

Learning JPEG Compression Artifacts for Image Manipulation Detection and Localization

Myung-Joon Kwon · Seung-Hun Nam · In-Jae Yu · Heung-Kyu Lee · Changick Kim

Received: 30 Aug 2021 / Accepted: 18 Apr 2022

Abstract Detecting and localizing image manipulation are necessary to counter malicious use of image editing techniques. Accordingly, it is essential to distinguish between authentic and tampered regions by analyzing intrinsic statistics in an image. We focus on JPEG compression artifacts left during image acquisition and editing. We propose a convolutional neural network (CNN) that uses discrete cosine transform (DCT) coefficients, where compression artifacts remain, to localize image manipulation. Standard CNNs cannot learn the distribution of DCT coefficients because the convolution throws away the spatial coordinates, which are essential for DCT coefficients. We illustrate how to design and train a neural network that can learn the distribution of DCT coefficients. Furthermore, we in-

troduce Compression Artifact Tracing Network (CAT-Net) that jointly uses image acquisition artifacts and compression artifacts. It significantly outperforms traditional and deep neural network-based methods in detecting and localizing tampered regions.

Keywords image forensics · multimedia forensics · image manipulation detection · double JPEG detection · image processing

1 Introduction

With the advance of mobile devices and image editing software, image editing has become easy and popular. Together with social networking services, edited images can be spread quickly. These changes enable people to create more beautiful selfies, reduce camera shake, place an unaccompanied friend in a group photo, remove undesired objects, and share these edited images with others. However, these advances cause social problems when edited images are used as false evidence or fake news. An object-removed surveillance camera image might falsely confirm that a criminal was not at a crime scene or vice versa. A fabricated photo suggesting a celebrity scandal might damage the celebrity's reputation. Therefore, to prevent malicious image manipulation, it is critical to detect them and localize forged regions.

Among many image manipulation types, copy-and-pasting some regions onto an image either from the same image (**copy-move**) or another image (**splicing**) is one of the most popular and straightforward image editing techniques. Because these manipulations are applied to local regions, analyzing them is more challenging than kernel-based or pixel-level manipulation (*e.g.*,

Myung-Joon Kwon
School of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea
E-mail: mjkwon2021@gmail.com
ORCID: 0000-0002-9784-8440

Seung-Hun Nam
NAVER WEBTOON AI, Seongnam, South Korea
E-mail: shnam1520@gmail.com
ORCID: 0000-0002-2576-7342

In-Jae Yu
Visual Display Business, Samsung Electronics Co., Ltd., Suwon, South Korea
E-mail: injae.yu@samsung.com
ORCID: 0000-0001-9865-2194

Heung-Kyu Lee
School of Computing, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea
E-mail: heunglee@kaist.ac.kr

Changick Kim
School of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea
E-mail: changick@kaist.ac.kr

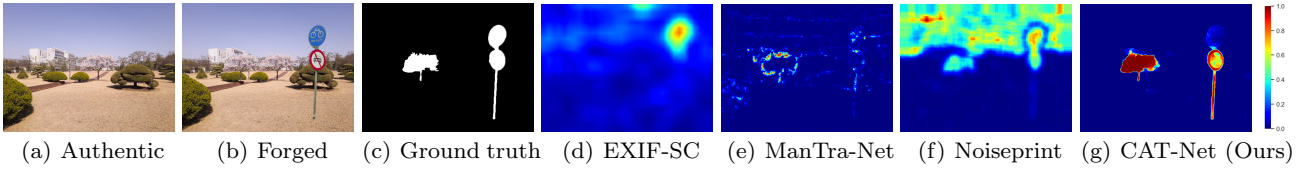


Fig. 1 Challenge of localizing manipulated regions from a JPEG image. Although many neural networks can trace noise precisely to detect manipulation, they are not ideal for capturing compression artifacts. The proposed approach considers RGB and DCT domains jointly to track visual clues and compression traces accurately. Given a possibly manipulated image (Fig. 1(b)), this study predicts the manipulated region (Fig. 1(g)). This study significantly outperforms state-of-the-art methods (Figs. 1(d), 1(e), and 1(f)) in detecting and localizing forged regions.

hue modification, blurring, contrast enhancement, or brightness adjustment) applied to the global region. Furthermore, splicing and copy-move may not leave visual clues visible to the human eyes that consider the harmony between the pristine image and the objects to be pasted (Figs. 1(a) and 1(b)). Consequently, in the last decade, many forensic approaches have been proposed to detect and localize image manipulation (Verdoliva, 2020; Korus, 2017).

A fundamental assumption underlying manipulated region detection and localization is that the image acquisition artifacts (Lukas et al., 2006) or JPEG compression artifacts (Wang and Zhang, 2016) of manipulated regions have different statistical properties from those of the pristine regions. An image acquired from a digital camera undergoes inherent internal processes. Thus, intrinsic statistical characteristics are left in digital images for each device and shooting setting. Moreover, most camera-equipped devices apply lossy compression (conventionally, JPEG) to the digital image for storage efficiency, leaving compression artifacts in the image. Characteristics of image acquisition and compression artifacts are consistently maintained within the media data if no manipulation occurs. Furthermore, the statistical characteristics of these artifacts can be changed when manipulation is applied. Image forensics aims to classify manipulated regions with different statistical fingerprints from pristine regions, so it is essential to understand detailed processes of image acquisition and JPEG compression.

First, image acquisition artifacts refer to traces from the processes applied when creating a digital image from shooting a scene. The types of representative image acquisition artifacts are as follows: lens aberration (Yerushalmy and Hel-Or, 2011), sensor pattern noise (Lukas et al., 2006), interpolation traces from the color filter array (CFA) (Bammey et al., 2020), and post-processing artifacts caused by color correction, white balance adjustment, and gamma correction (Swaminathan et al., 2008). These artifacts are device- and setting-dependent fingerprints that accompany the im-

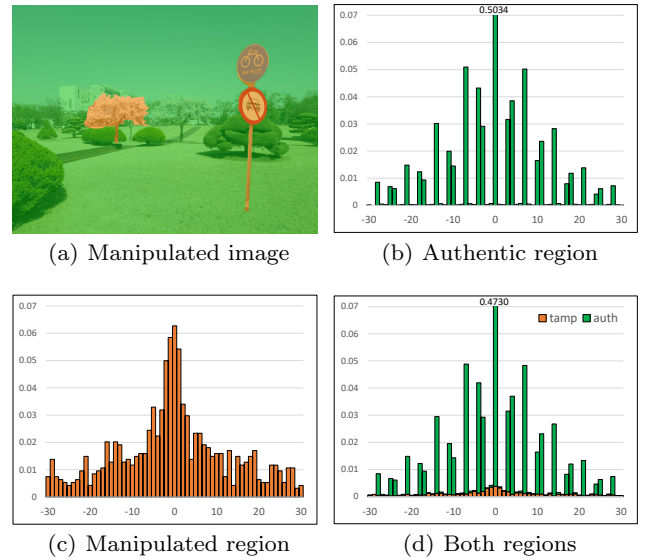


Fig. 2 Statistical differences between tampered and authentic regions. DCT histograms are obtained from Y-channel DCT coefficients at the frequency (1,1) for tampered and authentic regions separately. The x-axis is the DCT coefficient value and the y-axis is the relative frequency. The manipulated region follows a Laplacian distribution, whereas this distribution is interrupted for the authentic region. More in Sect. 2.

age acquisition process and are difficult to distinguish with the human eye. In the field of multimedia forensics, rule-based, handcrafted feature-based, and data-driven approaches to capture changes in the statistical properties of each acquisition artifact have been studied (Verdoliva, 2020). These approaches are designed to detect and expose manipulated regions by revealing inconsistencies in fine acquisition artifacts.

Second, JPEG is the most actively used compression standard to reduce storage space, leaving subtle but distinct artifacts due to quantization-based compression applied to the discrete cosine transform (DCT) domain (Barni et al., 2017). In image forensic research, double JPEG detection, *i.e.*, determining if a JPEG image has been compressed once or twice, is being actively stud-

ied (Wang and Zhang, 2016; Park et al., 2018; Verma et al., 2020). This task helps localize the manipulation regions. A region pasted onto another image likely has a statistically different distribution of Y-channel DCT coefficients compared to the authentic region (Figs. 2(b) and 2(c)). The authentic region is doubly compressed, first in a camera and again as part of the forgery, leaving periodic patterns in the histogram. The manipulated region follows a singly compressed distribution, based on the secondary quantization table (Sect. 2.2 and Popescu and Farid (2004)). Therefore, the ability to explore these compression artifacts helps in inferring and localizing the manipulated region. However, it is difficult to know in advance which region has been tampered with, *i.e.*, what we observe is the sum of two histograms (Fig. 2(d)).

Based on these observations of the two types of artifacts left in the manipulated image, we use both RGB and DCT domain information to detect and localize image manipulation. We propose an end-to-end trainable neural network-based image manipulation detector named Compression Artifact Tracing Network (CAT-Net). It traces image acquisition artifacts and JPEG compression artifacts accurately. The RGB domain enables the network to explore and learn fine-grained visual artifacts such as sensor pattern noise, block artifacts, and other acquisition artifacts. The DCT domain is used to explore compression artifacts.

However, supplying DCT coefficients directly to a convolutional neural network (CNN) is inadequate because the convolution throws away the spatial coordinates, which are crucial for DCT coefficients. Recently, Yousfi and Fridrich (2020) try to solve this problem using DCT volume representation in a steganalysis classification task. We adopt this representation in our network to learn the distribution of DCT coefficients. We demonstrate that this representation is also adequate for forgery localization tasks. Furthermore, the designed network includes only specially chosen network components to learn the image compression artifacts. Moreover, we propose a new pretraining method that uses double JPEG detection.

This paper extends our previous study (Kwon et al., 2021), which introduced a JPEG compression artifact-tracing method for image splicing detection. Whereas previous research only targeted splicing forgery, this paper also deals with copy-move forgery. New custom datasets are added to improve the performance further. More extensive experiments are performed with ten comparative methods, whereas only two methods were used in previous research. The results are reported with various metrics and newly added heatmaps. Fi-

nally, we released our code and trained weights publicly at <https://github.com/mjkwon2021/CAT-Net>.

Our main contributions are summarized as follows:

- We propose CAT-Net that learns compression artifacts based on DCT volume representation. This approach outperforms previous state-of-the-art networks using histogram representation in detecting double JPEG compression. Furthermore, we successfully transferred these weights to image manipulation detection and localization.
- CAT-Net learns the distribution of DCT coefficients without losing spatial information to finely localize tampered regions. In contrast, previous histogram approaches lose spatial information and function only for classification. CAT-Net is the first neural network that accepts DCT coefficients directly into a segmentation network.
- For the first time, CAT-Net localizes manipulated regions considering RGB and DCT domains jointly. The network captures image acquisition artifacts in the RGB domain and compression artifacts in the DCT domain. Extensive experiments with diverse benchmark datasets demonstrate that CAT-Net significantly outperforms state-of-the-art manipulation detectors.

The remainder of this paper is organized as follows. Section 2 explains forensic clues and reviews relevant previous studies. Section 3 proposes our forensic approach. Section 4 explains a double JPEG pretraining scheme and evaluates CAT-Net in terms of learning compression artifacts. Section 5 describes the main experiments, image manipulation detection and localization, and demonstrates the performance of CAT-Net. Section 6 concludes the paper.

2 Related Work

In this section, we review forensic clues including image acquisition artifacts and JPEG compression artifacts. We then introduce previous forensic approaches related to this study. Two types of inevitable artifacts remain in the digital image without manipulation: image acquisition artifacts and compression artifacts. These artifacts are essential forensic clues because their intrinsic properties differ before and after the manipulation process.

2.1 Image Acquisition Artifacts

Image acquisition artifacts denote fine artifacts generated when camera-equipped devices obtain a digital im-

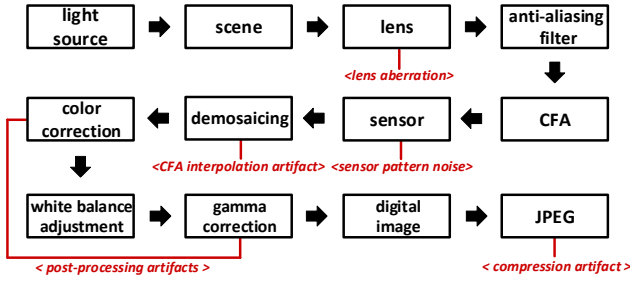


Fig. 3 Process of acquiring images from a digital camera. Red words illustrate artifacts exploited as forensic fingerprints.

age. We can detect manipulated regions by distinguishing between authentic and tampered regions. Accordingly, we should partition the image into areas from the same image. Therefore, understanding camera-specific and capture-setting artifacts caused by the image acquisition process is advantageous.

Figure 3 illustrates the detailed processes of image acquisition from a digital camera. The terms in the arrow brackets refer to acquisition artifacts generated from a specific acquisition process. Before the light from the photographed scene reaches the sensor on the digital camera, it passes through the lenses, an anti-aliasing filter (*i.e.*, optical low-pass filter), and a CFA. Because of the minor defects in the manufacturing process of a lens, the lens produces several types of image aberrations: spherical aberration, field curvature, lens radial distortion, and chromatic distortion. For source camera identification (Choi et al., 2006) and forgery detection (Yerushalmy and Hel-Or, 2011), these lens aberrations can be used as forensic fingerprints. The light passed through the lens passes through an anti-aliasing filter, which reduces aliasing and moire patterns. Then, the light passes through a CFA before reaching the sensor. The CFA is a mosaic of color filters that block out a particular portion of the spectrum, inducing each pixel to detect only one specific color among red, green, and blue (Piva, 2013).

The sensor, composed of minimal addressable elements that collect photons and convert them into electrical signals, is a critical component of a digital camera. With an analog-to-digital converter on the imaging sensor, the voltages can be sampled to digital signals (Lukas et al., 2006). The two sensor types include the charge-coupled device (CCD) and the complementary metal-oxide-semiconductor (CMOS), which both leave fine traces of sensor pattern noise (Piva, 2013). The sensor pattern noise is caused primarily by imperfections during imaging sensor manufacturing. The main types of sensor pattern noise are fixed pattern and photo-

response nonuniformity (PRNU). Because the pattern noise is an inherent property dependent on a specific camera model, it is actively used for image forensics as a distinct feature (Lukas et al., 2006; Chierchia et al., 2014; Korus and Huang, 2016).

For a CFA-based sensor (*e.g.*, CCD or CMOS), the digitized sensor output is interpolated, exploiting the color interpolation process (*i.e.*, demosaicing) to obtain the missing pixel values for the three-color layers (Piva, 2013). In this process, CFA interpolation artifacts are applied to the image, and these traces can be used to detect image forgery (Bammey et al., 2020; Choi et al., 2013). The output signal is then further processed based on post-processing, such as color correction, white balance adjustment, and gamma correction. These post-processing processes are fine corrections for perceptual quality, during which acquisition artifacts are added to the digital image.

Finally, the digital image is written to the camera memory device in a user-selected image format. JPEG is a representative lossy compression technique for digital images that mitigates or removes the high-frequency components. This paper uses *compression* to denote lossy compression (instead of loss-less compression, which is irrelevant for compression artifacts). The details of JPEG compression are introduced in the following subsection.

2.2 JPEG Compression Artifacts

In this subsection, we review the JPEG compression process and observe the double quantization artifacts left in the DCT domain. An input image is divided into non-overlapping 8×8 blocks, each block individually transformed using the DCT. In this paper, we consider only Y-channel DCT coefficients because chroma channels (*i.e.* C_b and C_r) are less useful for forensics. The DCT coefficients are then quantized using a single 8×8 quantization matrix. Quantization is an element-wise operation described as:

$$Q_{q_1}(u) = \left\lceil \frac{u}{q_1} \right\rceil, \quad (1)$$

where q_1 is the quantization step, u is a value in the DCT domain, and $\lceil \cdot \rceil$ is a rounding operator. The quantized coefficients and the quantization table — not the spatial domain pixels — are saved in a JPEG file. The coefficients are dequantized when the image file is opened (*i.e.*, during the JPEG decoding process):

$$Q_{q_1}^{-1}(v) = q_1 v, \quad (2)$$

where v is the quantized DCT coefficient. Q_{q_1} is not mathematically invertible, so the loss of information occurs here. Double quantization can then be described as:

$$Q_{q_1, q_2}(u) = Q_{q_2}(Q_{q_1}^{-1}(Q_{q_1}(u))) = \left\lfloor \left\lceil \frac{u}{q_1} \right\rceil \frac{q_1}{q_2} \right\rfloor, \quad (3)$$

where q_1 is the primary quantization step and q_2 the secondary quantization step.

We then investigate the relationship between an initial DCT coefficients histogram and a double compressed histogram. Assume a DCT coefficient in the u_1 -th bin in the former is relocated in a u_2 -th bin in the latter, *i.e.*, $Q_{q_1, q_2}(u_1) = u_2$. Then, the number of original histogram bins $n(u_2)$ contributing to bin u_2 in the double quantized histogram can be expressed as follows (Lin et al., 2009):

$$n(u_2) = q_1 \left(\left\lfloor \frac{q_2}{q_1} \left(u_2 + \frac{1}{2} \right) \right\rfloor - \left\lfloor \frac{q_2}{q_1} \left(u_2 - \frac{1}{2} \right) \right\rfloor + 1 \right), \quad (4)$$

where $\lfloor \cdot \rfloor$ is the flooring operator and $\lceil \cdot \rceil$ is the ceiling operator. Based on Eq. (4), $n(u)$ is periodic with period $q_1/\text{gcd}(q_1, q_2)$ where gcd is the greatest common divisor. Therefore, the double compressed region has periodic patterns in the histograms of quantized DCT coefficients. For example, Fig. 2 illustrates a double compressed image with quality factor 70 followed by 90 and the effects of double quantization at frequency (1, 1). Then, $q_1 = T_{70}(1, 1) = 7$ and $q_2 = T_{90}(1, 1) = 2$ where $T_x(i, j)$ is the value of the (i, j) component of the quantization table with the quality factor x , where $i, j = 0, \dots, 7$. Thus, based on the Eq. (4), $n(7k) = 7, n(7k+1) = 0, n(7k+2) = 0, n(7k+3) = 7, n(7k+4) = 7, n(7k+5) = 0$, and $n(7k+6) = 0$ where k is an integer. This values coincide with the observation that specific bins are empty in Fig. 2(b). This periodic pattern is an example of many double compression effects (More in Sect. 2.4).

The above reasoning assumed that quantization uses rounding to the nearest integer with tie-breaking toward positive infinity. However, different operations such as rounding toward zero can be used depending on camera manufacturers or image editing software (Agarwal and Farid, 2018; Butora and Fridrich, 2020). Furthermore, information loss occurs during decoding by rounding after inverse DCT is applied and truncating to the proper image pixel range $[0, 255]$. The precision of the DCT transform also impacts the distribution of coefficients (Lukáš and Fridrich, 2003). Accordingly, quantization artifacts in real-world implementations are diverse and should be handled with care.

Table 1 Summary of image manipulation detection and localization methods. Top: methods not using deep learning, bottom: methods using deep learning.

Method	Final Decision	Forensic Clue	Localization
Lukáš and Fridrich (2003)	2-layer neural network	DCT histogram	Image-level
Ye et al. (2007)	Rule-based algorithm	DCT histogram	Block-level
Fu et al. (2007)	SVM	First digit distribution of DCT coef.	Image-level
Lin et al. (2009)	SVM	Double compression artifacts	Block-level
Mahdian and Saic (2009)	Block merging	Noise inconsistency	Block-level
Amerini et al. (2011)	Clustering	SIFT descriptor	Object-level
Bianchi and Piva (2012)	Mathematical modeling	Non-aligned requantization artifact	Block-level
Ferrara et al. (2012)	Mathematical modeling	Demosaicing artifact	Block-level
Lyu et al. (2014)	Mathematical modeling	Noise inconsistency	Block-level
Iakovidou et al. (2018)	Rule-based algorithm	JPEG grid inconsistency	Block-level
Nikoukhan et al. (2019)	Rule-based algorithm	Number of zeros in the 8×8 DCT blocks	Block-level

Method	Backbone Network	Forensic Clue	Localization
Wang and Zhang (2016)	CNN	DCT histogram	Image-level
Barni et al. (2017)	CNN	DCT histogram	Image-level
Park et al. (2018)	CNN	DCT histogram + quantization table	Image-level
Bayar and Stamm (2018)	CNN	Noise residual with constrained layer	Image-level
Zhou et al. (2018)	Faster R-CNN	Visual tampering artifact+Noise	Object-level
Boroumand et al. (2018)	CNN	Noise residual with unpooled layer	Image-level
Huh et al. (2018)	SiameseNet	EXIF metadata inconsistency	Block-level
Bi et al. (2019)	U-Net	Image essence property	Pixel-level
Wu et al. (2019)	VGG+ConvLSTM	Anomalous feature	Pixel-level
Kniaz et al. (2019)	GAN	Semantic inconsistency	Pixel-level
Cozzolino and Verdoliva (2019)	SiameseNet	Camera model fingerprint	Pixel-level
Bamney et al. (2020)	CNN	Local CFA inconsistency	Block-level
Marra et al. (2020)	Xception	Spatial anomalies with noise residual	Image-level
Hu et al. (2020)	VGG	Anomalous feature	Pixel-level
Liu and Pun (2020)	DenseNet	Noise and JPEG discrepancies	Image-level
Ours	HRNet	Acquisition and compression artifacts	Pixel-level

The manipulated and authentic portions exhibit different statistical distributions in the DCT histogram. The authentic regions are compressed twice. The tampered region is treated as single compression because the 8×8 grid used in the second compression is likely misaligned with the primary compression grid (with probability $\frac{63}{64}$). Even when the two grids align, blocks containing the boundary of the pasted object have both authentic and tampered pixels, so these blocks do not follow the double compression rules (Wang and Zhang, 2016).

2.3 Image Forensics Using Image Acquisition Artifacts

Image forensics aims to verify the authenticity of media content by detecting and exploring manipulation artifacts. It is challenging to localize and detect manipulation applied to local regions (*e.g.*, splicing or copy-move) and related studies are steadily progressing. Table 1 summarizes historic image forensic approaches. The forensic clues used by each approach are categorized primarily into image acquisition and JPEG compression artifacts. This subsection reviews previous studies that explore traces of image acquisition. These studies either use acquisition artifacts directly as forensic features or explore low-level features to detect statistical changes on acquisition artifacts caused by image manipulation.

Mahdian and Saic (2009) propose a forgery localization method using local noise standard deviation estimated based on tiling the high-pass wavelet coefficient. Amerini et al. (2011) use a scale-invariant feature trans-

form (SIFT) to detect copy-move forgery; the pairs of SIFT descriptors between the pristine and manipulated regions are selected using a clustering algorithm. Ferrara et al. (2012) perform block-based forgery detection exploring demosaicing artifacts, a subtle deformation applied to the original CFA pattern during the forgery process. Lyu et al. (2014) formulate blind noise estimation as an optimization problem and detect local noise inconsistency to localize region forgery.

Inspired by computer vision tasks that have achieved significant progress after adopting CNNs, CNNs are actively exploited in image forensics to localize forged areas and detect fine-grained manipulation clues (Verdoliva, 2020; Nam et al., 2020; Yu et al., 2020). Bayar and Stamm (2018) propose a constrained layer-based network that jointly suppresses the content of a given image and adaptively learns features from noise-like signals generated by image manipulation. Zhou et al. (2018) place SRM kernel (Fridrich and Kodovsky, 2012) as a pre-processing layer and uses the Faster R-CNN (Ren et al., 2015) architecture to detect the manipulated area in units of object-level. Boroumand et al. (2018) place unpooled layers at the early part of the network, which extracts rich features of low-level signals and illustrates excellent performance for steganalysis. Huh et al. (2018) propose a self-supervised approach to train a model and explores the inconsistency of EXIF metadata. Their research exhibits outstanding performance in localizing manipulation but requires significant computation to compute the consistency for every patch pair. Bi et al. (2019) frame manipulation localization as a segmentation problem. They design a neural network based on U-Net (Ronneberger et al., 2015) and analyze the conventional semantic property by providing an RGB pixel image as input to the network.

Wu et al. (2019) design a ManTra-Net that extracts features using the SRM kernel and constrained layer for preprocessing and performs pixel-wise anomaly detection. Their research classifies various types of manipulation successfully. However, the performance of the forgery localization is not robust to JPEG compression because it uses compression as one of the manipulation types. Kniaz et al. (2019) propose a generative adversarial network-based framework for training a discriminative segmentation model to localize manipulated regions. Cozzolino and Verdoliva (2019) propose an approach to extract intrinsic noise of a camera model (*i.e.*, Noiseprint), where the content of a given image is suppressed and acquisition artifacts are enhanced. Their study explores anomalies with respect to the dominant pristine model for localizing manipulated parts.

Bammey et al. (2020) exploit an unsupervised CNN that learns to explore the underlying pattern of CFA interpolation artifacts and detects suspicious regions by identifying local mosaic inconsistencies. Marra et al. (2020) present a framework comprising of three phases — patch-wise feature extraction, image-wise feature aggregation, and global decision — that enables the use of rich features gathered at full resolution from the whole image. Hu et al. (2020) present a local self-attention block-based CNN that models and establishes the spatial relationship between patches at multiple scales to capture forensic fingerprints in forgery localization. Liu and Pun (2020) propose a fusion network that concentrates on learning low-level features and explores forensic hypotheses such as noise and JPEG discrepancies.

2.4 Image Forensics Using JPEG Compression Artifacts

In this subsection, we review previous forensic methods exploiting forensic clues caused by JPEG compression. Lam and Goodman (2000) mathematically illustrate that histograms of JPEG DCT coefficients follow the Laplacian distribution. Image editing conventionally involves additional compression and breaks the distribution, leaving compression traces in the image. Therefore, JPEG compression artifacts have been used as important fingerprints in image forensics.

Lukáš and Fridrich (2003) observe a fundamental characteristic left in the DCT domain when an image is forged. A pasted portion of a forged image likely exhibits traces of single compression, while the rest of the authentic region exhibits signs of double compression. The researchers present properties of missing values and double peaks in a histogram of DCT coefficients to detect double compression and estimate the primary quantization table. Fu et al. (2007) observe that the distribution of the first digits of the DCT coefficients follows Benford’s law, which is violated if the image is double compressed. This law is used for Q-factor estimation and double JPEG detection. Ye et al. (2007) use inconsistency of JPEG blocking artifacts to detect forgeries. A blocking artifact measure is calculated based on the estimated quantization table using the power spectrum of the DCT coefficient histogram. Lin et al. (2009) use JPEG double quantization effects such as periodic peaks and valleys in the DCT histogram to detect image forgeries automatically at the scale of 8×8 blocks. A block posterior probability map computed from histograms is thresholded to differentiate the tampered and authentic regions. Bianchi and Piva (2012) design a unified statistical model characterizing the DCT coefficients for both aligned and nonaligned double JPEG

compression. The model is used to compute a likelihood map indicating the probability of each DCT block being doubly compressed. Iakovidou et al. (2018) use JPEG grid alignment abnormalities for forgery detection. The method evaluates multiple grid positions using a fitting function, where lower contribution areas are identified as grid discontinuities. Nikoukhah et al. (2019) perform global grid detection via determining the likeliest JPEG blocks containing the largest number of zero coefficients to detect grid non-alignment caused by splicing.

In the deep learning era, there have been subsequent studies on double JPEG detection using DCT histograms to generate neural network features. Wang and Zhang (2016) are the first to use histogram features as input to a CNN for double JPEG detection. They also achieve forgery localization by integrating image-level classification results using an overlapping stride of 8 pixels. Barni et al. (2017) integrate histogram computation as part of a CNN, allowing GPU to construct histogram features in parallel. They also improve the structure of the CNN using two-dimensional (2D) convolutions instead of one-dimensional (1D) convolutions. Park et al. (2018) improve classification performance by appending a reshaped quantization table in fully connected layers.

CNN-based approaches with a DCT histogram are confined to image-level classification, primarily because using a DCT histogram requires a fixed size input and removes spatial information for localization. Previously, using a DCT histogram was mandatory because CNNs could not learn from naive DCT coefficients due to their predominantly decorrelated and locally heterogeneous nature. This study is the first to use a segmentation model based on DCT coefficients, which is possible due to the DCT volume replacing the DCT histogram and carefully designed network components. Moreover, we pretrain our network in a double JPEG detection task to produce rich initialization for learning compression artifacts left in the DCT coefficients.

3 Proposed Method

We describe how to extract features over DCT coefficients to learn their distributions using standard CNN components. Accordingly, we propose a JPEG artifact learning module (Fig. 4) that can be placed at the starting point of a CNN. Furthermore, we propose CAT-Net, a complete end-to-end image manipulation detection network. CAT-Net comprises an RGB stream, DCT stream, and fusion stage (Figs. 5 and 6). It accepts RGB pixels, DCT coefficients, and a quantization table as network inputs and outputs a probability map of each pixel being tampered with. We first describe four

key points that enable a CNN to learn the distribution of DCT coefficients: DCT volume representation, frequency-wise operations, grid-aligned cropping, and transfer learning from double JPEG detection. We then describe the detailed network architecture. Finally, we describe how CAT-Net processes non-JPEG images.

3.1 DCT Volume Representation

As explained in Sect. 1, CNNs cannot automatically learn the compression artifacts from raw DCT coefficients because the convolution assumes a translation-invariant property and handles every coefficient the same. However, the spatial coordinates are critical for DCT coefficients. Thus, we convert the input array of DCT coefficients, \mathbf{M} , to a binary volume (Yousfi and Fridrich, 2020) using a transformation $f : \mathbb{Z}^{H \times W} \rightarrow \{0, 1\}^{(T+1) \times H \times W}$ such that

$$f(\mathbf{M})_{t,i,j} = \begin{cases} 1, & \text{if } \text{abs}(\text{clip}(\mathbf{M}))_{i,j} = t \\ 0, & \text{otherwise} \end{cases}, \quad (5)$$

where $\text{clip}(\cdot)$ clips the array element-wise into the interval $[-T, T]$ and $\text{abs}(\cdot)$ takes element-wise absolute values. The DCT coefficients are recorded in channel indices with 0 or 1.

The $\text{clip}(\cdot)$ function is due to memory constraints. A larger T enables capturing a broader range of histogram bins (Fig. 2) but requires more GPU memory. We chose T to be 20, experimentally. The $\text{abs}(\cdot)$ function is due to the symmetry of the DCT histogram as depicted in Eq. 4. With the identity $\lfloor -x \rfloor = -\lceil x \rceil$, we obtain $n(-u) = n(u)$. Thus, information loss caused by taking absolute values is negligible, but the feature map size becomes almost half.

For manipulation localization in JPEG images, the DCT volume representation is more accurate than the DCT histogram, which detects double JPEG compression (Wang and Zhang, 2016; Barni et al., 2017; Park et al., 2018). Whereas the DCT histogram merges information patch-wise and loses its visual representation, the DCT volume maintains image resolution suitable for prediction at the pixel level. Nevertheless, the ability to extract statistical information is an improvement over DCT histograms (Sect. 4).

The DCT histogram is the result of applying global average pooling to the DCT volume. Thus, the DCT volume is a feature before losing location information. Furthermore, the convolution on this representation produces much richer statistical features such as co-occurrence. For example, consider a 3×3 kernel $K \in \mathbb{R}^{(T+1) \times 3 \times 3}$, where all the elements in K are zero except $K[m, 1, 1] = 1$, $K[n, 1, 2] = 1$, and

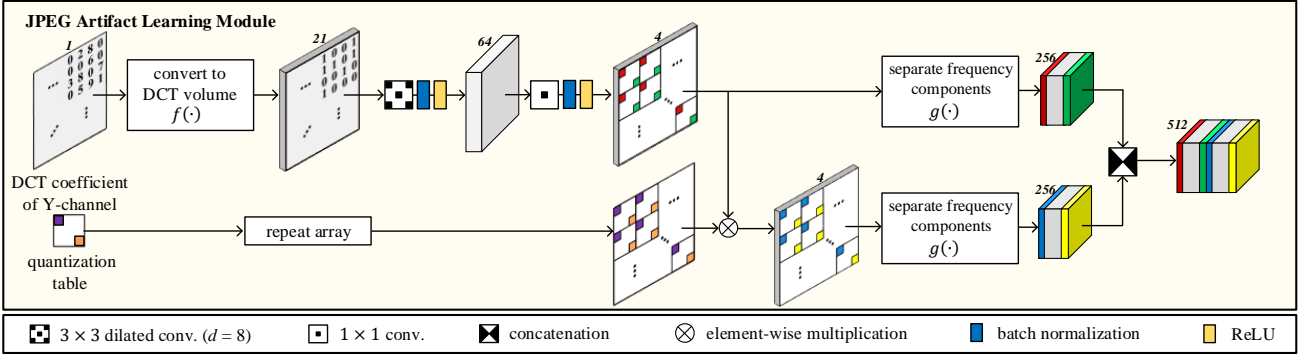


Fig. 4 Proposed JPEG artifact learning module architecture. The DCT volume conversion $f(\cdot)$ and the frequency component separation $g(\cdot)$ are depicted in Eqs. 5 and 6, respectively.

$m, n \in \{0, 1, \dots, T\}$. If a convolution using the kernel K is applied to the DCT volume and global average pooling is followed, the horizontal co-occurrence for coefficient pairs (m, n) is computed. In contrast to JPEG image steganalysis, which evaluates the probability of whole image manipulation, our goal is localizing manipulation. Therefore, we aim to extract features among different DCT blocks and use convolutions with a dilation of 8, which enables frequency-wise operations.

3.2 Frequency-wise Operations

In contrast to RGB pixels, DCT coefficients represent different frequencies depending on where they are located. The DCT coefficient at (x, y) represents frequency $(x \bmod 8, y \bmod 8)$ of the $(\lfloor \frac{x}{8} \rfloor, \lfloor \frac{y}{8} \rfloor)$ image subblock. Conventional convolutions (with stride one) mix these frequency components. All operations should be performed on a frequency-wise basis to avoid this. Namely, these include an 8×8 convolution with a dilation of 8, a 1×1 convolution, quantization table multiplication, and frequency component separation (Fig. 4). The 8×8 convolution with a dilation of 8 operates on the same frequencies because DCT coefficients consist of 8×8 blocks. The 1×1 convolution is also valid because it does not mix frequency components. A quantization table is used to help the network learn compression history.

Park et al. (2018) are the first to use a quantization table in fully connected layers. However, because our network is fully convolutional, we cannot follow their approach. We solve this problem by mimicking the JPEG decoding process. The quantization table is multiplied element-wise with a feature map. This approach uses the role of quantization tables for dequantizing quantized coefficients (Eq. 2). Quantized and dequantized feature maps are both used in our module.

Frequency component separation $g : \mathbb{R}^{C \times H \times W} \rightarrow \mathbb{R}^{64C \times \lfloor \frac{H}{8} \rfloor \times \lfloor \frac{W}{8} \rfloor}$ is an index changing mapping that can be implemented using only reshaping and permutation:

$$g(\mathbf{M})_{c,i,j} = \mathbf{M}_{\lfloor \frac{c}{64} \rfloor, 8i + \lfloor \frac{c \bmod 64}{8} \rfloor, 8j + (c \bmod 8)}, \quad (6)$$

where c, i, j starts from 0. After frequency component separation, the feature maps can be used without special care, *i.e.*, conventional 3×3 convolutions may follow.

3.3 Grid-aligned Cropping

Deep neural networks take input images of the same fixed size when training to construct a batch. Conventional computer vision networks use resizing or random cropping to satisfy this constraint. With DCT coefficients, these two methods cannot be used because they destroy position information. We propose a new cropping method that can be used for DCT coefficients. Images should be cropped in a grid-aligned scheme to enable these components to function correctly. Given an input image \mathbf{M} and crop size $h \times w$, the conventional cropped image $h(\mathbf{M})$ can be represented as:

$$h(\mathbf{M}) = \mathbf{M}[i : i + h, j : j + w], \quad (7)$$

where NumPy index slicing is used. Grid-aligned cropping requires h, w, i, j to be the multiples of eight. This simple remedy enables the neural network components described in Sect. 3.2 to function. With grid-aligned cropping, each feature map channel after frequency component separation represents one frequency. For example, the first channel corresponds to frequency $(0, 0)$, the second channel corresponds to frequency $(0, 1)$, and so on. If conventional random cropping were used instead, frequency components would be distributed over

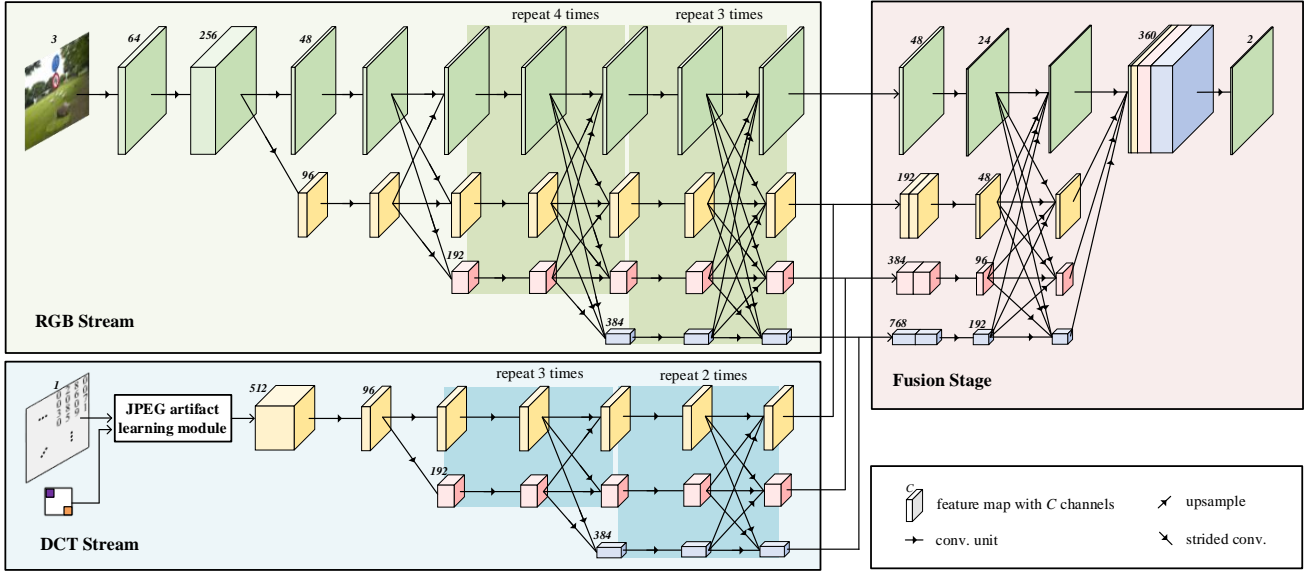


Fig. 5 Proposed CAT-Net architecture, including an RGB stream, a DCT stream, and a final fusion stage. The RGB stream takes RGB pixels and the DCT stream takes Y-channel DCT coefficients and a Y-channel quantization table as inputs. JPEG artifact learning module is depicted in Fig. 4.

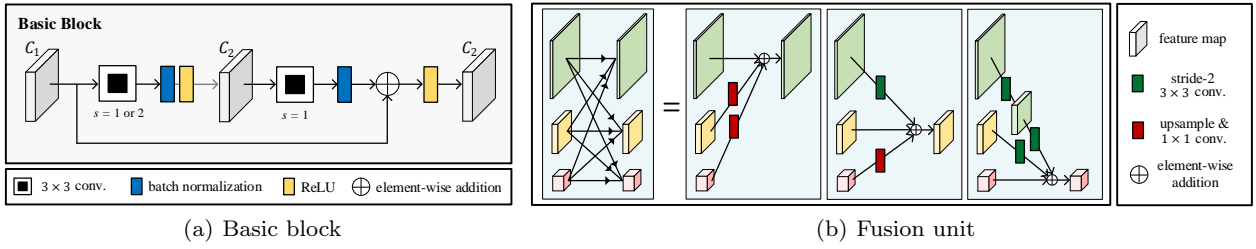


Fig. 6 Elements in the proposed network. A convolutional unit in Fig. 5 mostly consists of four consecutive basic blocks. Fusion unit fuses multi-resolution feature maps by summing them after matching resolutions.

all channels and the subsequent convolution could not distinguish frequencies.

3.4 Pretraining on Double JPEG Detection

It is common practice in the deep learning literature to start training using pretrained weights from a similar task. People often use pretrained weights from image classification for semantic segmentation, especially from ImageNet (Krizhevsky et al., 2012). We also initialize the RGB stream pretrained on ImageNet to extract visual clues more efficiently. However, being the first study to use DCT coefficients as input to a segmentation network, there is no “common practice” to pretrain the DCT stream. We introduce a new pretraining scheme on a double JPEG detection task, classifying single and double compressed JPEG images.

The DCT stream is pretrained with various quantization tables on a double JPEG detection task to

learn to handle real-world compression artifacts. The pretrained weights are transferred to the image manipulation detection task. Obtaining datasets with various compression parameters for double JPEG detection is much easier than obtaining forgery images with ground truth masks. Ablation studies demonstrate that this pretraining scheme helps the network train faster and achieve higher detection performance (Sect. 5.5).

3.5 Network Architecture

This subsection describes how we design the CAT-Net structure. CAT-Net consists of the RGB stream, DCT stream, and fusion stage. The RGB stream takes an RGB image as input and learns image acquisition artifacts, such as sensor pattern noise, EXIF metadata, blocking artifacts, or visual content itself. The DCT stream takes raw DCT coefficients and a quantization table obtained from the JPEG header as inputs and

learns compression artifacts. The network is built on top of HRNet (Wang et al., 2020). The structure of the RGB stream is HRNet itself. The DCT stream is a three-resolution variant of HRNet, replacing the first stage with our JPEG artifact learning module.

We adopt HRNet in a forensic task for the first time because it maintains high-resolution representations through the entire process, enabling us to capture the overall picture without losing fine artifacts essential for forensic investigations. With HRNet as the backbone, CAT-Net can acquire fine-grained forensic clues and learn the correlation among different regions. In addition, the HRNet feature map sizes are well suited to tracing JPEG artifacts. Because DCT is applied in 8×8 blocks, the minimum resolution the DCT stream can predict is 8×8 , which is $\frac{1}{8}$ th that of the input size (depicted in the yellow feature map in Fig. 5). This resolution can be easily joined by concatenation with the second resolution in the RGB stream. Furthermore, HRNet uses stride-2 convolution to downsample feature maps and does not use pooling layers. Recent studies have demonstrated that pooling is undesirable for tasks that require subtle signals because pooling reinforces content and suppresses noise-like signals (Boroumand et al., 2018). Although this behavior is desirable for computer vision tasks, it is inappropriate for forensic tasks because the noise-like low-level feature is an important clue.

3.6 Processing Non-JPEG Images

Although JPEG is one of the most widely used formats for storing image data, many other formats are commonly used. Non-JPEG images may not contain DCT coefficients or quantization tables required for the DCT stream. CAT-Net permits these images (assuming they are uncompressed, *e.g.*, PNG) to be its inputs. In this case, CAT-Net computes DCT coefficients by applying DCT to RGB pixel values and assumes the quantization table is filled with ones. This straightforward approach can be implemented with little difference by initially compressing the image with JPEG quality 100 without chroma subsampling. Therefore, CAT-Net can also process non-JPEG images.

A forged image with uncompressed image format does not imply that the original image is uncompressed. It only implies that the final forged image is saved without compression. Thus, compression artifacts during the image acquisition may exist, so analyzing DCT coefficients may be advantageous. However, due to the use of DCT coefficients and quantization tables, the DCT stream is not suitable for analyzing forensic clues

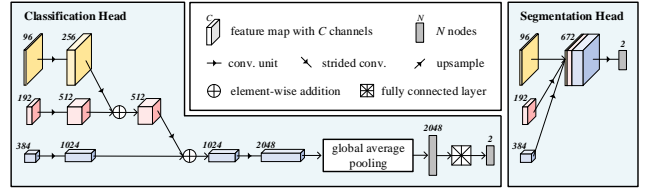


Fig. 7 DCT stream classification and segmentation head architecture. Each is attached at the end of the DCT stream to classify double JPEG images for pretraining (Sect. 4) and localize the forgery using the DCT stream for the ablation study (Sect. 5.5), respectively. RGB stream heads used for ablation study can be similarly constructed using four resolutions.

in images with other compression formats (*e.g.*, HEIC). In this case, CAT-Net must rely on the RGB stream.

4 Double JPEG Detection

Double JPEG detection is a binary classification task to determine whether a given JPEG image is JPEG compressed once or twice. This task requires the ability to analyze the compression artifacts in an image. Therefore, the DCT stream, a sub-network of CAT-Net, is pretrained on this task to capture rich compression artifacts. The primary purpose of this task is to initialize the image manipulation detection network more efficiently. A classification head is attached at the end to convert the segmentation network to a classification network (Fig. 7).

4.1 Datasets

We used 1.054 million singly and doubly-compressed JPEG images provided by Park et al. (2018). They compressed raw images (Gloe and Böhme, 2010; Bas et al., 2011; Dang-Nguyen et al., 2015) with 1,120 distinct quantization tables including 51 standard tables (Q50-Q100) and additional custom tables obtained from requested images to their public forensic web service. Consequently, their dataset closely represents real-world compression parameters. We used 21 thousand images for testing and the rest for training.

4.2 Evaluation Metrics

Because this is a binary classification task, accuracy (Acc), true positive rate (TPR), and true negative rate (TNR) are measured. We treat doubly compressed images as positives.

Table 2 Double JPEG detection performance comparison (%). Our DCT stream had the highest classification accuracy. Accordingly, the DCT stream learns the compression artifacts accurately. Thus, its weights are used as the initial weights in the image manipulation localization task.

Method	Input Type	Acc	TPR	TNR
ResNet152	RGB pixels	54.08	0.00	100.00
HRNet	RGB pixels	54.08	0.00	100.00
ManTra-Net IMTFE	RGB pixels	54.08	0.00	100.00
SRNet	RGB pixels	54.08	0.00	100.00
ResNet152	raw DCT	54.08	0.00	100.00
HRNet	raw DCT	54.08	0.00	100.00
Wang and Zhang (2016)	DCT hist. [-5, 5]	73.05	67.74	78.37
Barni et al. (2017)	DCT hist. [-60, 60]	84.46	78.35	90.53
Park et al. (2018)	DCT hist. [-60, 60] + QT	92.76	90.90	94.59
ResNet152	DCT vol. [-20, 20]	90.19	81.97	97.17
HRNet	DCT vol. [-20, 20]	91.56	84.60	97.47
DCT Stream w/o QT	DCT vol. [-20, 20]	91.71	84.97	97.42
DCT Stream (Proposed)	DCT vol. [-20, 20] + QT	93.93	89.43	97.75

$$\text{Acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \quad (8)$$

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{TNR} = \frac{\text{TN}}{\text{TN} + \text{FP}}. \quad (9)$$

4.3 Results

Table 2 illustrates double JPEG detection results. The first mega row presents four methods taking RGB pixels as inputs. Two general computer vision networks, ResNet (He et al., 2016) and HRNet (Wang et al., 2020), cannot learn compression artifacts at all. Neither can the ManTra-Net (Wu et al., 2019) feature extractor part (Image Manipulation Trace Feature Extractor) learn, as reported by the researchers (Fig. 3 of their paper). SRNet (Boroumand et al., 2018) is a steganalysis network designed to trace minute signals, but it cannot learn either. The RGB domain is not suitable for detecting JPEG double compression. The next mega row reveals that two general-purpose networks cannot learn the compression artifacts, supporting our previous claim that CNNs cannot learn the compression artifacts when raw DCT coefficients are supplied directly to them.

The third and last mega rows are the methods using the DCT histogram and DCT volume, respectively. The results reveal that the DCT histogram was indeed a suitable feature representing the distribution of DCT coefficients. The DCT volume is also a highly effective representation of compression artifacts. The DCT stream without the quantization table (*DCT Stream w/o QT*) differs from the normal DCT stream in which

Table 3 Forgery datasets used in the experiments (Sect. 5). These consist of nine publicly available datasets and five custom datasets.

Dataset		Images	JPEGs	QTs
CASIA v2	auth.	7,491	7,437	50
	tamp.	5,105	2,057	7
Fantastic Reality	auth.	16,592	16,592	153
	tamp.	19,423	19,423	1
IMD2020	auth.	414	414	58
	tamp.	2,010	1,813	73
NC16 SP	tamp.	288	288	3
Carvalho	auth.	100	0	-
	tamp.	100	0	-
Columbia	auth.	183	0	-
	tamp.	180	0	-
GRIP	auth.	80	0	-
	tamp.	80	0	-
CoMoFoD	auth.	200	0	-
	tamp.	200	0	-
COVERAGE	auth.	100	0	-
	tamp.	100	0	-
SP COCO	tamp.	200,000	200,000	41
CM COCO	tamp.	200,000	200,000	41
CM RAISE	tamp.	200,000	200,000	41
CM-JPEG RAISE	tamp.	200,000	200,000	41
JPEG RAISE	auth.	24,462	24,462	41

the quantization table path and concatenation in Fig. 4 are removed. *DCT Stream w/o QT* (52.6M) has 30% fewer parameters than HRNet (75.4M) but a higher accuracy. The proposed DCT stream achieved the highest performance among all methods. The results also demonstrate that adding a quantization table to the network increased the ability to analyze the compression artifacts. The full results confirm that the DCT stream is well designed to capture JPEG compression artifacts.

4.4 Implementation Details

The third mega row of Table 2 is from Park et al. (2018), the dataset provider. We performed all other experiments. We used a stochastic gradient descent optimizer with Nesterov momentum (0.9) and weight decay (10^{-4}). The learning rate started from 0.05 and decreased by a factor of 0.1 every 10 epochs. We trained until 30 epochs and report the test result for the highest-performing epoch.

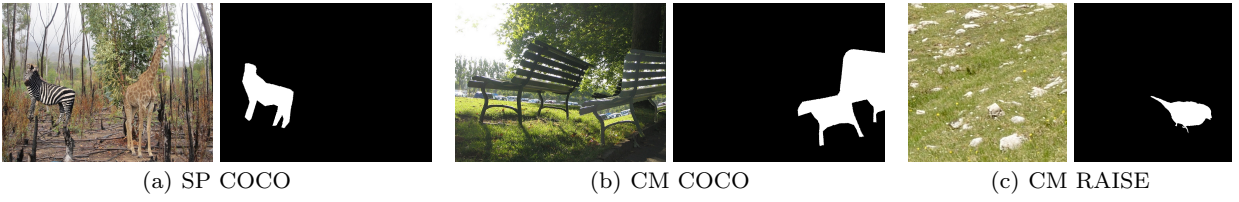


Fig. 8 Sample images of created datasets. SP COCO adds an object from another COCO image. CM COCO adds an object from the same COCO image. CM RAISE copies and pastes some random regions.

5 Image Manipulation Detection

This section illustrates the experiments of image manipulation detection and localization. After CAT-Net is initialized using ImageNet classification and double JPEG detection, it starts end-to-end training using authentic and tampered images in a supervised manner. First, we describe datasets used for training and testing. Then, the detailed training methods, baseline methods, and localization evaluation metrics are described. We then present the quantitative and qualitative localization results. Finally, the ablation studies, effect of compression quality, and robustness tests on additional compression are illustrated.

5.1 Datasets

Table 3 summarizes the datasets used in the experiments. We collected nine publicly available datasets. **CASIA v2** (Dong et al., 2013) is a popular dataset for copy-move and splicing forgery, images collected from several sources. We used masks provided by a third-party user (Pham et al., 2019) because ground truth masks are not provided officially. **Fantastic Reality** (Kniaz et al., 2019) includes many spliced images for various scenes along with ground truth masks. Although authentic images have diverse (153) quantization tables, the tampered images have only one quantization table. **IMD2020** (Novozamsky et al., 2020) includes real-life manipulated images and manually created ground truth masks. This dataset contains the most diverse quantization tables because images were collected from the Internet and hence reflect real-world compression schemes. **NC16 SP** (Guan et al., 2019) is a subset of NC16 provided by the National Institutes of Standards and Technology (NIST). NC16 contains high resolution and challenging manipulated images. Although there are several forgery types, we only use splicing forgery. **Carvalho** (De Carvalho et al., 2013) DSO-1 contains images of people. Forgeries were created by adding one or more individuals from one image to another with postprocessing to increase photo-realism. Blocking artifacts are evident when zoomed

in, indicating that although the images are not finally saved in JPEG format, the source images were JPEG compressed, leaving compression artifacts. **Columbia** (Ng et al., 2004) is a historic dataset for manipulation detection. Ground truth masks are obtained by calculating the difference between authentic and forged images followed by post-processing. The images in this dataset were not compressed in a camera, so they left no compression artifacts. **GRIP** (Cozzolino et al., 2015) contains realistic copy-move forgery images. In that dataset, the ground truth mask contains not only the tampered object region but also the source object region. Thus, we manually remove the source object region so the masks are consistent with the masks in other datasets. **CoMoFoD** (Tralic et al., 2013) contains copy-move forgeries carefully designed to make forgery detection challenging. The original images were obtained in an uncompressed format. The forged images were heavily postprocessed with JPEG compression, noise adding, image blurring, brightness change, color reduction, and contrast adjustments to hide tampering traces. Manual mask removal of the source object region was also performed on this dataset. **COVERAGE** (Wen et al., 2016) contains copy-move images designed to counter a similarity-based copy-move forgery detector. Similar but genuine objects are included deliberately, inducing many false positives for those detectors.

We distinguished training and test datasets to measure the ability to generalize over real-world data, *i.e.*, not splitting the same dataset into training and test components. We used the six smallest datasets for testing and the remaining three for training. Those three datasets contain a limited number of images and limited kinds of quantization tables, insufficient to represent real-world image distribution and compression artifacts. Thus, we created five custom datasets and used them for training (Fig. 8). **SP COCO** was constructed using the COCO 2017 dataset (Lin et al., 2014). Similar to Wu et al. (2018) and Zhou et al. (2018), spliced images were automatically created by selecting one or more arbitrary objects in one image and pasting them onto another image at random posi-

tions, with random rotation and resizing. These images were then compressed. In this paragraph, *compression* refers to JPEG compression at random quality factor ranges from 60–100. We did not apply additional post-processing, such as blurring the spliced boundary, because that might mislead the network to act like a blur detector. **CM COCO** was constructed similarly, but the copied objects came from the same image. **CM RAISE** was constructed using RAISE (Dang-Nguyen et al., 2015) as an image source but COCO as an object mask. First, the RAISE image was compressed. Then, an arbitrary region was selected using unrelated random polygon annotation from COCO. That region was then pasted within the same image, and finally, the whole image was compressed. This process often creates removal-like forgeries when the background region is selected and copy-pasted. **CM-JPEG RAISE** was constructed by simply applying additional compression to CM RAISE. This approach mimics the scenario where a forged image is sent through SNS, inducing one more compression. **JPEG RAISE** is an authentic dataset, created by simply compressing RAISE.

5.2 Implementation Details

We initialized CAT-Net weights by pretraining on ImageNet (Krizhevsky et al., 2012) classification for the RGB stream and double JPEG classification for the DCT stream (Sect. 4). The network was trained end-to-end with authentic and tampered image data. We sampled the balanced number of images in each dataset to construct one epoch and efficiently manage the wide variety of dataset sizes. Accordingly, each epoch did not include all training images but only a subset of them. Training images were cropped to 512×512 patches aligned with an 8×8 grid (Sect. 3.3). Full-resolution images were used for testing, which was possible because the proposed network was fully convolutional. The network was implemented in PyTorch (Paszke et al., 2019) using a stochastic gradient descent optimizer with a momentum of 0.9. The batch size was 22. We trained for 200 epochs. The learning rate started from 0.005 and decayed exponentially to 0 at the end. The objective was to minimize the pixel-wise binary cross-entropy loss with fivefold more weight on the tampered class. The experiments were performed using two NVIDIA TITAN RTX graphic cards.

We compared our model performance with ten other methods. The code for seven traditional methods was obtained from MKLab (Zampoglou et al., 2017): **DBA** (Ye et al., 2007), **NOI1** (Mahdian and Saic, 2009), **ADQ** (Lin et al., 2009), **NADQ** (Bianchi and Piva, 2012), **CFA** (Ferrara et al., 2012), **NOI2** (Lyu

et al., 2014), and **CAGI** (Iakovidou et al., 2018). We converted output maps in the range $[0, 255]$ to probability maps in the range $[0, 1]$. The code for three deep neural networks and the trained weights were obtained from official public repositories: **EXIF-SC** (Huh et al., 2018), **ManTra-Net** (Wu et al., 2019), and **Noiseprint** (Cozzolino and Verdoliva, 2019). For EXIF-SC, mean-shift was used for output aggregation. ManTra-Net could not infer some extra-large NC16 SP images with full resolution due to GPU memory constraints (24GB). We cropped these 268 images and their corresponding ground truth images to 2560×1440 (QHD) to test ManTra-Net.

5.3 Evaluation Metrics

Our task is a binary segmentation, labeling each pixel in the input image as tampered (positive, 1) or authentic (negative, 0). Thus, each output pixel can be marked as true positive ($G:1, P:1$), true negative ($G:0, P:0$), false positive ($G:0, P:1$), or false negative ($G:1, P:0$), where G is the ground truth mask and P is the prediction output. G and P are 2D binary arrays with the same size as the input image.

We evaluate network performance using accuracy (Acc), F1 score, and average precision (AP). The accuracy is defined as:

$$\text{Acc}(G, P) = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \quad (10)$$

However, the problem of accuracy in forensics is that there are much more negative (authentic) pixels than positive (tampered) pixels in the ground truth image. Thus, outputting all pixels as negative produces high accuracy. The F1 score is used to emphasize the positive class — it is the harmonic mean of precision and recall:

$$\text{F1}(G, P) = \frac{2}{\frac{\text{TP}}{\text{TP} + \text{FP}} + \frac{\text{TP}}{\text{TP} + \text{FN}}} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}. \quad (11)$$

Accuracy and F1 score only measure the binary decision map with a fixed threshold. Although the fixed threshold is indeed essential, we also use average precision to measure the performance free from the threshold. The average precision is the area under the precision-recall curve, which measures an average performance among all thresholds.

For forgery localization tasks, it is sometimes ambiguous which of the two segments is tampered with.

Thus, based on Huh et al. (2018), we also use the permuted metrics for evaluation, defined as:

$$\text{p-Acc}(G, P) = \max(\text{Acc}(G, P), \text{Acc}(G, P^{\complement})) \quad (12)$$

$$\text{p-F1}(G, P) = \max(\text{F1}(G, P), \text{F1}(G, P^{\complement})) \quad (13)$$

$$\text{p-AP}(G, P) = \max(\text{AP}(G, P), \text{AP}(G, P^{\complement})) \quad (14)$$

where \complement negates (flips) the prediction. Permuted metrics measure how well a model can distinguish authentic and tampered regions, not its ability to identify which is which. Some studies use permuted metrics without explicitly specifying ‘p-’ (Cozzolino and Verdoliva, 2019) or use a similar flipping strategy depending on ground truth (Wu et al., 2019) or prediction (Huh et al., 2018). Furthermore, some papers use varying thresholds per image and report the best value, resulting in much higher numbers (Cozzolino and Verdoliva, 2019; Huh et al., 2018). The authors claim a varying threshold is used to measure the performance without threshold selection ability. However, because that performance is measured via AP, we chose to use the fixed threshold (0.5) for accuracy and F1 score, to strictly measure the detection performance. Each metric is calculated per image and averaged over a dataset.

5.4 Results

Table 4 presents a performance comparison among eleven methods: seven traditional approaches, three state-of-the-art deep neural networks, and our CAT-Net. The results are depicted for six independent datasets: three splicing datasets and three copy-move datasets. All test datasets are completely unseen during training, *i.e.*, not test splits, to measure the general performance for real-world forgeries. We also report accuracy for authentic images in each dataset, if provided, to observe the false positive rate for untampered images. Figure 9 illustrates some prediction outputs of the six highest-performing methods.

Among eleven methods, CAT-Net achieves the highest localization performance for five out of six forgery datasets in terms of both p-F1 and p-AP. In particular, the CAT-Net results for GRIP are surprising (Table 4). CAT-Net achieves 76.45% p-F1 and 91.87% p-AP, while the second-best methods are 10.98% p-F1 (Noiseprint) and 15.12% p-AP (CAGI). The other ten methods could not localize the forgeries, unlike CAT-Net, which significantly outperformed those methods. Although GRIP creators tried not to leave any forgery traces, the acquisition-level compression artifacts remained and those are detected by our detector. However, CAT-Net could not detect forgeries well in CoMo-

FoD (14.01% p-F1, 21.46% p-AP), defeated by ManTra-Net (19.28% p-F1, 22.06% p-AP). The result was caused by the absence of initial compression, which indicates the DCT stream could find no traces. In contrast, although Columbia also does not contain initial compression traces, CAT-Net achieves excellent performance due to the RGB stream. CAT-Net attains 93.97% p-F1 and 95.87% p-AP, while the second-highest performing method, EXIF-SC, achieves 78.05% p-F1 and 94.50% p-AP. Their method is suitable for Columbia because this dataset is claimed to be uncompressed, so EXIF metadata is unharmed.

The permuted accuracy of authentic images suggests that DBA has the lowest false positives for untampered images. However, because its localization performance for tampered images is low in many cases, we conclude that this method predicts tampered regions relatively less often. CAT-Net achieves a high score for authentic and tampered images, implying that it could be used for image integrity verification. CAT-Net achieves state-of-the-art performance in image manipulation detection and localization.

5.5 Ablation Studies

Table 5 and Fig. 10 present the ablation study results. Two substreams are separately trained using the same training settings to observe the contribution of each stream. In some cases, the DCT stream outperforms the RGB stream and, in others, the opposite — depending on the existence of compression artifacts. If the compression artifacts exist, the DCT stream outperforms the RGB stream and vice versa if the DCT stream cannot trace meaningful compression traces. For the datasets without initial compression, the joint performance sometimes decreases because the DCT stream produces unhelpful features, negatively impacting the entire network. Nevertheless, full CAT-Net exhibits the highest overall performance using both streams.

The last row of Table 5 illustrates the effect of double JPEG pretraining. *CAT-Net w/o D.P* indicates CAT-Net started training from random initialization for the DCT stream, not from the pretrained weights using double JPEG detection. For the datasets with compression traces, pretraining on double JPEG detection improves the localization performance significantly. For example, *CAT-Net w/o D.P* attains 24.60% p-F1 and 43.43% p-AP for GRIP. The performance increases to 76.45% p-F1 and 91.87% p-AP when the training starts from double JPEG initialization. Furthermore, Fig. 11 illustrates that double JPEG pretraining produces faster training. These results are likely due to the various quantization tables used in double JPEG

Table 4 Image manipulation detection and localization performance for completely unseen datasets (%). Among eleven localization methods, CAT-Net attains the highest localization performance for five out of six forgery datasets in terms of both p-F1 and p-AP.

Method	NC16 SP		Carvalho			Columbia		
	Tamp. (SP)		Auth.	Tamp. (SP)		Auth.	Tamp. (SP)	
	p-F1	p-AP	p-Acc	p-F1	p-AP	p-Acc	p-F1	p-AP
DBA	12.60	21.13	100.00	24.48	31.87	100.00	40.87	41.48
NOI1	17.66	25.51	86.16	36.27	37.23	75.26	48.13	54.77
ADQ	17.01	14.74	81.32	40.84	37.89	83.83	41.22	37.71
NADQ	12.69	7.91	99.58	24.54	14.87	99.04	48.14	36.21
CFA	16.59	18.60	81.98	27.87	25.88	97.52	72.54	75.02
NOI2	14.26	13.18	88.32	25.84	23.74	93.42	43.28	46.70
CAGI	14.45	24.81	91.81	34.87	50.05	95.34	48.28	56.99
EXIF-SC	40.72	51.60	96.96	43.98	53.01	98.92	78.05	94.50
ManTra-Net	27.85	33.38	98.65	41.68	52.86	95.66	50.97	64.66
Noiseprint	21.51	39.89	98.58	42.12	76.79	94.07	50.42	80.85
CAT-Net (ours)	55.62	68.76	99.91	78.79	86.41	99.61	93.97	95.87

Method	GRIP			CoMoFoD			COVERAGE		
	Auth.	Tamp. (CM)		Auth.	Tamp. (CM)		Auth.	Tamp. (CM)	
	p-Acc	p-F1	p-AP	p-Acc	p-F1	p-AP	p-Acc	p-F1	p-AP
DBA	100.00	4.24	3.78	90.24	5.32	5.76	99.94	19.57	17.80
NOI1	75.32	6.07	4.50	98.20	6.28	6.67	92.15	21.88	19.97
ADQ	87.90	5.75	4.22	90.90	5.89	3.69	78.18	22.57	16.14
NADQ	99.96	4.25	2.18	99.12	5.28	2.81	99.90	19.71	11.40
CFA	70.64	8.81	12.83	78.85	6.33	5.71	75.16	22.93	17.23
NOI2	88.92	6.01	4.64	80.21	7.77	5.26	71.89	39.47	17.91
CAGI	97.73	4.83	15.12	86.50	6.92	7.23	83.36	22.58	22.92
EXIF-SC	99.71	4.26	7.94	99.78	5.15	7.29	99.98	19.57	22.15
ManTra-Net	99.08	5.47	3.92	99.12	19.28	22.06	99.22	33.58	50.24
Noiseprint	93.32	10.98	10.23	89.87	8.99	9.33	81.09	25.46	25.45
CAT-Net (ours)	99.46	76.45	91.87	98.68	14.01	21.46	93.00	41.27	53.76

Table 5 Ablation Studies (%). D.P stands for double JPEG pretraining (Sect. 4).

Method	NC16 SP		Carvalho			Columbia		
	Tamp. (SP)		Auth.	Tamp. (SP)		Auth.	Tamp. (SP)	
	p-F1	p-AP	p-Acc	p-F1	p-AP	p-Acc	p-F1	p-AP
CAT-Net	55.62	68.76	99.91	78.79	86.41	99.61	93.97	95.87
RGB Stream	42.82	55.74	99.80	40.68	61.10	99.96	94.26	96.59
DCT Stream	41.75	49.87	99.53	64.85	74.99	99.93	71.88	85.93
CAT-Net w/o D.P	47.43	60.76	99.90	51.33	68.87	99.98	95.15	98.21

Method	GRIP			CoMoFoD			COVERAGE		
	Auth.	Tamp. (CM)		Auth.	Tamp. (CM)		Auth.	Tamp. (CM)	
	p-Acc	p-F1	p-AP	p-Acc	p-F1	p-AP	p-Acc	p-F1	p-AP
CAT-Net	99.46	76.45	91.87	98.68	14.01	21.46	93.00	41.27	53.76
RGB Stream	99.97	9.46	17.63	99.66	17.48	32.99	92.50	43.03	54.84
DCT Stream	99.90	65.94	81.96	99.53	8.89	13.12	94.06	34.96	42.36
CAT-Net w/o D.P	99.97	24.60	43.43	99.66	15.53	29.95	97.84	39.46	57.04

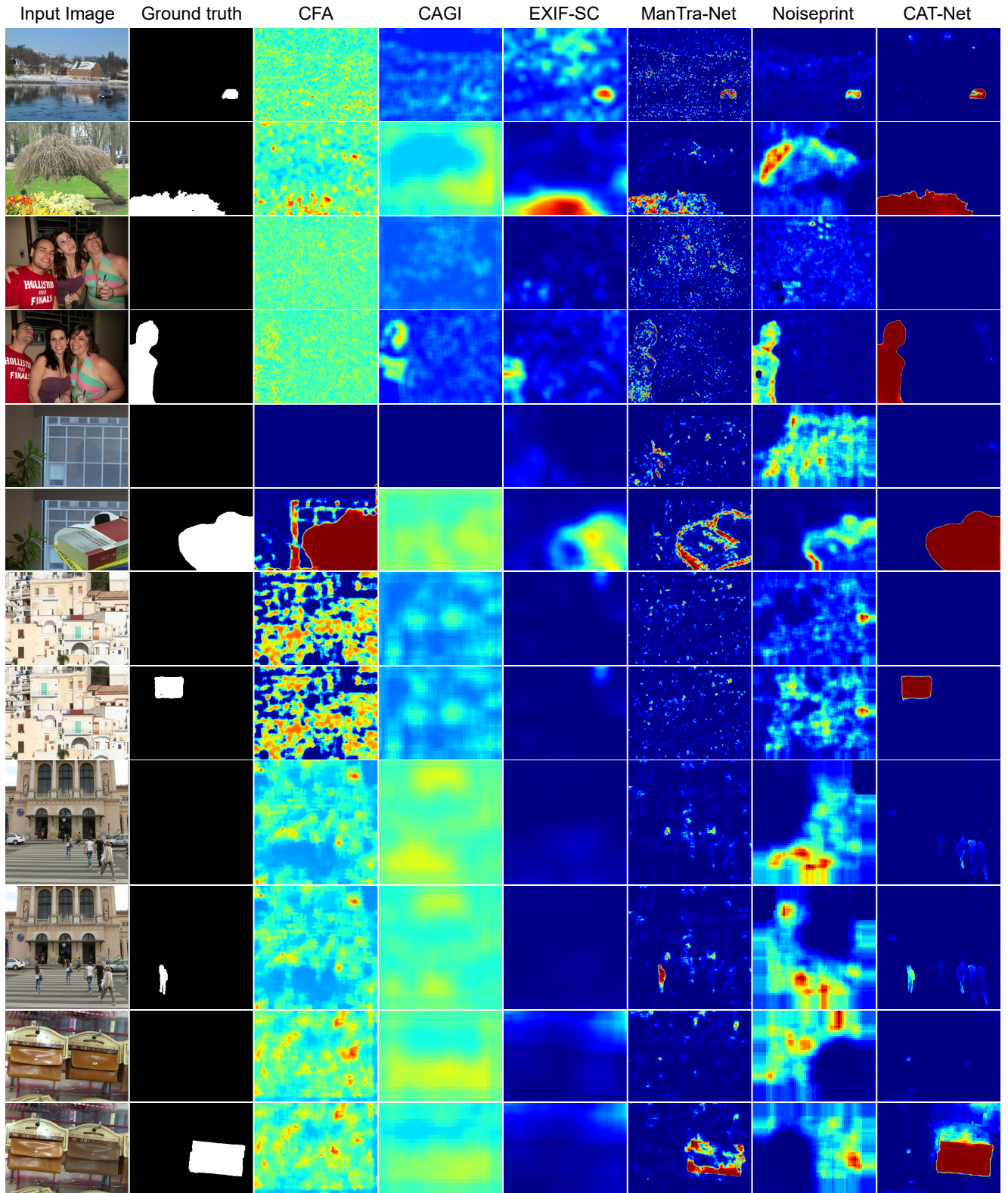


Fig. 9 Image manipulation detection and localization results. The colors indicate confidence of being tampered with. The color bar is depicted in Fig. 1. From top to bottom: 2x NC16 SP, Carvalho (auth.), Carvalho (tamp.), Columbia (auth.), Columbia (tamp.), GRIP (auth.), GRIP (tamp.), CoMoFoD (auth.), CoMoFoD (tamp.), COVERAGE (auth.), COVERAGE (tamp.).

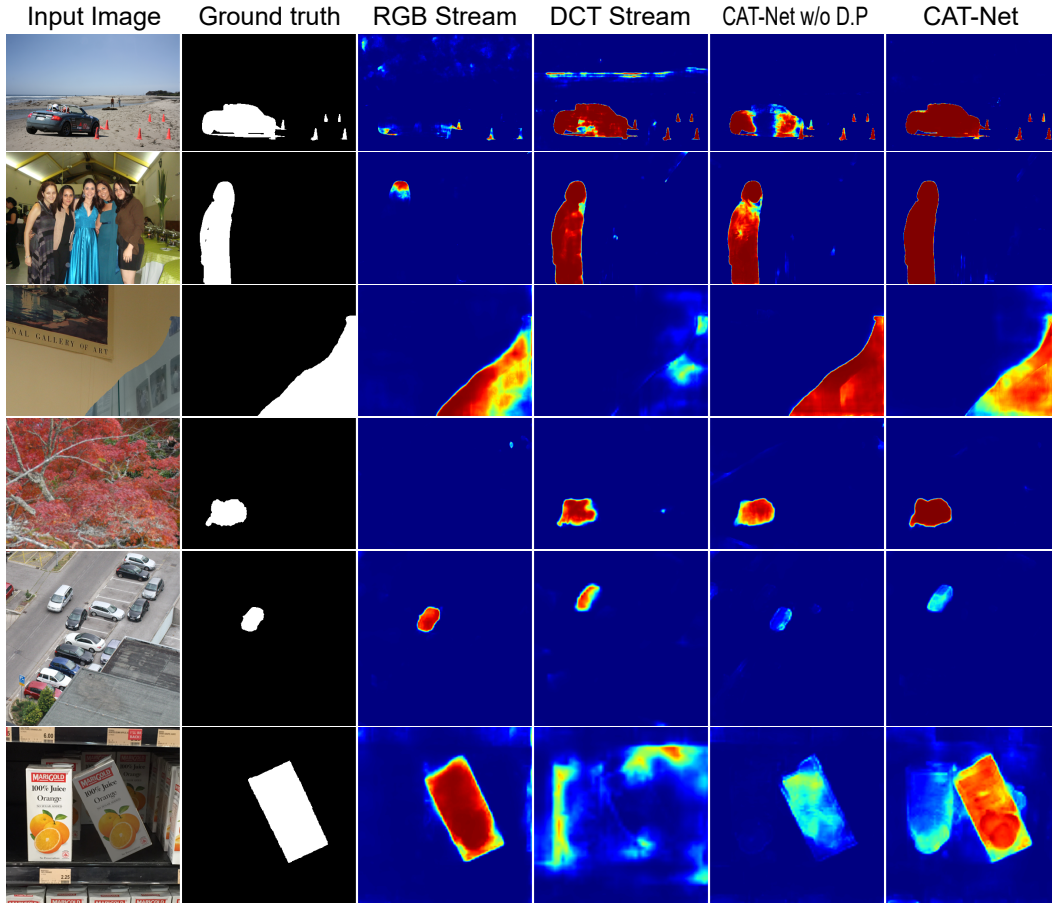


Fig. 10 Ablation studies (Sect. 5.5). From top to bottom: NC16 SP, Carvalho (tamp.), Columbia (tamp.), GRIP (tamp.), CoMoFoD (tamp.), COVERAGE (tamp.).

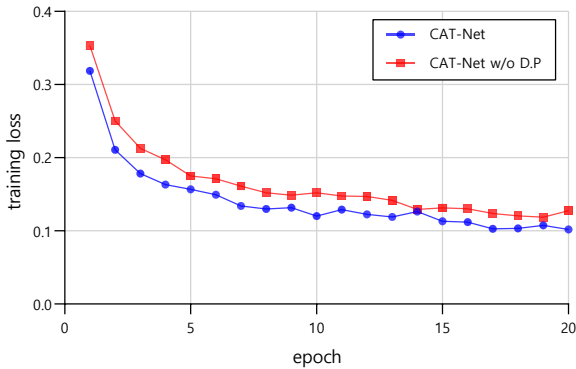


Fig. 11 Effect of double JPEG detection pretraining on the DCT stream. Pretraining on double JPEG detection produced a faster drop in training loss. The first 20 epochs out of 200 epochs are represented.

pretraining (1,120 types). It is challenging to acquire a forged image and a ground truth mask pairs for diverse quantization tables. In contrast, it is easy to ob-

tain singly and doubly-compressed images with diverse quantization tables because we may obtain raw images and compress them once or twice. Therefore, we recommend using double JPEG pretraining for future research.

However, when tested on the datasets without useful compression traces, the overall performance decreases when pretrained. For example, *CAT-Net w/o D.P* exhibits 95.15% p-F1 and 98.21% p-AP for Columbia, which decrease to 93.97% p-F1 and 95.87% p-AP when started from double JPEG pretraining. When we use pretraining, the DCT stream produces more accurate predictions during training, causing the entire network to focus more on the DCT stream than the RGB stream. When evaluating images without compression traces, this stream tries to use unavailable compression traces more frequently, reducing performance. Hence, we conclude that pretraining the DCT stream with double JPEG detection enables rich initialization of the forgery localization task where compression artifacts remain.

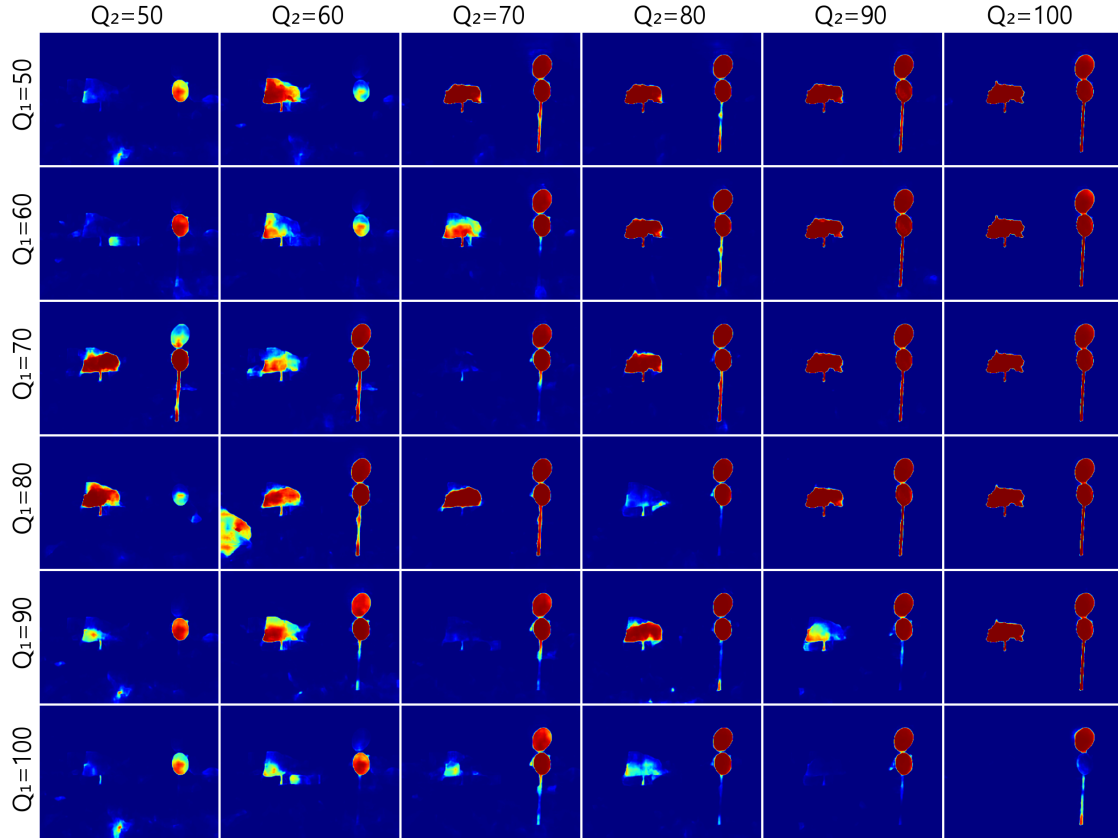


Fig. 12 Effect of first and second compression qualities on localization performance (Sect. 5.6). Forged image and ground truth are depicted in Figs. 1(b) and 1(c), respectively.

5.6 Effect of Compression Quality

This subsection analyzes the effect of first and second compression quality. We created forged images similar to Fig. 1(b) but with different JPEG compression qualities. Recall that the forged image contains the copy-moved object (the cherry blossom tree) and the spliced object (the sign). The authentic image (Fig. 1(a)) was first compressed using JPEG quality Q_1 , and another authentic image, the sign, was compressed using JPEG quality 70. Then copy-move and splicing manipulation were applied, followed by the second compression using JPEG quality Q_2 .

Figure 12 depicts CAT-Net’s localization results with diverse compression qualities. The smaller Q_1 and the larger Q_2 , the higher the localization performance. The diagonal images from the top-left to the bottom-right illustrate the special cases when $Q_1 = Q_2$. In these cases, CAT-Net had a lower chance of detecting the copy-moved object because the same quantization produces significantly fewer compression artifacts. In contrast, the spliced object was accurately detected even when the same quantization was used in the first and second compression ($Q_2 = 70$). This is mainly be-

cause the spliced object was from a different image, unlike the copy-moved object. The RGB stream detected that the object had different acquisition artifacts from the other area, so CAT-Net could localize the sign as tampered. Furthermore, both objects were hard to detect for images with low second compression qualities ($Q_2 \leq 60$) because the strong final compression extensively destroys the forensic clues.

5.7 Robustness Tests on Additional Compression

Figure 13 illustrates the localization performance when images are JPEG compressed once more. Additional JPEG compression conventionally occurs when the manipulated images are transmitted through the Internet, like posting on social media or sending via messengers. These services often use JPEG compression to reduce storage or bandwidth. Thus, detectors should maintain their performance with additional JPEG compression.

CAT-Net achieves the highest performance in 16 out of 24 settings in terms of p-F1 score and in 23 out of 24 settings in terms of p-AP. Consequently, CAT-Net is robust to additional JPEG compression for various



Fig. 13 Robustness tests on additional JPEG compression (Sect. 5.7). CAT-Net achieves the best performance in 16 out of 24 settings in terms of p-F1 score and in 23 out of 24 settings in terms of p-AP.

quality factors compared to other neural network approaches and is suitable for detecting real-world forgeries.

6 Conclusion

We presented a new approach using image compression artifacts to detect and localize image manipulation. This study is the first to accept DCT coefficients directly into a segmentation network, which was possible due to DCT volume representation and specially chosen neural network components. We also introduced a new pretraining method that uses double JPEG detection. Our neural network approach was the first to use both RGB and DCT domain information for forgery localization. Proposed CAT-Net significantly outperformed state-of-the-art forgery detectors. This study is a starting point for using compression artifacts in deep learning-based image forensics. We hope that many future studies will build upon this idea.

Acknowledgements This research was partially supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2021R1I1A1A01043600).

Conflict of interest

The authors declare that they have no conflict of interest.

References

- Agarwal S, Farid H (2018) A jpeg corner artifact from directed rounding of dct coefficients. Tech. Rep. TR2018-838, Dartmouth College
- Amerini I, Ballan L, Caldelli R, Del Bimbo A, Serra G (2011) A sift-based forensic method for copy-move attack detection and transformation recovery. *IEEE transactions on information forensics and security* 6(3):1099–1110
- Bammey Q, Gioi RGv, Morel JM (2020) An adaptive neural network for unsupervised mosaic consistency analysis in image forensics. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp 14194–14204
- Barni M, Bondi L, Bonettini N, Bestagini P, Costanzo A, Maggini M, Tondi B, Tubaro S (2017) Aligned and non-aligned double jpeg detection using convolutional neural networks. *Journal of Visual Communication and Image Representation* 49:153–163
- Bas P, Filler T, Pevný T (2011) “break our steganographic system”: the ins and outs of organizing boss. In: *International workshop on information hiding*, Springer, pp 59–70
- Bayar B, Stamm MC (2018) Constrained convolutional neural networks: A new approach towards general purpose image manipulation detection. *IEEE Transactions on Information Forensics and Security* 13(11):2691–2706
- Bi X, Wei Y, Xiao B, Li W (2019) Rru-net: The ringed residual u-net for image splicing forgery detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp 0–0
- Bianchi T, Piva A (2012) Image forgery localization via block-grained analysis of jpeg artifacts. *IEEE Transactions on Information Forensics and Security* 7(3):1003–1017
- Boroumand M, Chen M, Fridrich J (2018) Deep residual network for steganalysis of digital images. *IEEE Transactions on Information Forensics and Security* 14(5):1181–1193
- Butora J, Fridrich J (2020) Steganography and its detection in jpeg images obtained with the “trunc” quantizer. In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp 2762–2766
- Chierchia G, Poggi G, Sansone C, Verdoliva L (2014) A bayesian-mrf approach for prnu-based image forgery detection. *IEEE Transactions on Information Forensics and Security* 9(4):554–567
- Choi CH, Lee HY, Lee HK (2013) Estimation of color modification in digital images by cfa pattern change. *Forensic science international* 226(1-3):94–105
- Choi KS, Lam EY, Wong KK (2006) Source camera identification using footprints from lens aberration. In: *Proceedings of SPIE*, vol 6069, pp 172–179
- Cozzolino D, Verdoliva L (2019) Noiseprint: A cnn-based camera model fingerprint. *IEEE Transactions on Information Forensics and Security* 15:144–159
- Cozzolino D, Poggi G, Verdoliva L (2015) Efficient dense-field copy-move forgery detection. *IEEE Transactions on Information Forensics and Security* 10(11):2284–2297
- Dang-Nguyen DT, Pasquini C, Conotter V, Boato G (2015) Raise: A raw images dataset for digital image forensics. In: *Proceedings of the 6th ACM Multimedia Systems Conference*, pp 219–224
- De Carvalho TJ, Riess C, Angelopoulou E, Pedrini H, de Rezende Rocha A (2013) Exposing digital image forgeries by illumination color classification. *IEEE Transactions on Information Forensics and Security* 8(7):1182–1194

- Dong J, Wang W, Tan T (2013) Casia image tampering detection evaluation database. In: 2013 IEEE China Summit and International Conference on Signal and Information Processing, IEEE, pp 422–426
- Ferrara P, Bianchi T, De Rosa A, Piva A (2012) Image forgery localization via fine-grained analysis of cfa artifacts. *IEEE Transactions on Information Forensics and Security* 7(5):1566–1577
- Fridrich J, Kodovsky J (2012) Rich models for steganalysis of digital images. *IEEE Transactions on Information Forensics and Security* 7(3):868–882
- Fu D, Shi YQ, Su W (2007) A generalized benford’s law for jpeg coefficients and its applications in image forensics. In: Security, Steganography, and Watermarking of Multimedia Contents IX, International Society for Optics and Photonics, vol 6505, p 65051L
- Gloe T, Böhme R (2010) The ‘dresden image database’ for benchmarking digital image forensics. In: Proceedings of the 2010 ACM Symposium on Applied Computing, pp 1584–1590
- Guan H, Kozak M, Robertson E, Lee Y, Yates AN, Delgado A, Zhou D, Kheyrkhah T, Smith J, Fiscus J (2019) Mfc datasets: Large-scale benchmark datasets for media forensic challenge evaluation. In: 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW), IEEE, pp 63–72
- He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
- Hu X, Zhang Z, Jiang Z, Chaudhuri S, Yang Z, Nevatia R (2020) Span: Spatial pyramid attention network for image manipulation localization. In: European Conference on Computer Vision, Springer, pp 312–328
- Huh M, Liu A, Owens A, Efros AA (2018) Fighting fake news: Image splice detection via learned self-consistency. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 101–117
- Iakovidou C, Zampoglou M, Papadopoulos S, Kompatsiaris Y (2018) Content-aware detection of jpeg grid inconsistencies for intuitive image forensics. *Journal of Visual Communication and Image Representation* 54:155–170
- Kniaz VV, Knyaz V, Remondino F (2019) The point where reality meets fantasy: Mixed adversarial generators for image splice detection. In: Advances in Neural Information Processing Systems, pp 215–226
- Korus P (2017) Digital image integrity—a survey of protection and verification techniques. *Digital Signal Processing* 71:1–26
- Korus P, Huang J (2016) Multi-scale analysis strategies in prnu-based tampering localization. *IEEE Transactions on Information Forensics and Security* 12(4):809–824
- Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp 1097–1105
- Kwon MJ, Yu IJ, Nam SH, Lee HK (2021) Cat-net: Compression artifact tracing network for detection and localization of image splicing. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp 375–384
- Lam EY, Goodman JW (2000) A mathematical analysis of the dct coefficient distributions for images. *IEEE transactions on image processing* 9(10):1661–1666
- Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft coco: Common objects in context. In: European conference on computer vision, Springer, pp 740–755
- Lin Z, He J, Tang X, Tang CK (2009) Fast, automatic and fine-grained tampered jpeg image detection via dct coefficient analysis. *Pattern Recognition* 42(11):2492–2501
- Liu B, Pun CM (2020) Exposing splicing forgery in realistic scenes using deep fusion network. *Information Sciences* 526:133–150
- Lukáš J, Fridrich J (2003) Estimation of primary quantization matrix in double compressed jpeg images. In: Proc. Digital forensic research workshop, pp 5–8
- Lukas J, Fridrich J, Goljan M (2006) Digital camera identification from sensor pattern noise. *IEEE Transactions on Information Forensics and Security* 1(2):205–214
- Lyu S, Pan X, Zhang X (2014) Exposing region splicing forgeries with blind local noise estimation. *International journal of computer vision* 110(2):202–221
- Mahdian B, Saic S (2009) Using noise inconsistencies for blind image forensics. *Image and Vision Computing* 27(10):1497–1503
- Marra F, Gragnaniello D, Verdoliva L, Poggi G (2020) A full-image full-resolution end-to-end-trainable cnn framework for image forgery detection. *IEEE Access* 8:133488–133502
- Nam SH, Ahn W, Yu IJ, Kwon MJ, Son M, Lee HK (2020) Deep convolutional neural network for identifying seam-carving forgery. *IEEE Transactions on Circuits and Systems for Video Technology*
- Ng TT, Chang SF, Sun Q (2004) A data set of authentic and spliced image blocks. Columbia University, ADVENT Technical Report 203-2004-3
- Nikoukhah T, Anger J, Ehret T, Colom M, Morel JM, Grompone von Gioi R (2019) Jpeg grid detection based on the number of dct zeros and its application to automatic and localized forgery detection. In:

- Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp 110–118
- Novozamsky A, Mahdian B, Saic S (2020) Imd2020: A large-scale annotated dataset tailored for detecting manipulated images. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision Workshops, pp 71–80
- Park J, Cho D, Ahn W, Lee HK (2018) Double jpeg detection in mixed jpeg quality factors using deep convolutional neural network. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 636–652
- Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, et al. (2019) Pytorch: An imperative style, high-performance deep learning library. In: Advances in neural information processing systems, pp 8026–8037
- Pham NT, Lee JW, Kwon GR, Park CS (2019) Hybrid image-retrieval method for image-splicing validation. *Symmetry* 11(1):83
- Piva A (2013) An overview on image forensics. *International Scholarly Research Notices* 2013
- Popescu AC, Farid H (2004) Statistical tools for digital forensics. In: international workshop on information hiding, Springer, pp 128–147
- Ren S, He K, Girshick R, Sun J (2015) Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems, pp 91–99
- Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention, Springer, pp 234–241
- Swaminathan A, Wu M, Liu KR (2008) Digital image forensics via intrinsic fingerprints. *IEEE transactions on information forensics and security* 3(1):101–117
- Tralic D, Zupancic I, Grgic S, Grgic M (2013) Comofod—new database for copy-move forgery detection. In: Proceedings ELMAR-2013, IEEE, pp 49–54
- Verdoliva L (2020) Media forensics and deepfakes: an overview. *IEEE Journal of Selected Topics in Signal Processing* 14(5):910–932
- Verma V, Singh D, Khanna N (2020) Block-level double jpeg compression detection for image forgery localization. *arXiv preprint arXiv:200309393*
- Wang J, Sun K, Cheng T, Jiang B, Deng C, Zhao Y, Liu D, Mu Y, Tan M, Wang X, et al. (2020) Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*
- Wang Q, Zhang R (2016) Double jpeg compression forensics based on a convolutional neural network. *EURASIP Journal on Information Security* 2016(1):23
- Wen B, Zhu Y, Subramanian R, Ng TT, Shen X, Winkler S (2016) Coverage—a novel database for copy-move forgery detection. In: 2016 IEEE international conference on image processing (ICIP), IEEE, pp 161–165
- Wu Y, Abd-Almageed W, Natarajan P (2018) Image copy-move forgery detection via an end-to-end deep neural network. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, pp 1907–1915
- Wu Y, AbdAlmageed W, Natarajan P (2019) Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 9543–9552
- Ye S, Sun Q, Chang EC (2007) Detecting digital image forgeries by measuring inconsistencies of blocking artifact. In: 2007 IEEE International Conference on Multimedia and Expo, Ieee, pp 12–15
- Yerushalmy I, Hel-Or H (2011) Digital image forgery detection based on lens and sensor aberration. *International journal of computer vision* 92(1):71–91
- Yousfi Y, Fridrich J (2020) An intriguing struggle of cnns in jpeg steganalysis and the onehot solution. *IEEE Signal Processing Letters*
- Yu IJ, Nam SH, Ahn W, Kwon MJ, Lee HK (2020) Manipulation classification for jpeg images using multi-domain features. *IEEE Access* 8:210837–210854
- Zampoglou M, Papadopoulos S, Kompatsiaris Y (2017) Large-scale evaluation of splicing localization algorithms for web images. *Multimedia Tools and Applications* 76(4):4801–4834
- Zhou P, Han X, Morariu VI, Davis LS (2018) Learning rich features for image manipulation detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1053–1061