ORIGINAL ARTICLE



Jumping into the artistic deep end: building the catalogue raisonné

Todd Dobbs¹ · Aileen Benedict¹ · Zbigniew Ras^{1,2}

Received: 12 August 2021 / Accepted: 23 November 2021 / Published online: 9 January 2022 © The Author(s) 2022, corrected publication 2022

Abstract

The catalogue raisonné compiled by art scholars holds information about an artist's work such as a painting's image, medium, provenance, and title. The catalogue raisonné as a tangible asset suffers from the challenges of art authentication and impermanence. As the catalogue raisonné is born digital, the impermanence challenge abates, but the authentication challenge persists. With the popularity of artificial intelligence and its deep learning architectures of computer vision, we propose to address the authentication challenge by creating a new artefact for the digital catalogue raisonné: a digital classification model. This digital classification model will help art scholars with new artwork claims via a tool that authenticates a proposed artwork with an artist. We create this tool by training a machine learning model with 90 artists having at least 150 artworks and achieve an accuracy of 87.31%. We use the ResNet Convolutional Neural Network to improve accuracy and number of artist classification by providing a consistent method to recreate our dataset and providing a consistent method to calculate performance metrics based on imbalanced data.

Keywords AI \cdot Painting/artist classification \cdot Deep learning \cdot Convolutional neural networks \cdot ImageNet \cdot ResNet \cdot Catalogue raisonné

1 Introduction

In 1751, Edme François Gersaint created the first catalogue raisonné for Rembrandt. This creation signifies the beginnings of a process to improve genuine art commerce and protect the amateur art collector. Figure 1 shows an image of Rembrandt's catalogue raisonné (Friedenthal 2020). Since the late eighteenth century, the catalogue raisonné has served as a complete record of an artist's work. Appraisers, artists, auction houses, collectors, curators, scholars, and students use the catalogue raisonné as a tool in their daily activities. A catalogue raisonné consists of a unique combination of

☑ Todd Dobbs btdobbs@uncc.edu

> Aileen Benedict abenedi3@uncc.edu

Zbigniew Ras ras@uncc.edu

¹ College of Computing and Informatics, The University of North Carolina at Charlotte, Charlotte, NC, USA

² Polish-Japanese Academy of Information Technology, Warsaw, Poland information of each piece of an artist's work. Information can include, for example, image, medium, provenance, and title. Recently, advances in technology such as the digitizing of materials and cloud computing spiked an interest in recompiling and augmenting the dated and nonexistent catalogue raisonné to address the issue of art and impermanence and to add new capabilities (Rogers 2015). Issues with impermanence and art surface when the medium of art deteriorates and is no longer restorable. For example, art made of organic materials may preserve indefinitely under ideal conditions. However, ideal conditions may not be possible during exhibitions or extreme happenstance such as fire or theft. Faulty painting techniques and materials can create conditions where a piece of art cracks or becomes discolored. Some artists create art such as David Medalla's columns of foam with limited life. These are but a few of many examples of the impermanence of art (Cannon-Brookes 1983).

Regardless of the medium of the catalogue raisonné, the issue of authenticity is pervasive due to questionable artworks resulting from loss due to theft or negligence. Documentation such as certificate of authenticity, past ownership, artist signature, and other physical attributes such as



Fig. 1 Rembrandt's catalogue raisonné from 1751 (Friedenthal 2020)

dimension, medium, and title of artwork represent artifact provenance. Such attributes evolve with progress and are supporting factors for the account of the artwork authenticity process. Authenticity is important for the account of process because the value of artwork is directly proportional to proper authentication. The opinion of the scholar who compiles a catalogue raisonné forms over time from their extensive research of an artist. In the end, it is paramount to the decision on whether a piece of art makes the cut to be included as an authentic piece of a collection. The market influences scholars with powerful clients who pressure this decision via legal action. These pressures surfaced in such events as the Warhol Authentication Board closing and the Knoedler Gallery forgeries scandal (Rogers 2015). New technological capabilities enable the existence of the online catalogue raisonné and digital storage of catalogue raisonné artifacts. With the popularity of modeling digital assets with machine learning algorithms in the past ten years, we believe a unique fingerprint or model that characterizes an artist's work is a useful, novel addition to an artist's catalogue raisonné. Such a model could further support the decision to authenticate or not to authenticate a piece of art with a collection. Modeling an artist based on their work is an image classification problem. Recent advances in machine learning and imaging have outperformed humans in tasks of image classification. A key project contributing to these advances is the ImageNet project (Deng et al. 2009).

The ImageNet project organizes a vast number of online images using an ontology of images built on the WordNet lexical database. The dataset produced from ImageNet lays a state-of-the-art foundation for image classification and training (Deng et al. 2009). In 2010, the ImageNet project formed

Year ImageNet Error rate 2010 Lin et al 28.2 2011 Sanchez and Perronnin 25.8 2012 Krizhevsky et al. (AlexNet) 16.4 2013 Zeiler and Fergus 11.7 2014 Simonyan and Zisserman (VGG) 7.3 2014 Szegedy et al. (GoogLeNet) 6.7 Human 5.1 2015 He et al. (ResNet) 3.6 2016 Shao et al. 3.0 2017 Hu et al. (SENet) 2.3

 Table 1
 Human and ImageNet error rates (Bianco et al. 2018; Russakovsky et al. 2015)

The bold font change calls out the human classification accuracy reference point

the basis needed to start the ImageNet challenge competition, which includes a variety of classification tasks for 1000 classes. The goal is for teams to compete to create deep neural networks that can outperform expert human annotators. The baseline human classification error to target is 5.1% (Russakovsky et al. 2015). In 2015, the ResNet architecture won the competition with a 3.57% error rate, thus surpassing expert human capability with image classification (He et al. 2016). While the ImageNet challenge continued through 2016, we focus on the ResNet architecture due to the combination of its simplicity and the minor performance increase of ensuing winners. Table 1 shows ImageNet winners from 2010 to 2017 (Bianco et al. 2018).

In this paper, we propose to make four contributions. First, we propose to increase the classification accuracy of artwork authentication for paintings using more classes than earlier experiments and a deeper ResNet architecture. Second, we propose to use the ResNet architecture to create a model for inclusion in an artist's catalogue raisonné to aid in the artwork authentication problem. Third, we address inconsistencies in the way scholars approach artist classification by providing a consistent method to recreate our dataset. Fourth, we address inconsistencies in the way scholars approach artist classification by providing a consistent method to calculate performance metrics based on imbalanced data.

The academic contribution of this paper is increased classification accuracy and class count using state-of-theart deep learning techniques for objects that a typical human observer would find difficult to discern. We also provide standard methods for recreating the data source and measuring results from imbalanced data. The paper is important to an interdisciplinary audience of art scholars and computer scientists. For art scholars, a born digital model of an artist's artwork is available to help with artwork authentication claims. For computer scientists, the complexities of an algorithm map an objective measure to the abstract nature of art. How this mapping works and can be improved supplies opportunities for continued research. For art scholars and computer scientists, we support continued research by providing standard methods for database recreation and result measurement.

In the next section, we review works relating to the problem of art identification. This includes an exploration of various art datasets used, as the data itself is critical, and existing methods of artist classification. In Sect. 3, we discuss the methods we use in our experiments to create a state-of-theart model to include in a catalogue raisonné. In Sect. 4, we explore our experiments in detail and show that our results outperform the current state-of-the-art models for WikiArt by artist count and accuracy. In the last section, we conclude with a discussion of future research related to this work.

2 Related work

Our approach to artwork authentication for the catalogue raisonné is to create and associate an artwork model generated from a convolutional neural network (CNN). We generate the artwork model with 90 artists to strengthen the binary class authentication claim for the artist in question (Abramovich and Pensky 2019). A catalogue raisonné is a comprehensive listing of an artist's known works. In a traditional sense, think of a catalogue raisonné like a book of art found on a coffee table, in a bookcase, or for purchase in a gift store of an art museum. A CNN is a complex computer algorithm inspired by visual biological processes that classify visual input. The output of a CNN is a mathematical model of classification. In this paper, we ascribe this model as a digital asset with a catalogue raisonné. This model must supply state-of-the-art accuracy and number of classifiers. The rest of this section discusses the historical effort and critical importance of this paper of compiling a digital art database and identifying artists based on their artwork using machine learning techniques.

2.1 Artist database

Our work uses the WikiArt dataset, a public source of data for artists and their artworks, including high-resolution images of art (Pirrone et al. 2009). All artwork from the WikiArt dataset has an associated artist so no anonymous or unknown artworks exists in the dataset. The dataset contains approximately 290 different artwork styles ranging from abstract to surrealism. We discuss related work using the WikiArt dataset as our primary focus. To a lesser extent, we explore related work using the Rijksmuseum dataset, which contains images of cutlery, furniture, maps, newsprint, paintings, sculptures, text, and other pieces of art (Mensink and van Gemert 2014). We also discuss work sourced from OmniArt, which combines data from WikiArt, Rijksmuseum Museum, and other sources (Strezoski and Worring 2018), and anime image datasets as a way of comparison of methodology and experimentation.

OmniArt combines data from WikiArt, Rijksmuseum Museum, and other sources. While this dataset is one of the most comprehensive datasets reviewed, the related experiments performed thus far fall short. For example, a seven artist classifier with 70.9% accuracy using a CNN similar to VGG, the 2014 ImageNet winner (Strezoski and Worring 2018). Likewise, experiments involving anime images yield a 93% classification rate for only five artists using the ResNet50 CNN (Kondo and Hasegawa 2020).

2.2 Artist classification

According to Johnson et al. (2008), the availability of highresolution images prompted more research utilizing van Gogh paintings in 2008, thus forming the art authentication problem's foundations. This study of 101 paintings revealed that classification through machine learning is possible using the fluency, geometry, style, and texture of a painting. Of these 101 paintings, 82 are well-known van Gogh, 13 are questionable van Gogh according to experts, and six are not van Gogh. Comparing all paintings' textures using a Gabor wavelet decomposition and support vector machine (SVM) classification, four of the six non-van Gogh classified as van Gogh. Moreover, two van Gogh paintings were classified as non-van Gogh. Art experts consider this analysis of texture to detect enough dissimilarity in brushstrokes to support authenticity assessment (Johnson et al. 2008). This binary experiment is for van Gogh and a group of six artists, and the accuracy of classification is 94%.

Soon after the van Gogh experiments by Johnson et al. (2008) and the WikiArt dataset creation by Pirrone et al. (2009), related research continued for multiple artists in 2010 and 2011. Blessing and Wen (2010) ran experiments on seven artists and achieved 85.13% using histogram of oriented gradients (HOG) for feature extraction and SVM for classification. The data for this experiment sources from Google image search (Blessing and Wen 2010). We consider this source of data closely tied to WikiArt because all artists were publicly available through WikiArt at the time of this experiment. Moreover, all these artists are part of the experiments conducted in this paper. Influenced by Blessing and Wen's work, Jou and Agrawal conducted similar

experiments using histogram of oriented gradients (HOG) for feature extraction and Naïve Bayes for classification. This approach leads to a reduced accuracy of 65% with less artists. It is important to note that the data for this experiment sources from specific websites for each artist, and two artists are not part of the Blessing and Wen experiments (Jou and Agrawal 2011). We consider this source of data closely tied to WikiArt because all artists were publicly available through WikiArt at the time of this experiment. Moreover, all these artists except one are part of the experiments conducted in this paper.

The number of artists in experiments using data from WikiArt greatly increases with the use of convolutional neural networks (CNN) after the ImageNet challenge starts in 2015 (Russakovsky et al. 2015). In 2017, Viswanathan produced the most notable of these experiments using the ResNet 18 algorithm with transfer learning to achieve 77.7% accuracy with 57 artists. For this experiment, the artists have at least 300 paintings each (Viswanathan and Stanford 2017). While this experiment does not supply an exact list of artists, the 300-painting threshold places these artists in a subset of the artists used in this experiment. Moreover, the method used in Viswanathan's experiment is closely related to the experiment in this paper. We mention two related experiments using WikiArt and CNNs. While the results we are interested in pale in comparison to Viswanathan's results, they are important to mention to show the varied research in the area. First, using a variation of Viswanathan's CNN design, a 15 artist classifier with 74.7% accuracy. It is important to note that the experiments in Viswanathan's paper are geared toward a comparison between using CNNs and SVMs and the setup involved for both (Chen 2018). Cetinic et al. (2018) develop an experiment using 23 artists and CaffeNet, a CNN derived from the 2012 winner of the ImageNet challenge called AlexNet. This method achieves a 79.1% accuracy, and the team explores more classification experiments of genre, style, timeframe, and nationality (Cetinic et al. 2018). The last two experiments explicitly list the artists of which all exist within the domain of artists used in this paper.

Similar experiments using data from the Rijksmuseum Museum produce promising results. In 2013 the Rijksmuseum Museum started a series of challenges to name the artist, type, material, and creation year of their art using computer science techniques. In 2014, the first experiment used SVM to classify 100 artists with 76.3% accuracy using a 96-dimension Fisher vector based on scale-invariant feature transform (SIFT). It's important to note that the algorithm uses the top 100 performing artists from an initial pool of 374 artists and an initial classification accuracy of 59.1% (Mensink and van Gemert 2014). In 2015, Van Noord et al. extends this work with a focus on paintings. Using PigeoNet, a CNN derived from CaffeNet and AlexNet, the

top 78 artists in the dataset that are the least likely to be confused are classified with 73.3% accuracy (van Noord et al. 2015). In 2017, OmniArt developed a multi-task deep learning method that, when applied to the Rijksmuseum Museum challenge, produced 81.9% accuracy for the top 52 artists in the dataset that are the least likely to be confused (Strezoski and Worring 2017).

Experiments with data sourced from OmniArt and anime produce results with good accuracy but few classes. Using OmniArt, a seven artist classifier with 70.9% accuracy using a CNN similar to VGG, the 2014 ImageNet winner (Strezoski and Worring 2018). Performance and number of classifiers are improved using the OmniArt multi-task deep learning method. Experiments yield 80.8% accuracy for 87 artists (Strezoski and Worring 2017). Likewise, experiments involving anime images yield a 93% classification rate for only five artists using the ResNet50 CNN (Kondo and Hasegawa 2020).

2.3 Summary

Our research aims to improve on existing work that uses a subset of the WikiArt data in our experiment. Given the related work, we take on the task of producing an experiment that will improve upon Viswanathan's work. This will create a model for inclusion in an artist's catalogue raisonné to aid in the artwork authentication problem.

3 Methods

Our goal is to build a system that inputs a single image of a painting and outputs an artist label. Our system must be able to handle red, green, and blue additive color model (RGB) images. Our target is to classify twice as many artists with 250 + paintings with comparable accuracy to Viswanathan's experiment, which reports a 77.7% accuracy with 57 artists having 300 + paintings. We do not plan to cherry-pick artists based on their performance to maximize accuracy because we aim to generically show the style of an artist with a random sample of artists with base proliferation. The goal is to maximize the number of artists and accuracy because both metrics strengthen the model to add to the catalogue raisonné. We carry out this goal using a state-of-the-art CNN architecture. We show a pictorial of our method in Fig. 2. In this figure, an artist's paintings feed a CNN to create a model. An art scholar attaches the model to an artist's catalogue raisonné and uses it as a tool to aid future claims for adding new art to the catalogue raisonné.

Specifically, we implement ResNet 101 CNN with ImageNet transfer learning. ResNet 101 is the 2015 ImageNet winner and grants ease of implementation and solid performance on detection, localization, and segmentation aspects



Fig. 2 A pictorial of creating an artist's model from a CNN and associating it with a catalogue raisonné

of the challenge. We decide to bypass the implementation of the 2016 and 2017 ImageNet winners due to the increased implementation complexity, which would theoretically only allow a gain of 0.6–1.3% (Bianco et al. 2018; He et al. 2016).

The ResNet CNN introduces the concept of residual learning. Residual learning addresses the accuracy degradation problem that arises as the depth of CNNs increase. Research found that accuracy can diminish after making a change that should logically produce better results. On the one hand, extra layers increase the performance of the CNN. On the other hand, research shows that blindly adding layers diminishes accuracy because enduring discoveries fade due to a vanishing gradient. Residual learning addresses this problem by ensuring these discoveries persist as layers of the network are added (He et al. 2016).

By way of comparison, consider the activities associated with the classic shape sorting child's toy. In this activity, a child receives a variety of colored, three-dimensional, wooden shapes. The goal is to fit these shapes into a wooden box via a two-dimensional opening. There are a variety of things to consider when fitting each shape into the corresponding box opening. For example, objective considerations like shape type, shape size, shape orientation, shape velocity, shape acceleration, hole type, hole size, and hole orientation determine a fitting outcome. Other subjective considerations like color or pattern matching may exist for an added challenge. If we use a robot to perform this activity, we can map these considerations to separate learning layers of a CNN. Obviously, scenarios exist where we do not want to lose residual accomplishments as learning progresses, and a model begins to form. For example, we do not want to lose key residual learning with respect to what is known about placing a cube into a square hole when learning the subjective measure of color as a blue cube is placed into a square blue hole rather than a square red hole.

How does this shape sorting activity relate to classifying art? Like the shapes in the sorting activity, a painting consists of color and shape or texture. Research shows that an artist's style alone contributes a significant amount to art classification. For example, through feature learning of a CNN versus feature engineering and clustering, artist classification for single and dual authorship show that a distinctive visual texture is present even in areas that appear empty to the human eye (van Noord et al. 2015). The notion that more CNN layers increase performance supports the mapping needed for the multitude of layers necessary to represent the vast number of ways to approach the style of a painting. Therefore, the concept of deeper CNNs and therefore deeper residual learning is necessary to yield greater CNN performance for art authentication.

4 Experiment

We benchmark our ResNet 101 implementation with a previously published ResNet 18 implementation (Viswanathan and Stanford 2017). Both Resnet implementations use the MATLAB deep learning toolbox (Kim 2017) and use the same data from WikiArt, which uses artists with 250 or more paintings (Pirrone et al. 2009). We compare precision, recall, F1 score, accuracy, and mean class accuracy (MCA) overall and at the class level for both our experiment and the baseline to evaluate the performance.

4.1 Data

We acquired data for our experiments from WikiArt using a custom download tool and the WikiArt API. We query all artists and download an artist's artworks if they have 250 or more paintings. We only download RGB formatted images. In some cases, we retrieve less than 250 artworks due to invalid formats. Overall, we downloaded 45,974 paintings for 90 artists. The most paintings downloaded are for Vincent Van Gogh, with a total of 1931. The fewest paintings downloaded are for George Grosz, with a total of 158. A

 Table 2
 Select artist artwork distribution along with the training, validation, and test counts used in experiments

Artist	Artwork count	Training count	Valida- tion count	Test count
Claude Monet	1366	956	205	205
Francisco Goya	284	199	43	42
Henri Matisse	999	699	150	150
Pablo Picasso	1139	797	171	171
Rembrandt	765	536	115	114
Salvador Dali	1164	815	175	174
Vincent Van Gogh	1931	1352	290	289

select and full distribution of artists is shown in Tables 2 and 4, respectively (Pirrone et al. 2009). We share our work on GitHub to recreate our WikiArt data source and verify the artwork used in our experiments.

One challenge with this dataset is the class imbalance. The ImageNet dataset does not declare class balance as a prevailing property, but its designers mention the importance of class balance when comparing their dataset to related datasets (Deng et al. 2009). For the ImageNet challenge, the focus is on the accuracy of classification and object detection. There are no class balance measures, which leaves the responsibility of handling class imbalance to competitors (Russakovsky et al. 2015). Moreover, the topic of balancing input for CNNs remains an active area of research since larger numbers of observations are encouraged for each class for performance (Johnson and Khoshgoftaar 2019). We can address the class imbalance through input data modification or out measure calculation. From an input perspective, research shows that oversampling handles class imbalance optimally with respect to multi-class true-positive rate (TPR) and false-positive rate (FPR) (Buda et al. 2018). From a measurement perspective, research shows that micro-balanced accuracy based on true-positive rate (TPR) and falsepositive rate (FPR) is a good predictor when there is a concern for under-represented classes (Grandini et al. 2020). For our research, we choose to handle class imbalance using the micro-balanced accuracy measurement. We choose this approach to learn as much as possible from each artist and for simplicity of implementation. Moreover, we found no research showing oversampling outperforms micro-balanced accuracy for the CNN multi-class imbalance problem. We share our work on GitHub to recreate result measures for our experiments.

A common technique to maximize experiment results is to select the top *n* true-positive artists from a larger class experiment. These top-performing artists feed later experiments, which boosts accuracy metrics (van Noord et al. 2015). For this experiment, we refrain from this tactic and use all artists selected for the experiment regardless of performance. We do this to explore the opportunities presented from the analysis of weaker performing artists.

4.2 Training details

Training details are identical for the baseline and proposed experiment. We use default hyperparameter values from MATLAB for the first experiment. These default hyperparameter values end up producing solid results. The only default that we change is the data split between training, validation, and test. The default splits the data set into 70% training and 30% validation. We change this to 70% training to allow for test data, and the rest splits into 15% validation and 15% testing. Training data creates a model by learning from the data. Validation data checks for accuracy during training. Test data tests model accuracy once validation accuracy is acceptable. The full distribution of artists, training, validation, and test split counts are shown in Table 2 (Pirrone et al. 2009).

Input painting images are resized to match the network's input size, which is $224 \times 224 \times 3$. We randomly rotate paintings between – 90 degrees and 90 degrees, randomly scale paintings between one to two times the original size, and randomly reflect paintings on the x-axis. The solver used is stochastic gradient descent with momentum (SGDM) with a learning rate of 0.01 and a momentum of 0.9. Training passes through the data set 30 times (30 epochs), with validation occurring after 50 iterations. The epoch count of 30 is the default of MATLAB and gives ample iterations for validation accuracy saturation. If experimentation shows a monotonic increase of accuracy with each epoch, repeating the experiment with a higher epoch count is necessary. With each iteration, a mini batch size of 128 images processes through the CNN. The mini batch corresponds to the subset of the training data that evaluates the gradient of the loss function and updates the weights through backpropagation. After each epoch, training data shuffles paintings to handle the situation where the mini batch size does not equally partition the data. To reduce overfitting, a weight decay regularization term with a value of 0.0001 adds to the loss function.

To give an example of how paintings train and crossvalidate, it is helpful to review the processing of an epoch. Given that we have 45,974 paintings, we use 70% of these data or 32,181 paintings for training. Given that we process paintings in batches of size 128, the training process crossvalidates every 250 iterations. Note that 250 iterations multiplied by a batch size of 128 is 32,000 paintings. However, there are 32,181 paintings for training. To account for the remaining 181, we shuffle paintings after each epoch. We continue this process for 30 iterations. We visualize this entire process, displaying the accuracies and losses over the iterations, in Figs. 4, 5, 7, and 8.

The execution environment is set to parallel, which takes advantage of multiple CPU cores and GPUs. The environment is set to process on one node in a high-performance cluster (HPC) using four cores, each of which has two GPUs. The specific hardware for this node is dual 8-Core Intel Xeon Silver 4215R CPU @ 3.20 GHz (16 cores total) with 192 GB RAM (12 GB/core) and 8×Titan V GPUs (12 GB HBM2 RAM per GPU).



Fig. 3 ResNet 18 CNN Architecture with a focus on convolutions



Fig. 4 Progress of ResNet 18 model plotting the training and validation curves for accuracy by iteration

4.3 Baseline experiment

The baseline experiment uses a ResNet 18 CNN architecture. This architecture has 71 layers and 78 connections. We show a visual of the layers and connections with a focus on convolutions of the architecture in Fig. 3. Note that we group similar convolutions by color and scale up in the number of convolutions performed with respect to the depth in the stack. We display residual convolutions with a dashed box and transition convolutions with a dotted box. The convolutions in Fig. 3 couple batch normalization and ReLU activation function steps, both of which remain hidden to conserve space. It took 6 h and 5 min to train the model. The training and validation accuracy and loss are in Figs. 4 and 5, respectively. The aim of training is to maximize accuracy and minimize loss. The accuracy represents



Fig. 5 Loss of ResNet 18 model plotting the training and validation curves for loss by iteration

how well predications are made, and the loss represents the errors in prediction. The blue curve represents training accuracy in Fig. 4 and training loss in Fig. 5, and the red curve represents validation accuracy in Fig. 4 and validation loss in Fig. 5. The training accuracy is a result of the specific iteration while the validation accuracy takes all iterations into account. We report on the validation numbers. We perform this experiment to compare with Viswanathan's experiment, which uses a ResNet 18 CNN architecture on 57 artists, and our proposed experiment, which uses a ResNet 101 CNN architecture on 90 artists.

4.4 Proposed experiment

The proposed experiment uses a ResNet 101 CNN architecture. This architecture has 347 layers and 379 connections. From a network layer perspective, the ResNet 101



Fig. 6 ResNet 101 CNN Architecture with a focus on convolutions



Fig. 7 Progress of ResNet 101 model plotting the training and validation curves for accuracy by iteration

architecture has 276 more layers than ResNet 18. A visual of the layers and connections with a focus on convolutions of the architecture are in Fig. 6. Note, we do not repeat the details on the architecture because they are the same as the ResNet 18 CNN architecture mentioned above. Other than the number of convolutions, the major difference between the ResNet 18 and ResNet 101 CNN architecture is the grouping of multiple convolutions and the combination of residual and transition convolutions, which we show with a dashed and dotted box. It took seven hours and 46 min to train the model. The training and validation accuracy and loss are in Figs. 7 and 8, respectively. The aim of training is to maximize accuracy and minimize loss. The accuracy represents how well predications are made, and the loss represents the errors in prediction. The blue curve represents training accuracy in Fig. 7 and training loss in Fig. 8, and



Fig. 8 Loss of ResNet 101 model plotting the training and validation curves for loss by iteration



Predicted Class

Fig. 9 Confusion matrix for baseline experiment showing the blue diagonal of true-positive predictions and red points of false negatives and false positives



Fig. 10 Confusion matrix for proposed experiment showing the blue diagonal of true-positive predictions and red points of false negatives and false positives

the red curve represents validation accuracy in Fig. 7 and validation loss in Fig. 8. The training accuracy is a result of the specific iteration while the validation accuracy takes all iterations into account. We report on the validation numbers. We perform this experiment to show both improvement

in accuracy and artist count with respect to Viswanathan's experiment.

4.5 Results

Tests using the baseline and proposed experiment models produce the two confusion matrices shown in Figs. 9 and 10. Both matrices have total-normalized artwork counts to account for the fact that some artists have more artwork than others (i.e., the data's imbalanced nature). The saturation of blue on the diagonal stands for the number of a true-positive predictions. The saturation of red outside of the diagonal stands for the number of a false negative predictions on the horizontal axis and false-positive predictions on the vertical axis. These confusion matrices supply a high-level visual that supports the fact that our results produce more true-positive results versus false negative and positive results. From the raw values of these confusion matrices, we calculate measures for all the baseline and proposed experiments listed in Table 3. We set the alpha or significance level to a typical value of 0.05 stating that we would like to be 95% confident that our analysis is correct. Given the micro-balanced accuracy of the 90 artists using ResNet 18 and ResNet 101, we arrive at a p value of 0.01001732672. Since our observed p value is lower than alpha, we conclude that our results are statistically significant. By way of comparison, the unbalanced accuracy of the 90 artists using ResNet 18 and ResNet 101 provides a p value of 0.03491554592. This p value is lower than alpha and is statistically significant. Using balanced data calculations provides a stronger p value for our experiments. We compare these measures calculated from the confusion matrices with Viswanathan's experiment in the analysis section.

We calculate measures for multi-class classification based on a generalization of binary measures from a confusion matrix generated from the test data set and training model. Macro measures are an average of the class measures. Micro measures are a sum of the class measures before measure calculation. We add measures for error rate and the macro and micro versions of precision, recall, F1 score, and accuracy (Sokolova and Lapalme 2009). Furthermore, we add Grandini's macro and micro versions of the balanced accuracy measure to address class imbalance (Grandini et al. 2020). All future measure references will be at the micro level. We leave the macro calculations for reference. We list all the formulas used in this paper in the next section.

4.6 Result formulas

$$\begin{aligned} fp &= false \ positive & fn = false \ negative & tp = true \ positive & tn = true \ negative \\ \mu &= micro \ calculation & M = macro \ calculation \\ \end{aligned} \\ From Rate_{\mu,M} &= \frac{\sum_{i=1}^{l} \frac{fp_i + fn_i}{tp_i + fn_i + fp_i + tn_i}}{l} & Average \ Accuracy_{\mu,M} &= \frac{\sum_{i=1}^{l} \frac{tp_i + tn_i}{tp_i + fn_i + fp_i + tn_i}}{l} \\ Balanced \ Accuracy_{\mu} &= \frac{\sum_{i=1}^{l} tp_i}{2} + \frac{fn_i}{tp_i + fn_i} + \frac{\sum_{i=1}^{l} tn_i}{2} \\ Balanced \ Accuracy_{\mu} &= \frac{\sum_{i=1}^{l} tp_i + fn_i}{2} + \frac{\sum_{i=1}^{l} tn_i}{2} \\ Precision_{\mu} &= \frac{\sum_{i=1}^{l} tp_i}{tp_i + fp_i} \\ Recall_{\mu} &= \frac{\sum_{i=1}^{l} tp_i}{\sum_{i=1}^{l} tp_i + fn_i} \\ F1 \ Score_{\mu} &= \frac{2 \times Precision_{\mu} \times Recall_{\mu}}{Precision_{\mu} + Recall_{\mu}} \\ \end{aligned}$$

4.7 Analysis

Our results indicate that there is an 87.31% chance to identify one of the 90 artists given one of the 45,974 paintings in our dataset. The probability of randomly guessing an artist is 1.11%. The best chance to randomly guess an artist is 4.2% for Vincent Van Gogh. There are 290 different styles of art in our dataset, we are confident that our proposed algorithm will produce similar results for a different set of 90 artists with their own style of creative curiosity. The algorithm works because it learns the texture and colours produced from an artist's imagination, brush strokes, and colour selection.

We analyze the results in Table 3. First, we review accuracy. Next, we compare the ResNet 18 baseline versus ResNet 101 proposed experiments. We then show improvement from Viswanathan's work with our ResNet 18 baseline and ResNet 101 proposed experiments with a focus on performance and class count. Lastly, we look at artists with the best and worst performance with respect to artwork count, image similarity, and mean-squared error.

Measure	ResNet 18 (Baseline) (%)	ResNet 101 (Proposed) (%)	Improvement (%)
Macro/micro error rate	0.67	0.56	- 19.64
Macro/micro accuracy	99.33	99.44	0.11
Micro balanced accuracy	84.68	87.31	3.01
Macro balanced accuracy	84.48	86.89	2.77
Micro precision	69.70	74.90	6.94
Macro precision	67.78	72.96	7.10
Micro recall	69.70	74.90	6.94
Macro recall	69.29	74.06	6.44
Micro F1 score	69.70	74.90	6.94
Macro F1 score	68.53	73.50	6.76

Table 3An increase for all
measures from the baseline to
proposed experiments using test
data (Sokolova and Lapalme
2009)

4.7.1 Accuracy

We note that the macro and micro accuracy of the ResNet 18 baseline and ResNet 101 proposed artists are unusually high and rival the best accuracies in the ImageNet challenge. These accuracies are high because we are using unbalanced data, and this further supports the need to use the micro-balanced accuracy measures in our analysis. Moving forward in our analysis, we use the term accuracy in place of micro-balanced accuracy for brevity.

4.7.2 ResNet 18 baseline vs. ResNet 101 proposed

With this comparison, we see that all measures improved from our baseline ResNet 18 experiment to our proposed ResNet 101 experiment. This experiment is new for 90 WikiArt classes of artists, and the problem of classifying a painting is much more open-ended than that of the specific images in ImageNet. However, we expected improved results since we increase the depth of the CNN and use residual learning, both of which work together to allow for the performance increase. According to He et al. (2016), the increase of 3.01% in accuracy is on par with similar depth increases shown in residual learning research.

4.7.3 Viswanathan vs. ResNet 18 baseline

Accuracy improves by 8.24% from Viswanathan's experiment to the baseline ResNet 18. However, precision and recall decrease by 11.48% and 11.05%, respectively. This discrepancy is because the former experiment has 63.33% of the latter experiment's artists' classes. Moreover, the former experiment uses a random sample of balanced data, while the latter experiment uses all samples and is imbalanced. The source of both experiments is WikiArt, and we verify that the 57 artists used in Viswanathan's experiment is a subset of the 90 artists used in our experiment. We are unable to find the specific pieces of art to reproduce Viswanathan's experiment exactly, but the extra learning from the increased classes with the increased accuracy as evidence shows an overall improvement. Moreover, in his research, Viswanathan concludes that a future experiment using the method we implement in this paper should yield an increased accuracy.

4.7.4 Viswanathan vs. ResNet 101 proposed

Accuracy improves by 11.01% from Viswanathan's experiment to the proposed ResNet 110 experiment. However, precision and recall decrease to 3.74% and 3.33%, respectively. We explain this discrepancy using the same rationale in Sect. 4.7.3. The only difference we see here is that accuracy increases, and the precision and recall difference shrink. These measure improvements are a direct result of using a



Fig. 11 Artwork count vs. accuracy

deeper CNN with residual learning. We use this analysis of our results as the final basis to satisfy the state-of-the-art method to provide a CNN model to assist with the artwork authentication problem for the catalogue raisonné.

4.7.5 Calculating accuracy analysis measures

To rule out the correlation between accuracy and simple engineered features of an artist's artworks, we analyze our accuracy results for each artist by comparing with their artwork count, similarity, and estimator measures. It is important to show no correlation to support the viability of our learned models. For artwork count, we count the number of artworks used in our experiments for each artist. For similarity, we calculate the average structured similarity index (SSIM) between all the combinations of two artworks for an artist. For the estimator, we calculate the average mean-squared error (MSE) between all the combinations of two artworks for an artist. Before analysis, we augment the artwork images to the same as the input size of the experiment CNN network, which is $224 \times 224 \times 3$.

For the similarity and estimator measures, we use the binomial coefficient formula to figure out the number of calculations needed for each artists' artworks taken two at a time. We use the following formula for each artist where n is the number of their paintings and k is 2:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

The sum of these combinations results in 17,018,158 calculations needed for each SSIM and MSE. For this number of calculations, we need to use an HPC. The calculations take 17 h and 6 min to process, and the execution environment is set to process on one node using 12 cores and





Fig. 12 Mean SSIM vs. accuracy

128 GB of RAM. The specific hardware for this node is Dual 24-Core Intel Xeon Gold 6248R CPU @ 3.00 GHz (48 cores/node) 384 GB RAM (8GBs/core).

4.7.6 Artwork count vs. accuracy

The purpose of this analysis is to verify that there is no major impact on artist accuracy based on an artists' number of artworks. Moreover, we want to verify that our minimum number of 158 pieces of artwork for learning is sufficient. We display the result of the artwork count versus accuracy analysis in Fig. 11. To compare artwork counts with the accuracy of each artist in the same pictorial, we normalize artwork counts. We also sort by artwork counts to aid in the visualization between the artwork count and accuracy curves. Due to space restrictions, we do not list all artist names, but we do call out the artists' minimum and maximum accuracies with a black dot on both curves. The accuracy moves between the minimum and maximum accuracy values independent from artwork count, thus visually showing no correlation between the two measures. From this analysis, we are confident that there is no major impact on accuracy based on the number of artworks for each artist. Moreover, we are confident that 158 pieces of artwork are sufficient for learning an artist's style.

4.7.7 Mean SSIM vs. accuracy

Li et al. define SSIM as a measure that assesses the visual impact of the luminance, contrast, and structure characteristics of an image (Li and Bovik 2009). The formula used to calculate SSIM is as follows where $\mu_x, \mu_y, \sigma_x, \sigma_y$, and σ_{xy} are the local means, standard deviations, and cross-covariance for images x and y. C_1 and C_2 are constants to prevent division by zero:

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$
(Li and Bovik 2009)

Fig. 13 Mean MSE vs. accuracy

An SSIM between two images with an upper bound value of one specifies that the images are the same. The minimum value of SSIM is zero, showing a maximum difference between two images. Our goal is to obtain an average SSIM value for an artist, given all the possible combinations of an artist's paintings. We aim to show that the similarity of an artist's paintings does not significantly impact artist accuracy. We display the result of SSIM versus accuracy analysis in Fig. 12. We do not need to normalize SSIM for our analysis because the domain of SSIM values is in proportion to accuracy. We sort by average SSIM to aid in the visualization between the average SSIM and accuracy curves. Due to space restrictions, we do not list all artist names. However, we do call out the artists' minimum and maximum accuracies with a black dot on both curves. Like artwork count, the accuracy moves between the minimum and maximum accuracy values independent from average artwork SSIM, thus visually showing no correlation between the two measures. There is one exception in that our artist with the highest accuracy correlates to the artist with minimum similarity. However, we note at least five other artists with high accuracy and similarity scores spaced out amongst the whole spectrum of similarity. From this analysis, we are confident that there is no major impact on accuracy based on each artist's similarity of artworks.

4.7.8 Mean MSE vs. accuracy

Pishro-Nik defines MSE as a measure that assesses the quality of an estimator (Pishro-Nik 2014). The formula used to calculate MSE is as follows where x and y are the images to compare and *n* is the number of pixels to compare: $MSE(x, y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - y_i)^2$ (Pishro-Nik 2014)

An MSE between two images with a value closer to zero is better because it shows an overall smaller difference in the image's pixel values. Our goal is to obtain an average MSE value for an artist, given all the possible combinations of an artist's paintings. We aim to show that the estimator of an artist's paintings does not have a major impact on artist accuracy. We display the result of MSE versus accuracy analysis in Fig. 13. We normalize MSE for our analysis because the domain of MSE values is not in proportion to accuracy, which makes the visual comparison of curves impossible. We sort by average MSE to aid in the visualization between the average MSE and accuracy curves. Due to space restrictions, we do not list all artist names. However, we do call out the artists minimum and maximum accuracies with a black dot on both curves. Like artwork count and average SSIM, the accuracy moves between the minimum and maximum accuracy values independent from average artwork MSE, thus visually showing no correlation between the two measures. From this analysis, we are confident that there is no major impact on accuracy based on an estimator of artworks for each artist.

4.7.9 Best-performing artist

Kenneth Noland has the best classification accuracy measure of 98.52%. We downloaded 271 of his artworks. Our model trains from 190 (70%) of his artworks, and we calculate the accuracy measure from the test data of 40 (15%) of his artworks. Given our analysis, the number of artworks, artwork similarity, and estimator does not affect accuracy in a significant way. Kenneth Noland was an American abstract painter who was one of the best-known color field painters. Kenneth Noland has many more false negatives than false positives, meaning that these paintings are classified with other artists. Of the false negative artists, none are abstract color field painters (Pirrone et al. 2009). However, several of these artists have many false negatives with the other artists in our research, which leads us to believe that there are either common missed opportunities for learning by the ResNet 101 architecture or intractable situations for learning for these artists.

4.7.10 Worst performing artist

Alfred Sisley has the worst classification accuracy measure of 72.04%. We downloaded 471 of his artworks. Our model trains from 330 (70%) of his artworks, and we calculate the accuracy measure from the test data of 70 (15%) of his artworks. Like our best performing artists, our analysis does not show that the number of artworks or artwork similarity and estimator impact accuracy in a significant way. According to Pirrone et al. (2009), Alfred Sisley was a French impressionist landscape painter who rarely deviated from painting landscapes. Reviewing our experiment confusion matrix for Alfred Sisley, he was predominately confused as false-positive and false negative

with Camille Pissarro and Claude Monet, who are both French impressionists (Pirrone et al. 2009). Out of the false classifications, Alfred Sisley's false positives are more prominent, which means that paintings by Camille Pissarro and Claude Monet classify incorrectly as Alfred Sisley rather than the other way around. Both false classifications have two to three times as many artworks. However, Pyotr Konchalovsky has a similar artwork count to Camille Pissarro and Pierre Auguste Renoir has a similar artwork count to Claude Monet, and both artists have two false classifications with Alfred Sisley. Therefore, there is no correlation between the number of artworks for an artist and false classification count.

5 Conclusion

In this paper, we introduce the idea to include a born digital classification model to the catalogue raisonné to aid art scholars with the artist authentication and impermanence problems. We improve artist classification using WikiArt data with a model that improves on earlier work from an accuracy and number of class perspective. Specifically, we increase accuracy by 11.01-87.31% and the number of classes by 57.89-90%. We use the ResNet 101 CNN to carry out this increase in performance. We also show that the number, similarity, and estimator characteristics of an artist's artworks do not have a major influence of the accuracy of our trained models. These improvements supply an academic contribution for art scholars and computer scientists to use and extend. Art scholars obtain an object born digital which will bolster the denial or support of claims, and the computer scientist discovers a new application and research opportunity for an algorithm which improves classification accuracy and class count measures for otherwise indiscernible objects. Lastly, we share code artifacts and methods to recreate our data source and result performance measurements.

In future work, we would like to aid art scholars with improved accuracy and number of classes using a deeper CNN and a CNN with augmented layers beneficial to painting classification. In showing how artwork count, similarity, and estimator aspects do not have a major impact on accuracy, we would like to conduct experiments that give a better understanding of the salient features that aid in learning. We also believe adding style as a decision attribute in addition to our model attribute will increase classification performance. Moreover, we believe that it is possible to increase accuracy by creating a binary classifier for each artist with respect to the remaining group and adding these binary classifiers into a composition for classification. In addition to using the WikiArt collection, we would like to apply our work to Rijksmuseum data and contemporary art collections. Lastly, we would like to work closely with art historians to figure out the best number of artists' classes and classification accuracy for model usefulness in a catalogue raisonné.

The work in this paper supports the future trends we see emerging as AI applies to art history collections. Primarily, we see applications to authentication, generative art, style transfer, and born digital artefacts. From an authentication perspective, we believe further analysis of results as the accuracy and class count increases will help explain what aspects of an artist's paintings are most helpful with classification. As the confidence of generated models increases with art historians, we expect these models to be ubiquitous as part of an art scholar's decision, but not as a full replacement. As we glean a better understanding of artist classification, we expect aspects of the generated model as useful with addressing issues when generating new art or transferring the style of artist to an existing piece of art. Lastly, we are optimistic that authentication through AI will foster the catalogue raisonné as a born digital artefact by default.

Appendix

See Table 4

Table 4All artist artworkdistribution along with the
training, validation, and test
counts used in experiments

Artist	Artwork count	Training count	Validation count	Test count
Albert Bierstadt	336	235	51	50
Albrecht Durer	707	495	106	106
Alexander Roitburd	264	185	40	39
Alfred Freddy Krupa	598	419	90	89
Alfred Sisley	471	330	71	70
Amedeo Modigliani	349	244	53	52
Boris Kustodiev	645	452	97	96
Byzantine Mosaics	255	179	38	38
Camille Corot	498	349	75	74
Camille Pissarro	881	617	132	132
Charles M Russell	278	195	42	41
Chicote Cfc	307	215	46	46
Childe Hassam	550	385	83	82
Claude Monet	1366	956	205	205
David Burliuk	400	280	60	60
Edgar Degas	625	438	94	93
Egon Schiele	299	209	45	45
Ernst Ludwig Kirchner	387	271	58	58
Eugene Boudin	560	392	84	84
Eyvind Earle	422	295	64	63
Felix Vallotton	314	220	47	47
Ferdinand Hodler	256	179	39	38
Fernand Leger	446	312	67	67
Francis Bacon	312	218	47	47
Francisco Goya	284	199	43	42
George Grosz	158	111	24	23
George Stefanescu	264	185	40	39
Giovanni Battista Piranesi	1353	947	203	203
Gustave Courbet	270	189	41	40
Gustave Dore	389	272	59	58
Gustave Loiseau	258	181	39	38
Henri De Toulouse Lautrec	372	260	56	56
Henri Fantin Latour	289	202	44	43
Henri Martin	408	286	61	61
Henri Matisse	999	699	150	150
Honore Daumier	254	178	38	38
Hryhorii Havrylenko	408	286	61	61
Ilya Repin	541	379	81	81

Table 4 (continued)

Artist	Artwork count	Training count	Validation count	Test count
Isaac Levitan	449	314	68	67
Ivan Aivazovsky	579	405	87	87
Ivan Shishkin	522	365	79	78
Jacek Yerka	308	216	46	46
James Tissot	432	302	65	65
Jean Auguste Dominique Ingres	259	181	39	39
Joaqu N Sorolla	365	256	55	54
John Henry Twachtman	255	179	38	38
John Singer Sargent	800	560	120	120
Katsushika Hokusai	265	186	40	39
Kazimir Malevich	360	252	54	54
Kenneth Noland	271	190	41	40
Konstantin Korovin	317	222	48	47
Konstantin Makovsky	366	256	55	55
Konstantin Somov	254	178	38	38
Konstantin Yuon	293	205	44	44
Louis Comfort Tiffanv	261	183	39	39
Lucian Freud	283	198	43	42
M C Escher	469	328	71	70
Marc Chagall	1018	713	153	152
Martiros Sarian	551	386	83	82
Mary Caseatt	304	213	46	45
Mauriaa Brandargaat	370	215	4 0 57	
Max Ernst	269	203	55	57
Niakolas Booriah	1824	1284	55 275	55 275
Odilan Badan	1654	1264	213	213
	433	319	171	171
Pablo Picasso	1139	/9/	1/1	1/1
Paul Cezanne	587	411	88	88 77
Paul Gauguin	512	358	77	11
Peter Paul Rubens	395	211	59	59
Pierre Auguste Renoir	1409	986	212	211
Pyotr Konchalovsky	925	648	139	138
Raphael Kirchner	525	368	/9	/8
Rembrandt	765	536	115	114
Rene Magritte	371	260	56	55
Robert Henri	263	184	40	39
Roger Weik	502	351	76	75
Salvador Dali	1164	815	175	174
Samuel Peploe	252	176	38	38
Stanley Spencer	270	189	41	40
Theodor Severin Kittelsen	375	263	56	56
Theophile Steinlen	1136	795	171	170
Thomas Eakins	306	214	46	46
Titian	245	172	37	36
Utagawa Kuniyoshi	418	293	63	62
Vasily Surikov	267	187	40	40
Veletanlic Darmin	668	468	100	100
Vincent Van Gogh	1931	1352	290	289
William Adolphe Bouguereau	259	181	39	39
William Merritt Chase	377	264	57	56
Zdislav Beksinski	708	496	106	106
Zinaida Serebriakova	415	291	62	62

Acknowledgements Not applicable.

Author contributions Conceptualization: TD and ZR. Data curation: TD. Formal analysis: TD. Funding acquisition: not applicable. Investigation: TD. Methodology: TD. Project administration: TD. Resources: ZR. Software: TD. Supervision: TD and ZR. Validation: TD and AB. Visualization: TD. Writing—TD. Writing—review and editing: TD, AB, and ZR.

Funding The University of North Carolina at Charlotte, College of Computing and Informatics Department of Education (DOE).

Availability of data and material Artworks available from WikiArt as public domain or copyright protected by fair use principle as used for informational and educational purposes.

Code availability https://github.com/btdobbs/cnn-catalogue-raisonne-

Declarations

Conflict of interest Not applicable.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

- Abramovich F, Pensky M (2019) Classification with many classes: challenges and pluses. ArXiv:1506.01567 [Math, Stat]. https:// arxiv.org/abs/1506.01567
- Bianco S, Cadene R, Celona L, Napoletano P (2018) Benchmark analysis of representative deep neural network architectures. IEEE Access 6:64270–64277. https://doi.org/10.1109/ACCESS.2018. 2877890
- Blessing A, Wen K (2010) Using machine learning for identification of art paintings. Tech report, 5
- Buda M, Maki A, Mazurowski MA (2018) A systematic study of the class imbalance problem in convolutional neural networks. Neural Netw 106:249–259. https://doi.org/10.1016/j.neunet.2018.07.011
- Cannon-Brookes P (1983) Impermanence: a curator's viewpoint. Int J Mus Manage Curatorship 2(3):283–285. https://doi.org/10.1080/ 09647778309514884
- Cetinic E, Lipic T, Grgic S (2018) Fine-tuning convolutional neural networks for fine art classification. Expert Syst Appl 114:107– 118. https://doi.org/10.1016/j.eswa.2018.07.026
- Chen J (2018) Comparison of machine learning techniques for artist identification. https://www.semanticscholar.org/paper/Compa rison-of-Machine-Learning-Techniques-for-Chen/8cf70f05a9 582fb3383c6c465bd4e7a608631dc2
- Deng J, Dong W, Socher R, Li L-J, Kai Li, Li Fei-Fei (2009) ImageNet: a large-scale hierarchical image database. 2009 IEEE Conference

on Computer Vision and Pattern Recognition, 248–255. https://doi.org/10.1109/CVPR.2009.5206848

- Friedenthal A (2020) John Smith's Rembrandt research project: an art dealer establishes the first catalogue raisonné of the paintings (1836). Netherlands Yearbook for History of Art/Nederlands Kunsthistorisch Jaarboek Online 69(1):212–247. https://doi.org/ 10.1163/22145966-06901008
- Grandini M, Bagli E, Visani G (2020) Metrics for multi-class classification: an overview. ArXiv:1506.01567 [Cs, Stat]. http://arxiv. org/abs/2008.05756
- He K, Zhang X, Ren S, Sun J (2016) Deep Residual Learning for Image Recognition. 770–778. https://openaccess.thecvf.com/content_ cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_ paper.html
- Johnson JM, Khoshgoftaar TM (2019) Survey on deep learning with class imbalance. J Big Data 6(1):27. https://doi.org/10.1186/ s40537-019-0192-5
- Johnson C, Hendriks E, Berezhnoy I, Brevdo E, Hughes S, Daubechies I, Li J, Postma E, Wang J (2008) Image processing for artist identification. IEEE Signal Process Mag 25(4):37–48. https://doi.org/ 10.1109/MSP.2008.923513
- Jou J, Agrawal S (2011) Artist identification for renaissance paintings. https://www.semanticscholar.org/paper/Artist-Identification-for-Renaissance-Paintings-Jou-Agrawal/399c6be15a90e94cc361 0121566147e29520c9ea
- Kim P (2017) MATLAB deep learning: with machine learning, neural networks and artificial intelligence. Apress, Berkeley, California
- Kondo K, Hasegawa T (2020) CNN-based Criteria for Classifying Artists by Illustration Style. Proceedings of the 2020 2nd International Conference on Image, Video and Signal Processing, 93–98. https://doi.org/10.1145/3388818.3389163
- Li C, Bovik AC (2009) Three-component weighted structural similarity index. In: Farnand SP, Gaykema F (eds) Electronic imaging. San Jose, CA. http://proceedings.spiedigitallibrary.org/proce eding.aspx?doi=10.1117/12.811821. Accessed 22 Nov 2020
- Mensink T, van Gemert J (2014) The Rijksmuseum challenge: museum-centered visual recognition. In: Proceedings of international conference on multimedia retrieval. Association for Computing Machinery, Glasgow, United Kingdom. https://doi.org/10. 1145/2578726.2578791
- Pirrone R, Cannella V, Gambino O, Pipitone A, Russo G (2009) Wiki-Art: An Ontology-Based Information Retrieval System for Arts. 2009 Ninth International Conference on Intelligent Systems Design and Applications, 913–918. https://doi.org/10.1109/ISDA. 2009.219
- Pishro-Nik H (2014) Introduction to probability, statistics, and random processes. Kappa Research LLC, Blue Bell, PA
- Rogers K (2015) Viewpoint: The Catalogue Raisonné Scholars Association (CRSA) - ProQuest. https://search.proquest.com/docview/ 1693347821?fromopenview=true&pq-origsite=gscholar
- Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L (2015) ImageNet Large Scale Visual Recognition Challenge. ArXiv:1409.0575 [Cs]. http://arxiv.org/abs/1409.0575
- Sokolova M, Lapalme G (2009) A systematic analysis of performance measures for classification tasks. Inf Process Manage 45(4):427– 437. https://doi.org/10.1016/j.ipm.2009.03.002
- Strezoski G, Worring M (2017) OmniArt: Multi-task Deep Learning for Artistic Data Analysis. ArXiv:1708.00684 [Cs]. http://arxiv. org/abs/1708.00684
- Strezoski G, Worring M (2018) OmniArt: a large-scale artistic benchmark. Assoc Comput Mach. https://doi.org/10.1145/3273022
- van Noord N, Hendriks E, Postma E (2015) Toward discovery of the artist's style: learning to recognize artists by their artworks. IEEE

Signal Process Mag 32(4):46–54. https://doi.org/10.1109/MSP. 2015.2406955

Viswanathan N, Stanford (2017) Artist Identification with Convolutional Neural Networks. https://www.semanticscholar.org/paper/ Artist-Identification-with-Convolutional-Neural-Viswanathan-Stanford/dafe87bf57c4413d769de46af78f7e4305087838 **Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.