Painting Authorship and Forgery Detection Challenges with AI Image Generation Algorithms: Rembrandt and 17th Century Dutch Painters as a Case Study

Marcelo Fraile-Narváez*, Ismael Sagredo-Olivenza, Nadia McGowan

Universidad Internacional de La Rioja (Spain)

Received 2 October 2022 | Accepted 15 November 2022 | Published 28 November 2022

Abstract

Image authorship attribution presents many challenges and difficulties which have increased with the capabilities presented by synthetic image generation through different artificial intelligence algorithms available today. The hypothesis in this research considers the possibility of using artificial intelligence as a tool to detect forgeries through the usage of a deep learning algorithm. The proposed algorithm was trained using a dataset comprised of paintings by Rembrandt and other 17th century Dutch painters. Three experiments were performed with the proposed algorithm. The first was to build a classifier able to ascertain whether a painting belongs to the Rembrandt or non-Rembrandt category, depending on whether it was painted by this author or not. The second tests included other 17th century painters in four categories. Artworks could be classified as Rembrandt, Eeckhout, Leveck or other Dutch painters. The third experiment used paintings generated by Dall-e 2 and attempted to classify them using the prior categories. Experiments confirmed the hypothesis with best executions reaching accuracy rates of more than 90%. Future research with extended datasets and improved image resolution are suggested to improve the obtained results.

Keywords

Artificial Intelligence, Authentication, Image Generation, Machine Learning, Neural Network.

DOI: 10.9781/ijimai.2022.11.005

I. Introduction

SYNTHETIC image generation algorithms based on deep learning have become increasingly popular in recent years. Advances in this field have brough surprising results, such as new neural networks capable of generating images with incredible precision. Among them are Dall-e 2 [1], based on CLIP [2], Midjouney [3] or Imagen, by Google [4]. These algorithms are focused on the generation of artificial images according to a set of keywords. They combine language understanding, classification, and image generation systems.

Essentially, we are faced with a new technology with significant potential, but which has also led to concern among contemporary artists as they fear their work may be endangered by these algorithms. There has even been speculation regarding creating new works by deceases artists [5]. An example of this is the case presented by AI expert Carlos Santana [6], who published on social media some results obtained using Dall-E 2 and famous artworks, creating painting astonishingly faithful to the artists' works. AI synthetic image generation has generated controversy among artists. Some extremes dismiss this technology, considering the images it generates lack artistic value [7] while others fully embrace them [8].

This paper would like to question the possibility of using AI as a tool to protect artworks and their attribution by using it as a tool to detect possible forgeries created by AI algorithms. The working hypothesis of this research is that it is possible to formulate a deep learning algorithm capable of detecting with a high degree of accuracy whether an image is authentic, or a digital falsification created by an AI. Such an algorithm would have other potential applications, such as, for example, supporting specialists in the first stage of attributing artworks whose creators are anonymous or unknown.

The question that drives this article is not foreign to academic research. Similar research has been carried out regarding deepfakes [9]. These are videos where deep learning models are used to substitute the face of a person (usually well-known) for someone else's. These videos can have humorous intent but can also harbor malicious purposes [10]-[14]. The possible implications of attacks with modified images to hack learning systems that have images as an input, such as adversarial attacks in classification systems, have also been studied [15].

A new reality is emerging, where machines are now able to learn, to compose music, and to paint like Rembrandt. Through a deep learning algorithm and facial recognition, with devices within everyone's reach, it is possible for a program to create its own artwork. This process seems to blur the boundaries between art and technology.

A relevant example of the use of AI art creation technology is that of The Next Rembrandt. This project aimed to generate a painting by the Dutch Baroque painter Rembrandt van Rijn (1606-1669) that was indistinguishable from his originals. This project involved Microsoft, the Dutch bank ING, the advertising agency J. Walker Thompson, the Technical University of Delft, the Mauritius Museum, and the Museum Het Rembrandthuis. In addition to the artificial intelligence used to

* Corresponding author.
E-mail address: marcelo.fraile@unir.net

create the painting, the project also sought to recreate the techniques used by the artist using a 3D printing technology that replicated the texture of oil painting.

Specialists question whether it would be possible to establish the authenticity of a work of art in the midst of conflicting expert opinions [16]-[17].

Two considerations must be taken into account when working with these paintings. Firstly, many artists benefited from the help of their disciples. At times, they left in their hands the execution of some details of their work. One well-known case is that of Leonardo Da Vinci and his intervention in different pieces by the master Verrocchio, to the point where today it is debated whether the latter intervened in some of these paintings. Secondly, many unsigned images have been attributed to painters such as Rembrandt and later studies have dismissed or questioned these attributions.

Secondary studies on the work of art are important to ascertain attribution. An artwork is comprised of several layers, other than the final painting seen. There is a primer, guides or sketches drawn on the canvas and modifications made during the painting process which can encompass variations in the background or other details.

An interesting project in this field is that developed by the Zhejiang University of Technology in China, which seeks to create a machine learning model capable of describing and classifying works of art by genre and style. In their results [18] the authors highlighted the importance of using convolutional neural networks to classify art.

This study experimented with seven different models applied to three different datasets under the same experimental setup. The algorithm initially categorizes images according to style and genre, and then classifies them by looking for similarities. A growing number of projects use CNN models to solve classification problems. However, training CNNs requires datasets containing a large number of labeled entries. Their success depends on the availability of large datasets such as ImageNet. For this project, the authors used Cafenet, a slightly modified version of the AlexNet model, to evaluate the fine-tuning process using five pre-trained networks.

Their dataset images were obtained from WikiArt, an organized collection of more than 80,000 images, with more than 1000 artists, 27 different styles and 45 genres separated into different categories. The size of each sample was set at 450 x 450, neither too small so as not to limit the analysis of fine details, nor too large so as not to overfit the CNN with the training data. This resulted in a success rate of more than 90 %. The CNN initially took only color information to classify the painting, later it included spatial information to help the model distinguish portraits from landscapes. It does, however, present problems to identify individual painters from styles.

Steven Frank, in turn, developed a CNN capable of identifying painters such as Picasso, Van Gogh or Rembrandt. It generates a probably map through the division of an imagen divided into a mosaic of small square fragments that can be handled by the CNN, while increasing the number of images used for its training [19]. In his research, Frank developed a CNN to identify authentic Rembrandts from forgeries. He selected 50 portraits by Rembrandt and 50 by randomly selected artists. Some had a very similar style to his and others, while similar, were clearly distinguishable from his work. This choice was made due to the fact that if they were too much alike, the CCN would over-fit and not generalize its training.

Other machine learning techniques have also been used for the classification of works of art, as for example in the work of Wu [20], Xu [21] or Blessing and Wen [22]. In the latter, the authors try to perform two-by-two classifiers among a set of painters including Cezanne, Dali, Durer, Monet, Picasso, Rembrandt, and Van Gogh. As feature extractors, they used different algorithms such as GIST [23], HOG2x2 [24], Dense SIFT [25], etc. And for classification Support Vector Machine was used, with results ranging between 90% and 95%.

Narang and Soriano [26] used the Gray-Level Co-occurrence Matrix (GLCM) to extract the characteristics of a painting and to classify a neural network or a Support Vector Machine with a gaussian kernel, obtaining results of around 83% and 85% accuracy in the detection of paintings created by Juan Luna. For training, this project used 13 high resolution paintings created by Luna, and another 13 by other Filipino artists.

In this study, a CNN will be trained with a dataset developed based on similar artistic styles. This is intended to make the network more sensitive to small perturbations in the artists' styles. These considerations are key to detect forgeries more effectively between paintings that are already remarkably similar. Thus, seeking to focus this research, artworks by seventeenth-century Dutch artists contemporary with Rembrandt's academy are selected.

Rembrandt was chosen as the main painter of this study due to two conditioning factors: firstly, the number of works sufficiently large to build an acceptable training corpus and, secondly, having enough imitators, disciples, attributed paintings and the existence of The Next Rembrandt, a painting generated using Microsoft's AI. All these resources implied an abundant amount of information to assess the algorithm.

II. Methods

To test our hypothesis, a machine learning model capable of identifying Rembrandt's paintings among authors sharing aesthetic similarities was created to identify the details that differentiate Rembrandt from his imitators, in the hope that the network would also learn to distinguish forgeries. To this end, and drawing on the literature, a convolutional network was chosen as an image feature extractor. Specifically, a feature extractor encompassed within the MobileNet V2 family of algorithms proposed by Howard in 2018 [27] was chosen. In particular, the TF-Hub module used the TF-Slim implementation of mobilenet_v2 with a depth multiplier of 1.0 and an input size of 224x224 pixels. All images used for network training are scaled to this resolution.

This feature extractor has been trained using the ILSVRC-2012-CLS image dataset used in the google ImageNet competition. This network has been trained with an image corpus of 1.2 million images. As an unsupervised algorithm, it does not take into account the classes to which each of these images belong since the purpose of its training is to obtain a feature extractor from the image. Specifically, it uses DeepLabv3 as the feature extractor of the model proposed by Chen et al. [28] where it is explained in greater detail, but which we will briefly describe hereafter. This feature extractor uses the 3x3 Atrous convolution originally developed for the efficient computation of the undecimated wavelet transform [29] and which has been widely used for object detection [30]. A series of convolutional filters typical in these feature extraction algorithms are applied in this model, specifically the layers are based on successive copies of the blocks proposed by ResNet [31] placed in a cascade of up to 6 levels, replacing the fifth level by an Atrous Spatial Pyramid Pooling [32] with four parallel Atrous convolutions with different Atrous rates that are applied on top of the feature map because this layer has been shown to be effective to resample features at different scales allowing a more accurate classification. The module generates an output of 1280 features extracted from the original image that are then used to perform a classification.

To perform the subsequent classification, we use a pair of dense layers, the first one of 2560 neurons with ReLU activation and the

second one, a layer with N neurons, being N the number of classes that we are going to establish with SoftMax or sigmoidal activation depending on the number of classes used and the purpose of these. Further on in the specific experiments we will detail which is used in each case.

Out of the entire network, only the two dense layers were trained, leaving the convolutional feature extraction layer pre-trained with the ImageNet dataset. Next, we attempted to retrain the layer in order to detect the presence of other authors. However, the main problem found when retraining the features extractor is mainly due to the fact that the number of existing paintings by a single artist is very limited and the network lacks sufficient examples to learn correctly, even when applying augmentation techniques. Therefore, in this experiment it was decided to keep the feature extractor trained on a set of generic images and not applied to paintings. Future research would need to explore a mechanism to combine this general feature extraction provided by the module used with some other features extractor trained only on paintings, to enhance the input of dense layers with more information.

All experiments were data augmented by generating batches of 32 images with 1/255 rescaling, 50 degrees rotation, 0.25 horizontal and vertical displacement, 15 shear and zoom from 0.25 to 1.55. All experiments also used 80% of the examples for training and 20% for validation. The tests have been performed using the Keras library from Google Collaborate with GPU access.

Several experiments have been carried out using this model and variants in the dense layers, which are detailed below.

A. Rembrandt and Non-Rembrandt Detection

Taking into account the existing literature on the subject, the first step taken was to build a classifier of painting belonging to Rembrandt and those not painted by him. To maximize fake detection, the approach chosen was to sort in a binary classification between Rembrandt and non-Rembrandt paintings.

To this end, a training corpus consisting of 280 images of different resolutions was created. This was due to the fact that they were obtained from online web scrapping. All images were rescaled to a resolution of 224x224 using the OpenCV library for processing by the feature extractor.

One third of the images were paintings by Rembrandt, two thirds of the images belonged to Rembrandt's disciples and the rest were paintings by 17th century Dutch artists who influenced or were influenced by Rembrandt.

As discussed in the introduction and hypothesis, similar paintings have been selected to try to make the network generalize and learn the characteristics of the artist (Rembrandt) among similar paintings, to improve the detection of forgeries. The assumption is that the network will learn to differentiate the small subtleties of the feature vector between Rembrandt's paintings and those of his contemporaries and disciples, to then be able to generalize to AI-generated paintings. While these will present certain characteristics similar to those of Rembrandt, they should be closer to the non-Rembrandt class than to the Rembrandt class in the classification.

The goal would be to create a model capable of detecting fake images from any image generator and not only known ones. This is a basic principle of adversarial attack systems and other fake detection systems. Fake examples are used to train the network, but it is important for the system to have a good detection rate without the need to retrain the model with fake examples since models tend to overfit the data entered during training and lose their ability to generalize. It has been shown that, in the context of adversarial attacks, it is difficult to train a network with examples of attacks for learning. One will always find new examples for which the network does not behave as one expected [33]. Something similar happens in this field. If the network is trained to detect fakes produced by Dall-e 2, it does not necessarily correctly detect images generated by Imagen and vice versa. Therefore, the aim of this study is to establish specific training for already known image generation algorithms. They are an aid to the main detection algorithm, but they should not replace a more generalist model that maintains a good rate of detection of fakes, regardless of the algorithm that generates these fakes. It is because of this that the initial training corpus has not included paintings generated by these algorithms, instead it has used others that are similar but generic.

B. 17th Century Painter Detection

The Rembrandt and non-Rembrandt classification was extrapolated to 17th century artists, to include other artists. This second test has been carried out as a trial and would require further research to improve its results. This could be accomplished with a more complex network or more training examples, but results are deemed sufficient to raise within this article the possibility to extend this multiclass classification system where each painter is a specific class. A "miscellaneous" class has been used to identify other authors that are not included among the training options.

To detect 17th century Dutch painting among other artworks, two classes could be created. One would contain examples of Dutch painting from the 17th century while the second would hold examples of other authors from that century. For example, these could be non-Dutch painters. There are endless possibilities, although obviously the more complex the classifier, the machine learning model will need to be more complex, include more training examples and classification results will worsen.

As noted by Steven J. Frank, the basic problem around this domain should also be noted. Unlike image classification or object detection in images, the number of examples used in training is limited to the works produced by different painters. Rembrandt was a prolific artist with a corpus of several hundred works. Other authors do not have such a corpus available for training.

This experiment builds on the basis of the previous one. In this occasion, the network includes four classes: Rembrandt, Eeckhout, Leveck and other 17th century Dutch painters. The dataset for each class is composed of seventy sample images. Eeckhout and Leveck have been chosen due to the similarity of their style to that of Rembrandt, given that they were disciples of the renowned artist. In this experiment there is a single discard group, that of 17th century Dutch painters. After training, different tests were performed with paintings from the selected period as well as others.

The network used was based on the same feature extractor, but the classification layers are modified. A pair of dense layers are used, one of 2560 neurons with ReLU activation and the second, a layer with 4 neurons with soft max activation function.

C. Validation With Images Generated Using Dall-e 2

In the third experiment, images in the style of Rembrandt will be generated with Dall-e 2. The goal is to validate the models generated in the previous experiments in order to determine which best classifies images generated using Dall-e 2. The approach was not only to verify whether the classifier detects if a painting is a Rembrandt or not. The network output can be used to interpret the probability a processed painting has to belong to one of the categories the model classifies.

As an example, suppose we use the model from experiment 2, where there is a network that can classify 4 groups of painters. The output of the network is therefore constituted by a vector of cardinality 4. Let us imagine that once the network is trained, we expose it to a Rembrandt

painting, generating the following output [0.975, 0.750, 0.619, 0.120] with the first value being the probability of being it being painted by Rembrandt, the second the probability of being painted by Eeckhout and so on.

This output has two possible interpretations. One is to assume that the most probable class is the one that the network classifies. With this interpretation we can say that in this example the network classifies the painting as a Rembrandt. However, we can also interpret the output as a set of probabilities. Using this interpretation, we can say that the painting has a 97% chance of being a Rembrandt, a 75% chance of being an Eeckhout and a 69% chance of being a Leveck. Through this interpretation we can give more information to the network user, since with this information we can estimate the degree of confidence that the network gives to the classification. With this information we can further refine the result.

Let us imagine another scenario where the output is [0.75, 0.70, 0.40, 0.1]. In this scenario with interpretation 1 we would say that it is a Rembrandt, but if we look at the data displayed by the network, we can infer that there is a high probability of it not being a Rembrandt. The level of confidence of the network in this classification is very low for the class with the highest probability.

Three classes for the detection of fakes are proposed: Rembrandt, Fake and Doubtful.

When a painting is classified as Doubtful, the system provides a recommendation of which alternate category it leans towards, i.e., whether it is more likely to belong to the Rembrandt or Fake category. To do this the system calculates the outputs of the network in the form of probability. There is a parameter θ that determines which is the margin of doubt that best classifies the frames according to probability. To detect this parameter, we perform a study of its continuity to determine if the classification improves or worsens by applying slight changes in the parameter.

The classification function can be defined as a piecewise function as follows:

$$f(x) = \begin{cases} 2 \;\; if \quad \left(\left|c_0 - \sum_{i=1}^{N-1} C_i\right|\right) < \theta \\ 1 \;\; if \;\; C_o < \left(\sum_{i=1}^{N-1} C_i\right) \; and \; \left|c_0 - \sum_{i=1}^{N-1} C_i\right| > \theta \\ 0 \;\; if \;\; C_o > \left(\sum_{i=1}^{N-1} C_i\right) \; and \; \left|c_0 - \sum_{i=1}^{N-1} C_i\right| > \theta \end{cases} \quad (1)$$

Where C_0 is the neuron that represents the probability of a painting being a Rembrandt. C_i is the output of the network for the remaining neurons of the output layer. They indicate the probability estimated by the network of an image belonging to one of the classes predicted by the network. N is the number of classes predicted by the network.

If the result of the function is 0, it would indicate that the image is a Rembrandt. If the result is 1, it is a Fake. If the output is 2, the case is Doubtful. In this case, the system will then display the probability and also show what it predicts to be the most likely option. The expert using the system will thus have information on the reliability of the network's prediction. This is important to be able to perform a subsequent review of the doubtful cases and as we will see in the results, it improves the results obtained by the network, detecting many false positives as doubtful.

III. Results and Discussion

The results from the previously described experiments are detailed below, separated into different subsections. Results presented in this article are the best values obtained in different executions performed, as it has not been possible to minimize experiment randomness since they were executed in GPU. Due to floating point precision errors, it is extremely difficult to achieve exactly the same results in each run.

A. Results of Rembrandt and Non-Rembrandt Detection

In the first experiment performed we sought to classify the training data and its validation between two classes. These were Rembrandt and non-Rembrandt artworks. In this experiment, the hyperparameters of the dense layers have kept their default values.

The dataset was comprised of 280 images. 70 were classified as Rembrandt and 210 as non-Rembrandt. 80% were used for training and 20% for validation.

A final dense layer with sigmoidal activation was used, which is described in the literature as working very well for this type of binary classification. As a measure of loss calculation, Binary Cross-Entropy has been used, as per the following equation:

$$H_p(q) = 1 - \frac{1}{N} \sum_{i=1}^{N} y_i \dot{\log}(p(y_i)) \dot{\log}(1 - p(y_s i)) \quad (2)$$

Where y_i is the output (1 or 0) and $p(y_i)$ is the probability predicted by the network. This measure is also typical of binary classification within the literature.

The network was trained using EarlyStopping with the validation loss measure as a stopping metric.

According to these parameters, the training result produced a convergence around epoch 20 and 0.8929 accuracy.

The following step was to substitute the last layer with a SoftMax layer and change the loss metric to categorical cross entropy, as shown in Equation (3):

$$Loss = 1 - \sum_{i=1}^{N} y_i \dot{\log} p(y_i) \quad (3)$$

The output is then treated as a probability that will be used when detecting doubtful cases.

Training with these changes produced a convergence around epoch 16 with an accuracy of 0.8929 in the model validation. The results of both models are similar, but in the third experiment we will test the effectiveness of both in detecting fakes, interpreting the output as a binary classification and as a probability.

By testing the algorithm with the painting "The Next Rembrandt", a painting developed by AI, the network has yielded a 99% attribution of the painting as an authentic Rembrandt. And although a first impression would suggest that the system has failed since the painting is not really a Rembrandt, the fidelity of the result obtained by Microsoft is very high and its detection as a fake is very complex. It would probably be quite complicated to create a model that would classify it as not-Rembrandt.

B. Results of 17th Century Painter Detection

This experiment seeks to develop a classification by categories. In order to do so, the configuration from the second part of the first experiment is kept, with a last dense layer of four neurons instead of two, with SoftMax activation function. Similarly, the stopping criterion and the loss measure used are also maintained.

The results obtained from this second experiment produced a convergence around epoch 26 with an accuracy of 0.7621 in the validation of the model.

Given low image resolution and the small number of paintings used to generate the dataset, the obtained result of 0.7621, although not optimal, is acceptable for an approximation to the problem. It follows that, to optimize results, a larger dataset would be necessary, and the images used would require greater resolution. This leaves an open path for future work related to the subject through high resolution images obtained from the Rembrandt Museum in Amsterdam.

C. Results of the Validation With Images Generated Using Dall-e 2

The third experiment used the three models generated by the two prior experiments (Rembrandt and non-Rembrandt detection, 17th century painter detection) that obtained the best accuracy metrics to attempt to detect fake Rembrandt images created by the Dall-e 2 platform.

52 images were generated in the Dall-e 2 platform with the prompts "Rembrandt", "Knight painting painted by Rembrandt", "Rembrandt painted", "Rembrandt painted portrait", "Rembrandt oil portrait", and "Rembrandt-type painted portrait". Some examples of images generated by these prompts can be found in Fig. 1 and Fig. 2.

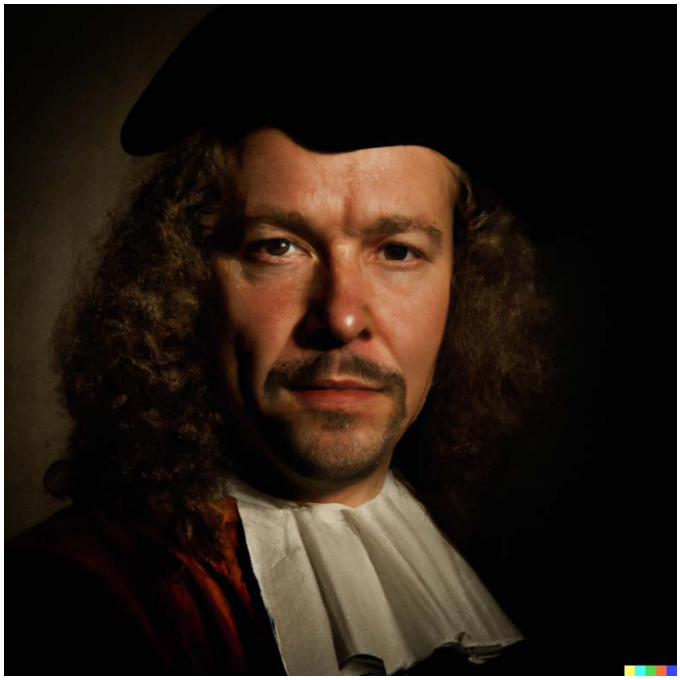

Fig. 1. Image generated by Dall-e 2 using as a prompt "Rembrandt".

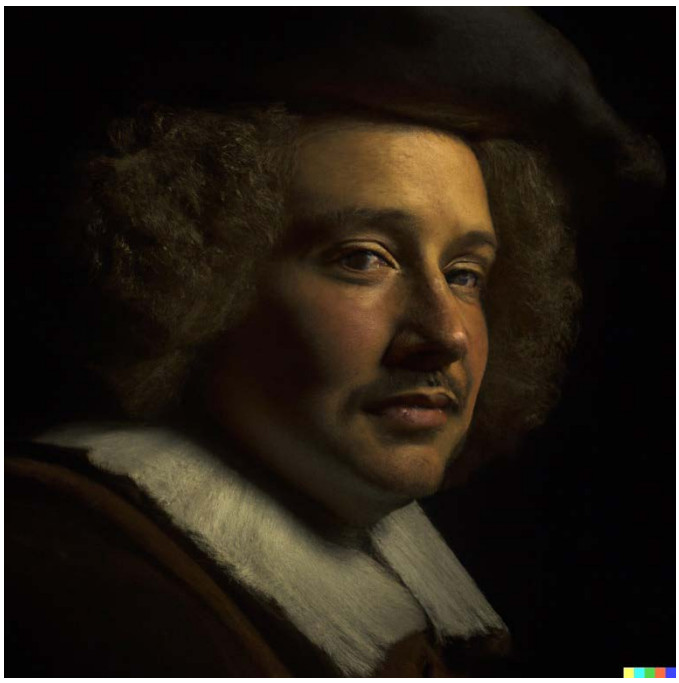

Fig. 2. Image generated by Dall-e 2 using as a prompt "Knight painting painted by Rembrandt".

The two binary classifiers from the first experiment obtained an accuracy of 0.9038 and 0.8846 respectively, classifying the paintings produced by Dall-e 2 in the non-Rembrandt class.

However, as expected due to its reduced performance, the accuracy of the multi-class classifier in detecting fakes has decreased to 0.6153. To calculate this value, we considered artworks as non-Rembrandt if the algorithm attributed the painting to any of the three classes (Eeckhout, Leveck, and other 17th century Dutch painters) except Rembrandt.

These results indicate that, at least with the number of paintings in the dataset (70 per artist), the binary classifier obtains better results than a multi-class classifier when trying to detect fakes produced by Dall-e 2.

The next test used the probability generated by the second network from experiment 1 to check if the detection of fakes improved following Equation (1). In this case, the parameter θ chosen as the optimum at the authors' discretion was 0.1. This leaves us with 5% of doubtful paintings and improves the detection of fakes considering fake those classified as non-Rembrandt and doubtful to an accuracy of 0.9423. It is important not to have a high percentage of doubtful artworks since then the network would detect all paintings as doubtful. In other words, while the accuracy would be maximized, the network would not have practical sense. A value of θ that would produce more than 10% of doubtful artworks would not be useful, considering that the values obtained are already quite high.

The final experiment consisted of introducing half of the examples generated with Dall-e 2 as part of the training cases in the non-Rembrandt class. This was performed to check whether, as we supposed, the fake detection rate improved when introducing examples of the image generation algorithm in the training. However, as previously discussed in this study, caution is recommended, as the network may become overtrained when trying to detect examples generated by the algorithm as fakes and lose its ability to generalize with other unknown algorithms. Although this seems intuitively logical, it has not been tested in this work and is proposed as future research.

The results of the latter experiment converged at epoch 11 with an accuracy of 0.9655 using binary classification. In this case, the detection of dubious cases did not lead to any improvements as only one case was detected. However, it belonged to the non-Rembrandt class. As we can see in this experiment, introducing the Dall-e 2 examples in the training improves the detection of fakes obtaining better accuracy (0.9655) than the best result training without Dall-e 2 images (0.9038).

These results described in the last subsection were summarized in the Table I, where Binary classifier corresponding to the result obtained with the best model in the first experiment where the last layer had a sigmoidal activation function. The classifier Rembrandt Non-Rembrandt (R-NR) corresponds to the second part of the experiment 1, where the last layer had a SoftMax activation function.

TABLE I. Summary of the Results Obtained by the Models in the Detection of False Rembrandt With Dall-e 2

Model	Accuracy
Binary classifier	0.9038
Classifier Rembrandt – Non-Rembrandt (R-NR)	0.8846
Classifier with 4 categories	0.6153
Classifier (R-NR) witch doubtful	0.9423
Classifier (R-NR) trained with Dall-e 2 images	0.9655

IV. Conclusions

Several experiments have been performed in this study to confirm the starting hypothesis. In it, it was stated that it was possible to create a deep learning algorithm capable of detecting with a high degree of

accuracy false images generated by AI algorithms. This study focused on Dall-e 2. In this context, the best executions reached more than 90% accuracy rates.

Images generated by algorithms (Dall-e 2 in this case) were tested as a part of the training. Results improve but further work is needed to ensure that there is no loss of its generalization capacity against other algorithms that can develop this kind of forgeries or fakes. In the future, tests with other image generators such as Google Imagen or Midjourney could be performed.

The study presented has several limitations, such as those posed by the number of images available by some artists and their resolution. These elements have limited the scope of this work. This has been present in the results obtained by the second and third experiments.

Image authorship attribution presents many problems for specialists, which is why an algorithm such as the one presented in this study could potentially become a useful tool for early identification of works by an artist. However, while this would help optimize the work time of researchers, it could not replace experts when attributing authorship.

Acknowledgments

This paper was developed as part of the Quantification and prediction studies of cinematographic works (*Estudios de cuantificación y predicción de parámetros estéticos de obras cinematográficas*) research project funded by Universidad Internacional de La Rioja.

References

[1] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, M. Chen, "Hierarchical text-conditional image generation with clip latents," 2022, doi: 10.48550/arXiv.2204.06125.

[2] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., "Learning transferable visual models from natural language supervision," in International Conference on Machine Learning, 2021, pp. 8748–8763, PMLR.

[3] Midjourney, https://www.midjourney.com/. [Online; accessed 01-October-2022].

[4] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, et al., "Photorealistic text-to-image diffusion models with deep language understanding," arXiv preprint arXiv:2205.11487, 2022.

[5] A. Cullins, "Star Wars' and the legal issues of dead but in-demand actors," https://www.hollywoodreporter.com/movies/movie-news/carrie-fisher-star-wars-legal-issues-dead-but-demand-actors-997335/, 2017. [Online; accessed 01-October-2022].

[6] C. Santana [@DotCSV], https://twitter.com/DotCSV/status/1544959141004861441, 7 July 2022. [Online; accessed 01-October-2022].

[7] B. Edwards, "Flooded with AI-generated images, some art communities ban them completely," https://arstechnica.com/information-technology/2022/09/flooded-with-ai-generated-images-some-art-communities-ban-them-completely/, 12 September 2022. [Online; accessed 01-October-2022].

[8] Christies, https://www.christies.com/features/A-collaboration-between-two-artists-one-human-one-a-machine-9332-1.aspx, 12 December, 2018. [Online; accessed 01-October-2022].

[9] S. Lyu, "Deepfake detection: Current challenges and next steps," in 2020 IEEE international conference on multimedia & expo workshops (ICMEW), 2020, pp. 1–6, IEEE.

[10] B. Chesney, D. Citron, "Deep fakes: A looming challenge for privacy, democracy, and national security," California L. Rev., vol. 107, p. 1753, 2019.

[11] R. Delfino, "Pornographic deepfakes—revenge porn's next tragic act–the case for federal criminalization," 88 Fordham L. Rev., vol. 887, 2019.

[12] H. B. Dixon Jr, "Deepfakes: More frightening than photoshop on steroids," Judges J., vol. 58, p. 35, 2019.

[13] S. Feldstein, "How Artificial Intelligence Systems Could Threaten Democracy," The Conversation, 2019.

[14] P. Rey-García, N. McGowan, La amenaza híbrida: la guerra imprevisible, ch. El deepfake como amenaza comunicativa: diagnóstico, técnica y prevención. Madrid, Spain: Ministerio de Defensa, 2020.

[15] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, "Intriguing properties of neural networks," arXiv preprint arXiv:1312.6199, 2013.

[16] S. J. Frank, "This ai can spot an art forgery." https://spectrum.ieee.org/this-ai-can-spot-an-art-forgery, 23 AUG 2021. [Online; accessed 01-October-2022].

[17] S. J. Frank, A. M. Frank, "Rembrandts and robots: Using neural networks to explore authorship in painting," arXiv preprint arXiv:2002.05107, 2020.

[18] W. Zhao, D. Zhou, X. Qiu, W. Jiang, "Compare the performance of the models in art classification," Plos one, vol. 16, no. 3, p. e0248414, 2021. https://doi.org/10.1371/journal.pone.0248414

[19] S. J. Frank, A. M. Frank, "Salient slices: Improved neural network training and performance with image entropy," Neural Computation, vol. 32, no. 6, pp. 1222–1237, 2020. https://doi.org/10.1162/neco_a_01282

[20] Y. Wu, Q. Wu, N. Dey, S. Sherratt, "Learning Models for Semantic Classification of Insufficient Plantar Pressure Images," International Journal of Interactive Multimedia and Artificial Intelligence, vol. 6, no. 1, pp. 51-61, 2020. https://doi.org/10.9781/ijimai.2020.02.005

[21] F. Xu, T. Wu, S. Huang, K. Han, W. Lin, S. Wu, S. CB, S. R. Dinesh Jackson, "Extensive Classification of Visual Art Paintings for Enhancing Education System using Hybrid SVM-ANN with Sparse Metric Learning based on Kernel Regression," International Journal of Interactive Multimedia and Artificial Intelligence, vol. 7, no. 2, pp. 224-231, 2021. https://doi.org/10.9781/ijimai.2021.10.001

[22] A. Blessing, & K. Wen, Using machine learning for identification of art paintings. Technical report. 2010.

[23] A. Oliva, A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," International journal of computer vision, vol. 42, no. 3, pp. 145–175, 2001.

[24] N. Dalal, B. Triggs, "Histograms of oriented gradients for human detection," in 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), vol. 1, 2005, pp. 886–893, IEEE.

[25] S. Lazebnik, C. Schmid, J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in 2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06), vol. 2, 2006, pp. 2169–2178, IEEE.

[26] M. J. G. Narag, M. N. Soriano, "Identifying the painter using texture features and machine learning algorithms," in Proceedings of the 3rd International Conference on Cryptography, Security and Privacy, 2019, pp. 201–205.

[27] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.C. Chen, "Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation," arXiv preprint https://arxiv.org/abs/1801.04381v2, 2018.

[28] L.-C. Chen, G. Papandreou, F. Schroff, H. Adam, "Rethinking atrous convolution for semantic image segmentation," arXiv preprint arXiv:1706.05587, 2017.

[29] M. Holschneider, R. Kronland-Martinet, J. Morlet, P. Tchamitchian, "A real-time algorithm for signal analysis with the help of the wavelet transform," in Wavelets, Springer, 1990, pp. 286–297.

[30] J. Dai, Y. Li, K. He, J. Sun, R. Fcn, "Object detection via region-based fully convolutional networks," arXiv preprint arXiv:1605.06409, 2016.

[31] K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[32] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," IEEE transactions on pattern analysis and machine intelligence, vol. 40, no. 4, pp. 834– 848, 2017.

[33] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, P. Frossard, "Universal adversarial perturbations," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1765–1773, 2017.

Marcelo Fraile-Narváez

Marcelo Fraile-Narváez received a PhD in architecture from the University of Buenos Aires. He is a specialist in new technologies and biodigital design. He has taught undergraduate and postgraduate courses at different institutions including the University of Costa Rica, the International University of Catalonia, the Complutense University of Madrid, and the King Juan Carlos University. He is currently a lecturer at the School of Engineering and Technology of the International University of La Rioja (UNIR) where he is also the Academic Coordinator of the master's degree in Multimedia Design and Production. His research interests include the use of artificial intelligence methods and biology to aid the architectural design process.

Ismael Sagredo-Olivenza

Ismael Sagredo-Olivenza received a Ph.D. degree from the Complutense University in Madrid for his research in artificial intelligence applied to video game design and development. He worked as a Programmer in the video game industry in studies, such as Pyro Studios and Padaone Games. He is a currently a Professor with the High School of Engineering and Technology, International University of La Rioja (UNIR). He is also the Director of the MSC Program in videogames design and development with UNIR. In the last one, he developed some serious and educational games. His research interest includes use of artificial intelligence methods to help the video game design process.

Nadia McGowan

Nadia McGowan holds a degree in Cinematography (ECAM), Bachelor's in Art History (UNED), Master's in Screenwriting (UNIR), and Doctorate Audiovisual communication, advertising, and public relations (Universidad Complutense). She has worked at Notre Dame University and the Lebanese German University in Lebanon and currently is part of the Design Department of the School of Engineering and Technology at Universidad Internacional de La Rioja. Her research is focused on technical aspects of filmmaking.