Pre-trained Models on Scanned Historic Watermarks

A Comparative Analysis Exploring Pre-Trained Models on Scanned Historic Watermarks

Marin Alexandru-Remus[1]

Supervisor(s): Dr. Martin Skrodzki[1]**, Dr. Jorge Martinez Castaneda**[1]

[1]**EEMCS, Delft University of Technology, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 22, 2024

Name of the student: Marin Alexandru-Remus
Final project course: CSE3000 Research Project
Thesis committee: Dr. Martin Skrodzki, Dr. Jorge Martinez Castaneda, Dr. Christoph Lofi

Abstract

This paper tackles the problem of evaluating the task of finding similar scanned historical watermarks - small images embedded in historical paper that have been digitized to be processed on a computer - using pre-trained neural networks. This research aims to identify an efficient and accurate alternative to the traditional, time-consuming manual detection methods for finding similar watermarks. The primary issue addressed is the inefficiency of these manual methods. The evaluation focuses on finding similar watermarks for a specific query watermark, assessing the efficacy of neural networks in comparison to a prior art system that employs traditional image processing techniques. This comparison aims to determine how well these neural networks perform in the task of watermark similarity detection. The study involves a dataset of 500 labeled images tested in two distinct contexts: one using unprocessed images and another using images processed to keep only the watermark outline. The results show that pre-trained models achieve higher accuracy and time efficiency compared to the prior art system that uses image processing. These models demonstrate significant effectiveness in watermark recognition and comparison, with each network achieving over 80% accuracy for traced watermarks. EfficientNetB0 achieved 94.66%, VGG16 89.33%, ResNet50 86.67%, and InceptionV3 84%, while the prior art system gets 64,8%. These results conclude that these models are valuable tools in the field of watermark recognition and comparison.

1 Introduction

The watermarks studied in this work are small images embedded in historical paper, that can only be seen when shining a light from a specific angle. These watermarks were embedded into the paper during the paper-making process by impressing a design onto the paper while it was still wet and malleable, using a wire mold or dandy roll. This information is important to historians and researchers because watermarks can provide details about the origin, place, or time of a document [1]. In the past, watermarks served as an identifier for the paper mill that made the sheet [2]. Nowadays, scientists are trying to identify unique watermarks to know the evolution of commercial and cultural exchanges between countries [3]. To discern a watermark a specialist must be contacted and will manually search in the archives to find that specific watermark, but this process might take a considerable amount of time, from hours to days.

Traditional methods of watermark detection are very time-consuming and effortful. Although accurate, manual search is not scalable when managing large data archives. Consequently, integrating automated techniques is not only a way to streamline the process, but also a way to unlock historical insights on a much larger scale. An example of an automated technique is represented by pre-trained models, which are machine learning models that have already undergone training on large datasets.

The pre-trained models demonstrate efficacy in image comparison across various datasets. Even though some work has been done on applying these models to watermark detection, a comprehensive performance comparison in an accessible pipeline has yet to be fully explored. Specifically, the research question

"*How effective are the pre-trained models, VGG16, ResNet50, EfficientNet, and InceptionV3 in improving watermark recognition results?*"

will be answered.

The task involves analyzing a system for watermark recognition and comparison with pre-trained deep learning models, such as VGG16 [4], ResNet50 [5], EfficientNet [6], and InceptionV3 [7]. These models aim to assess how effective they are in improving the results of watermark comparison, as they demonstrated remarkable performance across various datasets. These models — VGG16, ResNet50, EfficientNet, and InceptionV3 — were chosen due to their proven performance in feature extraction and image recognition tasks. Their robust architectures, combined with the availability of pre-trained weights for transfer learning, make them great candidates for improving watermark detection accuracy. These models represent a balance of performance and efficiency, making them particularly well-suited for this task. While other models might offer unique advantages, the exceptional track records and widespread of these four provide a reliable foundation for improving the effectiveness of watermark recognition systems.

To conduct a comprehensive analysis on watermarks, it is necessary to examine the following: effectiveness on traced watermarks (tracings of watermarks), effectiveness on untraced watermarks (raw watermarks), effectiveness in comparison to a system made by Banta et al. [8], and speed of the process. This process is done in two distinct contexts, one which uses the raw watermarks (in their unchanged form) and another context in which a set of pre-processing techniques are applied to the watermarks, to enhance their visibility.

This research paper aims to explore the potential of deep learning pre-trained models in the field of watermarks, in comparison to this prior art system that uses traditional image processing techniques [8]. This study tries to bridge the gap between traditional manual methods and modern automated techniques. Integrating deep learning into watermark recognition aims to improve both speed and accuracy, making it feasible to process vast historical archives. So, this study investigates the potential of VGG16, ResNet50, EfficientNet, and InceptionV3 to overcome unique challenges posed by watermark imagery, such as varying paper textures, degradation over time, and subtle differences in watermark designs.

The ideal outcome would provide a clear indication of how pre-trained models can contribute to the effectiveness of watermark recognition and comparison. Demonstrating significant improvements in accuracy, efficiency, or adaptability, might establish the utility of the models in this domain, so it might validate pre-trained models as valuable tools for water-

mark recognition.

2 Background

Deep learning, a subset of machine learning that significantly advanced the field of artificial intelligence is used in this research. This method "teaches" computers to process data in a way that is inspired by the human brain. Deep learning allows for the computation of models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction [9] and it can recognize complex patterns in, for example, pictures, texts, sounds and more, to produce accurate insights and predictions. However, there is a significant amount of time required for training these models, especially when dealing with large datasets and complex architectures. This can lead to substantial computational costs and longer development cycles.

Each deep learning model trains a network composed of numerous layers, wherein each layer contributes to the model's capacity to learn and generalize. A layer in a network represents a structure that takes information from the previous layers and passes it to the next layer in a different form.

The concepts of depth, width, and resolution are fundamental in the architecture of neural networks. Depth denotes the number of layers in a network. The width refers to the number of neurons or filters in each layer, where broader networks can encapsulate a greater number of features at every level. Resolution signifies the spatial dimensions of feature maps (representations of the input data that highlight specific features detected by the filters) generated by a layer (convolutional layer).

Zeiler et al. [10] observed that neural networks can localize the objects within a scene and that those objects are crucial for image comparison. If the object is occluded, there is a strong decrease in the network's ability to classify the scene accurately. Even though the importance of detecting the object is very high, Zeiler et al. observed that the parts of the object are analyzed only in the higher layers of the network, while in the lower ones are examined: edges, textures, shapes, and patterns. These conclusions were also drawn by Zhou et al.[11; 12]. Although objects are a key part of the classification, the other representations used in combination with them (textures, materials) impact this process.

When working with large datasets, deep learning networks require a period of several days to weeks to train, and to avoid such a situation we use pre-trained models [13]. Training involves feeding the network with data to learn the patterns and adjust parameters for accurate predictions. It takes a considerable amount of time due to the complexity of the models, large dataset size, high computational demands, and the need to optimize millions of parameters. Pre-trained models are machine learning algorithms that reuse the knowledge achieved from one or more tasks on a new task [14]. This process is known as transfer learning, the reuse of knowledge to solve a new task in a faster or more efficient way [15]. In this case, zero-shot transfer learning is employed, as the weights used for image classification were not explicitly trained on data from this specific category of watermarks. This aligns with the definition of zero-shot learning [16], wherein the model recognizes and makes predictions for tasks that have never been encountered during the training phase.

The ImageNet competition, initiated in 2009, has been the foundation in the field of computer vision and deep learning, playing a pivotal role in the emergence of various algorithms [17]. ImageNet provides an extensive image database structured based on the WordNet hierarchy, a system that organizes English words into sets of synonyms. Each node of the hierarchy is associated with hundreds and thousands of images. The weights from ImageNet are used because this dataset contains a wide variety of objects and scenes and has proven efficiency for a vast range of tasks. This database will enhance the transfer learning process for VGG16, ResNet50, EfficientNet, and InceptionV3 models by pre-training them with these weights, thereby optimizing their performance.

The pre-trained models demonstrated efficacy in image comparison across various datasets: natural disasters [18; 19], plants [20; 21], medicine [22; 23; 24], sports [25] etc. Although some work has been done on applying these models to watermark detection (see Section 3), a comprehensive performance comparison in a broadly accessible pipeline has yet to be fully explored.

As previously mentioned, the deep learning models that are used for watermark comparison in this paper are VGG16 [4], ResNet50 [5], EfficientNet [6], and InceptionV3[7]. Those are convolutional neural networks, considered to be some of the best for image recognition and comparison. In this context, "recognition" refers to the capacity of the system to identify the watermark and to translate it into numerical data, while the "comparison" process means detecting similar watermarks based on this numerical data (Fig. 1).

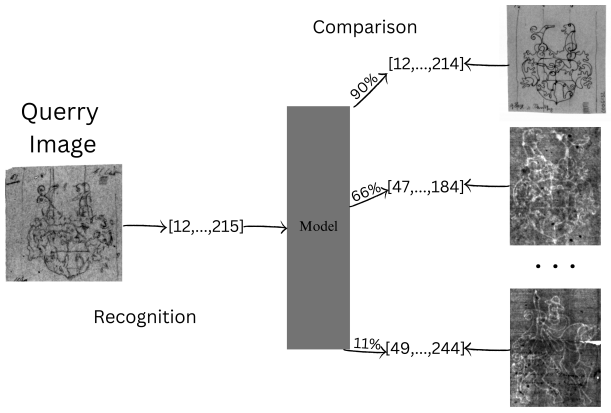

Figure 1: Visual explanation of Recognition and Comparison processes. Recognition refers to the capacity of the system to translate the image into numerical data, while comparison refers to the process of detecting similar watermarks based on this numerical data. Similar images usually have a higher similarity percentage.

VGG16 is unique because it has 16 layers with weights in its architecture and does not rely on many hyperparameters [4]. It is widely appreciated in the field of deep learning for its simplicity and effectiveness.

ResNet50 is considered innovative because it uses residual

connections, that allow the network to learn a set of functions that map the input to the output. The "50" in its name represents the total number of layers, demonstrating its depth and capacity to learn complex patterns [5]. This depth allows the model to progressively extract and combine features from simpler to more abstract levels, facilitating the recognition of complex patterns in the input data.

EfficientNet earned acclaim for its performance and efficiency in the field of images. It is considered a powerful convolutional neural network architecture because it scales the width, depth, and resolution of the layers in a systematic and compound manner rather than randomly [6]. EfficientNet has several variants, which means the number of layers can vary.

InceptionV3 stands out in the field of convolutional neural networks for its remarkable ability to execute deep processing with a low computational cost. This is achieved by balancing the width and the depth of the layers, making it an efficient and accurate solution for image comparison [7]. In its architecture, it has 48 weighted layers.

Those pre-trained deep learning models will be tested on both traced and untraced watermarks. "Untraced" elements refer to scans that were taken directly from the watermark paper (Fig. 2). "Traced" ones are scans of tracings of watermarks (Fig. 3). From Figure 3, it is evident that traced watermarks generally exhibit more clarity in comparison to direct scans. Those tracings have been used to create several catalogs such as the Briquet dataset [26], the Piccard dataset [27], or the Bernstein dataset [28], which can also be used for testing the pre-trained models.

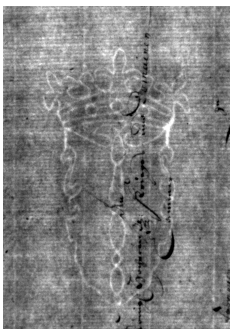

Figure 2: Untraced watermark.

Figure 3: Traced watermark.

3 Related Work

Over the past decade, after the widespread of artificial intelligence for image recognition, researchers gained a new tool for automating the process of comparing and identifying watermarks. There has been some research that uses deep learning algorithms, but not only pre-trained models.

Shen et. al [29] proposed a watermark recognition solution that matches using a neural network, which needs to be retrained for every image added to the dataset. Continuous database updates would make such implementation impractical, highlighting the inefficiency of frequent model training and the utility of transfer learning. To this date, a significant portion of watermark collections has not been digitized, and the digitization process remains slow. Therefore, any viable solution designed to work with digitized watermarks must be capable of integrating newly acquired data.

On this approach, Bounou et al. [30] built a web application for watermark recognition. To speed up the process, they adjusted how they measure similarity, but this change also made the algorithm less precise. As of June 2024, the application is no longer available.

Pondenkandath et al. [31] introduced a matching algorithm that consists of a classification that uses a ResNet convolutional neural network initialized with the weights from the ImageNet dataset and a similarity matching algorithm. They performed this algorithm on a database with 106,000 images and 96,000 tracings from the dataset compiled by Gerhard Piccard [27]. This dataset contains almost entirely quality tracings, so it does not compare with this case where also untraced watermarks are under evaluation.

A project that uses pre-trained models in the same way as this research aims to is the one from TU Munich by Beriozchin et al. [32]. This project uses ResNet18 to extract the feature vectors and uses the Spotify Annoy [33] algorithm to find similar watermarks, in a dataset of 6600 images. It has an accuracy of more than 50% for the 25 nearest neighbors and 68% for the 50 nearest neighbors. In this context, accuracy refers to the proportion of correctly identified similar watermarks among the nearest neighbors.

In this work, the system compared with the pre-trained deep learning models is a watermark recognition and comparison system [8] based on the database from the German Museum of Books and Writings [1], that got decent outcomes, 42% for a random dataset (with many unclear watermarks) and 82% for a dataset with watermarks visible to the human eye. These accuracy values indicate the proportion of correctly recognized watermarks within the datasets.

Banta et al. [8] built a system to automatically identify similar watermarks using traditional image processing techniques. The pipeline takes as input an image and tries to extract useful information about the watermark and then calculates its similarity with the comparison watermarks. This pipeline contains a harmonization step, that is used in one of the contexts of this research in which the watermarks are analyzed using pre-trained deep-learning models. The harmonization aims to take the input image and isolate the watermark from it, to keep only the outline of the watermark. It involves the following steps: pre-processing (enhance image contrast, remove shadows, remove lines), denoising (remove noise), thresholding (segmentation based on intensity levels), and post-processing (remove any elements that do not belong to the watermark.) [8]. The effects can be seen step by step in the following figure (Fig. 4).

As features, this system uses the following three: SIFT [34], Hu Moments [35] and Zernike Moments [36] and calculates a geometric mean to get a similarity score. In comparison, this research uses the above-mentioned models as feature extractors.

[1] German Museum of Books and Writings, "Gutenberg Bible Exhibit," accessed May 10, 2024, http://www.dnb.de/EN/museum.

(a) (b) (c) (d)

Figure 4: Harmonization steps on a traced watermark (a): traced watermark after pre-processing (b), traced watermark during post-processing (c) and traced watermark after post-processing (d). These images are an example of how the harmonization should work. In this case, harmonization is used to remove the background and keep only the outline of the watermark

4 Methodology

The applicability of the pre-trained models in this domain is evaluated, to intensify the comprehension through empirical analysis. A set of 500 images, randomly chosen from the dataset provided by the German Museum of Books and Writings [37] is used. For diversity, the subset contains equally divided traced and untraced watermarks. To facilitate the evaluation process, the dataset is manually annotated by categorizing those images into different classes. Each image is labeled according to that specific class. Each class has four images, including two traced and two untraced watermarks. An illustrative example of such a class is depicted in Figure 5.

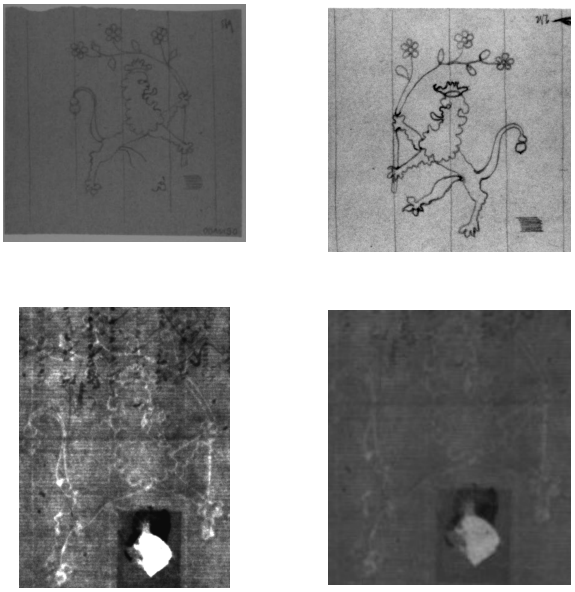

Figure 5: Class 23 from the dataset with 2 traced watermarks(first row) and 2 untraced watermarks(second row). The watermarks have been provided by the German Museum of Books and Writing.[2]

The images are tested within 2 distinct contexts. In both of them, the pre-trained models - VGG16, ResNet50, EfficientNet, InceptionV3 - classify the images using the weights from ImageNet.

Each model, pre-trained on the ImageNet dataset, serves as a feature extractor, enabling the comparison of the images based on the visual content. Each image in the dataset will undergo pre-processing to meet the model's requirements. The models VGG16, ResNet50, EfficientNet expect a resized image to a specific dimension of 224 x 224 pixels, while InceptionV3 needs a resized image to a specific dimension of 299 x 299 pixels. Additionally, the images require to be in the RGB color space, meaning they consist of three color channels: red, green and blue.

Feature vectors, representing the visual characteristics of the images, are extracted from the processed image. These feature vectors are saved for subsequent comparison between individual images. A comparison of the feature vectors extracted from the models is conducted using cosine similarity to determine their similarity percentage. By analyzing the similarity scores and ordering images in decreasing order based on these scores, it identifies the images that share similar visual content.

4.1 Using raw images

First, the images are used in their "raw" form, without any pre-processing technique applied. Those images will be analyzed with each model to evaluate 3 cases: only traced watermarks, only untraced watermarks and both at the same time. This process is repeated for each pre-trained deep learning model to evaluate the 3 cases mentioned.

4.2 Using harmonization

As mentioned in Chapter 3, the second case has a pre-processing stage, the one used in the paper from Banta et al. [8]. The comparison of watermarks is here based only on the feature vectors of the watermark outlines, this is why the image undergoes a pre-processing step.

The process of applying harmonization produces a new set of 500 binary images with a black background and a white watermark, an example is given in Figure 6.

At this stage, the aim is to observe how pre-trained models behave when elements like background, texture, material, etc. are occluded, to make sure that the watermark is retained.

Figure 6: An untraced watermark after the Harmonization process. The outline of the watermark has some missing parts, but this should not significantly affect the comparison.

4.3 Evaluation

To evaluate the models, each image in the dataset is analyzed to find the most similar watermarks. Feature vectors representing the visual characteristics of each processed image are extracted. These vectors are then compared using cosine similarity to measure similarity percentages, which helps to identify images with similar visual content. Additionally, the rank of similar images for each watermark is recorded, where "similar images" refers to those in the same class as the query image.

The effectiveness of the models refers to the capacity of each model to classify the labeled images correctly. To understand how effective are those models, the following evaluation metrics will be analyzed: model accuracy, average lowest rank, and average similarity.

To calculate the accuracy of each model, the number of images with a match is calculated. This number is divided by the total number of images. A match is found if at least one similar watermark is found in the top 10% of returned similar watermarks arranged in descending order of similarity. This evaluation was used such that those pre-trained models can be compared to the model by [8]. This type of model accuracy was chosen because historians typically need to investigate up to 30-50 images closest watermarks for the searched watermark [32]. The formula for model accuracy can be observed in the subsequent figure.

$$M = \frac{\sum_{i=1}^{n} x_i}{n} \tag{1}$$

where n is the total number of images and

$$x_i = \begin{cases} 1, & \text{if there is a similar watermark in the top 10\%,} \\ 0, & \text{otherwise.} \end{cases}$$

The images are analyzed with each model to evaluate 3 cases: only traced watermarks, only untraced watermarks and both at the same time, for both scenarios presented. Also, the algorithms are compared on computation complexity. Historians should not wait longer than it would take to solve the search manually when receiving the list of similar watermarks from the models.

In addition, each image is analyzed to determine its average similarity within its class. For every image annotated with a specific class, the average similarity value of other images annotated with the same class is calculated. Subsequently, an overall mean is calculated for all images. This way of evaluation is inspired by the mAP metric [38] (mean average precision), which was also used in the system developed by Pondenkandath et al. [31]. The formula for average similarity can be observed right after this paragraph.

$$S = \frac{\sum_{i=1}^{n} a_i}{n} \tag{2}$$

where n is the total number of images and

$$a_i = \frac{\sum_{j=1}^{4} s_{ij}}{4}$$

where s_{ij} is the similarity percentage between the query image (the cosine similarity between them) and image *j* from class *i*.

A second evaluation metric is applied by Pondenkandath et al. [31]. It calculates the average of the lowest ranks of each image. These images are ordered in descending order based on their similarity score. It is important to note that an image with a higher similarity score, such as 90%, is closer to achieving a low rank, while an image with a lower similarity score, such as 10%, is closer to obtaining a high rank. Usually, historians aim to discover a similar watermark quickly, so the objective is to have the image rank as low as possible. The formula for the average lowest rank can be observed in the next figure.

$$R = \frac{\sum_{i=1}^{n} l_i}{n} \tag{3}$$

where n is the total number of images and

$$l_i = \min_{j=1}^{4} r_{ij}$$

where r_{ij} is the rank of the image *j* from class *i* in the entire hierarchy.

The models should return the most similar images among the first. The study also analyzes the distribution of the returned similar images for each model in both contexts.

5 Results

5.1 Model accuracy

In Table 1, the first column presents the results of the model accuracy on traced watermarks, in comparison to the prior art system created by Banta el al. [8]. Certain algorithms have different versions, for instance, EfficientNet has seven variations, ranging from B0 to B7. However, for this evaluation, only the version that proved to be the most accurate is selected, EfficientNetB0.

Model	Traced Watermarks	Untraced Watermarks	Overall Dataset
VGG16	81,33	80,12	67,80
ResNet50	84,00	76,51	70,40
EfficientNetB0	80,00	71,99	64,60
InceptionV3	79,67	**81,02**	**71,20**
Harmonization + VGG16	89,33	53,61	64,13
Harmonization + ResNet50	86,67	54,82	62,90
Harmonization + EfficientNetB0	**94,66**	53,61	63,64
Harmonization + InceptionV3	84,00	46,08	58,72
Prior Art System	64,80	56,80	57,60

Table 1: Model Accuracy for Traced and Untraced Watermarks in Scenario 1 (first 4 rows) and Scenario 2 (next 4 rows) compared to the Prior Art System (last row). Numbers are in percentages. The best values for each case are highlighted. Values over 80% show that the models are very efficient, since it is almost impossible to achieve 100%.

It is clear from the results that the pre-trained deep learning models yield superior performance compared to the prior art system. Among those models, in the first scenario, the optimal one is ResNet50, which boasts an accuracy of 84%, while the others are not trailing far behind.

Considering traced watermarks in the first scenario, InceptionV3 emerges as the least favorable solution from the category of the pre-trained models. However, the discussion shifts when the focus is on untraced watermarks. Similar to the results for traced watermarks, the pre-trained models surpass the accuracy of the prior art system. Furthermore, the results for the pre-trained models are not very different, except for EfficientNet, which seems to exhibit slightly lower accuracy.

The overall results, for a combined set of equally divided, traced, and untraced images, can be seen in Table 1, last column. It might be surprising that the overall results are lower than the ones for traced and untraced watermarks. This discrepancy is attributed to the fact that pre-trained models exhibit lower accuracy when comparing a traced watermark with an untraced one. The low accuracy in this context can be explained by the nature of these models, which also consider the background, affecting performance [39]. This challenge is addressed by applying harmonization such that background becomes irrelevant.

By applying harmonization to the dataset, for traced images, the accuracy for each model is increased and this increase can be seen in the lower part of Table 1. The highest gain in accuracy, compared to the first scenario, can be seen at EfficientNet, which is the most reliable in this case. This difference is explained by the fact that the background becomes invariable and only the outline of the watermark is compared. The method with the lowest gain compared to the first scenario is InceptionV3.

For untraced watermarks (second column, the lower part in Table 1), the results are lower than in the first scenario but comparable with the ones from the prior art system, which proves to be slightly more accurate. A reason for this decrease in accuracy might be the low quality of the untraced watermarks and the impossibility, in some cases, for the harmonization process to identify the watermark. Except for InceptionV3, the other pre-trained deep learning models provide similar results, with ResNet50 featuring a slight edge in accuracy.

In comparison to the previous scenario, where the overall results are lower than both traced and untraced cases, in this context the results suggest a balance between them (last column in Table 1). All models exhibit comparable accuracy, with none discernibly standing out.

5.2 Average Similarity

The average performance of each model on the images that belong to the same class as the query image is also compared through the average similarity evaluation. This approach assesses how well the pre-trained models can recognize and match images that are essentially the same, providing a measure of consistency and accuracy.

In Table 2, the results for this evaluation are given. On traced watermarks, it can be seen that the most accurate pre-trained model is ResNet50, which intersects with the conclusion drawn in the previous subsection, for the same characteristics, which also selects this model as the most precise.

In the case of untraced watermarks, the results are quite different. With respect to model accuracy, EfficientNet is the least accurate algorithm, but in terms of average model probability, this model is the most accurate one.

The results are not the same for untraced watermarks, a decrease in overall model probability can be seen in this case, which could be because pre-trained models are imprecise in matching a traced watermark with an untraced one.

The model, that performs well overall is EfficientNet, getting over 70% accuracy, being first for 2 categories and second for another one.

The results for the second scenario can be seen in the lower part of Table 2. An increase in accuracy for all models can be seen in the case of traced watermarks. This increase means that the removal of the background enhances the ability of the pre-trained models to compare traced watermarks. ResNet50 is the algorithm that provides the highest accuracy and it is also the most precise in the first context.

In the case of untraced watermarks, the results decrease for all models, except ResNet50 where there is a negligible increase. The same trend is also observed for model accuracy. It indicates that deep learning models are ineffective when they apply harmonization to untraced watermarks. Similar to the results for traced watermarks, ResNet50 is the most

Model	Traced Watermarks	Untraced Watermarks	Overall Dataset
VGG16	71,56	69,66	65,43
ResNet50	75,82	71,27	67,91
EfficientNetB0	73,50	**74,68**	70,05
InceptionV3	73,31	71,52	68,01
Harmonization + VGG16	75,94	66,76	63,92
Harmonization + ResNet50	**84,82**	71,49	**70,76**
Harmonization + EfficientNetB0	74,36	60,76	59,73
Harmonization + InceptionV3	78,97	64,27	63,27

Table 2: Model Similarity for Traced and Untraced Watermarks in Scenario 1 (first 4 rows) and Scenario 2 (next 4 rows). Numbers are in percentages. The best values for each case are highlighted. The decrease in accuracy by applying harmonization expresses the incapacity of the models to detect the watermark in the raw untraced images, due to the low quality of the watermarks.

accurate algorithm.

It is interesting to note that in the first scenario, EfficientNet proves to be the most precise. Moving to untraced watermarks in the second scenario, there is a big decrease in accuracy. This observation suggests that the other elements, except the watermark outline, influences more the results for this model.

The overall results are lower compared to both traced and untraced cases from the same scenario and with the results from the previous context. ResNet50 proves to be the best option, with an overall increase in accuracy from 67,91% to 70,76%. For untraced watermarks, EfficientNet delivers poor results, with an accuracy of 59,73%.

ResNet50 is the model that performs well in all cases from the second context. It gets over 70% accuracy, being first for all categories. Also, ResNet50 did not exhibit as poorly in the first scenario, compared to how EfficientNet did in the second one.

No single model visibly outperforms the others, all of them demonstrate decent comparable results across all cases and scenarios. This indicates that each model has its strengths and weaknesses, but overall they prove to be effective in watermark comparison.

5.3 Average Lowest Rank

The evaluation of the average lowest rank of the most relevant similar watermark is essential in this analysis. Historians examine 30-50 watermarks to uncover the specific information they seek. An average rank lower than 30-50 may help researchers to streamline their effort, reducing the time of the process.

Table 3 shows the results for the average lowest rank evaluation for traced and untraced watermarks. The overall results are neglected because those represent the average of the results for traced and untraced watermarks.

The values, in the first context, are visibly lower than 30-50. It means that the pre-trained models can serve as a tool that accelerates the process of watermark comparison. The best convolutional neural network for both traced and untraced watermarks when the images do not undergo a preprocessing step is VGG16. Even though for traced watermarks, the average highest rank is approximately 6 for all of them, for untraced watermarks, VGG16 is the only one that provides a result under 20.

By looking at the distribution of ranks (see Appendix 1, Scenario 1 - Traced Watermarks), it can be observed that VGG16 is also the model in which the number of traced watermarks whose rank is greater than 50 is the lowest one, only 4/250, in comparison to 7/250 for the others. The number of traced watermarks ranked with 1 is the highest for VGG16, 170/250. The other models deliver also good results which prove the utility of the pre-trained models for watermark recognition and comparison. Despite VGG16 being the most effective in the first context, for untraced watermarks (see Appendix 1, Scenario 1 - Untraced Watermarks) the model that ranks the most images in the first place is InceptionV3 with 144/250 (VGG16 - 134/250).

In the second context, the harmonization stage contributes

Model	Traced Watermarks	Untraced Watermarks
VGG16	5,98	**17,95**
ResNet50	6,5	21,64
EfficientNetB0	6,08	27,54
InceptionV3	6,33	20,43
Harmonization + VGG16	3,57	44,92
Harmonization + ResNet50	5,4	43,46
Harmonization + EfficientNetB0	**2,89**	42,25
Harmonization + InceptionV3	6,12	54,43
Prior Art System	23,43	47,77

Table 3: Average Lowest Rank for Traced and Untraced Watermarks in Scenario 1 (first 4 rows) and Scenario 2 (next 4 rows) compared to the Prior Art System. The best values for each case are highlighted. A low rank means that a similar image is between the firsts in the ranking based on the cosine similarity.

positively to getting a lower rank for traced watermarks. In this case, EfficientNet gets a 2,89 average rank. The greatest number of traced watermarks ranked in first place, 185/250 (see Appendix 1, Scenario 2 - Traced Watermarks) is produced by EfficientNet. VGG16 is the best in the first scenario, while in this context it is only second, ranking only one watermark over 50.

In the case of untraced watermarks, harmonization decreases the accuracy because the average rank goes from approx. 21 to approx. 44. These results suggest that for untraced watermarks either the harmonization step is not very effective or the quality of some images is very poor and the outline of the watermark is not easily distinguishable. No model outperforms the others, but EfficientNet ranks the most images in first place with 66 out of 250 (see Appendix 1, Scenario 2 - Untraced Watermarks). On the other side, InceptionV3 is the most inaccurate with 54,43% accuracy and it ranks the most watermarks after 50, 98/250.

In the case of the system built by Banta et al. [8], the average rank for traced watermarks represents a decent outcome because it is less than what was aimed to surpass(30-50). It gets 23,43, which is worse than what the pre-trained models achieved. For untraced watermarks, the result is comparable to what pre-trained models achieved with harmonization. It appears to be less efficient than when harmonization is not used.

5.4 Computation complexity

In terms of computation complexity, the first scenario is more effective than the second one since it does not include any processing step. The system built by Banta et al. [8] contains a pre-processing step and the classification for one watermark takes approximately 4.5 seconds.

The pre-trained deep learning models outperform this system, delivering the results in approximately 0.2 seconds for one watermark. EfficientNet is the fastest network and delivers the result in 0.05 seconds, for one watermark comparison. This result was anticipated since the structure of EfficientNet differs from the one of VGG16, ResNet50, or InceptionV3, because it has fewer parameters. In comparison to those, the scaling of the width, depth, or resolution of the layers of EfficientNet is not random, these are scaled using a set of fixed coefficients achieving lower computational cost.

The efficacy of pre-trained models, delivering rapid results compared to manual search, underscores their potential to streamline the watermark analysis and the archival research. This suggests their wider applicability in various domains requiring pattern recognition. Overall, pre-trained deep learning models present an efficient approach to watermark recognition and comparison, with ongoing research expected to refine their capabilities further.

6 Responsible Research

This section reflects on the ethical aspects of the research and discusses how it can be reproduced. Two main ethical concerns must be considered, the first regards the dataset of images this research uses and the second refers to bias and fairness.

The data used in this project consists of a subset of 500 images, that represent documents from hundreds of years ago. The entire set of images has been provided by the German Museum of Books and Writing [37]. Due to their age, the documents can not be copyrighted. Even if the data is legally obtained, it can still present ethical challenges, since it might include personal documents, and private correspondences, which raise privacy concerns. Also, it is not known if each owner gave their consent to share or use those watermarks. It is presumed that the museum has ethically obtained those documents. This assumption is substantiated by the museum's reputation as a reputable and reliable institution.

To ensure reproducibility, a viable approach would be to acquire a new dataset of watermarks. Preferably, every document in the dataset would have explicit consent to be used. Given the wide array of watermark documents needed and their historical nature, locating either owners or descendants would be impractical. Consequently, there will be a trade-off between dataset size and reliability. To reproduce the results exclusively for traced watermarks, images from the Piccard and Briquet datasets can be used.

To ensure fairness, the weights used for pre-training the models should originate from a dataset without any watermarks. As mentioned in the Methodology section, this research involves a process of zero-shot transfer learning, which assumes that the model analyzes the data without it being seen before. For reproducibility, the weights that should be used are from the ImageNet dataset. Other weights used from a dataset that might contain watermarks, might change the results of this investigation.

The analysis has been done in Python language and the architecture of each model can be found in the Keras library [40], in the application module. The pre-processing step for images to be resized and in the necessary format can be found in the same module. For calculating the similarity between the images, it has been used the cosine distance, which can be found in the Scipy library. The harmonization step used to outline only the watermark from the image is explained in the paper of Banta et al. [8].

7 Conclusions

In this work, the recognition and comparison of watermarks using pre-trained deep learning models, namely VGG16, ResNet50, EfficientNet, and InceptionV3 were evaluated, in contrast to a system that utilizes traditional image processing methods for the same purpose. These models were evaluated using a dataset of 500 images in two distinct contexts: one where no watermark pre-processing is involved, and another where such pre-processing is applied. The evaluation encompassed the determination of the model accuracy, the average similarity, the average lowest rank, and the computation complexity for each model. Based on these metrics, pre-trained deep learning models prove to be slightly more effective than the prior art system.

Incorporating harmonization into the identified watermarks increased the accuracy across all models, achieving a 94.66% model accuracy for EfficientNetB0 with an average rank of 2.89. Conversely, the application of harmonization to un-

traced watermarks resulted in lower values, highlighting the low quality of certain untraced watermarks and the limitations of harmonization in keeping only the watermark in specific scenarios.

However, since the results are not perfect, future work may imply the training of those models using a watermark dataset, not only the weights from ImageNet. Also, a larger dataset can be used for a more comprehensive evaluation.

Based on the findings, pre-trained models serve as valuable tools for watermark recognition. Models such as VGG16, ResNet50, InceptionV3, and EfficientNet show promising advancements in both accuracy and efficiency. These models demonstrate proficiency in comparing traced watermarks and achieve the best results when harmonization is included. The broad applicability of these models suggests their potential utility in various contexts, ranging from historical document authentication to modern digital watermarking.

References

[1] L. Müller, "Understanding paper: Structures, watermarks, and a conservator's passion," *Harvard Art Museum*, May 2021, accessed: Apr. 23, 2024.

[2] G. Brunner and H. Burkhardt, "Classification and retrieval of ancient watermarks," in *Data Analysis, Machine Learning and Applications*, C. Preisach, H. Burkhardt, L. Schmidt-Thieme, and R. Decker, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 237–244.

[3] F. Yahya and R. Jones, "Malay manuscripts: a guide to paper and watermarks. the collected works of russell jones 1972–2015," *Indonesia and the Malay World*, vol. 49, no. 144, pp. 139–160, 2021. [Online]. Available: https://doi.org/10.1080/13639811.2021.1939521

[4] S. Tammina, "Transfer learning using vgg-16 with deep convolutional neural network for classifying images," vol. 9, 10 2019, p. p9420.

[5] S. Agrawal, V. Rewaskar, R. Agrawal, S. Chaudhari, Y. Patil, and N. Agrawal, "International journal of intelligent systems and applications in engineering advancements in nsfw content detection: A comprehensive review of resnet-50 based approaches," vol. 11, pp. 41–45, 10 2023.

[6] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," *CoRR*, vol. abs/1905.11946, 2019. [Online]. Available: http://arxiv.org/abs/1905.11946

[7] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," *CoRR*, vol. abs/1512.00567, 2015. [Online]. Available: http://arxiv.org/abs/1512.00567

[8] D. Bantă, S. Kho, A. Lantink, A.-R. Marin, and V. Petkov, "A watermark recognition system: An approach to matching similar watermarks," 2023, last accessed 28 May 2024. [Online]. Available: http://resolver.tudelft.nl/uuid:e8dfbd63-ae54-4159-b786-d1d8c64dc827

[9] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[10] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," *CoRR*, vol. abs/1311.2901, 2013. [Online]. Available: http://arxiv.org/abs/1311.2901

[11] B. Zhou, A. Khosla, À. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," *CoRR*, vol. abs/1512.04150, 2015. [Online]. Available: http://arxiv.org/abs/1512.04150

[12] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Object detectors emerge in deep scene cnns," in *International Conference on Learning Representations (ICLR)*, 2015. [Online]. Available: https://arxiv.org/abs/1412.6856

[13] S. Mascarenhas and M. Agarwal, "A comparison between vgg16, vgg19 and resnet50 architecture frameworks for image classification," in *2021 International Conference on Disruptive Technologies for Multi-Disciplinary Research and Applications (CENTCON)*, vol. 1, 2021, pp. 96–99.

[14] H. Wang, J. Li, H. Wu, E. Hovy, and Y. Sun, "Pre-trained language models and their applications," *Engineering*, vol. 25, pp. 51–65, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2095809922006324

[15] M. Hussain, J. J. Bird, and D. R. Faria, "A study on cnn transfer learning for image classification," in *UK Workshop on Computational Intelligence*, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:52066998

[16] H. Larochelle, D. Erhan, and Y. Bengio, "Zero-data learning of new tasks," p. 646–651, 2008.

[17] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," pp. 248–255, 2009. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2009.5206848

[18] P. Jain, B. Schoen-Phelan, and R. J. Ross, "Tri-band assessment of multi-spectral satellite data for flood detection," in *MACLEAN@PKDD/ECML*, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:228085706

[19] H. Jabnouni, I. Arfaoui, M. A. Cherni, M. Bouchouicha, and M. Sayadi, "Resnet-50 based fire and smoke images classification," in *2022 6th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, 2022, pp. 1–6.

[20] G. Penugonda, R. Singamaneni, and A. Kalyani, "A comparative study for monocot remembrance using vgg16, efficientnet, inceptionv3, and resnet50 on accuracy and response time," 12 2023, pp. 218–224.

[21] J. Yao, S. N. Tran, S. Garg, and S. Sawyer, "Deep learning for plant identification and disease classification from leaf images: Multi-prediction approaches," 2023.

[22] T. Kujani, S. Alex David, T. Sathya, P. Arivubrakan, and S. Shanmuga Priya, "Efficient brain tumor detection using vgg-16 and resnet50 transfer learning models," in *Soft Computing for Security Applications*, G. Ranganathan, Y. EL Allioui, and S. Piramuthu, Eds. Singapore: Springer Nature Singapore, 2023, pp. 455–467.

[23] N. N. Zahrani and R. Hedjar, "Comparison study of deep-learning architectures for classification of thoracic pathology," in *2022 13th International Conference on Information and Communication Systems (ICICS)*, 2022, pp. 192–198.

[24] Z.-P. Jiang, Y.-Y. Liu, Z.-E. Shao, and K.-W. Huang, "An improved vgg16 model for pneumonia image classification," *Applied Sciences*, vol. 11, no. 23, 2021. [Online]. Available: https://www.mdpi.com/2076-3417/11/23/11185

[25] K. Joshi, V. Tripathi, C. Bose, and C. Bhardwaj, "Robust sports image classification using inceptionv3 and neural networks," *Procedia Computer Science*, vol. 167, pp. 2374–2381, 2020, international Conference on Computational Intelligence and Data Science. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050920307560

[26] C.-M. Briquet, "Briquet database," https://briquet-online.at/, 2021, accessed: May. 2, 2024.

[27] G. Piccard, "Piccard database," https://www.piccard-online.de/start.php, 2021, accessed: Apr. 22, 2024.

[28] Institute for Medieval Research, "Bernstein project," https://memoryofpaper.eu/BernsteinPortal, 2009, accessed: May 2, 2024.

[29] X. Shen, I. Pastrolin, O. Bounou, S. Gidaris, M. Smith, O. Poncet, M. Aubry, "Large-scale historical watermark recognition: dataset and a new consistency-based approach," *CoRR*, 2019, accessed: Apr. 22, 2024.

[30] O. Bounou, T. Monnier, I. Pastrolin, X. Shen, C. Benevent, M.-F. Limon-Bonnet, F. Bougard, M. Aubry, M. Smith, O. Poncet, and P.-G. Raverdy, "A web application for watermark recognition," *Journal of Data Mining Digital Humanities*, vol. Atelier DigitHum, 07 2020.

[31] V. Pondenkandath, M. Alberti, N. Eichenberger, R. Ingold, and M. Liwicki, "Identifying cross-depicted historical motifs," *CoRR*, vol. abs/1804.01728, 2018. [Online]. Available: http://arxiv.org/abs/1804.01728

[32] E. Beriozchin, S. Pfaff, P. Weyh, "Detection and classification of historic watermarks using neural networks and nearest neighbor search," *Book of Abstracts - DHd2024*, February 2024. [Online]. Available: https://doi.org/10.5281/zenodo.10698260

[33] E. Bernhardsson, "Spotify Annoy: A hypothetical integration of spotify and annoy library," 2024. [Online]. Available: https://github.com/example/spotify_annoy

[34] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[35] M. Hu, "Visual pattern recognition by moment invariants," *IEEE Transactions on Information Theory*, vol. 8, no. 2, pp. 179–187, Feb. 1962.

[36] S. S. Z. Chen, "A zernike moment phase-based descriptor for local image representation and matching," *IEEE Transactions on Image Processing*, vol. 19, no. 1, pp. 205–219, Jan. 2010.

[37] G. M. of Books and Writing, "German museum of books and writing database," 2023.

[38] B. Wang, "A parallel implementation of computing mean average precision," *ArXiv*, vol. abs/2206.09504, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:249889360

[39] K. Y. Xiao, L. Engstrom, A. Ilyas, and A. Madry, "Noise or signal: The role of image backgrounds in object recognition," *ArXiv*, vol. abs/2006.09994, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:219721312

[40] F. Chollet *et al.* (2015) Keras. [Online]. Available: https://github.com/fchollet/keras

A Usage of LLMs

Artificial Intelligence tools such as ChatGPT[3] and Grammarly[4] were occasionally utilized to assist in the writing of this report. They were not used to generate sentences or paragraphs from scratch, but rather to rephrase certain sections. For example, there were requests along the lines of "Could you rephrase X?" or "Can you make Y more formal, succinct, or polished?" where X and Y represented the sentences requiring modification.

It is important to note that during the finalization of the report, several editing stages resulted in significant changes to the text to improve brevity, grammar, and clarity.

B Rank distribution

This section presents the distribution of the ranks of the first relevant similar watermark for each query watermark. The first 8 images represent the first scenario, while the subsequent 4 images represent the second scenario.

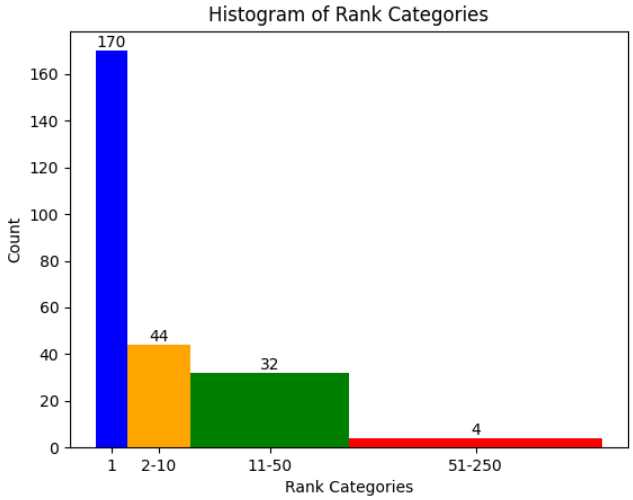

Figure 7: Histogram of rank categories for VGG16 in Scenario 1 - Traced Watermarks.

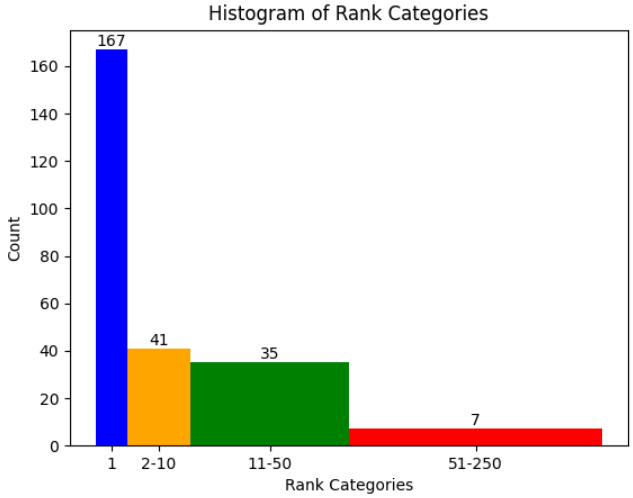

Figure 8: Histogram of rank categories for ResNet50 in Scenario 1 - Traced Watermarks.

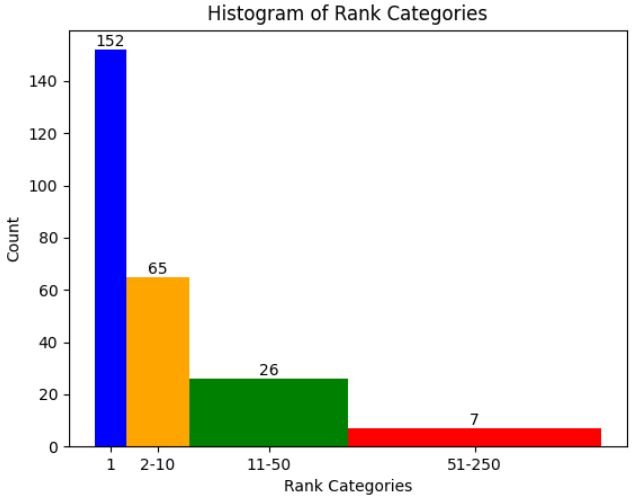

Figure 9: Histogram of rank categories for EfficientNetB0 in Scenario 1 - Traced Watermarks.

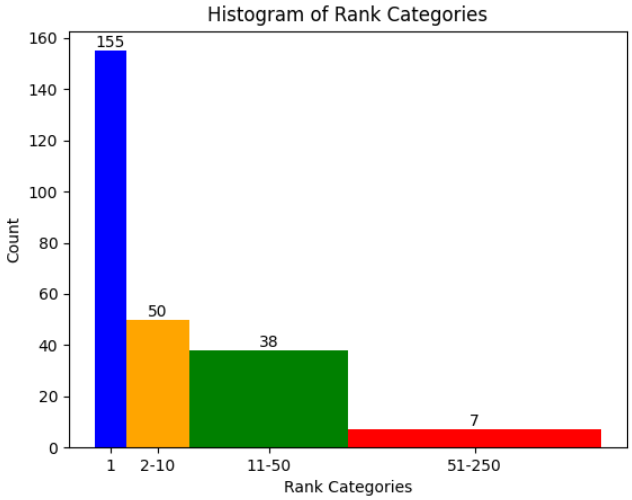

Figure 10: Histogram of rank categories for InceptionV3 in Scenario 1 - Traced Watermarks.

[3]https://openai.com/blog/chatgpt
[4]https://app.grammarly.com/

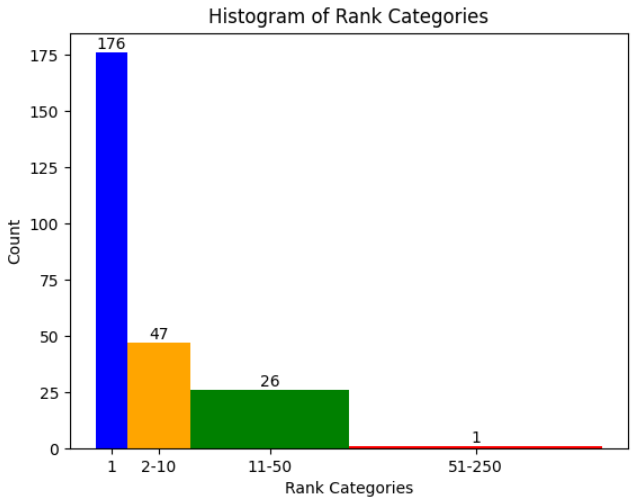

Figure 11: Histogram of rank categories for VGG16 in Scenario 2 - Traced Watermarks.

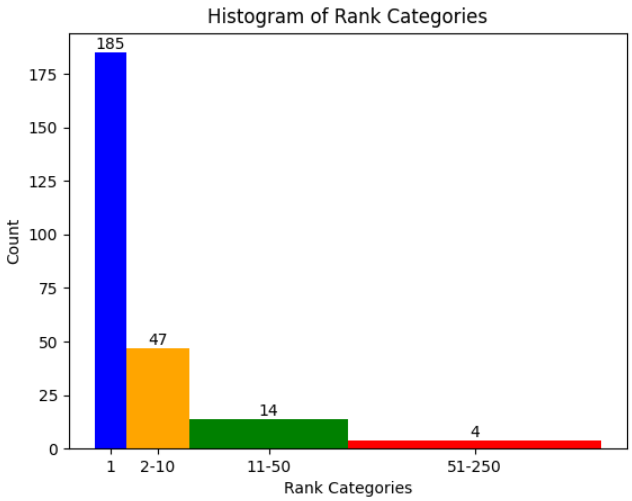

Figure 13: Histogram of rank categories for EfficientNetB0 in Scenario 2 - Traced Watermarks.

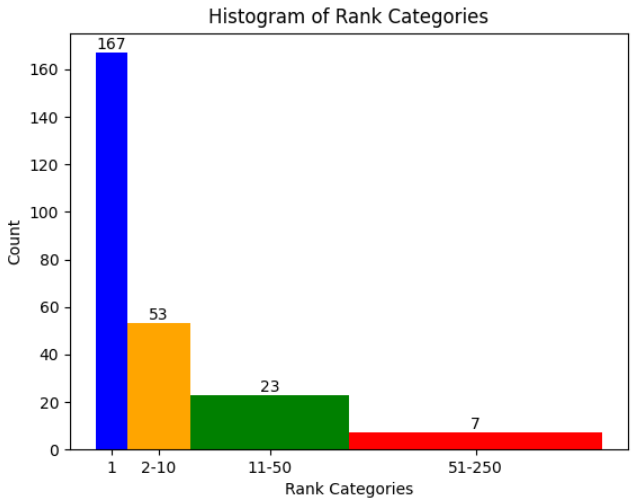

Figure 12: Histogram of rank categories for ResNet50 in Scenario 2 - Traced Watermarks.

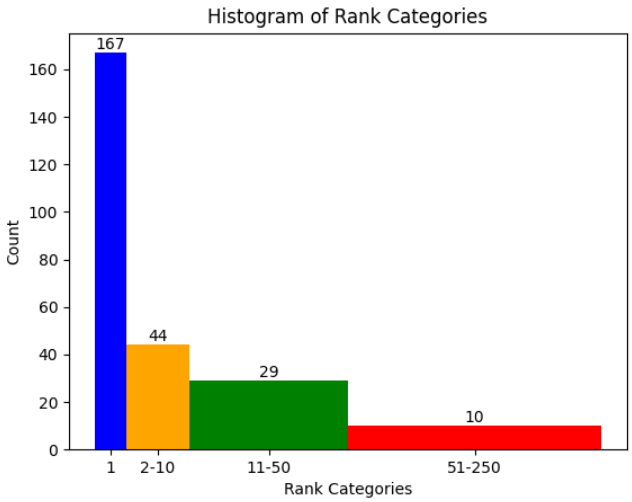

Figure 14: Histogram of rank categories for InceptionV3 in Scenario 2 - Traced Watermarks.

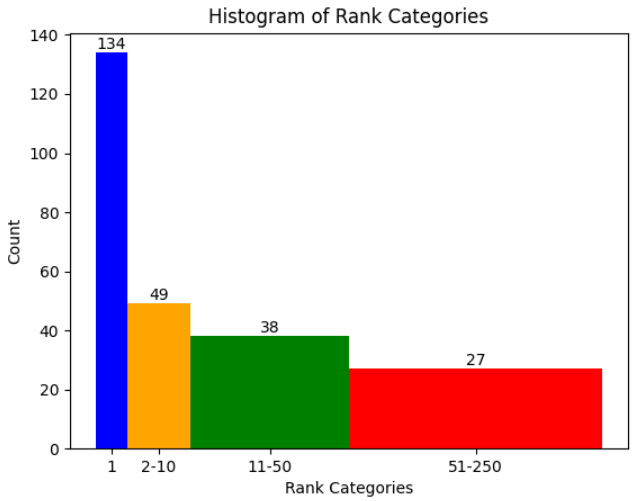

Figure 15: Histogram of rank categories for VGG16 in Scenario 1 - Untraced Watermarks.

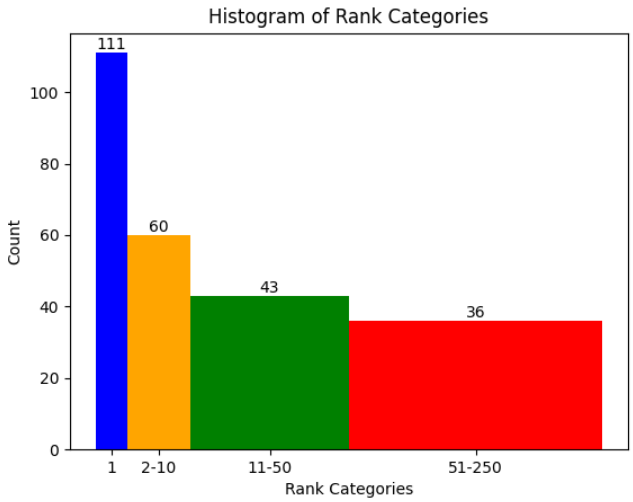

Figure 16: Histogram of rank categories for ResNet50 in Scenario 1 - Untraced Watermarks.

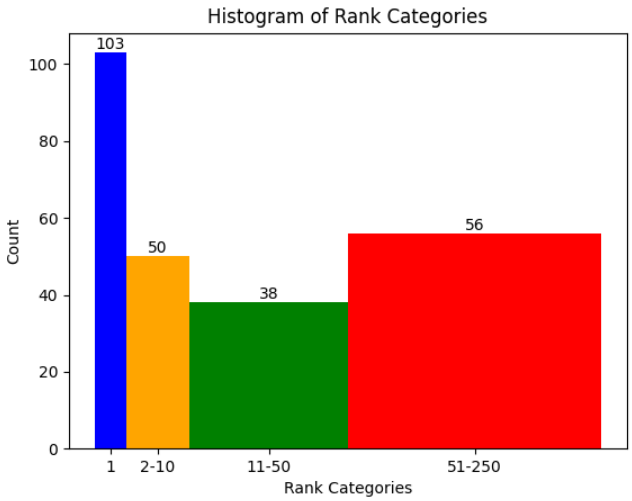

Figure 17: Histogram of rank categories for EfficientNetB0 in Scenario 1 - Untraced Watermarks.

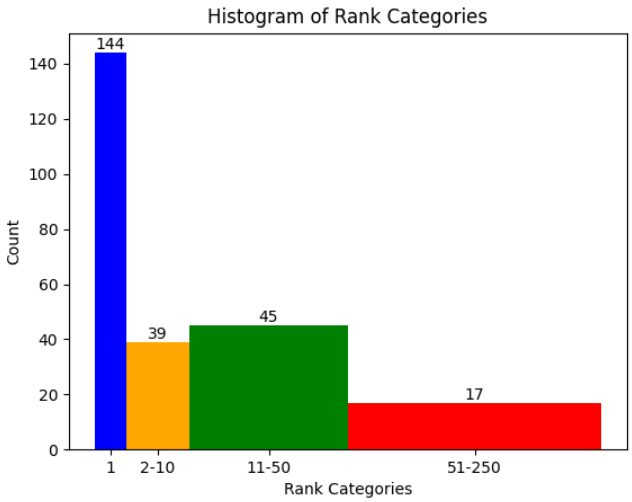

Figure 18: Histogram of rank categories for InceptionV3 in Scenario 1 - Untraced Watermarks.

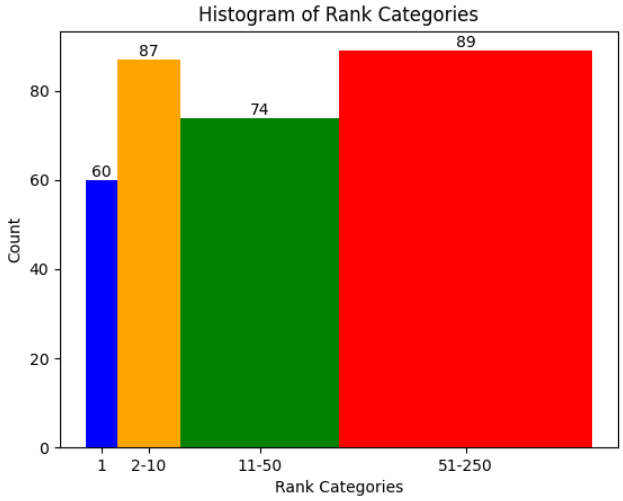

Figure 19: Histogram of rank categories for VGG16 in Scenario 2 - Untraced Watermarks.

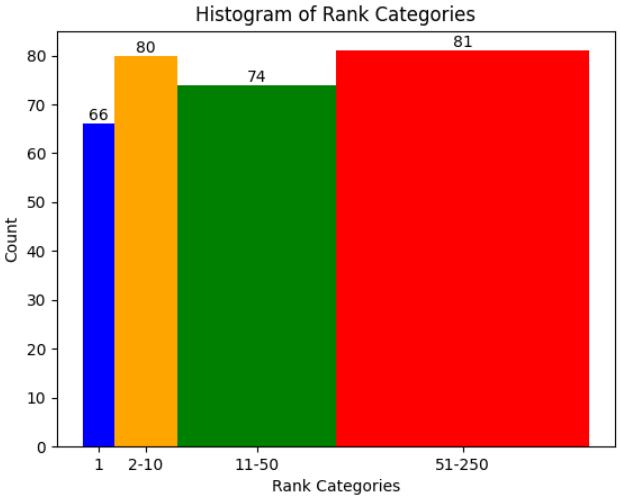

Figure 21: Histogram of rank categories for EfficientNetB0 in Scenario 2 - Untraced Watermarks.

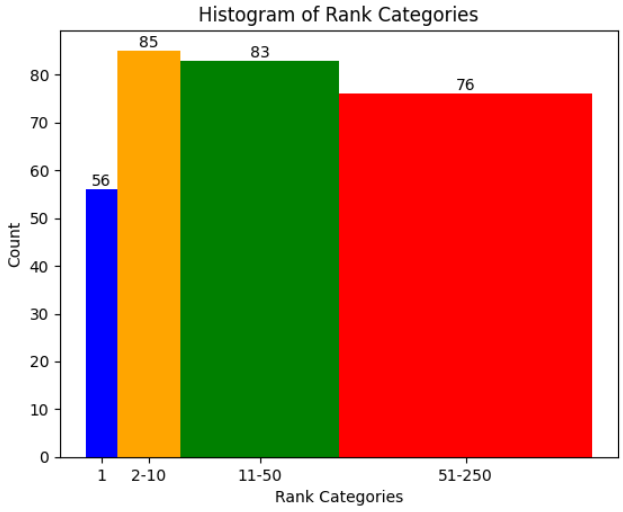

Figure 20: Histogram of rank categories for ResNet50 in Scenario 2 - Untraced Watermarks.

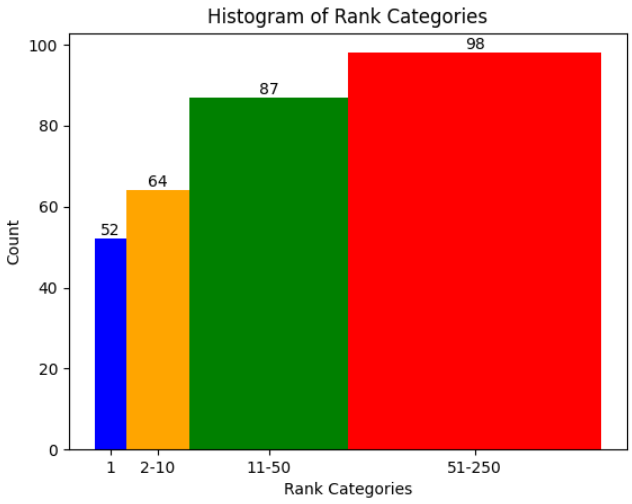

Figure 22: Histogram of rank categories for InceptionV3 in Scenario 2 - Untraced Watermarks.