PLOS ONE

Synthetic images aid the recognition of human-made art forgeries

Johann Ostmeyer[1]☯, Ludovica Schaerf[2]☯, Pavel Buividovich[1]*, Tessa Charles[3], Eric Postma[4], Carina Popovici[2]

1 Department of Mathematical Sciences, University of Liverpool, Liverpool, United Kingdom, 2 Art Recognition AG, Adliswil, Switzerland, 3 Australian Synchrotron, Australian Nuclear Science and Technology Organisation (ANSTO), Clayton, Australia, 4 Cognitive Science & AI, Tilburg University, Tilburg, The Netherlands

☯ These authors contributed equally to this work.
* Pavel.Buividovich@liverpool.ac.uk

Abstract

Previous research has shown that Artificial Intelligence is capable of distinguishing between authentic paintings by a given artist and human-made forgeries with remarkable accuracy, provided sufficient training. However, with the limited amount of existing known forgeries, augmentation methods for forgery detection are highly desirable. In this work, we examine the potential of incorporating synthetic artworks into training datasets to enhance the performance of forgery detection. Our investigation focuses on paintings by Vincent van Gogh, for which we release the first dataset specialized for forgery detection. To reinforce our results, we conduct the same analyses on the artists Amedeo Modigliani and Raphael. We train a classifier to distinguish original artworks from forgeries. For this, we use human-made forgeries and imitations in the style of well-known artists and augment our training sets with images in a similar style generated by Stable Diffusion and StyleGAN. We find that the additional synthetic forgeries consistently improve the detection of human-made forgeries. In addition, we find that, in line with previous research, the inclusion of synthetic forgeries in the training also enables the detection of AI-generated forgeries, especially if created using a similar generator.

Introduction

Forgeries are a serious threat to the artwork market, as illustrated for instance by the infamous Max Ernst forgery "La Horde". In 2006, the auction house Christie's announced the sale of the artwork, with an estimated value of about £3,000,000. However, it turned out that "La Horde" was a forgery created by the art forger Wolfgang Beltracchi [1]. Similarly, at the beginning of the 20th century, the Wacker case made the headlines globally. The German art dealer Otto Wacker, possibly with the help of his brother Leonhard, managed to sell over 30 fake Van Gogh paintings to public and private collectors, and many of the paintings were even included in the Catalogue Raisonné by Van Gogh expert Jacob de la Faille [2]. Despite experts' disagreement, the art dealer was charged with fraud in April 1932.

Recent developments in computer vision and machine learning techniques may contribute to the issue in several ways [3].

Starting from late 1990s, various computer vision and image analysis techniques such as fractal analysis [4], wavelets [5–7], sparse coding [8], clustering-based segmentation [9] and tight frame method [10] were applied to extract characteristic features of individual artist's style automatically. More recently, the development of efficient classifier neural networks such as Convolutional Neural Networks (CNNs) allowed reaching very high accuracies in artwork attribution [11–16].

While most of the studies concentrate on the attribution of an artwork to several predefined authors, similar machine-learning methods can also be used to distinguish between authentic artworks by a given author and forgeries. Due to the very close resemblance between original images and human-made forgeries (such as the Wacker forgeries), art authentication is generally a more challenging task than artwork attribution. In particular, authentication algorithms often have to learn very fine details such as brushstroke structure [7, 9, 17]. An additional challenge is that, for a given artist, forgeries are typically much less numerous than original artworks and often lack systematic documentation and high-resolution scans or photos. Despite such limitations, in recent years, NNs such as Convolutional Neural Networks (CNNs) or transformer-based architectures have shown promising results in both art attribution, when trained on datasets of authentic paintings and other stylistically similar artworks [17] as well as in artwork authentication, trained against forgeries [18].

In this context, the new trend of Generative Artificial Intelligence (GenAI) appears to present both threats and opportunities. On the one hand, GenAI might be adopted to create refined synthetic digital forgeries [19], which might populate the internet and diffuse misinformation. The possibility of creating 'fake' synthetic artworks using AI-based methods gained popularity with the publication of Neural Style Transfer (NSF) [20], which learns to decouple the style of an artwork from its content. This method is capable of creating synthetically styled images in the particular predisposition of a given artist to varying scales of accuracy and applicability. The successive publication of various Generative Adversarial Network enhanced architectures (e.g. StyleGANs [21, 22]) and powerful large-scale diffusion models (e.g. Stable Diffusion [23] and DALL-E 2 [24]) paved the way to the generation of realistic synthetic forgeries. In particular, the introduction of text conditioning using contrastive text-image models such as CLIP [25], created an accessible and quick interface for the creation of artworks 'in the style of'. Differently from NSF, the latter is not tied to an input natural image and therefore allows greater freedom of generation.

At the same time, the ability of GenAI to create synthetic forgeries may mitigate the limitation of AI-based art authentication of being hampered by the limited availability of known forgeries and imitations. The goal of this work is to explore to what extent the recent GenAI methods such as StyleGANs and Stable Diffusion are able to augment the training datasets of known forgeries and enhance the performance of AI-based art authentication. While most recent proposals to detect fake images mainly address photorealistic images [26], the use of synthetic forgeries in artwork authentication is a widely unexplored area. Specifically, we focus on paintings by Vincent van Gogh, which are frequently used as a benchmark dataset for machine-based art attribution methods [6, 7, 9, 10]. Van Gogh painted a sheer amount of artworks, now in the public domain, and was widely forged due to its enormous market value. Van Gogh datasets, therefore, serve as valuable case studies for art authentication.

We build on the already publicly available dataset VGDB-2016 on Van Gogh [27] available here, containing a set of 126 RGB original images by the artist and a set of 212 non-authentic RGB images by other Impressionist and Expressionist artists. The VGDB-2016 dataset does not contain forgeries, making it unsuitable for forgery detection. To address this, we enrich it

for the purposes of art authentication and add 11 RGB images of well-known forgeries created by the forger Otto Wacker into our dataset. We also include 8 forgeries by former art forger and now legitimate artist creating *genuine fakes*, John Myatt. The latter images are not in the Open Domain, therefore we only provide a pointer to those images. Furthermore, we release the artificially generated AI-based forgeries specifically generated for this paper. Finally, to reinforce our findings on van Gogh, we carry out the same analysis on datasets of Amedeo Modigliani and Raffaello Sanzio (Raphael), which are detailed in the supporting information S1 Appendix.

The outline of this paper is as follows. The next section details the Methodology we employ to generate synthetic images used to augment the training data set of known forgeries. We also briefly discuss the classifier model that we use for forgery detection. We then present our main findings on improved Classification methodology, leading up to the goals listed above. As a consistency check, we also discuss the authentication of synthetic forgeries created by Stable Diffusion and StyleGANs Detection of synthetic images. A brief summary is provided in the Discussion and Conclusions.

Methodology

In this section, we provide an overview of the methods we employed to generate synthetic images for art authentication. We first outline the process of creating synthetic images for the training datasets and provide details about the composition of the dataset. We also elaborate on our classification methodology for art authentication. Finally, we explain how we evaluate the authentication efficiency.

Methods for synthetic image generation

We use two fundamentally different GenAI methods to generate synthetic artwork: an image-to-image generative adversarial network (GAN) and a text-to-image diffusion model. The images generated both by the diffusion model and GAN are collectively referred to as synthetic data.

We used the NVlabs implementation `StyleGAN3` [22] which is one of the most recent and successful GANs. `StyleGAN3` was trained from scratch on 10380 portraits in various genres and by many different authors, including 126 portraits by van Gogh, 280 by Modigliani, and 157 by Raphael. The portraits by the three artists are sourced from Wikiart, and while there is considerable overlap, they do not entirely represent the sets of original artworks detailed in the subsequent datasets. The latter ones are not limited to portraits but, in turn, are filtered to include only artworks appearing in museum collections or Catalogue Raisonnés, ensuring a high level of certainty regarding their authenticity.

The training took 5M epochs on 4 GPUs. More details on the training procedure and the quality of the resulting images can be found in the supporting information S1 Appendix. With such training, `StyleGAN3` produces images in a mixture of styles by random authors. We used the trained `StyleGAN3` to produce a "raw" dataset of 2000 random portraits. In what follows, images picked at random from this "raw" dataset are referred to as the "raw GANs" image set. Furthermore, subsets of synthetic portrait images in the style of a specific artist (van Gogh, Modigliani, or Raphael) were created through further training for 50k epochs exclusively on original paintings by the respective artist. This yielded image sets of synthetic images that looked stylistically close to the works of van Gogh, Modigliani, and Raphael. We refer to these datasets as "tuned GANs" image sets. We remark that, due to the limited number of paintings available for each artist, prolonged training on the exclusive data sets often results in a decline in the quality of the `StyleGAN3` images. Rather than achieving the desired outcome

of generating a large variety of images in a given style, long specialized training tends to produce an almost exact reproduction of the training set. Specialized training for some time between 20k and 100k epochs has proven to be a good compromise, striking a balance between a wide variety of images and effective adaption of the desired style.

To create the text-to-image synthetic artworks, we use the Stable Diffusion [23] generative model. It relies on CLIP guidance [25] to semantically align the latent text representation and the latent image representation and a U-Net architecture [28] as a de-noising diffusion model. The quality of images generated using Stable Diffusion strongly depends on the text prompt. We generate images in the style of each artist using a simple prompt indicating the style, the content, and the artist; for example: 'Post-impressionist painting of a young boy, by Vincent van Gogh'. We adopt the Stable Diffusion version 2.1 (`v2-1_768-ema-pruned.ckpt`), with 60 inference steps, 8 guidance scale, and 512 × 512 pixels resolution. The resulting synthetic dataset is referred to as "diffusion". Note that Stable Diffusion has been trained on subsets of the very broad open-source dataset `LAION-2B(en)` collected *in the wild* and using the large contrastive model OpenCLIP while the GAN was trained on the controlled `WikiArt` dataset. We used the second version of Stable Diffusion because it is trained on fully open data and models.

Composition of training datasets

AI-based art authentication is a binary classification task where the model learns to differentiate between authentic and non-authentic artworks (including known forgeries). This requires training on two sets of artworks for each artist, an authentic and a contrast set. Our experiments are centered around the van Gogh dataset which contains 126 original artworks from the VGDB-2016 dataset [27]. The dataset was gathered from Wikimedia Commons, and it contains artworks with a similar chronology or artistic movement to van Gogh and with a density of at least 196.3 PPI (Pixels Per Image), the dataset also contains two artworks with debated attribution for testing. Here we note that the number of images does not exactly match those mentioned in the original paper [27], we provide the number of images that were actually downloaded through the dataset link provided in [27].

In addition, in supporting information S1 Appendix we provide two further tests of our approach on the artworks by Modigliani (100 original artworks) and Raphael (206 original artworks) and imitations/forgeries thereof. The latter datasets were collected from museum collections or sourced from *Catalogue Raisonnées*, which are expert-curated lists documenting all verified authentic artworks by the respective artists.

The contrast set includes artworks that were not made by the artist, but that resemble it closely and are helpful in detecting forgeries of the artist's work. Normally, this includes artworks of similar artists, referred to as 'proxies', and forgeries or explicit imitations of the artist, referred to as 'imitations'. Proxies are paintings by different human authors who painted in a similar style to the artist (i.e. artists pertaining to the same artistic movement) and/or were collaborators, pupils, and teachers. The word imitation is used here as an umbrella term to encompass human-made non-autograph copies of authentic works, artworks explicitly made in the style of the artist, and known forgeries of the artist. To these elements, we add synthetic fakes generated by Stable Diffusion 2.1 and StyleGAN3, and we test whether their addition increases the performance of the models.

The contrast set of Vincent van Gogh contains 212 artworks by similar artists, 19 forgeries (11 by Otto Wacker and 8 by John Myatt), 30 Stable Diffusion generated images, 30 GANs fine-tuned on the artist, and 30 random GANs (the 'raw GANs'). The set of 'raw GANs' contains the same exact images across all three datasets used in this work.

Table 1. Composition of the van Gogh dataset.

Image set	Number of images	Patches
authentic	126	2582
imitations	19	271
proxies	212	4208
tuned GANs	30	150
raw GANs	30	150
diffusion	30	150

https://doi.org/10.1371/journal.pone.0295967.t001

All images are pre-processed according to the procedure detailed in [18]. Specifically, we generate sub-images of paintings, i.e., RGB images normalized to a fixed size of 256×256 pixels, with channel values normalized to the unit interval. These sub-images are created by dividing the entire image into $2^P \times 2^P$ equally sized units, where p is determined by the resolution of the original image. If the smaller side of an image is larger than 1024 pixels, then $p = 2$; if the smaller side is larger than 512 pixels and smaller than 1024, then $p = 1$. For all images, irrespective of resolution, we also include the sub-image obtained by center-cropping a square from the full image. Therefore, depending on the resolution of the original image, the images are patched in 21, 5, or 1 adjacent non-overlapping patches. Using bi-cubic resampling, we reshape all the images to either 224×224 pixels or 256×256 pixels depending on the input supported by the model.

The patches are split randomly into training (72%), validation (11%), and test (17%) sets, ensuring that patches belonging to the same original image feature in the same set. We randomly sample the split 10 times, obtaining 10 bootstrapped splits for cross-validation.

For the sake of clarity, we will present the results for the van Gogh dataset [29] in the remainder of this paper. The outcomes for Modigliani and Raphael are available in the supporting information S1 Appendix.

Table 1 provides a detailed overview of the van Gogh dataset. The rows represent the six image sets, while the columns show the number of images and the corresponding number of patches. Representative images of each class (authentic, imitation, GAN, and diffusion) are shown in Fig 1.

Classification methodology

After preparing the training and testing datasets of human-made and synthetic artwork forgeries as described in the previous Section, we proceed to explain the classification methodology on the gathered dataset. In line with the approach outlined in [18], we employ transformer-based classification methods (specifically, Swin Base [31]) and state-of-the-art Convolutional Neural Networks (EfficientNet B0 [32]) to distinguish between authentic artworks and forgeries. These models have been adopted in [18] and proved to outperform the canonical ResNet101 model, the typical baseline model for art authentication. The Swin Base is an image transformer model that uses a hierarchical structure with shifting windows to reduce the computational complexity of transformer models, it accepts inputs of size 224×224 and has $88M$ parameters. On the other side, EfficientNet B0 is a CNN-based model belonging to the class of EfficientNets which adopts an optimal width, depth, and resolution scaling for the architecture. It accepts slightly larger inputs of shape 256×256 and has only $5.3M$ parameters. We note that Swin Base is a larger model version compared to EfficientNet B0. We will present the classification outcomes for both the Swin Base and EfficientNet models, however, we do not directly compare the performance of these models. Rather, the purpose is to demonstrate

Fig 1. Illustration of real (top row) and synthetic (bottom row) van Gogh images. "Self Portrait with a Straw Hat", Vincent van Gogh (1887) [2] (square-cropped, top left), "Self-portrait with a Bandaged Ear and Pipe", sold by Otto Wacker, previously attributed to van Gogh [30] (top right), fine-tuned GAN generated image in the style of van Gogh (bottom left), and Stable Diffusion generated image in style of van Gogh (bottom right).

https://doi.org/10.1371/journal.pone.0295967.g001

that incorporating synthetic data into training datasets enhances classification reliability, regardless of the classifier architecture.

We use the Swin Base and EfficientNet models pre-trained on ImageNet data [33] and fine-tune them for the art authentication task. To do so, we substitute the final activation layer with one dense layer converging in a single node with *sigmoid* activation and train using the *binary*

cross-entropy loss without freezing the weights. For both models, we use a learning rate of 10^{-5}, a batch size of 32. We train the models on binary classification, where class 1 contains the authentic artworks by the artist (authentic set) and class 0 refers to the non-authentic artworks (proxies, imitations, and synthetic images).

To investigate how the addition of synthetic images in the training set improves classification accuracy we run the following experiments. First, we test whether the addition of each of the synthetic sets separately, as well as the combination of Stable Diffusion and fine-tuned GANs, improves the classification accuracy of the human-made forgeries against a baseline trained using 'proxies' and 'imitations'. This baseline also agrees with the previous work [18].

Secondly, we investigate the extent to which synthetic images can increase the detection of human-made forgeries while never training the models on any human forgeries, thus relying solely on 'proxies' and excluding 'imitations'. The setup of the experiments is schematized in Fig 2. We note that the second task is inherently much harder than the first one, as it tests whether synthetic images **alone** can substitute the need to train on human-made forgeries to detect such forgeries. This case scenario addresses the situations in which there are no known forgeries of an artist but it is still desirable to be able to flag possible forgeries. This case is common in art connoisseurship as less well-known artists are rarely forged. To evaluate the performance of our classifiers, we compute the confusion matrices and classification accuracy on the test datasets aggregated at the image level, separately for authentic artworks, imitations, and synthetic images.

Statistical significance of the results is guaranteed by the use of cross-validation with 10 different splits and subsequent uncertainty estimation. All quoted results are the median of the joint distributions and the uncertainties are the (symmetrized) 68%-quantiles of the median (equivalent to the 1σ-standard-error for normally distributed data). We use the concise parenthesis notation in Tables 2–4. Thus, an entry like 0.710(46) means that the central value of the distribution we obtained during the cross-validation process is 0.710 and it is unlikely (at most 32%) that the true value does deviates from this central value by more than 0.046. Similarly, in Figs 3 and 4 this would correspond to a main bar at 0.710 and error-bars of lengths 0.046 both up and down.
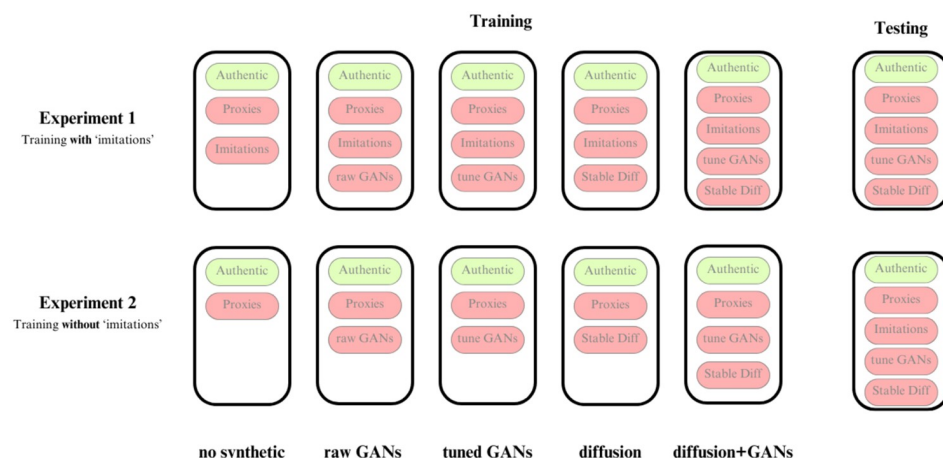

Fig 2. Composition of the training and testing sets for the different experiments. Each box in the training represents a training configuration. The configuration names on the bottom row are used throughout the following sections. Green sub-boxes indicate the original set, red indicates the contrast set.

https://doi.org/10.1371/journal.pone.0295967.g002

Table 2. Performance on different tests after training with forgeries. The composition of the underlying van Gogh data set is detailed in Table 1 and visualised in Fig 2. The best result for each test is highlighted in **bold**. Values are medians with respective uncertainties in parentheses.

training contrast set	model architecture	accuracy forgeries	accuracy originals
no synthetic	Swin Base	0.710(46)	0.960(13)
no synthetic	EfficientNet B0	0.621(72)	0.867(20)
raw GANs	Swin Base	0.738(53)	0.965(9)
raw GANs	EfficientNet B0	0.665(62)	0.863(20)
tuned GANs	Swin Base	0.842(64)	0.963(10)
tuned GANs	EfficientNet B0	0.614(50)	0.850(19)
diffusion	Swin Base	**0.866(44)**	0.960(10)
diffusion	EfficientNet B0	0.723(45)	0.857(17)
diffusion+GANs	Swin Base	0.742(62)	**0.968(7)**
diffusion+GANs	EfficientNet B0	0.785(48)	0.855(18)

https://doi.org/10.1371/journal.pone.0295967.t002

Table 3. Performance on different tests after training without forgeries. The composition of the underlying van Gogh data set is detailed in Table 1 and visualised in Fig 2. The best result for each test is highlighted in **bold**. Values are medians with respective uncertainties in parentheses.

training contrast set	model architecture	accuracy forgeries	accuracy originals
no synthetic	Swin Base	0.269(57)	0.960(14)
no synthetic	EfficientNet B0	0.241(44)	0.857(20)
raw GANs	Swin Base	0.219(63)	**0.965(10)**
raw GANs	EfficientNet B0	0.250(44)	0.873(18)
tuned GANs	Swin Base	0.337(62)	0.959(9)
tuned GANs	EfficientNet B0	0.313(37)	0.876(17)
diffusion	Swin Base	**0.660(73)**	0.943(17)
diffusion	EfficientNet B0	0.535(69)	0.857(23)
diffusion+GANs	Swin Base	0.603(114)	0.953(36)
diffusion+GANs	EfficientNet B0	0.538(68)	0.860(18)

https://doi.org/10.1371/journal.pone.0295967.t003

Table 4. Accuracy of synthetic forgery detection. The composition of the underlying van Gogh data set is detailed in Table 1 and visualised in Fig 2. The best result for each test is highlighted in **bold**. Values are medians with respective uncertainties in parentheses.

training contrast set	model architecture	accuracy Stable Diffusion	accuracy tuned GANs
no synthetic	Swin Base	0.014(7)	0.895(25)
no synthetic	EfficientNet B0	0.015(13)	0.732(64)
raw GANs	Swin Base	0.011(3)	0.972(15)
raw GANs	EfficientNet B0	0.019(8)	0.823(48)
tuned GANs	Swin Base	0.062(16)	**0.989(5)**
tuned GANs	EfficientNet B0	0.034(20)	0.937(25)
diffusion	Swin Base	**0.964(14)**	0.961(13)
diffusion	EfficientNet B0	0.838(37)	0.862(28)

https://doi.org/10.1371/journal.pone.0295967.t004

Results

The outcomes of our classification experiments for van Gogh are shown in Tables 2 and 3. The results are also visually depicted in Fig 3. The evaluation of the classification performance is based on two main criteria: the accuracy in classifying human forgeries (accuracy 'forgeries')
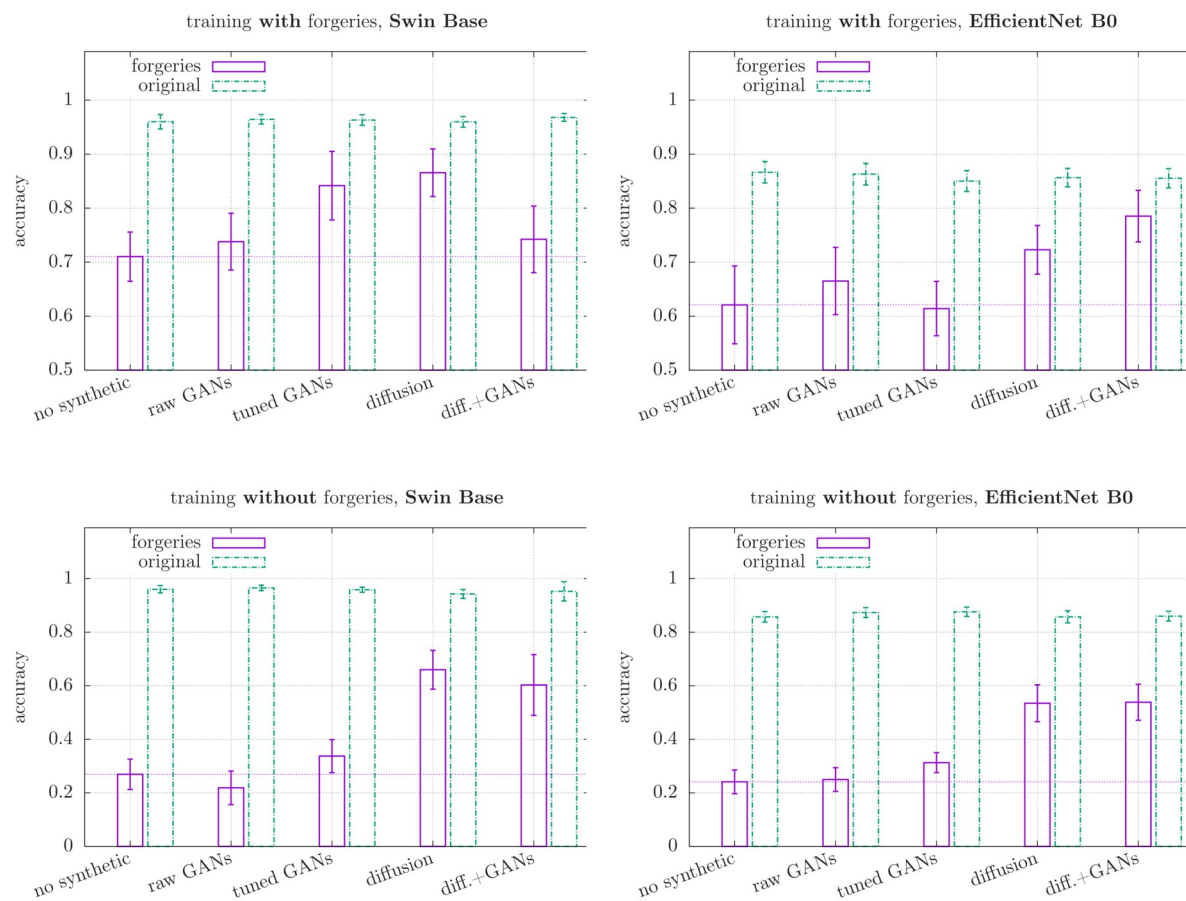
Fig 3. Accuracies of different models for originals and forgeries. Based on the results presented in Tables 2 and 3 with the composition of the underlying van Gogh data set as detailed in Table 1 and visualised in Fig 2. The horizontal dotted line shows the baseline without synthetic images in the training data. Similar results for the artists Modigliani and Raphael can be found in S1 Appendix.
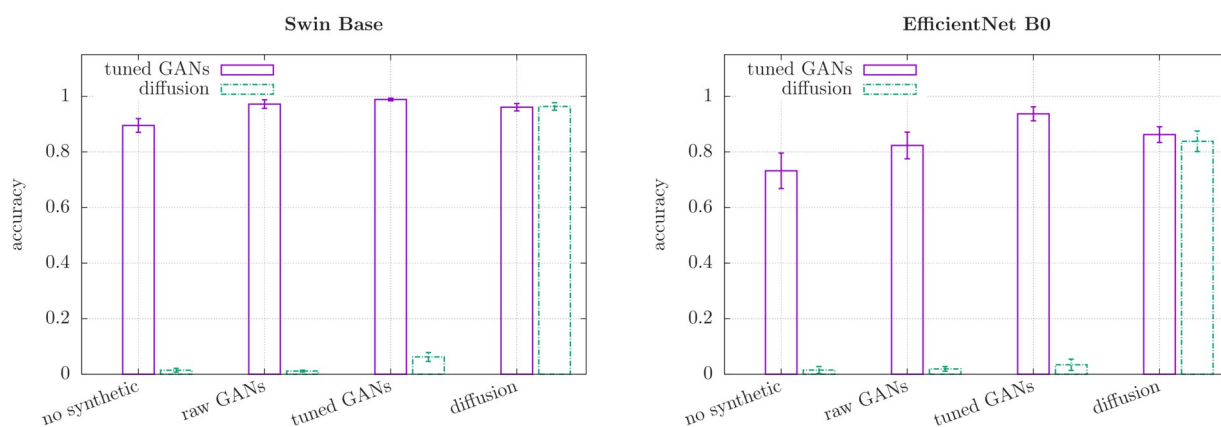
https://doi.org/10.1371/journal.pone.0295967.g003

Fig 4. Accuracies of different models for synthetic data. Based on the results shown in Table 4 with the composition of the underlying van Gogh data set as detailed in Table 1 and visualised in Fig 2.

https://doi.org/10.1371/journal.pone.0295967.g004

and the accuracy in classifying authentic paintings (accuracy 'originals'). Note that what we here refer to as 'forgeries' is synonymous with the set of 'imitations', as those imitations are, in this case, indeed forgeries.

Detection of human forgeries

The classification accuracy for authentic paintings reveals consistently high levels, approximately 90% or higher with the 'Swin Base' model classifier and at least 80% with 'EfficientNet B0', across all training sets, as indicated by the green bars in Fig 3. We observe, moreover, the reproducible improvement in the classification accuracy of human-made forgeries when synthetic forgeries are added to the training datasets. This finding is, to the best of our knowledge, yet unobserved in the literature. On the mixed synthetic training datasets, we were able to achieve accuracies approaching 80% (see Table 2). Images generated by Stable Diffusion appear to be particularly beneficial, leading to accuracy improvements of 10% to 20%. These improvements are evident in Fig 3, where the purple bars (forgeries) associated with "no synthetic" values (this baseline is also extended as a dotted horizontal line) are consistently lower or equal to the values associated with the synthetic ("diffusion" and "tuned GANs") counterparts. This result is particularly impressive considering that the forgeries were often painted by professionals with the goal of avoiding detection.

In addition to augmenting the human-made forgeries with synthetic ones, we also investigated the case where no human-made forgeries were included in the contrast set at all (experiment 2). All classification accuracies are bound to be lower in this case, which is what we observe, and we certainly cannot recommend using this approach in practice if any human-made forgeries are available. However, it allows to resolve the benefit of synthetic forgeries with higher statistical significance. As can be seen in Table 3 and the lower two panels of Fig 3, the addition of synthetic data allowed to improve the forgery detection accuracy by almost 40% and 30% for 'Swin Base' and 'EfficientNet B0', respectively, both corresponding to a significance of about 4σ.

Finally, the quality of synthetic data plays a crucial role in training success, as expected. As seen in Tables 2 and 3, training solely on "raw GAN" images resulted in minor or no improvement in authentication accuracy. Nevertheless, it is interesting to note that in some cases, the addition of "raw GAN" datasets without any author-specific features led to slight enhancements in authentication capabilities. Consistent with previous findings by Schaerf et al. [18] and the inherent differences in model sizes, the transformer-based Swin Base classifier demonstrated slightly superior overall performance.

All findings are consistent across the two models 'Swin Base' and 'EfficientNet B0'. This observation is also supported by a similar analysis of Modigliani and Raphael's datasets described in supporting information S1 Appendix. While the numerical values of the classification results may vary between models and artists, the qualitative conclusions remain consistent across all six combined cases.

Detection of synthetic images

Machine learning methods for the detection of synthetic artwork and synthetic images is currently a very active research topic (see e.g. [34–37]). While the primary focus of this paper is the detection of art forgeries created by humans, in this Section we demonstrate that, in agreement with previous studies, our classifier Neural Networks (Swin Base and EfficientNet) are also capable of detecting synthetic artwork forgeries created by GenAI. A novel aspect of our approach is that, unlike in most previous studies, the classifier is trained on both human-made and synthetic forgeries.

We assess the efficiency of the detection of synthetic forgeries using the synthetic sets listed in Table 1. Tuned GANs and Stable Diffusion images are tested independently with central values (medians) and uncertainties of the cross-validation results are computed in the same way as in teh previous Section Detection of human forgeries. The results are summarized in Table 4 and in Fig 4, which show a considerable improvement of synthetic images detection when integrating synthetic images in the training set.

Our findings are consistent with the typical conclusion in literature (e.g. [35–39]), in that the training on synthetic artwork (tuned GANS, diffusion and raw GANs) is crucial for the classifier to detect forgeries created by GenAI.

We have to distinguish the two cases here of training the classifier with a similar GenAI as has to be detected versus training with a different one. In agreement with the literature, the highest authentication accuracy is obtained if the classifier had already seen synthetic images by the same generator architecture during the training [36, 37]. As one can infer from Table 4, the best results (all above 80%) for tuned GANs detection are achieved when tuned GANs are also included in the training and equivalently training on Stable Diffusion images allows the highest accuracy for diffusion detection. We also observe that including tuned GANs in the training helps to some extent with the detection of images generated by Stable Diffusion, and vice versa. This is an interesting observation given that most previous studies on synthetic forgery detection concentrated on generator-specific image features and visual inconsistencies [39–41].

In the van Gogh based studies presented in this main manuscript it turned out that our classifier could detect GAN images relatively well even without training on synthetic data. The trends described above are visible for both, tuned GANs and diffusion, but they are significantly more pronounced for the latter. When performing the same analysis with the artists Modigliani and Raphael (see S1 Appendix), we found that in some cases tuned GANs also eluded detection very effectively (some accuracies below 10%) as long as no `StyleGAN3` images had been included in the training. In all of these cases training on the given architecture readily improved the accuracy.

Discussion and conclusions

In this work, we demonstrated that additional training on synthetic forgeries can help classifier neural networks to detect human-made art forgeries. To this end, we generated synthetic images using `Stable Diffusion 2.1` [23] as well as `StyleGAN3` [22], and we added these images to our training contrast sets. We discovered that the introduction of synthetic data into the training set significantly improves the accuracy of the detection of human-made forgeries. This holds for all the artists and classifier architectures we tested, both with GAN-generated and Stable Diffusion images, though the latter is clearly preferable. The results are particularly significant when the training set does not include any human-made forgeries at all. This result is particularly novel as it hints at the possibility that synthetic forgeries are similar enough to human forgeries to be able to teach something to the classifier.

In agreement with previous research, we also confirmed that the training on synthetic images expectedly improves the authentication efficiency on synthetic forgeries, especially when the same GenAI architecture was used to produce the training dataset.

Further exploration should be dedicated to quantifying the optimal ratio of human-made forgeries to synthetic data. Moreover, it might be of interest to investigate the influence of image resolutions on the performance of both generators and classifiers. However, significantly larger computing resources than currently available to us would be needed for this type of analysis. With more computational resources, an additional improvement to this work

might be the dedicated post-training of generators like Stable Diffusion on authentic artworks of given artists in order to further enhance the quality of synthetic data.

A notable limitation of our study is that the synthetic GAN-based forgeries are limited to portrait paintings due to the poor convergence of other types of images. While the artistic styles of van Gogh, Modigliani and Raphael are without doubt very different, further tests should be carried out to generalize the findings to a variety of genres and even more different artists.

Supporting information

S1 Appendix. 1. Image generation using StyleGAN. This section presents details regarding the training configurations employed by StyleGAN3. Additionally, we list the numbers of images used for training, along with the corresponding quality of the generated results, sorted by categories. Furthermore, we include visual representations of sample images generated after the training. **2. Classification results for Modigliani and Raphael**. The entire workflow presented for the artist Vincent van Gogh in the main manuscript has also been performed for Amedeo Modigliani and Raphael (Raffaello Sanzio da Urbino). The results can be found in this section.
(PDF)

Acknowledgments

Author Contributions

Conceptualization: Johann Ostmeyer, Ludovica Schaerf, Pavel Buividovich, Tessa Charles, Eric Postma, Carina Popovici.

Data curation: Johann Ostmeyer, Ludovica Schaerf, Carina Popovici.

Formal analysis: Johann Ostmeyer, Ludovica Schaerf, Pavel Buividovich, Eric Postma.

Funding acquisition: Pavel Buividovich, Tessa Charles, Carina Popovici.

Investigation: Johann Ostmeyer, Ludovica Schaerf, Pavel Buividovich, Eric Postma, Carina Popovici.

Methodology: Johann Ostmeyer, Ludovica Schaerf, Pavel Buividovich, Tessa Charles, Eric Postma, Carina Popovici.

Project administration: Pavel Buividovich, Tessa Charles, Carina Popovici.

Resources: Johann Ostmeyer, Ludovica Schaerf, Carina Popovici.

Software: Johann Ostmeyer, Ludovica Schaerf, Carina Popovici.

Supervision: Pavel Buividovich, Tessa Charles, Eric Postma, Carina Popovici.

Validation: Johann Ostmeyer, Ludovica Schaerf, Pavel Buividovich, Carina Popovici.

Visualization: Johann Ostmeyer, Ludovica Schaerf.

Writing – original draft: Johann Ostmeyer, Ludovica Schaerf, Pavel Buividovich, Tessa Charles, Eric Postma, Carina Popovici.

Writing – review & editing: Johann Ostmeyer, Ludovica Schaerf, Pavel Buividovich, Carina Popovici.

References

1. Alberge D. Christie's caught up as £30m forgeries send shock waves through the art world; 2010. Available from: https://www.theguardian.com/artanddesign/2010/oct/17/christies-forger-art-scam.

2. Bailey M. At least 45 Van Goghs may well be fakes: The Art Newspaper Investigates; 2021. Available from: https://www.theartnewspaper.com/1997/07/01/at-least-45-van-goghs-may-well-be-fakes-the-art-newspaper-investigates.

3. Bell P, Offert F. Reflections on connoisseurship and computer vision. Journal of Art Historiography. 2021; 24:BO1.

4. Taylor RP, Micolich AP, Jonas D. Fractal analysis of Pollock's drip paintings. Nature. 1999; 399. https://doi.org/10.1038/20833

5. Lyu S, Rockmore D, Farid H. A digital technique for art authentication. Proc Natl Acad Sci USA. 2004; 101:17006–10. https://doi.org/10.1073/pnas.0406398101 PMID: 15563599

6. Johnson CR, Hendriks E, Berezhnoy IJ, Brevdo E, Hughes SM, Daubechies I, et al. Image processing for artist identification. IEEE Signal Processing Magazine. 2008; 25(4):37–48. https://doi.org/10.1109/MSP.2008.923513

7. Qi H, Taeb A, Hughes SM. Visual stylometry using background selection and wavelet-HMT-based Fisher information distances for attribution and dating of impressionist paintings. Signal Processing. 2013; 93(3):541–553. https://doi.org/10.1016/j.sigpro.2012.09.025

8. Hughes JM, Graham DJ, Rockmore DN. Quantification of artistic style through sparse coding analysis in the drawings of Pieter Bruegel the Elder. Proceedings of the National Academy of Sciences. 2010; 107(4):1279–1283. https://doi.org/10.1073/pnas.0910530107 PMID: 20080588

9. Li J, Yao L, Hendriks E, Wang JZ. Rhythmic Brushstrokes Distinguish van Gogh from His Contemporaries: Findings via Automated Brushstroke Extraction. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2012; 34(6):1159–1176. https://doi.org/10.1109/TPAMI.2011.203 PMID: 22516651

10. Liu H, Chan RH, Yao Y. Geometric tight frame based stylometry for art authentication of van Gogh paintings. Applied and Computational Harmonic Analysis. 2016; 41(2):590–602. https://doi.org/10.1016/j.acha.2015.11.005

11. van Noord N, Hendriks E, Postma E. Toward Discovery of the Artist's Style: Learning to recognize artists by their artworks. IEEE Signal Processing Magazine. 2015; 32(4):46–54. https://doi.org/10.1109/MSP.2015.2406955

12. Dobbs T, Ras Z. On art authentication and the Rijksmuseum challenge: A residual neural network approach. Expert Systems with Applications. 2022; 200:116933. https://doi.org/10.1016/j.eswa.2022.116933

13. Zhu Y, Ji Y, Zhang Y, Xu L, Le Zhou A, Chan E. Machine: The New Art Connoisseur. ArXiv preprint. 2019.

14. Cetinic E, Lipic T, Grgic S. Fine-tuning Convolutional Neural Networks for fine art classification. Expert Systems with Applications. 2018; 114:107–118. https://doi.org/10.1016/j.eswa.2018.07.026

15. Pancaroglu D. Artist, Style And Year Classification Using Face Recognition And Clustering With Convolutional Neural Networks. Preprint. 2020.

16. Varshney S, Lakshmi CV, Patvardhan C. Madhubani Art Classification using transfer learning with deep feature fusion and decision fusion based techniques. Engineering Applications of Artificial Intelligence. 2023; 119:105734. https://doi.org/10.1016/j.engappai.2022.105734

17. David LO, Pedrini H, Dias Z, Rocha A. Authentication of Vincent van Gogh's Work. In: Tsapatsoulis N, Panayides A, Theocharides T, Lanitis A, Pattichis C, Vento M, editors. Computer Analysis of Images and Patterns. Lecture Notes in Computer Science. Cham: Springer International Publishing; 2021. p. 371–380. Available from: https://dx.doi.org/10.1007/978-3-030-89131-2_34.

18. Schaerf L, Popovici C, Postma E. Art Authentication with Vision Transformers. Neural Comput & Applic. 2023. https://doi.org/10.1007/s00521-023-08864-8

19. Shahid M, Koch M, Schneider N. Paint it Black: Generating paintings from text descriptions; 2023. Available from: https://arxiv.org/abs/2302.08808.

20. Gatys LA, Ecker AS, Bethge M. Image Style Transfer Using Convolutional Neural Networks. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016. p. 2414–2423. Available from: https://dx.doi.org/10.1109/CVPR.2016.265.

21. Karras T, Laine S, Aittala M, Hellsten J, Lehtinen J, Aila T. Analyzing and Improving the Image Quality of StyleGAN. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020. p. 8107–8116. Available from: https://dx.doi.org/10.1109/CVPR42600.2020.00813.

22. Karras T, Aittala M, Laine S, Härkönen E, Hellsten J, Lehtinen J, et al. Alias-Free Generative Adversarial Networks; 2021. Advances in Neural Information Processing Systems 34 (NeurIPS 2021). Available from: https://arxiv.org/abs/2106.12423.

23. Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. High-Resolution Image Synthesis with Latent Diffusion Models. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022. p. 10674–10685. Available from: https://dx.doi.org/10.1109/CVPR52688.2022.01042.

24. Ramesh A, Pavlov M, Goh G, Gray S, Voss C, Radford A, et al. Zero-Shot Text-to-Image Generation. In: Meila M, Zhang T, editors. Proceedings of the 38th International Conference on Machine Learning. vol. 139 of Proceedings of Machine Learning Research. PMLR; 2021. p. 8821–8831. Available from: https://proceedings.mlr.press/v139/ramesh21a.html.

25. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning Transferable Visual Models From Natural Language Supervision. In: Meila M, Zhang T, editors. Proceedings of the 38th International Conference on Machine Learning. vol. 139 of Proceedings of Machine Learning Research. PMLR; 2021. p. 8748–8763. Available from: https://proceedings.mlr.press/v139/radford21a.html.

26. Sha Z, Li Z, Yu N, Zhang Y. DE-FAKE: Detection and Attribution of Fake Images Generated by Text-to-Image Generation Models; 2022. Available from: https://arxiv.org/abs/2210.06998.

27. Folego G, Gomes O, Rocha A. From Impressionism to Expressionism: Automatically Identifying Van Gogh's Paintings. In: 2016 IEEE International Conference on Image Processing (ICIP); 2016. p. 141–145. Available from: https://dx.doi.org/10.1109/icip.2016.7532335.

28. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF, editors. Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015. Cham: Springer International Publishing; 2015. p. 234–241. Available from: https://dx.doi.org/10.1007/978-3-319-24574-4_28.

29. Schaerf L, Ostmeyer J, Popovici C, Buividovich P. Vincent van Gogh Authentication Dataset. Zenodo. 2023.

30. Harvard. Self-portrait with a bandaged ear and Pipe; 2019. Available from: https://harvardartmuseums.org/collections/object/230942?position=0&context=exhibition&id=707.

31. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Los Alamitos, CA, USA: IEEE Computer Society; 2021. p. 9992–10002. Available from: https://dx.doi.org/10.1109/ICCV48922.2021.00986.

32. Tan M, Le Q. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In: Chaudhuri K, Salakhutdinov R, editors. Proceedings of the 36th International Conference on Machine Learning. vol. 97 of Proceedings of Machine Learning Research. PMLR; 2019. p. 6105–6114. Available from: https://proceedings.mlr.press/v97/tan19a.html.

33. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition; 2009. p. 248–255. Available from: https://dx.doi.org/10.1109/CVPR.2009.5206848.

34. Islam A, Long C, Basharat A, Hoogs A. 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); 2020. p. 4675–4684. Available from: https://dx.doi.org/10.1109/CVPR42600.2020.00473.

35. Gragnaniello D, Cozzolino D, Marra F, Poggi G, Verdoliva L. Are GAN Generated Images Easy to Detect? A Critical Analysis of the State-Of-The-Art. In: 2021 IEEE International Conference on Multimedia and Expo (ICME). Los Alamitos, CA, USA: IEEE Computer Society; 2021. p. 1–6. Available from: https://dx.doi.org/10.1109/ICME51207.2021.9428429.

36. Ojha U, Li Y, Lee YJ. Towards Universal Fake Image Detectors that Generalize Across Generative Models. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2023; p. 24480–24489.

37. Guo X, Liu X, Ren Z, Grosz S, Masi I, Liu X. Hierarchical Fine-Grained Image Forgery Detection and Localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023. p. 3155–3165. Available from: https://openaccess.thecvf.com/content/CVPR2023/html/Guo_Hierarchical_Fine-Grained_Image_Forgery_Detection_and_Localization_CVPR_2023_paper.html.

38. Corvi R, Cozzolino D, Zingarini G, Poggi G, Nagano K, Verdoliva L. On The Detection of Synthetic Images Generated by Diffusion Models. In: ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2023. p. 1–5. Available from: https://dx.doi.org/10.1109/ICASSP49357.2023.10095167.

39. Wang SY, Wang O, Zhang R, Owens A, Efros AA. CNN-Generated Images Are Surprisingly Easy to Spot. . . for Now. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020. p. 8692–8701. Available from: https://dx.doi.org/10.1109/CVPR42600.2020.00872.

40. Farid H. Lighting (In)consistency of Paint by Text; 2022. Available from: https://arxiv.org/abs/2207.13744.

41. Farid H. Perspective (In)consistency of Paint by Text; 2022. Available from: https://arxiv.org/abs/2206.14617.