ROBUST SOURCE ATTRIBUTION OF SYNTHETICALLY GENERATED WESTERN BLOT IMAGES

A PREPRINT

Matthew A. Hyatt
Department of Computer Science
Loyola University Chicago
Chicago, IL
mhyatt@luc.edu

George K. Thiruvathukal
Department of Computer Science
Loyola University Chicago
Chicago, IL
gkt@cs.luc.edu

Daniel Moreira
Department of Computer Science
Loyola University Chicago
Chicago, IL
dmoreira1@luc.edu

November 29, 2023

ABSTRACT

Retracted papers commonly include manipulations of images and figures that are unfit for publication. While some manipulations are benign, like increasing contrast or zooming in, others are designed to fool the intended audience. Recent improvements in generative computer vision pose a security risk to scientific review since generated images are often indistinguishable from authentic bioscience evidence; even to field experts. In this work, we improve upon previous attempts to detect synthetic images by attending to their differences in the frequency domain. Additionally, we solve the multi-class classification of synthetic Western blots and attribute Western blot images to their respective generative model architecture. We demonstrate that our method outperforms previous methods for synthetic Western blot detection; including efforts to classify JPEG compressed images.

Keywords Computer Vision · Media Forensics · Western Blots · Generative Adversarial Networks

1 Introduction

Western blots are the outcome of laboratory techniques used in Biology-related fields to detect and visualize targeted proteins within a given tissue sample. Their image representations in the form of lane-aligned stains (a.k.a. the *blots*) on the top of uniform-looking substrates have been largely used in scientific manuscripts to relatively quantify the presence of proteins. In 2020, Moritz (1) estimated that 8–9% of protein-related publications leveraged Western blots to support their claims over the past three decades.

While Western blot images are indeed useful to diagnose diseases and attest to the effectiveness of substances or treatments by visually depicting the presence of proteins, their simple and elusive appearance makes them easy targets for improper reuse and manipulation (2). With simple editing tools (e.g., Photoshop), stains can be easily replicated from one lane to another, or hidden by cloning substrate portions over them. To make things worse, Mandelli et al. (3) have recently demonstrated the possibility of generating visually convincing synthetic Western blots out of noise by employing methods such as Generative Adversarial Networks (GANs) (4). They have even demonstrated that some GANs can be driven by *template masks*, which guide where the synthetic stains should be placed. This scenario asks for the quick development of solutions to mitigate the fabrication of synthetic Western blots.

In this manuscript, we extend upon Mandelli et al.'s and other works, moving from the *detection* of synthetically generated Western blot images to their *robust attribution*. While detection comprises the decision problem of classifying Western blot images as either genuine or synthetic, attribution takes one step further by recognizing what network was used to generate a given synthetic image. From the forensic science standpoint, source attribution is fundamental in the process of gathering evidence that links culprit and wrongdoing in an eventual formal investigation (5). Moreover, we want to perform *robust* attribution, in the sense that the herein proposed method can deal with the typical image

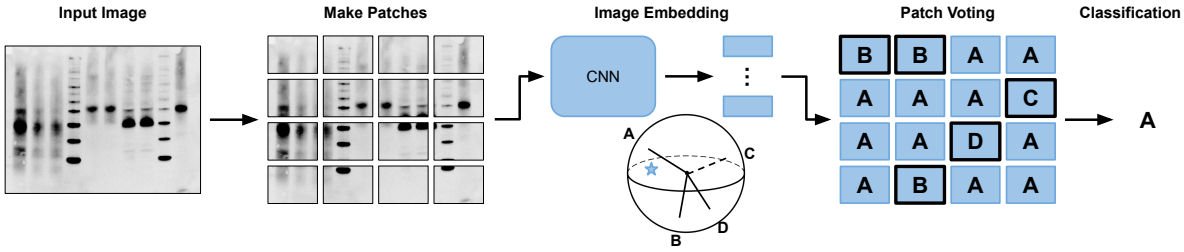

Figure 1: Proposed Western blot attribution pipeline for image patches. Following the dataset collected in (3), source attribution involves patch separation, image embedding, patch voting, and finally classification.

post-processing operations such as lossy compression. These operations are relevant because they frequently occur whenever an image is published online or embedded into the data stream of scientific manuscripts.

Fig. 1 summarizes how the herein proposed method works to attribute the source network of synthetically generated Western blot images. The images in question are first divided into patches. Each patch is then processed by a Fast Fourier Convolutional Network (FFC-ResNet) (6), whose aim is to describe the content of the patches with a focus on their texture. After that, we employ the cross-entropy loss function to learn patch-wise feature embeddings that represent the likelihood of an image belonging to one of four source model architectures.

Experimental results over the benchmark introduced by Mandelli et al. (3) show that the proposed solution advances the state of the art of the detection task. Additionally, it performs the attribution task with an average accuracy of 96.03% in the face of genuine and synthetic images obtained from four different generation techniques. The proposed method is performant even in the face of post-processing lossy compression at diverse rates.

In summary, this manuscript provides the following contributions:

- We improve the performance of synthetic image detection.
- To the best of our knowledge, we provide the first evaluation of closed-set source attribution of Western blot images.
- We discuss future directions to investigate solutions for open-set source attribution of synthetic images.

The remainder of this manuscript has four sections. In Sec. 2, we explain the proposed method. In Sec. 3, we detail the configuration of the experiments, followed by Sec. 4, where we report the obtained results. Lastly, we discuss the lessons learned from the experiments and potential future work in Sec. 5.

2 Proposed Method

Synthetic image *source attribution* is the problem of identifying the most probable neural network used to generate a synthetic image, including the *none* option in the case of authentic images. Source attribution of synthetic Western blots is harder than the detection of synthetic natural scenes (such as landscapes and everyday objects) because Western blots have a less pixel-value variation and simpler structure. Namely, lighter gray rectangular lanes depict the substrate, while darker gray stains depict the blots. Additionally, source attribution is harder than binary classification of synthetic images because it requires a classifier to pick up on subtle patterns to characterize different synthetic image generators.

To accomplish the source attribution task, we propose a data-driven neural network-based solution. We build off of previous improvements made by Mandelli et al. (3) that begin by segmenting Western blot images into $(k \times k)$-pixel patches. We input these Western blot patches into an FFC-ResNet (6) to describe important patch features, represented by a single N-dimensional feature vector.

As discussed in (7; 8; 9; 10), synthetic images differ in their frequency spectrum, even when they look indistinguishable from real images in the spatial domain. In the case of Western blots, we show some spectral differences in Fig. 2 for the sake of illustration. We attend to this phenomenon by opting for a classifier that looks for patterns in spectral images. In this work, we selected FFC-ResNet because of its recent success in computer vision, as well as its ability to incorporate spectral features into the learned model.

FFC-ResNet (6) improves upon traditional ResNet architectures by allowing the network to attend to local and global information simultaneously. This minimizes problems with convolutional receptive fields while also allowing the model to generalize to patterns in the spectral domain. The authors do this by replacing convolutions in ResNet's residual

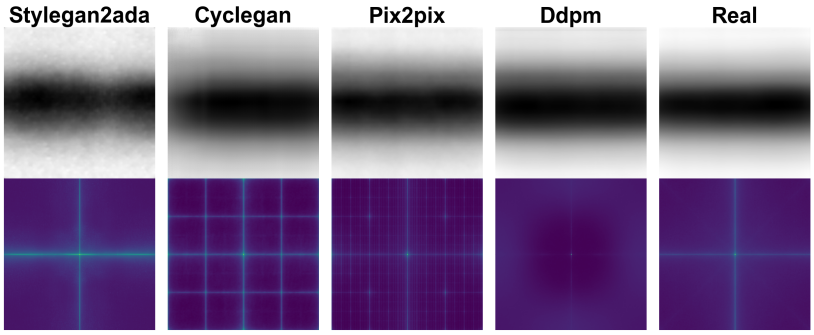

Figure 2: We show the average synthetic Western blot (above) and average spectral image (below) for each of four distinct generative model architectures as well as real Western blot patches. We follow (7; 12) by using a high pass filter to isolate noise. While Western blots look nearly indistinguishable, there exist patterns in the Fourier space that may indicate a signature left behind by the generative models.

block with their Fast Fourier Convolutions (FFC). Inspired by GoogLeNet (11), FFC works by maintaining four paths of information. Three of these paths use convolutional layers, while the fourth one (namely, the global) uses a Fast Fourier Transform (FFT) to produce a spectral image or spectral feature map and applies a learned convolution before transforming the feature map back to the spatial domain. The FFC block combines information from the local and global paths via pixel-wise summation and then channel-wise concatenation before passing the feature maps to the next block. In FFC-ResNet, FFC layers replace ResNet's usual 3x3 convolutional layers to compose a series of FFC blocks. For more details on this architecture, we refer the reader to Chi et al. (6).

Following traditional deep learning methods, we adopt a supervised learning regime, where the weights of FFC-ResNet are adjusted through backpropagation and optimization of a cross-entropy loss function. Cross-entropy loss is a multi-class loss function that produces a high penalty when the predicted categorical distribution does not match the target distribution. In optimizing a model to minimize this loss function, the model learns to predict a distribution similar to the target. Concretely, we optimize the probability that the target distribution reflects the generative neural network architecture used to create the synthetic image.

3 Experimental Setup

We compare the performance of various backbones in our experiments. Namely, we follow (10; 8; 13; 3) in using ResNet-based architectures. We compare the FFC-Resnet family from (6) with the methods from (3), namely SRNet (13) and a handcrafted classifier. SRNet was originally proposed in (13) for steganography and reused in (3; 8).

We train FFC-ResNet on the dataset of Western blot patches collected in (3). The dataset contains 14,200 images of *authentic* Western blot patches. The authentic patches were used to train four generative models: Pix2pix (14), CycleGAN (15), StyleGAN2Ada (16), and Denoising Diffusion Probabilistic Model (DDPM) (17). Each of these generative models, except for DDPM, is a GAN. The four trained models were each used to synthesize 6,000 images, resulting in a total dataset of 38,200 images.

Our model is trained for 5000 steps with a base learning rate of 0.04 and square-root learning rate scaling for multi-GPU training. We use a polynomial learning rate scheduler $p = 2$, and warm-up period equal to 10% of the training steps. We train models with both the Adam and LAMB optimizers and empirically find that LAMB converges faster – especially in the multi-GPU setting. The batch size is set to 128 images per GPU. We use four NVIDIA A100-SXM4-40GB GPUs in these experiments.

We follow (3) in testing our best method against images that have been perturbed by post-processing operations. We test the attribution robustness against:

- **Upscaling:** images are scaled up by 1.25 or 1.5 times their size, before being reduced back to their original size.
- **Downscaling:** images are scaled down by 50%, 75%, or 90% of their size. The remaining information is increased back to the original size.
- **JPEG compression:** images undergo JPEG compression to reduce the information to 70%, 80%, or 90% of their original size.

For fair comparison to previous works, we train our model on the same post-processing operation that we test it on. However, in a real-world scenario, the user may not know beforehand which operations were used to perturb the image. We also address this scenario in Sec. 4.2.2.

4 Results

We provide results for the tasks of synthetically generated Western blot detection (as a "real" versus "synthetic" binary decision problem) and closed-set multi-class source attribution.

4.1 Synthetic Western Blot Detection

Our FFC-ResNet solution performs comparably with the current state-of-the-art in this task. Table 1 summarizes the classification accuracy results reported in (3) for their binary classifier in the face of synthetic images coming from four different generative methods. To compare their method with ours, we simplified the output of our multi-class FFC-ResNet solution by paying attention to only the "real" class output. If this output presented the largest activation in the network forward propagation, we considered the fed image as a "real" one; otherwise, we considered it as "synthetic". We thus add to Table 1 the results of our FFC-ResNet solution following this setup.

	Mandelli et al. (3)	FFC-ResNet (ours)
Pix2pix (14)	92.58%	**98.77%**
CycleGAN (15)	88.99%	**100.00%**
StyleGAN2Ada (16)	88.93%	**99.65%**
DDPM (17)	94.69%	**100.00%**

Table 1: Real-versus-synthetic classification accuracy for Mandelli et al. (3) and FFC-ResNet (proposed).

Although our results seem to be superior, the solution proposed in (3) was not trained with samples collected from each one of the four generation methods but with a focus on the real class, in opposition to our approach. This may justify our superior results, in addition to the differences in classification method and network architectures.

4.2 Synthetic Western Blot Source Attribution

Table 2 summarizes the multi-class source attribution results of our FFC-ResNet solution over real and synthetic Western blots coming from the four generative methods. We report class-wise classification accuracy.

	FFC-ResNet
Pix2pix (14)	**93.83%**
CycleGAN (15)	**99.92%**
StyleGAN2Ada (16)	**90.43%**
DDPM (17)	**99.95%**
Real	**99.33%**

Table 2: Source attribution for FFC-ResNet (proposed). We report class-wise classification accuracies for each one of the four generative methods and the "real" class.

Our solution performs favorably in the multi-class source attribution task. In all scenarios, the model achieves above 90% classification accuracy. Our solution also recognizes authentic Western blot patches with an F1 score of 0.971, demonstrating that the model is not oversensitive to any particular generative technique.

In addition to the previous results, we also investigate the performance of FFC-ResNet in scenarios with post-processing operations, as described in Sec. 3.

4.2.1 Compression-Aware Source Attribution

During the compression-aware evaluation, we train eight FFC-ResNet-50 models on the source attribution task. Each model is trained to specialize in the source attribution of images perturbed by a specific compression algorithm. In Table 3, we report the class-wise classification accuracies for each one of these eight solutions, in the source attribution scenario.

	U-125	U-150	D-50	D-75	D-90	JPEG-80	JPEG-90	JPEG-100
Pix2pix (14)	99.83	99.66	99.66	99.94	99.72	78.91	76.16	99.83
CycleGAN (15)	99.78	99.78	99.73	99.62	99.67	97.87	94.0	98.91
StyleGAN2Ada (16)	99.54	99.6	99.31	99.71	99.19	89.39	90.43	95.27
DDPM (17)	100.0	100.0	100.0	100.0	99.95	99.31	99.63	99.95

Table 3: Multi-class source attribution of eight distinct compression-aware FFC-ResNet solution over post-processed images. Each model is trained to address a particular compression technique. U stands for Upscale, D stands for Downscale, JPEG implies JPEG Compression.

We see that in the compression-aware scenario, FFC-ResNet can classify upscaled and downscaled Western blots with above 99% accuracy. During JPEG-80 and JPEG-90 trials, which appear especially difficult, half of the source architectures are still correctly assigned with above 90% accuracy. It can be seen in these difficult cases that images from the Pix2pix architecture are the hardest to classify. We speculate that this is because Pix2pix is directly conditioned on template masks, which may help the model synthesize blots in near-optimal locations or shapes.

4.2.2 Compression-Agnostic Source Attribution

In opposition to the compress-aware approach, where we trained eight versions of FFC-ResNet to deal with distinct post-processing operations, we also try a compression-agnostic approach. During these trials, we aim to simulate a real-world source attribution scenario. We train one model on the classification task, where examined images are perturbed by one of eight compression algorithms – or the option of no-preprocessing – randomly selected with equal probability. Therefore, the same model weights are trained simultaneously to perform well under all eight compression scenarios. The model does not know beforehand which algorithm, if any, might be selected. We report the class-wise classification accuracies for the single trained FFC-ResNet solution in the face of each one of the investigated post-processing operations in Table 4.

	U-125	U-150	D-50	D-75	D-90	JPEG-80	JPEG-90	JPEG-100
Pix2pix (14)	58.27	46.78	58.22	64.78	66.74	**79.64**	**78.41**	61.69
CycleGAN (15)	92.03	86.74	95.36	95.52	96.02	**98.31**	**97.16**	96.78
StyleGAN2Ada (16)	81.15	76.83	86.05	86.11	85.71	**89.97**	88.01	86.46
DDPM (17)	99.89	99.89	99.84	99.79	99.79	**99.47**	**99.84**	99.84

Table 4: Multi-class source attribution of a single compression-agnostic FFC-ResNet over post-processed images. U stands for Upscale, D stands for Downscale, JPEG implies JPEG Compression. Scenarios that outperform the correlative compression-aware scenario are denoted in bold.

Interestingly, the compression-agnostic model observes reduced performance in all areas except JPEG-80 and JPEG-90 compression, which are the hardest tasks for the compression-aware models. Identifying the reason for this phenomenon will require further investigation. Images from Pix2pix are the hardest to classify by a wide margin and are consistent with findings from the compression-aware trials. In all aforementioned experiments, images from the DDPM model class are identified with over 99% accuracy. While there remains significant room for improvement, these results improve upon previous works.

5 Discussion and Conclusions

Our findings are preliminary and suggest the need for further investigation. Subsequent projects will investigate source attribution of full Western blot images. In this study we use Western blot patches, following (3), but using entire Western blot images is likely to represent a real-world scenario more closely. However, the proposed future study must address challenges associated with increased scale. Creation of the Western blot patches dataset was non-trivial in that it required Mandelli et al. (3) to train the four aforementioned model architectures to produce high-quality Western blot patches. Proper training of generative models requires a high number of authentic images. It is likely that splitting authentic Western blot images into patches allowed the authors to vastly increase the number of training samples – alleviating the burden of collecting several thousands of Western blots. We are currently investigating alternative methods of synthesizing high-quality full Westen blot images.

Additionally, it is important to design a solution that is robust to unseen generative models. As the field of generative deep learning advances, synthetic media detection algorithms become outdated. Our source attribution model operates in a closed-set scenario, making it vulnerable to Western blots images that have synthesized by unseen model architectures and even similar architectures that have been optimized in a different manner (18). We aim to study simulated open-set scenarios by learning a robust image embedding space and attributing images to the most probable model architecture occupying the same region of the learned embedding space.

In summary, in this work, we leveraged FFC-ResNet to get on par with the state-of-the-art in synthetic Western blot detection. To the best of our knowledge, we are the first to perform the multi-class classification task as source attribution of Western blot images. Additionally, we evaluated our model on images that have been corrupted by compression algorithms to show our model is robust to small amounts of corrupt data. Our findings motivate future studies on source attribution in order to promote transparent scientific practices.

References

[1] C. Moritz, "40 years western blotting: A scientific birthday toast," *Elsevier Journal of Proteomics*, vol. 212, pp. 1–4, 2020.

[2] A. Marcus and I. Oransky, "Can we trust western blots?," *Lab Times*, vol. 2, p. 41, 2012.

[3] S. Mandelli, D. Cozzolino, E. Cannas, J. Cardenuto, D. Moreira, P. Bestagini, W. Scheirer, A. Rocha, L. Verdoliva, S. Tubaro, and E. Delp, "Forensic analysis of synthetically generated western blot images," *IEEE Access*, vol. 10, pp. 59919–59932, 2022.

[4] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[5] Z. Li, Y. Liu, X. Hu, and G. Wang, "A new uniform framework of source attribution in forensic science," *Nature Humanities and Social Sciences Communications*, vol. 9, no. 1, pp. 1–11, 2022.

[6] L. Chi, B. Jiang, and Y. Mu, "Fast Fourier convolution," in *Conference and Workshop on Neural Information Processing Systems (NeurIPS)*, pp. 4479–4488, 2020.

[7] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "CNN-generated images are surprisingly easy to spot... for now," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8695–8704, 2020.

[8] D. Gragnaniello, D. Cozzolino, F. Marra, G. Poggi, and L. Verdoliva, "Are GAN generated images easy to detect? a critical analysis of the state-of-the-art," in *IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2021.

[9] F. Marra, D. Gragnaniello, D. Cozzolino, and L. Verdoliva, "Detection of gan-generated fake images over social networks," in *2018 IEEE conference on multimedia information processing and retrieval (MIPR)*, pp. 384–389, IEEE, 2018.

[10] X. Zhang, S. Karaman, and S.-F. Chang, "Detecting and simulating artifacts in gan fake images," in *2019 IEEE international workshop on information forensics and security (WIFS)*, pp. 1–6, IEEE, 2019.

[11] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

[12] F. Marra, D. Gragnaniello, L. Verdoliva, and G. Poggi, "Do GANs leave artificial fingerprints?," in *IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pp. 506–511, 2019.

[13] M. Boroumand, M. Chen, and J. Fridrich, "Deep residual network for steganalysis of digital images," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 5, pp. 1181–1193, 2018.

[14] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1125–1134, 2017.

[15] J.-Y. Zhu, T. Park, P. Isola, and A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *IEEE International Conference on Computer Vision (ICCV)*, pp. 2223–2232, 2017.

[16] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila, "Training generative adversarial networks with limited data," *Advances in neural information processing systems*, vol. 33, pp. 12104–12114, 2020.

[17] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[18] N. Yu, L. S. Davis, and M. Fritz, "Attributing fake images to GANs: Learning and analyzing GAN fingerprints," in *IEEE International Conference on Computer Vision (ICCV)*, pp. 7556–7566, 2019.