=1}^T$. Note that the LL^T captures low-frequency information, while the LH^i, HL^i and HH^i capture the horizontal, vertical and diagonal high-frequency information. Particularly, we set $T=2$ and obtain 7 corresponding sub-bands. LL^2 represents low-frequency information. LH^2, HL^2 and HH^2 are concatenated in the channel dimension to obtain high-frequency information at level 2. Similarly, LH^1, HL^1 and HH^1 are concatenated to represent high-frequency information at level 1. Then, as shown in Figure 2, they are fed into the corresponding CNNs to obtain multi-scale features. These feature are concatenated and encoded (ResNet [15]) to obtain the final compression representation.

Contrastive learning strategy. As described above, by utilizing the wavelet-based encoder, we encode the query, positive and negative patches into compression representations, respectively. Then the representations are further fed into the multi-layer perceptron (MLP) to obtain x, x^+ and x^- for loss calculation, as shown in Figure 2. Note that x and x^+ have the same JPEG compression quality factor, while x and x^- have different QFs. Therefore, x is encouraged to be similar to x^+ and dissimilar to x^-. Inspired by the methods [3, 33, 42], the compression representation learning (CRL) loss function is expressed as:

$$L_{CRL} = -\log \frac{\exp(x \cdot x^+/\tau)}{\sum_{n=1}^{N} \exp(x \cdot x_n^-/\tau)}, \tag{1}$$

where N represents the number of negative patches and τ is a hyper-parameter. Obviously, CRL loss encourages the extracted features

with the same QF to be similar, while the extracted features with different QFs to be away from each other. The well-trained CRL-Net can encode JPEG compressed forged images into distinguishable compression representations, which can be used for JPEG-resistant IFL network as described below.

3.3 JPEG Compression-Aware Image Forgery Localization Network (CA-IFL)

Based on the learned compression representations, we train the CA-IFL network to obtain accurate forgery localization for JPEG compressed images, as shown in Figure 3. The specific network architecture and training loss function are described below.

Network architecture. The architecture of the CA-IFL is shown in Figure 3. It takes JPEG compressed forged images y as input and aims to predict the corresponding forged region masks $\hat{m}$ as accurately as possible. We take ResBlocks [15] as backbone to extract basic features f_1, f_2, f_3, f_4 from y. Furthermore, we introduce two important branches to enhance localization accuracy against JPEG compression. The first one is the compression-aware branch, which consists of the well-trained wavelet-based compression representation encoder (shown in the orange box) and fusion TransBlock (shown in the dark blue box). The second one is the boundary correction branch (shown as tandem gray blocks: EB and CEBs), which aims to alleviate the edge artifacts caused by JPEG compression.

Specifically, for the first branch, y is encoded into compression representation r. Then r together with the basic feature f_4 are sent into a fusion TransBlock for forged region localization, denoted as

$$\hat{m} = Sigmoid\left(\mathrm{T}\left(z\right)\right), \tag{2}$$

where $z = concat(r, f_4)$ and $Sigmoid(\cdot)$ is activation. $T(\cdot)$ represents fusion TransBlock. To reduces the computational cost, we don't use global self-attention like the vanilla Transformer but perform a *Window-based Multi-head Self-Attention* [35] within non-overlapping local windows, abbreviated as $T(\cdot)$ in Eq. 2. For the second branch, the boundary correction branch consists of edge block (EB) and combined edge blocks (CEB) [4]. It takes basic features f_1, f_2, f_3, f_4 as input and gradually extracts edge information, so that the network can make refined boundary constraints. The structure of EB is at the bottom left in Figure 3. It take basic feature f_1 as input, denoted as:

$$c_1 = s_1 = EB(f_1), \tag{3}$$

where c_1 together with basic feature f_2 are the input of CEB, whose structure is at the bottom right in Figure 3, denoted as:

$$c_{i+1} = CEB(f_{i+1}, c_i), \tag{4}$$

where $i = 1, 2, 3$. The final edge-related feature c_4 is utilized to predict boundary of forged area for precise constraints. The predicted egde $\hat{e}$ can be denoted as:

$$\hat{e} = Sigmoid(c_4), \tag{5}$$

Loss Functions. As mentioned above, we obtain the predicted tampered region mask $\hat{m}$ and tampered region boundary $\hat{e}$. Then we apply pixel-level constraints on $\hat{m}$ and $\hat{e}$. Specifically, to obtain accurate localization, we conduct two loss functions, i.e., mask loss L_{mask} and edge loss L_{edge}. Inspired by the method [4], to learn from extremely imbalanced data, we adopt Dice loss to build L_{mask} and L_{edge} loss function, denoted as

$$L_{mask} = 1 - \frac{2 \cdot \sum_{i=1}^{W} \sum_{j=1}^{H} \hat{m}_{ij} \cdot m_{ij}}{\sum_{i=1}^{W} \sum_{j=1}^{H} \hat{m}_{ij}^2 + \sum_{i=1}^{W} \sum_{j=1}^{H} m_{ij}^2}, \tag{6}$$

$$L_{edge} = 1 - \frac{2 \cdot \sum_{i=1}^{W} \sum_{j=1}^{H} \hat{e}_{ij} \cdot e_{ij}}{\sum_{i=1}^{W} \sum_{j=1}^{H} \hat{e}_{ij}^2 + \sum_{i=1}^{W} \sum_{j=1}^{H} e_{ij}^2}, \tag{7}$$

where W and H are the spatial resolution of the forged image. $m_{ij} \in \{0, 1\}$ is a binary label indicating whether the (i, j) pixel is manipulated, while $\hat{m}_{ij}$ is the pixel value in the predicted mask $\hat{m}$. e_{ij} and $\hat{e}_{ij}$ indicate similar meanings as m_{ij} and $\hat{m}_{ij}$. Note that e presents the ground truth edge obtained from m. The above loss functions constitute the overall loss to train the CA-IFL network:

$$L = L_{mask} + \lambda L_{edge}, \tag{8}$$

where λ is used to balance loss functions.

Table 1: The datasets involved in our experiments. The corresponding numbers of these datasets are presented below (C-M: Copy-Move; Sp: Splicing; Re: Removal). We also annotate whether the post-processing (P-P) is applied.

Type	Dataset	Total	C-M | Sp | Re	P-P
Training	CASIAv2 [7]	5063	3235 | 1828 | 0	✓
Testing	COVER [36]	100	100 | 0 | 0	✓
	Columbia [16]	180	0 | 180 | 0	-
	NIST16 [13]	564	68 | 288 | 208	✓
	CASIAv1 [7]	920	459 | 461 | 0	✓

4 EXPERIMENTS

In this section, we conduct extensive experiments to evaluate the performance of the proposed JPEG compression-aware image forgery localization model based on the novel compression representations learning scheme. First, we introduce the training/testing datasets and experimental details. Then we investigate the effectiveness of each component involved in the proposed model. In subsection 4.3, we compare our method with the state-of-the-art methods on standard datasets. Last but not least, we evaluate the robustness of our our method to JPEG compression with multiple compression quality factors, as shown in subsection 4.4.

4.1 Experimental Setup

Datasets. The datasets involved in our experiments are summarized in Table 1. For fair comparison with the state-of-the art methods, we adopt five public datasets: CASIAv2 [7], COVER [36], Columbia [16], NIST16 [13] and CASIAv1 [7]. The CASIAv2 dataset is used for training, while the other four datasets are used for testing. Some datasets apply post processing like filtering and blurring to hide traces of tampering. Specifically, in CASIAv1 and CASIAv2, the tampered regions are carefully selected, which have various objects. Different from the other three datasets, CASIAv1 and CASIAv2 do not contain the ground truth masks. We threshold the difference between forged images and authentic images to obtain corresponding

Table 2: Ablation study on F_1 score. Our proposed framework contains CRL-Net and CA-IFL. Specifically, CRL-Net can be wavelet-based or spatial-based. CRL-Net can use contrastive learning strategy or directly regress the quality factor. While the CA-IFL contains two important components: the compression fusion block (transformer-based vs. CNN-based) and the boundary correction branch. We evaluate the effectiveness of these settings on COVER dataset. The best results are boldfaced.

Setting Number	Compression Representation Learning Network				Fusion Block		Boundary Correction	F_1
	Wavelet-based	Spatial-based	Contrastive-learning	Regression	Transformer-based	CNN-based		
1	-	-	-	-	-	-	✓	0.441
2	-	✓	✓	-	✓	-	✓	0.457
3	✓	-	-	✓	✓	-	✓	0.459
4	✓	-	✓	-	-	✓	✓	0.463
5	✓	-	✓	-	✓	-	-	0.460
6 (Ours)	✓	-	✓	-	✓	-	✓	**0.469**

GT masks. The COVER dataset covers similar objects as the pasted regions to conceal the tampering artifacts. The Columbia dataset is relatively small, which focuses on splicing manipulation. The NIST16 dataset is a challenging dataset, which contains all three tampering techniques (copy-move, splicing and removal).

Evaluation metrics. For quantitative assessment, we use pixel level F_1 as evaluation metric, which is more suitable for the image forgery localization (IFL) task. In specific, based on a threshold, the output mask $\hat{m}$ can be converted to a binary mask $\hat{m}_b$. Compared with the GT mask, we can compute pixel-level precision and recall. Their harmonic mean is pixel level F_1. Following [4], for comparison with the state-of-the-art methods on standard datasets, we adopt fixed threshold (0.5) and the optimal threshold to obtain F_1, respectively. Compared with the fixed threshold F1, the optimal threshold F1 sets different thresholds respectively. Corresponding to each threshold, we calculates the binary mask and the F1. Then we takes the maximum F1 value as optimal threshold F1, which is computationally expensive. As for the ablation study and the comparison on JPEG compressed forged images with multiple QFs, we only utilize F_1 with fixed threshold (0.5), which is more practical. In specific, the F_1 metric is defined as:

$$F_1(\hat{m}_b, m) = \frac{2TP}{2TP + FN + FP}, \tag{9}$$

Where $\hat{m}_b$ represents the binary system output mask. m represents the ground truth mask. TP represents the number of pixels classified as true positive. As similar definition, FN and FP represent false-negative numbers and false-positive numbers, respectively.

Implementation details. The proposed JPEG compression-aware image forgery localization model (CA-IFL) based on the novel compression representations learning (CRL-Net) scheme is implemented in PyTorch. The CRL-Net and CA-IFL are trained on CASIAv2 dataset. For model input, we obtain the JPEG compressed forged images I^c by compressing the forged images I with different JPEG quality factors ($QF = 50, 60, 70, 80, 90, 100$) through *Matlab* API function. We choose the range of $QF = 50 \sim 100$ because images from real social media are generally within this range and this setting is used by other comparison methods. Then, I^c is utilized to construct query, positive and negative patches to train CRL-Net. Based on the well-trained CRL encoder, I and I^c are used for CA-IFL training. The ResBlocks involved in our model are initialized with ImageNet-pretrained counterparts. For optimization, we use Adam optimizer [20] to train our model with a learning rate periodically decays from 10^{-4} to 10^{-7}. Following [4], We apply regular data augmentation for training, including flipping, blurring, etc.

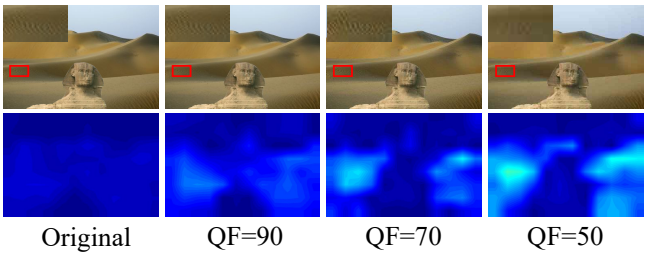

Figure 4: Visualization of the feature maps output by CRL-Net. As can be seen, CRL-Net has a low response to the clean image and high responses to JPEG compressed images, especially in the area where artifacts are serious.

4.2 Ablation Study

Our proposed network contains compression representation learning network (CRL-Net) and JPEG compression-aware image forgery localization network (CA-IFL). Specifically, CRL-Net can be divided into wavelet-based net and spatial-based net. Besides CRL-Net can use contrastive learning strategy or directly regress the quality factor. While the CA-IFL contains two important components: the compression fusion block, which can be be transformer-based or CNN-based [10], and the boundary correction branch. We conduct ablation experiments on the above settings to investigate the effectiveness, which are shown in the Table 2.

Effect of CRL-Net. Obviously, when removing the CRL-Net, compression fusion block no longer exists. As shown in the Table 2, comparing the performance of setting-1 and setting-6, F_1 score drops from 0.469 to 0.441, which indicates that, by introducing JPEG compression representation, CRL encoder can guide the CA-IFL network to effectively locate the tampered region. To prove that our CRL-Net is able to predict compression-related representations, we visualize the learned feature maps in Figure 4. Obviously, CRL-Net has a low response to the clean image and high responses to JPEG compressed images, especially in the area where artifacts

Table 3: F_1 score with fixed|optimal threshold on four standard datasets. We conduct on four public datasets without additional compression operation. Note that F_1 score with fixed threshold (0.5) is more practical. The best results are in bold.

Dataset	MFCN [31]	RGB-N [47]	HP-FCN [21]	ManTra-Net [38]	CR-CNN [1]	GSR-Net [46]	MVSS-Net [4]	Ours
COVER	n.a.|n.a.	n.a.|0.379	0.003|0.199	0.286|0.772	0.291|0.470	0.285|0.489	0.453|0.824	**0.469**|**0.826**
Columbia	n.a.|0.612	n.a.|n.a.	0.067|0.471	0.364|**0.709**	0.436|0.704	0.613|0.622	0.638|0.703	**0.657**|0.690
NIST16	n.a.|0.422	n.a.|n.a.	0.121|0.360	0.000|0.455	0.238|0.428	0.283|0.456	0.292|**0.737**	**0.298**|0.721
CASIAv1	n.a.|0.541	n.a.|0.408	0.154|0.214	0.155|0.692	0.405|0.662	0.387|0.574	0.452|0.753	**0.471**|**0.767**

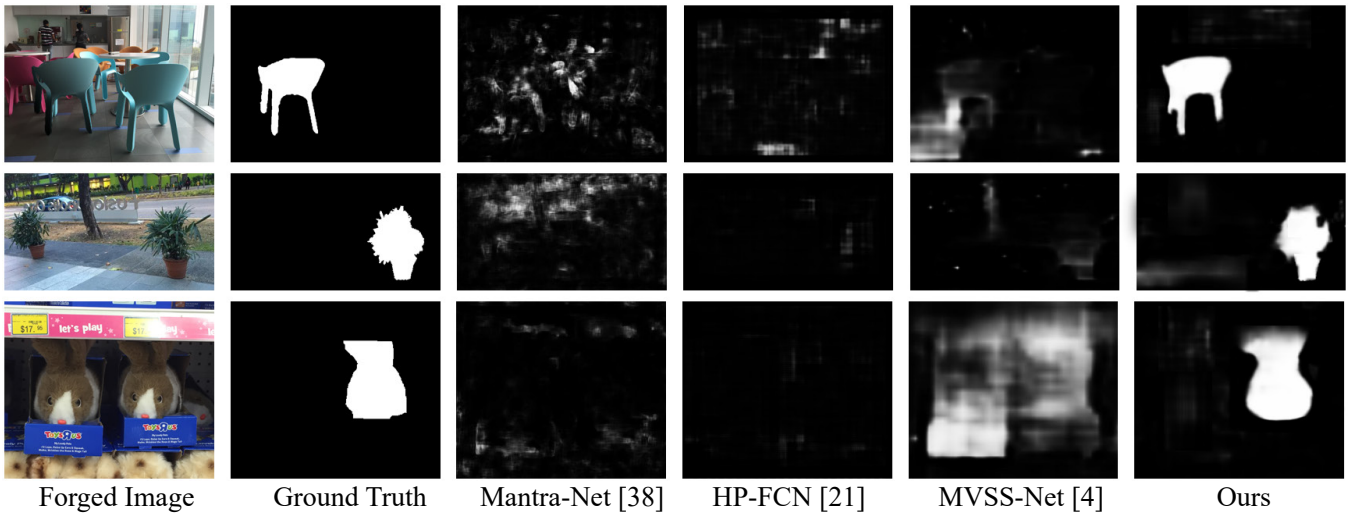

Figure 5: Hard cases. From top to bottom, these forged images have light/shadow effects, complex tampered boundaries or textures. The forged regions blend in with the surrounding scene and are indistinguishable even by human eyes.

are serious. Therefore, the learned representations, which is high related to JPEG compression artifacts, can be utilized to guide the subsequent CA-IFL network.

Effect of CRL-Net: wavelet-based vs. spatial-based. As shown in setting-2 and setting-6, when we replace wavelet-based net with spatial-based net [33], F_1 score drops from 0.469 to 0.457. This demonstrates the advantages of wavelet-based net on JPEG images which are highly correlated with frequency.

Effect of CRL-Net: contrastive learning vs. regression. To train the CRL-Net, we utilize the contrastive learning strategy instead of directly regressing the real QF values. To compare the two strategies, we train a QF-regression network with MSE loss. As setting-3 and setting-6 shown in Table 2, the F_1 score decreases from 0.469 to 0.459. It indicates that, the JPEG representation encoder training based on contrastive learning strategy is more suitable for IFL task. Because our goal is to learning a "good" compression representation rather than explicitly estimating the QF.

Effect of fusion block: transformer-based vs. CNN-based. The fusion block is utilized to introduce the learned representation into CA-IFL network. As shown in setting-4 and setting-6, the transformer-based block has a 0.003 performance improvement over the traditional CNN-based block.

Effect of boundary correction branch. As shown in Table 2, the F_1 metric of setting-5 is 0.009 lower than that of setting-6. In Figure 1, intuitively, as the compression quality factor decreases, the boundaries of the forged regions become unclear. Therefore, the refinement correction of the boundary is necessary.

4.3 Comparison with the State-of-the-Art Methods

Following the method [4], for a fair comparison, we compare our method with state-of-the-art methods: MFCN [31], RGB-N [47], HP-FCN [21], ManTra-Net [38], CR-CNN [1], GSR-Net [46] and MVSS-Net [4]. Note that all the models are trained on CASIAv2, expect for ManTra-Net [38] and HP-FCN [21], which are trained on private set. They are tested on four standard datasets (shown in Table 1) without additional compression operations. To comprehensively evaluate the superiority of our method, we evaluate F_1 scores with fixed and optimal thresholds (shown in Table 3) and show some forgery localization results on hard cases (depicted in Figure 5).

As shown in Table 3, we evaluate F_1 scores with "fixed|optimal" threshold, respectively. The description of the metric is described in subsection 4.1. As we can see, compared with other methods, our method achieves comparable or even better performance on the four standard datasets. Especially, our method reaches the best overall performance in F_1 score with the fixed threshold (0.5), which

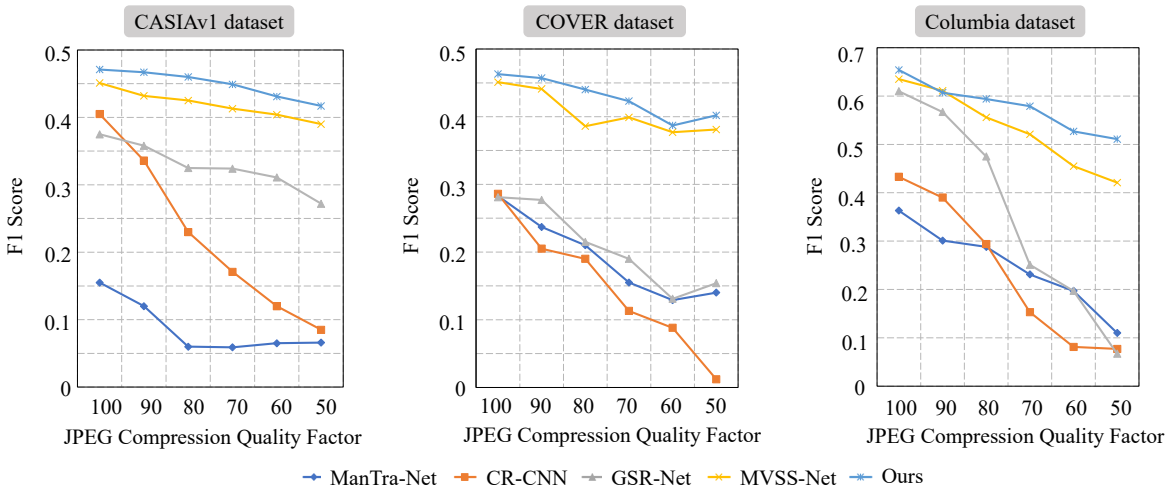

Figure 6: Robustness evaluation against JPEG compression in terms of F_1 score. We compare our method with the state-of-the-art approaches on three datasets. The JPEG compression quality factor ranges from 100 to 50.

is more commonly used in practice. In addition, we show some hard cases results in Figure 5. These forged images have complex tampered boundaries, textures, light and shadow effects, etc. The forged regions blend in with the surrounding scene and are indistinguishable even by the human eyes. As we can see, most methods fail, our method can still locate the correct regions.

Note that although our problem setting is for JPEG compressed forged images, our model also obtains superior performance on the standard datasets. This is because that JPEG compression can be regarded as a strong interference, which brings blocking artifacts, ringing effects, blurring and color distortion, etc. These complex compression artifacts overlap with the image tampering traces. While our specially designed CA-IFL network is trained to extract more intrinsic and essential features, which are highly correlated with tampering traces, so as to eliminate the effects of complex artifacts. Therefore, the forgery localization performance can be greatly improved even on hard samples that are indistinguishable. Our model can be extended to enhance forgery localization by introducing other reasonable perturbations.

4.4 Robustness on JPEG Compression Scenarios

In this section, we focus on evaluating the robustness performance of our method against JPEG compression with multiple quality factors. As shown in Figure 6, we compare our method with state-of-the-art approaches ManTra-Net [38], CR-CNN [1], GSR-Net [46] and MVSS-Net [4] on three datasetes: CASIAv1, COVER and Columbia. We show the performance (F_1 score with fixed threshold 0.5) under different compression quality factors: 100, 90, 80, 70, 60, 50. Note that ManTra-Net and MVSS-Net adopted a wide range of data augmentations including compression, while CR-CNN and GSR-Net did't use such data augmentation, which is unfair. So we mainly focus on the comparison with ManTra-Net and MVSS-Net.

Comparing Figure 6 with Table 3, it can be seen that the JPEG compression operation leads to performance drop in terms of F_1 score on every method, and as QF decreases, the performance is getting worse. While our method achieves comparable or better performance at all quality factors. Taking the CASIAv1 dataset as an example, as QF decreases from 100 to 50, the F_1 performance of CR-CNN drops from 0.405 to 0.085, ManTra-Net drops from 0.155 to 0.066, GSR-Net drops from 0.375 to 0.272 and MVSS-Net drops from 0.451 to 0.390. while our model performance drops from 0.471 to 0.417. Especially at $QF = 50$, our performance is generally much higher than other methods. Note that images are usually compressed with $QF \approx 70$ in most real social networks [27], e.g., Facebook ($QF = 71$) and Wechat ($QF = 70$). Our model can handle this situation very well. Besides, the Figure 1 at the beginning intuitively shows the impact of JPEG compression on forgery localization task. Our method can obtain more accurate localization results.

5 CONCLUSION AND DISCUSSION

In this paper, we propose a JPEG compression-aware image forgery localization network (CA-IFL) with the guidance of a wavelet-based compression representation learning network (CRL-Net). Specifically, CRL-Net utilizes contrastive learning strategy to obtain compression representations, which are highly related with JPEG compression. Based on the learned representations, we introduce the CA-IFL to flexibly handle forged images compressed with various JPEG quality factors. Additionally, a boundary correction branch is designed to alleviate the edge artifacts caused by JPEG compression. Extensive experimental results have demonstrated the superiority of the proposed method not only on standard datasets, but also on the JPEG forged images with multiple compression quality factors.

Further, other than compression, there are still some real-social-media conditions that have not been considered, such as noise, down-sampling and so on. They will bring more complex artifacts, and the tampered traces will be almost completely concealed. This will bring greater challenges to the image forgery localization task. It is harder but has practical application values. These extensions are considered as our future work.

REFERENCES

[1] Belhassen Bayar and Matthew C Stamm. 2018. Constrained convolutional neural networks: A new approach towards general purpose image manipulation detection. *IEEE Transactions on Information Forensics and Security* 13, 11 (2018), 2691–2706.

[2] Shenhao Cao, Qin Zou, Xiuqing Mao, Dengpan Ye, and Zhongyuan Wang. 2021. Metric Learning for Anti-Compression Facial Forgery Detection. In *Proceedings of the 29th ACM International Conference on Multimedia*. 1929–1937.

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 1597–1607.

[4] Xinru Chen, Chengbo Dong, Jiaqi Ji, Juan Cao, and Xirong Li. 2021. Image Manipulation Detection by Multi-View Multi-Scale Supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14185–14193.

[5] Vincent Christlein, Christian Riess, Johannes Jordan, Corinna Riess, and Elli Angelopoulou. 2012. An evaluation of popular copy-move forgery detection approaches. *IEEE Transactions on information forensics and security* 7, 6 (2012), 1841–1854.

[6] Davide Cozzolino and Luisa Verdoliva. 2019. Noiseprint: a CNN-based camera model fingerprint. *IEEE Transactions on Information Forensics and Security* 15 (2019), 144–159.

[7] Jing Dong, Wei Wang, and Tieniu Tan. 2013. Casia image tampering detection evaluation database. In *2013 IEEE China Summit and International Conference on Signal and Information Processing*. IEEE, 422–426.

[8] Max Ehrlich, Larry Davis, Ser-Nam Lim, and Abhinav Shrivastava. 2021. Analyzing and Mitigating JPEG Compression Defects in Deep Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2357–2367.

[9] Jessica Fridrich and Jan Kodovsky. 2012. Rich models for steganalysis of digital images. *IEEE Transactions on Information Forensics and Security* 7, 3 (2012), 868–882.

[10] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. 2019. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3146–3154.

[11] Xueyang Fu, Menglu Wang, Xiangyong Cao, Xinghao Ding, and Zheng-Jun Zha. 2021. A model-driven deep unfolding method for jpeg artifacts removal. *IEEE Transactions on Neural Networks and Learning Systems* (2021).

[12] Xueyang Fu, Xi Wang, Aiping Liu, Junwei Han, and Zheng-Jun Zha. 2021. Learning dual priors for jpeg compression artifacts removal. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4086–4095.

[13] Haiying Guan, Mark Kozak, Eric Robertson, Yooyoung Lee, Amy N Yates, Andrew Delgado, Daniel Zhou, Timothee Kheyrkhah, Jeff Smith, and Jonathan Fiscus. 2019. MFC datasets: Large-scale benchmark datasets for media forensic challenge evaluation. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*. IEEE, 63–72.

[14] Jing Hao, Zhixin Zhang, Shicai Yang, Di Xie, and Shiliang Pu. 2021. TransForensics: Image Forgery Localization with Dense Self-Attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15055–15064.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[16] J Hsu and SF Chang. 2006. Columbia uncompressed image splicing detection evaluation dataset. *Columbia DVMM Research Lab* (2006).

[17] Jie Huang, Xueyang Fu, Zeyu Xiao, Feng Zhao, and Zhiwei Xiong. 2022. Low-Light Stereo Image Enhancement. *IEEE Transactions on Multimedia* (2022).

[18] Jie Huang, Yajing Liu, Xueyang Fu, Man Zhou, Yang Wang, Feng Zhao, and Zhiwei Xiong. 2022. Exposure Normalization and Compensation for Multiple-Exposure Correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6043–6052.

[19] Jie Huang, Zhiwei Xiong, Xueyang Fu, Dong Liu, and Zheng-Jun Zha. 2019. Hybrid image enhancement with progressive Laplacian enhancing unit. In *Proceedings of the 27th ACM International Conference on Multimedia*. 1614–1622.

[20] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *Computer science* (2014).

[21] Haodong Li and Jiwu Huang. 2019. Localization of deep inpainting using high-pass fully convolutional network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8301–8310.

[22] Haodong Li, Weiqi Luo, Xiaoqing Qiu, and Jiwu Huang. 2017. Image forgery localization via integrating tampering possibility maps. *IEEE Transactions on Information Forensics and Security* 12, 5 (2017), 1240–1252.

[23] Siwei Lyu, Xunyu Pan, and Xing Zhang. 2014. Exposing region splicing forgeries with blind local noise estimation. *IJCV* 110, 2 (2014), 202–221.

[24] Xunyu Pan and Siwei Lyu. 2010. Region duplication detection using image feature matching. *IEEE Transactions on Information Forensics and Security* 5, 4 (2010), 857–867.

[25] Xunyu Pan, Xing Zhang, and Siwei Lyu. 2012. Exposing image splicing with inconsistent local noise variances. In *2012 IEEE International Conference on Computational Photography (ICCP)*. IEEE, 1–10.

[26] Zhihong Pan, Baopu Li, Dongliang He, Mingde Yao, Wenhao Wu, Tianwei Lin, Xin Li, and Errui Ding. 2022. Towards Bidirectional Arbitrary Image Rescaling: Joint Optimization and Cycle Idempotence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 17389–17398.

[27] Yuan Rao and Jiangqun Ni. 2021. Self-Supervised Domain Adaptation for Forgery Localization of JPEG Compressed Images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15034–15043.

[28] Yuan Rao, Jiangqun Ni, and Huimin Zhao. 2020. Deep learning local descriptor for image splicing detection and localization. *IEEE Access* 8 (2020), 25611–25625.

[29] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* 28 (2015), 91–99.

[30] Seung-Jin Ryu, Matthias Kirchner, Min-Jeong Lee, and Heung-Kyu Lee. 2013. Rotation invariant localization of duplicated image regions based on Zernike moments. *IEEE Transactions on Information Forensics and Security* 8, 8 (2013), 1355–1370.

[31] Ronald Salloum, Yuzhuo Ren, and C-C Jay Kuo. 2018. Image splicing localization using a multi-task fully convolutional network (MFCN). *Journal of Visual Communication and Image Representation* 51 (2018), 201–209.

[32] Luisa Verdoliva, Davide Cozzolino, and Giovanni Poggi. 2014. A feature-based approach for image tampering detection and localization. In *2014 IEEE international workshop on information forensics and security (WIFS)*. IEEE, 149–154.

[33] Longguang Wang, Yingqian Wang, Xiaoyu Dong, Qingyu Xu, Jungang Yang, Wei An, and Yulan Guo. 2021. Unsupervised Degradation Representation Learning for Blind Super-Resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10581–10590.

[34] Menglu Wang, Xueyang Fu, Zepei Sun, and Zheng-Jun Zha. 2021. JPEG artifacts removal via compression quality ranker-guided networks. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*. 566–572.

[35] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. 2022. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 17683–17693.

[36] Bihan Wen, Ye Zhu, Ramanathan Subramanian, Tian-Tsong Ng, Xuanjing Shen, and Stefan Winkler. 2016. COVERAGE—A novel database for copy-move forgery detection. In *2016 IEEE international conference on image processing (ICIP)*. IEEE, 161–165.

[37] Yue Wu, Wael Abd-Almageed, and Prem Natarajan. 2017. 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*. 1480–1502.

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

[39] Ruikang Xu, Mingde Yao, Chang Chen, Lizhi Wang, and Zhiwei Xiong. 2021. Continuous Spectral Reconstruction from RGB Images via Implicit Neural Representation. *arXiv preprint arXiv:2112.13003* (2021).

[40] Yanwu Xu, Shaoan Xie, Wenhao Wu, Kun Zhang, Mingming Gong, and Kayhan Batmanghelich. 2022. Maximum Spatial Perturbation Consistency for Unpaired Image-to-Image Translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18311–18320.

[41] Mingde Yao, Zhiwei Xiong, Lizhi Wang, Dong Liu, and Xuejin Chen. 2019. Spectral-depth imaging with deep learning based reconstruction. *Optics express* 27, 26 (2019), 38312–38325.

[42] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. 2018. Learning a discriminative feature network for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1857–1866.

[43] Yingchen Yu, Fangneng Zhan, Shijian Lu, Jianxiong Pan, Feiying Ma, Xuansong Xie, and Chunyan Miao. 2021. WaveFill: A Wavelet-based Generation Network for Image Inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14114–14123.

[44] Hai-Dong Yuan. 2011. Blind forensics of median filtering in digital images. *IEEE Transactions on Information Forensics and Security* 6, 4 (2011), 1335–1345.

[45] Xudong Zhao, Shilin Wang, Shenghong Li, and Jianhua Li. 2014. Passive image-splicing detection by a 2-D noncausal Markov model. *IEEE Transactions on Circuits and Systems for Video Technology* 25, 2 (2014), 185–199.

[46] Peng Zhou, Bor-Chun Chen, Xintong Han, Mahyar Najibi, Abhinav Shrivastava, Ser-Nam Lim, and Larry Davis. 2020. Generate, segment, and refine: Towards generic manipulation segmentation. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 13058–13065.

[47] Peng Zhou, Xintong Han, Vlad I Morariu, and Larry S Davis. 2018. Learning rich features for image manipulation detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1053–1061.