Text based Image Inpainting Detection

Bryan Plummer, Divya Spoorthy

Abstract

Image Inpainting is the process where a deteriorated image is modified either using manual or deep learning techniques to form a complete image. While image inpainting has it's advantages various advanced image inpainting tools could be put to malevolent use and could result in disastrous results like fake news reporting, image forgeries for court trials etc. DL based methods of inpainting are extremely difficult to detect and localize. Most inpainting techniques do not make use of the text data associated with the image on various platforms and here we propose a text based image inpainting detection where we leverage the text data for image localization and find the inpainted regions from the localized part of the image.

1. Introduction

Image Inpainting is the process where a deteriorated image is modified either using manual or deep learning techniques to form a complete image. The applications of image inpainting are widespread in the field of computer vision and which help to remove deteriorated object scenes, cut and crop unwanted parts of image and help enhance and beautify the image. But recently such tampered images are mistaken to be real ones and have been increasingly used in fake news generation, criminal fraud and other malevolent purposes. The aim of this paper is to detect and pinpoint the manipulated regions of an image at pixel level which could be caused by both manual and deep learning based image inpainting techniques. This process of localization and pixel level detection of image manipulation has been extensively studied in the past years and is very crucial to study because of the surge in fake news media and social media image forgeries. There are mainly three types of image inpainting techniques : Sequential-based, CNN-based and GAN-based methods. The object removal techniques can be divided into two methods: Inpainting and copy-move methods. There are multiple enhanced Image inpainting techniques which make the detection of forgery and inpainting of images extremely difficult. Current

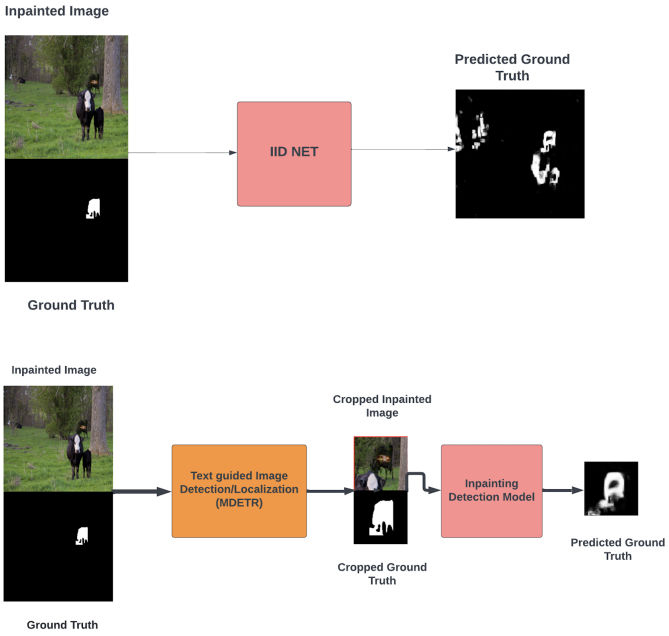

Figure 1. Fig a represents IID net, Fig b represents our new text guided detection model

inpainting methods [17] use the entire image for processing and detecting inpainted regions and usually the inpainted region is very small when compared to the total image size. This could add a lot of noise because the backgroud pixels are large in number compared to the inpainted pixels and the model will have harder time detecting the exact location of manipulation. Our method will focus on the specific regions on the image and will help guide detections which will reduce the noise of all the background pixels.

We often see that images in various contexts whether it be news, social media or any other platform usually comes with a text associated with it. It usually is a caption, header or could be a brief paragraph usually describing or related to the image present nearby. It is more likely that the objects mentioned in the caption are present in the associated image and there is high chance that the particular object

in the caption is manipulated. This will help us ignore the rest of the image thereby decreasing the noise associated. While there are multiple image forgery detection methods available in the machine learning community these days none of these methods exploit this text information which could be very vital in localization of the inpainted part of the images. Usually a very small portion of the image is manipulated and this will hinder the existing models for exact localization. Here we would like to leverage this information and build a model that takes in input as text data as well as image data to get an enhanced performance for image inpainting manipulation detection and localization task.

2. Previous Work

2.1. Image Inpainting

Reconstruction of deteriorated images has been studied for decades and more comprehensive methods have been proposed and developed in recent years for the automation of this process. Traditional Inpainting methods can be classified into two categories: Patch-based and Diffusion based methods. Patch Based methods use the existing undamaged part of the image as search space and try to fill the inpainted part with the best matching patch. Some of the earliest work consists of [2, 4, 7, 11, 16] While in Diffusion based methods we smoothly propagate the image content from the boundary pixels to the interior region of the missing pixel region for the reconstruction of the missing part of the image. Previous work on diffusion based methods include [1, 3, 8, 14]. These methods are robust for simple images but fail to exploit the complex structural information with images that mainly comprises one single object or images with complicated textures, in which case search for one patch becomes more difficult. After the dominance of CNNs in 2012 multiple image inpainting techniques have been proposed using CNNs and encoders which try to capture the complexities exhibited by images [15, 18, 19]. While CNN based methods have proven to show promising results in this challenging task where they can generate visually plausible results, they often create blurry patches and distorted regions which are inconsistent with neighboring areas. This occurs because CNNs are unable to replicate information from distant spatial regions. This is where GANs were proposed which can not only generate novel image structures but also capture the surrounding image regions during training phase and make better inpainting predictions. GANs are getting better and better everyday at inpainting tasks suggested by [6, 20–22, 24] and there is many more GAN based techniques currently being deployed which are not only visually plausible but also much harder to detect. While these GAN based and other inpainting techniques have various advantages in the field of computer vision, they could also be used for malicious purposes which is increasingly raising security concerns. It is extremely common for news media to use the manipulated images and deliver fake news to common public which could result in misleading the viewers. Photo manipulation is also very common in social media which not only mislead the audience but also can cause serious mental and emotional health issues. So there is an urgent need to find Deep Learning based image inpainting detection algorithms against such inpainting forgeries.

2.2. Image Inpainting Detection

There have been various attempts at detection of inpainting in images. Haiwei et al [17] used IID-net which consists of three blocks called the enhancement which consists of hierarchically combined input layers for enhancing the inpainting traces, extraction where they designed an NAS algorithm targeted to extract features for actual inpainting detection tasks and decision block to detect the inpainted regions at pixel accuracy. Xiru et al [5] proposed MVSS-net which trains a Deep Neural Network which is capable of learning semantic-agnostic features that are sensitive to manipulations whilst specific to prevent false alarms. Ang et al [13] made an attempt towards universal detection of deep inpainting which can detect and generalize well different deep inpainting methods. Here they proposed a Noise-Image Cross-fusion Network (NIX-Net) which effectively exploits the discriminative information present in the images and their noise patterns.

Jing et al [9] introduced TransForensics, a novel image forgery localization method inspired by Transformers. They used dense self-attention encoders which model global context and all pairwise interactions between local patches at different scales and dense correction modules for improving the transparency of the hidden layers and correcting the outputs from different branches.

There are other inpainting detection methods which are specific to one form of image manipulation and one of them is Xiuli et al who tackled the problem of image splicing forgery detection and localization. Here they proposed a fake-to-realistic transform generator GT to automatically suppress the tampering artifacts in the forgery images and adversarially trained it against another generator GM which detects the tampered regions and learns from the forgery images retouched by GT.

Then there are state of the art algorithms like SPAN which models the relationship between image patches at multiple scales by constructing a pyramid of local self-attention blocks. Some other methods which have proposed more general solutions that are not specific to manipulation types are RGB-N(noise) Net [10] where they modeled a Faster R-CNN based method to detect tampered

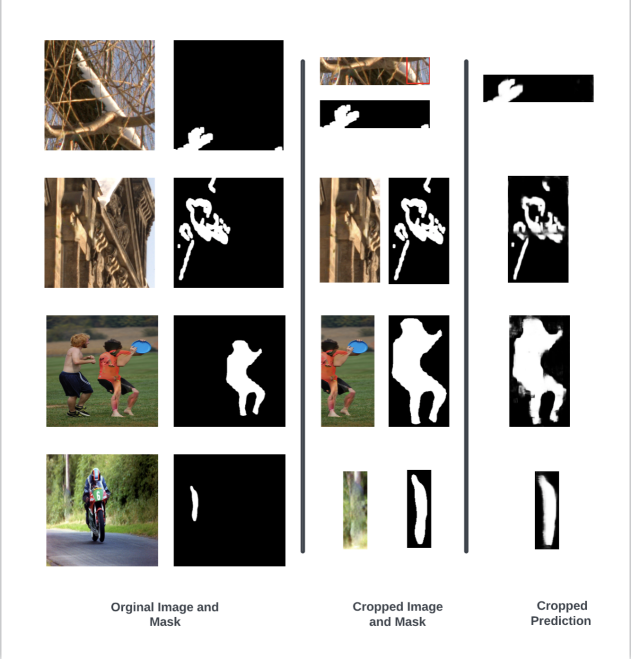

Figure 2. This Figure shows various experiment results for three different kinds of datasets. First two rows correspond yo results of GC dataset, third row corresponds to EC dataset and fourth row corresponds to DFV2 dataset

region whose predictions are limited to rectangular box, and Manipulation Tracing Network(ManTra-NET) [23] which achieves comparable results to RBG-N and makes pixel level predictions.

2.3. Text Guided Detection

In most of the real world applications where images are used to represent a form of data we usually see an associated text along with the posted image. This is true in most cases like news websites, social media applications like Facebook and Instagram etc, where the image usually is associated with some text or caption and is usually related to the image. So, here we focus on building a universal image inpainting detection system which can detect both manual and deep learning based inpainting and forgery which uses the associated text data corresponding to the image. While most of the existing work focuses on just the image data here we try to observe and leverage the correlation between the image data and the associated text data and boost the performance of image inpainting detection models.

3. Text Based Detection

In this section we present the various parts of details of our Text Guided Model for detecting the inpainted manipulations, which are altered using Deep Learning based algorithms. The Schematic Diagram of our model is shown in Fig2. Our approach consists of three parts: 1. Object Detection using the given caption/text summary. 2. Cropping out the objects along the bounding box predictions that we obtain in the first part 3. Pass these cropped images into IID-Net and get the inpainted regions. These steps are discussed in detail in the sections below.

3.1. Text Guided Object Detection

Here we rely on a pre-trained object detector and extract region of interest from the Image. As a part of our future work we are set to use MDETR(Modulated Detection for End-to-End Multi-Modal Understanding) [12] which helps in object detection in an image conditioned on a raw text query (caption or question). Here the authors have used a transformer based architecture which associates the object present in image with the text given. We could get multiple objects detected for each caption or text. There are two typed of multiple object detections 1. Overlapping 2. Non-overlapping. Handling non-overlapping object detections is simple, we could simply use both the detected regions and pass them onto the next stage. This gets tricy when there are overlapping detections. Here we could solve the problem by just taking an union of all the regions and consider that as one detected region. Then we can get one clubbed bounding box for all the proposed regions by the model. This is depicted in Figure 3. Since this is a part of future work the results and experiments are not yet presented here.

3.2. Image and Mask Crop

This is the part where our approach mainly differs from the other inpainting detection approaches. After getting the multiple detections from an image we crop out all the detected regions and pass these on to the next stage instead of the original image. Here the intention is to remove most of the noise that comes from background that comes from the non-inpainted parts of the image. This should considerably increase the inpainting detection accuracy. This helps us prove the point that the text associated with images could in general help us with better inpainting detection.

3.3. Inpainting Detection

After obtaining the cropped regions from the text based object detected images we send these smaller cropped regions into IID-Net which separately for each image gives the output of the inpainted regions. After detecting the inpainted regions we gather all the regions that belong to one image and club them all together based on their bounding box regions into one mask and send is as the final output.

Table 1. QUANTITATIVE COMPARISONS BY USING AUC AS CRITIC

Training Dataset	Test Data Set (AUC)					
	GC	GC Cropped	EC	EC Cropped	DFV2	DFV2 Cropped
GC	99.43	98.34	90.7	89.9	85.28	89.9
EC	84.2	45.3	98.23	85.64	67.96	84.5
DFV2	-	-	-	-	-	-
GC Cropped	-	-	-	-	-	-
EC Cropped	-	-	-	-	-	-
DFV2 Cropped	-	-	-	-	-	-

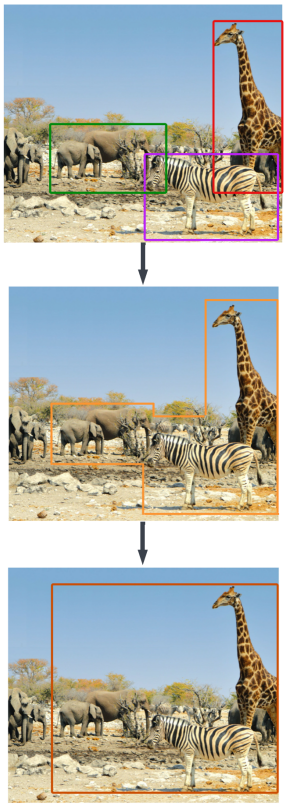

Figure 3. This Figure shows the stages after object detection in case of overlapping object detections. In the top most image we have multiple detections and multiple bounding boxes around the detected inpainted objects. The the middle image we have the union of all the bounding boxes representing one inpainted region. Then a final bounding box which represents all the regions is drawn surrounding the regions and passed on to the next stage

4. Experiments

Here we train two different models. The first one is text guided object detection and we use MDETR for this. After we obtain the cropped images we send the cropped images through IID net and obtain the ground truth masks. We use AUC and F1 score as evaluation criteria. The loss function is similar to the one used in [17] and is defined as follows:

$$\{L_B(M_g, M_o) = -\frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}(M_g(i,j)logM_o(i,j) + ([1 - M_g(i,j)][1 - logM_o(i,j)]))\} \quad (1)$$

Using the ground truth mask M_g and the predicted mask M_o we compute AUCs on various datasets of inapinted images. We train our models on four different datasets and obtained the results presented in table 1. The quantitative results of the inpainting pipeline is presented in Figure3.

5. Conclusion and Future Work

In this paper, we propose a novel DL based inpainting detection model for all kinds of inapinting manipulations which leverages the idea of noise reduction because of presence of excessive background present during inpainting region detection. As future work we aim to impement the stages I and III of the proposed algorithm completely and exhibit the results in this report.

References

[1] C. Ballester, M. Bertalmio, V. Caselles, G. Sapiro, and J. Verdera. Filling-in by joint interpolation of vector fields and gray levels. *IEEE Transactions on Image Processing*, 10(8):1200–1211, 2001. 2

[2] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. PatchMatch: A randomized correspondence algorithm for structural image editing. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 28(3), Aug. 2009. 2

[3] M. Bertalmio, L. Vese, G. Sapiro, and S. Osher. Simultaneous structure and texture image inpainting. *IEEE Transactions on Image Processing*, 12(8):882–889, 2003. 2

[4] A Criminisi, P Perez, and K Toyama. Region filling and object removal by exemplar-based image inpainting. *IEEE Trans. on Image Processing*, 13(9):1200–1212, 2004. 2

[5] Chengbo Dong, Xinru Chen, Ruohan Hu, Juan Cao, and Xirong Li. Mvss-net: Multi-view multi-scale supervised networks for image manipulation detection, 2021. 2

[6] Qiaole Dong, Chenjie Cao, and Yanwei Fu. Incremental transformer structure enhanced image inpainting with masking positional encoding, 2022. 2

[7] Iddo Drori, Daniel Cohen-Or, and Hezy Yeshurun. Fragment-based image completion. *ACM Trans. Graph.*, 22(3):303–312, jul 2003. 2

[8] SELIM ESEDOGLU and JIANHONG SHEN. Digital inpainting based on the mumford–shah–euler image model. *European Journal of Applied Mathematics*, 13(4):353–370, 2002. 2

[9] Jing Hao, Zhixin Zhang, Shicai Yang, Di Xie, and Shiliang Pu. Transforensics: Image forgery localization with dense self-attention, 2021. 2

[10] Xuefeng Hu, Zhihan Zhang, Zhenye Jiang, Syomantak Chaudhuri, Zhenheng Yang, and Ram Nevatia. SPAN: spatial pyramid attention network forimage manipulation localization. *CoRR*, abs/2009.00726, 2020. 2

[11] Jia-Bin Huang, Sing Bing Kang, Narendra Ahuja, and Johannes Kopf. Image completion using planar structure guidance. *ACM Trans. Graph.*, 33(4):129:1–129:10, 2014. 2

[12] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr – modulated detection for end-to-end multi-modal understanding, 2021. 3

[13] Ang Li, Qiuhong Ke, Xingjun Ma, Haiqin Weng, Zhiyuan Zong, Feng Xue, and Rui Zhang. Noise doesn't lie: Towards universal detection of deep inpainting, 2021. 2

[14] Dong Liu, Xiaoyan Sun, Feng Wu, Shipeng Li, and Ya-Qin Zhang. Image compression with edge-based inpainting. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(10):1273–1287, 2007. 2

[15] Jimmy SJ Ren, Li Xu, Qiong Yan, and Wenxiu Sun. Shepard convolutional neural networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. 2

[16] Jian Sun, Lu Yuan, Jiaya Jia, and Heung-Yeung Shum. Image completion with structure propagation. Association for Computing Machinery, Inc., August 2005. 2

[17] Haiwei Wu and Jiantao Zhou. Iid-net: Image inpainting detection network via neural architecture search and attention. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3):1172–1185, 2022. 1, 2, 4

[18] Junyuan Xie, Linli Xu, and Enhong Chen. Image denoising and inpainting with deep neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. 2

[19] Li Xu, Jimmy SJ Ren, Ce Liu, and Jiaya Jia. Deep convolutional neural network for image deconvolution. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. 2

[20] Raymond A. Yeh, Chen Chen, Teck Yian Lim, Alexander G. Schwing, Mark Hasegawa-Johnson, and Minh N. Do. Semantic image inpainting with deep generative models, 2016. 2

[21] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. Generative image inpainting with contextual attention, 2018. 2

[22] Yingchen Yu, Fangneng Zhan, Shijian Lu, Jianxiong Pan, Feiying Ma, Xuansong Xie, and Chunyan Miao. Wavefill: A wavelet-based generation network for image inpainting, 2021. 2

[23] Wael AbdAlmageed Yue Wu and Premkumar Natarajan. Mantra-net: Manipulation tracing network for detection and localization of image forgerieswith anomalous features. 2019. 3

[24] Haitian Zheng, Zhe Lin, Jingwan Lu, Scott Cohen, Eli Shechtman, Connelly Barnes, Jianming Zhang, Ning Xu, Sohrab Amirghodsi, and Jiebo Luo. Cm-gan: Image inpainting with cascaded modulation gan and object-aware training, 2022. 2