Benchmarking Scientific Image Forgery Detectors

João P. Cardenuto, Anderson Rocha, Senior Member, IEEE
RECOD Lab., Institute of Computing, University of Campinas
Email: {phillipe.cardenuto, anderson.rocha}@ic.unicamp.br

Abstract—The scientific image integrity area presents a challenging research bottleneck, the lack of available datasets to design and evaluate forensic techniques. Its data sensitivity creates a legal hurdle that prevents one to rely on real tampered cases to build any sort of accessible forensic benchmark. To mitigate this bottleneck, we present an extendable open-source library that reproduces the most common image forgery operations reported by the research integrity community: duplication, retouching, and cleaning. Using this library and realistic scientific images, we create a large scientific forgery image benchmark (39,423 images) with an enriched ground-truth. In addition, concerned about the high number of retracted papers due to image duplication, this work evaluates the state-of-the-art copy-move detection methods in the proposed dataset, using a new metric that asserts consistent match detection between the source and the copied region. The dataset and source-code will be freely available upon acceptance of the paper.

Index Terms—Scientific Integrity Benchmark, Image Forgery Library, Computational Scientific Integrity, Image Forensics, Tampering Detection, Algorithm Evaluation.

I. INTRODUCTION

Integrity researchers have been reporting a threat to scientific image integrity for a long time[1]. This improbity has achieved even areas like cancer[5], or more dramatically used on ‘paper mills’ services[6].

Although this serious problem has been an urge in the scientific community, to the best of our knowledge, few forensics works are dedicated to this topic. So far, no rich annotated forensic benchmark containing scientific tampering images was published. We believe that a large dataset would foster the forensic community to work more actively in this subject and assist state-of-the-art forensic techniques that might require large training datasets.

In this sense, we tried to collect known doctored scientific images, but we faced two main issues that make us avoid these methodology: legal and practical. To publish a dataset with real tampering cases, we would have to face copyrights and legal aspects of pointing third-party works that were retracted due to suspicious manipulation. Even if we decided to manage this legal aspect, we had to be guided by a retraction notice relative to suspicious manipulation. Even if we decided to manage this legal aspect, we had to be guided by a retraction notice relative to the issued images. However, after reading several retraction notices, we realized that many of them are not precise enough to pinpoint the issued region’s at a pixel level, which would not lead to an accurate ground-truth. The example of a real retraction notice due to an honest error, depicted in Fig. 1, shows this inaccuracy, in which the highlighted words ‘some lanes’ and ‘not the appropriate ones’ translate to an ambiguous region and cause.

On the other hand, we notice that a representative number of retracted images were due to duplication and basic image processing operations that could be automatically created. Therefore, this work presents the RECOD Scientific Image Integrity Library (RSIIL) that enables creating a synthetic scientific image tampering dataset with enriched pixel-wise ground-truth and without any associated legal issue. With this library, we created the RECOD Scientific Image Integrity Dataset (RSIID) with the most common image operations reported by scientific integrity researchers[2], [4]. To create this benchmark, we doctored 2,923 figures from creative common sources resulting in 39,423 tampered figures (26,496 for training and 12,927 for testing). In addition, we propose a new metric to evaluate copy-move forgery detection dedicated to scientific images using an enriched ground-truth map, that assert a consistent detection match between the cloned region and its source. Finally, using this new dataset and metric, we evaluate the performance of the state-of-the-art copy-move forgery detection[7], [8], [9], establishing a baseline and setting the ground for any future investigation.

We organize the remaining of this paper into seven sections: Section II presents related work while Section III details the proposed library, RSIIL. Section IV presents the dataset RSID while Section V brings out a new evaluation metric aimed at a more consistent copy-move detection evaluation. Section VI presents an analysis of state-of-the-art copy-move forgery detectors on the proposed dataset setting the ground for future research while Section VII presents the conclusions and future work directions.

II. RELATED WORK

To the best of our knowledge, few works try to design a tampering benchmark focused on scientific images. So far, we were only able to find two works that address scientific integrity image datasets. The first one is from Xiang and Acuna[10], which created a synthetic tampering dataset of scientific images from the web. They doctored microscopy and
western blot images using three types of manipulations that they claim to be the common cause of problems in scientific papers: cleaning of an image region with a single color or noise (Cleaning); copying an alien content region into the image (Splicing); and applying visual adjustments in the image content (Retouching). Their dataset contains 747 manually manipulated scientific images, of which 616 are dedicated to Removal. As we were only able to find the pre-print version of [10], we could not find any released data. Due to this, the quality of their manipulations and the dataset license is still unclear. Despite the authors manually constructed the dataset to create a more realistic scenario, their dataset is still limited to a small size that might not represent the diversity of scientific images. Besides this, the dataset is highly concentrated on Cleaning, preventing one to properly evaluate the robustness of a forensic method among all modalities.

The second one is the work of Koker et al. [11], named as Bio-Image Near-Duplicate Examples Repository (BINDER), which have the pioneering idea of using legal issue-less scientific images for an integrity dataset. This dataset is limited to finding near-duplicate images, aiming to find image re-use across scientific publications. Their dataset has 10,179 non-overlapping patches tiled in 256 × 256 or 128 × 128 pixels. To create their dataset, they gathered microscopy images from the following public repositories: NYU Mouse Embryo Tracking Database (METD), the Broad Bioimage Benchmark Collection (BBBC), the Adiposoft Image Dataset (AID), and the Open Microscopy Image Data Resource (IDR). Besides, they also applied some geometric, brightness/contrast, and compression transformations on some images. However, their dataset is still not as realistic as the figures presented in scientific publications. Despite scientific images often embed graphical elements and captions, hampering to detect re-use, the authors did not add these elements to the images. In addition, they did not apply any local tampering (region-level), which is also a typical manipulation in inappropriate image re-use [4].

In addition to these works, we also found two scientific integrity initiatives that collect real cases of retracted papers. The first is the Retraction Watch Database maintained by the Retraction Watch. This database has more than 20,000 metadata of retracted, corrected, or concerned papers. The metadata presents the paper’s title, retraction reason, authors, and Digital Object Identifier (DOI), among other fields. Although this database is not dedicated to image integrity issues, it is possible to filter the retracted papers to this category. However, only the paper’s metadata will be retrieved – due to legal aspects; it is not possible to retrieve the articles PDF, Figures, or Retraction notice –. which is a drawback of this database.

The second is the HEADT Centre Image Integrity Database[1] an initiative focused on researchers and developers working on scientific image manipulation detection. Their database contains more than 500 images’ metadata from retracted papers due to image manipulation. In addition to the basic information of a paper (Title, Authors, Publisher, Journal), they also added a text description of each manipulation, including the figure’s panel in which the manipulation occurs and its category (e.g., copy-move). This text description is based on the retraction notice associated with the figure; therefore, some text also presents ambiguity as depicted in the Figure [1]. Despite this text description, we could not find any manipulation map at pixel level in this dataset, which we believe is needed to evaluate a detection method properly.

III. RECOD SCIENTIFIC IMAGE INTEGRITY LIBRARY - RSIIIL

Before working with synthetic data, we tried to gather real-world problematic scientific images. To avoid any bias from our side, we relied upon retracted papers due to image problems given that they have a retraction notice resultant of an integrity investigation. However, to publish an accessible benchmark for forensic research, we possibly would have to deal with some legal aspects (e.g., figure copyright and causing possible defamation to someone).

As reported by Adam Marcus [12], a retracted paper could make their authors feel their reputation harmed and make them sue journals for defamation. Azoulay et al. [14] also indicate that retraction due to misconduct —which are the most important papers to be included in a forensic benchmark— has a significant reputation penalty to their authors. Even co-authors that might not be involved in the image manipulation, who already suffered severe consequences [14], could be affected by such benchmark since it would promote their association with the retracted paper.

Besides this legal aspect, we also faced some practical issues regarding data annotation. When manually annotating the problematic figures’ regions following their retraction notice, we experienced an absence of standard, including vague sentences (as illustrated in Figure [1]), resulting on unreliable ground-truths.

Because of these issues, we decide to avoid using real-world scientific problematic image and create a photorealistic dataset using the library introduced in this section. Thus, this Section presents the types of forgeries implemented in the library (Sub-Section A), explains how the library mimics realistic figures as they usually are presented in scientific documents (Sub-Section B), and addresses the manipulation ground-truth (Sub-Section C). Finally, the section also discusses how the proposed library is amenable to extensions of new image manipulations types (Sub-Section D).

A. Library functionalities

The goal of the library is to implement the most common image manipulations reported in the scientific community. Although we are aware of the possibilities of more complexity

1http://celltracking.bio.nyu.edu (Last access May, 2021)
2https://bbbc.broadinstitute.org (Last access May, 2021)
3https://imagej.net/Adiposoft (Last access May, 2021)
4https://idr.openmicroscopy.org (Last access May, 2021)
5http://retractiondatabase.org (Last access May, 2021)
6A non-profit organization affiliated with the Center for Scientific Integrity and dedicated to report and discuss cases of retracted papers and related issues.
7https://headt.eu/Image-Integrity-Database (Last access May, 2021)
tools for image manipulation, for example, the creation of scientific images using artificial intelligence (AI) algorithms. We suspect that these tools are not vastly used yet due to their complexities. Therefore, while designing the library, we adopt the forgery function based on the most common image processing operation accessible for a non-expert in AI or Computer Vision. We also design each block of the library to allow it to be extended to other more complex operations in the future.

Following the research from Bik et al. and Rossner and Yamada, we selected three main types of manipulations that can be recreated using common image processing software (e.g., Adobe Photoshop):

1) **Retouching**: The process of image beautification leading to an experiment misreading. This modality implements contrast, brightness, and blurring adjustments that highlight or obfuscate an image region. Figure 2 depicts an image that we applied retouching with our library. In Figure 2a, we used a Gaussian filter within the selected objects to obfuscate its content. Figure 2b illustrates an image with contrast and brightness adjustment, in which the method changes the selected object pixels intensity to cause an experimental misreading.

2) **Cleaning**: The result of obfuscating a foreground object using a background region. For this modality, we use inpainting and a brute-force routine. For the inpainting, we use the method of Criminisi et al. implemented by Moura. For the brute-force routine, we develop an in-house method to mimic the forgery procedure of a person seeking to cover an object using the background. To implement this routine, we select a foreground object \( \mathcal{FO} \); then, using brute force, we fit \( \mathcal{FO} \) on a background region \( \mathcal{BR} \) that has the most similar color histogram of this object; finally, we copy \( \mathcal{BR} \) into \( \mathcal{FO} \) and blur the border of \( \mathcal{FO} \), smoothing (feather edges) the difference from the copied \( \mathcal{BR} \) and the neighborhood of \( \mathcal{FO} \).

\(^8\)Code available at https://github.com/igorcmoura/inpaint-object-remover. (Last access March, 2021)

3) **Duplication**: The action of copying and pasting a region of an image within the same or another image, using or not post-processing operations. Note that this definition includes both copy-move and splicing. We organized this category into three sub-categories:

a) **Copy-Move Forgery**: Duplication of a region within the same image using geometric transformations (translation, rotation, flip, and scaling) and post-processing (e.g., retouching). All transformations can be combined with another. Due to the intrinsic result of scaling, we always combined it with another operation, otherwise it would cover the source object region. Besides these transformations, we also implemented a random object-to-background copy-move (that we named Random). This routine copies a random object \( \mathcal{RO} \) to a background region \( \mathcal{BR} \) that has the same shape as \( \mathcal{RO} \).

b) **Overlap Forgery**: Creation of two images with an overlap region from a single one. From a source image \( \mathcal{I} \), we select different regions of \( \mathcal{I} \) that share an overlap area to create two images from these regions. Any of these new regions can suffer post-processing to obfuscate its source. Figure 3 depicts the creation of an image with an overlap area.

c) **Splicing**: Creation of an image composition that uses a donor figure’s elements into a host one. Figure 5 depicts an Splicing forgery.

Despite all cases are generated without any human interac-
Fig. 4. Example of Copy-Move Forgery implemented in the library. The object of image (e) containing an arrow is duplicated with (a) translation, (b) rotation, (c) flip, and (d) scaling and pasted within the same image.

Fig. 5. Example of Overlap forgery included in the library. (a) represent a source image that is divided in overlapping regions A and B, and then presented as unique images in (b) and (c). The region A (b) suffer a post-processing brightness adjustment to make harder to compare with region B (c).

Fig. 6. Example of Splicing forgery function included in the library. The object highlighted with a red arrow from the donor image (a) is placed in a background region of the host image (b) resulting in (c).

Fig. 7. Example of a compound figure with different levels of indicative letters verbosity. From top to bottom, a compound figure without any indicative letter receives different levels of indicative verbosity, depicted by one of its panels; at the bottom, the result figures for each level.

To produce forgeries as realistic as possible, some functions from the library require as input an object map (segmentation map). The object map locates each object inside the image and assists a forgery function to execute the falsification more likely as a human would do.

B. Realistic Scientific Figures

As our key objective is to create scientific figures, we include two features (frequently present in such figures) in the library: captions/indicative letters and compound figures.

1) Caption/indicative letters: Scientific figures often present indicative letters or captions that overlay the image’s content. As a result, this overlay is a splicing operation between a letter or a word within the experiment image that could raise a false alarm during forgery detection. Therefore, we add to the library the possibility to mimic this overlap behavior as it appears on scientific papers. We include three different levels of indicative verbosity. Level 1 includes only indicative letters around each panel of the figure. Level 2 includes the features of Level 1 and a random word around each panel. Level 3 includes all features from Level 2 and an indicative letter inside each panel. Figure 7 depicts all these levels of verbosity.

2) Compound figures: A Compound figure is a composition of multiple images that are organized in panels. These figures usually appear in articles to represent an overview of an experiment. To avoid creating unrealistic compound figures, we make use of figure templates based on real cases. These templates are image masks that can inform each panel’s location to the method, as well as their type (e.g., graphs, photos).

To create Compound figures, we implemented a routine that has as input a set of realistic compound figures templates $T$, a dataset of scientific images $D$ (to be included in the
compound figure), a source image $S$ ($S$ is not in $D$), and a forgery function $f$ (to be applied in $S$). Figure 8 illustrates this routine. Thus, the method selects a template $t$ from $T$ with at least one panel whose aspect ratio is similar to the aspect ratio of $S$. Then, the routine applies the forgery function $f$ in $S$, creating $S_f$ (a forged version of $S$). Later, a figure with the same size of $t$ is created, and $S_f$ is resized and placed in the panel of $t$ with the most similar aspect ratio of $S_f$. Finally, all other panels of $t$ are filled with different images from $D$ that have similar aspect ratio to those panels or with fake graphics.

C. Data Annotation

In spite of the importance of reliable ground-truth to evaluate a forensic method, to the best of our knowledge, there is no scientific forgery image dataset that presents an enriched ground-truth. Hence, all tampered operations implemented in the library provide detailed maps to indicate the manipulated regions. Each object involved in the tampering operation is indicated with a different ID in the ground-truth, which helps pinpoint the object’s exact location before and after the forgery, as depicted in Figure 8. The library also enables the creation of a JSON file containing metadata related to the forgery. This metadata includes the source images, the method and arguments used, and the location of each panel inside the Compound figure. As the metadata includes the source images and the forgery methods applied, one can evaluate provenance analysis [17] using these information as reference.

D. Library Extension

Given that scientific image tampering improves over time to convince even researchers [15], the benchmark of tampering detection also should include cutting-edge forgery techniques. In this sense, to facilitate the inclusion of new manipulation in RSIL, we implemented a high-level routine that receives as one of its arguments a forgery function and applies it to an image. This routine is responsible for regulating the application of any new manipulation, asserting its guidelines to the ground-truth and the metadata associated with the forgery. Because of this, any new forgery function capable of returning the forged image along with its manipulation map—a pixel-wise map locating the forgery inside the image—can be easily added to the library to generate Simple or Compound tampering figures.

IV. RECOD SCIENTIFIC IMAGE INTEGRITY DATASET - RSIIID

In addition to the library we introduced in the previous section, we created a dataset to serve as future benchmark for the area. For that, we selected the most frequent retracted types of images from the biomedical area. For this, we followed the orientation of Bucci [18] and Bik et al. [4] that report a high image manipulation rate on images from Western Blot techniques and Microscopy imagery. With this in mind, we downloaded real scientific images collected from diverse sources to apply the forgeries.

To avoid any legal aspect of creating manipulated images and aiming to publish the dataset with a common creative license, we only downloaded data available under public domain (PD) or common creative attributed (CC-BY) licenses. These license allow us to remix, transform, and reuse the images without asking for the author’s authorization.

We use the following data source to gather the image collection:

1) **Broad Bioimage Benchmark Collection** [11] (BBBC), : A collection of freely downloadable microscopy image sets. From this source, we selected the datasets BBBC038, BBBC039, and BBBC019. The first two are dedicated to segmented nuclei images and have object-mask—that are needed for forgeries at object-level–. The last dataset (BBBC019) is dedicated to cell migration which we use for Overlap forgeries.

2) **PubMed Central (PMC)** [12]

PMC is a free article archive of biomedical and life sciences. To download each figure from this repository, we use an API available by PubMed, in which we could select images that only have PD or CC-BY licenses. We choose to include published western blots images. To include the western blots to the source dataset, after downloading the PMC figures, we manually extracted

---

9https://creativecommons.org/publicdomain/zero/1.0 (Last access May, 2021)
10https://creativecommons.org/ licenses/by/4.0 (Last access May, 2021)
11https://bbbc.broadinstitute.org (Last access May, 2021)
12https://www.ncbi.nlm.nih.gov/pmc (Last access May, 2021)
13https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist (Last access May, 2021)
the panels that had western blots for each figure. We based the templates images used for creating Compound Figures based real figures retrieved from this repository.

3) **TNBC [19]:**

This dataset, announced in Naylor et al. [19], was designed aiming at nuclei segmentation of cells by deep neural networks. Therefore, this dataset has a high-quality object map that we make use of to assists the forgery operations at object-level.

Table I shows the number of source figures for each collection by type (Microscopy or Western blot) pointing if they have object-mask annotation, and the number of template images created based on the figures from PMC.

### A. Dataset Construction

While designing the dataset, we project it to evaluate a forensic tool in different tasks with different complexities. Because of this, the dataset is organized so that a user can easily find the data and its annotation for each forgery operation in different tasks with different complexities. Thus, we divided the dataset into two types of figure complexities: Simple and Compound.

1) **Simple Scientific Figures**: Figures with this complexity are represented by a single experiment image, (e.g., Figure 2). To include a tampering figure in this complexity type, we forge an original figure using Retouching, Cleaning, or Duplication techniques, implemented in our library. To avoid unexpressive forgeries, we only included doctored figures in the dataset that have at least 500 manipulated pixels. In addition to the tampered figures, we also reserved a pristine directory in which we include the original images, so that a user can easily evaluate false positives. Figure 3 illustrates the organization of Simple figures in our dataset.

2) **Compound Scientific Figures**: This complexity type is described in Section III-B and depicted in Figure 8. We divided the Compound Figures into two types of tampering: Intra-Panels and Inter-Panels.

   a) **Intra-Panels** are forgeries that are present in just one panel of the figure. To create this tampering type, we add a Simple forgery as one of the figure’s panels. Forgeries that need more than one source image (e.g., splicing) or that generate more than one doctored figure (e.g., overlap) were not included in this modality.

   b) **Inter-Panels** are forgeries that have two or more panels involved in the manipulation process. This modality aims to evaluate duplications among two or more panels within the same figure. These duplications can be object-level, region-level, or panel-level. At object-level, the objects from a donor panel are copied into a host, using splicing operation. At region-level, an overlap forgery operation creates two panels with overlapping areas. At panel-level, the entire panel is duplicated with or without post-processing (e.g., retouching, cleaning, or geometric transformations).

For each Compound figure, we generated the three levels of indicative letters verbosity, as described in Section III-B. Figure 10 illustrates the organization of Compound figures in our dataset.

To assist machine learning forensics techniques, we further divided the dataset into training/test sets. Tables II and III express the number of manipulated figures included in each modality. Note that, overlap forgery appears only in the test set, since this modality is similar to the copy-move, and this protocol will force the generalizability of a forensic tool among the methods.

### V. COPY-MOVE FORGERY DETECTION PROPOSED METRIC

Popular metrics used on Copy-Move Forgery Detection (CMFD) (e.g., F1-score and Precision) make use of True Pos-

---

**Table I:** RSSID Source Image and Template Collection

| Collection | Microscopy | Western Blot | Object-Mask |
|------------|------------|--------------|-------------|
| BBCC019    | 809        | 0            | ✓           |
| BBCC038    | 592        | 0            | ✓           |
| BBCC019    | 165        | 0            | ✓           |
| TNBC       | 50         | 0            | ✓           |
| PMC        | 382        | 1,069        | ✓           |

* Templates created based on the figures from the collection.

**Table II:** Number of Simple Figures per modality in the dataset

| Modality      | Train Number of Figures | Test Number of Figures |
|---------------|-------------------------|------------------------|
| Source of Forgeries |                        |                        |
| Simple        |                        |                        |
| Pristine      | 1,932                  | 991                    |
| Duplication   |                        |                        |
| Copy-Move     | 3,761                  | 1,629                  |
| Splicing      | 604                    | 274                    |
| Total         | 4,365                  | 2,903                  |
| Cleaning      |                        |                        |
| Retouching    |                        |                        |
| Inter-Panels  |                        |                        |
| Brute-force   | 961                    | 412                    |
| Total         | 1,232                  | 529                    |
| Total         | 1,927                  | 829                    |

**Table III:** Number of Compound Figures per modality in the dataset

| Modality      | Train Number of Figures | Test Number of Figures |
|---------------|-------------------------|------------------------|
| Source of Forgeries |                        |                        |
| Compound      |                        |                        |
| Inter-Panels  |                        |                        |
| Duplication   |                        |                        |
| Copy-Move     | 9,316                  | 4,094                  |
| Splicing      | 904                    | 274                    |
| Total         | 10,120                 | 4,368                  |
| Cleaning      |                        |                        |
| Inter-Panels  |                        |                        |
| Duplication   |                        |                        |
| Copy-Move     | 3,701                  | 1,629                  |
| Splicing      | 904                    | 274                    |
| Total         | 4,605                  | 1,903                  |
| Cleaning      |                        |                        |
| Inter-Panels  |                        |                        |
| Retouching    |                        |                        |
| Inter-Panels  |                        |                        |
| Duplication   |                        |                        |
| Copy-Move     | 904                    | 274                    |
| Splicing      | 904                    | 274                    |
| Total         | 1,808                  | 549                    |
| Cleaning      |                        |                        |
| Inter-Panels  |                        |                        |
| Total         | 1,927                  | 829                    |

**TABLE II**

| Source of Forgeries | Train Number of Figures | Test Number of Figures |
|---------------------|-------------------------|------------------------|
| Simple              |                         |                        |
| Pristine            | 1,932                   | 991                    |
| Duplication         |                         |                        |
| Copy-Move           | 3,761                   | 1,629                  |
| Splicing            | 604                     | 274                    |
| Total               | 4,365                   | 2,903                  |
| Cleaning            |                         |                        |
| Retouching          |                         |                        |
| Intra-Panels        |                         |                        |
| Duplication         |                         |                        |
| Copy-Move           | 3,761                   | 1,629                  |
| Splicing            | 604                     | 274                    |
| Total               | 4,365                   | 2,903                  |
| Cleaning            |                         |                        |
| Inter-Panels        |                         |                        |
| Duplication         |                         |                        |
| Copy-Move           | 3,701                   | 1,629                  |
| Splicing            | 904                     | 274                    |
| Total               | 4,605                   | 1,903                  |
| Cleaning            |                         |                        |
| Inter-Panels        |                         |                        |
| Duplication         |                         |                        |
| Copy-Move           | 904                     | 274                    |
| Splicing            | 904                     | 274                    |
| Total               | 1,808                   | 549                    |
| Cleaning            |                         |                        |
| Inter-Panels        |                         |                        |
| Total               | 1,927                   | 829                    |
detection in which the copied objects with \( id = 1 \) and \( id = 2 \) are inconsistent with the ground-truth. For both objects, only the source or its copy overlaps with the ground-truth, which is inconsistent, since both (object and its copy) are expected to be included during detection. However, when evaluating this detection map with traditional true positive score, both regions would be considered as true positive hit.

To mitigate this drawback, this section introduces a new metric that takes advantage of the enriched pixel-wise ground-truth of the dataset. The proposed metric is a variation on how to consider a pixel as true positive in a detection map named as the Consistent True Positive score (CTP) and defined as:

Given a ground-truth map \( GM \) with \( n \geq 1 \) manipulated regions, a detection map \( DM \) with \( m \geq 0 \) copy-pasted regions, each one of the \( n \) regions included in \( GM \) has \( c_{gm} \geq 2 \) connected components (the source object and all its copies), and each region in \( DM \) has \( c_{dm} \geq 1 \) connected components. Let \( R_{dm} \) be a detected region from \( DM \) and \( R_{gm} \) a tampered region indicated by the ground-truth. Also, let \( p \) be a pixel from \( DM \), such that \( p \in R_{dm} \). Thus, \( p \) is a consistent true positive if exists \( R_{gm} \in GM \), such that, at least two connected components from \( R_{gm} \) intersects \( R_{dm} \).

In other words, to consider a region \( R_{dm} \) from the detection map as a consistent true positive, at least two components from \( R_{dm} \) (the source and at least one of its copies) have to intersect the ground-truth.

As Figure 12 depicts, a region from the detection map can overlap with two or more region from the ground-truth. Given that the goal of \( CTP \) is consistency, we only consider as \( CTP \) the region of the ground-truth that has the maximum intersection area with the detected region. Hence, \( CTP \leq TP \). As a result, \( FN \) will have higher penalty on Precision and \( F1 \)-score metrics, if calculated with \( CTP \).

Thus, the equation of \( F1 \)-score and Precision using \( CTP \) become:

\[
F1\text{-score}_{CTP} = \frac{2CTP}{2CTP + FN + FP} \quad (3)
\]

\[
\text{Precision}_{CTP} = \frac{CTP}{CTP + FP} \quad (4)
\]

VI. Evaluating CMFD Methods

Duplication of scientific images is one of the threats highly reported and studied in the literature \[4], \[2], \[20\] which includes copy-move, a well-studied forgery in the digital forensic field. Although this field presents multiples CMFD solutions for natural images, we could not find any study that evaluates their performance in the scientific image domain. In this sense, to assist any future forensic method with a baseline, we investigated the performance of popular CMFD solutions on natural images applied to the RSIID. In addition, we checked the difference of \( F1 \)-score using the proposed consistent true positive metric and the regular true positive one. For this, we choose the following CMFD methods:

1) Efficient Dense-Field from Cozzolino et al. \[7\]. During the evaluation, we use the implementation of Ehret \[21\]. Ehret released two versions of \[7\] using Zernike
and SIFT features. To distinguish this method from the others, we named them Zernike-PM and SIFT-PM, since these detectors use the PatchMatch algorithm \cite{22} to match similar blocks contents.

2) CMFD library implemented by Christlein et al. \cite{8}. We selected SIFT and SURF methods from this library since the others were not efficient enough to be explored on such a large dataset. To distinguish them from the previous CMFD detectors, we named them SIFT-NN and SURF-NN since they use a regular approximate nearest-neighbor approach to match similar blocks.

3) Busternet from Wu et al. \cite{9}. This method is a deep neural architecture for CMFD. During the evaluation, we use the pre-trained version of the model released by Wu et al. \cite{9}.

To evaluate SIFT-PM, Zernike-PM, SIFT-NN, and SURF-NN using CTP, we modified their implementation, including a routine that assigns each detected object and its copies a unique ID. For the sake of reproducibility, we released the methods source-code with this modification in the same repository of RSIIL. On the other hand, to evaluate Busternet, we normalized its output [0,255], then binarized all pixels greater than 100 to 1, otherwise 0. As Busternet is based on neural networks, we could not find an explainable methodology that would track the matching among different objects and their copies. Thus, to the CTP metric, all detected and ground-truth objects are set with the same ID = 1. Consequently, CTP would not be able to properly check inconsistencies on figures with more than one tampered object for Busternet’s output; however, CTP is still valid and useful to check if Busternet’s output overlaps with two or more connected components from the ground-truth.

As a baseline approach, the evaluation protocol consists of running all methods without any training or fine-tuning and measuring their output with F1-score\textsubscript{CTP}. During the evaluation, we use all figures from the test set applicable for CMFD (i.e., images with duplicated areas within the same image). We group the baseline results into Simple and Compound scientific figures, which were divided by modalities. All copy-move modalities presented in Figures\textsuperscript{13} and \textsuperscript{14} can also include scaling. Since scaling cannot be applied alone, we did not indicate when this operation is combined with others.

Figure\textsuperscript{13} presents a radar graph visualization in which the forgery modalities are arranged in the radius axes. Each CMFD methods’ result is represented with a different color in the radar chart. In this visualization, we insert the score of each method along the modality axis (e.g., copy-move with flip) which start from the radar center (score zero) until its border (highest score); thus, as farther a method point (color point) is from the center as better is the method for the axis copy-move modality. After inserting all points of a method for each copy-move modality, we connected those points resulting in a polygon. The larger the polygon area, the better the method’s robustness among different forgery modalities. Also, comparing each detector’s robustness to the operations, this type of visualization helps identifying possible complementary behaviors among different methods. As an example, consider Figure\textsuperscript{13} left panel. In this case, we have five modalities being compared (e.g., Copy-Move with Flip, Cleaning with Brute-Force, and Copy-Move with Translation). This chart shows the results of five methods represented by different polygon color (see legend on the right of the figure). The best method in this figure is Busternet (in orange) while the two worse methods (SURF-NN and Zernike-PM) are in superposition at the center (smaller areas). In the following, we discuss the forgery evaluation for each modality.

A. Simple Figure Forgery Baseline

In this modality, we tested the chosen methods on Simple figures, forged with Cleaning (Brute-Force) and Copy-Move. Although the chosen CMFD detectors have high efficacy on natural image benchmarks, their performance drastically decrease when applied to our scientific dataset. As Figure
(a) CMFD Simple Figure Evaluation. The best method is Busternet because its polygon contains all the others, indicating that it performs better for all operations. The shrinking polygons area from $F1$-score$_{TP}$ to $F1$-score$_{CTP}$ indicates that all methods show inconsistencies in their detection map.

(b) CMFD Inter-Panel Figure Evaluation on different levels of indicative letters verbosity. In this plot, each method fares differently for each modality. The polygons from SIFT-NN and SURF-NN have larger area than the others methods, indicating that they are robust to more operations than the others methods. The shrinking polygons area from $F1$-score$_{CTP}(V1)$ to $F1$-score$_{CTP}(V3)$ indicates the higher the caption Level, the lower is the method’s effectiveness.

(c) CMFD Inter-Panel Figure Evaluation on different levels of indicative letters verbosity. All methods show low performance and concentrate in the center of the radar, indicating that this is a challenging modality.

Fig. 13. Evaluation Baseline Results. Inside the parenthesis of each copy-move modality, there is the transformation used during the copy-move forgery. All F1-scores presented in this figure are normalized $[0, 100]$. The best result for each duplication modality is indicated with the color of the respective detector. 

(a) Result for Single Figure Evaluation using $F1$-score$_{TP}$ and $F1$-score$_{CTP}$. (b) Result for Inter-Panel Figure Evaluation using $F1$-score$_{CTP}$ across all levels of indicative letters verbosity, indicated by the number in its subtitle (i.e., V1 for verbosity Level 1). (c) Result for Intra-Panel Figure Evaluation using $F1$-score$_{CTP}$ across all levels of indicative letters verbosity.
Fig. 14. Comparative duplication detection output per modality. The purple color represents a pristine/non-suspect region, and each other color in the ground-truth and detection maps represents a different ID assigned to each object and its copies. Inside the parenthesis of each method, we insert the F1-score\textsubscript{CTP} metric normalized into [0, 100]. All compound figures are with level 1 of indicative letters verbosity.

\textsuperscript{13a} shows, the best CMFD method in the Simple Figure Evaluation was Busternet \cite{9}, despite its modest scores.

For this modality, we also compare each methods’ performance between F1-score using \( TP \) and \( CTP \). We notice a difference in these scores for all methods, represented by the area reduction from their polygon chars, indicating the existence of copy-move inconsistencies on their detection maps, which is also depicted in Figure \textsuperscript{14}. The second row of this figure shows an example of an inconsistent detection map, in which Busternet activates just one of the connected components involved in the manipulation, resulting in F1-score\textsubscript{CTP} = 0. On the other hand, in the same row, SIFT-NN detects both regions (object and copy), resulting in F1-score\textsubscript{CTP} = 81.51.

\textbf{B. Inter-Panel Figure Baseline}

We evaluate the \textit{Inter-Panel} tampered figures for all indicative verbosity levels in Copy-Move (at panel-level), Splicing, and Overlap forgeries. Figure \textsuperscript{13b} shows the result for the \textit{Inter-Panel} forgery evaluation using F1-score\textsubscript{CTP}. In this modality, the radar visualization allowed us to notice some complementary performance among the chosen detectors. For instance, SURF-NN and Zernike-PM show a complementary behavior to copy-move with rotation and retouching. We believe that this complementary aspect indicates that
a fusion/ensemble technique might enhance their individual robustness. For the Inter-Panel scenario, the flipped copy-move, Splicing, and Overlap forgery showed to be the most challenging forgeries. In addition, the indicative letters are shown to have a perceptible impact in this scenario, reducing by up to seven points from level 1 to level 3 for some detectors. Although Busternet achieves the best performance in Simple Figure modality, when applied to compound figures, it leads to a higher false-positive rate, as depicted in Figure 14 by activating the entire image.

We also noticed that graphs and indicative letters are the most common causes of false alarms in the Compound Figures scenario, as illustrated by the third, fourth, and fifth rows of Figure 14 which SIFT-PM wrongly activates letters and graphs regions.

These findings help us to see where researchers should focus on when dealing with the scientific image forgery detection problem.

C. Intra-Panel Figure Baseline

We evaluate the Intra-Panel tampered figures for all levels of indicative verbosity in Cleaning (Brute-Force) and Copy-Move (at object-level).

As presented by Figure 13, this is the most challenging scenario, in which the detectors scored lower than four on F1-score CTP for all evaluated operations. A possible explanation for this is the lower percentage number of doctored pixels in these figures than in other modalities. The detectors’ low performance does not allow us to measure the impact of verbosity levels in the figures properly. However, as Figure 14 shows, graphs and indicative letters would also be one cause of false alarms in this modality.

VII. Conclusions

In addition to the daunting scenario of fraud in science — due to the increase of image misconduct cases—, there is a legal issue related to copyrights and judicial aspects that prevents one from creating a large collection of fraudulent scientific images, even for an in-depth forensic study to benchmark and drive the development of appropriate detection methods.

Therefore, this work introduced a library and a dataset to assist the scientific integrity and forensic community to overcome this legal hurdle. We believe that by presenting a large dataset to the forensic community, we are fostering the development of more complex and robust detection tools (e.g., AI-based models).

The proposed library implements the most common image manipulation forgeries described by scientific integrity researchers. Also, it is extendable to more complex tampering operations. As a special feature, the library generates an enriched ground-truth addressing all regions affected before and after applying a tampering function, assigning a unique ID for the regions involved (when applicable). Using this library on creative common scientific images, we created a dataset with 39,423 manipulated figures freely available.

Leveraging the dataset’s enriched ground-truth, we proposed a metric that avoids inconsistent detection during CMFD evaluation. Using this metric, we evaluate popular CMFD methods on our dataset. Although we choose high-cited and effective CMFD tools for natural images, all solutions presented a lower performance when transferred to the scientific image domain. This is not a fault of such algorithm as they were not designed for this specific setup. However, these findings show an important lack of methods and a tremendous research opportunity for new specialized detectors aiming at finding forgeries in scientific-related images. In addition, we notice that some of the chosen algorithms present complementary performance and might benefit from a fusion approach.

Notwithstanding the large size and diversity of the proposed dataset, we believe that science will report more sophisticated tampering operations in the near future, as warned by [15]. Thus, we are also concerned about more issue-less and freely available scientific integrity datasets with complex, enhanced, and realistic tampering modalities, aiding the design of more robust detectors.

Therefore, as future work, in addition to investigating robust forensic solutions using AI-based or fusion-based methods, we believe that studies on automated realistic scientific forgeries would also assist the forensic community in fighting scientific misconduct. Furthermore, we believe that the detailed pixel-wise ground-truth of RSIIID opens a research opportunity to explore eXplainable AI solutions that might assist analysts on sensitive cases, such as misconduct investigations.

ACKNOWLEDGMENTS

This research was supported by São Paulo Research Foundation (FAPESP), under the thematic project DéjàVu, grants 2017/12646-3 and 2020/02211-2.

REFERENCES

[1] J. Krueger, “Forensic examination of questioned scientific images,” Accountability in Research, vol. 9, no. 2, pp. 105–125, Apr. 2002. [Online]. Available: https://doi.org/10.1080/08989620212970
[2] M. Rossner and K. Yamada, “What’s in a picture? The temptation of image manipulation,” The Journal of Cell Biology, vol. 166, no. 1, pp. 11–15, 2004.
[3] N. Gilbert, “Science journals crack down on image manipulation,” Nature, Oct. 2009. [Online]. Available: https://doi.org/10.1038/news.2009.991
[4] E. Bik, A. Casadevall, and F. Fang, “The prevalence of inappropriate image duplication,” Biomedical Research publications, vol. 7, no. 3, 2016.
[5] Q. Schiermeier, “German task force outraged by changes to science fraud report,” Nature, vol. 415, no. 6867, pp. 3–3, Jan. 2002. [Online]. Available: https://doi.org/10.1038/415003a
[6] D. Chawla, “A single ‘paper mill’ appears to have churned out 400 papers, sleuths find.” Science, Feb. 2020. [Online]. Available: https://doi.org/10.1126/science.abd4930
[7] D. Cozzolino, G. Poggi, and L. Verdoliva, “Efficient dense-field copy-move forgery detection,” IEEE Transactions on Information Forensics and Security, vol. 10, no. 11, pp. 2284–2297, 2015.
[8] V. Christlein, C. Riess, J. Jordan, C. Riess, and E. Angelopoulou, “An evaluation of popular copy-move forgery detection approaches,” IEEE Transactions on Information Forensics and Security, vol. 7, no. 6, pp. 1841–1854, Dec. 2012. [Online]. Available: https://doi.org/10.1109/tifs.2012.2218397
[9] Y. Wu, W. AbdAlmageed, and P. Natarajan, “Busternet: Detecting image copy-move forgery with source/target localization,” in European Conference on Computer Vision (ECCV). Springer, 2018.
[10] Z. Xiang and D. Acuna, “Scientific image tampering detection based on noise inconsistencies: A method and datasets,” arXiv preprint arXiv:2001.07799, 2020.

[11] T. E. Koker, S. S. Chintapalli, S. Wang, B. A. Talbot, D. Wainstock, M. Cicconet, and M. C. Walsh, “On identification and retrieval of near-duplicate biological images: A new dataset and protocol,” in International Conference on Pattern Recognition (ICPR), IEEE, 2021. [Online]. Available: https://ailb-web.ing.unimore.it/icpr/author/3517

[12] A. Marcus, “Pitt researchers sue journal for defamation following retraction,” Dec 2019. [Online]. Available: https://retractionwatch.com/2019/12/02/pitt-researchers-sue-journal-for-defamation-following-retraction/

[13] P. Mongeon and V. Larivière, “The collective consequences of scientific fraud: An analysis of biomedical research,” in Proceedings of ISSI 2013: Proceedings of the International Conference on Scientometrics and Informetrics. Vienna: Austrian Institute of Technology, 2013, pp. 1897–1899.

[14] C. Qi, J. Zhang, and P. Luo, “Emerging concern of scientific fraud: Deep learning and image manipulation,” bioRxiv, 2020.

[15] A. Criminisi, P. Pérez, and K. Toyama, “Region filling and object removal by exemplar-based image inpainting,” IEEE Transactions on image processing, vol. 13, no. 9, pp. 1200–1212, 2004.

[16] A. Marcus, “Pitt researchers sue journal for defamation following retraction,” Dec 2019. [Online]. Available: https://retractionwatch.com/2019/12/02/pitt-researchers-sue-journal-for-defamation-following-retraction/

[17] D. Moreira, A. Bharati, J. Brogan, A. Pinto, M. Parowski, K. Bowyer, P. Flynn, A. Rocha, and W. Scheirer, “Image provenance analysis at scale,” IEEE Transactions on Image Processing, vol. 27, no. 12, pp. 6109–6123, 2018.

[18] E. Bucci, “Automatic detection of image manipulations in the biomedical literature,” Nature Cell death & disease, vol. 9, no. 3, p. 400, 2018.

[19] P. Naylor, M. Lae, F. Reyal, and T. Walter, “Nuclei segmentation in histopathology images using deep neural networks,” in 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017). IEEE, Apr. 2017. [Online]. Available: https://doi.org/10.1109/isbi.2017.7950669

[20] M. Wjst, “Scientific integrity is threatened by image duplications,” American Journal of Respiratory Cell and Molecular Biology, vol. 64, no. 2, pp. 271–272, Feb. 2021. [Online]. Available: https://doi.org/10.1165/rcmb.2020-0419le

[21] T. Ehret, “Automatic detection of internal copy-move forgeries in images,” Image Processing On Line, vol. 8, pp. 167–191, Jul. 2018. [Online]. Available: https://doi.org/10.5201/ipol.2018.213

[22] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman, “Patchmatch: A randomized correspondence algorithm for structural image editing,” in ACM Transactions on Graphics (ToG), vol. 28, no. 3, 2009, p. 24.