INFOPRINT: INFORMATION THEORETIC DIGITAL IMAGE FORENSICS

Aurobrata Ghosh[⋆], *Zheng Zhong*[⋆], *Steve Cruz*[†], *Subbu Veeravasarapu*[⋆], *Maneesh Singh*[⋆], *Terrance E Boult*[†]

[⋆] Verisk AI, Verisk Analytics [†] University of Colorado, Colorado Springs

ABSTRACT

Tampered images pose a serious predicament since digitized media is a ubiquitous part of our lives. These are facilitated by the availability of image editing software and recent advances in deep Generative Adversarial Networks (GANs). We propose an innovative method to formulate the problem of localizing manipulated regions in fake images as a deep representation learning problem using the Information Bottleneck (IB) principle. We devise a convolutional neural net-based architecture, InfoPrint (IP), that uses variational inference to approximate the IB formulation. Testing on three standard datasets, we demonstrate that InfoPrint outperforms the state-of-the-art by 3% points or more. Additionally, we demonstrate that it has the ability to to detect alterations made by inpainting GANs.

***Index Terms*—** Digital image forensics, Information Bottleneck, deep representation learning, variational inference.

1. INTRODUCTION

With our increased reliance on digital images and videos as trustworthy sources of information, the ability to photo-realistically alter their contents is a grave danger. Here, we focus on localizing an important class of manipulations that introduce foreign material into a target image, e.g., splicing, where part(s) of other image(s) are inserted, or inpainting, where part(s) are hallucinated by specialized algorithms.

Unlike in most computer-vision problems, non-semantic pixel-level statistics have proven to be more successful for solving forensic tasks than semantic information since attackers use semantics to hide their modifications. Low-level statistics contain camera model-specific distortions and noise patterns, which differ between the tampered regions and the host image because they originate from different sources.

Over the years, numerous hand engineered statistical features, including sensor noise, demosaicing traces, and compression artefacts, have been explored [1, 2, 3]. Others have improved over these by modelling non process-specific *noise-residuals* using high-pass filters like Wavelet transforms and spatial Rich Filters (RFs) [4, 5]. Recent advances come from Convolutional Neural Networks (CNNs) and learned RFs with constrained convolutions [6]. However, such high-pass filters also capture semantics – like edges – therefore, forensic algorithms may not function optimally [7]. SpliceRadar (SR) [7], attempted to suppress these semantic edges by regularizing the Mutual Information (MI) between the input and the features but used a numerically unstable method. Here, we tackle this issue with an information theoretic framework.

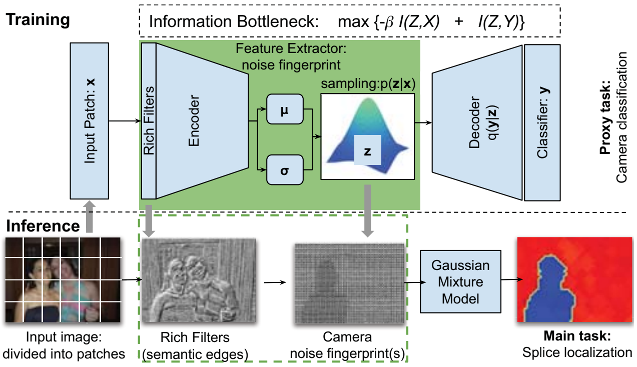
Training
Information Bottleneck: max {-β I(Z,X) + I(Z,Y)}
Input Patch: x
Rich Filters
Encoder
Feature Extractor: noise fingerprint
μ
σ
sampling:p(z|x)
z
Decoder q(y|z)
Classifier: y
Proxy task: Camera classification
Inference
Input image: divided into patches
Rich Filters (semantic edges)
Camera noise fingerprint(s)
Gaussian Mixture Model
Main task: Splice localization

Fig. 1. InfoPrint leverages the Information Bottleneck (IB) to improve forgery-localization in digital images. The encoder (with stochastic latent-space) is trained to learn low-level noise-residual features (fingerprints) on a proxy task of camera model classification. At inference, a Gaussian mixture model segregates these noise fingerprints to identify the genuine patches from the forged ones. Rich filters, which have been used previously, are inaccurate since they are contaminated by semantic edges of the input image. IB improves over rich filters by suppressing the semantic contents and amplifying the *true* camera noise fingerprint(s) in an image.

Information theory is a powerful framework that is increasingly adopted to improve various aspects of deep machine learning, e.g., representation learning, generalizability, and regularization [8], and for interpreting how deep neural networks function [9]. Here, we consider the Information Bottleneck (IB) [10], a framework for learning *compressed representations* that allows us to control the information flow between the input and the representation layer.

In this work, we propose a novel IB-based CNN architecture for localizing splicing/inpainting image forgeries. Our solution overcomes the limitations of state-of-the-art approaches by learning to extract low-level camera model noise-residuals uncontaminated by semantic edges. The main contributions of this work are: *(i)* we cast the problem

978-1-7281-6395-6/20/$31.00 ©2020 IEEE

of modelling distinguishing camera model fingerprints as a data-driven representation learning problem based on IB; *(ii)* our application of IB is unique in the sense that it is the complete reverse of why it has been typically applied [11, 12] – instead of garnering semantic contents, the information compression helps ignore the semantics and focuses on low-level noise-residuals useful for segregating camera models; and *(iii)* we demonstrate our method's ability to detect signatures of deep generative models by pitting it against three recent inpainting GANs. We call our proposed method InfoPrint (IP) because the learned noise-residual representation is like a camera model's fingerprint.

2. INFORMATION BOTTLENECK

Learning a predictive model $p(\mathbf{y}|\mathbf{x})$ is hampered when a model overfits nuisance detractors in the input data X, irrelevant for a task Y. This is crucial in deep learning when the input is high dimensional (e.g., an image), the task is a simple low dimensional class label, and the model is a flexible neural network. The goal of IB is to overcome this problem by learning a compressed representation Z, of X, which is optimal for the task Y. The IB Lagrangian [10], based on the MI values $I(Z,X)$, $I(Z,Y)$, is $\mathcal{L} = I(Z,Y) - \beta \cdot I(Z,X)$. Intuitively IB extracts the relevant information that X contains about Y and discards non-informative signals. β regulates the information flow: a larger β results in a greater constriction of information from X to Z.

However, MI is hard to compute with high dimensional variables. A variational approximation to IB, applicable to neural networks, is proposed in [11], where the IB Lagrangian is bounded below by the approximate objective function:

$$J_{IB}(p,q) = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\mathbf{z}\sim p(\mathbf{z}|\mathbf{x}_i)}\left[-\log q(\mathbf{y}_i|\mathbf{z})\right] \\ + \beta\,\mathrm{KL}[p(\mathbf{z}|\mathbf{x}_i)||r(\mathbf{z})] \geq -\mathcal{L}, \quad (1)$$

where Z is a stochastic latent layer, $p(\mathbf{z}|\mathbf{x})$ is an encoder network, $q(\mathbf{y}|\mathbf{z})$ is a decoder network that approximates the intractable $p(\mathbf{y}|\mathbf{z})$, and $r(\mathbf{z})$ is a prior distribution that replaces the unknown $p(\mathbf{z})$. The first term on the r.h.s. is the average cross-entropy loss (with stochastic sampling over $\mathbf{z}$), while the second term is a regularization. Eq 1 can be minimized using the reparameterization trick [13].

According to the rate-distortion interpretation of IB [12], the loss term is denoted as distortion D that approximates $-I(Z,Y)$. The unweighted regularization term is denoted as rate R, which approximates $I(Z,X)$ and measures the number of bits required to encode the representations. The RD plane allows to visualize solutions to the IB Lagrangian for different values of β. Inspecting this RD *curve* helps in selecting β to trade off between the distortion, which affects task accuracy, and the rate, which affects compression and hence the generalization capacity. This results in a principled information theoretic regularization, which can be measured in quantities of *information* (in units of bits).

3. INFOPRINT ARCHITECTURE

Our goal is to localize a forgery where foreign material has been inserted into a host image to alter its contents. Since semantic structures can be misleading, our strategy is to extract low-level noise fingerprints of an image and then pinpoint inconsistencies in these (Fig. 1), where ideal noise fingerprints are high frequency contents uncorrelated to the semantic structures in an image.

The optimal network, therefore, needs to learn representations of noise-residuals that (a) are not contaminated by semantic information and (b) can distinguish camera models. To achieve this, first we design an architecture that extracts relevant high frequency contents and then uses IB to squeeze out semantic correlations with the input. Second, we select a proxy training task of classifying the source camera model. As the semantics cannot be related to the camera model it forces the network to focus on non-semantic noise statistics. Additionally, we train our network with a large number of camera models to improve its ability to segregate even unseen devices. This allows our network to generalize effectively in a blind test setting with images acquired on unknown cameras.

Architecture We consider a deep encoder-decoder architecture with a stochastic representation layer that takes in an RGB image-patch X, computes features Z of the patch's noise-residuals, and from these classifies the source camera model Y during training.

In the first layer, we include constrained convolutions of the form [7]: $\mathcal{R}^{(k)} = \mathbf{w}_k(0,0) + \sum_{i,j\neq 0,0}\mathbf{w}_k(i,j) = 0$, which computes noise-residuals: the mismatch between a pixel's true value and its interpolated value from its $S \times S$ neighbours. These can be trained end-to-end by including the penalty $\mathcal{R}_{RF} = (\sum_k(\mathcal{R}^{(k)})^2)^{\frac{1}{2}}$ in the cost function.

For our encoder $p(\mathbf{z}|\mathbf{x})$, we consider a CNN architecture inspired by ResNet-18v1 [14], where we discard operations that quickly shrink the input and encourage learning high-level (semantic) features. Namely, we discard the initial max-pooling layer, all convolutions with stride greater than one, and the final global pooling layer. We found these to be detrimental to our task. Instead, we include additional 7×7 and 5×5 convolutions to reduce the input patch to a single *feature-pixel* with a large bank of filters, thus avoiding fully connected layers. The final architecture is a 27-layers deep CNN (Table 1). Every convolution is followed by batch normalization and ReLU activation. To get a stochastic encoding, we split the CNN's output vector of 72 filters into $\boldsymbol{\mu}_{\mathbf{x}}$ and $\boldsymbol{\sigma}_{\mathbf{x}}$ and model $p(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}_{\mathbf{x}}, diag(\boldsymbol{\sigma}_{\mathbf{x}}))$ [13].

We adopt an extremely simple decoder $q(\mathbf{y}|\mathbf{z})$ to deter our model from degenerating to the auto-decoder limit [12], an issue we also observed. We select a logistic regression model: a

constrained conv.	residual conv.	encoding conv.
$3 \times 3, 64$	$\underbrace{7 \times 7, 64, \underbrace{\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix}}_{\text{res.block}}, 5 \times 5, 64, \underbrace{\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix}}_{\text{res.block}}}_{\times 4}, 7 \times 7, 64$	$1 \times 1, (\underbrace{36 + 36}_{\boldsymbol{\mu},\, \boldsymbol{\sigma}})$

Table 1. The InfoPrint encoder is a CNN with 27 layers. The input patch is $49 \times 49 \times 3$ and the output encoding is $1 \times 1 \times 72$.

dense (logit generating) layer that is connected to the stochastic code layer Z and is activated by the softmax function.

Such an architecture, however, would suffer from the same limitations as the state-of-the-art approaches, since the constrained convolutions would also capture semantic edges. To squeeze out such semantic correlations, we consider the IB objective function rather than a simple cross-entropy loss for training because Eq. 1 allows us to control the information flow between the input X and the latent code Z. We select the regularization parameter β by plotting the RD curve and finding the right balance between training task accuracy and the amount of information throughput. Excess throughput would result in Z being contaminated by semantic edges, while insufficient throughput would result in training failure.

However, we have two tasks. Our main task is splice localization but we train our model on a proxy task of camera model identification. Therefore, we employ the RD curve of the training task to first narrow the potential range for β before determining the optimal β('s) through empirical testing.

Inference Assuming that the untampered region is the largest part of the image, we simplify the forgery localization problem to a two-class feature segmentation problem. First, we compute our network's representation $(\boldsymbol{\mu}, \boldsymbol{\sigma})$ for all juxtaposed patches in the test image. Then, we segment these 72-dimensional features using a Gaussian mixture model with two components using expectation maximization (EM). The Gaussian distributions are only approximate statistics of the two classes that separate them probabilistically.

(F1)	DSO-1	NC16	NC17-dev1
IP-NoRF	0.50 (0.40)	0.38 (0.27)	0.40 (0.29)
IP-NoIB	0.68 (0.56)	0.40 (0.29)	0.41 (0.31)
IP1e-3	0.71 (0.55)	**0.42** (0.29)	**0.44** (0.31)
IP5e-4	**0.72** (0.58)	0.40 (0.29)	0.42 (0.31)
SR	0.69 (0.59)	0.39 (0.28)	0.40 (0.32)
EX-SC	0.57 (0.49)	0.38 (0.31)	**0.44** (0.37)
SB	0.66 (0.54)	0.37 (0.26)	0.43 (0.36)

Table 2. F1 scores. Black/blue: optimal/Otsu threshold.

(MCC)	DSO-1	NC16	NC17-dev1
IP-NoRF	0.44 (0.32)	0.37 (0.24)	0.33 (0.22)
IP-NoIB	0.64 (0.53)	0.38 (0.27)	0.35 (0.25)
IP1e-3	0.67 (0.53)	**0.40** (0.28)	**0.38** (0.25)
IP5e-4	**0.69** (0.55)	0.38 (0.27)	0.35 (0.24)
SR	0.65 (0.55)	0.37 (0.26)	0.33 (0.25)
EX-SC	0.52 (0.43)	0.36 (0.29)	**0.38** (0.30)
SB	0.61 (0.48)	0.34 (0.25)	0.36 (0.25)

Table 3. MCC scores. Black/blue: optimal/Otsu threshold.

Implementation We consider input patches of size $49 \times 49 \times 3$ (empirically); and $k = 64$ constrained convolutions with $S = 3$ in the first layer. Also, our encoder has a fixed number of 64 filters in every layer. For the prior distribution, we use the factorized standard Gaussian $r(\mathbf{z}) = \prod_i \mathcal{N}_i(0, 1)$ proposed in [11] and train our network using the loss $J = J_{IB} + \lambda \mathcal{R}_{RF}$, where we found $\lambda = 1$ gives the best results.

We train on the Dresden Image Database [15] that contains 17,000+ images from 27 camera models. For each camera model, we randomly select 70%, 20% and 10% of the images for training, validation and testing. We use a mini-batch of 200 patches, and train for 700 epochs with 100,000 randomly selected patches every epoch. We maintain a learning rate of 10^{-4} for 100 epochs, then decay it linearly to 5×10^{-6} in the next 530 epochs and then finally decay it exponentially by a factor 0.9 over the last 70 epochs. This allows us to achieve a camera model prediction accuracy of $\sim 80\%$ on the validation/test sets.

4. EXPERIMENTS & RESULTS

To evaluate our method, we test it on three standard datasets. However, first, we tune our model by plotting the RD curve and selecting the optimal parameter β. Then, we conduct an ablation study to assess the relevance of IB and RFs and lastly, we compare InfoPrint to state-of-the-art algorithms.

The standard datasets we employ are DSO-1 [19], Nimble Challenge 2016 (NC16) and the Nimble Challenge 2017 (NC17-dev1) [20]. DSO-1 consists of 100 spliced images in PNG format. NC16 & NC17 contain 564 & 237 spliced images respectively, mostly in JPEG format. All three datasets are state-of-the-art, contain hard to detect forgeries and are accompanied by the ground truth masks. Additionally, we generate manipulations created by three inpainting GANs, namely Yu et al. [17], Nazeri et al. [18], Liu et al. [16], which represent the state-of-the-art in image inpainting.

We use the F1 score and Matthews Correlation Coefficient (MCC) to score the performances. These are widely used for evaluating splice localization [1]. However, F1 and MCC

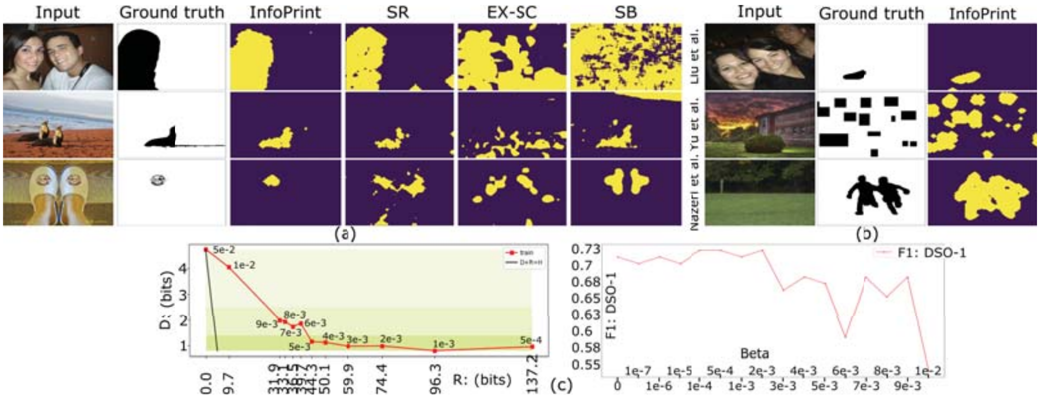

Fig. 2. (a) Qualitative results showing the superiority of InfoPrint over published methods. (b) InfoPrint vs inpainting GANs: Liu et al. [16], Yu et al. [17], Nazeri et al. [18]. (c) β selection. Left: RD curve in red. Right: F1 metric on DSO-1. Low distortion values are observed for $\beta \leq$ 5e-3, while a peak in the F1 is observed from 2e-3 to 1e-4.

require a binarized output mask, while our method predicts probabilities. Although in the forensic literature it is customary to report the scores for optimal thresholds, computed from the ground truth masks, we additionally report scores from automatic thresholding using Otsu's method which performs the best from eight other tested methods of scikit-image.

β selection To select β we plot the RD curve (Fig. 2c). We observe that our model achieves low distortion values for $\beta \leq$ 5e-3 for the training task. To select β for the forensic task, we compute F1 scores on DSO-1 for all values of β and find a peak from 2e-3 to 1e-4 (1e-3 is an anomaly we attribute to stochastic training). Hence, we conduct our experiments for two central values, $\beta =$ 1e-3, 5e-4. Note that zero information throughput is achieved for $\beta =$ 5e-2, hence $\beta >$ 5e-2, e.g. $\beta = 1$ (VAE [13]) cannot perform better.

Ablated models To evaluate the importance of IB and RFs, we consider i) IP-NoIB: an InfoPrint model with β=0, which results in only the cross-entropy loss during training, and ii) IP-NoRF: an InfoPrint model with the RFs replaced by regular convolution layers with identical sizes.

SOTA models We consider three state-of-the-art (SOTA) algorithms to compare against InfoPrint. SpliceBuster (SB) [5], is a top-performer of the NC17 challenge [20] that uses one fixed RF and the co-occurrence of its values to discern forged regions. SR [7] is a recent approach that has been discussed above. EX-SC [21] is a recent deep learning-based algorithm that predicts meta-data self-consistency to localize tampered regions. We also attempt to include ManTraNet [22], however, as acknowledged by the authors, it crashed often with large images. Therefore, we do not include it here. While SB is a model-based approach with few trainable parameters, SR has 1,322,942 trainable weights and EX-SC: 76,088,020 weights. In comparison, InfoPrint has 2,029,584 weights ($\sim$ size of SR and 37x smaller than EX-SC).

Results Quantitative results are presented in Tables 2 & 3. Comparing the ablated models and InfoPrint indicates that dropping RFs has detrimental effects on performance, while dropping IB is comparable to SOTA models. However, RFs with IB has the best performance.

For SOTA, all scores presented in these tables indicate improved results over published methods. The F1 scores indicate improvements up to 3% points over SR, 6% points over SB, and 15% points over EX-SC on DSO-1, with best scores on NC16 & NC17. The MCC scores are again high for InfoPrint in comparison to the other methods, with a margin of up to 4% points on DSO-1 in comparison to SR.

Qualitative results are presented in Fig. 2a and 2b (also supp. materials). Fig. 2a compares InfoPrint's predicted manipulation mask to the ground truth mask and masks predicted by SR, SB, and EX-SC. The examples come from all three test datasets. Fig. 2b demonstrates the ability of InfoPrint to detect the signatures of top-of-the-line inpainting GANs. Unfortunately, no standard datasets exist to report quantitative results. Furthermore, most of the examples are sourced from the internet and have been already processed, e.g. resized or compressed. This destroys camera model traces. However, InfoPrint is still able to localize the manipulations correctly.

Failures (supp. mat.) All methods fail when the input image is small, e.g. 300×300 pixels or contains saturated regions.

5. CONCLUSION

We presented a novel information theoretic formulation to address the issue of localizing tampered regions in digital images. Using IB, we proposed an approach to learn distinguishing low-level statistics uncontaminated by semantic contents and showed that it outperformed published forensic methods. Our IB formulation was also unique because we used it to learn noise-residual patterns and suppress semantics rather than the other way around. Additionally, we demonstrated our method's potential to detect inpainting operations by recent deep generative methods.

6. REFERENCES

[1] M. Zampoglou, S. Papadopoulos, and I. Kompatsiaris, "Large-scale evaluation of splicing localization algorithms for web images," *Multimedia Tools and Applications*, 09 2016.

[2] A. C. Popescu and H. Farid, "Exposing digital forgeries in color filter array interpolated images," *IEEE Transactions on Signal Processing*, vol. 53, no. 10, pp. 3948–3959, 10 2005.

[3] M. Barni, E. Nowroozi, and B. Tondi, "Higher-order, adversary-aware, double JPEG-detection via selected training on attacked samples," in *25th European Signal Processing Conference (EUSIPCO)*, 08 2017, pp. 281 – 285.

[4] B. Mahdian and S. Saic, "Using noise inconsistencies for blind image forensics," *Image and Vision Computing*, vol. 27, no. 10, pp. 1497 – 1503, 2009, Special Section: Computer Vision Methods for Ambient Intelligence.

[5] D. Cozzolino, G. Poggi, and L. Verdoliva, "Splicebuster: A new blind image splicing detector," in *2015 IEEE International Workshop on Information Forensics and Security (WIFS)*, 11 2015, pp. 1–6.

[6] B. Bayar and M. C. Stamm, "Constrained convolutional neural networks: A new approach towards general purpose image manipulation detection," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 11, pp. 2691–2706, Nov 2018.

[7] A. Ghosh, Z. Zhong, T. E. Boult, and M. Singh, "SpliceRadar: A learned method for blind image forensics," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[8] X. B. Peng, A. Kanazawa, S. Toyer, P. Abbeel, and S. Levine, "Variational Discriminator Bottleneck: Improving Imitation Learning, Inverse RL, and GANs by Constraining Information Flow," in *International Conference on Learning Representations (ICLR)*, May 2019.

[9] R. Shwartz-Ziv and N. Tishby, "Opening the Black Box of Deep Neural Networks via Information," *arXiv e-prints*, p. arXiv:1703.00810, Mar 2017.

[10] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," in *Proc. of the 37-th Annual Allerton Conference on Communication, Control and Computing*, 1999, pp. 368–377.

[11] A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, "Deep variational information bottleneck," in *International Conference on Learning Representations (ICLR)*, 04 2017.

[12] A. Alemi, B. Poole, I. Fischer, J. Dillon, R. A. Saurous, and K. Murphy, "Fixing a broken ELBO," in *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 07 2018.

[13] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *2nd International Conference on Learning Representations, (ICLR) 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 06 2016, pp. 770–778.

[15] T. Gloe and R. Böhme, "The 'Dresden Image Database' for benchmarking digital image forensics," in *Proceedings of the 25th Symposium On Applied Computing (ACM SAC)*, 2010, vol. 2, pp. 1585–1591.

[16] G. Liu, F. A. Reda, K. J. Shih, T-C. Wang, A. Tao, and B. Catanzaro, "Image inpainting for irregular holes using partial convolutions," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 85–100.

[17] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 5505–5514.

[18] K. Nazeri, E. Ng, T. Joseph, F. Qureshi, and M. Ebrahimi, "Edgeconnect: Generative image inpainting with adversarial edge learning," *arXiv preprint arXiv:1901.00212*, 2019.

[19] T. J. D. Carvalho, C. Riess, E. Angelopoulou, H. Pedrini, and A. d. R. Rocha, "Exposing digital image forgeries by illumination color classification," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 7, pp. 1182–1194, 07 2013.

[20] J. Fiscus, H. Guan, Y. Lee, A. Yates, A. Delgado, D. Zhou, D. Joy, and A. Pereira, "The 2017 Nimble Challenge Evaluation: Results and Future Directions," 2017.

[21] M. Huh, A. Liu, A. Owens, and A. A. Efros, "Fighting fake news: Image splice detection via learned self-consistency," in *Proceedings of the European Conference on Computer Vision (ECCV)*, Cham, 2018, pp. 106–124, Springer International Publishing.

[22] Y. Wu, W. AbdAlmageed, and P. Natarajan, "ManTra-Net: Manipulation Tracing Network For Detection And Localization of Image ForgeriesWith Anomalous Features," 2019.