Abstract

This thesis presents a comprehensive analysis of existing research and improvements in numerous aspects in image inpainting forgery detection, addressing important difficulties and novel solutions.

Since image processing methods first emerged, there has been much research on the alteration of digital images. Image editing software has become increasingly powerful over time, enabling users to modify images for artistic or commercial purposes. Image inpainting is a significant advancement in this domain since it enables the smooth reconstruction of absent or deleted sections of an image. Initially developed for simple purposes such as repairing damaged photographs or eliminating unwanted elements, inpainting has evolved into a sophisticated tool with positive as well as negative capabilities. The accelerated progress of inpainting technologies creates an increasing demand for trustworthy detecting systems. As inpainting techniques advance, they challenge traditional forensic tools, requiring the creation of novel methods to detect manipulated content.

To advance the research field, I have designed and implemented architectures that address known issues of current techniques as well as introduce novel algorithms for inpainting forgery detection. The thesis begins with a comprehensive evaluation of both inpainting methods and the detection of these methods. To address the limitations of current state-of-the-art datasets, we propose a new, extensible dataset that encompasses multiple inpainting methods and allows for a robust evaluation of detection techniques. Unlike existing datasets that often focus on a single inpainting approach, our dataset is designed to assess the performance of detection methods across various techniques. A key innovation of our dataset lies in its construction methodology: we leverage a public semantic segmentation dataset (Google V7), manually analyze each image, and carefully select a single object for removal. Crucially, the selection process was designed to favor inpainting performance rather than aiding detection. Specifically, we prioritized objects surrounded by textures and patterns that are relatively easier for inpainting algorithms to reconstruct, rather than those that inherently challenge inpainting techniques. This ensures that any detection difficulty is not artificially introduced by choosing particularly complex removal scenarios but rather stems from the inpainting method's ability to realistically synthesize missing regions. By applying multiple inpainting methods to the resulting image-mask pairs, our approach enables a comprehensive assessment of detection robustness across different techniques. Additionally, it allows us to systematically identify and categorize inpainting artifacts, as detailed in Chapter Two. Initial extensive experiments revealed that classical inpainting techniques, whether diffusion-based or patch-based, are relatively easy to detect. However, images altered using machine learning-based methods present a greater challenge. Our findings indicate that even the most effective detection methods achieve an average IoU and F1 score below 35%, highlighting significant opportunities for improvement in detection mechanisms.

The second focus of the thesis, following the setting up of the mathematical model for inpainting artifacts, was to explore various methods for feature extraction of these artifacts. Our research demonstrated that one of these methods relies on complex wavelets. With this information, the subsequent step involved developing multiple detection methods that utilize complex wavelets, alongside semantic segmentation and analysis of noise level inconsistencies. We initially proposed a novel method that integrates wavelet feature extraction with semantic segmentation. The results appeared superior to those achieved by the current leading detection

method, exhibiting an average IoU of 0.5. The observed variances in performance, particularly in recall and IoU, across different inpainting techniques for all detection methods suggest a potential limitation. The initial proposed method may lack consistent effectiveness in detecting artifacts produced by various inpainting processes. With this relevant data considered, the next step involved re-evaluating the proposed method to include a noise inconsistency module. The detection process was modified. Semantic segmentation was conducted with complex wavelet feature extraction, followed by operations performed at each individual segment level. Each segment's area is improved using various methods, with the median-modified Wiener filter being the most significant. The final steps involve an analysis of noise consistency. This improvement increased the overall IoU to approximately 63%.

With the increase in this value, we have continued with several ablation studies. The results were not as favorable as anticipated. In a separate dataset, IID, the average IoU results decreased to 0.49. We postulate that several factors may contribute to this issue, such as the small image size of 256x256 pixels and the random generation of the inpainted area, which does not remove a real object from the images. Nonetheless, there remains room for improvement. Furthermore, our ablation study examines the effects of different post-processing methods applied to the image and their influence on detection outcomes. As anticipated, altering the distribution of pixels through operations such as resizing and blurring will significantly reduce detection results.

The prior results provided confidence to further investigate machine learning methods, incorporating our findings: complex wavelets can be utilized for feature extraction in conjunction with semantic segmentation to improve the enhancement of forged areas. Consequently, and informed by a literature review, we have developed custom neural networks to address the process of inpainting detection. We have examined approaches ranging from classical UNET architecture to LSTM-based models. The optimal results were achieved with the network referred to as StackDeepAll. For each individual channel, a subnetwork, UNet like network, is developed from the complex coefficients. Each subnet is tasked with improving the forged area in their corresponding complex wavelet channel and ultimately generating an average score for each pixel, distinguishing between forged and real. A supplementary sub-network is appended at the end, integrating all outputs from preceding networks as input. This approach yields favorable results, achieving an average Intersection over Union (IoU) greater than 0.8 in both validation and testing phases. This approach has the disadvantage of many parameters in the overall architecture.

To address this limitation and draw from recent research, we have developed a neural network architecture that integrates wavelet scattering within a UNET-like structure, incorporating an additional noise inconsistency module to reduce the incidence of false positives. The proposed architecture can achieve results with an IoU exceeding 0.8. The proposed architecture exhibits a reduction in the number of parameters relative to its predecessor. The core idea of the proposed method is that, for a specific texture, wavelet coefficients are expected to align with a particular statistical distribution, typically a standard distribution, particularly in high-level, noise-free regions. Deviations from the expected distribution suggest potential abnormalities or inconsistencies. This expectation aligns with the general presumption that in a coherent texture, the response of wavelet coefficients should exhibit stability, conforming to a near-normal distribution as outlined by the central limit theorem in the statistical analysis of natural images. Substantial deviations of the coefficients from the expected distribution suggest possible inconsistencies, indicating either noise or textural irregularities that differ from the original structure. This method ensures that inconsistencies are not merely variations in pixel color but statistically significant deviations in the texture's structure. The method appears stable; however, issues may occur when processing images with overlapping textures or minor inconsistencies that do not significantly affect wavelet coefficient distributions. Similar to the preceding chapter, we have conducted ablation studies for the proposed method. The results appear to correlate with our prior findings: increased disturbance in the post-processing method leads to reduced adaptation by the method.