# Searching for visual patterns in a children's drawings collection

*Links to famous artworks*

by Ravinithesh Annapureddy

Master Thesis

Approved by the Examining Committee:

Prof. Frédéric Kaplan
Thesis Advisor

Prof. Aurélien Bénel
External Expert

Dr. Julien René Pierre Fageot
Thesis Supervisor

Digital Humanities Laboratory (DHLAB)

École Polytechnique Fédérale de Lausanne (EPFL)
Lausanne, Switzerland

July, 2022

Cultivation of mind should be
the ultimate aim of human existence.
— B. R. Ambedkar

To all the beautiful minds...

# Acknowledgments

First of all, I am deeply indebted to my supervisors, Frédéric Kaplan and Julien Fageot, for their guidance and discussion and for supporting me throughout the project.

I would like to thank Carolina Suarez and her team at the IMAJ-UNESCO center for providing access to their digitized drawings collection, without which the project would not have been in its present form. Thanks again to Julien for arranging and making this project possible.

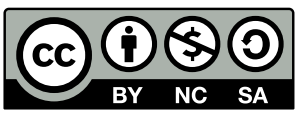
I am particularly grateful to Ludovica, who allowed me to use some parts of her work and the numerous coffee breaks.

I cannot thank you enough, Jithendra, for all of your unconditional support and constant encouragement.

Last but not least, I am grateful to my parents for their support and solace. Thank you, S.G., for being there.

*Lausanne, CH, July, 2022* Ravinithesh Annapureddy





# Abstract

The success of large-scale digitization projects at museums, archives, and libraries is pushing other cultural institutions to embrace digitization to preserve their collections. By juxtaposing digital tools with digitized collections, it is now possible to study these cultural objects at a previously unknown scale. This thesis is the first attempt to explore a recently digitized children's drawings collection while developing a system to identify patterns in them linked with popular cultural objects. Artists, as young as three and as old as 25, created nearly 90,000 drawings in the span of three decades from most countries in the world. The preliminary examination unveils that these drawings mirror a solid cultural ethos by using specific iconographic subjects, objects, and colors, and the distinction between children of different parts of the globe is visible in their works. These factors not only make the dataset distinct from other sketch datasets but place it distantly from them in terms of size and multifariousness of creations and the creators. The essential and another dimension of the project is matching the drawings and the popular cultural objects they represent. A deep learning model that learns a metric to rank the visual similarity between the images is used to identify the drawing-artwork pairs. Though the networks developed for image classification perform inadequately for the matching task, networks used for pattern matching in paintings show good performance. Fine-tuning the models increases the performance drastically. The primary outcomes of this work are (1) systems trained with a few methodically chosen examples perform comparably to the systems trained on thousands of generic samples and (2) using drawings enriched by adding generic effects of watercolor, oil painting, pencil sketch, and texturizing mitigates the situation of network learning examples by heart.

***Keywords*—** digitization, drawings, child art, visual similarity, pattern search, artworks, deep learning, transfer learning, style augmentation

# Résumé

Le succès des projets de numérisation à grande échelle dans les musées, les archives et les bibliothèques pousse d'autres institutions culturelles à adopter la numérisation pour préserver leurs collections. En juxtaposant les outils numériques aux collections numérisées, il est désormais possible d'étudier ces objets culturels à une échelle jusqu'alors inconnue. Cette thèse est la première tentative d'explorer une collection de dessins d'enfants récemment numérisée tout en développant un système permettant d'identifier dans ces dessins des modèles liés à des objets culturels populaires. Des artistes, âgés de trois ans à 25 ans, ont créé près de 90 000 dessins en l'espace de trois décennies, provenant de la plupart des pays du monde. L'examen préliminaire dévoile que ces dessins reflètent un ethos culturel solide en utilisant des sujets iconographiques, des objets et des couleurs spécifiques, et la distinction entre les enfants de différentes parties du globe est visible dans leurs ceuvres. Ces facteurs non seulement rendent l'ensemble de données distinct des autres ensembles de croquis, mais le placent loin d'eux en termes de taille et de multiplicité des créations et des créateurs. Une autre dimension essentielle du projet consiste à faire correspondre les dessins et les objets culturels populaires qu'ils représentent. Un modèle d'apprentissage profond qui apprend une métrique pour classer la similarité visuelle entre les images est utilisé pour identifier les paires dessin-objet. Bien que les réseaux développés pour la classification d'images ne soient pas assez performants pour la tâche de mise en correspondance, les réseaux utilisés pour la correspondance de motifs dans les peintures montrent de bonnes performances. Le réglage fin des modèles permet d'augmenter considérablement les performances. Les principaux résultats de ce travail sont (1) les systèmes formés avec quelques exemples choisis méthodiquement ont des performances comparables à celles des systèmes formés sur des milliers d'échantillons génériques et (2) l'utilisation de dessins enrichis par l'ajout d'effets génériques d'aquarelle, de peinture à l'huile, de croquis au crayon et de texturation atténue la situation du réseau qui apprend des exemples par cœur.

***Mots-clés—*** numérisation, dessins, art enfantin, similarité visuelle, recherche de motifs, œuvres d'art, apprentissage profond, apprentissage par transfert, augmentation de style

# Contents

# Chapter 1

# Introduction

## 1.1 Visual Arts

Art stimulates thoughts, emotions, beliefs, or ideas through the senses in an individual. Various forms of art deal with different sensibilities of the human body. For example, paintings, sculptures, and photographs concern the visual organs. Music is consumed through auditory organs. Then there are art forms such as theatre, dance, drama, and movies that involve both senses. Since the dawn of humankind, art forms have evolved in diverse ways. Yet from cave paintings to today's metaverse, the visual aspect of the art remains constant.

While creating arts constitutes a critical part of the human experience and portrays the world around us in the present moment, studying them throws light on our culture, lives, and the past experiences of fellow humans. Besides, the study can inspire, reflect and serve us in designing the future. A wide range of activities, from fine arts to product designs, constitute visual arts; this work refers to paintings and drawings as visual arts. Studying the relations between artworks and comparing them across time and space has been a fundamental activity in Art History. Besides, many studies have already established the importance of visual arts in developing a creative mindset at a young age [1], understanding the learning outcomes in children [2], and shaping humans' personalities [3].

## 1.2 Digitization and Analysis

Digitization and the online availability of art collections make the physical study of the art pieces a thing of the past. The markedly improved photographic and scanning technologies and their availability at comparatively inexpensive costs made digitization take a front seat at many libraries, galleries, museums, and archives. These digitized collections, especially publicly

available ones like the WikiArt [4], MET collection [5], or the Rijks Museum collection [6], have enabled all kinds of people to explore these art databases.

Viewing, remembering, and identifying thousands of images, if not millions, is an impractical task for humans. In parallel to the hardware advancement (photographic devices), progress in software and algorithms related to machine vision helped to create tools to view, analyze and improve the understanding of visual arts. Among the computer vision algorithms, algorithms that use neural networks have achieved and beaten human performance on many tasks such as classification and recognition in the last decade. A subset of algorithms that use multiple layers of neural networks, dubbed Deep Learning, which requires large volumes of data, takes credit for achieving unimaginable improvements. Among the deep learning techniques, *Convolutional Neural Networks* (CNN) pushed the limits of machines in solving vision-related problems. The ability of machine learning algorithms and the digital availability of vast art collections - due to digitization - opened a new avenue in the study of art.

The conjuncture of Art History and CNNs from deep learning became an active research area in the recent past. Multiple improved architectures of CNNs gave the state of the art performance in predicting the attributes of paintings [7, 8] and specifically on style [9], genre [10], and artist [11]. Shen et al., [12] developed systems for object and near-duplicate detection of artworks. In the latest round of developments, new art was produced [13, 14] using the art made by humans in the Generative Adversarial Networks [15]. The examination of similarity between artworks and artists has received profound interest in the last five years. Seguin et al., [16] proposed a supervised deep learning technique to search for similar visual patterns through metric learning. Further, methods to search for visual patterns in an unsupervised setting were proposed [17, 18] and all these works can broadly fit into the problem of clustering. Although there are comparatively fewer studies, other research areas in this conjuncture are fake art detection [19], art to photo translation (and vice-versa) [20, 21], emotion detection in paintings [22] and generating descriptions of artworks [23]. Santos et al. [24] provides a review of deep learning applications in visual arts.

## 1.3 Children and Visual Arts

Visual subjects like the people, animals, buildings, or surroundings serve as entry into this world for a child. Later, in school and private life, children view and produce various arts, which provides a rich learning domain for a young child [25]. Often, many might not regard the young children's drawings as novel creations and even compare them with the works of great artists or artworks. It is almost impossible to pinpoint the factors that influenced the child to make a particular drawing. There will be spontaneous instinct, vivid imagination, the influence of daily life, friends, family, teachers, and other unknown elements. While it is difficult to quantify many of these personal factors, it is possible to find the influence of artists and art forms in the drawings. Art viewing and the abundance of information available about renowned artworks

and artists could explain the impact of famous works in their drawings. At the same time, art schools use the famous works in teaching children and practice recreating them. These art classes could also be another reason that explains the resemblance between the drawings and famous artworks.

## 1.4 Artistic Analysis of Children's Drawings

Humans perceive objects through patterns, and as mentioned earlier, art historians try and find morphological, stylistic, and semantic similarities in the artworks. Along with conveying meaning, these patterns can provide insights into artists and the influences they might have undergone in producing the work. Again, in the context of children, the drawings (even scribbles by babies) were used in psychological [26], pediatric, and education studies [27]. In addition to the medical insights, investigating children's drawings helps to learn the impressions of historical, cultural, geopolitical, and socio-economic situations on the young children alongside exploring the themes, objects, artistic styles, and time period influences in their drawings. The unavailability of a large number of drawings produced by children makes such a study on a vast scale strenuous. This work attempts to move research in that direction.

Children move from scribbling to schematic representation and drawing realism to crises of adolescence stages during the growth and development of art in them [28] (as cited in [29]). In this process, the youngsters with the influence of popular cultural objects such as movies, music, and paintings try to recreate them[1]. Analyzing and mining such references provides an understanding of cross-cultural influences on children and the propagation of art beyond borders. Finding the connections in a drawing also helps move from simply looking at it as a sketch to taking a deeper look into it and unraveling the hidden connotations. To this end, this project endeavors to create a system to identify the popular artwork references in the children's drawings.

## 1.5 UNESCO Center in Troyes

The *Institut Mondial d'Art de la Jeunesse - Centre pour l'UNESCO Louis François* (World Youth Art Institute - Louis François Center for UNESCO) in the French city of Troyes aims to "promote creativity, develop artistic practices, and enhance the diversity of cultural expressions among young people." Located 140 kilometers from Paris and in the Champagne wine region, originally started as Cercle UNESCO de Troyes in 1978, *Louis François Center for UNESCO* is the only UNESCO center in France. It has undergone several name changes until its current name in 2019 and became a UNESCO center in 1994. With an objective to *inscribe childhood and youth*

[1]This work refers to the collections of popular paintings, posters of music albums, movies, dances, photographs of buildings, landmarks, people, and other famous cultural subjects collectively as popular/famous artworks.

*in the Memory of Humanity*, the center provides educational resources and organizes multiple workshops and competitions [30]. Their *Concours international d'arts plastiques* (International visual art competition) is the largest and most popular activity among the activities organized by the center to promote its objectives.

## 1.6 Graines d'artistes du monde entier

The international visual arts competition, titled *Graines d'artistes du monde entier* (Seeds of artists from around the world), started locally in 1985 and became a national competition in 1992, and within two years, it evolved into an annual international competition. Since 1994, the center has accumulated the artistic expressions of young people under 25 years old, under varying themes each year, and rewards medals and diplomas to a hundred laureates.

As part of the World Art Institute of Youth's mission, they conserve the works of all young artists who have participated in the competition. All the submissions received as part of the competition, starting from 1994, are now part of their *Mémoires du Futur* (Memories of the Future) art library. *Mémoires du Futur* museum already contains more than 100,000 artistic productions by young people aged 3 to 25. The drawings are spread over a quarter-century and originated from 150 countries, making the collection spatially and temporally diverse. The UNESCO center started digitizing this unique and invaluable collection to take them to a broad audience [30].

Making the drawings available on a digital platform accessible to everyone provides recognition to the children. On the other hand, it gives a chance to revisit the research on psychological, sociological, artistic, and historical aspects to improve the understanding of the creative notion of the child and its evolution on a broader scale with a large set of drawings. In addition, the digital copies provide an opportunity to study them using computer vision techniques. To this day, they have digitized more than 80% of the collection they have acquired in the past 28 years. Chapter 3 provides a detailed description of the collection.

## 1.7 Motivation

Discovering the artwork references, even in small numbers, with high confidence is essential in providing a cultural context for the drawings. Experts can quickly point out the popular connections among a handful of drawing pieces. The task becomes tedious and impractical in the case of comparing the works of different aged artists rooted across the world with the famous cultural objects across countries.

Children's drawings greatly differ from paintings or standard image datasets containing photographs. Children use diverse styles, materials, and techniques and do not necessarily use the

same method to recreate famous paintings. In addition, the detailing and morphological closeness of the drawing and artwork differ among children of different ages, making the comparison between them difficult. Learning specific patterns in images could lead to an unsuccessful comparison as the similarity between the drawing and the famous works are about correspondences which are neither among the low-level features such as color, texture, nor exact shapes and objects. At the same time, due to the minuscule number of examples available, the system can learn those pairings by heart. The comparision also differs from the sketch retrieval systems that commonly use pencil sketches for training and evaluation [31] lacking the diversity in the different genres of drawing. This uniqueness characterizes the task at hand not as a standard image retrieval problem or seldom a typical machine learning task and places it in the territory of cross-domain drawing retrieval. Only a limited amount of work is available on such a subject (discussed in Section 2.5), and the existing ones require heavy computation.

In light of these scenarios, cross-domain drawing retrieval is a clearly defined complex problem. However, the examples of drawing - artwork pairs show that they cannot fit into a singularly defined concept but needs a fluid boundary. In addition, the interpretations of similarity differ with an individual and the context, making it a necessary step to visit the primary sources and codify the constraints. These are typical traits of a Digital Humanities problem. The artwork matching is considered in the Digital Humanities domain and attempts to solve such an intricate problem using the digital computational tools, in this case, deep computer vision tools.

## 1.8 Thesis Goals

This thesis

- primarily aims to devise techniques to identify famous artworks that are visually similar to the children's drawings and
- provide the first insights into the recently digitized database of children's drawings

The recent boom in digitization pushed researchers to create efficient pipelines to digitize documents containing text, art, images, and many other types of information. Then the focus shifts to providing functional access to digitized data, and this thesis form part of such ongoing efforts. However, it attempts to extract similar images in two different datasets different from the access based on traditional indexing using the metadata. Matching images that vary in texture, colors, method, and techniques pose a challenge to current visual image search methods. It is more pressing when one set of images are creations of children. The earlier works are also limited in dataset size and identifying the exact paintings children have recreated. The recently digitized massive collection of drawings will provide a chance to look at the problem coming over previous constraints and a case to furnish the first impressions about the dataset.

## 1.9 Organization of the thesis

The thesis constitutes seven chapters. The next chapter documents the related works, and the third chapter introduces and discusses more than 80,000 digitized children's drawings from the *Louis François center for UNESCO*. The fourth chapter presents formal modeling of the similarity detection problem between drawings and paintings and the methods used in this study. The fifth and penultimate chapter describes the experiments and provides the results. The sixth chapter discusses the results, analyzes the limits, provides suggestions for the improvement of the methods used to identify the similarity, and examines the possible potential future work. Lastly, the seventh chapter summarizes the thesis while highlighting the contributions of this work.

# Chapter 2

# Background and Previous Works

This chapter presents a general background to the concepts and methods used in the project and reviews the literature related to the problem this project is trying to solve. The first section discusses image feature representation and its evolution, followed by a summary of neural networks, an overview of image processing in deep learning using convolutional neural networks, and ideas of transfer learning and metric learning. The last section presents the previous work on finding similar artworks using visual features.

## 2.1 Image Representation: Hand-Engineered to Learned Features

An image, seen through human eyes, can be an ensemble of various elements such as landscape, animals, people, text, tools, or any object with shape and form. However, this is not the case with computers, where an image is an array of pixels in Red, Green, and Blue channels. Although the same arrangement of pixels in the image enables humans to interpret the elements, it is not sufficient for computer vision tasks. Thus the *features* that can convey the information present in the image are used to represent it in computer vision tasks.

Earlier methods involved crafting problem-specific features. Some known handcrafted features useful in object detection and localization include simple histograms and *Histograms of Oriented Gradients* (HOG) [32]. The *Scale-invariant Feature Transform* (SIFT) [33] or *Speeded-Up-RobustFeature* (SURF) [34] features based on local descriptors are beneficial in comparing images. Artificial neural networks (described in the next section) can benefit from these features. However, this feature engineering step requires expert knowledge.

More often than not, these feature descriptions are based on only one property (like texture, edge, or color) and sensitive to minor variations, and generalize poorly. Convolutional neural networks address this problem of feature engineering where the *convolution* layers transform the

input (image) before passing through a feed-forward network to perform classification, detection, or localization. These convolutional layers *learn* the *features* of the image and eliminate the need to hand-engineer them.

## 2.2 Artificial Neural Networks

*Artificial Neural Networks* (ANN), inspired by the neuron connections in the brain, were ideas proposed nearly seven decades ago. With the increasing computing power and algorithms to train, they have been an active research topic since the 1980s [35]. An ANN is composed of layers of neurons, the first layer accepts the inputs and transforms them non-linearly in the hidden (or middle) layers, and the last layer provides the output. The number of neurons in the first and last layers is the number of input and output features of the problem, respectively. The number of hidden layers and the neurons in each layer are hyperparameters. Each neuron accepts the input from all the neurons in the previous layer, multiplies it by a certain weight, and transforms it using non-linear functions such as Sigmoid, TanH, or ReLU. If there are no hidden layers, an ANN is a non-linear function applied to the input, and adding multiple hidden layers makes it a Deep Neural Net. The weights between the layers are initialized randomly and optimized until the network meets the desired criteria. The networks are commonly optimized using the backpropagation algorithm that propagates the error from the output to the input layer and updates the weights using gradient descent methods.

Input to an ANN is one dimension vector and thus makes it difficult to process images that are three-dimensional tensors with Height, Width, and Color (usually three) dimensions. Although it is possible to flatten a three-dimensional image into a one-dimensional vector, it increases the size of the network. Hence, the images are represented using a handcrafted feature vector and presented as an input to the ANN. The previous section mentions the drawbacks of using hand-engineered features, and Convolutional Neural Networks are used to overcome them.

## 2.3 Convolutional Neural Networks

*Convolutional Neural Networks* (CNN) accept the image tensor in its original form and passes it through convolutional layers before flattening and passing it through a fully connected layer to perform classification or prediction. The operation in the convolutional layers is similar to the conventional image processing convolution operation. However, the weights of filters are not predetermined but updated when training the network to perform the desired task. Although a tensor of any shape can be an input to the CNNs, as they primarily process images, the rest of the thesis assumes an image as an input to the convolutional network.

The *Alexnet* [36] proposed in 2012 brought back the attention to CNNs when it used Graphics

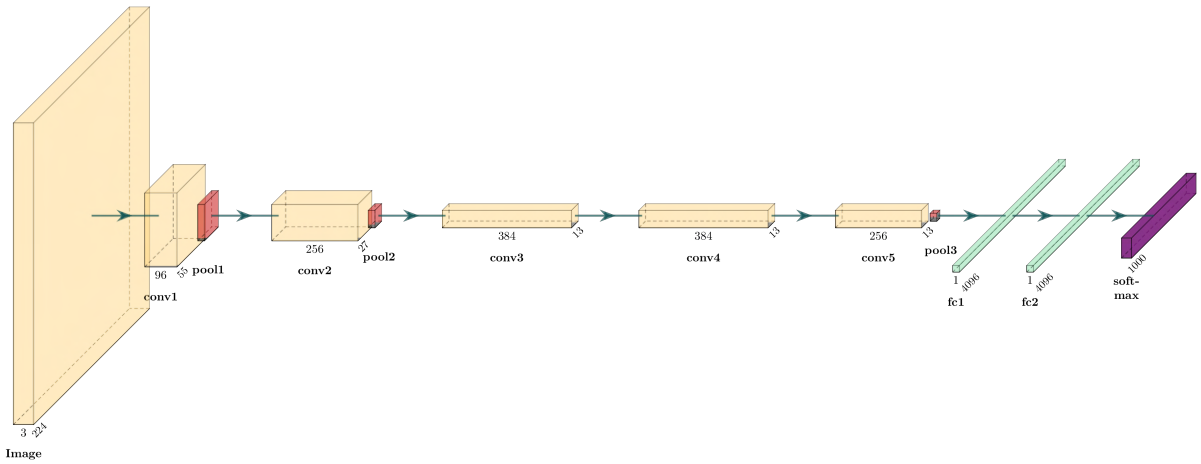



Figure 2.1: Alexnet [36] architecture, revisualized using [37]

Processing Units to train the model and won the *ImageNet Large Scale Visual Recognition Challenge* [38] by a margin of nearly eleven percentage points compared to the next best performance. *Alexnet* (Figure 2.1) has five convolutional layers (prefix *conv*), and in between them, pooling layers, indicated with prefix *pool*, are used to reduce the dimension of the output. The output of the last convolutional layer is flattened and passed through two fully connected layers (prefix *fc*) before predicting the class of the image (out of 1000 possible classes) in the last layer. Thus, the intermediate convolutional layers on the CNN *learn* hierarchal features while solving the problem, and the convolutional layers act as feature extractors.

### 2.3.1 Residual Networks

After *Alexnet*, many variations of the CNNs - Residual, Inception, Dense, Highway, Attention-based networks, and various learning tricks - Data Augmentation, Dropout, Rectified Linear Unit activation, Batch Normalization came into existence. Out of such improvements, residual networks proposed by Kaiming et al. [39] became popular to fight the problem of *vanishing gradients* in deep networks. During multi-layered neural network training, the gradient (change in weight with respect to the error) used to modify the weight becomes negligible (vanishes) before updating the initial layers because of the depth. The residual networks try to solve this problem by using a skip connection that adds a link between two points in a network to skip any modification on the input. Figure 2.2 shows a residual block (presented in [39]). This technique retains the original properties of the input image and the already learned transformations on it.

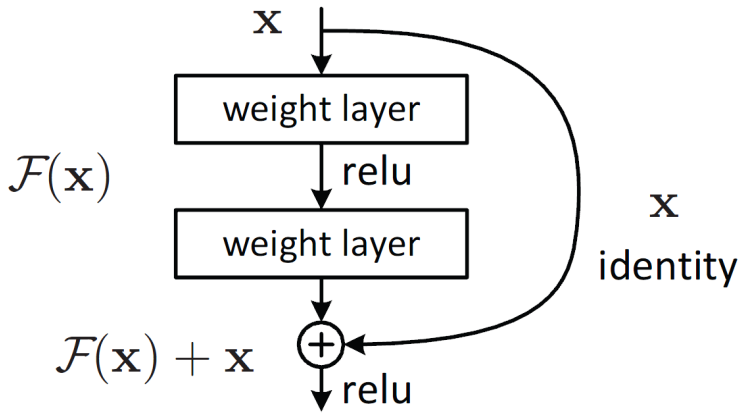



Figure 2.2: Residual block with a skip connection. Adapted from [39].

**ResNeXt**

Another family of CNNs are Inception models that use a split-transform-merge strategy. The inception network splits the input into a few lower dimensional embeddings, transforms them with specialized filters for each embedding (called path), and concatenates through depth to merge them. Although Inception modules are known to have lower computational complexity, their adaption for new tasks/datasets is not clearly defined [40]. ResNeXt network architecture proposed by Xie et al. [40] combines the split strategy of the Inception network and the idea of depth from ResNet [39] that merges the transformed low dimensional inputs by summation instead of concatenation. Additionally, transformations applied to each input embedding are the same. As this architecture is more easily adaptable to new tasks than the Inception networks [40], the current project uses a model based on residual connections, specifically the ResNeXt architectures.

## 2.4 Few-shot learning

The availability of large datasets plays a central role in the success of Deep Learning methods, including other factors like network architecture, learning techniques, and graphic card computation. However, the possibility of obtaining hundreds or thousands of samples to train a neural net is not always high. *Few-Shot Learning* (FSL) approaches are used in training (fine-tuning) classification models when data is scarce. One-shot learning is a sub-category of few-shot learning methods where only one example is available per class. One-shot learning is widely used in computer vision tasks and gained popularity with face recognition studies [41]. Wang et al. [42] divide FSL methods into three categories. Based on the prior knowledge of the dataset, the first category deals with augmenting the training data, the constraints on the model hypothesis space fall in the second category, and algorithms to provide good initialization or instruct the parameter update.

The current work also suffers from the problem of fewer examples per class as a drawing generally refers to only one artwork, and all three types of methods mentioned earlier are employed to circumvent it. First, increasing the drawings made by the children by augmenting them into various styles like watercolors, pencil sketches, and oil paintings. Second, utilizing the Task-Invariant Embedding model (Metric Learning through Triplets) and the same model to process drawings and paintings (enables parameter sharing). Lastly, initializing the parameters of the CNN model obtained in training it for a classification task with Imagenet data [43] and fine-tune it from there (Transfer Learning).

### 2.4.1 Transfer Learning

Transfer Learning refers to using models trained for one task either for a second task or as a starting point to fine-tune the model for the second task. Generally, transfer learning uses the state of the art models, and the motivation for that is two-fold; training state-of-the-art models use large datasets and heavy computational power, and deep neural nets, especially CNNs, are known to learn the generic features in the convolutional layers [44].

There are two most common ways of using CNNs in transfer learning. First, CNNs act as feature extractors to obtain the feature representation of the new images through the convolutional layers of the pre-trained network. These features are then used in conjunction with other Machine Learning algorithms to compare, classify or visualize the data. Many works utilizing the CNNs trained on the ImageNet dataset [43] have achieved satisfactory performances even when the domain of images is different from the ImageNet [45–47]. The second way uses the pre-trained CNN as a starting point, and the network is fine-tuned on the specific dataset instead of end-to-end training the CNN for the new problem. The complexity of the problem and the similarity between the new dataset and the pre-trained CNN dataset determines the layers to retrain.

### 2.4.2 Metric Learning

Clustering, classification, or retrieval of images requires transforming them from their original space into a latent space that uses the feature representation and operations on these feature vectors fulfill the required tasks. Examples include using the euclidean distance between the feature vectors in K-means clustering of images or transforming them into a lower dimension (than the feature vector) space using Principal Component Analysis. However, the distances do not directly affect the creation of the feature vectors. Besides, unlike the classification or regression problems, it is hard to quantify the similarity between images, but it is possible to rank them based on their similarity. The idea of Metric Learning is to use image similarity or dissimilarity directly in training a feature generation model. Metric learning tries to reduce the distance between similar samples and increase the distance between dissimilar ones. Lu et al.

[48] provide an overview of using metric learning in deep neural nets for vision-related tasks.

### 2.4.3 Triplet Networks

Siamese and triplet networks are the popular variants of neural networks that use metric learning. The former network requires pairs of positive (similar) and negative (dissimilar) samples for training and it minimizes a contrastive loss. The siamese network accepts only positive or negative pairs at a time, and the training might not converge, or the network could overfit the training data. Triplet networks attempt to overcome this shortcoming, and as implied in the name, the network requires three images instead of two at a time for the training [47]. A triplet is composed of an anchor image, a positive sample similar to the anchor image, and a negative one dissimilar to the anchor image. Due to the triplets, the network simultaneously tries to minimize the distance with the positive sample while increasing it with the negative instance. For a set $T$ of triplets, the triplet network minimizes the triplet loss [48] defined as

$$L = \sum_{(an,p,n)\epsilon T} h(\tau + dist(\boldsymbol{x}_{an}, \boldsymbol{x}_p) - dist(\boldsymbol{x}_{an}, \boldsymbol{x}_n)) \tag{2.1}$$

where $\boldsymbol{x}_{an}, \boldsymbol{x}_p$ and $\boldsymbol{x}_n$ are the feature vectors of the anchor, positive and negative samples respectively, $dist(\boldsymbol{x}_{an}, \boldsymbol{x}_p)$ is the distance between anchor and the positive sample, $dist(\boldsymbol{x}_{an}, \boldsymbol{x}_n)$ is the distance between anchor and the negative sample, and $h(x) = max(0, x)$ is the hinge loss with $\tau$ as a positive threshold that makes the constraint valid with a margin or in other words the minimum distance between the positive and the negative samples is $\tau$.

#### Mining of Triplets

The selection of positive and negative samples is crucial to ensure convergence of the model training process. The choice for positive examples is straightforward but not easy to get as it requires manual annotation. On the other hand, it is easy to obtain negative samples (all the non-positive instances can be negative samples). However, they could lead to overfitting as their number is far greater than the positives, and at the same time, selecting the negatives that are already far from the anchor will disallow learning in the network. References [47, 49] suggest a hard sampling strategy to counter the under and over fitting, where non-positive artworks proximate to the anchor, obtained using the distance function $dist$, act as negatives in the triplets.

Using metric learning in training a network implies optimizing the distance measure for the task that uses the specific metric. Therefore, using the triplets chosen at the start of the training (offline selection) for the entirety of the training poses an issue where the model could fit triplets, and the loss vanishes, effectively stopping the training. The online triplet selection process deals with this problem, by selecting a new set of triplets after updating the model in each round of

training [47]. This work uses the online method of triplet selection.

## 2.5 Related work

Literature, methods, and engineering techniques related to image retrieval are abundant. This section does not provide a survey of those works but discusses the relevant ones in developing the solutions to the current task.

### 2.5.1 Image Retrieval using Computer Vision

Following the digital camera revolution and the recent digitization, massive databases of images are a reality. Browsing and searching through these images apiece is difficult even if they are indexed based on name, creation date, or other structural metadata, as people usually wish to explore the databases based on their content. The text queries that operate on the descriptive metadata such as titles, tags of contents, and other image-related keywords aid up to some extent. However, some elements of an image, like its morphological characteristics, cannot be described in words. *Content-Based Image Retrieval* (CBIR) tries to overcome these shortcomings using the visual image contents (color, shape, texture, spatial arrangement).

CBIR systems using the traditional features like GIST, affine-invariant Hessian regions, and SIFT/SURF [50] and learned features using CNNs as image descriptors [51] are popularly known. As examined by Zhou et al. [52], the datasets used in the experiments are photos of landmarks (*Holidays, Paris*), buildings (*Oxford-5K, ZuBuD dataset*), or logos (*FlickrLogos-32 dataset*). The difference in artwork images and photos from these datasets makes it difficult to use the models to retrieve artworks. Only recently, CBIR methods have been developed and evaluated for visual art objects [16–18] that use features obtained through CNNs. This thesis frames the problem of identifying the artwork referenced in a drawing as a retrieval task. The artworks similar to the queried drawing will be retrieved and ranked based on their closeness to the query.

### 2.5.2 Cross-domain Image Matching

Shrivastava et al.'s work [53] that searches for the same photograph in diverse lighting conditions or images similar to a painting or a sketch is the closest to the task of this project. Their system involves computing the HOG descriptor for each image and using one similar image and many dissimilar images to minimize a convex objective function of the Support Vector Machine (SVM) with hinge loss. While the method achieves adequate results, it has two drawbacks. First, their retrieval system does not search for exact matches of the sketches with the images. Second, due to the need to train an SVM at the time of querying, the computational cost of the method

is high as each iteration requires more time than the previous one. Also, they evaluate the sketch-to-image matching with 50 sketches of cars and bicycles and 50 paintings of outdoor scenes.

In distinction to the Shrivastava et al.'s [53] proposition, the current project attempts to find an exact match of the artwork present in the diverse set of drawings, and the use of CNN ensures that the computation time remains the same across iterations.

### 2.5.3 Visually Linked Paintings

The work of Seguin et al. [16] was among the first to use modern deep learning-based computer vision techniques to help Art Historians track the pattern propagation in paintings. They have created a system to visually search a digitized photo archive and detect photographs containing the same objects. As it is difficult to define and quantify similarity strength between photos, they have used the concept of partially ordering. By sorting the connections based on their relative similarity, they avoid explicit quantification of the similarity. They train a CNN to estimate the image similarity and find new replications in the archive using the trained CNN.

The system proposed by Seguin [16] inputs the image to a CNN to get a three-dimensional feature map - the output of the last convolutional layer - and uses spatial matching similar to SIFT-matching to obtain a similarity. Using the manually annotated connections between paintings and spatial matching as the similarity (distance) function, they fine-tune the CNN using the metric learning process. Their experiments have shown that systems using pre-trained CNNs perform significantly better than the Bag-of-Words approach used in traditional retrieval systems, and fine-tuning the CNNs further improves the performance.

# Chapter 3

# Children's Drawings Dataset

This chapter presents digitized children's drawings. The first section provides additional context about the art competition, and the following section details the dataset quantitatively. The last section describes the insights obtained through the exploration of drawings and possible directions of studies using this dataset.

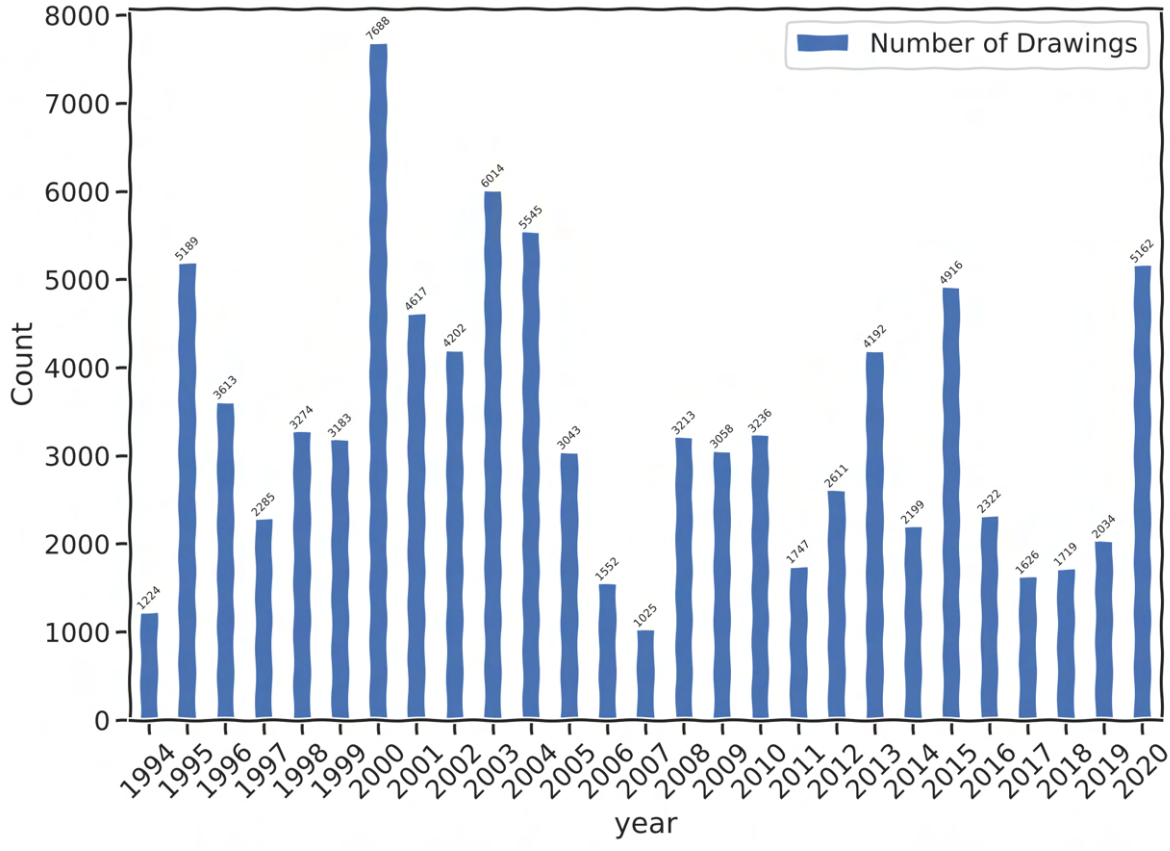



Figure 3.1: Number of digitized drawings per year

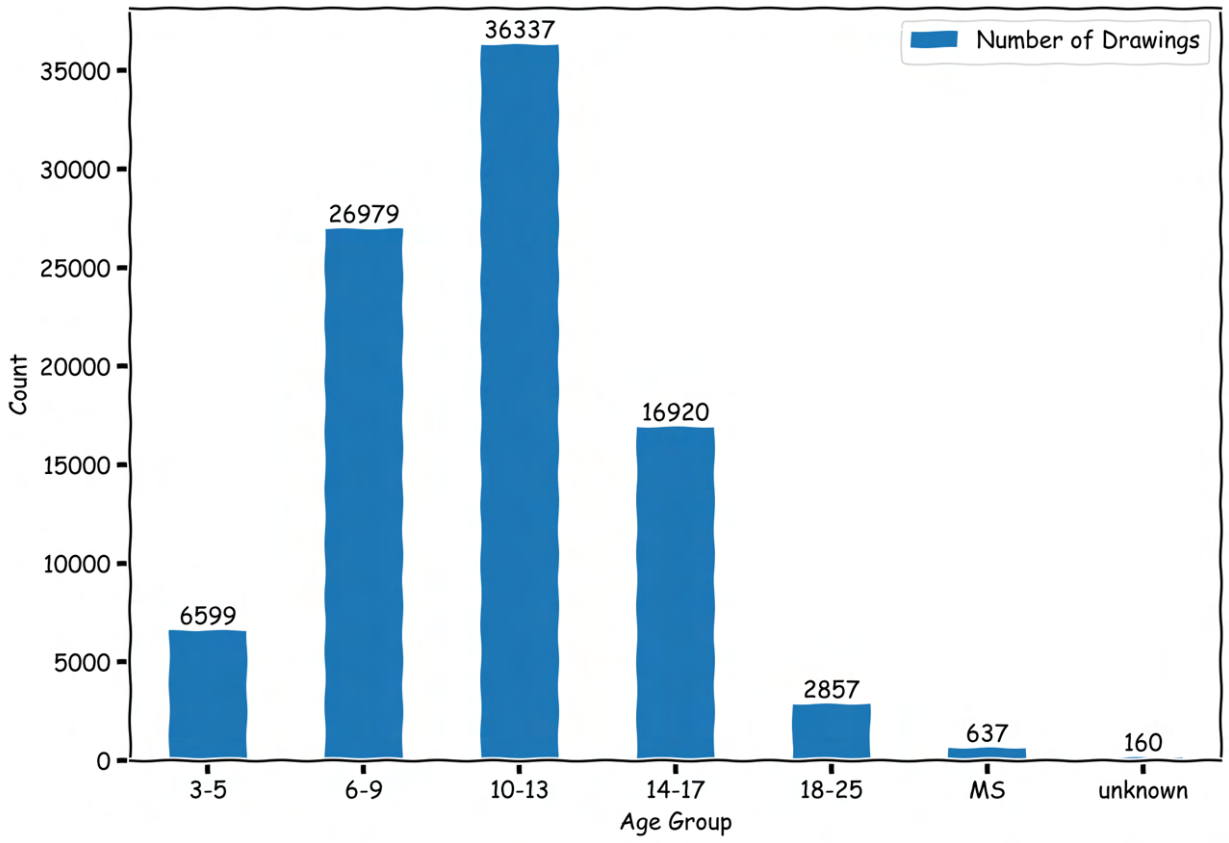



Figure 3.2: Number of digitized drawings per age group

## 3.1 Graines d'artistes du monde entier

The *Seeds of artists from around the world* competition provides a space to discuss, deliberate, and develop ideas about the happening issues of this world. The *Institut Mondial d'Art de la Jeunesse* or *World Youth Art Institute* (IMAJ, in short) selects a topic that challenges the planet as a theme and centers the competition around that theme. IMAJ conducts the competition in collaboration with more than 1000 institutions from 150 countries around the world. As mentioned previously, IMAJ continues to preserve more than 100,000 submissions they received since the competition's inception at the *Mémoires du Futur*. To further preserve the multi-dimensional diverse cultural heritage, IMAJ initiated the digitization process of the drawings collection in 2017, and more than 90,000 works are already digitized. Besides conservation, other digitization objectives include allowing digital consultation, facilitating research, and enabling analysis using computational tools of the collection [30].

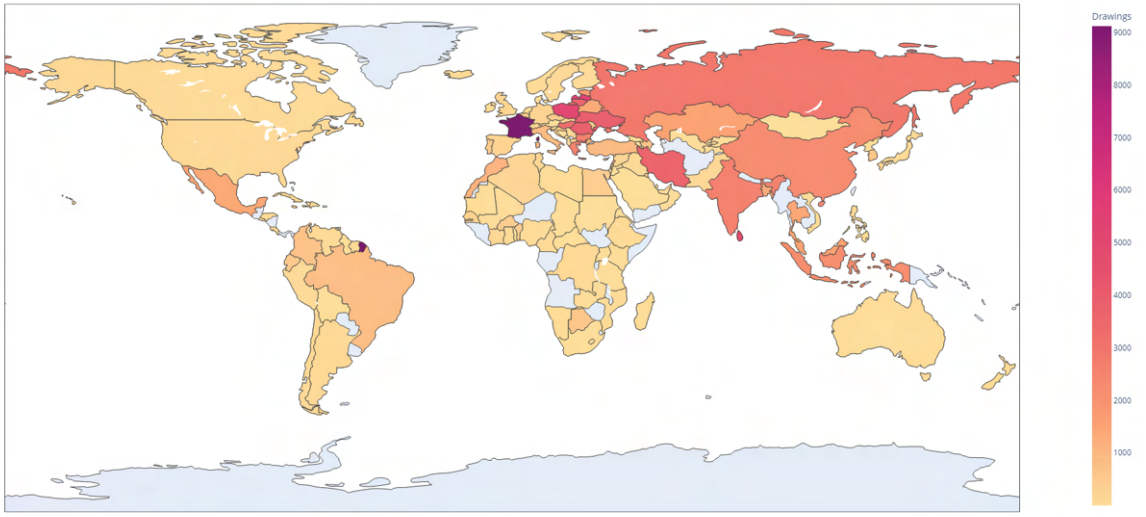



Figure 3.3: Number of digitized drawings per country

## 3.2 Digitized Drawings Dataset

Any submission to the art competition has at least two sides, the illustration on the front and some metadata on the back. The metadata of each drawing is almost always handwritten and very few times in print. At the time of writing this thesis, 179,427 scans (including the front and the back) were available, and nearly 90,000 contained an illustration[1]. The drawings are categorized based on young artists' country and age group. The file name of the scan, made up of six parts, provides this information. The six parts are the year of submission, age group, unique id, country of origin, side of the scan, and the quality of the scan (high-quality scan for archiving and standard quality for consultation).

Elemental analysis of this metadata reveals that the collection has 26 years of drawings from 1994 to 2020 (including both years). Also, there are six age groups: 3-5, 6-9, 10-13, 14-17, 18-25, and Medico-Social (MS) for children with special needs. On average, 3352 drawings are available per year, with a significant hike in a few years, and Figure 3.1 shows their spread over the years[2]. Figure 3.2 shows the drawings available per age category, and they predominantly belong to the 10-13 (40%), 6-9 (30%), and 14-17 (18%) groups. Table A.1 shows drawings distribution per year across each age category. Lastly, Figure 3.3 shows the drawings per country on a map. France is the highest contributor with 10% (9120), followed by Latvia (5.7%), Poland (5.4%), Sri Lanka (5.1%), and Ukraine (4.2%) in the top five. The age and country are not available in the filename for nearly 985 drawings. Figures 3.4 - 3.9 show the sample drawings from 6 different

[1]The number of drawings is not exactly half of the scans because some of them had more than one front side

[2]As some of the drawings are yet to be available in digital format, it is too early to remark on the low number of drawings in specific years.

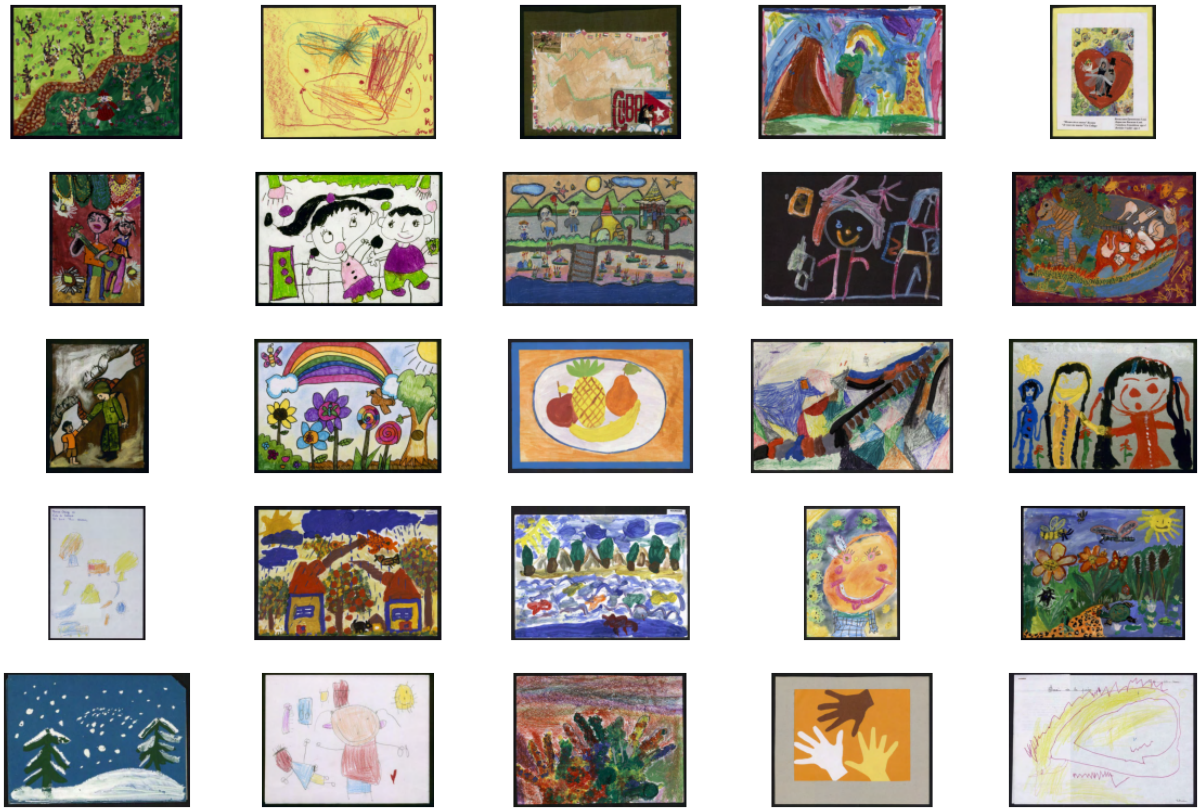


Figure 3.4: Sample drawings by children between 3 to 5 years

age categories.

## 3.3 Insights and Research Directions

Before zeroing on the current research question, the drawing collection was explored manually. While the rest of the thesis discusses identifying drawings similar to well-known artworks, this section summarizes the insights obtained during the exploration and provides directions for further research.

Every year, the children base their illustrations on the theme provided by the IMAJ, producing a diverse set of drawings that vary in technique and substance. The techniques include crayons, pencil sketching, watercolours, gauche, pastels, acrylic, ballpoint pen, ink, photos, and computer graphics. Parallelly, the children use a diverse set of styles, Fauvist, Cubist, Surrealist, Figurative, Hyperrealist, and Abstract, to name a few. Unsurprisingly, the drawings of children under five are primarily scribbles and re-creation of common subjects around them like houses, rivers, and mountains. This novelty steadily changes through the ages of 6 to 18, and substantial new creations appear in older individuals. A clear distinction exists between the works of young

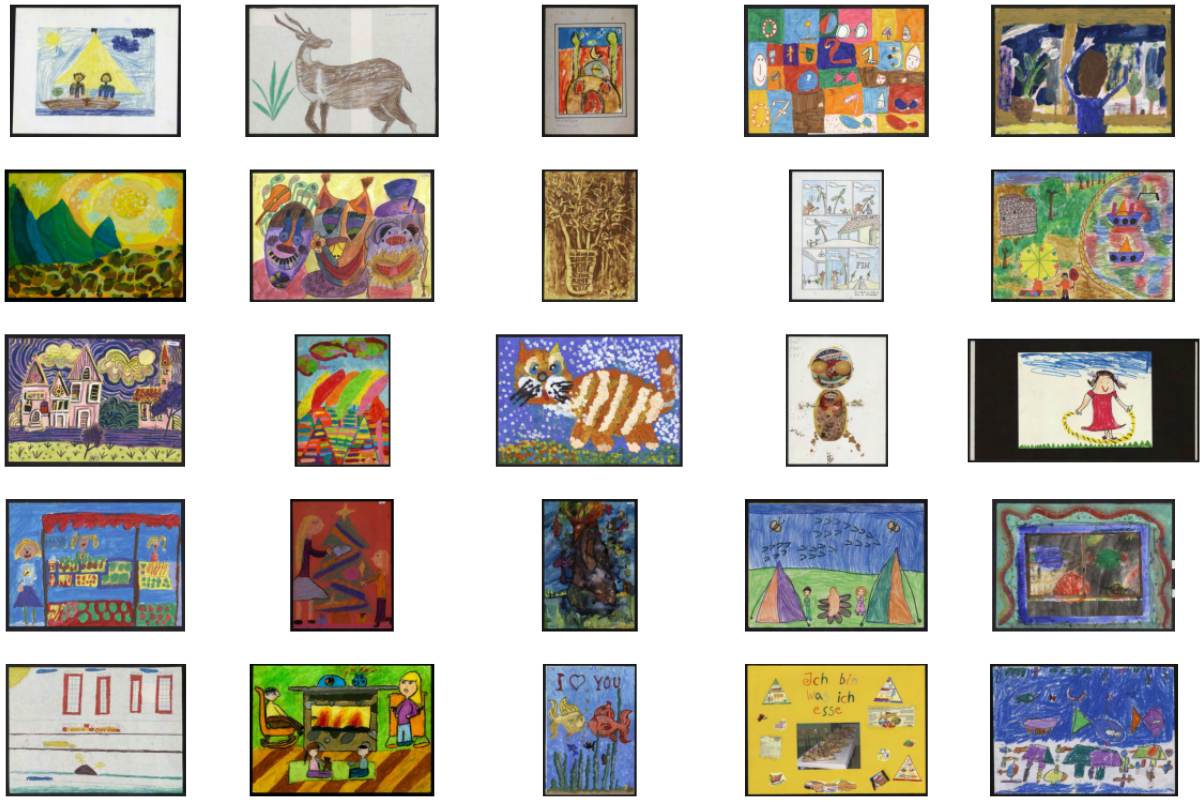



Figure 3.5: Sample drawings by children between 6 to 9 years

artists based on their country in terms of style. Nonetheless, the references in the drawings cross the nation's borders, and the artistic response to international issues is at the same level as domestic ones.

### 3.3.1 Metadata Extraction

The metadata about the artist and the school is available on the verso (back) side of the drawing. With some minor differences, overall, the drawings contain the metadata detailed below. However, only the age and country information was extracted during the digitization process and stored in the file name.

- About the Artist
  - Name
  - Address
  - Phone number
  - Birthday/Age

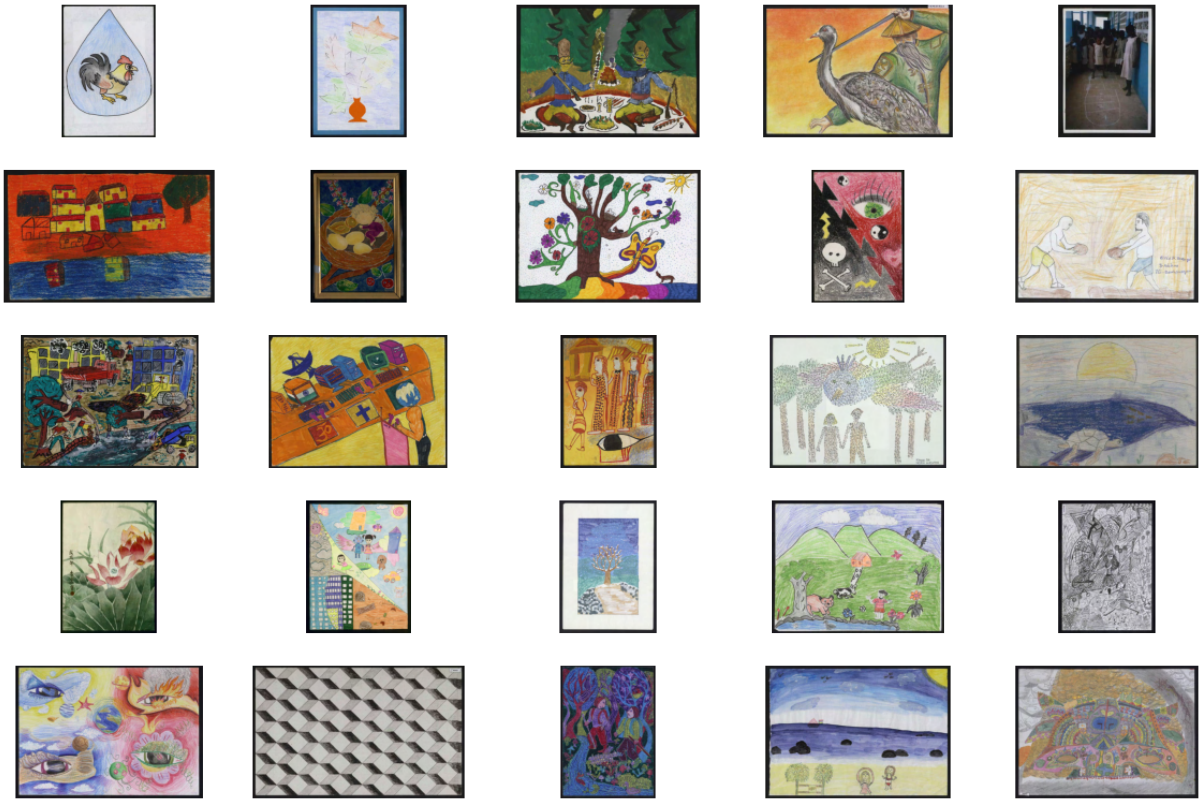


Figure 3.6: Sample drawings by children between 10 to 13 years

  - Sex
  - Disability status

- About the Drawing
  - Title/Description
  - Technique
  - Size

- About the institution
  - Name of the institute
  - Address of the institute
  - Name of the teacher/professor

Extracting this multi-lingual data that is partially handwritten and partially in print is challenging. At the same time, this metadata offers an opportunity to identify the artists and schools and analyze the works at a more granular level. In addition, this metadata will aid the ongoing efforts at the IMAJ to create a web platform to access the collection. Lastly, the description of the work

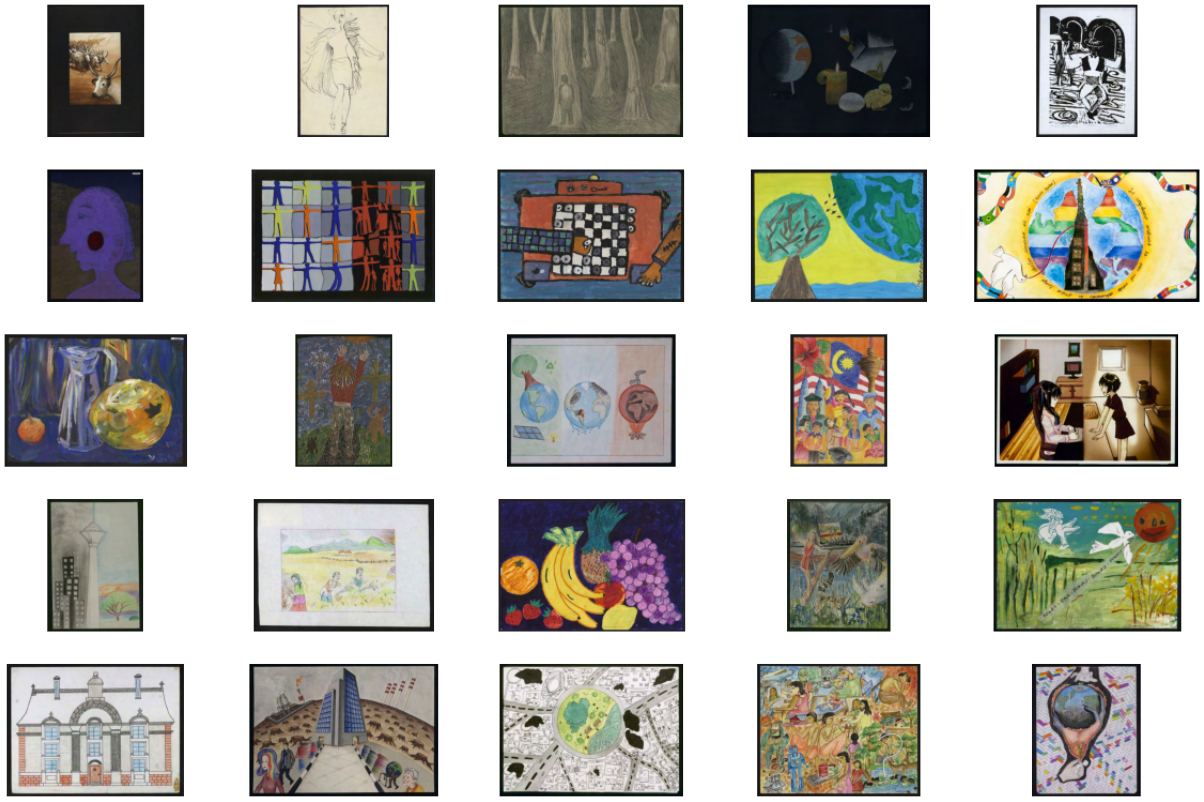


Figure 3.7: Sample drawings by children between 14 to 17 years

by the young artists themselves gives a clear and unfiltered explanation of their work and helps us understand their interpretation of the theme.

### 3.3.2 Iconographic analysis

The influence of culture, especially the traditions at various levels, is clearly reflected in the drawings. These ideas are represented as subjects or background objects or through colors. For example, religious symbols such as *Stupas* or Christian art often appear in drawings from some regions. Some works commonly contain animals and houses. These icons appear irrespective of the themes, and a correlation between age and place of origin was observed. In line with the current work, deep learning models can help in detecting and localizing icons in the drawings. Exploring this treasure of iconography aids in understanding the evolution of their usage, tracing the circulation of symbols in drawings, and the social patterns they reveal.

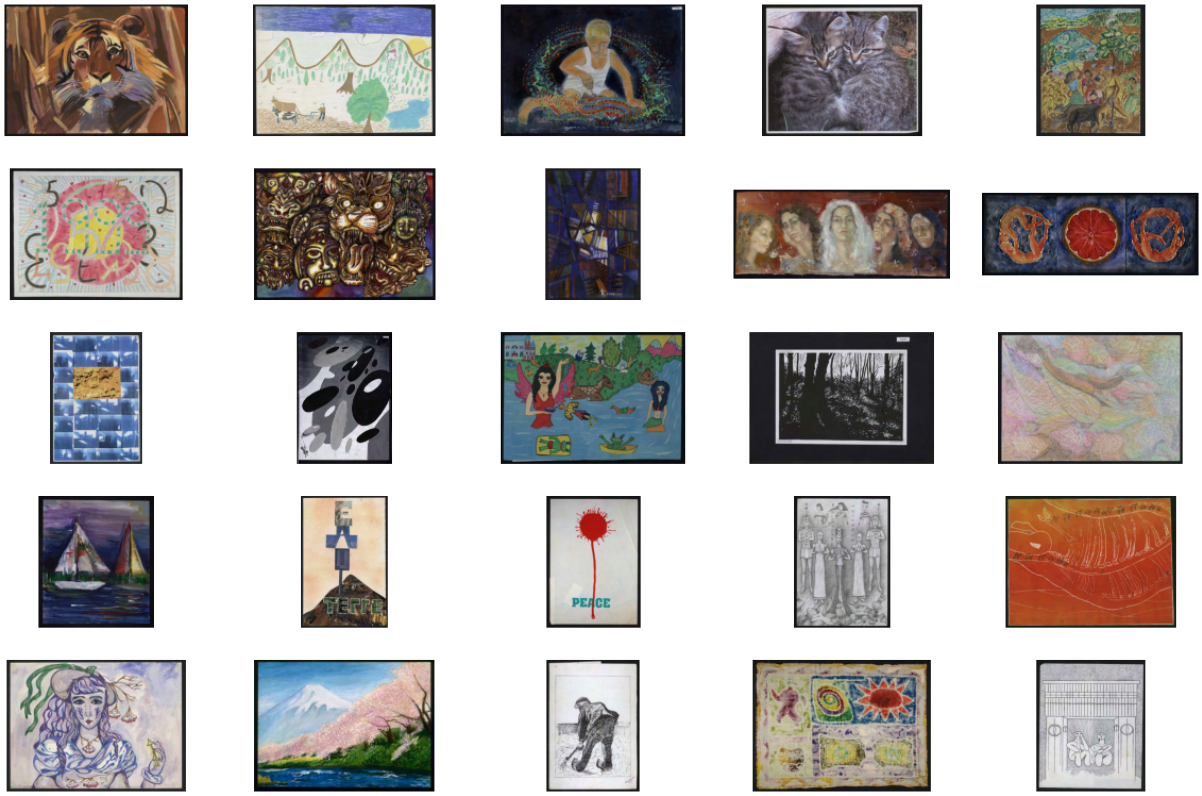



Figure 3.8: Sample drawings by children between 18 to 25 years

### 3.3.3 Realization of themes by young minds

Although each child receives an identical description of the competition theme, their realization of it as a drawing differs from one another. Each one uses backgrounds, elements, and objects differently from the others. Extracting the information about these entities present in the drawings across countries, age groups, and themes and looking through the historic date calendar provides an opportunity to understand the effect of socio-economic and political events on the children. It could also unearth classic cross-cultural references used to represent the same or similar ideas by children across the globe.

### 3.3.4 Clustering and artistic signatures

For some years, the themes of the competition were repeated with a different title, and most submissions were through art schools or institutions. These two factors might have influenced the children to reuse patterns, and ideas, improve existing work, or work together with a friend. In fact, some drawings are complete only when combined with others. This last direction of work proposes to group the creations by the children that share a visual and thematic similarity

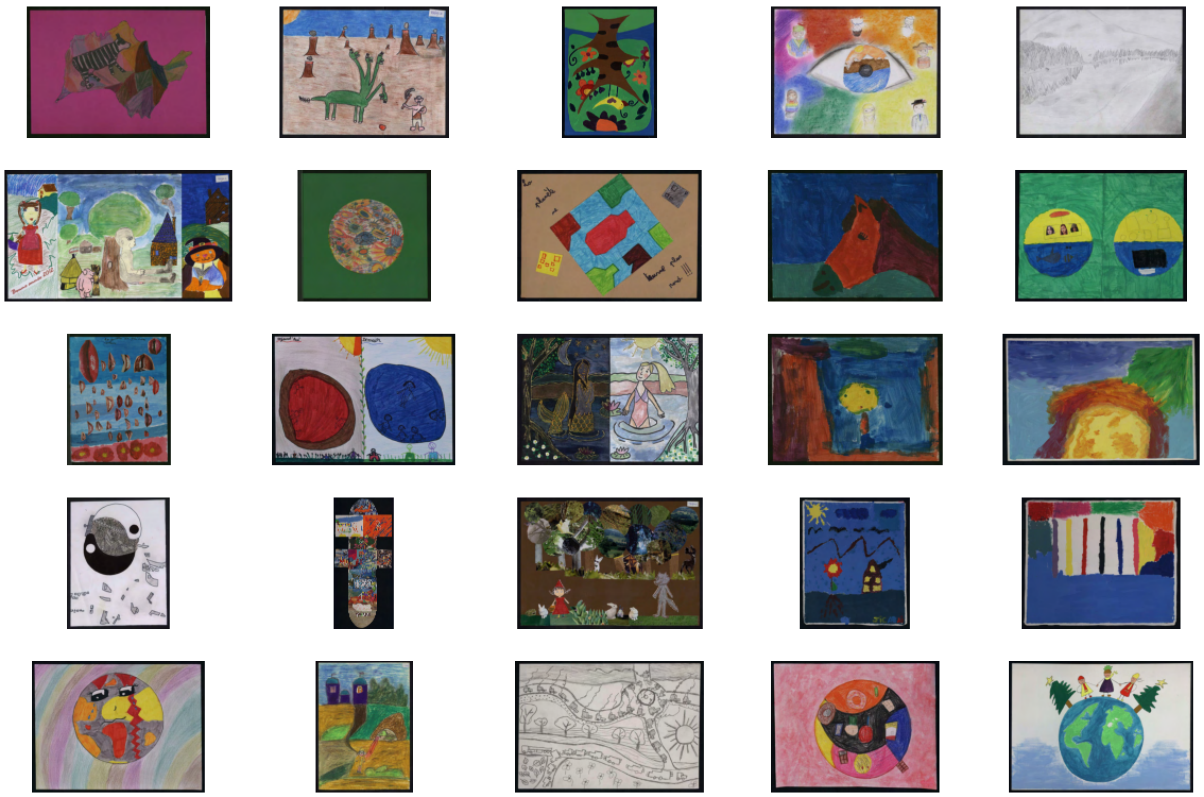


Figure 3.9: Sample drawings by children with special medical conditions

separately and together. This investigation can help declutter up to what extent the drawings from an institute are homogenous and the tendency to repeat patterns.

# Chapter 4

# Formulation and Methods

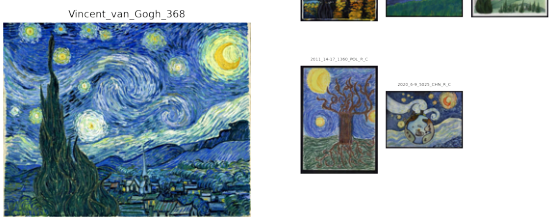
This chapter describes the goals for the project and the methods employed to achieve them. The first two sections explain the class of similarities between artworks and drawings considered in this project and present examples. Then, the third and fourth section provides a formulation and describes the model architecture along with the data annotation process. The last section details metrics used to evaluate different models.

## 4.1 Problem Statement - Objectives

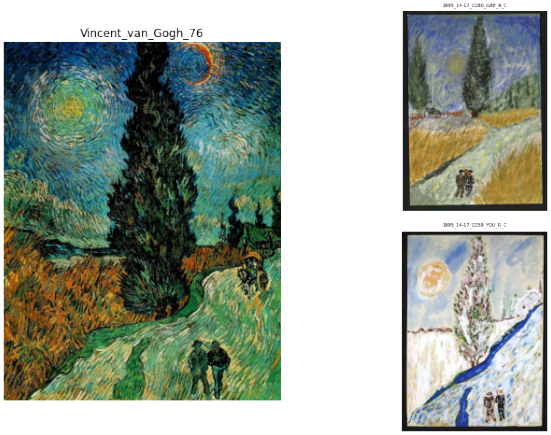
Children's drawings contain a rich treasure that gives a peek into their world. Unlike the traditional medical analysis on children's drawings, this project explores them from an Art History point of view by mining the patterns in the drawings that are similar to the historically famous artwork. The digitized collection by the IMAJ - UNESCO center is a miscellany of illustrations from all over the world that vary in drawing techniques and themes. Thus, the problem of matching patterns of renowned works, equally diverse as drawings, in children's drawings translates to a problem of cross-domain image matching or retrieval.

Image similarity can mean a spectrum of things ranging from being exact copies to having similar shapes, colors, or subjects of the reference work. Children's drawings and famous artworks are comparable in terms of style, semantic content, nearness of forms, and colors. One can choose and use similarity depending on the research interest, and this work focuses on the re-creation of famous artwork by children. The drawing-artwork pair is said to be similar if the former is a reproduction of the latter, even if the techniques, colors, or domain differ. Figures 4.1 - 4.3 shows some example pairs that fall into this category of comparison. The image on the left side of the figure is the artwork similar to the drawing(s) on the right side of the image. Mining

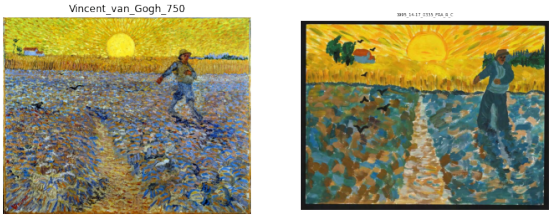
[1]The first drawing is indeed a copy of the photograph. The drawing competition accepted photos and other digital creations as well.




(a) The Starry Night




(b) Road with Cypresses

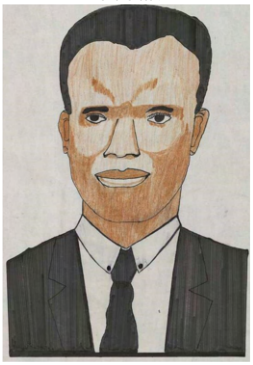



(c) The Sower


(d) Self-Portrait with Grey Felt Hat

Figure 4.1: Drawings similar to various works of Vincent van Gogh


(a) Samori Ture

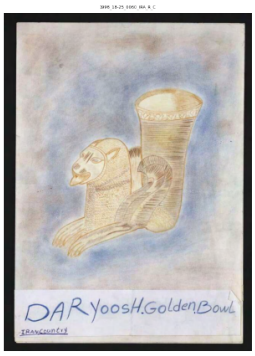



(b) James Emman Kwegyir Aggrey


(c) Marilyn Monroe

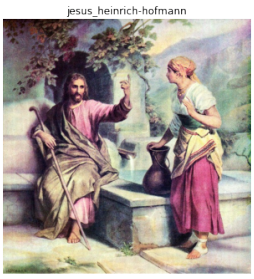

(d) Audrey Hepburn[1]

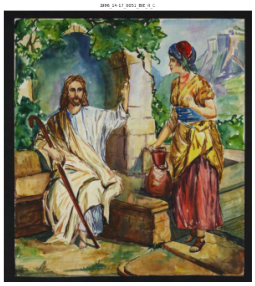
Figure 4.2: Drawings recreating portraits of famous people


(a) Achaemenid Persian Lion Rhyton


(b) Efrosinya Polotskaya

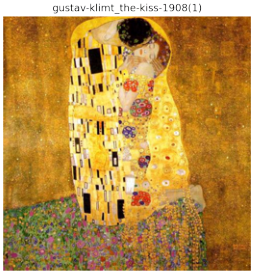


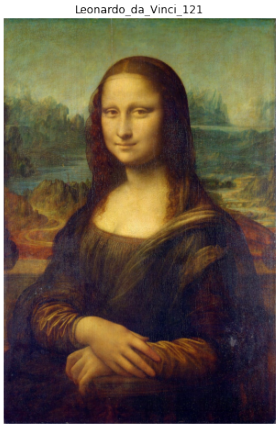
(c) Jesus and the woman of Samaria



(d) Poster of Nossa Senhora Da Oliveira




(e) The Kiss by Gustav Klimt




(f) Mona Lisa (La Gioconda or La Joconde) by Leonardo da Vinci

Figure 4.3: Other sample drawings and the corresponding artworks

these connections can unearth the process of creative creation in children. Namely, it could provide artistic impressions in youngsters and how it varies with age, time, and country.

Additionally, because of the broad-ranging age of children and different levels of expertise and purposes in creating an illustration, the drawings might not replicate the original artwork, requiring to enforce softer thresholds when matching patterns and shapes. All in all, a versatile definition of likeness distinct from the conventional visual searches characterizes similarity in this project, and the corresponding search engine should be invariant to style, medium, color, or shape variations when searching for similar artworks to the children's drawings.

## 4.2 Dataset

The drawings of children and the popular cultural objects those drawings might have recreated are the two datasets required for this project. While the IMAJ-UNESCO center provided the children's drawings, the cultural objects dataset is primarily composed of the best artworks of 50 influential artists and a few hand-picked examples that were not present in the best artworks dataset.

### 4.2.1 Famous Artworks Dataset

A dataset of reference cultural entities is equally essential as the dataset of children's drawings for this project. As mentioned in Section 1.4, these cultural objects include multiple kinds of visual arts. Nevertheless, due to the nature of the competition and the type of institutes that participate in the competition, historically famous paintings constitute the primary component of the reference set. Table 4.1 lists a few online paintings archives (presented in [16]) suitable for this project.

| Dataset | Mini-Description |
|---|---|
| **Web Gallery Of Art** [54] | Database of more than 52,000 European fine arts from the Baroque, Gothic, and Renaissance periods. |
| **Wikiart** [4] | Online visual arts database that contains around 250,000 artworks from museums, universities, and civic buildings. |
| **Rijksdata** [55] | Dataset of nearly 112,000 photographs of artworks at the Rijksmuseum. Mensik and Germet presented the dataset and baseline scores for predicting the creator, material, type, and year of artworks. |
| **Best Artworks Of All Time** [56] | Hand-curated dataset of 8355 works of the most influential artists. The dataset was scraped from the Art Challenge website [57] and published on Kaggle by a user. |

Table 4.1: Descriptions of datasets of reference artworks

Each dataset in Table 4.1 contains a great set of works in terms of quantity and variety, but the current project uses the Best Artworks of All Time (BAAT) dataset to compare the children's drawings. The reasons for choosing the BAAT dataset are

1. it contains the paintings of influential artists, increasing the probability of recreating their works, and
2. using a smaller dataset makes the solution computationally faster. Nevertheless, the solution is independent of the reference dataset.

BAAT dataset contains photographs of 8446 artworks created by the 50 most influential artists[2] from 18 nations. The genres of the works include Renaissance, Byzantine Baroque, Realism, Impressionism, Pop Art, Primitivism, Surrealism, Social Realism, Muralism, Expressionism, Cubism, and Neoplasticism. After going through the children's drawings, 131 photos of people, places, animals, and paintings that were not present before supplemented the BAAT dataset, increasing the total count to 8557. Section 4.3.3 describes the steps in obtaining these additional images, and they are similar to a few drawings in the dataset.

## 4.3 Approach

The basic idea to solve the problem of pattern matching in children's drawings (hereon drawings) and famous artworks (hereon artworks) was to use the distances between their feature vectors and perform a nearest neighbor search. A deep neural network system using a CNN to extract features of the images and a metric learning strategy to train the network implements the core solution. The decision to use a CNN stems from the previous results on their capabilities in image retrieval tasks [51] and particularly from the findings of [16], where the authors reported that CNNs perform significantly better than the classical Bag-of-Words methods that use local descriptors like SIFT for a similar type of task.

Popular CNNs architectures trained for image classification tasks are publicly available as pre-trained models/networks. Although pre-trained models learn general features in images, their performance in image retrieval tasks compared to object classification is minimal. The retrieval capability further worsens in the cross-domain search of drawings and artworks, providing scope for improvement. However, the lack of a large labeled dataset of drawing and artwork pairs, such as ImageNet for classification, makes it hard to train a deep learning model for the specific task. At the same time, the standard distance measures do not capture the class of similarity this project aims to find between the images, and it is also challenging to design a suitable metric for such a task. Metric learning helps construct this distance metric, and training a CNN model using that metric optimizes the image embeddings enabling a comparison between them.

[2]The choice of influential artists is not universal and could differ among individuals. This thesis considers the dataset as it is without delving into the debate of the correctness of the artists' selection.

Therefore, using the Transfer learning process, a pre-trained model is fine-tuned by employing a triplet learning approach.

### 4.3.1 Feature Extraction

ResNeXt CNN architecture [40] acts as the base network for feature extraction in this project. In particular, the ResNeXt-101 architecture, where 101 implies the number of layers, is used. The selection of ResNeXt-101 results from the analysis in [58], that suggests newer architectures and high-dimensional feature vectors provide better performance compared to low-dimensional feature vectors. ResNeXt is a relatively modern architecture that combines the features of ResNet and Inception Net and produces a feature vector of size 2048. The intuition of using a high-dimensional feature vector is that it could hold more information about the image.

Although it is possible to create a new CNN architecture, reusing already proven existing architectures is preferred, and it gives a possibility for Transfer Learning. For each image, a fixed-length vector of dimension 2048 is obtained by average pooling the activation output of the last convolutional layer in ResNeXt-101. The fixed-length vector is then $l^2$-normalized to create a feature vector.

### 4.3.2 Quantification of Similarity

While CNNs can produce image feature vectors, a distance function is needed to compare them. This distance (or similarity) function should accept the feature vectors as input and output the distance (or similarity) between them. A commonly used metric to compare high dimensional vectors is Cosine Similarity, as it is simple and independent of the magnitude of the values in the vector.

Cosine Similarity is the cosine of the angle between the vectors and lies in the range $[-1,+1]$. With this definition, the similarity is $+1$ for proportional vectors as the angle between them is zero, $-1$ for vectors that are $180^\circ$ (degrees) apart, and $0$ for orthogonal vectors - perpendicular to each other. If two vectors are similar, their cosine similarity is $1$, and the distance should be $0$. Thus, the cosine distance is defined as $1 - Cosine\ Similarity$. This work uses the cosine distance to estimate the nearness between the images, i.e., the *distance function* (See Section 2.4.2)

$$\begin{aligned} dist(\boldsymbol{x},\boldsymbol{y}) &= 1 - cos(\boldsymbol{x},\boldsymbol{y}) \\ &= 1 - \frac{\boldsymbol{x}\cdot\boldsymbol{y}}{||\boldsymbol{x}||\cdot||\boldsymbol{y}||} \\ &= 1 - \boldsymbol{x}\cdot\boldsymbol{y} \end{aligned}$$

where $\boldsymbol{x},\boldsymbol{y}$ are the $l^2$-normalized feature vectors of two images. Therefore, the artworks similar to the drawings should have lower distances than the unrelated artworks.

### 4.3.3 Data Annotation

A crucial step in the training process is to map the drawings and similar artworks to create an anchor and positive pairs to use in triplet learning. All the children's creations do not necessarily recreate a well-known art. Moreover, there is no pre-existing mapping between the IMAJ UNESCO center drawings and BAAT dataset, or for that matter, any reference dataset, resulting in the manual annotation of drawings and artworks pairs. After a visual inspection, the current project uses the works of only the 14-17 and 18-25 age groups, as the drawings in the very young ages are scribbles and other age categories have few references to famous artworks. There are 19,777 drawings in the combined 14-25 age group, and they were surveyed to identify a reference artwork using a three-step process.

The first step deals with the clear cases by either pairing them with familiar artworks or discarding them if there is no apparent reference. The majority of the drawings fall into the latter category. In the second step, a few ambiguous drawings are searched online using the reverse image search on Google Images[3]. The reverse image search revealed very few references. Lastly, a text-based search using the title or description or the location of the artist available on the verso side of the non-obvious drawings helped uncover a couple of references/inspirations. Examining 8800 drawings (44%) in steps 2 and 3 resulted in identifying 208 possible references between 194 (2.2% of 8800) drawings and 151 artworks. Out of the 208 possible pairs, 77 were not forthright and overreaching in some cases. Scrapping the 77 ambiguous pairs leaves with 131 anchor and positive duos between 127 drawings and 100 artworks.

### 4.3.4 Algorithm

Figure 4.4 lays out the model schema where CNN processes the images and is optimized based on the difference in distance between the images. The triplet learning method requires three inputs (Anchor (an), Positive (p), and Negative (n)), three images in this case. The anchor image is a child's drawing, and the positive and negative images are artworks similar and dissimilar to the anchor, respectively. The dependency between the mining of triplets and the distance function makes the model training an iterative process. Its summary is in Pseudocode 1.

## 4.4 Usage

Deploying the model to explore the connections between a fixed set of drawings and artworks is simple and available offline as all the feature vectors can be computed once and stored. However, an iterative loop of adding new connections, fine-tuning the model, and exploring and identifying further similar pairs is possible. First, the trained CNN is used to compute the feature

[3]https://images.google.com/

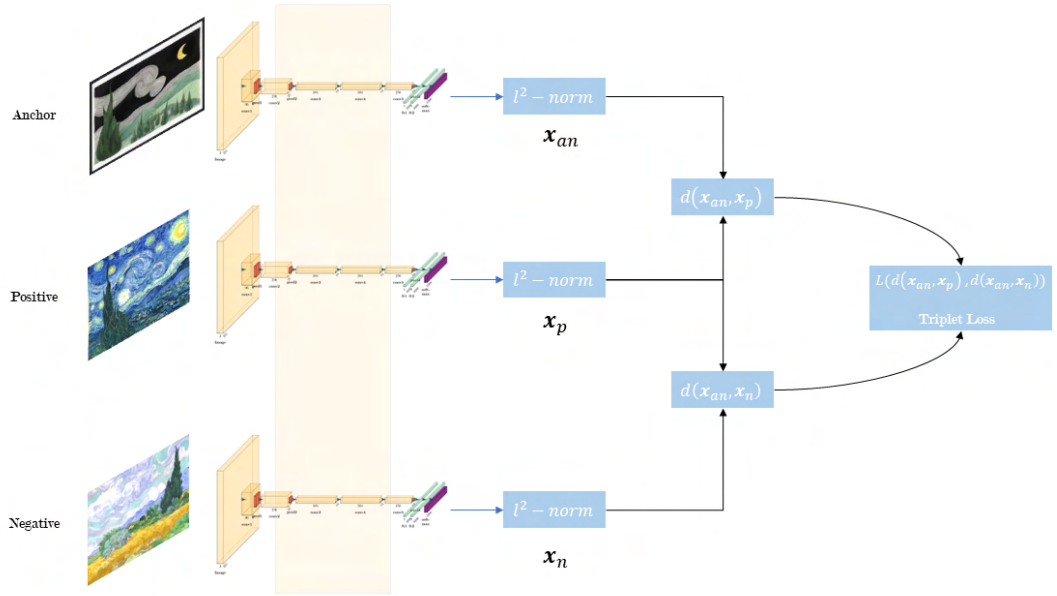



Figure 4.4: Schema for the model training using triplets. Adapted from [58]. The same network process all the images; thus, all the networks share the same parameters. Each image is processed, converted into a 2048 dimension vector, and normalized. The network parameters are updated using the triplet loss ($L$), where $d$ is the distance function.

vectors for all images in both datasets. Arranging the artworks in the ascending order of their cosine distance from the drawing produces a ranking of similarities, and the top ones should be the relevant artwork for the drawing.

## 4.5 Evaluation Metrics

The metrics used to compare and evaluate the model's performances fall into two categories. The position of the relevant artwork for a drawing in the retrieved list will assist in tracking the learning of the models. On the other hand, the recall and precision values at different positions assess the performance of the models. Due to the limited availability of data, there is a possibility of obtaining a result that fits specific drawing-artwork pairs. Thus, it is necessary to estimate the uncertainty of the model and the quality of the results. This work uses a 11-Fold Cross-Validation to assess a model, where eleven different train, validation, and test data splits evaluate the model. The reported metrics result from averaging the results of all eleven splits.

**Algorithm 1** Pseudo algorithm for the model training

1: Initialize a pre-trained CNN (ResNeXt) truncating the fully connected layers at the end to produce a feature vector to the input.
2: The CNN network symbolizes the distance function $dist$.
3: **while** $dist(\boldsymbol{x}_{an}, \boldsymbol{x}_n) + \tau < dist(\boldsymbol{x}_{an}, \boldsymbol{x}_p)$ for any $(an, p)$ pair **do**
4: For each $an$ and $p$, mine for $N_T$ number of $n$'s using the feature vectors obtained from the CNN network to create a Triplets set
$T = \{(an, p, n) : an \in D, p \in A, and\{n \in A : |\{n' \in A : dist(\boldsymbol{x}_{an}, \boldsymbol{x}_n) < dist(\boldsymbol{x}_{an}, \boldsymbol{x}_{n'}))\}| < N_T\}\}$
5: Calculate the Triplet loss for $T$ using Equation (2.1)
6: Compute an average of the loss over all $(an, p)$ pairs.
7: Backpropagate and update the CNN network using the triplet loss
8: **end while**

### 4.5.1 Mean Position

The aim of training a retrieval model is that when queried with a drawing, the relevant artwork(s) should have the least distance and appear in the first place(s). Monitoring the change in the retrieved position of artwork enables seeing how early the relevant artworks appear in the retrieval list and how it changes as the model is updated. An average of such rank for all drawings in a set gives a single metric to track the change. A low average position implies that the model rates the similar artworks closer to the query than the dissimilar ones.

If $a$ is an artwork in the set of artworks ($A$), $d$ is a drawing in the set of drawings $D$, and $A_d$ is the set of relevant artworks for drawing d. The average position of a set of drawings ($MP$) is defined as

$$MP = \frac{1}{|D|} \sum_{d \varepsilon D, d_a \varepsilon A_d} (Rank(d_a, A_{Md}))$$

where $A_{Md}$ is the list of artworks ($A$) ordered in increasing order of distance with respect to $d$ and $Rank$ is function that accepts the ordered artworks and the relevant artwork to output the position of $d_a$ in $A_{Md}$ in the range of $[0, |A| - 1]$.

### 4.5.2 Recall

Recall facilitates assessing the relevance of the results to the query. It provides the percentage of relevant artworks in the retrieved list. However, since the model ranks all the artworks based on the distance, the recall reaches $100\%$. Hence, recall at different ranks is computed by limiting the number of results taken into account to calculate it. Recall@k is the share of relevant artworks in the top-k-ranked artworks.

The recall up to a rank $k$ for a set of drawings ($Recall@k$) is the average recall up to a rank $k$

for each drawing in the set

$$Recall@k = \frac{1}{|D|} \sum_{d \varepsilon D} Recall@k(d)$$

and

$$Recall@k(d) = \frac{|\{\{A_{Mdi} | i \leq k\} \cap A_d\}|}{|A_d|}$$

### 4.5.3 Mean Average Precision

While recall gives the share of relevant artworks in the retrieved, precision provides the percentage of relevant retrieved artworks. As all the artworks are ranked using the cosine distance to the drawing, similar to the $Recall@k$, $Precision@k$ is more relevant, and the Average Precision (AP) aggregates the precisions at all positions.

The Mean Average Precision (MAP) for a set of drawings is the average of the average precision for each drawing in the set,

$$MAP = \frac{1}{|D|} \sum_{d \varepsilon D} AP_d$$

$$AP_d = \sum_{k=1}^{|A_d|} Precision@k(d) * (Recall@k(d) - Recall@k - 1(d))$$

$$Precision@k(d) = \frac{|\{\{A_{Mdi} | i \leq k\} \cap A_d\}|}{|\{A_{Mdi} | i \leq k\}|}$$

# Chapter 5

# Experiments and Results

This penultimate chapter reports the experimentations and results in training and evaluating the models. The chapter presents the experimental setup in the first section and the preprocessing steps in the second section. The metrics of baseline models are in the third section. The last section describes the fine-tuning experiments to optimize the baseline models and their assessments, including the generalization ability of the trained models.

## 5.1 Setup

The project uses Python, particularly the PyTorch library, for the implementation, as the library provides pre-trained models and wrapper functions that simplify the training of the deep learning models. Two machines having NVIDIA GeForce GPUs (GTX TITAN X and RTX 2070 with Max Q design with a memory of 12GB and 8GB, respectively) were employed to train the models[1].

### 5.1.1 Data Split

In the training of neural networks, data is split randomly, except for time series data, into the train, validation, and test sets without overlapping. The model weights are updated, and hyper parameters are adjusted using the train and validation data, respectively. The test data determines the generalization capability of the model. Ensuring an artwork appears only in one of the three sets is desired because the model learns an artwork pattern in training, and reusing it to test is not necessarily a challenge. Consequently, while splitting the data for the experimentation, a conscious effort was made to ensure that the train, validation and test datasets have different artworks.

[1] Note: The machines were operated separately to train models parallelly but not together for a single model

The data splitting process starts with creating a bipartite graph using the drawings and artworks as the two disjoint vertices and the 131 drawing-artwork pairs as the edges. The second stage extracts the weakly connected components of the bipartite graph as they will not contain any overlapping nodes. Finally, using a random assignment, 60% of these components form the train set, 25% for the validation set, and the remaining 15% for the test set.

### 5.1.2 Hyperparameters

Adaptive Moment Estimation algorithm, popularly known as Adam [59], was used in the model optimization with an exponentially decaying learning rate of $0.1$ that starts at $10^{-6}$. A weight decay coefficient regularizes the adam optimizer, and this project uses a value equal to $10^{-4}$. The training updates all layers of the pre-trained model and uses a batch size of $4$ triplets. Each model variant was trained for ten epochs with the early stopping criteria on the metrics to halt the optimization process. Lastly, the triplet loss margin ($\tau$ in Equation (2.1)) was set to $0.7$).

## 5.2 Preprocessing

### 5.2.1 Transformations

As a limited amount of data (artwork-drawing pairs) was available, it was essential to preprocess the images to improve the generalization capacity of the model and avoid overfitting the finite data. All the (pixel values of the) images were normalized using the mean and standard deviation of the images in ImageNet. During the training, the images were subject to the following set of other standard data-augmentation procedures on the fly:

- Resize to $330 \times 330$ size.
- Crop a $280 \times 280$ random part of the image. Thus the size of the image processed by the networks is $280$ on each side.
- Flip the image horizontally with a $20\%$ probability.
- Rotate the image randomly in the range $[-5, +5]$ degrees.
- Randomly jitter the image's brightness, contrast, and saturation by a factor in the range $[0.9, 1.1]$.
- Lastly, convert it into a grayscale image with a $40\%$ chance.

During the validation and testing, only resize the images to $280 \times 280$ without any other transformation. Due to a constraint on the GPU memory, an optimal batch size of $4$ was possible by scaling the images to $280$.

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

Figure 5.1: Examples of drawings (on left) after apply oil painting effect (on right)

Figure 5.2: Examples of drawings (on left) after apply watercolor effect (on right)

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

Figure 5.3: Examples of drawings (on left) after apply textured effect (on right)

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

Figure 5.4: Examples of drawings (on left) after apply pencil sketch effect (on right)

### 5.2.2 Style Augmentation

As discussed in earlier chapters, the children do not necessarily recreate the famous artworks using the same material and technique. Therefore, augmenting the drawings in the training set helps to consider the diversity to some extent. This augmentation step was performed offline in addition to the preprocessing or transformations mentioned above. The training drawings were expanded with five styles using the effects available in the OpenCV library, increasing the training dataset $6$ times.

#### Oil Painting

An oil painting effect was applied using the OpenCV's *oilPainting* function in the *xphoto* module by setting *size* and *dynRatio* parameters to $7$ and $1$, respectively. Figure 5.1 shows the result of converting some drawings.

#### Water Color

OpenCV's stylization function applies the watercolor effect to the image with *sigma_s* set to $60$ and *sigma_r* as $0.6$. *sigma_s* controls the size of the neighborhood, and *sigma_r* controls averaging of colors in the chosen neighborhood. The drawings with the watercolor effect are shown in Figure 5.2.

#### Textured

The drawings were converted into textured sketches (Figure 5.3) using the *detailEnhance* function in OpenCV with *sigma_s* as $60$ and *sigma_r* as $0.4$.

#### Grayscale

The grayscaled versions of the drawings obtained through OpenCV's standard *cvtColor* function were also part of the training.

#### Pencil Sketch

A pencil sketch effect was applied to the drawings using a set of custom operations:

1. Invert the grayscaled version of the image to reverse its colors.
2. Apply a gaussian blur to the inverted image to reduce the noise and detail.
3. Blend the grayscale and blurred images by dividing the former with the latter.

The examples of drawings after applying this effect are shown in Figure 5.4.

## 5.3 Pre-trained Models - Baseline solution

| Model | Baseline (ResNeXT-101) |
|---|---|
| *MP* | 1039.94 ± 178.51 |
| *MAP* | 12.5 ± 4.48 |
| *R@400* | 40.35 ± 3.63 |
| *R@200* | 32.98 ± 7.06 |
| *R@100* | 26.07 ± 7.51 |
| *R@50* | 23.53 ± 7.41 |
| *R@20* | 17.45 ± 6.1 |

Table 5.1: Metrics on the pre-trained ResNeXt-101 model

The CNN models trained on image classification act as baseline models, and Table 5.1 shows the results of drawing-artwork pattern matching using a pre-trained ResNeXt-101 architecture. For nearly 60% of drawings, the corresponding artwork does not appear in the top 400 results, and only for less than 20% of drawings, the match is found in the first 20 ranked artworks. The mean appearance position of 1039 and the 12% MAP suggests that the model has trouble finding the matching artwork and ranking it at the top. These results indicate room for improvement.

## 5.4 Fine Tuning Experiments

Training a Neural network requires adjusting various parameters and conditions. This section presents the results of experimentation in fine-tuning the model using style augmented images in the training dataset, followed by experiments using CNNs other than pre-trained models as the starting point.

### 5.4.1 Style Augmentation

This experiment determines whether the presence of style-transferred drawings in the training data improves the model performance in identifying the matching artworks. Figure 5.5 compares

the distribution of the Mean Position (MP) of the retrieved artworks, and Table 5.2 provides the MAP and Recall measures. Baseline refers to the pre-trained ResNeXt-101 model, whereas FT Aug and FT No-Aug are the fine-tuned ResNeXt-101 models with and without using the style augmented drawings. The MPs in the fine-tuned models are always lower compared to the baseline. On average, the artworks appear 620 and 655 spots earlier in the FT No-Aug and FT Aug models.

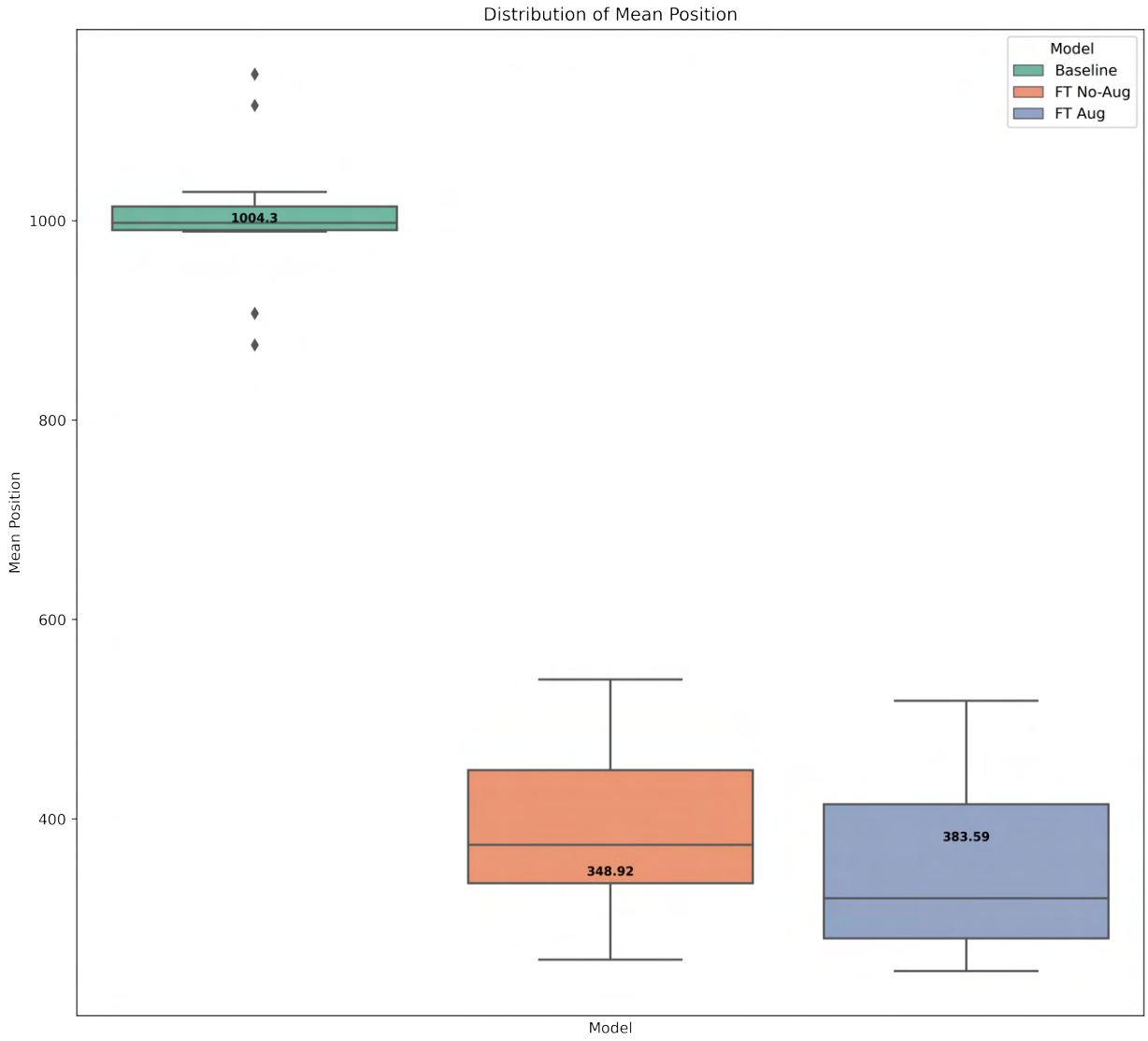



Figure 5.5: Mean Position comparison fine-tuning models using style augmented drawings

The FT No-Aug model improves recall by *39.34*, *39.31*, *38.35*, *30.79*, and *27.54* percentage points from R@400 to R@50 compared to the baseline. Similarly, the recall values in FT Aug improved by *46.83*, *45.97*, *41.31*, *33.58*, and *32.98* percentage points. The model trained using the augmented drawings improves the retrieval performance significantly. While the FT No-Aug model also improves from baseline. However, analyzing the retrievals of FT No-Aug shows that

| Model | Baseline | FT No-Aug | FT Aug |
|---|---|---|---|
| *MAP* | 12.5 ± 4.48 | 26.82 ± 7.24 | 31.9 ± 10.03 |
| *R@400* | 40.35 ± 3.63 | 79.69 ± 8.11 | 87.18 ± 5.28 |
| *R@200* | 32.98 ± 7.06 | 72.29 ± 5.87 | 78.95 ± 7.26 |
| *R@100* | 26.07 ± 7.51 | 64.42 ± 8.44 | 67.38 ± 6.38 |
| *R@50* | 23.53 ± 7.41 | 54.32 ± 8.26 | 57.11 ± 11.26 |
| *R@20* | 17.45 ± 6.1 | 44.99 ± 9.02 | 50.43 ± 11.7 |

Table 5.2: Evaluation metrics of fine-tuning models using style augmented drawings

the artworks in the training set populate the top ranks irrespective of the drawing, hinting at overfitting.

Therefore, fine-tuning the model with style-augmented drawings provides better performance without overfitting the training data than the model fine-tuned otherwise.

### 5.4.2 Models trained for detection of pattern propagation

Transfer learning methods in deep computer vision usually refer to using a CNN trained for image classification data using ImageNet data for other tasks. However, there is no restriction on using models other than those trained on ImageNet. This subsection presents the results of evaluating the drawing-artwork retrieval problem with a model developed to find duplicate photographs of artworks.

Digitization of photo collections opens a window to perform a wide range of operations to process and extract information. In a photo collection of artworks, it is possible to have multiple photos of the same subject (painting, sculpture, carvings) either captured by the same person or a different one. Conversely, two paintings can have the same global composition with different local elements or vice versa, or they can be thematically related by depicting the same scene differently. Seguin [58] developed a system (known as *Replica*) to identify images sharing identical patterns in photos of artworks (Section 2.5.3). Ludovica [60] expands this work further to cluster photographs of paintings sharing patterns where duplication or re-creation of a visual object with an inspiration amounts to pattern sharing irrespective of the factors causing it.

Two CNN models using ResNeXt-101 architecture made available by the extension form part of the experimentation in feature vector generation for the current drawing-artwork retrieval problem. The first model modernizes the Replica system through the latest CNN architecture to rank artwork photos based on their visual similarity. It is referred to as *Mini-Replica* as it trains on a subset of data used in the original Replica project using triplet loss with cosine distance. The second model advances the Replica by clustering photos that share patterns. The training of the second model (*Clus-Replica*) uses a modified triplet loss that includes the distance between the positive and negative sample of the triplet and an additional constraint to minimize the distance between the anchor and positive sample. This modified loss ensures that the distance

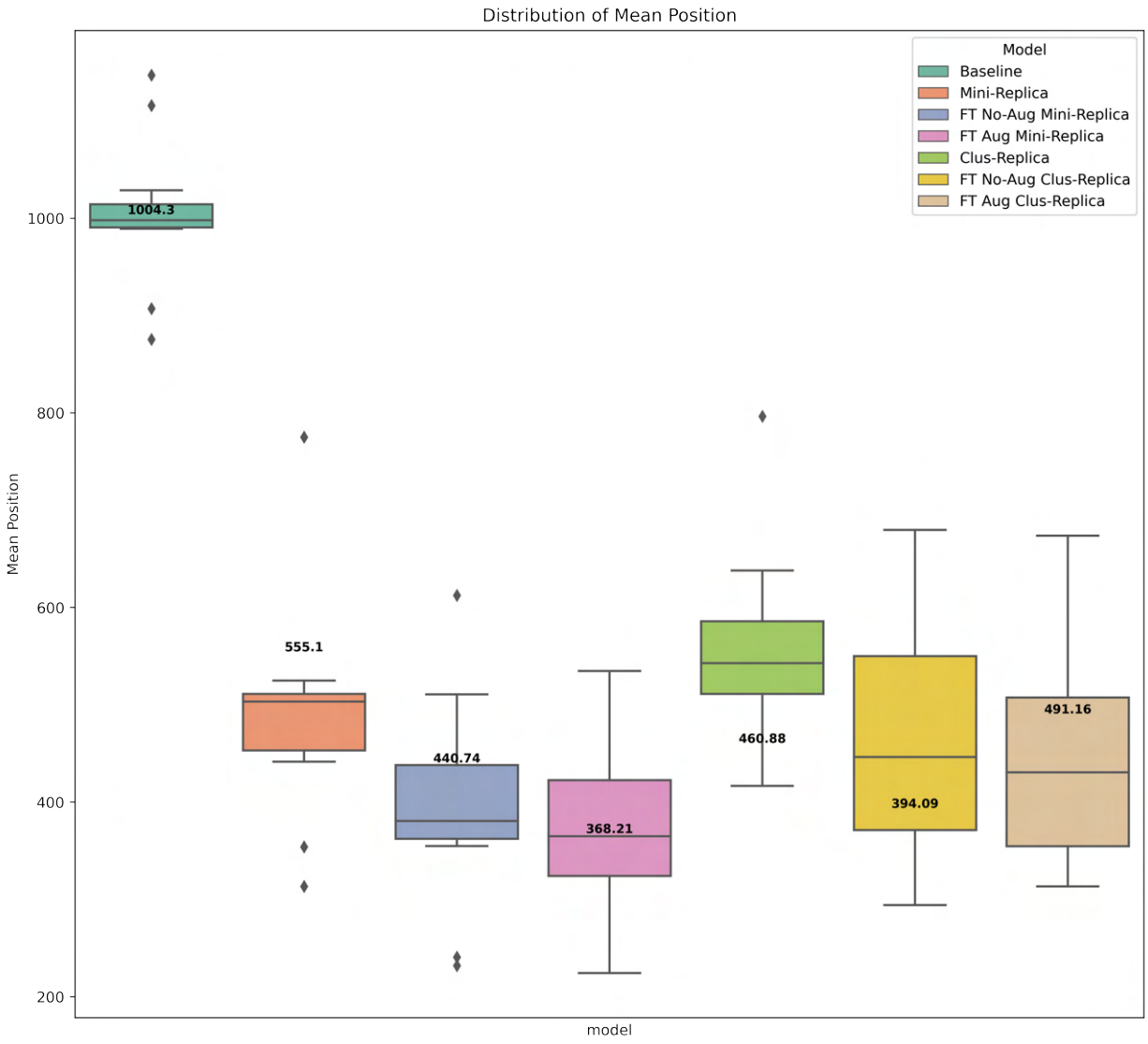



Figure 5.6: Mean Position comparison of Evaluation metrics of Replica variants

| Model | MAP | R@400 | R@200 | R@100 | R@50 | R@20 |
|---|---|---|---|---|---|---|
| *Baseline* | 12.5 ± 4.48 | 40.35 ± 3.63 | 32.98 ± 7.06 | 26.07 ± 7.51 | 23.53 ± 7.41 | 17.45 ± 6.1 |
| *Mini-Replica* | 28.73 ± 11.77 | 70.99 ± 5.94 | 60.4 ± 9.88 | 51.47 ± 10.25 | 45.1 ± 13.32 | 39.34 ± 12.55 |
| *FT No-Aug Mini-Replica* | 33.26 ± 11.05 | 84.89 ± 5.02 | 76.46 ± 7.44 | 67.93 ± 9.0 | 62.23 ± 9.71 | 53.87 ± 10.45 |
| *FT Aug Mini-Replica* | **35.26 ± 11.83** | **86.51 ± 5.9** | **80.62 ± 8.1** | **74.11 ± 7.59** | **65.48 ± 8.24** | **55.08 ± 6.94** |
| *Clus-Replica* | 27.35 ± 9.6 | 68.86 ± 6.24 | 61.6 ± 7.49 | 56.22 ± 7.23 | 50.81 ± 8.6 | 41.11 ± 11.4 |
| *FT No-Aug Clus-Replica* | 31.42 ± 6.14 | 77.82 ± 5.91 | 71.17 ± 7.4 | 66.14 ± 7.94 | 58.91 ± 10.57 | 53.85 ± 6.5 |
| *FT Aug Clus-Replica* | 30.63 ± 6.69 | 76.76 ± 4.71 | 70.75 ± 5.69 | 65.46 ± 5.46 | 58.55 ± 9.46 | 49.9 ± 10.36 |

Table 5.3: Evaluation metrics of Replica variants

between the anchor and positive samples is lower than between the anchor and negative, and the distance between positive and the anchor is lower than between the positive and negative samples. Additionally, it limits the distance between the anchor and positive sample to a maximal value. Apart from the loss function, the training data of Mini-Replica includes a mix of color and grayscale images, while Clus-Replica only uses grayscale images (8,900 images with 4,900 connections, i.e., annotated pairs sharing a visual pattern).

The experiments involving Mini-Replica and Clus-Replica are of two kinds:

- First, retrieve and compare the artworks for the drawings using the feature vectors generated by the models trained on the Replica dataset.
- Second, fine-tune the models with the drawing-artwork pairs. As in Section 5.4.1, the experimentation involved the effect of style-augmented drawings in the dataset.

Table 5.3 and Figure 5.6 show and visualize the evaluation metrics using the Replica variants compared with the baseline model (other nomenclature remains the same as before). Even before fine-tuning, the Mini-Replica and Clus-Replica models have lower mean positions and better recall than the pre-trained ResNeXt-101 model without fine-tuning. Furthermore, fine-tuning the model for the specific task boosts performance. The effect of augmented data is the same as earlier - training the model using the drawings alone overfits the train data. However, for the Clus-Replica model, training with the stylized drawings drops the model's ability to rank better.

### 5.4.3 Models Comparision

A deep learning model trains on examples, and in the cross-validation process, the model processes the samples in one of the two ways that impact the model. The training samples directly influence the step of updating the model parameters and those in the validation set to aid in adjusting the hyperparameters. Thus the performance measures of the model on the test set are central in providing an unbiased evaluation of the model. Consequently, the models discussed in Sections 5.4.1 and 5.4.2 are evaluated on the test sets to obtain an impartial assessment. Table 5.4 presents the MAP and Recalls of the models averaged on 11 different data splits during the cross-validation, while Figure 5.7 shows the distribution of the mean positions.

Fine-tuning the models significantly improves the ranking of the artworks similar to the drawings. While the ResNeXt-101 model fine-tuned only using the artwork drawing pairs reaches the lowest mean position, Mini-Replica achieves the highest recall and precision. The style augmentation of the drawings in training shows a notable effect on the results, and the generalization ability of such models is better than those trained otherwise. However, this effect is not apparent in the Clus-Replica model, and Chapter 6 discusses it. Finally, these results on the test

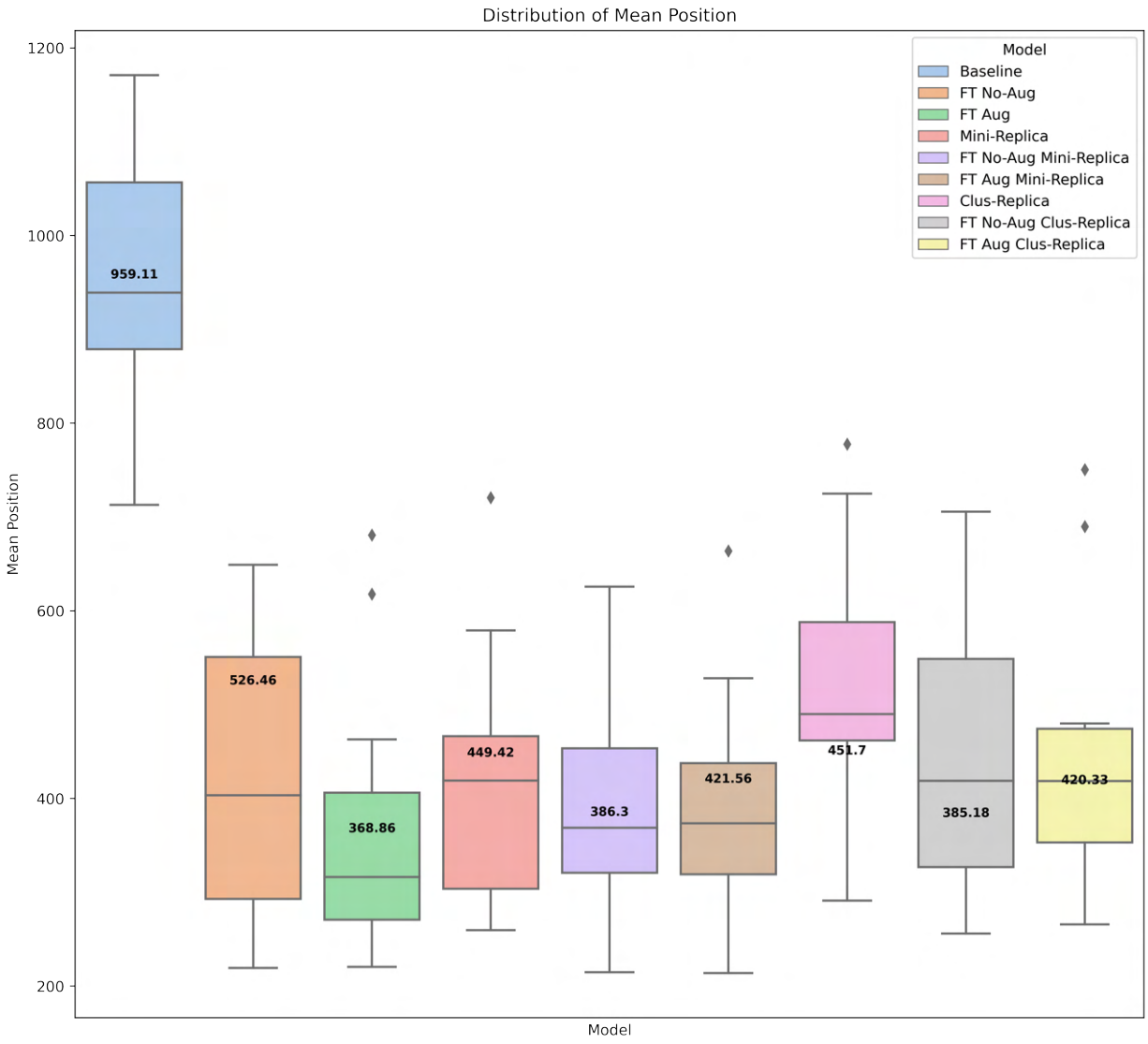



Figure 5.7: Mean Position comparison on Test data

set establish that the models do not fit specific examples but learn low and high-level features that help construct a connection between the drawing and artwork.

| Model | MAP | R@400 | R@200 | R@100 | R@50 | R@20 |
|---|---|---|---|---|---|---|
| *Baseline* | 17.51 ± 6.39 | 47.29 ± 7.39 | 39.79 ± 7.26 | 35.08 ± 6.78 | 33.78 ± 7.35 | 27.13 ± 8.65 |
| *FT No-Aug* | 32.58 ± 8.86 | 82.46 ± 10.39 | 74.6 ± 10.77 | 66.9 ± 13.94 | 58.61 ± 14.83 | 51.99 ± 15.53 |
| *FT Aug* | 37.09 ± 12.96 | 85.12 ± 8.61 | 76.65 ± 7.41 | 68.33 ± 12.62 | 61.38 ± 16.25 | 53.89 ± 16.25 |
| *Mini-Replica* | 32.67 ± 10.78 | 75.83 ± 9.27 | 69.43 ± 13.49 | 61.4 ± 13.76 | 58.16 ± 16.5 | 50.04 ± 15.93 |
| *FT No-Aug Mini-Replica* | 33.26 ± 11.05 | 84.89 ± 5.02 | 76.46 ± 7.44 | 67.93 ± 9.0 | 62.23 ± 9.71 | 53.87 ± 10.45 |
| *FT Aug Mini-Replica* | 37.42 ± 9.97 | **88.13 ± 7.05** | **80.08 ± 9.39** | **73.22 ± 10.84** | 64.42 ± 10.14 | 54.74 ± 8.3 |
| *Clus-Replica* | 33.09 ± 8.47 | 76.37 ± 8.65 | 67.11 ± 5.78 | 63.04 ± 6.7 | 58.1 ± 8.99 | 49.34 ± 10.43 |
| *FT No-Aug Clus-Replica* | **37.84 ± 9.83** | 79.03 ± 7.79 | 72.64 ± 7.85 | 69.05 ± 8.74 | **64.93 ± 10.33** | **55.42 ± 11.93** |
| *FT Aug Clus-Replica* | 34.79 ± 9.02 | 79.65 ± 6.7 | 74.02 ± 5.88 | 68.86 ± 7.96 | 64.46 ± 10.2 | 54.18 ± 11.55 |

Table 5.4: Evaluation metrics on Test data

# Chapter 6

# Discussion

This chapter examines the findings and critiques various steps of this thesis. The first two sections compare the models' performances both quantitatively and qualitatively. The third section elucidates the limitations and possible improvements and proposes potential directions for research on children's drawings.

## 6.1 Quantitative Performance Analysis

FT Aug Mini Replica, the Mini-Replica fine-tuned using the examples of drawing-artwork pairs, is the best model for retrieving artworks similar to a drawing. The model finds the relevant artwork in the first 400 results with a 88% probability. Although the likelihood of finding the relevant artwork falls to 54% in the first 20 results, it is still nearly two times more probable than using a CNN model trained for image classification. FT Aug, the model fine-tuned on the pre-trained ResNeXt-101, is the second-best model. FT Aug's performance is similar to that of Mini-Replica - the fine-tuned CNN to retrieve duplicate artwork photographs and fine-tuning Mini-Replica enhances the performance. FT No-Aug Clus-Replica does not retrieve visually relevant results, albeit showing the third-best performance.

The examples of drawings shown in this thesis portray their variousness; some drawings are pencil sketches of artworks that are photographs, some drawings used colored pencils to re-create the paintings, or in some cases, the artworks were in grayscale, and the drawing was in color. These differences make it challenging to create correspondences between the drawing and artwork only with one example. At the same time, there could be an unidentified drawing in the current dataset or a new drawing in the future that uses different techniques and materials than the artwork. Experimentation shows that the style augmentation step made it possible to overcome this hurdle.

The mean average precision comprises precision and recall. It becomes a vital measure indicating how well the model detects the correct artwork for a drawing, and it increases from close to 17% to little more than 37% in the top three models.

Unlike the MAP and Recall, the trend in the mean positions of the artworks is different. The models that do not use style augmented drawings for training have relatively lower mean positions than those that use them. The mean position averages the rank of the relevant artwork across all possible artworks, and it is skewed if there is more than one possible match, and one appears early while the other is ranked low. This contradiction between the mean position and other measures results from overfitting the models with limited examples (No-Aug cases). In some examples of drawings, it is possible to have more than one similar artwork. Table 6.1 contains a new Minimum Mean Position (MMP) measure that computes the mean position using one lowest rank per drawing on the test dataset (Excluding the No-Aug models). The best-performing models remain the same with a slight change in their order. FT Aug has the lowest MMP, followed by FT Aug Mini-Replica and FT Aug Clus-Replica. Earlier measures suggest that the Mini-Replica achieves a performance similar to that of a fine-tuned pre-trained model (FT Aug) without task-specific fine-tuning. Nonetheless, the MMP depicts that the FT Aug ranks the artworks better than the Mini-Replica except for a few cases discussed in the next section.

| Model | MMP |
|---|---|
| *Baseline* | 1491.63 ± 360.31 |
| *FT Aug* | **231.48 ± 134.63** |
| *Mini-Replica* | 416.90 ± 263.74 |
| *FT Aug Mini-Replica* | 285.21 ± 160.47 |
| *Clus-Replica* | 536.80 ± 214.68 |
| *FT Aug Clus-Replica* | 429.87 ± 214.67 |

Table 6.1: Minumum Mean position of drawings in Test data

All these experiments use a minimal set of examples and achieve satisfactory performances. Nevertheless, the lower standard deviation in the Replica models is a striking difference between the FT Aug and the Replica variants. The large datasets used in training could explain the low variation in the metrics of the Mini-Replica, Clus-Replica, and their fine-tuned versions. The performance of Replica models is better than the pre-trained model because the knowledge of identifying duplicate paintings is close and related problem to the drawing-artwork problem.

## 6.2 Qualitative Performance Analysis

Evaluation metrics convey a model's performance and do not capture the results' quality. This section discusses the ranking grade by comparing six models: Baseline, FT Aug, Mini-Replica, FT Aug Mini-Replica, Clus-Replica, and FT Aug Clus-Replica. The best variant of each model during the 11-fold validation was used to create the feature vectors and compare the images.

Each comparison image contains three parts where the drawing is at the top left, and expected similar artwork is at the bottom left of the image. Lastly, the top 10 retrievals using the six models mentioned above are on the right side of the image, along with the rank of the expected artwork. The rank is helpful, especially if the expected artwork does not appear in the top 10 results.

### 6.2.1 Closely Imitated Drawings

Many computer vision problems use CNN architectures developed for ImageNet image classification as a starting point as the layers have already learned to detect features such as the color and shape of objects. Although the overall performance of these models is poor in drawing-artwork retrieval, they identify the relevant artwork when the drawing is close to the compared artwork. Figure 6.1 shows such examples; in each of them, the drawing replicates the artwork very closely. Even when there is a slight color change, the baseline model ranks the appropriate artwork at the top.

### 6.2.2 Domain and Technique Constrainted Retrieval

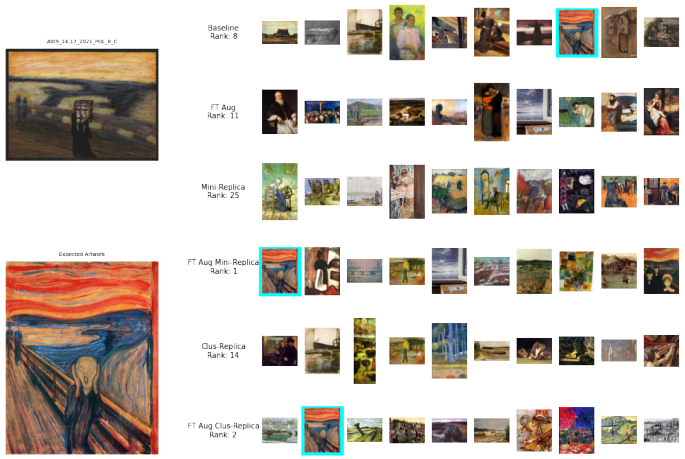
Another problem with the baseline model is its inability to retrieve artworks produced using a different technique than the drawing. Figure 6.2 shows some examples. The Replica variants retrieve these examples even before fine-tuning, and fine-tuning ranks them better. In the case of Figures 6.2a and 6.2c, the baseline model only retrieves artworks made using pencil sketches, while the relevant artwork is not a pencil sketch. The consequence of fine-tuning is visible in the results, making the artworks appear in the top positions with FT Aug and FT Aug Mini-Replica, overcoming the difficulty of recognizing the artwork produced using a different technique.

The artwork's material distinguishes it from a sketch, photograph, computer-generated image, or painting. Figure 6.3 compares the retrieval of artworks in examples where the method and material of the drawings are in contrast to the artwork. Similar to the previous discussion, the Mini-Replica and FT Aug models can overcome the domain limitations.

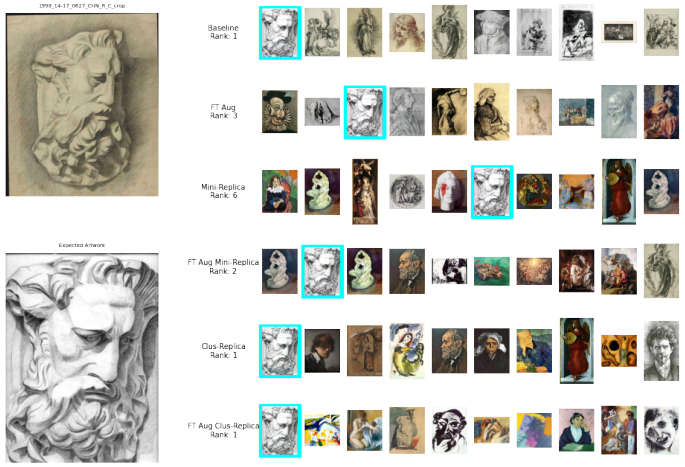
All in all, the Replica variants do not suffer from the difference in techniques between the artwork and drawings. However, their performance is slightly affected when the domain is different. Nevertheless, fine-tuning the baseline or the Replica variants helps confound it.

### 6.2.3 Differences between the Replica variants

Qualitative analysis of the results clearly shows that the results of the Clus-Replica are inferior to that of the Mini-Replica and FT Aug in many cases. Notwithstanding the evaluation metrics, the retrievals of Clus-Replica and its fine-tune are not up to the mark. Inspection of the results




(a) Drawing: `2009_14-17_2021_POL_R_C` part of training set

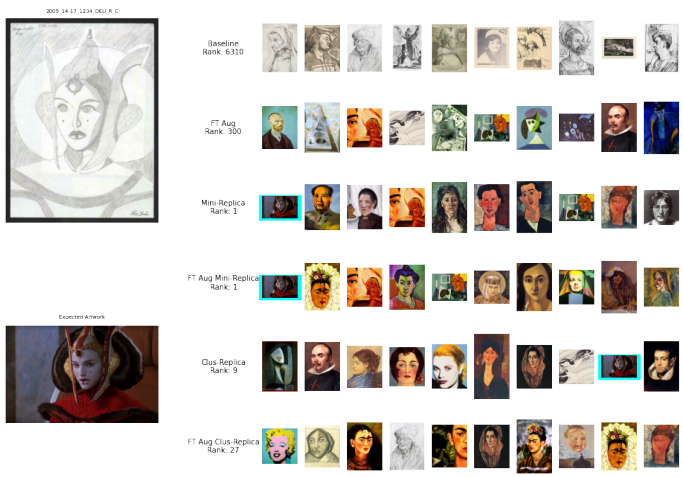



(b) Drawing: `1998_14-17_0627_CHN_R_C` part of training set


(c) Drawing: `1997_14-17_0014_LBA_R_C` part of test set

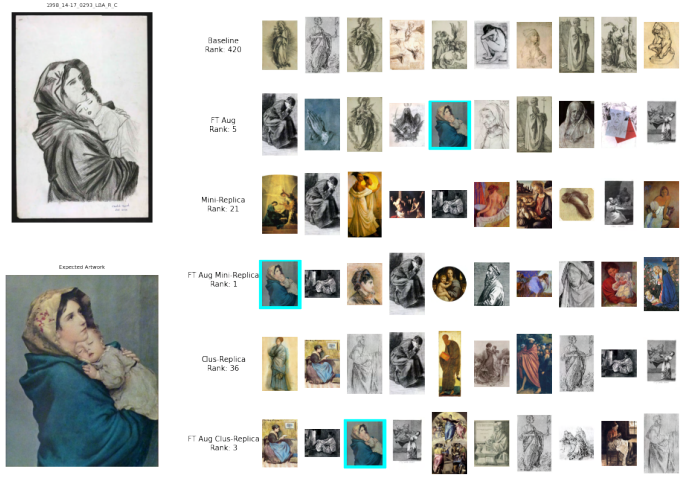

(d) Drawing: `2000_14-17_1345_KIR_R_C` part of validation set

Figure 6.1: Top 10 retrievals of drawings close to original artwork

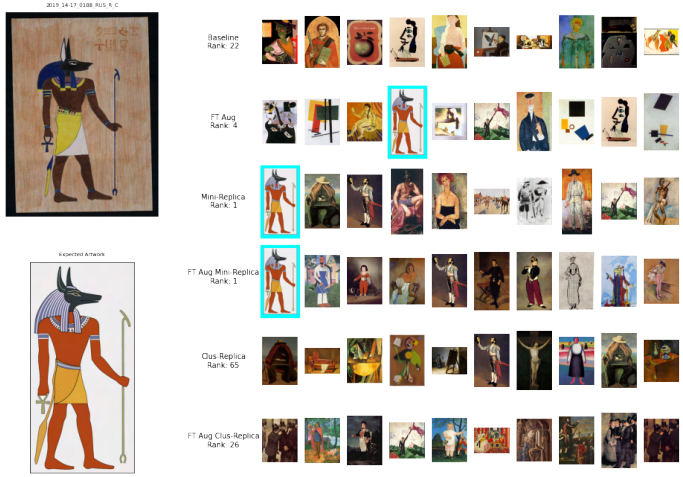



(a) Drawing: `2009_14-17_1234_DEU_R_C` part of training set


(b) Drawing: `2000_14-17_1583_COL_R_C` part of training set




(c) Drawing: `1998_14-17_0293_LBA_R_C` part of training set

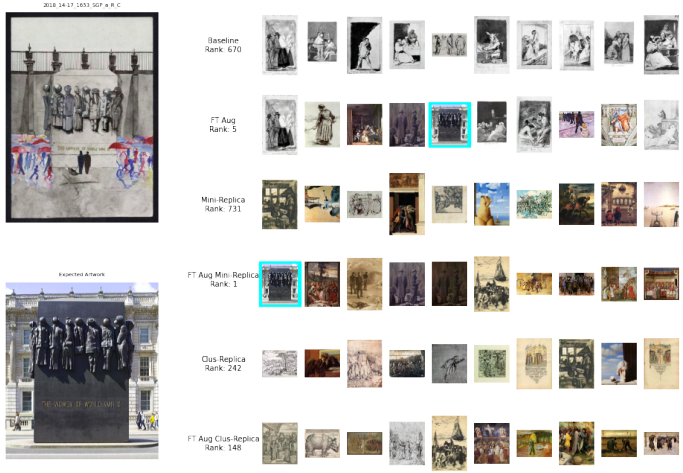



(d) Drawing: `2019_14-17_0188_RUS_R_C` part of validation set

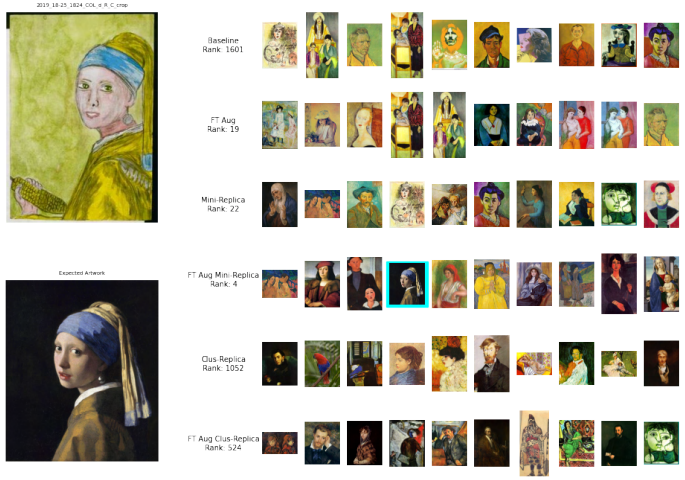
Figure 6.2: Top 10 retrievals of drawings using a different drawing technique than the original artwork


(a) Drawing: `2020_18-25_4587_CHN_R_C` part of training set


(b) Drawing: `2020_14-17_2051_UKR_R_C` part of training set




(c) Drawing: `2018_14-17_1653_SGP_a_R_C` part of training set




(d) Drawing: `2019_18-25_1824_COL_d_R_C` part of validation set

Figure 6.3: Top 10 retrievals of drawings in a different domain than the original artwork

reveals that FT Aug and Mini-Replica models learn the object localization and the semantic meaning of the drawing while Clus-Replica focuses on global average information (shape, color). The retrieval examples in Figure 6.4 shows these differences.

Although Mini-Replica and the Clus-Replica models start with the same base, train using metric learning, and have heavily overlapping training data, their objectives differ. The former model aims to find photos sharing a pattern, while the latter tries to cluster all such artwork photographs. The loss function in metric learning translates these differences in the objectives to training. The Mini-Replica model, whose training is similar to the drawing-artwork problem, moves the artwork similar to a drawing close to it than the dissimilar ones. In addition to Mini-Replica, Clus-Replica also ensures that images in a cluster are within a specified limit. Thus, optimizing one cluster of works impacts the other clusters and increases the probability of outlier images overflowing into other clusters and producing semantically undesirable results in this project.

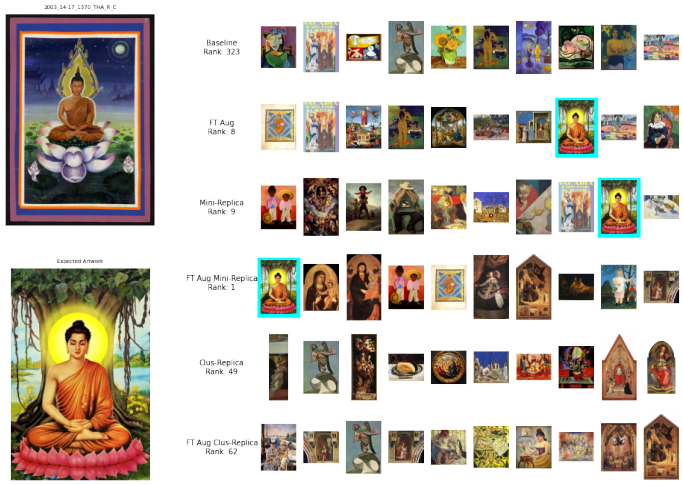
### 6.2.4 Failure modes

The examples in the previous sections show drawings whose artwork appears in the top results. There are three categories of failures, and Figure 6.5 showcases them. The first mistake is when a single element dominates the drawing. In Figures 6.5a and 6.5c, the retrieved artworks have predominantly white and black backgrounds, owing to the color of the paper and cardboard used to make those drawings. Figure 6.5b shows an example of the second category of mistakes where fine-tuned models push the artwork backward in the ranks. All these examples are in validation or test set, and they did not directly affect the parameters of the models. The last category is those with weak visual similarities. Although the relevant artwork does not appear in the first ten results in Figure 6.5d, it moves closer to artworks as reflected in the ranks.

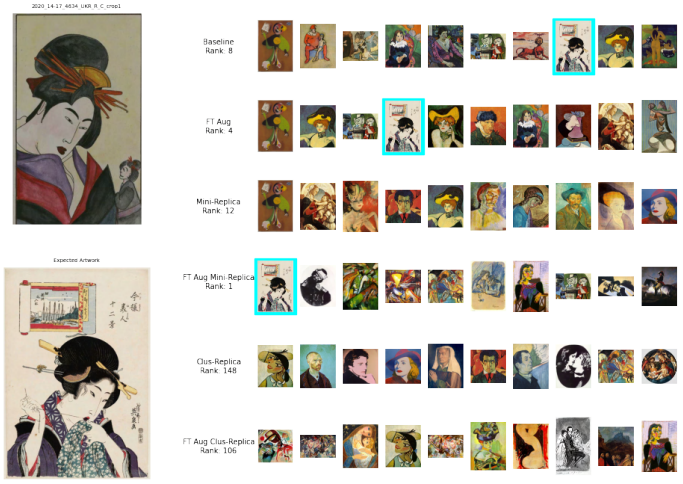
## 6.3 Limitations and Future Work

This project suffers from three drawbacks. Firstly, the choice of the artworks dataset. The artworks and the artists primarily fall in the Europe region without sufficient representation of works from the rest of the world, especially from the Eastern nations where a sizable number of drawings originate. While using the artworks from the digitized collections of museums of WikiArt would partially solve this problem, it increases the computational complexity involved in the online mining of the triplets. Additionally, many drawings refer to other pop culture objects, which are not necessarily paintings or drawings by famous artists. Supplementing the paintings dataset with popular cultural objects such as posters, shots in movies, portraits of celebrities, and others would unearth more drawing-artwork pairs.

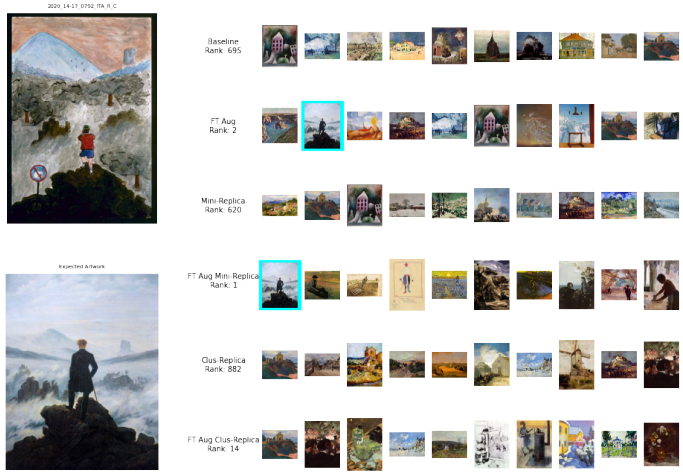
Secondly, it was only possible to annotate a hundred examples of valid drawing-artworks



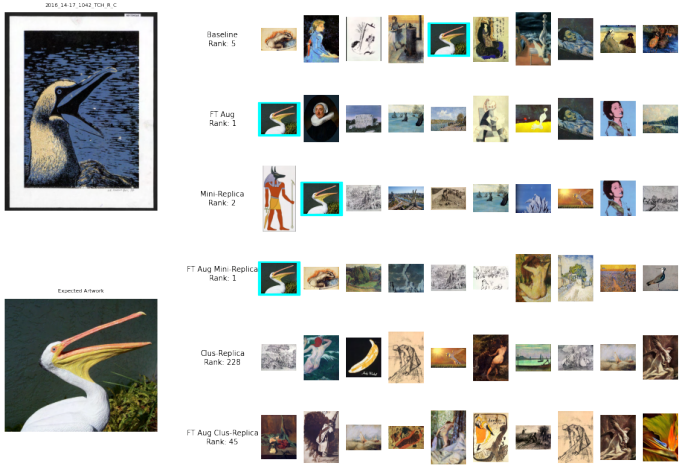

(a) Drawing: `2003_14-17_1370_THA_R_C` part of training set




(b) Drawing: `2020_14-17_4634_UKR_R_C` part of training set

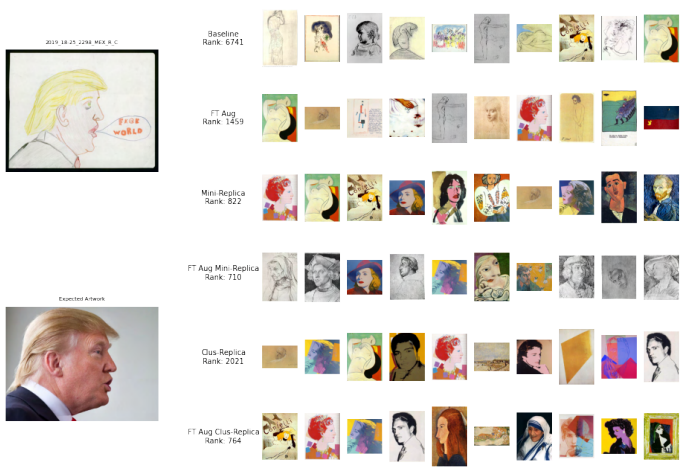



(c) Drawing: `2020_14-17_0792_ITA_R_C` part of training set

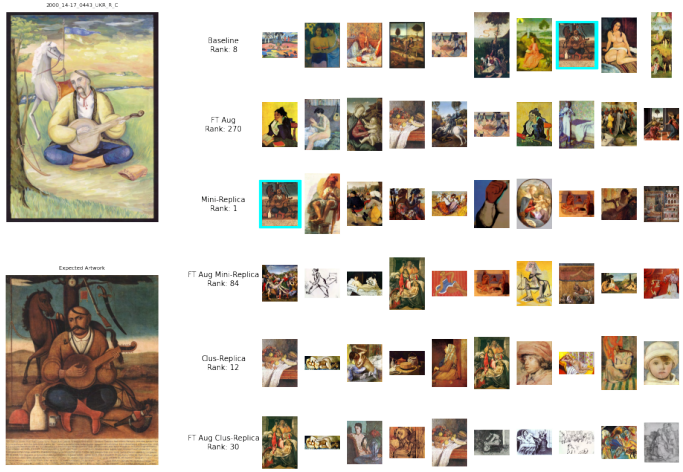



(d) Drawing: `2016_14-17_1042_TCH_R_C` part of test set

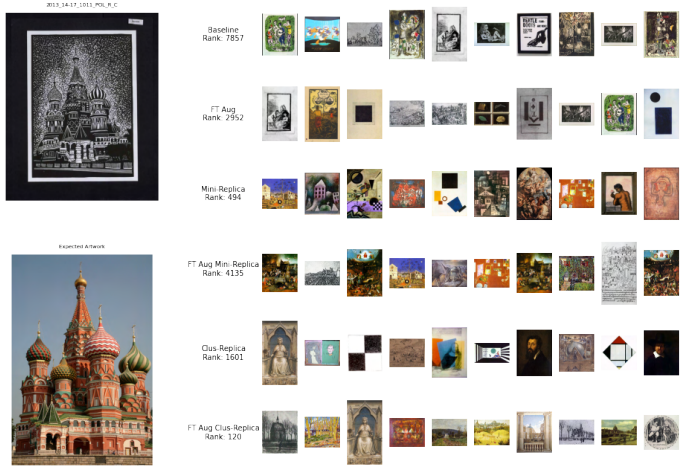
Figure 6.4: Examples depicting Clus-Replica's poor quality retrievals




(a) Drawing: `2019_18-25_2298_MEX_R_C` part of validation set




(b) Drawing: `2000_14-17_0443_UKR_R_C` part of validation set




(c) Drawing: `2013_14-17_1011_POL_R_C` part of test set


(d) Drawing: `2010_MS_0981_FRA_R_C` part of test set

Figure 6.5: Breakdown examples

during the project term. Increasing the number of examples does not constantly improve the models' performance. Nonetheless, having more annotated data helps to understand whether the relatively high variation in the metrics is inherent to the data or is only due to the low number of samples. The models developed during the project comes in handy in this aspect as they can aid in identifying new drawing-artwork pairs, subject to the availability of the relevant artwork in the comparison dataset.

Experimentation on the image's resolution was not part of the project, as the choice of 280 resulted from the trade-off between memory and computation time. Also, a comparison with sketch retrieval models is missing, particularly in using existing Sketch-Based Image Retrieval (SBIR) models as the starting point and fine-tuning them for the specific task. Thus, a future investigation of the image size and comparison with SBIR models would provide more cognizance.

Future work could expand the drawings dataset to include the drawing of 6-13-year-olds to the current 14-25-year-olds. In terms of the model architecture, an investigation to use more recent computer vision deep learning models is possible, significantly as the number of annotated examples increases. Another interesting exploration avenue is creating a retrieval system based on the similarity in the style or semantic meaning.

Lastly, the current system helps identify the drawings similar to other drawings in the drawings dataset. Reading the identical drawings with their metadata and temporality shall characterize the propagation of ideas across years. The previous suggestions are mainly technical directions, but this research also has a social aspect. The artwork ranking model, along with metadata about the young artists' country and that of the relevant artwork, provides insights into the cross-cultural influences of the famous artists and their work. In addition, Section 3.3 lists a set of possible explorations on the drawings dataset itself.

# Chapter 7

# Conclusion

The main objectives of this work were to explore the children's drawings dataset to provide preliminary insights and to develop a system that retrieves a famous work (painting, sculpture, poster, portrait, or photo of a landmark/building) that is similar to a child's drawing.

Drawings datasets are usually associated with sketch datasets [61–64], which are useful for tasks of image retrieval, semantic clustering using sketches, generating sketches using deep neural networks or in education, investigating psychology or crime. However, these datasets suffer some drawbacks. First, most of the datasets contain only grayscale pencil sketches. Second, the artist's age is available only in one or two datasets. Due to these deficiencies, along with not being child-centric and focusing on a singular object, they lack the diversity contained in the digitized IMAJ-UNESCO children's drawings dataset. The children's drawings dataset has drawings created by artists aged between 3 and 25 from 1994 till today based on a distinctive theme each year, making it a spatially, temporally, and conceptually diverse dataset.

The drawings have deeply rooted cultural references that vary geographically. Although children sometimes use a different drawing technique or reflect on global issues, their drawings contain elements of local culture. Few drawings also indicate that children do not hesitate to discuss uncomfortable or sensitive topics in their drawings. Many drawings have their references implanted in modern-day cultural objects. An important observation that stays valid across all the drawings in the datasets is that children tend to reuse a style/technique of a famous work than completely replicating it.

Digitizing and producing digital copies of the physical works is only one side. On the other side, these novel datasets pose challenges regarding storage (indexing), access, and search and retrieval. The exploration of the dataset identified some challenges specific to it. The partly solved first challenge is the correctness of the current metadata. The information in the file name does not always corroborate with the drawings' information. This discrepancy can lead to incorrect conclusions about the creator's country or age and the interpretation of the work. The second challenge lies in extracting the metadata. At the current state of digitization, it

is impossible to attribute the drawing to an artist. While most drawings contain the name, OCRizing non-uniform handwritten text is exigent.

The experiments using various CNN models explored the feasibility of retrieving artworks similar to drawings. The findings show that the CNN-based deep image retrieval techniques can efficiently solve the problem of matching drawings and artworks that share a visual similarity. Experimentation involved the CNN models trained for image classification (Baseline), pattern matching (Mini-Replica), and pattern clustering (Clus-Replica) in paintings. The models trained for pattern recognition in paintings achieve good performance than the model trained for classification on ImageNet. Fine-tuning these models enhance their retrieval capability. The best model is the fine-tuned version of a model trained to retrieve artwork photos that contain a similar visual pattern. It achieves an average recall of 88% at a threshold of the top 400 results, almost double that of the model trained for image classification and more than 12 percentage points compared to its non-fine-tuned version. The recall and mean average precision measures double in the best model compared to the baseline image classification model. Fine-tuning the baseline model with relevant examples of drawings and artworks gives metrics similar to the best model.

The style augmentation process that adds watercolor, oil painting, pencil sketch, and texture has a consequential effect on the model's performance. It boosts the evaluation metrics and, most importantly, increases the models' generalization ability. This style transfer helpfully communicates the visual similarity between the drawings and artworks in an otherwise disparate set of images and guides the CNN model in learning those differentiating features.

Qualitative evaluation of the models provides insights into their routine. The baseline model results are as good as the other models only when the drawing is an identical duplicate of the artwork. However, it retrieves images that use the same technique as the drawing. While the Replica variant models get over this, as often as not, the top-ranked artworks by the Clus-Replica model are semantically different from the drawings. The Mini-Replica and the task-specific fine-tuned baseline models learn the object localization and retrieve visually and semantically similar artworks.

Finally, analogizing the comparable performance of the fine-tuned baseline model with Mini-Replica and its fine-tuned version with the size of the training data uncloaks the need for quality data. It shows that even in small quantities, meticulously curated data helps create a deep learning system that performs up to the mark. This thought broadly falls into the realm of Data-Centric AI. As introduced by Andrew Ng[1], Data-Centric AI shifts the focus from improving the code (model) to improving systems for creating efficient and high-quality datasets to improve the overall system's performance.

In conclusion, this work opens a new direction of work in exploring children's drawings, different from using those drawings for psychological or developmental analysis. The project

[1] `https://landing.ai/data-centric-ai/`

demonstrates the possibility of identifying visually similar images by crossing the barriers of domains, techniques, and creators' age and hopefully inspires exploring the children's drawings through an artistic lens and using computer vision techniques for such tasks. The works searching for similarities in style, subjects, and objects of the drawings and famous cultural objects can come together and create a tool to extract artistic inspirations from a drawing.

# Bibliography


[1] Marie-Thérèse van de Kamp, Wilfried Admiraal, Jannet van Drie, and Gert Rijlaarsdam. "Enhancing divergent thinking in visual arts education: Effects of explicit instruction of meta-cognition". In: *British Journal of Educational Psychology* (2015), pp. 47–58. DOI: 10.1111/bjep.12061.

[2] Christopher Tyler and Lora Likova. "The Role of the Visual Arts in Enhancing the Learning Process". In: *Frontiers in Human Neuroscience* (2012). DOI: 10.3389/fnhum.2012.00008.

[3] Anna M. Kindler. "Visual Culture, Visual Brain, and (Art) Education". In: *Studies in Art Education* (2003), pp. 290–296. DOI: 10.1080/00393541.2003.11651745.

[4] *WIKIART - Visual Art Encyclopedia*. URL: https://www.wikiart.org/.

[5] *The Met Collection*. URL: https://www.metmuseum.org/art/collection.

[6] *RIJKS STUDIO*. URL: https://www.rijksmuseum.nl/en/rijksstudio.

[7] Babak Saleh and Ahmed Elgammal. "Large-scale Classification of Fine-Art Paintings: Learning The Right Metric on The Right Feature". In: *International Journal for Digital Art History* (2016). DOI: 10.11588/dah.2016.2.23376.

[8] Wei Ren Tan, Chee Seng Chan, Hernán E. Aguirre, and Kiyoshi Tanaka. "Ceci n'est pas une pipe: A deep convolutional network for fine-art paintings classification". In: *2016 IEEE International Conference on Image Processing (ICIP)*. 2016, pp. 3703–3707. DOI: 10.1109/ICIP.2016.7533051.

[9] Sergey Karayev, Matthew Trentacoste, Helen Han, Aseem Agarwala, Trevor Darrell, Aaron Hertzmann, and Holger Winnemoeller. "Recognizing Image Style". In: *Proceedings of the British Machine Vision Conference*. 2014. DOI: 10.5244/C.28.122.

[10] Eva Cetinic, Tomislav Lipic, and Sonja Grgic. "Fine-tuning Convolutional Neural Networks for fine art classification". In: *Expert Systems with Applications* (2018), pp. 107–118. DOI: 10.1016/j.eswa.2018.07.026.

[11] Nanne van Noord, Ella Hendriks, and Eric Postma. "Toward Discovery of the Artist's Style: Learning to recognize artists by their artworks". In: *IEEE Signal Processing Magazine* (2015), pp. 46–54. DOI: 10.1109/MSP.2015.2406955.

[12] Xi Shen, Alexei A. Efros, and Mathieu Aubry. “Discovering Visual Patterns in Art Collections With Spatially-Consistent Feature Learning”. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 9270–9279.

[13] A. Elgammal, Bingchen Liu, Mohamed Elhoseiny, and Marian Mazzone. “CAN: Creative Adversarial Networks, Generating "Art" by Learning About Styles and Deviating from Style Norms”. In: *ArXiv* (2017). DOI: 10.48550/ARXIV.1706.07068.

[14] Wei Ren Tan, Chee Seng Chan, Hernán E. Aguirre, and Kiyoshi Tanaka. “Improved ArtGAN for Conditional Synthesis of Natural Image and Artwork”. In: *IEEE Transactions on Image Processing* 28 (2019), pp. 394–409.

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. “Generative Adversarial Nets”. In: *Advances in Neural Information Processing Systems*. Vol. 27. 2014.

[16] Benoit Seguin, Carlotta Striolo, Isabella diLenardo, and Frederic Kaplan. “Visual Link Retrieval in a Database of Paintings”. In: *Computer Vision – ECCV 2016 Workshops*. 2016, pp. 753–767.

[17] Eren Gultepe, Thomas Edward Conturo, and Masoud Makrehchi. “Predicting and Grouping Digitized Paintings by Style using Unsupervised Feature Learning.” In: *Journal of cultural heritage* 31 (2018), pp. 13–23.

[18] Giovanna Castellano, Eufemia Lella, and Gennaro Vessio. “Visual Link Retrieval and Knowledge Discovery in Painting Datasets”. In: *Multimedia Tools Appl.* (2021), pp. 6599–6616. DOI: 10.1007/s11042-020-09995-z.

[19] Ahmed Elgammal, Yan Kang, and Milko Den Leeuw. “Picasso, Matisse, or a Fake? Automated Analysis of Drawings at the Stroke Level for Attribution and Authentication”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 32 (2018). URL: https://ojs.aaai.org/index.php/AAAI/article/view/11313.

[20] Matteo Tomei, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. “Art2Real: Unfolding the Reality of Artworks via Semantically-Aware Image-To-Image Translation”. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 5842–5852.

[21] Matteo Tomei, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. “What Was Monet Seeing While Painting? Translating Artworks to Photo-Realistic Images”. In: *Computer Vision – ECCV 2018 Workshops*. Springer International Publishing, 2019, pp. 601–616. ISBN: 978-3-030-11012-3.

[22] Xuanyu He and Wei Zhang. “Emotion recognition by assisted learning with convolutional neural networks”. In: *Neurocomputing* 291 (2018), pp. 187–194. ISSN: 0925-2312. DOI: 10.1016/j.neucom.2018.02.073.

[23] Noa Garcia and George Vogiatzis. “How to Read Paintings: Semantic Art Understanding with Multi-modal Retrieval”. In: *Computer Vision – ECCV 2018 Workshops*. Springer International Publishing, 2019, pp. 676–691. ISBN: 978-3-030-11012-3.

[24] Iria Santos, Luz Castro, Nereida Rodriguez-Fernandez, Álvaro Torrente-Patiño, and Adrián Carballal. "Artificial Neural Networks and Deep Learning in the Visual Arts: a review". In: *Neural Computing and Applications* (2021), pp. 121–157.

[25] Angela Eckhoff. "The Importance of Art Viewing Experiences in Early Childhood Visual Arts: The Exploration of a Master Art Teacher's Strategies for Meaningful Early Arts Experiences". In: *Early Childhood Education* (2008), pp. 463–472. DOI: 10.1007/s10643-007-0216-1.

[26] Masami Toku. "Cross-cultural Analysis of Artistic Development: Drawing by Japanese and U.S. Children". In: *Visual Arts Research* 27 (2001), pp. 46–59. URL: http://www.jstor.org/stable/20716021.

[27] Hui-Chin Yang and Andrea M Noel. "The developmental characteristics of four-and five-year-old pre-schoolers' drawing: An analysis of scribbles, placement patterns, emergent writing, and name writing in archived spontaneous drawing samples". In: *journal of early childhood literacy* 6 (2006), pp. 145–162.

[28] Viktor Lowenfeld. *Creative and mental growth*. 1957.

[29] Larry A. Kantner and Diane C. Gregory. *The Stages of Artistic Development*. 2002. URL: https://makingartwork.files.wordpress.com/2012/12/stages_of_art_development.pdf.

[30] *Institut Mondial d'Art de la Jeunesse*. Accessed on 01 Jun, 2022. URL: https://www.centre-unesco-troyes.org/.

[31] Yi Li and Wenzhao Li. "A Survey of Sketch-Based Image Retrieval". In: *Mach. Vision Appl.* (2018), pp. 1083–1100. DOI: 10.1007/s00138-018-0953-8.

[32] Navneet Dalal and Bill Triggs. "Histograms of oriented gradients for human detection". In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)* 1 (2005), pp. 886–893.

[33] David G Lowe. "Distinctive image features from scale-invariant keypoints". In: *International journal of computer vision* 60.2 (2004), pp. 91–110.

[34] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. "Speeded-Up Robust Features (SURF)". In: *Computer Vision and Image Understanding* 110 (2008), pp. 346–359. DOI: https://doi.org/10.1016/j.cviu.2007.09.014.

[35] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[36] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems* 25 (2012).

[37] Haris Iqbal. *HarisIqbal88PlotNeuralNet v1.0.0*. Version v1.0.0. Dec. 2018. DOI: 10.5281/zenodo.2526396.

[38] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. "ImageNet Large Scale Visual Recognition Challenge". In: *Int. J. Comput. Vision* 115 (2015), pp. 211–252. DOI: 10.1007/s11263-015-0816-y.

[39] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. "Deep Residual Learning for Image Recognition". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 770–778.

[40] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. "Aggregated Residual Transformations for Deep Neural Networks". In: *arXiv preprint arXiv:1611.05431* (2016).

[41] Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. "Siamese neural networks for one-shot image recognition". In: *ICML deep learning workshop*. Vol. 2. 2015.

[42] Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni. "Generalizing from a Few Examples: A Survey on Few-Shot Learning". In: *ACM Comput. Surv.* 53 (2020). DOI: 10.1145/3386252.

[43] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. "ImageNet: A large-scale hierarchical image database". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848.

[44] Matthew D Zeiler and Rob Fergus. "Visualizing and understanding convolutional networks". In: *European conference on computer vision*. Springer. 2014, pp. 818–833.

[45] Jonathan Long, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3431–3440.

[46] Hoo-Chang Shin, Holger R Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura, and Ronald M Summers. "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning". In: *IEEE transactions on medical imaging* 35.5 (2016), pp. 1285–1298.

[47] Florian Schroff, Dmitry Kalenichenko, and James Philbin. "FaceNet: A unified embedding for face recognition and clustering". In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), pp. 815–823.

[48] Jiwen Lu, Junlin Hu, and Jie Zhou. "Deep Metric Learning for Visual Understanding: An Overview of Recent Advances". In: *IEEE Signal Processing Magazine* (2017), pp. 76–84.

[49] Edgar Simo-Serra, Eduard Trulls, Luis Ferraz, Iasonas Kokkinos, P. Fua, and Francesc Moreno-Noguer. "Discriminative Learning of Deep Convolutional Feature Point Descriptors". In: *2015 IEEE International Conference on Computer Vision (ICCV)* (2015), pp. 118–126.

[50] James Philbin, Ondřej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. "Object retrieval with large vocabularies and fast spatial matching". In: *2007 IEEE Conference on Computer Vision and Pattern Recognition* (2007), pp. 1–8.

[51] Ji Wan, Dayong Wang, Steven Chu Hong Hoi, Pengcheng Wu, Jianke Zhu, Yongdong Zhang, and Jintao Li. "Deep Learning for Content-Based Image Retrieval: A Comprehensive Study". In: *Proceedings of the 22nd ACM International Conference on Multimedia*. 2014, pp. 157–166. DOI: 10.1145/2647868.2654948.

[52] Wen-gang Zhou, Houqiang Li, and Qi Tian. "Recent Advance in Content-based Image Retrieval: A Literature Survey". In: *ArXiv* (2017). DOI: 10.48550/ARXIV.1706.06064.

[53] Abhinav Shrivastava, Tomasz Malisiewicz, Abhinav Gupta, and Alexei A. Efros. "Data-Driven Visual Similarity for Cross-Domain Image Matching". In: *Proceedings of the 2011 SIGGRAPH Asia Conference*. 2011. DOI: 10.1145/2024156.2024188.

[54] *Web Gallery of Art*. URL: www.wga.hu.

[55] Thomas Mensink and Jan van Gemert. "The Rijksmuseum Challenge: Museum-Centered Visual Recognition". In: Association for Computing Machinery, 2014, pp. 451–454. DOI: 10.1145/2578726.2578791.

[56] Icaro. *Best Artworks of All Time*. Accessed on 07 Apr, 2022. 2019. URL: https://www.kaggle.com/datasets/ikarus777/best-artworks-of-all-time.

[57] Ruben and Anna. *Art Challenge - A quiz game of famous painters*. 2014. URL: https://artchallenge.ru/?lang=en.

[58] Benoît Seguin. "Making large art historical photo archives searchable". In: (2018), p. 169. DOI: 10.5075/epfl-thesis-8857.

[59] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *3rd International Conference on Learning Representations, ICLR*. 2015. URL: http://arxiv.org/abs/1412.6980.

[60] Ludovica Schaerf. "Semi-supervised clustering of learned visual signatures of artworks". In: (2022).

[61] Mathias Eitz, James Hays, and Marc Alexa. "How Do Humans Sketch Objects?" In: *ACM Trans. Graph. (Proc. SIGGRAPH)* 31 (2012), 44:1–44:10.

[62] Jonas Jongejan, Henry Rowley, Takashi Kawashima, Jongmin Kim, and Nick Fox-Gieg. "The Quick, draw! - A.I. Experiment". In: (2016). URL: https://quickdraw.withgoogle.com/.

[63] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. "The Sketchy Database: Learning to Retrieve Badly Drawn Bunnies". In: *ACM Trans. Graph.* 35 (2016). DOI: 10.1145/2897824.2925954.

[64] Ksenia Konyushkova, Nikolaos Arvanitopoulos, Zhargalma Dandarova Robert, Pierre-Yves Brandt, and Sabine Süsstrunk. "God(s) Know(s): Developmental and Cross-Cultural Patterns in Children Drawings". In: *ArXiv* abs/1511.03466 (2015).

# Appendix A

# Age and Category wise drawings

Table A.1 shows the distribution of drawings across different age groups for each year between 1994 and 2020. The drawings are classified as unknown when extracting the age information from the drawing was not feasible. Additionally, there are no drawings in the 18-25 years group for the year 2006 and in the 3-5 and 6-9 age groups of 2007. The probable reason for nonexistent drawings in those categories could be because some part of the collection is yet to be digitized.

| **Age Group → Year ↓** | **3-5** | **6-9** | **10-13** | **14-17** | **18-25** | **MS** | **unknown** | **Year Total** |
|---|---|---|---|---|---|---|---|---|
| **1994** | 2 | 99 | 711 | 340 | 27 | 45 | 0 | 1224 |
| **1995** | 139 | 1507 | 2064 | 1199 | 211 | 78 | 0 | 5198 |
| **1996** | 253 | 1207 | 1230 | 615 | 111 | 59 | 142 | 3617 |
| **1997** | 172 | 767 | 755 | 439 | 111 | 41 | 3 | 2288 |
| **1998** | 258 | 924 | 1107 | 733 | 291 | 2 | 1 | 3316 |
| **1999** | 264 | 839 | 1396 | 518 | 139 | 29 | 0 | 3185 |
| **2000** | 366 | 2012 | 3462 | 1810 | 255 | 89 | 0 | 7994 |
| **2001** | 424 | 1307 | 2008 | 946 | 144 | 19 | 2 | 4850 |
| **2002** | 329 | 1458 | 1754 | 538 | 89 | 43 | 0 | 4211 |
| **2003** | 415 | 1574 | 2377 | 1416 | 186 | 47 | 0 | 6015 |
| **2004** | 391 | 1915 | 2248 | 830 | 157 | 5 | 0 | 5546 |
| **2005** | 325 | 935 | 1248 | 534 | 1 | 0 | 0 | 3043 |
| **2006** | 208 | 642 | 512 | 190 | 0 | 0 | 0 | 1552 |
| **2007** | 0 | 0 | 732 | 295 | 0 | 0 | 0 | 1027 |
| **2008** | 385 | 141 | 2091 | 447 | 149 | 0 | 0 | 3213 |
| **2009** | 119 | 842 | 1168 | 746 | 118 | 65 | 0 | 3058 |
| **2010** | 438 | 1165 | 1096 | 444 | 73 | 20 | 0 | 3236 |
| **2011** | 179 | 530 | 696 | 270 | 45 | 13 | 14 | 1747 |
| **2012** | 208 | 877 | 1078 | 366 | 65 | 17 | 0 | 2611 |
| **2013** | 262 | 1209 | 1706 | 806 | 210 | 0 | 0 | 4193 |
| **2014** | 109 | 610 | 911 | 497 | 72 | 0 | 0 | 2199 |
| **2015** | 206 | 1528 | 2160 | 820 | 182 | 21 | 0 | 4917 |
| **2016** | 206 | 785 | 811 | 466 | 54 | 0 | 0 | 2322 |
| **2017** | 113 | 647 | 645 | 177 | 44 | 0 | 0 | 1626 |
| **2018** | 181 | 538 | 569 | 352 | 70 | 9 | 0 | 1719 |
| **2019** | 199 | 703 | 550 | 508 | 74 | 0 | 0 | 2034 |
| **2020** | 473 | 2288 | 1537 | 787 | 40 | 37 | 0 | 5162 |
| **Category Total** | 6624 | 27049 | 36622 | 17089 | 2918 | 639 | 162 | 91103 |

Table A.1: Number drawings present in each age group for every year

# Appendix B

# Discovered Drawing-Artwork Pairs

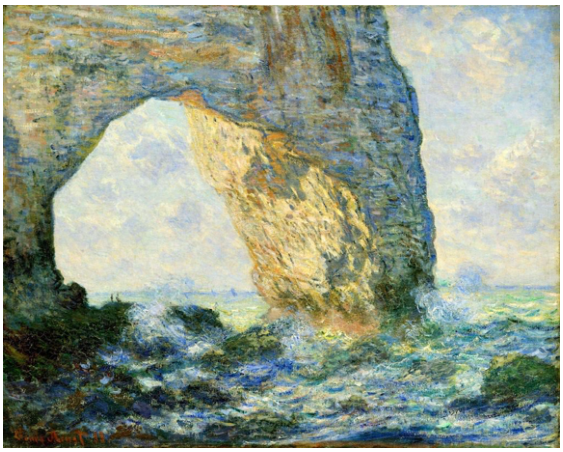
This appendix contains a few examples of new drawing-artwork pairs found using the FT Aug and FT Aug Mini-Replica models. These pairs were not known earlier and were not part of the group of manually annotated pairs.


(a) Drawing: `2004_14-17_1013_BLG_R_C`


(b) Artwork: Cyril and Methodius

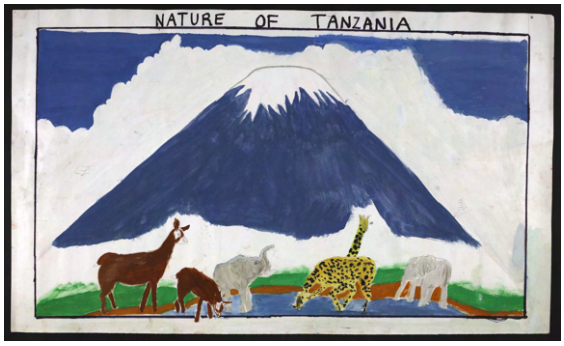
Figure B.1: Discovered Drawing-Artwork Pair: 1

(a) Drawing: `2011_14-17_2059_UKR_R_C`



(b) Artwork: La Manneporte (Étretat) by Claude Monet

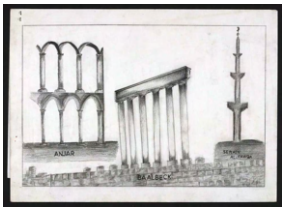
Figure B.2: Discovered Drawing-Artwork Pair: 2



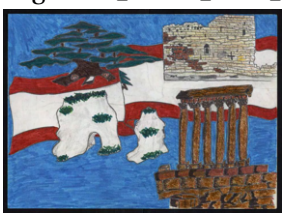

(a) Drawing: `1998_14-17_0441_TAN_R_C`


(b) Artwork: A stamp of Mount Kilimanjaro

Figure B.3: Discovered Drawing-Artwork Pair: 3



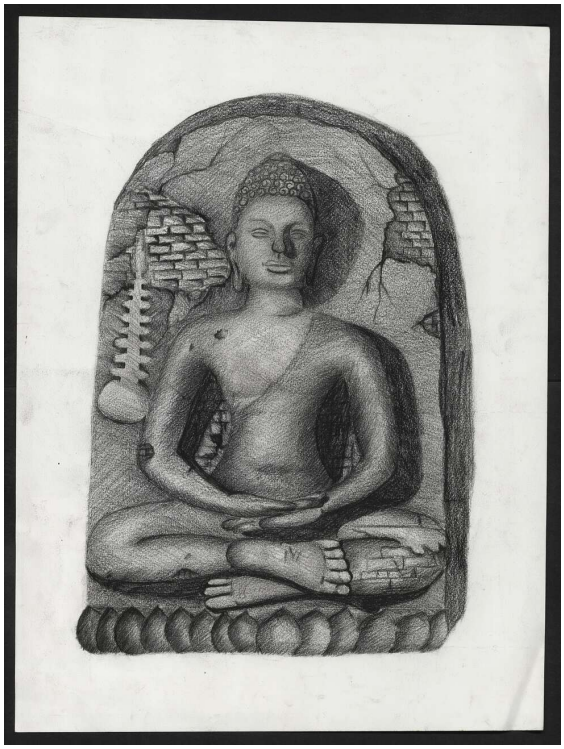
(a) Drawing: `1998_18-25_0080_LBA_R_C`



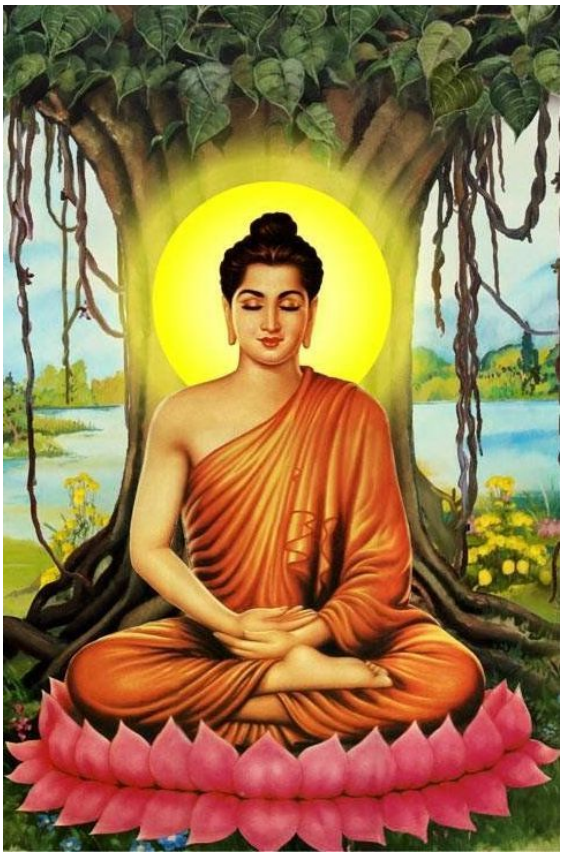
(b) Drawing: `1998_14-17_0125_LBA_R_C`

(c) Artwork: Ruins of Temple of Jupiter in Baalbek, Lebanon

Figure B.4: Discovered Drawing-Artwork Pair: 4



(a) Drawing: 2003_14-17_1243_THA_R_C



(b) Artwork: Buddha sitting under a tree

Figure B.5: Discovered Drawing-Artwork Pair: 5