Image Inpainting Detection via Enriched Attentive Pattern with Near Original Image Augmentation

Wenhan Yang
Rapid-Rich Object Search Lab
Nanyang Technological University
wenhan.yang@ntu.edu.sg

Rizhao Cai
Rapid-Rich Object Search Lab
Nanyang Technological University
rzcai@ntu.edu.sg

Alex Kot
Rapid-Rich Object Search Lab
Nanyang Technological University
eackot@ntu.edu.sg

ABSTRACT

As deep learning-based inpainting methods have achieved increasingly better results, its malicious use, *e.g.* removing objects to report fake news or to provide fake evidence, is becoming threatening. Previous works have provided rich discussions on network architectures, *e.g.* even performing Neural Architecture Search to obtain the optimal model architecture. However, there are rooms in other aspects. In our work, we provide comprehensive efforts from data and feature aspects. From the data aspect, as harder samples in the training data usually lead to stronger detection models, we propose near original image augmentation that pushes the inpainted images closer to the original ones (without distortion and inpainting) as the input images, which is proved to improve the detection accuracy. From the feature aspect, we propose to extract the attentive pattern. With the designed attentive pattern, the knowledge of different inpainting methods can be better exploited during the training phase. Finally, extensive experiments are conducted. In our evaluation, we consider the scenarios where the inpainting masks, which are used to generate the testing set, have a distribution gap from those masks used to produce the training set. Thus, the comparisons are conducted on a newly proposed dataset, where testing masks are inconsistent with the training ones. The experimental results show the superiority of the proposed method and the effectiveness of each component. All our codes and data will be online available.

CCS CONCEPTS

• **Applied computing** → **Evidence collection, storage and analysis**.

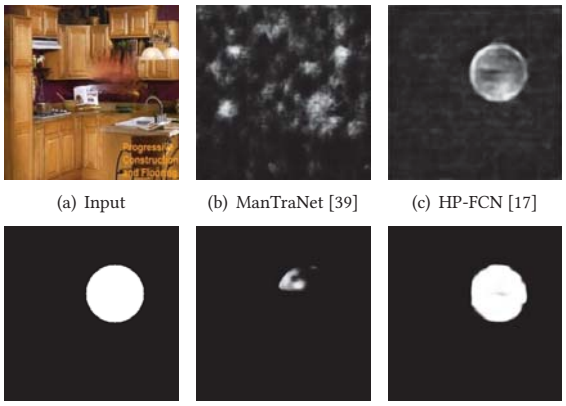
KEYWORDS

Image Inpainting Detection, Attentive Pattern, Data Augmentation, Near Original Image, Inpainting Mask

ACM Reference Format:
Wenhan Yang, Rizhao Cai, and Alex Kot. 2022. Image Inpainting Detection via Enriched Attentive Pattern with Near Original Image Augmentation. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), Oct. 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3503161.3547921

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

ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3547921



(a) Input (b) ManTraNet [39] (c) HP-FCN [17]
(d) GT (e) IIDNet [36] (f) Ours

Figure 1: Our proposed method significantly surpasses previous work for inpainting detection. The adopted inpainting method is [10].

1 INTRODUCTION

Image inpainting aims to fill the missing areas of given images based on the surrounding undamaged area with authentic contents and correct semantics. This technology has a wide range of real-world applications, such as restoring old photographs and removing unwanted people or objects. Conventional inpainting methods are built based on internal image statistics, such as fluid dynamics [3], image gradient [32], planar structure guidance [11], pixmix [9], *etc.* However, they are built on the assumption that the missing regions share the same statistical information with the surrounding undamaged ones, therefore, cannot create some complex missing object or semantic components. Later deep-learning based works make use of large-scale datasets to learn meaningful semantic features of images. Benefiting from the fascinating generative adversarial networks, these methods are capable of generating new visually authentic and semantically similar contents, providing state-of-the-art inpainting results.

However, image inpainting acts as a two-edged sword, also posing a threat about being used to remove some critical objects in images. As the inpainting methods can generate realistic inpainting results, they can be easily used to maliciously erase some important objects with newly added fake contents to obtain the visually authentic images as real ones. Hence, it is quite critical to look into how to accurately detect the inpainted images and locate the edited regions for revealing the adopted inpainting operations.

It is quite hot to detect and localize the modified regions in recent years for image forensics. The early works extract hand-crafted

patterns to detect edited regions, *e.g.* noise pattern [27], color consistency [6], compression artifact [1], copy-move trace [21], and EXIF consistency [12], *etc.* Some recent works focus on the detection of deep-learning (DL) based inpainting manipulation. As deep-learning-based inpainting methods are more capable of generating visually authentic and semantically correct new content, they poses great challenges for related detection. In [18], Li and Huang propose the seminar work on deep inpainting detection. A deep network is trained to detect specific deep inpainting traces for a specific inpainting scheme. In [36], Wu and Zhou evaluate several conventional handcrafted patterns in the convolutional neural network (CNN) framework and derive the more powerful combination. On top of that, the neural architecture search is adopted to find the optimal network architecture in a given architecture space for inpainting forgeries. In [16], Li *et al.* propose a novel data generation approach to generate a universal training dataset based on the image reconstruction conducted by the generative adversarial network (GAN) to imitate the noise discrepancies that exist in real inpainted image contents. With the wealth of existing CNN methods and related powerful tools, *e.g.* neural architecture search (NAS) and GAN, these detection methods can achieve higher accuracy compared with previous traditional methods, bringing light to the detection of deep-learning-based methods.

However, previous works neglect the following issues. First, most of the existing inpainting detection methods build their training sets based on existing inpainting methods. It is hard to synthesize less visible inpainting traces to simulate the trace generated by more competitive inpainting methods, *e.g.* the future proposed ones. Second, most of the existing methods mainly consider using the knowledge from a single inpainting method [16, 36]. The rich knowledge from diverse inpainting methods is not fully utilized. Third, in [16, 36], the transferability/generalization of inpainting traces among different inpainting methods during training is discussed briefly. However, the distributions of the inpainting masks between the training and testing phases are the same. In real applications, we seldom know what kinds of masks might be used in real inpainted images. This domain gap is not visited.

To address these issues, our work provides a comprehensive exploration from the data and feature aspects. From the data aspect, a novel near original image augmentation is proposed to pull the inpainted images close to the original image to generate the input images. Via the augmentation, we add harder samples into the training data, which leads to improved detection accuracy. From the feature aspect, we design the attentive pattern, which learns to reveal the inpainting trace with the knowledge from multiple inpainting methods. The attentive pattern learning with optimization of other parts of the network constructs a bi-level optimization problem. We provide an approximated solution and obtain desirable performance. We conduct extensive experiments on our new dataset, where testing masks are inconsistent with the training ones. The experimental results demonstrate the superiority of the proposed method and the effectiveness of each component.

Our contributions are summarized as follows,

- We propose a novel near original image augmentation. By pushing the inpainted images closer to the original images, the ones without distortion and inpainting, as the augmented training samples for the network input, the augmented training samples become harder and as a result help improve the accuracy of the detection models.
- We propose to extract the attentive pattern from inpainted images. Rather than being manually designed, the designed attentive pattern is trained in a data-driven way. As such, the knowledge of different inpainting methods can be better exploited during training.
- We are the first to evaluate the inpainting detection problem in the scenario where the masks adopted to synthesize the testing set are from an inconsistent distribution with those masks during the training phase, which is more practical for real applications.

2 RELATED WORK

2.1 DL-Based Image Inpainting

Image inpainting aims to fill the missing regions of images with visually authentic and semantically correct contents. Different from conventional methods [3, 9, 11, 32], which make use of the internal statistic information within the image, deep learning-based inpainting methods train deep networks on external large-scale diverse training data to own the capacity to infer the missing regions based on the sounding contexts. The work proposed in [30] is the pioneering work and adopts deep generative adversarial networks (GAN) for image inpainting. A later work [13] proposes a GAN injected with two context discriminators to judge the faithfulness based on global coherence and local consistency. In [42], the contextual attention is proposed to reconstruct the foreground feature patches with the background feature patches based on matching similarity scores. In [40], the features in the known region are shifted to serve as an estimation of the missing regions. In [33], the multi-scale image contextual attention is proposed to make reasonable use of the background information, along with the style loss and the perceptual loss to achieve style consistency. Some later works improve the performance of image inpainting via designing specified modules, including gated convolution [26, 43], region normalization [44], feature equalization [10], onion convolution [28], *etc.* Some methods introduce the side prior information to benefit the inference of image signals in the missing regions, including structure knowledge [14], edge information [29], semantic guidance [23], internal coherence [34], joint utilization of multiple clues [8, 24, 48], sketch tensor space [5], *etc.* There are also efforts on designing multi-scale or multi-stage networks [20, 31, 35, 45, 46], which augment the networks' capacities and help generate more visually pleasing results. In our work, we stand on the other side of the issue and try to identify to the inpainted regions to prevent malicious exploitation.

2.2 Inpainting Forensics

To fight against the malicious use of inpainting methods, some recent efforts are put into detecting and localizing inpainted regions for image forensics. The early researches design hand-crafted patterns to identify the edited/inpainted regions. Those patterns include color consistency [6], compression artifact [1], noise pattern [27], copy-move trace [21], and EXIF consistency [12], *etc.* Some works focus on the exploitation of nonlocal similarity patterns [22, 25, 38, 49] in inpainting. Namely, some inpainted blocks

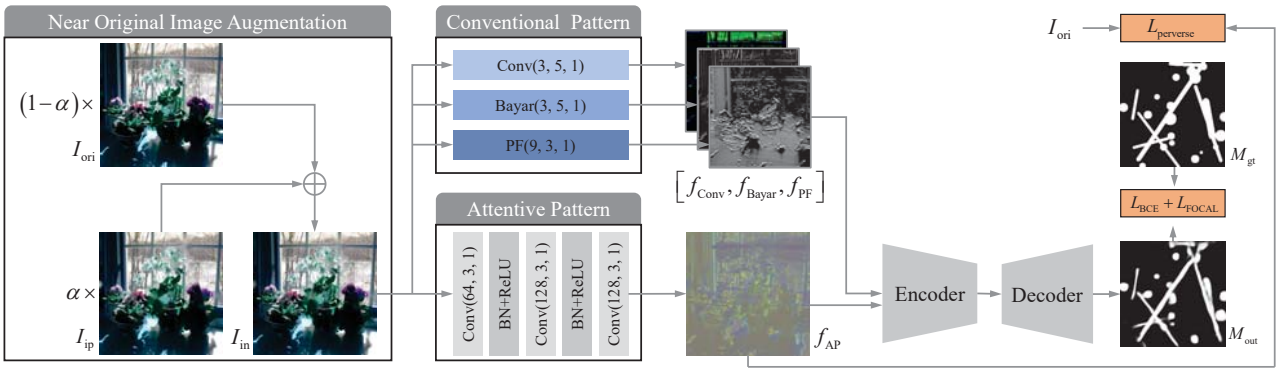


Figure 2: The framework of our proposed inpainting detection network. At the data end, the near original image augmentation is applied to create harder samples to increase the difficulty of the training set and make the model precept more diverse inpainting traces. At the feature end, a novel attentive pattern is extracted to enrich the inpainting traces with the knowledge distilled from diverse inpainting methods. (x, y, z) in each module denotes the output channel, kernel size, and stride, respectively.

are more similar to the combination of their surrounding blocks. Later works make efforts in mining statistical information, *e.g.* local variance of image Laplacian [19], and building neural networks [39].

Some recent works focus on the detection of deep-learning based inpainting manipulation. In [18], the pioneering work on deep inpainting detection is proposed to handle a specific inpainting scheme via training to detect specific deep inpainting traces. In [36], Wu and Zhou select the optimized combination of several conventional handcrafted patterns as the input of a CNN. Then, they use the neural architecture search to obtain the optimal network architecture for inpainting forgeries. In [16], a novel data generation approach is developed to generate a universal training dataset based on GAN's reconstruction to simulate the inpainting noise of real inpainted results. These detection methods achieve higher accuracy compared with traditional ones. Comparatively, in our work, we pay more attention to the data and feature aspects of the problem.

3 THE PROPOSED INPAINTING DETECTION METHOD

3.1 Overall Architecture

The proposed network follows the general pixel-wise detection architecture as shown in Fig. 2. First, the training pair is directly obtained from the training set or augmented by some data augmentation methods. First, several filters (conventional or learned) are applied to reveal the manipulation traces to provide richer clues. Then, convolutional layers are adopted to transform the preprocessed clues into meaningful features for detection. Finally, features are further augmented and projected back into the detection mask.

The previous work [36] has given full consideration to the model design, *i.e.* performing the neural architecture search to obtain the optimal model architecture in a specified module space. Comparatively, our work provides more comprehensive considerations from the data and feature aspects.

3.2 Near Original Image Augmentation

One critical issue in inpainting detection lies in the difficulty to obtain very hard examples as it is often observed that, all existing inpainting methods' results are visually unpleasant and contain obvious traces. Thus, it is quite rare to obtain training samples that include visually invisible traces. However, these traces might contribute to detecting the inpainting results generated by some well-performed methods and potentially better inpainting methods, *e.g.* some methods proposed in the future with better performance.

Therein, we propose a simple yet very effective method to generate the training samples including arbitrary small inpainting traces to benefit inpainting detection. As shown in Fig. 2, during the training, as the training set is synthesized, we have the inpainting result I_{ip} and the corresponding original image I_{ori}. Then, the training pair $(I_{\text{ip}}, M_{\text{gt}})$ can be augmented with a series of the input samples $(I_{\text{in}}, M_{\text{gt}})$, where

$$I_{\text{in}} = \alpha \times I_{\text{ip}} + (1-\alpha) \times I_{\text{ori}}, \quad \alpha \sim U(a, b), \tag{1}$$

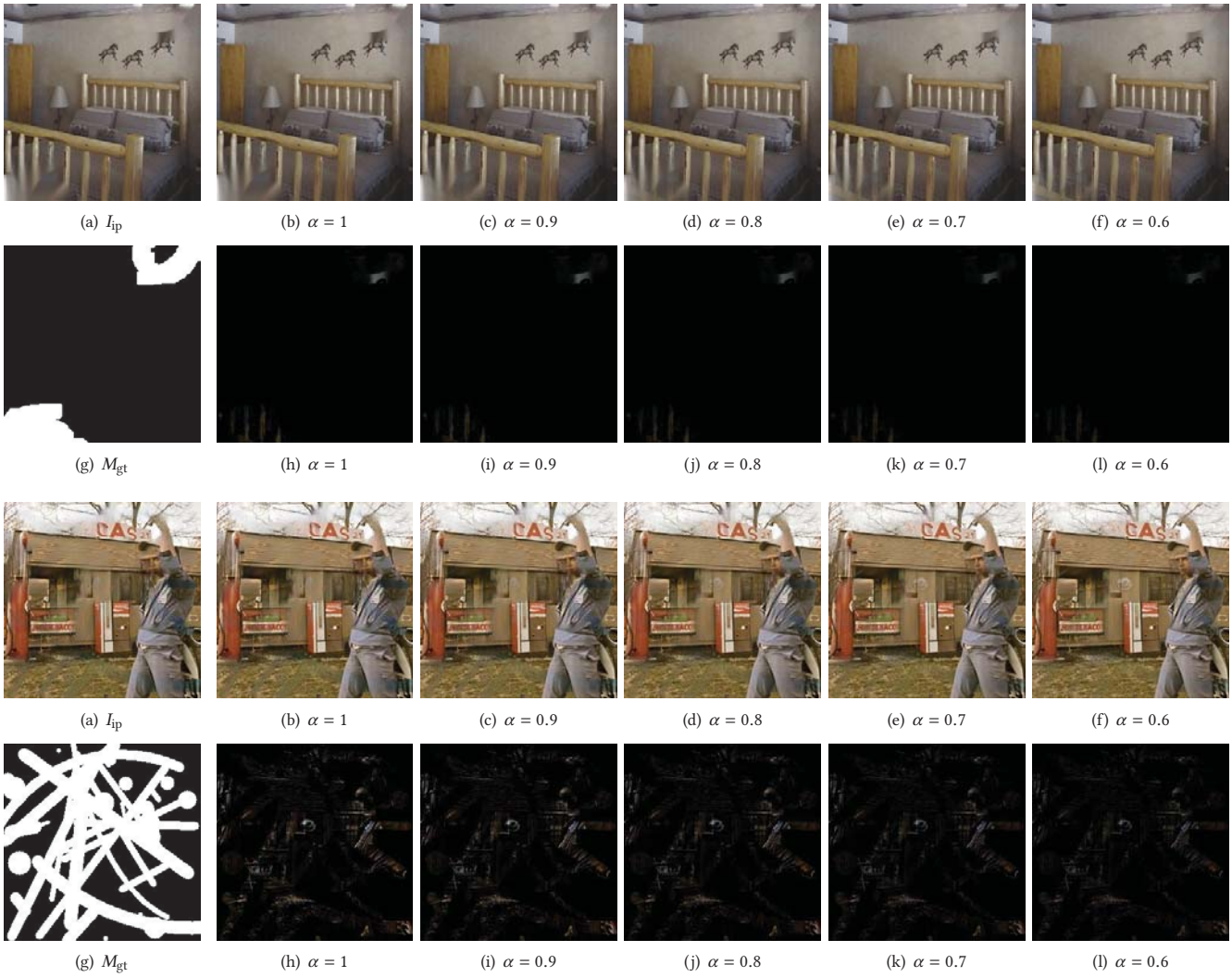
where α is sampled from a uniform distribution $U(a, b)$, the upper bound is $b = 1$, and the lower bound a defines the intensity of the augmentation, *i.e.* how many inpainted signals remaining in the augmented image.

Our proposed augmentation is simple but meaningful for two reasons. First, it enriches the training samples, especially providing abundant hard examples. Second, to some extent, it can create arbitrary small inpainting traces and is capable of simulating any future proposed methods from the statistical sense. Fig. 3 shows some visual results of our data augmentation method. The results show that our method creates the samples closer to the original image, which increases the sample hardness and trace diversities.

3.3 Inpainting Trace Extraction with Attentive Pattern

When obtaining the training pair, we first extract the inpainting trace, *i.e.* revealing the clues that might significantly contribute to the detection.

3.3.1 Conventional Pattern. Based on the analysis in the previous work [36], we adopt several conventional filters to initialize the corresponding first several convolutional layers to obtain the residual signal, which exhibits notable differences to distinguish real and inpainted details. Three conventional patterns f_{PF}, f_{Bayer} and f_{Conv}



(a) I_{ip} (b) $\alpha = 1$ (c) $\alpha = 0.9$ (d) $\alpha = 0.8$ (e) $\alpha = 0.7$ (f) $\alpha = 0.6$

(g) M_{gt} (h) $\alpha = 1$ (i) $\alpha = 0.9$ (j) $\alpha = 0.8$ (k) $\alpha = 0.7$ (l) $\alpha = 0.6$

(a) I_{ip} (b) $\alpha = 1$ (c) $\alpha = 0.9$ (d) $\alpha = 0.8$ (e) $\alpha = 0.7$ (f) $\alpha = 0.6$

(g) M_{gt} (h) $\alpha = 1$ (i) $\alpha = 0.9$ (j) $\alpha = 0.8$ (k) $\alpha = 0.7$ (l) $\alpha = 0.6$

Figure 3: Visual results of the near original image augmentation with different intensities.

are extracted by the layers initialized by Pre-Filtering layer [17], Bayar layer [2], and convolution, respectively. f_{PF} obtains the filtered residuals via calculating the gradient defined in the three directions:

$$f_{\text{PF}} = W_{\text{PF}} * I_{\text{in}}, \tag{2}$$
$$W_{\text{PF}} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 1 & 0 \end{bmatrix}; \begin{bmatrix} 0 & 0 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & 0 \end{bmatrix}; \begin{bmatrix} 0 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

f_{Bayer} adaptively predicts residual with the given constraint as follows,

$$f_{\text{Bayer}} = W_{\text{Bayer}} * I_{\text{in}}. \tag{3}$$

W_{Bayer} is learned with the following constraint:

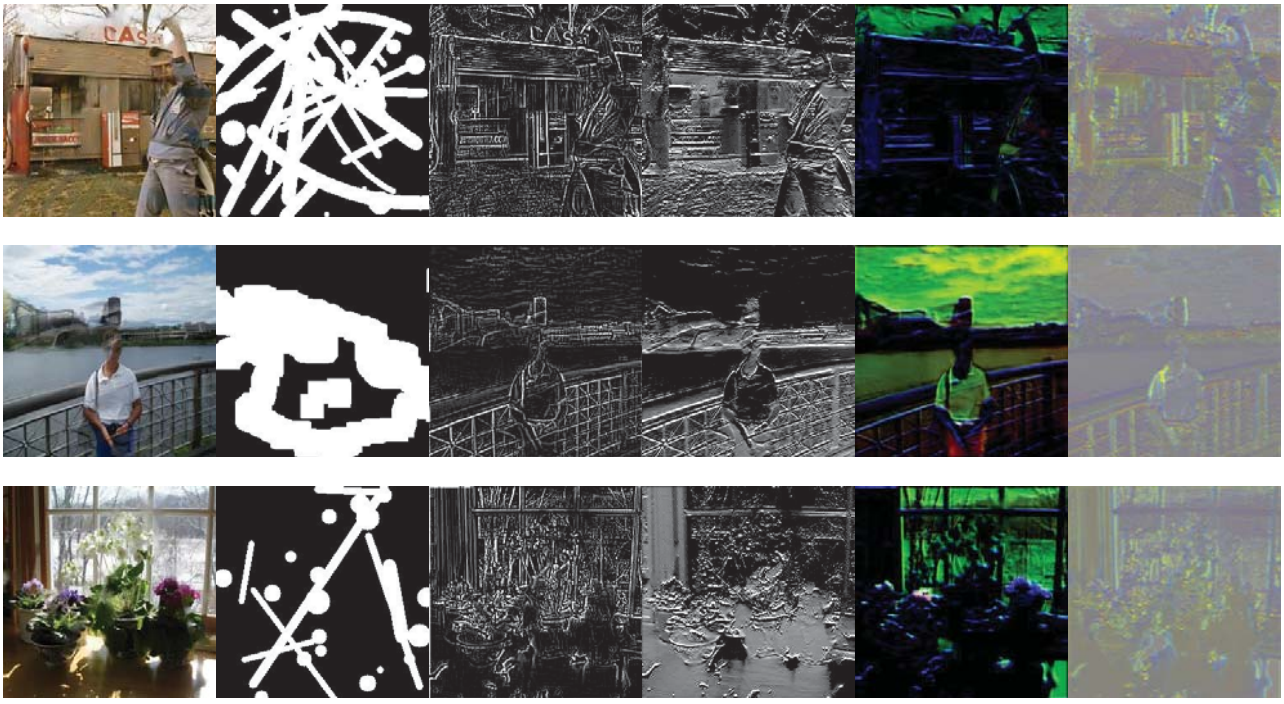
$$\begin{cases} W^c_{\text{Bayer}}(1,1) = -1 \\ \sum_{i,j} W^c_{\text{Bayer}}(i,j) = 0, \end{cases} \tag{4}$$

where c indexes the channel of W_{Bayer}. The summation of W^c_{Bayer} is zero, while its first element is -1, which is consistent with the general form of the gradient definition.

For the convolutional layer, it is set as a learnable 5×5 convolution, which learns to extract clues in a fully data-driven way without any human knowledge guidance.

We visualize the patterns extracted by the well-trained models in Fig. 4 (c)-(e). It is observed that, Pre-Filtering and Bayar layers extract the features visually similar to edges or gradients, while the learned convolutional layer changes the colors of the input images.

3.3.2 Attentive Pattern. The previous work [36] only considers training with the data generated by one inpainting method, *i.e.* gated convolution [41]. The analysis shows that, the training data mixed with the ones generated by other methods might not lead to generally better performance as the transferability among inpainting methods might be weak.



(a) I_{ip} (b) M_{gt} (c) f_{Bayer} (d) f_{PF} (e) f_C (f) f_A

Figure 4: Visual results of different trace patterns. Pre-Filtering and Bayar layers extract the features visually similar to edges or gradients, while the learned convolutional layer changes the colors of the input images. Comparatively, our proposed attentive pattern reflects more high-level clues and attends the most discriminative regions in the inpainted images for inpainting detection.

In our work, we make the first attempt to make full use of the knowledge information from various inpainting methods and propose an attentive pattern extraction method to improve the transferability among different methods. We build our attentive pattern extractor step by step.

1) Preliminary End-to-End Learned Model. As shown in Fig. 5 (a), the preliminary end-to-end trainable model F_D parameterized by θ_D can directly take the I_{in} as its input to predict the mask M_{out} as follows,

$$\begin{aligned} M_{\text{out}} &= F_D\left(I_{\text{in}}|\theta_D\right), \\ \theta &= \{\theta_D\}, \end{aligned} \tag{5}$$

where θ denotes all parameters in the whole inpainting network. This paradigm ignores existing knowledge and techniques about revealing the inpainting traces before applying a learned detection method, and cannot provide desirable performance.

2) Conventional Pattern. Another solution might first extract conventional hand-crafted features f_C by the module F_C parameterized by θ_C, then utilize the learnable F_D to predict the inpainting mask as follows,

$$\begin{aligned} f_C &= F_C(I_{\text{in}}|\theta_C), \\ M_{\text{out}} &= F_D(f_C|\theta_D), \\ \theta &= \{\theta_C, \theta_D\}. \end{aligned} \tag{6}$$

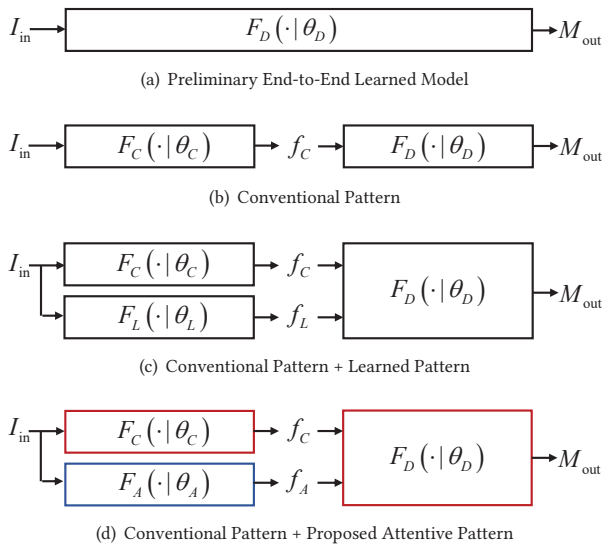
However, in this framework, the modeling capacity is limited by the design of conventional hand-crafted features.

3) Conventional Pattern + Learned Pattern. Some recent works [16, 36] introduce the hand-crafted pattern f_C and learned pattern f_L extracted by the modules F_C parameterized by θ_C and F_L parameterized by θ_L simultaneously. Based on the two kinds of clues, the inpainting mask is predicted as follows,

$$\begin{aligned} f_C &= F_C(I_{\text{in}}|\theta_C), \\ f_L &= F_L(I_{\text{in}}|\theta_L), \\ M_{\text{out}} &= F_D([f_C, f_L]\,|\theta_D), \\ \theta &= \{\theta_C, \theta_L, \theta_D\}, \end{aligned} \tag{7}$$

where $[\cdot]$ is the channel concatenation. With the wealth of both handcrafted and learned patterns, the detection performance can be significantly improved. However, as revealed in [36], the detection transferability among different inpainting methods might be weak. Namely, training with the data generated by some inpainting methods might lead to degraded detection performance. Therefore, it comes with an obstacle when we hope to make full use of the knowledge from different inpainting methods. The key issue lies in the fact that the parameters of the whole network θ are trained jointly on the same dataset. The fine-grained mechanism to control which parts of the network learn from which information, *i.e.* the knowledge of which inpainting method, is absent.

4) Conventional Pattern + Attentive Pattern. To address the issue and make full use of the knowledge from diverse inpainting methods, inspired by the meta-learning mechanism, we propose



(a) Preliminary End-to-End Learned Model

(b) Conventional Pattern

(c) Conventional Pattern + Learned Pattern

(d) Conventional Pattern + Proposed Attentive Pattern

Figure 5: The basic ideas with related formulations of image inpainting detection. (a) Preliminary end-to-end learned method. (b) Some conventional hand-crafted filters are applied before using a deep network to perform the inpainting detection. (c) The hand-crafted and learned filters are applied before using a deep network to perform the inpainting detection. (d) The proposed attentive pattern augments the modeling capacity of the inpainting network. Red and blue colors denote that, the components are trained on the respective specific datasets.

an attentive pattern extraction paradigm. The proposed attentive pattern naturally disentangles the training of pattern extraction and inpainting detection mapping, which enables more fine-grained control of the information flow during the training phase. Specifically, we consider that our network parameterized by θ consists of three parts: conventional pattern extractor F_C parameterized by θ_C, attentive pattern extractor F_A parameterized by θ_A, and detection module F_D parameterized by θ_D. We have:

$$f_C = F_C(I_{\text{in}}|\theta_C), \tag{8}$$
$$f_A = F_A(I_{\text{in}}|\theta_A),$$
$$M_{\text{out}} = F_D([f_C, f_A]\,|\theta_D), \tag{9}$$
$$\theta = \{\theta_C, \theta_A, \theta_D\}.$$

Instead of training the whole θ in a uniform way, we split the parameters into two sets as follows,

$$\theta_1 = \{\theta_{\text{C}}, \theta_{\text{D}}\}, \tag{10}$$
$$\theta_2 = \{\theta_{\text{A}}\},$$

Here, we assume that two datasets D_d and D_t can be used for training. D_d is generated by diverse inpainting methods while D_t is created by the typical, well-transferred gated convolution inpainting method [41]. The former provides rich knowledge about the inpainting traces provided by different methods. The latter provides empirically good material to train the whole inpainting method for detecting the results generated by different methods.

In previous methods, the optimization of θ is solved by Empirical Risk Minimization (ERM) over the whole training data D_g. Taking the case in Fig. 5-(c) as an example, the formulation is given as follows,

$$\arg\min_{\theta} \mathbb{E}_{(I_{\text{in}}, M_{\text{gt}})\sim D_g} \mathcal{L}\left(M_{\text{out}}, M_{\text{gt}}|\theta\right), \tag{11}$$

where $\mathcal{L}(\cdot|\theta)$ is the loss function that measures the distance between the input pair. We also list the involved parameter θ here for more clarity of presentation.

Comparatively, we hope to train different components of our network with D_d and D_t as shown in Fig. 5-(d) as follows,

$$\theta_1 = \arg\min_{\theta_1} \mathbb{E}_{(I_{\text{in}}, M_{\text{gt}})\sim D_t} \mathcal{L}\left(M_{\text{out}}, M_{\text{gt}}|\theta_2^*, \theta_1\right),$$
$$\text{s.t. } \theta_2^* = \arg\min_{\theta_2^*} \mathbb{E}_{(I_{\text{in}}, M_{\text{gt}})\sim D_d} \mathcal{L}\left(M_{\text{out}}, M_{\text{gt}}|\theta_2, \theta_1\right), \tag{12}$$

Eq. (12) illustrates a bi-level optimization problem. It is difficult to directly solve the optimization based on the gradient as the related gradient calculation is implicit and complicated.

To solve the bi-level optimization problem, following [4], we design a simple approximation solution. Instead of using the accurate global optimal θ_2^* to optimize θ_1, we adopt an approximated $\hat{\theta}_2$, which is a local optimal of θ_2 achieved by several back-propagation iterations. Then, the gradient used to update θ_1 can be replaced with an approximate one as follows,

$$\nabla_{\theta_1} \mathcal{L}\left(M_{\text{out}}, M_{\text{gt}}|\theta_2^*, \theta_1\right)$$
$$\approx \nabla_{\theta_1} \mathcal{L}\left(M_{\text{out}}, M_{\text{gt}}|\hat{\theta}_2, \theta_1\right). \tag{13}$$

This approximation helps us gets rid of the complex high-order calculation of the bi-level problem.

To summarize, we regard the optimization of our network as a bi-level optimization problem in (12), where the optimization of θ_1 on D_t given the optimal θ_2 constructs the outer layer optimization, and the optimization of θ_2 on D_d given the optimal θ_1 constructs the inner layer optimization. We use an approximation in (13) to solve the problem. A more detailed training algorithm description is provided in the supplementary material.

3.4 Network Implementation

For the CNN models, we directly adopt the network searched by NAS presented in [36]. The encoder shown in Fig. 2 includes several gradually down-sampled cascaded convolutional layers. On the decoder side, the features are first enriched by local and global attention. Then, several gradually up-sampled cascaded convolutional layers are adopted to project the feature into the predicted detection mask. A detailed network architecture about the encoder, decoder and loss functions will be presented in the supplementary material.

4 EXPERIMENTS

4.1 Implementation Details

We set the batch size to 4 and allow 18,000 batches per epoch. The Adam optimizer [15] is adopted in our training. We adopt two commonly used measures, Area Under the receiver operating characteristic Curve (AUC) and F-measure score. The learning rates

Table 1: F-measure results of different methods. The best and second-best results are denoted in red and blue.

Method	GC	CA	SH	EC	LB	RN	TE	NS	LR	PM	SG	CTSDG	MEDFE	Aver.
ManTraNet (Pretrained)	14.06	29.54	72.35	68.95	60.14	34.88	81.27	90.17	81.26	62.78	66.05	57.29	34.04	57.90
ManTraNet (Re-trained)	42.88	26.38	53.40	35.46	48.06	41.40	87.40	85.86	48.70	20.72	36.65	17.65	14.00	42.97
HP-FCN (JPG75)	24.82	23.77	24.94	0.28	29.98	3.87	4.14	21.60	3.19	1.76	8.97	2.70	23.90	13.38
HP-FCN (JPG96)	6.87	0.27	0.44	5.48	29.71	5.88	9.63	0.51	1.19	6.11	7.49	2.89	1.10	5.96
HP-FCN (Re-trained)	81.01	73.12	94.81	75.91	92.09	86.50	82.52	82.68	92.13	46.65	95.11	73.62	59.56	79.67
IID-Net	83.40	74.63	83.99	61.09	38.48	91.27	72.28	71.87	71.51	31.63	94.59	56.29	38.47	66.88
Proposed	78.66	76.92	93.13	79.12	95.55	90.47	81.65	82.12	85.99	61.62	96.16	77.53	77.21	82.46

Table 2: AUC results of different methods. The best and second-best results are denoted in red and blue.

Method	GC	CA	SH	EC	LB	RN	TE	NS	LR	PM	SG	CTSDG	MEDFE	Aver.
ManTraNet (Pretrained)	73.49	82.12	93.57	89.88	92.57	84.94	97.32	99.10	98.82	96.66	96.19	92.26	83.36	90.79
ManTraNet (Re-trained)	81.02	76.20	89.69	78.65	88.49	85.86	98.94	98.27	92.09	76.53	87.10	67.17	62.95	83.30
HP-FCN (JPG75)	50.46	50.55	50.23	49.76	51.60	49.90	50.06	51.79	50.70	49.58	53.33	49.88	51.59	50.73
HP-FCN (JPG96)	49.99	49.96	49.97	49.94	50.58	49.02	49.94	50.05	50.11	49.01	47.93	49.68	49.31	49.65
HP-FCN (Re-trained)	87.80	81.32	95.84	84.26	93.22	89.63	90.25	89.86	95.44	67.56	98.23	81.59	70.96	86.61
IID-Net	96.69	92.51	98.42	88.96	98.40	99.48	93.80	94.28	99.12	92.31	99.94	89.96	84.71	94.51
Proposed	89.87	90.70	99.45	91.69	99.75	99.05	96.71	97.04	99.60	98.12	99.95	90.05	88.70	95.44

Table 3: Ablation studies using f-measure as the measure. The best and second-best results are denoted in red and blue.

Method	GC	CA	SH	EC	LB	RN	TE	NS	LR	PM	SG	CTSDG	MEDFE	Aver.
Ours (GC)	83.40	74.63	83.99	61.09	38.48	91.27	72.28	71.87	71.51	31.63	94.59	56.29	38.47	66.88
Ours (Group)	75.81	73.14	91.14	75.01	92.28	83.10	81.77	82.45	83.99	45.38	93.10	72.62	70.69	78.50
Ours (Att)	79.31	77.87	92.85	78.86	94.64	88.57	82.11	82.49	87.99	44.15	95.57	77.36	75.70	81.34
Ours (Att+Aug)	78.66	76.92	93.13	79.12	95.55	90.47	81.65	82.12	85.99	61.62	96.16	77.53	77.21	82.46

Table 4: Ablation studies using AUC as the measure. The best and second-best results are denoted in red and blue.

Method	GC	CA	SH	EC	LB	RN	TE	NS	LR	PM	SG	CTSDG	MEDFE	Aver.
Ours (GC)	96.69	92.51	98.42	88.96	98.40	99.48	93.80	94.28	99.12	92.31	99.94	89.96	84.71	94.51
Ours (Group)	90.02	88.29	99.12	90.61	99.40	97.83	97.75	98.19	99.47	95.24	99.91	87.90	88.02	94.75
Ours (Att)	90.59	90.37	99.40	90.39	99.70	98.95	96.82	97.16	99.70	96.71	99.93	88.83	87.13	95.05
Ours (Att+Aug)	89.87	90.70	99.45	91.69	99.75	99.05	96.71	97.04	99.60	98.12	99.95	90.05	88.70	95.44

for learning the attentive patterns and other modules of the network are set to $1e-4$. The learning rate will be multiplied by 0.5 if the loss does not decrease for 10 epochs until the convergence. All images are cropped into 256×256 patches in the training stage.

Three state-of-the-art inpainting detection methods are compared in our comparison to detect the inpainted results generated by a series of inpainting methods. The involved detection methods include ManTraNet [39], HP-FCN [17] and IID-Net [36]. The involved inpainting methods include GC [41], CA [42], SH [40], EC [29], LB [37], RN [44], TE [32], NS [3], LR [7], PM [9], SG [11], CTSDG [8], and MEDFE [10]. IID-Net[1] is trained on the training data synthesized with Places [47] by GC, as recommended in [36]. The attentive pattern extractor of the proposed method is trained on the data synthesized with Places [47] by GC, TE, and NS, while other components are trained on the one synthesized only by GC. For ManTraNet[2] and HP-FCN[3], we compare with their pretrained versions and the retrained versions on the training data synthesized by GC, TE, and NS, for a fair comparison. For the testing set, the inpainting mask is generated following [26]. For the training set, the half inpainting mask is generated following [26] while others are generated by random lines, circles, and ellipses, to create a gap with the testing set.

4.2 Quantitative Evaluation

We first compare the F-measure and AUC scores of different methods in Tables 1 and 2. It shows that, our method surpasses previous methods by large margins, which is more significant in F-measure. Our method also achieves superior AUC results on MEDFE, PM, and EC, whose corresponding methods do not get involved in the creation of the training set, which further shows our model's excellent generality.

4.3 Qualitative Evaluation

We also compare the visual results of different methods in Fig. 6. The results show that our proposed method can obtain the most accurate results. ManTraNet provides globally scattered predictions. IID-Net tends to achieve locally over or under-detection results. Comparatively, our method obtains the most reasonable visual results.

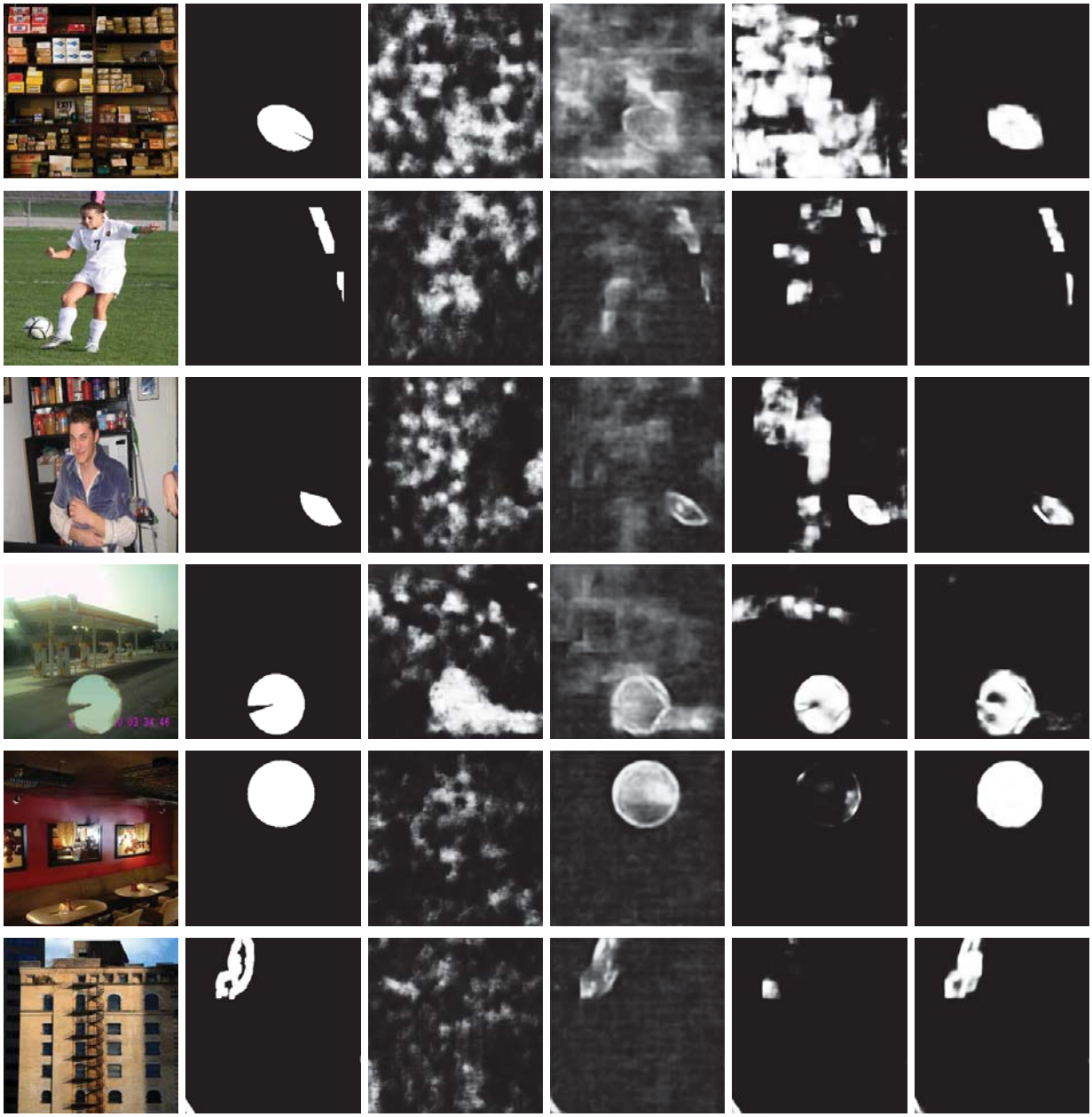
4.4 Ablation Study

We also conduct the ablation studies in Tables 4 and 3. Ours (GC) denotes our baseline, trained on GC generated dataset. Ours (Group) denotes our baseline, trained on the dataset generated by GC, TE, and NS. Ours (Att) denotes our network added with attentive pattern learning. Ours (Att+Aug) denotes our network added with attentive pattern learning, trained with our augmentation method. The results clearly show that our designed components, including attentive pattern learning and near original image data augmentation, leads to higher performance in AUC and F-measure.

5 CONCLUSION

In our work, we propose a novel data augmentation and learned feature extraction method for image inpainting detection. A near original image augmentation is proposed to push the inpainted

[1]https://github.com/HighwayWu/InpaintingForensics
[2]https://github.com/ISICV/ManTraNet
[3]https://github.com/lihaod/Deep_inpainting_localization



(a) Input (b) GT (c) ManTraNet (d) HP-FCN (e) IID-Net (f) Ours

Figure 6: Visual results for inpainting detection. The first two rows' inputs are made by EC. The third and forth rows' inputs are made by PM. The last two rows' inputs are made by MEDFE.

images closer to the original images (without distortion and inpainting) as the input images to generate harder samples in the training data for higher performance. Besides, the attentive pattern is designed to make full use of the knowledge of different inpainting methods during the training phase to obtain a more generalized detection capacity. The experimental results demonstrate the superiority of the proposed method and the effectiveness of each component.

ACKNOWLEDGMENTS

Corresponding author: Rizhao Cai. This work was done at Rapid-Rich Object Search (ROSE) Lab, Nanyang Technological University. This research is supported by the NTU-PKU Joint Research Institute (a collaboration between the Nanyang Technological University and Peking University that is sponsored by a donation from the Ng Teng Fong Charitable Foundation). Wenhan Yang's research is supported by Wallenberg-NTU Presidential Postdoctoral Fellowship.

REFERENCES

[1] Irene Amerini, Tiberio Uricchio, Lamberto Ballan, and Roberto Caldelli. 2017. Localization of JPEG Double Compression Through Multi-domain Convolutional Neural Networks. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 1865–1871.

[2] Belhassen Bayar and Matthew C. Stamm. 2018. Constrained Convolutional Neural Networks: A New Approach Towards General Purpose Image Manipulation Detection. *IEEE Trans. on Information Forensics and Security* 13, 11 (2018), 2691–2706.

[3] M. Bertalmio, A.L. Bertozzi, and G. Sapiro. 2001. Navier-stokes, fluid dynamics, and image and video inpainting. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, Vol. 1. I–I.

[4] Rizhao Cai, Zhi Li, Renjie Wan, Haoliang Li, Yongjian Hu, and Alex C. Kot. 2022. Learning Meta Pattern for Face Anti-Spoofing. *IEEE Trans. on Information Forensics and Security* 17 (2022), 1201–1213.

[5] Chenjie Cao and Yanwei Fu. 2021. Learning a Sketch Tensor Space for Image Inpainting of Man-made Scenes. In *Proc. IEEE Int'l Conf. Computer Vision*. 14489–14498.

[6] Yu Fan, Philippe Carré, and Christine Fernandez-Maloigne. 2015. Image splicing detection with local illumination estimation. In *Proc. IEEE Int'l Conf. Image Processing*. 2940–2944.

[7] Qiang Guo, Shanshan Gao, Xiaofeng Zhang, Yilong Yin, and Caiming Zhang. 2018. Patch-Based Image Inpainting via Two-Stage Low Rank Approximation. *IEEE Transactions on Visualization and Computer Graphics* 24, 6 (2018), 2023–2036.

[8] Xiefan Guo, Hongyu Yang, and Di Huang. 2021. Image Inpainting via Conditional Texture and Structure Dual Generation. In *Proc. IEEE Int'l Conf. Computer Vision*. 14134–14143.

[9] Jan Herling and Wolfgang Broll. 2014. High-Quality Real-Time Video Inpainting-with PixMix. *IEEE Trans. on Visualization and Computer Graphics* 20, 6 (2014), 866–879.

[10] Yibing Song Wei Huang Hongyu Liu, Bin Jiang and Chao Yang. 2020. Rethinking Image Inpainting via a Mutual Encoder-Decoder with Feature Equalizations. In *Proc. IEEE European Conf. Computer Vision*.

[11] Jia-Bin Huang, Sing Bing Kang, Narendra Ahuja, and Johannes Kopf. 2014. Image Completion Using Planar Structure Guidance. *ACM Trans. Graphics* 33, 4, Article 129 (jul 2014), 10 pages.

[12] Minyoung Huh, Andrew Liu, Andrew Owens, and Alexei A. Efros. 2018. Fighting Fake News: Image Splice Detection via Learned Self-Consistency. In *Proc. IEEE European Conf. Computer Vision*.

[13] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. 2017. Globally and Locally Consistent Image Completion. *ACM Trans. Graphics* 36, 4, Article 107 (jul 2017), 14 pages.

[14] Yong Shi Jie Yang, Zhiquan Qi. 2020. Learning to Incorporate Structure Knowledge for Image Inpainting. In *Proc. AAAI Conf. on Artificial Intelligence*, Vol. 34. 12605–12612.

[15] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proc. Int'l Conf. Learning Representations*.

[16] Ang Li, Qiuhong Ke, Xingjun Ma, Haiqin Weng, Zhiyuan Zong, Feng Xue, and Rui Zhang. 2021. Noise Doesn't Lie: Towards Universal Detection of Deep Inpainting. In *IJCAI*.

[17] Haodong Li and Jiwu Huang. 2019. Localization of Deep Inpainting Using High-Pass Fully Convolutional Network. In *Proc. IEEE Int'l Conf. Computer Vision*. 8300–8309.

[18] Haodong Li and Jiwu Huang. 2019. Localization of Deep Inpainting Using High-Pass Fully Convolutional Network. In *Proc. IEEE Int'l Conf. Computer Vision*. 8300–8309.

[19] Haodong Li, Weiqi Luo, and Jiwu Huang. 2017. Localization of Diffusion-Based Inpainting in Digital Images. *IEEE Trans. on Information Forensics and Security* 12, 12 (2017), 3050–3064.

[20] Jingyuan Li, Ning Wang, Lefei Zhang, Bo Du, and Dacheng Tao. 2020. Recurrent Feature Reasoning for Image Inpainting. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*.

[21] Yuanman Li and Jiantao Zhou. 2019. Fast and Effective Image Copy-Move Forgery Detection via Hierarchical Feature Point Matching. *IEEE Trans. on Information Forensics and Security* 14, 5 (2019), 1307–1322.

[22] Zaoshan Liang, Gaobo Yang, Xiangling Ding, and Leida Li. 2015. An Efficient Forgery Detection Algorithm for Object Removal by Exemplar-Based Image Inpainting. *J. Vis. Comun. Image Represent.* 30, C (jul 2015), 75–85.

[23] Liang Liao, Jing Xiao, Zheng Wang, Chia-Wen Lin, and Shin'ichi Satoh. 2020. Guidance and Evaluation: Semantic-Aware Image Inpainting for Mixed Scenes. In *Proc. IEEE European Conf. Computer Vision*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). 683–700.

[24] Liang Liao, Jing Xiao, Zheng Wang, Chia-Wen Lin, and Shin'ichi Satoh. 2021. Image Inpainting Guided by Coherence Priors of Semantics and Textures. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*. 6535–6544.

[25] Guo-Shiang Lin, Min-Kuan Chang, and You-Lin Chen. 2011. A Passive-Blind Forgery Detection Scheme Based on Content-Adaptive Quantization Table Estimation. *IEEE Trans. on Circuits and Systems for Video Technology* 21, 4 (2011), 421–434.

[26] Guilin Liu, Fitsum A. Reda, Kevin J. Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. 2018. Image Inpainting for Irregular Holes Using Partial Convolutions. In *Proc. IEEE European Conf. Computer Vision*.

[27] Siwei Lyu, Xunyu Pan, and Xing Zhang. 2014. Exposing Region Splicing Forgeries with Blind Local Noise Estimation. *Int'l Journal of Computer Vision* 110, 2 (nov 2014), 202–221.

[28] Shant Navasardyan and Marianna Ohanyan. 2020. Image Inpainting with Onion Convolutions. In *Proc. IEEE Asia Conf. Computer Vision*.

[29] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Qureshi, and Mehran Ebrahimi. 2019. EdgeConnect: Structure Guided Image Inpainting using Edge Prediction. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*.

[30] Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. 2016. Context Encoders: Feature Learning by Inpainting. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*. 2536–2544.

[31] Jialun Peng, Dong Liu, Songcen Xu, and Houqiang Li. 2021. Generating Diverse Structure for Image Inpainting With Hierarchical VQ-VAE. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*. 10775–10784.

[32] Alexandru Telea. 2004. An Image Inpainting Technique Based on the Fast Marching Method. *J. Graphics, GPU, & Game Tools* 9, 1 (2004), 23–34.

[33] Ning Wang, Jingyuan Li, Lefei Zhang, and Bo Du. 2019. MUSICAL: Multi-Scale Image Contextual Attention Learning for Inpainting. 3748–3754.

[34] Tengfei Wang, Hao Ouyang, and Qifeng Chen. 2021. Image Inpainting with External-internal Learning and Monochromic Bottleneck. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*. 5120–5129.

[35] Wentao Wang, Jianfu Zhang, Li Niu, Haoyu Ling, Xue Yang, and Liqing Zhang. 2021. Parallel Multi-Resolution Fusion Network for Image Inpainting. In *Proc. IEEE Int'l Conf. Computer Vision*. 14539–14548.

[36] Haiwei Wu and Jiantao Zhou. 2021. IID-Net: Image Inpainting Detection Network via Neural Architecture Search and Attention. *IEEE Trans. on Circuits and Systems for Video Technology* (2021), 1–1.

[37] Haiwei Wu, Jiantao Zhou, and Yuanman Li. 2021. Deep Generative Model for Image Inpainting with Local Binary Pattern Learning and Spatial Attention. *IEEE Trans. on Multimedia* (2021), 1–1. https://doi.org/10.1109/TMM.2021.3111491

[38] Qiong Wu, Shao-Jie Sun, Wei Zhu, Guo-Hui Li, and Dan Tu. 2008. Detection of digital doctoring in exemplar-based inpainted images. In *International Conference on Machine Learning and Cybernetics*, Vol. 3. 1222–1226.

[39] Yue Wu, Wael AbdAlmageed, and Premkumar Natarajan. 2019. ManTra-Net: Manipulation Tracing Network for Detection and Localization of Image Forgeries With Anomalous Features. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*. 9535–9544.

[40] Zhaoyi Yan, Xiaoming Li, Mu Li, Wangmeng Zuo, and Shiguang Shan. 2018. Shift-Net: Image Inpainting via Deep Feature Rearrangement. In *Proc. IEEE European Conf. Computer Vision*, Vol. 11218. 3–19.

[41] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas Huang. 2019. Free-Form Image Inpainting With Gated Convolution. In *Proc. IEEE Int'l Conf. Computer Vision*. 4470–4479.

[42] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. 2018. Generative Image Inpainting with Contextual Attention. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*. 5505–5514.

[43] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. 2019. Free-Form Image Inpainting With Gated Convolution. In *Proc. IEEE Int'l Conf. Computer Vision*.

[44] Tao Yu, Zongyu Guo, Xin Jin, Shilin Wu, Zhibo Chen, Weiping Li, Zhizheng Zhang, and Sen Liu. 2020. Region Normalization for Image Inpainting.. In *Proc. AAAI Conf. on Artificial Intelligence*. 12733–12740.

[45] 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 *Proc. IEEE Int'l Conf. Computer Vision*. 14094–14103.

[46] Yu Zeng, Zhe Lin, Jimei Yang, Jianming Zhang, Eli Shechtman, and Huchuan Lu. 2020. High-Resolution Image Inpainting with Iterative Confidence Feedback and Guided Upsampling. In *Proc. IEEE European Conf. Computer Vision*.

[47] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2018. Places: A 10 Million Image Database for Scene Recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 40, 6 (2018), 1452–1464.

[48] Yuqian Zhou, Connelly Barnes, Eli Shechtman, and Sohrab Amirghodsi. 2021. TransFill: Reference-guided Image Inpainting by Merging Multiple Color and Spatial Transformations. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*. 2266–2276.

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