BioFors: A Large Biomedical Image Forensics Dataset

Ekraam Sabir, Soumyaroop Nandi, Wael AbdAlmageed, Prem Natarajan
USC Information Sciences Institute, Marina del Rey, CA, USA
{esabir, soumyarn, wamageed, pnataraj}@isi.edu

Abstract

Research in media forensics has gained traction to combat the spread of misinformation. However, most of this research has been directed towards content generated on social media. Biomedical image forensics is a related problem, where manipulation or misuse of images reported in biomedical research documents is of serious concern. The problem has failed to gain momentum beyond an academic discussion due to an absence of benchmark datasets and standardized tasks. In this paper we present BioFors[1] *– the first dataset for benchmarking common biomedical image manipulations. BioFors comprises 47,805 images extracted from 1,031 open-source research papers. Images in BioFors are divided into four categories – Microscopy, Blot/Gel, FACS and Macroscopy. We also propose three tasks for forensic analysis – external duplication detection, internal duplication detection and cut/sharp-transition detection. We benchmark BioFors on all tasks with suitable state-of-the-art algorithms. Our results and analysis show that existing algorithms developed on common computer vision datasets are not robust when applied to biomedical images, validating that more research is required to address the unique challenges of biomedical image forensics.*

1. Introduction

Multimedia forensic research has branched off into several sub-domains to tackle various forms of misinformation and manipulation. Popular forensic research problems include detection of digital forgeries such as deepfakes [31, 41], copy-move and splicing manipulations [52, 53, 51] or semantic forgeries [40, 23]. These forensic-research areas essentially deal with social media content. A related but distinct research domain is biomedical image forensics; i.e. detection of research misconduct in biomedical publications [4, 13, 5]. Research misconduct can appear in several forms such as plagiarism, fabrication and falsification. Scientific misconduct has consequences beyond ethics and leads to retractions [5] and by one estimate $392, 582 of financial loss for each retracted article [46]. The general scope of scientific misconduct and unethical behavior is broad. In this paper we focus on detection of manipulation or inappropriate duplication of scientific images in biomedical literature.

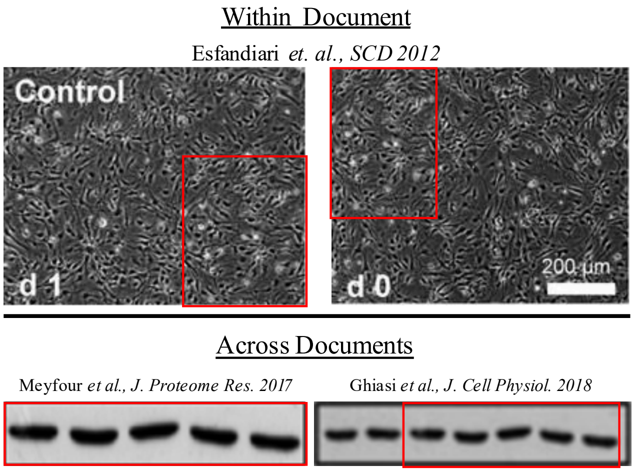
Within Document
Esfandiari et. al., SCD 2012
Control
d 1
d 0
200 µm
Across Documents
Meyfour et al., J. Proteome Res. 2017
Ghiasi et al., J. Cell Physiol. 2018

Figure 1. Real world examples of suspicious duplications in biomedical images. Top and bottom rows show duplications between images in the same and different documents respectively.

Duplication and tampering of protein, cell, tissue and other experimental images has become a nuisance in the biomedical sciences community. As the description suggests, duplication involves reusing part of images generated by one experiment to misrepresent results for unrelated experiments. Tampering of images involves pixel- or patch-level forgery to hide unfavorable aspects of the image or to produce favorable results. Biomedical image forgeries can be more difficult for a human to detect than manipulated images on social media due to the presence of arbitrary and confusing patterns and lack of real-world semantic context. Detecting forgeries is further complicated by manipulations involving images across different documents. Figure 1 shows reported examples[2] of inappropriate duplications in

[1] https://github.com/ISICV/BioFors

[2] https://scienceintegritydigest.com/2020/11/11/46-papers-from-a-royan-institute-professor/

different publications. The difficulty of noticing such manipulations coupled with a high paper-per-reviewer ratio often leads to these manipulations going unnoticed during the review process. It may come under scrutiny later leading to possible retractions [5]. While the problem has received the attention of the biomedical community, to the best of our knowledge there is no publicly available biomedical image forensics dataset, detection software or standardized task for benchmarking. We address these issues by releasing the first biomedical image forensics dataset (BioFors) and proposing benchmarking tasks.

The objective of our work is to advance biomedical forensic research to identify suspicious images with high confidence. We hope that BioFors will promote the development of algorithms and software which can help reviewers identify manipulated images in research documents. The final decision regarding malicious, mistaken or justified intent behind a suspicious image is to be left to the forensic analyst. This is important due to cases of duplication/tampering that are justified with citation, explanation, harmlessness or naive mistake as detailed in [4]. BioFors comprises 47,805 manually cropped images belonging to four major categories — (1) Microscopy, (2) Blot/Gel, (3) Macroscopy and (4) Flow-cytometry or Fluoroscence-activated cell sorting (FACS). It covers popular biomedical image manipulations with three forgery detection tasks. The dataset and its collection along with the forgery detection tasks are detailed in Section 3.

The contributions of our work are:

- A large scale biomedical image forensics dataset with real-world forgeries
- A computation friendly taxonomy of forgery detection tasks that can be matched with standard computer vision tasks for benchmarking and evaluation
- Extensive analysis explaining the challenges of biomedical forensics and the loss in performance of standard computer vision models when applied to biomedical images

2. Related Work

2.1. Computer Vision in Biomedical Domain

Machine learning and computer vision have made significant contributions to the biomedical domain involving problems such as image segmentation [27, 28, 49], disease diagnostics [35], super-resolution [34] and biomedical image denoising [55]. While native computer vision algorithms have existed for these problems, ensuring robustness on biomedical data has always been a challenge. This is partly due to domain shift and also due to the difficulty of training data-intensive deep learning models on biomedical datasets which are usually small.

2.2. Natural-Image Forensics

Image forensics is a widely studied problem in computer vision with standard datasets and benchmarking [48]. Common forensics problems include deepfake detection [31, 41], splicing [51, 15], copy-move forgery detection (CMFD) [52, 14, 39], enhancement and removal detection [53, 56]. While some forms of manipulation such as image enhancement may be harmless, others have malicious intent. Recently, deepfakes – a class of forgeries where a person's identity or facial experession is manipulated, has gained notoriety. Other malicious forms of forgeries are copy-move and splicing which involve pasting an image patch from within the same image and from a donor image respectively. For the manipulations mentioned, forgery detection methods have been developed to flag suspicious content with reasonable success. A critical step for the development of these algorithms has been the curation and release of datasets that facilitated benchmarking. As an example, FF++ [37], DeeperForensics [25] and Celeb-DF [29] helped develop methods for deepfake detection. Similarly, CASIA [16], NIST16 [1], COLUMBIA [33] and COVERAGE [50] helped advance detection methods for a combination of forgeries such as copy-move, splicing and removal.

2.3. Biomedical-Image Forensics

Misrepresentation of scientific research is a broad problem [7] out of which image manipulation or duplication of biomedical images has been recognized as a serious problem by journals and the community in general [13, 4, 5]. Bik *et al.* [4] analyzed over 20,000 papers and found 3.8% of these to contain at least one manipulation. In continuing research [5], the authors were able to bring 46 corrections or retractions. However, most of this effort was performed manually which is unlikely to scale given the high volume of publications. Models and frameworks have been proposed for automated detection of biomedical image manipulations [10, 8, 2, 54, 26]. Koppers *et al.* [26] developed a duplication screening tool evaluated on three images. Bucci *et al.* [8] engineered a CMFD framework from open-source tools to evaluate 1,546 documents and found 8.6% of it to contain manipulations. Acuna *et al.* [2] used SIFT [30] image-matching to find potential duplication candidates in 760k documents, followed by human review. In the absence of a robust evaluation, it is unknown how many documents with forgeries went unnoticed in [8, 2]. Cardenuto *et al.* [10] curated a dataset of 100 images to evaluate an end-to-end framework for CMFD task. Xiang *et al.* [54] test a heterogenous feature extraction model to detect artificially created manipulations in a dataset of 357 microscopy and 487 western blot images. It is unclear how the images were collected in [10, 54]. In summary, none of the proposed datasets unify the community around biomedical image forensics with standard benchmarking.

3. BioFors Benchmark

As discussed in Section 2, a dataset with standardized benchmarking is essential to advance the field of biomedical image forensics. Additionally, we want BioFors to have image level granularity in order to facilitate image and pixel level evaluation. Furthermore, it is desirable to use images with real-world manipulations. To this end, we used open-source or retracted research documents to curate BioFors. BioFors is a reasonably large dataset at the intersection of biomedical and image-forensics domain, with 46,064 pristine and 1,741 manipulated images, when compared to biomedical image datasets including FMD [55] (12,000 images before augmentation) and CVPPP [42] (284 images) and also compared to image forensics datasets, including Columbia [33] (180 tampered images), COVERAGE [50] (100 tampered images), CASIA [16] (5,123 tampered images) and MFC [19] (100k tampered images). Section 3.1 details the image collection procedure. Image diversity and categorization is described in Section 3.2. Proposed manipulation detection tasks are described in Section 3.3. A discussion on ethics is included in supplementary material.

3.1. Image Collection Procedure

Most research publications do not exhibit forgery, therefore collecting documents with manipulations is a difficult task. We received a set of documents from Bik *et al.* [4] along with raw annotations of suspicious scientific images which will be discussed in Section 3.3. Of the list of documents from different journals provided to us, we selected documents from PLOS ONE open-source journal comprising 1031 biomedical research documents published between January 2013 and August 2014.

The collected documents were in Portable Document Format (PDF), however direct extraction of biomedical images from PDF documents is not possible with available software. Furthermore, figures in biomedical documents are compound figures [43, 47] i.e. a figure comprises biomedical images, charts, tables and other artifacts. Sadly, state-of-the-art biomedical figure decomposition models [43, 47] have imperfect and overlapping crop boundaries. We overcome these challenges in two steps: 1) automated extraction of figures from documents and 2) manual cropping of images from figures. For automated figure extraction we used deepfigures [44]. We experimented with other open source figure extractors, but deepfigures had significantly better crop boundaries and worked well on all the documents. We obtained 6,543 figure images out of which 5,035 figures had biomedical images. For the cropping step, in order to minimize human error in manual crop boundaries we performed cropping in two stages. We cropped sub-figures with a loose bounding box, followed by a tight crop around images of interest. We filtered out synthetic/computer generated images such as tables, bar plots, histograms, graphs, flowcharts and diagrams. Verification of numerical results in synthetic images is beyond the scope of this paper. The image collection process resulted in 47,805 images. We created the train/test split such that a document and its images belong to the test set if it has at least one manipulation. Table 1 gives an overview of the dataset. For more statistics on BioFors please refer to the supplementary material.

Modality	**Train**	**Test**	**Total**
Documents	696	335	1,031
Figures	3,377	1,658	5,035
All Images	30,536	17,269	47,805
Microscopy Images	10,458	7,652	18,110
Blot/Gel Images	19,105	8,335	27,440
Macroscopy Images	555	639	1,194
FACS Images	418	643	1,061

Table 1. Top rows give a high level view of BioFors. Bottom rows provide statistics by image category. Training set comprises pristine images and documents.

3.2. Dataset Description

We classify the images from the previous collection step into four categories — (1) Microscopy (2) Blots/Gels (3) Flow-cytometry or Fluoroscence-activated cell sorting (FACS) and (4) Macroscopy. This taxonomy is made considering both the semantics and visual similarity of different image classes. Semantically, microscopy includes images from experiments that are captured using a microscope. They include images of tissues and cells. Variations in microscopy images can result from factors pertaining to origin (e.g. human, animal, organ) or fluorescent chemical staining of cells and tissues. This produces images of diverse colors and structures. Western, northern and southern blots and gels are used for analysis of proteins, RNA and DNA respectively. The images look similar and the specific protein or blot types are visually indistinguishable. FACS images look similar to synthetic scatter plots. However, the pattern is generated by a physical experiment which represents the scattering of cells or particles. Finally, Macroscopy includes experimental images that are visible to the naked eye and do not fall into any of the first three categories. Macroscopy is the most diverse image class with images including rat specimens, tissues, ultrasound, leaves, etc. Table 1 shows the composition of BioFors by image class. Figure 2 shows inter and intra-class diversity of each class. The image categorization discussed here is easily learnable by popular image classification models as shown in Table 2.

3.3. Manipulation Detection Tasks in BioFors

The raw annotations provided by Bik *et al.* [4] contain freehand annotations of manipulated regions and notes explaining why the authors of [4] consider them manipulated.

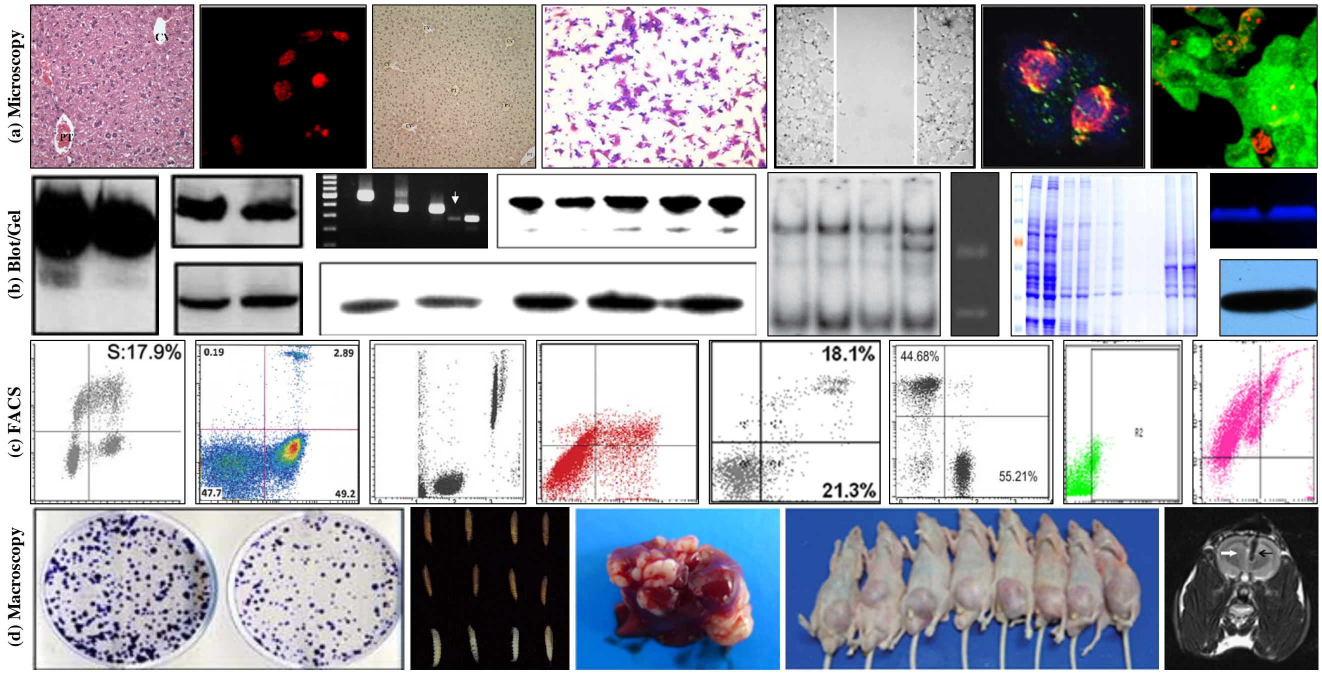

Figure 2. Rows of image samples representative of the following image classes: (a) Microscopy (b) Blot/Gel (c) FACS and (d) Macroscopy.

Model	Train	Test
VGG16 [45]	**99.79%**	97.11%
DenseNet [21]	99.25%	**97.67%**
ResNet [20]	98.93%	97.47%

Table 2. Accuracy of classifying BioFors images using popular image classification models is reliable.

Modality	EDD	IDD	CSTD
Documents	308	54	61
Pristine Images	14,675	2,307	1,534
Manipulated Images	1,547	102	181
All Images	16,222	2,409	1,715

Table 3. Distribution of pristine and tampered images in the test set by manipulation task.

However, the annotation format was not directly useful for ground truth computation. We inspected all suspicious images and manually created binary ground truth masks for all manipulations. This process resulted in 297 documents containing at least one manipulation. We also checked the remaining documents for potentially overlooked manipulations and found another 38 documents with at least one manipulation. Document level cohen's kappa (κ) inter-rater agreement between biomedical experts (raw annotations) and computer-vision experts (final annotation) is 0.91.

Unlike natural-image forensic datasets [33, 50, 1, 16] that include synthetic manipulations, BioFors has real-world suspicious images where the forgeries are diverse and the image creators do not share the origin of images or manipulation. Therefore, we do not have the ability to create a one-to-one mapping of biomedical image manipulation detection tasks to the forgeries described in Section 2.2. Consequently, we propose three manipulation detection tasks in BioFors — (1) external duplication detection, (2) internal duplication detection and (3) cut/sharp-transition detection. These tasks comprehensively cover the manipulations presented in [4, 13]. Table 3 shows the distribution of documents and images in the test set across tasks. We describe the tasks and their annotation ahead.

External Duplication Detection (EDD): This task involves detection of near identical regions between images. The duplicated region may span all or part of an image. Figure 3 shows two examples of external duplication. Duplicated regions may appear due to two reasons — (1) cropping two images with an overlap from a larger original source image and (2) by splicing i.e. copy-pasting a region from one image into another as shown in Figure 3a and b respectively. Irrespective of the origin of manipulation, the task requires detection of recurring regions between a pair of images. Further, another dimension of complexity for EDD stems from the orientation difference between duplicated regions. Duplicated regions in the second example of Figure 3 have been rotated by 180°. We also found orientation difference of 0°, 90°, horizontal and vertical flip. From an evaluation perspective, an image pair is considered one sample for EDD task and ground truth masks also oc-

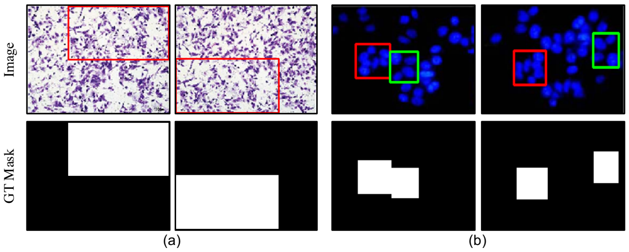

Figure 3. Two image pairs exhibiting duplication manipulation in EDD task. Duplicated regions are color coded to show correspondence. Bottom row shows ground truth masks for evaluation.

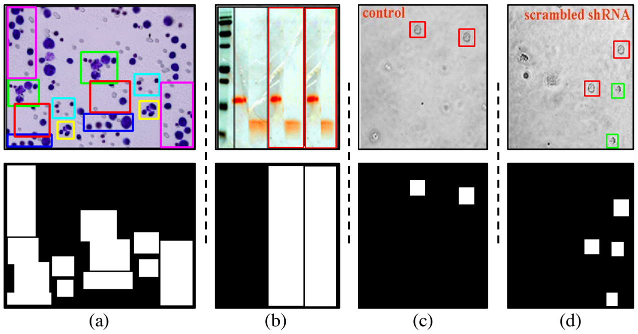

Figure 4. Manipulated samples in IDD task. Top row shows images and bottom row has corresponding masks. Repeated regions within the same image are color coded.

cur in pairs. The same image may have unique masks for different pairs corresponding to duplicated regions. Since, it is computationally expensive to consider all image pairs in a document, we drastically reduce the number of pairs to be computed by considering pairs of the same image class. This is a reasonable heuristic, since (1) we do not find duplications between images of different class and (2) automated image classification has reliable accuracy as shown in Table 2. For statistics on orientation difference and more duplication examples please refer to the supplementary material.

Internal Duplication Detection (IDD): IDD is our proposed image forensics task that involves detection of internally repeated image regions [52, 22]. Unlike a standard copy-move forgery detection (CMFD) task where the source region is known and is also from the same image, in IDD the source region may or may not be from the same image. The repeated regions may have been procured by the manipulator from a different image or document. Figure 4 shows examples of internal duplication. Notice that the regions highlighted in red in Figure 4c and d are the same and it is unclear which or if any of the patches is the source. Consequently from an evaluation perspective we treat all duplicated regions within an image as forged. Ground truth annotation includes one mask per image.

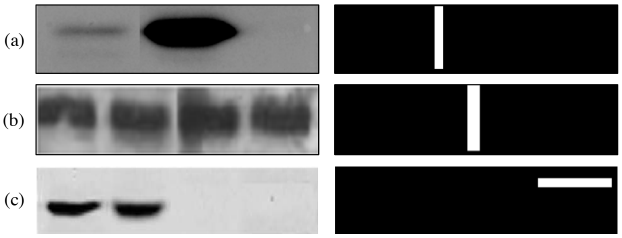

Figure 5. Examples of cuts/transitions. Noticeable sharp transition in (c) has been annotated, but the complete boundary is unclear.

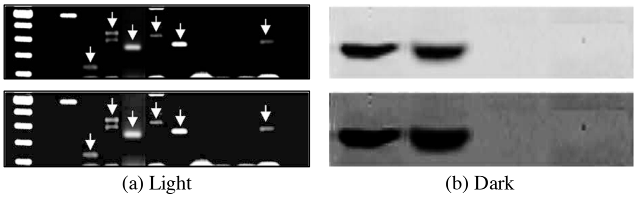
(a) Light
(b) Dark

Figure 6. Left and right examples show light and dark gamma correction of images making it easier to spot potential manipulations. The third arrow band in (a) appears to be spliced.

Cut/Sharp-Transition Detection (CSTD): A cut or a sharp transition can occur at the boundary of spliced or tampered regions. Unlike spliced images on social media, blot/gel images do not show a clear distinction between the authentic background and spliced foreground, making it difficult to identify the foreign patch. As an example, in Figure 5a and b it is not possible to identify if the left or right section of the western blot is spliced. Sharp transitions in texture can also occur from blurring of pixels or other manipulations of unknown provenance. In both cases, a discontinuity in image-texture in the form of a cut or sharp transition is the sole clue to detect manipulations. Accordingly we annotate anomalous boundaries as forged. From an annotation perspective, cuts or sharp transitions can be difficult to see, therefore we used gamma correction to make the images light or dark and highlight manipulated regions. Figure 6 shows examples of gamma correction. Ground truth is a binary mask for each image.

4. Why is Biomedical Forensics Hard?

Based on our insights from the data curation process and analysis of experimental results in Sec. 5, we explain potential challenges for natural-image forensic methods when applied to biomedical domain.

Artifacts in Biomedical Images: Unlike natural image datasets, biomedical images are scientific images presented in research documents. Accordingly, there are artifacts in the form of annotations and legends that are added to an image. Figure 7 shows some common artifacts that we found, including text and symbols such as arrows, scale and lines. The presence of these artifacts can create false positive matches for EDD and IDD tasks.

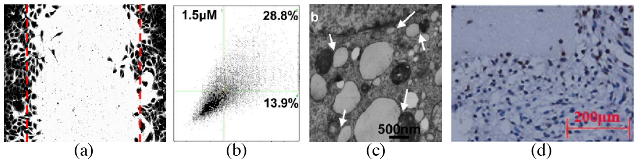

Figure 7. Examples of annotation artifacts in biomedical images: (a) dotted lines (b) alphanumeric text (c) arrows (d) scale.

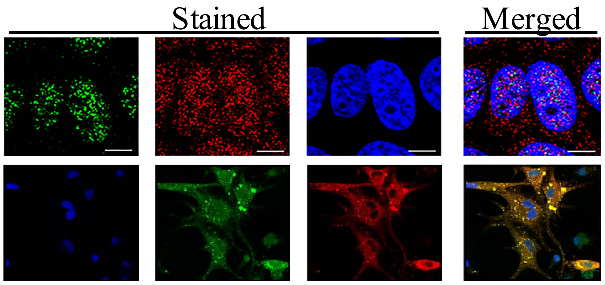

Figure 8. Left three columns show staining of microscopy images. Right column is an overlay of all stained images. Two or more images can be found tiled in this fashion.

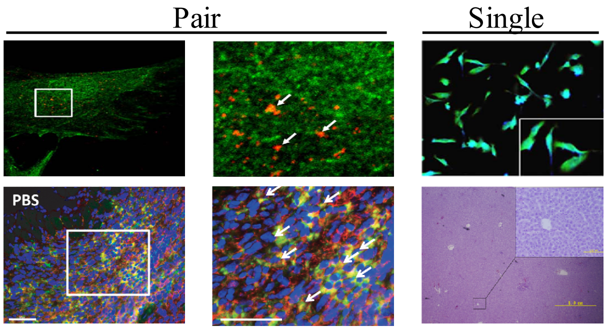

Figure 9. Images on the left show pairs of zoomed images. Right column has zoomed regions within the image. Rectangular bounding boxes are part of the original image.

Figure Semantics: Biomedical research documents contain images that are visually similar, but the figure semantics indicates that they are not manipulated. Two such statistically significant semantics are staining-merging and zoom. Forgery detection algorithms may generate false positive matches for images belonging to these categories. Stained images originate from microscopy experiments that involve colorization of the same cell/tissue sample with different fluorescent chemicals. This is usually followed by a merged/overlaid image which combines the stained images. The resulting images are tiled together in the same figure. Since the underlying cell/tissue sample is unchanged, the image structure is retained across images but with color change. Figure 8 shows some samples of staining and merging. The second semantics involves repeated portions of images that are magnified to highlight experimental results. Zoom semantics involves images that contain a zoomed portion of the image internally or themselves are a zoomed portion of another image. The zoomed area is indicated by a rectangular bounding box and images are adjacent. Figure 9 shows paired and single images with zoom semantics.

Image Texture: As illustrated in Figure 2, biomedical images tend to have a plain or pattern like texture with the exception of macroscopy images. This phenomena is particularly accentuated in blot/gel and microscopy images which are the largest two image classes and also contain the most manipulations. The plain texture of images makes it difficult to identify keypoints and extract descriptors for image matching, making descriptor based duplication detection difficult. We contrast this with the ease of identifying keypoints from two common computer vision datasets – Flickr30k [36] and Holidays [24]. Figure 10 shows the median number of keypoints identified in each image class using three off-the-shelf descriptor extractors: SIFT [30], ORB [38], BRIEF [9]. We resized all images to 256x256 pixels to account for differing images sizes. With the exception of FACS, other three image classes show a sharp decline in the number of extracted keypoints. We consider FACS to be an exception due to the large number of dots, where each dot is capable of producing a keypoint. However these keypoints may be redundant and not necessarily useful for biomedical image forensics.

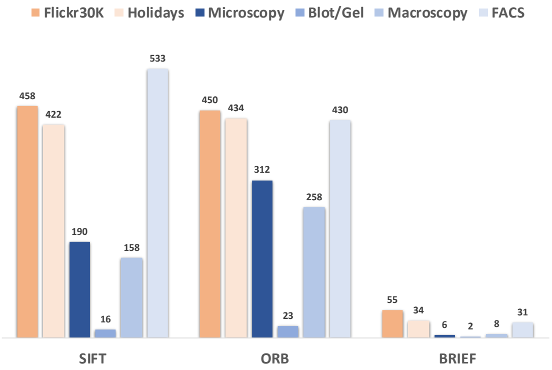

Figure 10. Median number of keypoints identified in images. Biomedical images have a relatively plain texture with the exception of FACS images, leading to fewer keypoints.

Hard Negatives: Scientific experiments often involve tuning of multiple parameters in a common experimental paradigm to produce comparative results. For biomedical experiments, this can produce very similar-looking images, which can act like hard negatives when looking for duplicated regions. For blot and gel images this can be true irrespective of a common experimental framework due to patterns of blobs on a monotonous background. Figure 11 shows some hard negative samples for each image class.

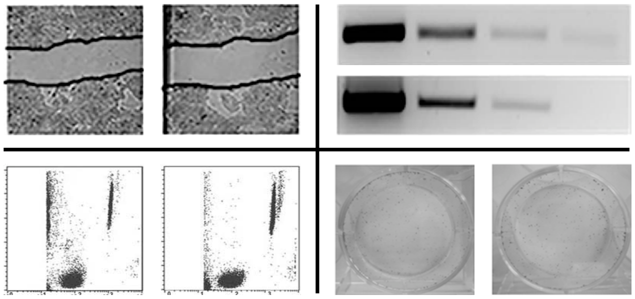

Figure 11. Hard negative samples from Blot/Gel, Macroscopy, FACS and Microscopy classes in clockwise order.

5. Evaluation and Benchmarking

5.1. Metrics

For all the manipulation tasks discussed in Section 3.3, detection algorithms are expected to produce a binary prediction mask of the same dimension as the input image. The predicted masks are compared against ground truth annotation masks included in the dataset. Manipulated pixels in images denote the positive class. Following previous work in forgery detection [52, 53, 51] we compute F_1 scores between the predicted and ground truth mask for all tasks. We also compute Matthews correlation coefficient (MCC) [32] between the masks since it has been shown to present a balanced score when dealing with imbalanced data [11, 6] as is our case with fewer manipulated images. MCC ranges from -1 to +1 and represents the correlation between prediction and ground truth. Due to space constraints, F_1 score tabulation is done in supplementary material. Evaluation is done both at the image and pixel-level i.e. true/false positives and true/false negatives are determined for each image and pixel. For image evaluation, following the protocol in [52], we consider an image to be manipulated if any one pixel has positive prediction. Pixel level evaluation across multiple images is similar to protocol A in [52] i.e. all pixels from the dataset are gathered for one final computation.

5.2. Baseline Models

We evaluate several deep learning and non deep learning models for our three tasks introduced in Section 3.3. Our baselines are selected from forensics literature based on model/code availability and task suitability. Deep-learning baselines require finetuning for weight adaptation. However, due to small number of manipulated samples, BioFors training set comprises pristine images only. Inspired by previous forgery detection methods [52, 51], we create synthetic manipulations on pristine training data to finetune models. Details of synthetic data and baseline experiments are provided in the supplementary material. To promote reproducibility, our synthetic data generators and evaluation scripts will be released with the dataset.

External Duplication Detection (EDD): Baselines for EDD should identify repeated regions between images. We evaluate classic keypoint-descriptor based image-matching algorithms such as SIFT [30], ORB [38] and BRIEF [9]. We follow a classic object matching approach, using RANSAC [17] to remove stray matches. CMFD algorithms can be used by concatenating two images to create a single input. We evaluated DenseField (DF) [14] with best reported transform – zernike moment (ZM) on concatenated images. Additionally, we evaluate a splicing detection algorithm, DMVN [51] to find repeated regions. DMVN implements a deep feature correlation layer which matches coarse image features at 16x16 resolution to find visually similar regions.

Internal Duplication Detection (IDD): Appropriate baselines for IDD should be suitable for identifying repeated regions within images. DenseField (DF) [14] proposes an efficient dense feature matching algorithm for CMFD. We evaluate it using the three circular harmonic transforms used in the paper: zernike moments (ZM), polar cosine transform (PCT) and fourier-mellin transform (FMT). We also evaluated the CMFD algorithm reported in [12], using three block based features – discrete cosine transform (DCT) [18], zernike moments (ZM) [39] and discrete wavelet transform (DWT) [3]. BusterNet [52] is a two-stream deep-learning based CMFD model that leverages visual similarity and manipulation artifacts. Visual similarity in BusterNet is identified using a self-correlation layer on coarse image features followed by percentile pooling.

Cut/Sharp-Transition Detection (CSTD): Unlike the previous two tasks, it is challenging to find forensics algorithms designed for detecting cuts or transitions. We evaluate ManTraNet [53], a state-of-the-art manipulation detection algorithm which identifies anomalous pixels and image regions. We also evaluated a baseline convolutional neural network (CNN) model for detecting cuts and transitions. The CNN was trained on synthetic manipulations in blot/gel images from the training set. For more details on the baseline please refer to supplementary material.

5.3. Results

Tables 4, 5 and 6 present baseline results for EDD, IDD and CSTD tasks respectively. We find that dense feature matching approaches (DF-ZM,PCT,FMT) are better than sparse (SIFT, SURF, ORB), block-based (DCT, DWT, Zernike) or coarse feature matching methods (DMVN and BusterNet) for identifying repeated regions in both EDD and IDD tasks. Dense feature matching is computationally expensive, and most image forensics algorithms obtain a viable quality-computation trade-off on natural images. However, biomedical images have relatively plain texture and

Method	Microscopy		Blot/Gel		Macroscopy		FACS		Combined	
	Image	Pixel	Image	Pixel	Image	Pixel	Image	Pixel	Image	Pixel
SIFT [30]	0.180	0.146	0.113	0.148	0.130	0.194	0.11	0.073	0.142	0.132
ORB [38]	0.319	0.342	0.087	0.127	0.126	0.226	0.269	0.187	0.207	0.252
BRIEF [9]	0.275	0.277	0.058	0.102	0.135	0.169	0.244	0.188	0.180	0.202
DF - ZM [14]	**0.422**	**0.425**	0.161	0.192	**0.285**	**0.256**	**0.540**	**0.504**	**0.278**	**0.324**
DMVN [51]	0.242	0.342	**0.261**	**0.430**	0.185	0.238	0.164	0.282	0.244	0.310

Table 4. Results for external duplication detection (EDD) task by image class. Image and Pixel columns denote image and pixel level evaluation respectively. All numbers are MCC scores. For corresponding F_1 scores, please refer to supplementary material.

Method	Microscopy		Blot/Gel		Macroscopy		Combined	
	Image	Pixel	Image	Pixel	Image	Pixel	Image	Pixel
DF - ZM [14]	**0.764**	0.197	**0.515**	0.449	0.573	0.478	0.564	0.353
DF - PCT [14]	**0.764**	**0.202**	0.503	**0.466**	**0.712**	**0.487**	**0.569**	**0.364**
DF - FMT [14]	0.638	0.167	0.480	0.400	0.495	0.458	0.509	0.316
DCT [18]	0.187	0.022	0.250	0.168	0.158	0.143	0.196	0.095
DWT [3]	0.299	0.067	0.384	0.295	0.591	0.268	0.341	0.171
Zernike [39]	0.192	0.032	0.336	0.187	0.493	0.262	0.257	0.114
BusterNet [52]	0.183	0.178	0.226	0.076	0.021	0.106	0.269	0.107

Table 5. Results for internal duplication detection (IDD) task by image class and a combined result. There are no IDD instances in FACS images. Image and Pixel columns denote image and pixel level evaluation respectively. All numbers are MCC scores.

Method	F_1		MCC	
	Image	Pixel	Image	Pixel
MantraNet [53]	**0.253**	**0.09**	**0.170**	**0.080**
CNN Baseline	0.212	0.08	0.098	0.070

Table 6. Results on the cut/sharp-transition detection (CSTD) task.

similar patterns, which may lead to indistinguishable features for coarse or sparse extraction. For the set of baselines evaluated, exchanging feature matching quality for computation is not successful on biomedical images. Furthermore, performance varies drastically across image classes for all methods, with models peaking across different image classes. The variation is expected since the semantic and visual characteristics vary by image category. However, as a direct consequence of this variance, image category specific models may need to be developed in future research. On CSTD, our simple baseline trained to detect sharp transitions produces false alarms on image borders or edges of blots. Both MantraNet and our baseline have similar performance, indicating that a specialized model design might be required to detect cuts and anomalous transitions. Finally, performance is low across all tasks which can be attributed to some of the challenges discussed in Section 4. In summary, it is safe to conclude that existing natural-image forensic methods are not robust when applied to biomedical images and also show high variation in performance across image classes. The results emphasize the need for robust forgery detection algorithms that are applicable to the biomedical domain. For sample predictions from reported baselines please refer to the supplementary material.

6. Conclusion and Future Work

Manipulation of scientific images is an issue of serious concern for the biomedical community. While reviewers can attempt to screen for scientific misconduct, the complexity and volume of the task places an undue burden on them. Automated and scalable biomedical forensic methods are necessary to assist reviewers. We presented BioFors, a large biomedical image forensics dataset. BioFors comprises a comprehensive range of images found in biomedical documents. We also framed three manipulation detection tasks based on common manipulations found in literature. Our evaluations show that common computer vision algorithms are not robust when extended to the biomedical domain. Our analysis shows that attaining respectable performance will require well designed models, as there are multiple challenges to the problem. We expect that BioFors will advance biomedical image forensic research.

7. Acknowledgement

We are deeply grateful to Dr. Elisabeth M. Bik, Dr. Arturo Casadevall and Dr. Ferric C. Fang for sharing with us the raw annotation of manipulations. Their contributions accelerated the creation of the dataset we have released. We also extend our special thanks to Dr. Bik for answering many of our questions and improving our understanding of the domain of biomedical images.

References

[1] Nimble challenge 2017 evaluation — nist. https://www.nist.gov/itl/iad/mig/nimble-challenge-2017-evaluation. (Accessed on 11/14/2020). 2, 4

[2] Daniel E Acuna, Paul S Brookes, and Konrad P Kording. Bioscience-scale automated detection of figure element reuse. *bioRxiv*, page 269415, 2018. 2

[3] M. Bashar, K. Noda, N. Ohnishi, and K. Mori. Exploring duplicated regions in natural images. *IEEE Transactions on Image Processing*, pages 1–1, 2010. 7, 8

[4] Elisabeth M Bik, Arturo Casadevall, and Ferric C Fang. The prevalence of inappropriate image duplication in biomedical research publications. *MBio*, 7(3), 2016. 1, 2, 3, 4

[5] Elisabeth M Bik, Ferric C Fang, Amy L Kullas, Roger J Davis, and Arturo Casadevall. Analysis and correction of inappropriate image duplication: the molecular and cellular biology experience. *Molecular and Cellular Biology*, 38(20), 2018. 1, 2

[6] Sabri Boughorbel, Fethi Jarray, and Mohammed El-Anbari. Optimal classifier for imbalanced data using matthews correlation coefficient metric. *PloS one*, 12(6):e0177678, 2017. 7

[7] Isabelle Boutron and Philippe Ravaud. Misrepresentation and distortion of research in biomedical literature. *Proceedings of the National Academy of Sciences*, 115(11):2613–2619, 2018. 2

[8] Enrico M Bucci. Automatic detection of image manipulations in the biomedical literature. *Cell death & disease*, 9(3):1–9, 2018. 2

[9] Michael Calonder, Vincent Lepetit, Christoph Strecha, and Pascal Fua. Brief: Binary robust independent elementary features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 778–792. Springer, 2010. 6, 7, 8

[10] JP Cardenuto, A Rocha, Relatório Técnico-IC-PFG, and Projeto Final de Graduação. Scientific integrity analysis of misconduct in images of scientific papers. 2019. 2

[11] Davide Chicco and Giuseppe Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):6, 2020. 7

[12] Vincent Christlein, Christian Riess, Johannes Jordan, Corinna Riess, and Elli Angelopoulou. An evaluation of popular copy-move forgery detection approaches. *IEEE Transactions on Information Forensics and Security*, 7(6):1841–1854, 2012. 7

[13] Jana Christopher. Systematic fabrication of scientific images revealed. *FEBS letters*, 592(18):3027–3029, 2018. 1, 2, 4

[14] Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva. Efficient dense-field copy–move forgery detection. *IEEE Transactions on Information Forensics and Security*, 10(11):2284–2297, 2015. 2, 7, 8

[15] Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva. Splicebuster: A new blind image splicing detector. In *2015 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE, 2015. 2

[16] Jing Dong, Wei Wang, and Tieniu Tan. Casia image tampering detection evaluation database. In *2013 IEEE China Summit and International Conference on Signal and Information Processing*, pages 422–426. IEEE, 2013. 2, 3, 4

[17] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 7

[18] A Jessica Fridrich, B David Soukal, and A Jan Lukáš. Detection of copy-move forgery in digital images. In *in Proceedings of Digital Forensic Research Workshop*. Citeseer, 2003. 7, 8

[19] Haiying Guan, Mark Kozak, Eric Robertson, Yooyoung Lee, Amy N Yates, Andrew Delgado, Daniel Zhou, Timothee Kheyrkhah, Jeff Smith, and Jonathan Fiscus. Mfc datasets: Large-scale benchmark datasets for media forensic challenge evaluation. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pages 63–72. IEEE, 2019. 3

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 4

[21] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4700–4708, 2017. 4

[22] Ashraful Islam, Chengjiang Long, Arslan Basharat, and Anthony Hoogs. Doa-gan: Dual-order attentive generative adversarial network for image copy-move forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4676–4685, 2020. 5

[23] Ayush Jaiswal, Yue Wu, Wael AbdAlmageed, Iacopo Masi, and Premkumar Natarajan. Aird: adversarial learning framework for image repurposing detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11330–11339, 2019. 1

[24] Herve Jegou, Matthijs Douze, and Cordelia Schmid. Hamming embedding and weak geometric consistency for large scale image search. In David Forsyth, Philip Torr, and Andrew Zisserman, editors, *Computer Vision – ECCV 2008*, pages 304–317, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg. 6

[25] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2886–2895. IEEE, 2020. 2

[26] Lars Koppers, Holger Wormer, and Katja Ickstadt. Towards a systematic screening tool for quality assurance and semiautomatic fraud detection for images in the life sciences. *Science and engineering ethics*, 23(4):1113–1128, 2017. 2

[27] Victor Kulikov and Victor Lempitsky. Instance segmentation of biological images using harmonic embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[28] Hong Joo Lee, Jung Uk Kim, Sangmin Lee, Hak Gu Kim, and Yong Man Ro. Structure boundary preserving segmentation for medical image with ambiguous boundary. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[29] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[30] David G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. 2, 6, 7, 8

[31] Iacopo Masi, Aditya Killekar, Royston Marian Mascarenhas, Shenoy Pratik Gurudatt, and Wael AbdAlmageed. Two-branch recurrent network for isolating deepfakes in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 1, 2

[32] Brian W Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451, 1975. 7

[33] Tian-Tsong Ng, Jessie Hsu, and Shih-Fu Chang. Columbia image splicing detection evaluation dataset. *DVMM lab. Columbia Univ CalPhotos Digit Libr*, 2009. 2, 3, 4

[34] Cheng Peng, Wei-An Lin, Haofu Liao, Rama Chellappa, and S. Kevin Zhou. Saint: Spatially aware interpolation network for medical slice synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[35] Fabio Perez, Sandra Avila, and Eduardo Valle. Solo or ensemble? choosing a cnn architecture for melanoma classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019. 2

[36] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2641–2649, 2015. 6

[37] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1–11, 2019. 2

[38] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2564–2571. Ieee, 2011. 6, 7, 8

[39] Seung-Jin Ryu, Min-Jeong Lee, and Heung-Kyu Lee. Detection of copy-rotate-move forgery using zernike moments. In *Proceedings of the 12th international conference on Information hiding*, pages 51–65, 2010. 2, 7, 8

[40] Ekraam Sabir, Wael AbdAlmageed, Yue Wu, and Prem Natarajan. Deep multimodal image-repurposing detection. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1337–1345, 2018. 1

[41] Ekraam Sabir, Jiaxin Cheng, Ayush Jaiswal, Wael AbdAlmageed, Iacopo Masi, and Prem Natarajan. Recurrent convolutional strategies for face manipulation detection in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 80–87, 2019. 1, 2

[42] Hanno Scharr, Massimo Minervini, Andreas Fischbach, and Sotirios A Tsaftaris. Annotated image datasets of rosette plants. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 6–12, 2014. 3

[43] Xiangyang Shi, Yue Wu, Huaigu Cao, Gully Burns, and Prem Natarajan. Layout-aware subfigure decomposition for complex figures in the biomedical literature. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1343–1347. IEEE, 2019. 3

[44] Noah Siegel, Nicholas Lourie, Russell Power, and Waleed Ammar. Extracting scientific figures with distantly supervised neural networks. In *Proceedings of the 18th ACM/IEEE on joint conference on digital libraries*, pages 223–232, 2018. 3

[45] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4

[46] Andrew M Stern, Arturo Casadevall, R Grant Steen, and Ferric C Fang. Financial costs and personal consequences of research misconduct resulting in retracted publications. *Elife*, 3:e02956, 2014. 1

[47] Satoshi Tsutsui and David J Crandall. A data driven approach for compound figure separation using convolutional neural networks. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 533–540. IEEE, 2017. 3

[48] Luisa Verdoliva. Media forensics and deepfakes: an overview. *arXiv preprint arXiv:2001.06564*, 2020. 2

[49] Dong Wang, Yuan Zhang, Kexin Zhang, and Liwei Wang. Focalmix: Semi-supervised learning for 3d medical image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[50] Bihan Wen, Ye Zhu, Ramanathan Subramanian, Tian-Tsong Ng, Xuanjing Shen, and Stefan Winkler. Coverage—a novel database for copy-move forgery detection. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 161–165. IEEE, 2016. 2, 3, 4

[51] Yue Wu, Wael Abd-Almageed, and Prem Natarajan. Deep matching and validation network: An end-to-end solution to constrained image splicing localization and detection. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1480–1502, 2017. 1, 2, 7, 8

[52] Yue Wu, Wael Abd-Almageed, and Prem Natarajan. Busternet: Detecting copy-move image forgery with source/target localization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 1, 2, 5, 7, 8

[53] Yue Wu, Wael AbdAlmageed, and Premkumar Natarajan. Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2019. 1, 2, 7, 8

[54] Ziyue Xiang and Daniel Acuna. Scientific image tampering detection based on noise inconsistencies: A method and datasets. *arXiv preprint arXiv:2001.07799*, 2020. 2

[55] Yide Zhang, Yinhao Zhu, Evan Nichols, Qingfei Wang, Siyuan Zhang, Cody Smith, and Scott Howard. A poisson-gaussian denoising dataset with real fluorescence microscopy images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11710–11718, 2019. 2, 3

[56] Peng Zhou, Xintong Han, Vlad I. Morariu, and Larry S. Davis. Learning rich features for image manipulation detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2