POLITECNICO DI MILANO

Facoltà di Ingegneria

Scuola di Ingegneria Industriale e dell'Informazione

Dipartimento di Elettronica, Informazione e Bioingegneria

Master of Science in

Computer Science and Engineering

Identification of salient iconography features in artwork analysis

Supervisor:

PROF. PIERO FRATERNALI

Co-supervisors:

PROF. RICARDO DA SILVA TORRES

FEDERICO MILANI

Master Graduation Thesis by:

NICOLÒ ORESTE PINCIROLI VAGO

Student Id n. 939782

Academic Year 2020-2021

ACKNOWLEDGMENTS

I would like to thank, first and foremost, all my supervisors for having guided me during this path and for their valuable suggestions. In particular, I would like to thank Professor Piero Fraternali for having introduced me to this thesis and for having made constructive critics. I would like to thank Professor Ricardo Torres, for his encouragement and his availability. I would like to thank Federico Milani for having always supported me when I was struggling.

I would like to thank Møre og Romsdal fylkeskommune for having funded my research and for having discussed useful applications of this work for the county.

I would like to thank all the professors I have worked with during the last two years since they have allowed me to develop my passions and to learn something in diverse fields. In particular, I deeply thank, in addition to my supervisors, Professors Hameed and Kachelriess from NTNU, Professor Rossi from Politecnico di Milano and the team of researchers from Brazil for having worked together on interesting research topics. Thank you also to Professor Arne Styve for having reviewed the Norwegian version of the thesis abstract.

I would like to thank my class in Norway since everybody has been supportive at any time, and the International Offices in Milan and Ålesund, since they have always been helpful and kind in case of need.

Last, but not least, I want to thank my friends for their everlasting friendships, and my family, which has always been supporting me.

Thank you.

FUNDING

This research is funded by Møre og Romsdal fylkeskommune, in the context of the *Masterstipend innan kultur*, whose purpose is "facilitating research and competence development in the field of culture."[1]

[1] https://mrfylke.no/tilskot-prisar-og-stipend/stipend/masterstipend-kultur-2021 (As of July 2021).

ABSTRACT

Iconography studies the visual content of artworks by considering the themes portrayed in them and their representation. Computer Vision has been used to identify iconography subjects in paintings and Convolutional Neural Networks (CNN) enabled the effective classification of characters in Christian art paintings. However, it still has to be demonstrated if the classification results obtained by CNNs rely on the same iconographic properties that human experts exploit when studying iconography. A suitable approach for exposing the process of classification by neural models relies on Class Activation Maps, which emphasize the areas of an image contributing the most to the classification. This work compares state-of-the-art algorithms (CAM, Grad-CAM, Grad-CAM++, and Smooth Grad-CAM++) in terms of their capacity of identifying the iconographic attributes that determine the classification of characters in Christian art paintings. Quantitative and qualitative analyses show that Grad-CAM, Grad-CAM++, and Smooth Grad-CAM++ have similar performances while CAM has lower efficacy. Smooth Grad-CAM++ isolates multiple disconnected image regions that identify small iconography symbols well. Grad-CAM produces wider and more contiguous areas that cover large iconography symbols better. The illustrated analysis is a step towards the computer-aided study of the variations of iconography elements positioning and mutual relations in artworks and opens the way to the automatic creation of bounding boxes for training detectors of iconography symbols in Christian art images.

SOMMARIO

L'iconografia studia il contenuto visivo delle opere d'arte considerando i temi ritratti in esse e la loro rappresentazione. La visione artificiale è stata utilizzata per identificare i soggetti dell'iconografia nei dipinti e le reti neurali convoluzionali (CNN) hanno consentito l'effettiva classificazione dei personaggi nei dipinti d'arte cristiana. Tuttavia, deve ancora essere dimostrato se i risultati di classificazione ottenuti dalle CNN si basano sulle stesse proprietà iconografiche che gli esperti umani sfruttano quando studiano l'iconografia. Un approccio adeguato per esporre il processo di classificazione mediante modelli neurali si basa sulle Class Activation Map, che enfatizzano le aree di un'immagine che contribuiscono maggiormente alla classificazione. Questo lavoro confronta algoritmi allo stato dell'arte (CAM, Grad-CAM, Grad-CAM++ e Smooth Grad-CAM++) in termini di capacità di identificare gli attributi iconografici che determinano la classificazione dei personaggi nei dipinti d'arte cristiana. Le analisi quantitative e qualitative mostrano che Grad-CAM, Grad-CAM++ e Smooth Grad-CAM++ hanno prestazioni simili mentre CAM ha un'efficacia inferiore. Smooth Grad-CAM++ isola più regioni dell'immagine disconnesse che identificano bene i piccoli simboli iconografici. Grad-CAM produce aree più ampie e contigue che coprono meglio i grandi simboli iconografici. L'analisi illustrata è un passo verso lo studio assistito da computer delle variazioni del posizionamento degli elementi iconografici e delle relazioni reciproche nelle opere d'arte e apre la strada alla creazione automatica di bounding box per il training di detector di simboli iconografici nelle immagini dell'arte cristiana.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

ADL	Attention Dropout Layer
ANN	Artificial Neural Network
CAM	Class Activation Map
CIFAR	Canadian Institute for Advanced Research
cIoU	Component Intersection over Union
CNN	Convolutional Neural Network
COCO	Common Objects in Context
CV	Computer Vision
FC	Fully Connected
GAP	Global Average Pooling
GC	Grad-CAM

GPU	Graphics Processing Unit
Grad-CAM++	Gradient-weighted Class Activation Map ++
Grad-CAM	Gradient-weighted Class Activation Map
IoU	Intersection Over Union
LSTM	Long short-term memory
mAP	Mean Average Precision
MIL	Multiple-Instance Learning
MS	Microsoft
NWC	Negative Weight Clamping
PASCAL	Pattern Analysis, Statistical Modelling and Computational Learning
ResNet	Residual Network
R-CNN	Region Based Convolutional Neural Networks
RGB	Red Green Blue
ReLU	Rectified Linear Unit
Smooth Grad-CAM++	Smooth Gradient-weighted Class Activation Map ++
TAP	Thresholded Average Pooling
VOC	Visual Object Classes

1 INTRODUCTION

Iconography is the discipline that concerns itself with the subject matter of artworks, as opposed to their form [53]. It is studied to understand the meaning of artworks and to analyze the influence of culture and beliefs on art representations across the word, from the Nasca [59] to the Byzantine [54] civilization. Iconography is a prominent topic of the art history studied through centuries [39, 62, 75]. The attribution of iconography elements (henceforth *classes*) is an important task in art history, related to the interpretation of meaning and to the definition of the geographical and temporal context of an artwork.

With the advent of digital art collections, iconography class attribution has acquired further importance, as a way to provide a significant index on top of digital repositories of art images, supporting both students and experts in finding and comparing works by their iconography attributes. However, the analysis of iconography requires specialized skills, based on the deep knowledge of the symbolic meaning of a very high number of elements and of their evolution in space and time.[1] This makes the manual attribution of iconography classes to image collections challenging, due to the tension between the available amount of expert work and the high number of items to be annotated.

[1] The WikiPedia page on Christian Saint symbolism (https://en.wikipedia.org/wiki/Saint_symbolism – As of June 2021) lists 257 characters with 791 attributes.

A viable alternative relies on the use of semi-automatic computer-aided solutions supporting the expert annotator in the task of associating iconography classes to art images. Computer Vision (CV) has already been used for artwork analysis tasks, such as genre identification [90], author identification [66], and even subject identification and localization [13]. The field of computer-aided iconography analysis is more recent and addressed by few works [28, 47]. Borrowing the standard CV terminology, the problem of computer-aided iconography analysis can be further specialized into *iconography classification*, which tackles the association of iconography classes to an artwork image as a whole, and *iconography detection*, which addresses the identification of the regions of an image in which the attributes representing an iconography class appear.

Applying CV to the analysis of art iconography poses challenges in part general and in part specific to the art iconography field. As in general-purpose image classification and object detection, the availability of large high quality training data is essential. The natural image data set in use nowadays are very large and provided with huge numbers of annotations. Conversely, in the narrower art domain, image data sets are less abundant, smaller, and with less high quality annotations. Furthermore, unlike natural images, painting images are characterized by less discriminative features than natural ones. The color palette is more restricted and subject to artificial effects, such as colored shadows and chiaroscuro. Images of paintings may also portray partially deteriorated subjects (e.g., in frescoes) and belong to historical archives of black and white photos.

Despite the encouraging results of applying CNNs for iconography classification [47], it remains unclear how such a task is performed by artificial models. Depending on the class, the human expert may consider the whole scene portrayed in the painting or instead focus on specific hints. Considering Christian art iconography, an example of the first scenario occurs in paintings of complex scenes, such as the crucifixion or the visitation of the magi. The latter case is typical of the identification of characters, especially Christian saints, which depends on the presence of very distinctive attributes. When CNNs are used for the classification task, the problem of *explainability* arises, i.e., of exposing how the CNN has produced a given result. A widely used strategy to clarify CNN image classification results relies on the use of Class Activation Maps [55, 70, 82], which visualize the regions of the input images that have the most impact on the prediction of the CNN. Computing the most salient regions of an image with respect to its iconography can help automate the creation of bounding boxes around the significant elements of an artwork from only image-wide annotations. This result could reduce the effort of building training sets for the much harder task of iconography detection.

This work addresses the following research questions:

- Are Class Activation Maps an effective tool for understanding how a CNN classifier recognizes the iconography classes of a painting?
- Are there significant differences in the state-of-the-art CAM algorithms with respect to their ability to support the explanation of iconography classification by CNNs?
- Are the image areas highlighted by CAMs a good starting point for creating semi-automatically the bounding boxes necessary for training iconography detectors?

The contributions of the conducted research can be summarized as follows:

- We apply four state-of-the-art class activation map algorithms[2] (namely, CAM [88], Grad-CAM [65], Grad-CAM++ [15], and Smooth Grad-CAM++ [52]) to the CNN iconography classification model presented in [47], which exploits a backbone based on ResNet50 [33] trained on the ImageNet data set [23] and refined on the ArtDL[3] data set consisting of 42,479 images of artworks portraying Christian saints divided into 10 classes.

- For the quantitative evaluation of the different algorithms, a test data set has been built which comprises 823 images annotated with 2957 bounding boxes surrounding specific iconographic symbols. One such annotated image is shown in Figure 1.1. We measured the agreement between the areas of the image highlighted by the algorithm and the bounding boxes annotated manually as ground truth.

- We analyzed the class activation map area based on percentage of covered area that does not contain any iconographic symbol.

- We count the symbol-level bounding boxes and the symbols covered by at least 20% of Grad-CAM. This investigation shows that the majority of the symbols is covered and that there are no significant differences between the symbols spreading across several bounding boxes and the ones contained in single bounding boxes..

[2] Note that, in order to avoid ambiguity, we refer to the specific algorithm as "CAM" and to the generic output as "class activation maps

[3] http://www.artdl.org (As of June 2021).

- We generate Saint-level bounding boxes from the class activation maps and perform a qualitative and a quantitative evaluation, calculating the GT-known Loc metric and the mean Average Precision. This analysis confirms that the best results are achieved using Grad-CAM, which gives additional evidence of the fact that Grad-CAM is better at focusing on the iconographical symbols and on the saint bodies.

- We perform a qualitative evaluation by examining the overlap between the ground truth bounding boxes and the class activation maps. This investigation illustrates the strengths and weaknesses of the analyzed algorithms, highlights their capacity of detecting symbols that were missed by the human annotator and discusses cases of confusion between the symbols of different classes.

- We generate symbol-level bounding boxes from the class activation maps and perform a qualitative evaluation, that emphasizes the advantages and limitations of using class activation maps. The results show that Grad-CAM is able to locate relevant iconographical symbols.

- The comparisons show that Grad-CAM, Grad-CAM++, and Smooth Grad-CAM++ deliver better results than the original CAM algorithm in terms of area coverage and explainability. This finding confirms the result discussed in [52] for natural images. Smooth Grad-CAM++ produces multiple disconnected image regions that identify small iconography symbols quite precisely. Grad-CAM produces wider and more contiguous areas that cover well both large and small iconography symbols.

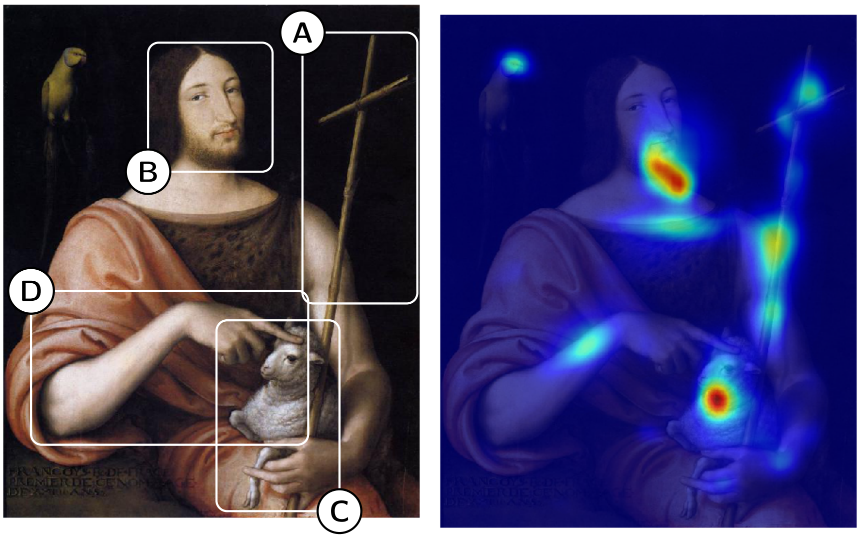


Figure 1.1: On the left: **Saint John the Baptist** image and iconography symbols identified manually (e.g., cross (A), face (B), and lamb (C), and hand pointing at lamb (D)). On the right: the CAM heat map associated with classification results of a CNN-based solution.

To the best of our knowledge, such a comparison has not been performed before in the context of artwork analysis.

This research work has also been published by the Journal of Imaging [57], and this thesis extends that publication.

Figure 1.1 shows an example of the assessment performed in this research. On the left, an image of Saint John the Baptist has been manually annotated with the regions (from A to D) associated with key symbols relevant for iconography classification. On the right, the same image is overlaid with the CAM heat map showing the regions contributing the most to the classification.

The rest of this document is organized as follows: Chapter 2 surveys related work; Chapter 3 describes the different CAM variants considered in our study; Chapter 4 describes the adopted evaluation protocol and the results of the quantitative and the qualitative analysis; Chapter 5 draws the conclusions and outlines possible future work;

finally, the Appendix A studies in more detail the behaviour of Smooth Grad-CAM++ for different hyper-parameters.

2

BACKGROUND CONCEPTS AND RELATED WORK

This chapter introduces relevant background concepts related to Artificial Intelligence and surveys the essential previous research in automated artwork analysis and CNN interpretability, the foundations of our work.

2.1 ARTIFICIAL INTELLIGENCE

This section introduces the most important characteristics and subfields of Artificial Intelligence since they serve as a foundation for this research. In particular, Section 2.1.1 introduces the most important characteristics of Machine Learning, Section 2.1.2 presents Artificial Neural Networks, Section 2.1.3 describes Convolutional Neural Networks in the context of image analysis, Section 2.1.4 presents the concept of Residual networks, and, finally, Section 2.1.5 gives an overview of Computer Vision techniques.

2.1.1 *Machine Learning*

This section introduces the concept of Machine Learning, a sub-field of Artificial Intelligence that can simulate a form of inductive reasoning given sets of sample data.

Machine Learning deals specifically with problems for which it is possible to draw conclusions from a set of examples. Such examples may be, for instance, images, videos, time series, text, or numerical data, depending on the problem. In particular, classification is one of the tasks of Machine Learning algorithms, and it aims at automatically assigning labels to unknown data given examples of similar data.

A data set may be labelled (i.e., each sample is associated with a class or a numerical quantity) or unlabelled. Labels, when present, are characterized by different levels of granularity, depending on the problem. For instance, given an image, it is possible to create a label referred to the image as a whole or a specific part of the image (e.g., a person, or an object). Labels can be created in different ways, depending on the nature of the data. In general, it is possible to create labels manually for each sample in a data set (e.g., for each image, indicating its content). In particular cases, it may be possible to create labels automatically or semi-automatically (e.g., using a heuristic procedure). The presence or absence of the labels determines the class of algorithms employed for learning from the data. In particular, it is possible to identify the four main branches of learning [18]:

- Fully supervised learning relies on labelled data (i.e., given a set of data labelled with a given granularity, a new datum is classified with the same level of granularity);
- Unsupervised learning relies on unlabelled data (i.e., it is possible to group data according, for instance, to similar characteristics, but it is not possible to assign a label to each group of data):

- Self-supervised learning does not rely on data annotated by humans;
- Reinforcement learning exploits the creation of adversary models with the purpose, for example, to develop stronger models (e.g., for improving the ability in playing a game)

Depending on the problem, a single datum may be associated with a variable number of labels, which in general may be greater than one. For instance, a sentence can be labelled concerning the emotions it conveys, and an image concerning the different objects it contains. Those are examples of multi-instance classification. Depending on the considered data, the approach may be different, where the most general approaches deal with multi-instance data sets.

Next, we introduce weakly supervised learning, a branch of machine learning not relying on full ground-truth labels.

Weakly supervised learning

Zhou has given an introduction to weakly supervised learning [89], emphasizing how it differs from fully supervised learning. Such difference is based on the concept of supervision. Supervised learning relies on training examples, which allow the creation of predictive models during the training phase. Such predictive models are able, given data unknown to the network (i.e., the test set), to assign them one or multiple labels. On the other hand, unsupervised learning allows, for instance, to group similar data, but not to assign a class to each group. Supervised learning, for this reason, is more precise than unsupervised learning, but requires labels both on the training set and on the test set. Since the annotation process is tedious and has a high cost, it was necessary to introduce an alternative kind of supervision, based on partial, inaccurate or coarse-grained labels. This kind of supervision is named "weakly supervised" and the associated learning technique is called "weakly supervised learning." There exist several types of weak supervision. Three of the most relevant are:

- Incomplete supervision (i.e., only a subset of the data is labelled);
- Inexact supervision (i.e., the training data labels are coarse-grained);
- Inaccurate supervision (i.e., the labels are not always ground-truth).

In case of incomplete supervision, two techniques can be employed:

- Active learning, which assumes that an oracle (e.g., a human expert) can label the missing data when necessary;
- Semi-supervised learning, which aims at exploiting the lack of labels o improve the learning performance, without the intervention of external oracles.

Weakly supervised learning showed promising results in diverse fields. Recent research by Ali-Dib et al. [2] proposed the application of this technique to the crater shape retrieval task. This example emphasizes the advantages of avoiding detailed manual labelling since it presents a particularly time-consuming task. Figure 2.1[1] shows the main challenge associated with their research, i.e., the number and variety of craters on the Moon.

Weakly supervised learning is also employed in the field of medicine. For example, Kanavati et al. recently proposed research on lung carcinoma, to differentiate between lung carcinoma and non-neoplastic [35],

[1] Underlying image by Nicolas Thomas on Unsplash (https://unsplash.com/photos/wKlqqfNTLsI).

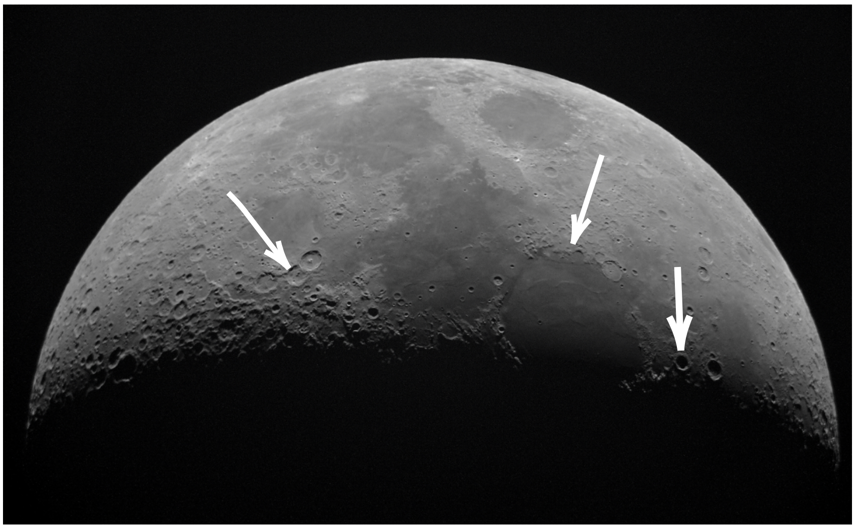


Figure 2.1: Weakly-supervised learning applied to crater detection – This figure shows the challenge in detecting craters. Only a few of them are indicated, but the results presented by [2] are able to find more of them automatically.

while Dong et al. applied weakly-supervised learning for endoscopic lesions segmentation [24].

2.1.2 *Artificial Neural Networks*

This section introduces Artificial Neural Networks, one of the main computational models employed in Machine Learning.

An **Artificial Neural Network** (ANN) is a collection of connected nodes (the *artificial neurons*) forming a structure loosely inspired by the biological brain structure. In particular, the principle behind the definition of Neural Networks is that the complexity of the data can be better tackled by creating a system constituted by several atomic structures (also indicated as nodes or artificial neurons), each with a simple and limited purpose. The complexity, therefore, emerges from the combined behaviour of those nodes. This idea is similar to the behaviour observed in a biological brain, where atomic structures (e.g., the neurons) manage to perform challenging tasks (e.g., image recognition, critical thinking, and motion) by establishing a network.

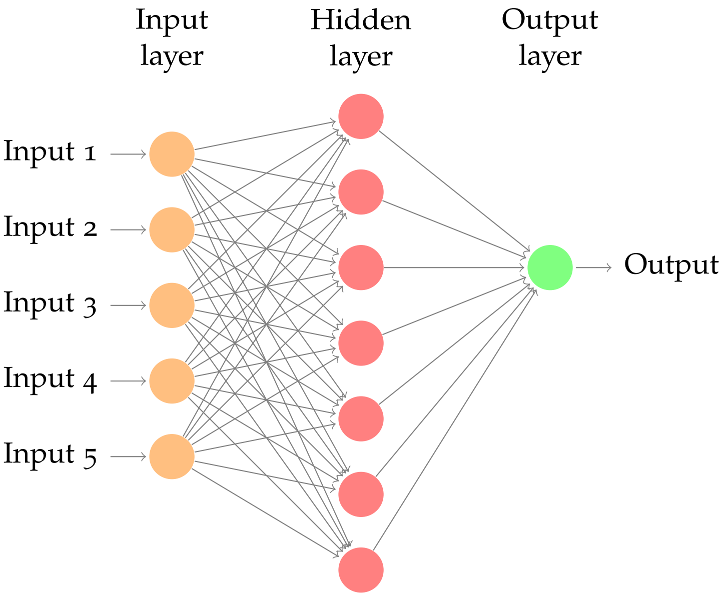


Figure 2.2: An example of Artificial Neural Network – This example shows a basic ANN, with an input (in orange), one hidden layer (in red) and an output (in green).

While the most elementary type of neural network (i.e., the *feed-forward* neural network) has a simple structure and does not contain cycles (being, indeed, a Directed Acyclic Graph), there exist more complex networks, which introduce a wider variety of substructures and possibly, as in the case of Recurrent Neural Networks, cycles. Figure 2.2 presents a basic example of feed-forward ANN, which consists of an input layer, only one hidden layer, and an output layer. During the process through which the network learns based on the given examples (i.e., the training), the network evolves. Each connection is associated with a weight, which changes during the learning process (or training). The nodes, on the other hand, do not change and perform the same operation.

The nodes can perform virtually any type of operation on the input. In particular, continuous and derivable functions are preferred since the training process relies on the progressive update of the arc weights, which happens by computing derivatives. For instance, ReLU is a common operation, defined as:

$$\mathrm{ReLU}(x) = max(0, x) \tag{2.1}$$

This function is continuous on the entire domain ($\mathbb{R}$), but it is derivable only in $\mathbb{R} \setminus \{0\}$. For this reason, there exist alternative functions with similar behaviour and derivable in all $\mathbb{R}$.

Another dimension of analysis considers the number of layers of a network rather than its nodes. If a neural network has more than one layer, it is a deep network and it is studied in the context of **Deep Learning**, a sub-field of Machine Learning, dealing specifically with more complex data and problems. In particular, one-layer ANNs can deal with a limited number of problems and cannot deal with data sets that are not linearly separable. In general, complex data are not linearly separable, therefore a multi-layer network is necessary to perform predictions. Shallow Learning, on the other hand, refers to one or two-layer networks [18].

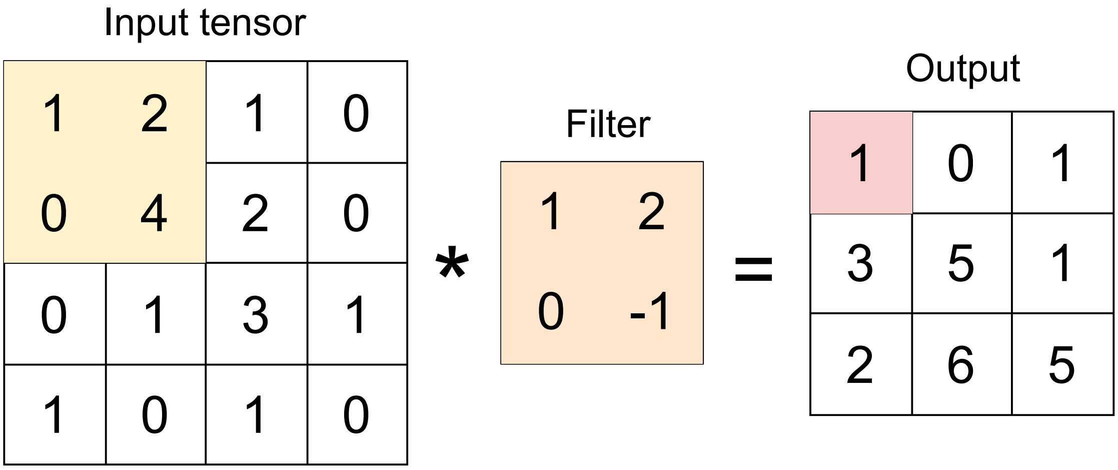
2.1.3 *Convolutional Neural Networks*

This section (extracted from [56, 74]) introduces the concept of Convolutional Neural Network, as a particular case of Artificial Neural Networks. Being used chiefly for image-based data, Convolutional Neural Networks are fundamental in this research work.

A **Convolutional Neural Network** (CNN) is a neural network that, given an input datum represented as a tensor and a set of classes, can be used to predict the class (or, more in general, the classes) to which the datum belongs. More precisely, CNNs are a class of artificial neural networks (ANN), which, differently from traditional ANNs, can perform convolutions in one or more dimensions using convolutional layers. In particular, convolutions are defined by introducing one or more filters, which are tensors of numbers. A filter is originally placed in correspondence of the top-left corner of the tensor representing an input datum (e.g., an image), and an element-wise multiplication between the filter and the underlying datum elements is performed. The multiplied elements are then summed and placed in an output tensor, in the same position as the filter top-left corner position. The filter, then, is moved by a quantity called *stride* along all the tensor elements, until the entire tensor has been covered. It is also possible to introduce *padding*, which consists of adding zeros to the tensor borders to obtain an output tensor with the same dimensions as the input tensor. Figure 2.3 presents the result of convolution on a bidimensional tensor with a single filter. The highlighted elements represent the first operation performed by the convolution, which in this example is given by:

$$\begin{pmatrix} 1 & 2 \\ 0 & 4 \end{pmatrix} * \begin{pmatrix} 1 & 2 \\ 0 & -1 \end{pmatrix} = 1 \cdot 1 + 2 \cdot 2 + 0 \cdot 0 + 4 \cdot (-1) = 1 \tag{2.2}$$

Additional to the convolution, many CNNs include pooling operations which output, for each position of the filter, the maximum element below the filter (*max pooling*) or the average of the elements



Input tensor

1	2	1	0
0	4	2	0
0	1	3	1
1	0	1	0

Filter

1	2
0	-1

Output

1	0	1
3	5	1
2	6	5

Figure 2.3: A convolution with a single filter.

below the filter (*average pooling*). In the final part of a CNN, it is necessary to insert fully connected layers, which consider all the inputs from the previous layer, perform a linear combination of such inputs, and output a vector whose length corresponds to the number of classes of the problem. The content of this vector consists of the probabilities associated with the different classes. A CNN, then, is formed by a sequence of convolutional layers, pooling layers, more complex layers based on them, and fully connected layers. This means that it encodes a complex transformation of the initial data into the labels associated with them. When images are considered, a convolution is defined over three dimensions (the width, the height, and the depth, which initially represent the colour channels encoded in the RGB format, while the subsequent layers represent the effect of the application of different filters over the input). This means that the input tensor has the shape $n \times m \times 3$ for a coloured image with $n \times m$ pixels. After the first convolution, the output tensor will have the shape $n' \times m' \times f$ where f is the number of filters, and n' and m' depend on the filter size and the presence of padding.

2.1.4 *Residual networks*

This section introduces the concept of Residual Networks, a class of ANNs whose purpose is contrasting the vanishing gradient phenomenon (i.e., the inability of deep networks to update their initial weights as the learning progresses) in Deep Residual Learning.

Residual Networks are implemented by introducing skip connections. Intuitively, their purpose is to create, in addition to the main path from the input layer to the output layer, shorter paths, which are meant to propagate promising results using fewer steps. The main building block of Residual Networks, differently from traditional CNNs, is a residual block. The concept of "residual block" is generic, hence there exist several kinds of residual blocks. Khan et al. recently presented a thorough survey about the architectures of deep convolutional neural networks [37], which analyses different residual blocks.

Deep Residual Networks have been exploited in the Image Recognition task. He et al. [33] applied plain (i.e., non-residual networks) to the CIFAR-10 data set [41], and observed that a consistent decrease in the training error was not associated with an equally consistent decrease in the test error, yielding to the network saturation as the depth increased. Moreover, increasing the number of layers yielded a higher error, both for the training and the test set.

The application of residual networks (e.g., ResNet50), instead, yields better results both in terms of training and test error. He et al. [33] also compared the behaviour of plain and residual networks for 20, 32, 44, 56, 110, and 1202 layers on CIFAR-10 and show that increasing the

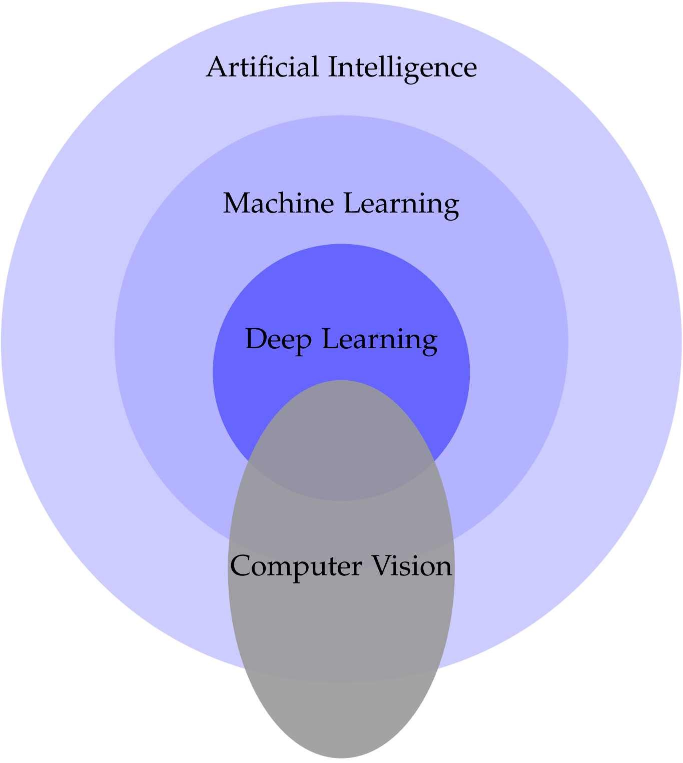


Figure 2.4: The relationship between Artificial Intelligence, Machine Learning, Deep Learning and Computer Vision – This diagram shows that Computer Vision is a generic technique, not necessarily implemented using Machine Learning.

number of layers in residual networks yields better results, different from plain networks.

Deep Residual Networks have been applied to images in several contests. Some prominent examples include Image Super-Resolution [43, 45, 49, 61, 72], which aims at up-scaling and improving low-resolution images quality, and Steganalysis [10, 80], which aims at discovering messages hidden using steganography [7].

2.1.5 *Computer Vision*

This section introduces **Computer Vision** (CV), whose aim is to study the content and the meaning of pictures or videos [51]. While this task is in general particularly easy for human beings, it is challenging for computers, since it requires abstracting complex and variable data (for instance, the task of recognizing a person in an image is made more difficult by the fact that the same individual can look different based on their facial expression).

The increasing availability of images, during the last years, has made a greater quantity of labelled data available. They are particularly important since they allow the application of Neural Networks to extract meaningful information from them. Moreover, analysing images requires the use of Deep Learning, rather than Shallow Learning. In this case, the recent development of new technologies (i.e., more powerful GPUs), combined with data availability, has contributed to the massive development of the field. The application of Neural Networks to Computer Vision is now the predominant research direction, even if not the only one. Figure 2.4 shows the relationship between Artificial Intelligence, Machine Learning, Deep Learning, and Computer Vision, emphasizing the extension of Computer Vision beyond the currently used methods.

2.2 INTERPRETABILITY AND ACTIVATION MAPS

In recent years, Deep Learning models have been treated as black-boxes, i.e. architectures that do not expose their internal operations to the user. These systems are used for various approaches and their interpretability is fundamental in many fields, especially when the outputs of the models are used for sensitive applications. Activation Maps are one of the techniques employed to explain the behaviour of neural models dealing with image data.

Section 2.2.1 presents the problem of interpretability, while Section 2.2.2 gives an overview about activation maps, discussed in further detail in Chapter 3.

2.2.1 *Interpretability*

Different from traditional algorithms, neural networks learn inductively from sets of examples. The black-box model developed as a result of the learning process, for this reason, is the result of sequences of complex operations on the input data and cannot be predicted in advance. For the same reason, understanding *why* a model works in a certain way is equally challenging. Interpretability deals with the understanding of the reasons why a model behaves in a certain way and is effective in detecting, for instance, possible biases in the model. Understanding the internal logic of the model is important also from an ethical point of view, especially for sensitive applications (e.g., medicine).

Guidotti et al. [31] surveyed the most prominent methods for explaining black-box models, considering critical applications of Machine Learning. In particular, they focused their analysis on the introduction of unconscious biases introduced by such models. In some cases, biases discriminate minorities and black people (e.g., by deeming them more likely to be repeat offenders). This is also the case of text translation. Prates et al. observed that the translation of sentences from gender-neutral languages (e.g., Hungarian) to English using Google Translate relied on assumptions based on the difference between traditionally male-dominated fields and female-dominated fields [58]. Table 2.1 presents an example, based on the observations

Table 2.1: An example of bias in language translation, based on the observations of [58]. This bias was likely introduced by the gender unbalance in the jobs in the example, and exemplifies how a black-box model can yield to biased results, making interpretability an indispensable cross field of research.

English	**Italian**	**Attributed gender**
The scientist obtained promising results.	Lo scienziato ha ottenuto risultati promettenti.	Male
The nurse is taking care of Bob.	L'infermiera si sta prendendo cura di Bob.	Female
The doctor won an important prize.	Il dottore ha vinto un premio importante.	Male
The babysitter is really loving.	La babysitter è davvero amorevole.	Female
The engineer is estimated by his colleagues.	L'ingegnere è stimato dai suoi colleghi.	Male
That student loves playing with dollies.	Quella studentessa adora giocare con le bambole.	Female
That student loves playing with trucks.	Quello studente adora giocare con i camion.	Male
The housekeeper works really well.	La governante lavora davvero bene.	Female

from Prates et al. paper [58], of this phenomenon, which was solved afterwards for translations to the English language. The table considers the results of the English to Italian translation, as of June 2021.

The topic of biases occurs also in images, and was surveyed more deeply by Buhrmester et al. [11]. The authors emphasized one of the major challenges related to interpretability, that are the trade-off between interpretability and accuracy. In general, the explainability decreases as the prediction accuracy increases, which constitute a major problem in critical applications, where a loss of accuracy is not acceptable. On the other hand, explainability is fundamental to help to understand whether a model shows good performances, for instance, because of biased data. In particular, a typical threat to Deep Learning models is the presence of adversarial examples, which are indistinguishable from the original examples for a human observer but make the model collapse. In this context, Moosafi-Dezfooli et al. [48] proposed DeepFool, which shows how some deep neural networks achieving impressive results were unstable to small perturbations of the images, even if imperceptible for human beings. Figure 2.5 illustrates a qualitative example of this phenomenon.

2.2.2 *Activation Maps*

In the literature, many techniques aim at explaining the behaviour of neural models [11, 31]. Saliency Masks are used to address the *outcome explanation problem* by providing a visualization of which part of the input data is mainly responsible for the network prediction. The most popular Saliency Masks are obtained with the Class Activation Map (CAM) approach. CAMs [88] have shown their effectiveness in

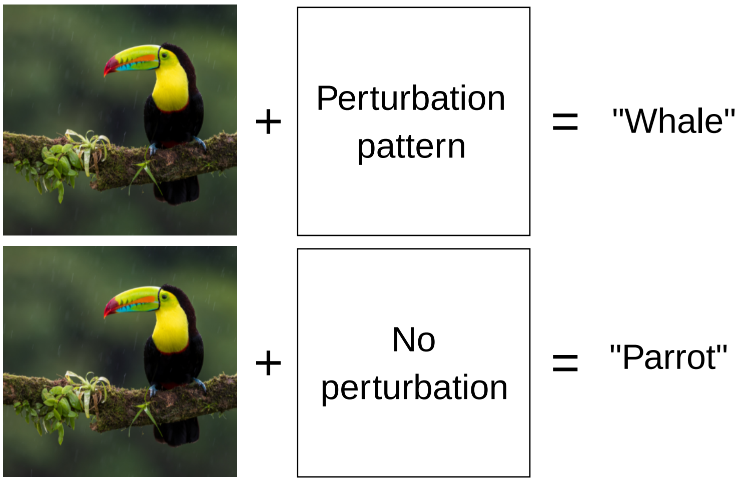


Figure 2.5: An example of adversarial perturbations – this image illustrates the effect of adversarial perturbation on images in terms of classification. The DeepFool method, presented in [48], generates a suitable perturbation.

highlighting the most discriminative areas of an image in several fields, ranging from medicine [32] to fault diagnostics [70]. The original formulation of CAMs has been subsequently improved. Selvaraju et al. [65] introduced Grad-CAM, which exploits the gradients that pass through the final convolutional layer to compute the most salient areas of the input. Chattopadhay et al. [15] introduced Grad-CAM++ which considers gradients too but is based on a different mathematical formulation that improves the localization of single and multiple instances. Smooth Grad-CAM++ [52] applies Grad-CAM++ iteratively on the combination of the original image and Gaussian noise.

The use of CAMs is not limited to the explainability of Deep Learning classification models but is the starting point for studies related to the weakly supervised localization of content inside the images [85]. CAMs have been also employed in several fields, including art [71, 81], food segmentation [77], and medicine [50].

2.3 AUTOMATED ARTWORK IMAGE ANALYSIS

The large availability of artworks in digital format has allowed researchers to perform automated analysis in the fields of digital humanities and cultural heritage using Computer Vision and Deep Learning methods. Several data sets containing various types of artworks have been proposed to support such studies [9, 22, 28, 36, 38, 46, 47, 69].

The performed analyses span several classification tasks and techniques: from style classification to artist identification, comprising also medium, school, and year classification [14, 64, 87]. These researches are useful to support cultural heritage studies and asset management, e.g., automatic cataloguing of unlabeled works in online and museum collections, but their results can be exploited for more complex applications, such as authentication, stylometry [26], and forgery detection [25].

A task that is more related to our proposal is artwork content analysis, which focuses on the automatic identification and, if possible, localization of objects inside artworks. The literature contains several state-of-the-art approaches [5, 13, 21, 28, 34, 47, 67]. Since there is an abundance of deep learning models trained with natural images but a deficiency of art-specific models, many studies focus on the transferability of previous knowledge to the art domain [5, 8, 19, 29, 47]. This approach is known as Transfer Learning and consists of fine-tuning a network, previously trained with natural images, using art images. The consensus is that Transfer Learning is beneficial for tasks related to artworks analysis.

The next sections present the main contributions in the field of automated artwork image analysis in greater detail, focusing on inexact and inaccurate supervision (Section 2.3.1), style recognition (Section 2.3.2), object retrieval (Section 2.3.3), interpretability (Section 2.3.4) and prominent uses of Class Activation Maps (Section 2.3.5).

2.3.1 *Inexact and inaccurate supervision in art*

This section focuses on the importance of inexact and inaccurate supervision in the context of artwork analysis.

Inexact supervision is the most interesting sub-field of weakly supervised learning in the case of Christian paintings. Different from other scenarios, Christian art is characterized by iconographical symbols associated with saints, hence labels can be defined hierarchically. In particular, the saints' labels are coarse-grained if compared with the symbols' labels, and the supervision may be limited to the saints' labels, to automatically find the symbols associated with the saints without indicating *which* symbol has been identified. In this way, a significantly lower amount of annotations is required.

Inaccurate supervision is also interesting in the case of artworks since the labels of the ArtDL data set were generated automatically [47], starting from basic information related to the painting (e.g., the title). For example, Figure 2.6 represents Saint Peter Martyr, characterized by a knife on the top of his head, differently from Saint Peter the Apostle.

In the ArtDL dataset, similarly to the situation presented in Figure 2.1, the same image may contain a variable number of objects, possibly belonging to the same class.

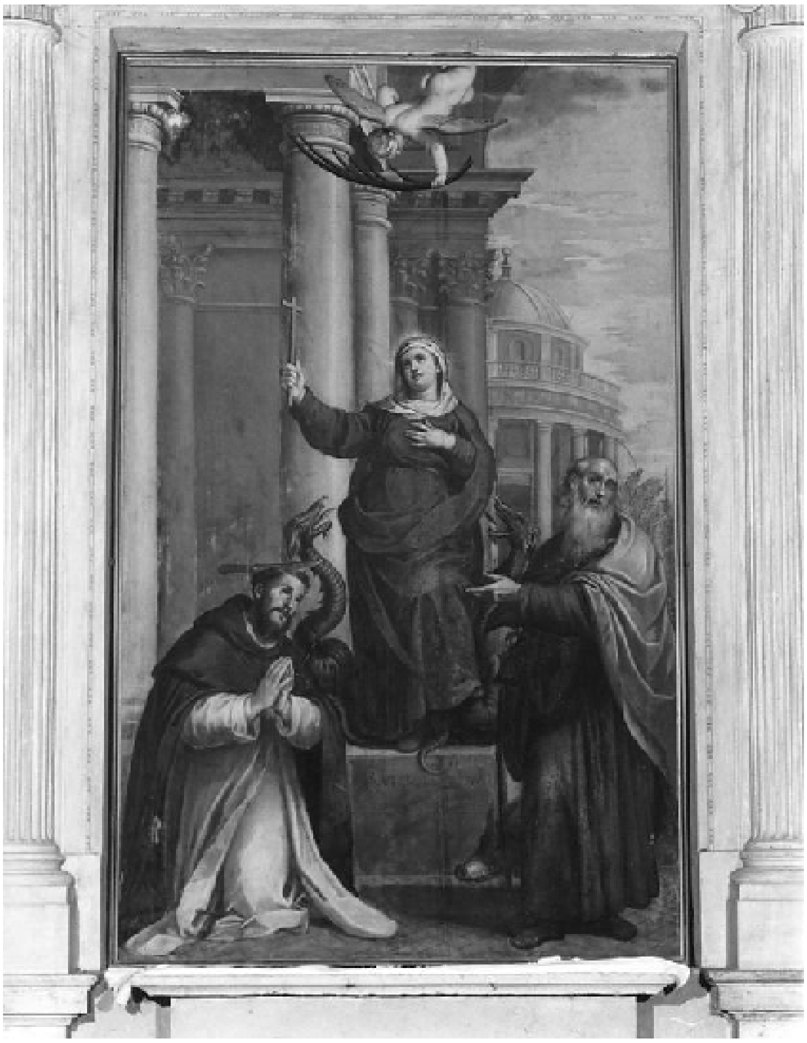
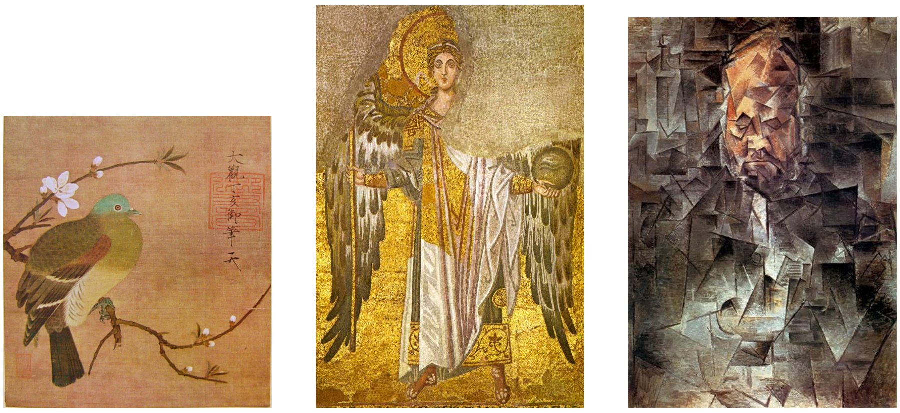


Figure 2.6: An example of inaccurate label – Saint Peter Martyr is different than Saint Peter the Apostle, and is characterized by a knife on the top of his head. During the automatic labelling process in the creation of the ArtDL data set [47], it was incorrectly assigned the "Saint Peter" label.



(a) Gongbi style[2] **(b)** Byzantine Art[3] **(c)** Cubism[4]

Figure 2.7: Examples of different styles from the WikiArt data set.

2.3.2 *Image style recognition*

One of the most common tasks in automated artwork image analysis is style classification [9, 36, 46], which consists of recognizing the style of a given painting. As highlighted by Karayev et al. [36], it is difficult to define visual style rigorously, even if recognizing different styles is an easy task for human beings. For this reason, it is also challenging to define different styles, which do not only characterize artworks but also photography, which may be considered a form of art as well. In particular, the work from Karayev et al. is an example of fully supervised learning, since both the WikiArt (formerly known as WikiPaintings) and the Flickr Style data sets contained labelled images. Figure 2.7 shows different paintings from the WikiArt data set and shows some of the diverse styles present in the data set.

To tackle this task, Karayev et al. implemented the Stochastic Gradient Descent method with adaptive subgradient and proposed the One

[1] Pigeon on a Peach Branch, Emperor Huizong, 1108.
[2] Angel Gabriel, nd, c. 867.
[3] Portrait of Ambroise Vollard, Pablo Picasso, 1910.

vs. All reduction to binary classifier to perform multi-class classification (i.e., when an image is described by more than one label).

Considering the results on a subset of the WikiArt data set comprising 85,000 images labelled with 25 different art styles, they obtain per-class accuracies ranging from 72% to 94% and show that their method can be used for performing style-based image search.

A more recent research by Mao et al. [46], instead, proposed the DeepArt framework, whose aim is to capture contents and styles of visual arts. Different from the contribution by Karayev et al., Mao et al. propose Art500k, a new data set including also WikiArt. The categories of Art500k allow the subdivision of artworks by artist, genre (e.g., interior, portrait, landscape), medium and art movement (e.g., Cubism, Realism, Expressionism). Even if this method improves previous results on the same data set in terms of the art movement and genre identification, a comparison with the WikiArt data set is missing, consequently, the results are not directly comparable with the ones proposed by Karayev et al. The abundance of different data sets is a typical characteristic of works focusing on automated artworks analysis and does not concern only the style recognition problem. For example, Khan et al. introduced Painting-91 [38] in 2014, while recent data sets in the Christian paintings sub-field were proposed by Gonthier et al. [28] and Milani and Fraternali [47].

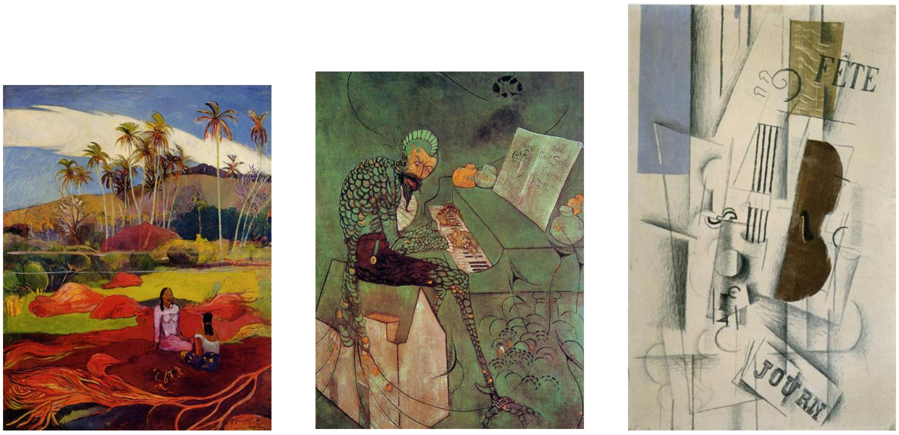
Similarly to Mao et al. [46], Bianco et al. proposed the use of a multi-task formulation for performing artist, style, and genre categorization [9], introducing a new data set, MultitaskPainting100k, based on WikiArt. Moreover, they applied state-of-the-art methods and their method to the Art500k data set introduced by Mao et al., showing advancement in the state of the art. This method is particularly interesting because it uses residual blocks, which have proven effective also in the work by Milani and Fraternali on Christian paintings classification [47].

Approaches similar to the ones used for style recognition were used for tackling other tasks. For instance, Strezoski and Worring [69] proposed a multi-task learning approach which, starting from a learned shared representation, was able to perform artist attribution, type prediction (e.g., painting, print, photograph), material prediction, and period estimation. Since WikiArt was insufficient for performing the required analysis, the authors introduced the OmniArt data set, which instead included chiefly artworks from the Rijksmuseum collection, the collection from the Met and the Web Gallery of Art collection.

2.3.3 *Object retrieval in art images*

Object retrieval, in general, consists of locating the object of research (e.g., an inanimate object, an animal or a person), typically inside an image. In the context of artwork analysis, the retrieval task can be performed, for example, in the photograph of a painting or of a sculpture. In the specific sub-field of Christian art, it is possible to establish a hierarchy of semantically interconnected objects. For instance, a Christian saint can be regarded as an object, and the iconographical symbols associated with them are additional objects dependent on the presence of the saint (e.g., the presence of a lion depends on the presence of Saint Jerome).

[4] Tahitian women under the palms, Paul Gauguin, 1892.
[5] The Gold Scab, James McNeill Whistler, 1879.
[6] Violin and Newspaper (Musical Forms), Georges Braque, c. 1912.



(a) High contrast[5] (b) Range of colours[6] (c) Innatural shapes[7]

Figure 2.8: Typical challenges faced in paintings analysis.

One of the main challenges in dealing with paintings is the difference in the depiction of paintings and photographs, since the first show, for instance, higher contrast, a more limited range of colours and, depending on the style, unnatural shapes. Figure 2.8 presents three examples of those challenges. Crowley and Zisserman [22] acknowledged this problem, but still recognized similarities between natural images and artworks. For this reason, they proposed the application of Transfer Learning on a network pre-trained on the PASCAL VOC natural images data set. For evaluating the images, they relied on the "Your Paintings" data set, comprising 210,000 medium-resolution oil paintings and now part of Art UK.[8] Moreover, they measured the spatial consistency between the objects in the natural images and the ones in the paintings with the purpose of re-ranking paintings with high classification scores. This approach, however, does not guarantee that an object in a painting can be matched consistently with an object in a natural image, since in the case of transformations (e.g., rotations), parts of the object may be hidden.

[8] https://artuk.org/ (As of June 2021).

Different from Crowley and Zisserman, Gonthier et al. proposed a novel approach [28], based on IconArt, a novel Christian Art data set, which makes this research particularly relevant for this thesis. Different from previous researches, they proposed a weakly supervised approach for detecting objects in the paintings, relying only on image-level labels rather than on detailed labels. Their purpose is the detection of iconographic elements in paintings, in addition to the main subjects (i.e., the saints). The need of relying on an unlabelled data set derives from the absence of a fine-grain labelled artworks data set, which makes this field different from the one of natural images. Moreover, the necessity of defining a new data set, rather than relying on the existing ones, stems from the specific sub-field of their research, similarly to Crowley and Zisserman researched on weakly-supervised learning applied to ancient Greek Gods and animals in pottery [21], where they defined a smaller data set in the bigger Beazley Art data set.[9] Christian paintings are characterized by symbols distinctly associated with an iconographical meaning. In other genres, instead, paintings may still be associated with symbolical meanings, but the associations between objects and symbolic meanings are more arbitrary, and for this reason, cannot be the object of comprehensive studies. The IconArt data set, indeed, includes labels referring to, for instance, "ruins" and "nudity," which are not peculiar to Christian paintings, together with peculiar labels, such as "Saint Sebastian" and "Jesus." In this data set, each image contains a variable number of labels, so Gothier et al. introduced new multiple-instance learning (MIL) technique. Their workflow consists of four main steps:

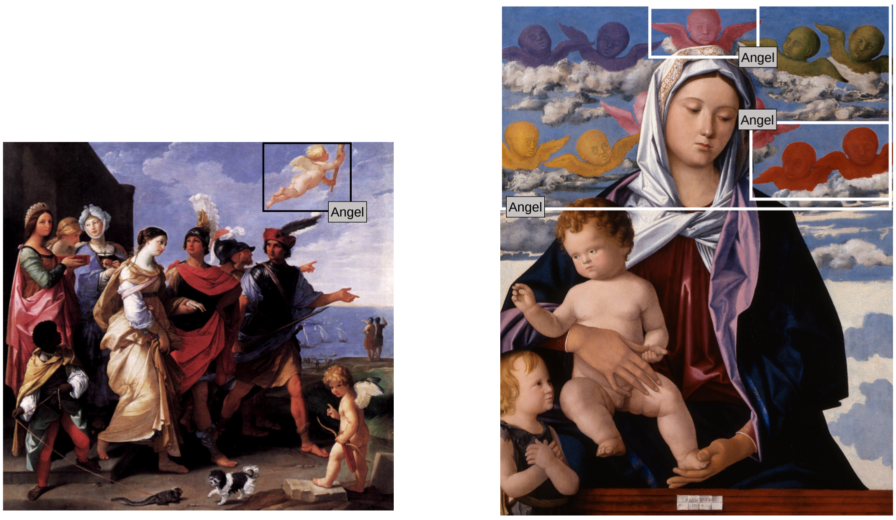
[9] https://www.beazley.ox.ac.uk/carc (As of June 2021).

- Application of Faster R-CNN as a feature extractor, extracting candidate bounding boxes, which are initially class-agnostic (i.e., a bounding box is not associated with a specific class);
- Given an image and a visual category (e.g., "angel"), the label associated with that category is $+1$ if the image contains the visual category and -1 otherwise;
- If an image contains a category and a set of candidate bounding boxes, it is possible to hypothesize that at least one bounding box is associated with the category, so the goal is finding this bounding box;
- Given an image, a set of bounding boxes proposals and a visual category, the authors apply gradient descent to find a "hyperplane separating the most positive element of each positive image from the least negative element of the negative image."

The proposed approach has the advantage of requiring only image-level annotations. At the same time, it relies on the strong hypothesis that an image containing a certain category must also contain a class-agnostic bounding box that can be associated with that category. This proposal, therefore, needs robust initially generated bounding boxes.

The evaluation process relies on partially overlapping labels, which make it particularly challenging. For instance, an angel is likely characterized by the labels "angel" and "nudity," which likely cover the entire figure or a relevant part of it. The initial selection of the candidate bounding boxes, however, undergoes a filtering process, which keeps only the most relevant ones in a given area. For this

[9] Rapimento di Elena, Guido Reni, 1631.
[10] Madonna col Bambino, Giovanni Bellini, c. 1490-1500.



(a) Single angel[10]

(b) Multiple angels[11]

Figure 2.9: Qualitative representation of the failure in detecting "angel" and "nudity", based on the results from [28].

reason, the presence of two nearly coincident bounding boxes is discouraged and yields to poor performances for some classes and high inter-class variability. Two examples of this phenomenon can be observed in Figure 2.9, based on the results presented in [28].

A promising research was proposed by Milani and Fraternali, that showed that the ResNet50 architecture is effective in classifying artworks in the ArtDL data set [47]. Their research lays the foundations of this thesis.

2.3.4 *Interpretability in art images*

This section presents the challenges and implications concerning interpretability in art images.

Section 2.2.1 presented researches concerning the importance of interpretability for tackling the ethical consequences of the use of Machine Learning. In the context of art, ethical considerations are negligible, but interpretability is crucial for understanding which parts of an image are the most prominent for determining the outcome of classification [47].

The introduction of biases is also one of the current challenges in Machine Learning. In Christian paintings analysis, biases are also a potential problem. From a Machine Learning point of view, the reasons are likely similar to the ones introducing biases penalizing black people and associating genders to professions. Some Saints, indeed, appear more often in Christian paintings, which makes data sets such as ArtDL [47] unbalanced and prone to biases.

A related field is the one of counterfeiting. Different from critical fields (e.g., security and health), counterfeit pictures of paintings are not considered as a threat for artworks analysis. Moreover, this thesis shows that promising classification results correspond to results coherent with Art History studies. Christian iconography, indeed, has been described clearly and, for the saints under analysis, followed strictly. A model, then, is expected to focus on well-known symbols. Other fields (e.g., astrophysics and medicine), on the other hand, may hide details currently unknown to the experts in the field, and a network may discover novel patterns. In such a case, it would be challenging to understand if the new patterns correspond to discoveries (e.g., scientific discoveries) or the inadequacy of the model.

2.3.5 *Using Class Activation Maps in artwork analysis*

This section presents some prominent uses of Class Activation Maps in artwork analysis.

Yang and Min [81] presented an approach based on CNNs for classifying the artistic media (e.g., pencil, pastel, etc.) used in artworks.

In their research, CAMs are used to identify the most prominent areas for determining the artistic medium, showing similar performance and recognition pattern with human, emphasizing that CAMs, in this case, focus on regions deemed relevant for a human classifier as well. Surapaneni et al. [71], employed Grad-CAM for exploring gender biases in artworks. This research is closer to the topic of this thesis since Class Activation Maps are used to identify the most relevant parts of an image, given a pre-trained model. The goal, instead, is opposite: while the objective of Christian artworks analysis is finding the saints and their symbols, exploiting a vast data set in which such symbols are repeated, Surapaneni et al. used Class Activation Maps to show that the repetition of some characteristics (e.g., long hair) leads to misclassifying artworks (e.g., concerning the gender), similarly to the case of biased translations presented in Table 2.1. Figure 2.10[12] shows an example of image candidate to misprediction caused by gender biases, similar to the one presented in [71].

[12] Image by Animesh Bhattarai on Unsplash (https://unsplash.com/photos/4ZuwfRD3c68).

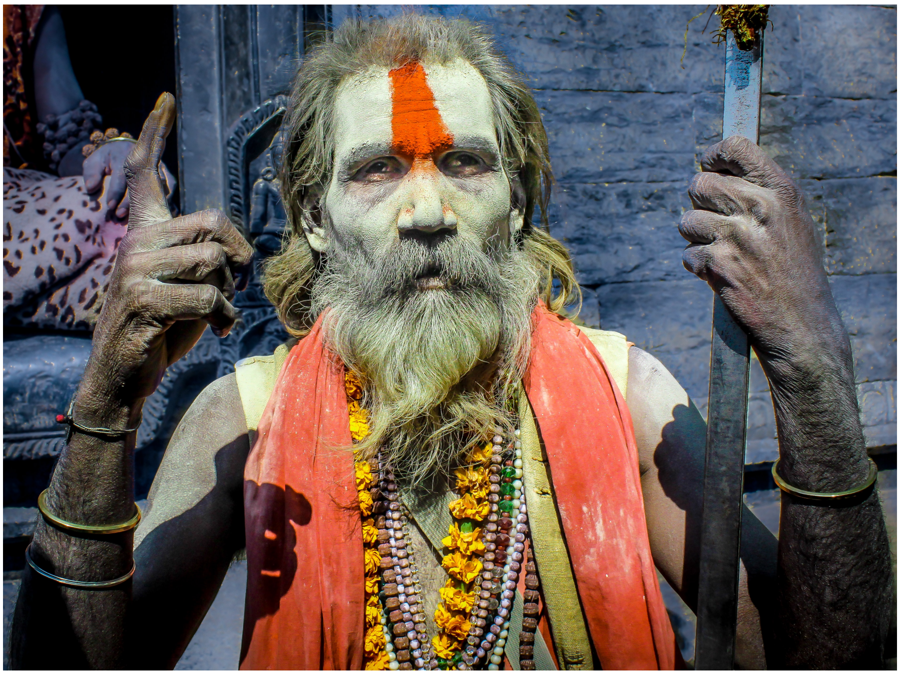


Figure 2.10: Gender misprediction – This example is similar to the one presented in [71], since this man has characteristics traditionally associated with women in some cultures (i.e., a necklace, long hair and long clothes). Such characteristics may confuse a biased model.

3

CLASS ACTIVATION MAPS FOR ICONOGRAPHY CLASSIFICATION

This thesis compares different CAM algorithms: Grad-CAM, Grad-CAM++, and Smooth Grad-CAM++. Their implementation is based on the mathematical definitions provided, respectively, by [88], [65], [15], and [52].

Figure 3.1 shows the ResNet50 classifier architecture used to compute the class activation maps. The input of the network is an image and the output is the set of probabilities associated with the different classes. In the evaluation, the input images portray artworks and the output classes denote 10 Christian Saints. ResNet50 contains an initial convolutional layer (`conv1`) followed by a sequence of convolutional residual blocks (`conv2_x` ... `conv5_x`). A Global Average Pooling (GAP) module computes the average value for each feature map obtained as an output of the last layer (`conv5_x`). The probability estimates are computed by the last component, which is typically a Fully Connected (FC) layer [44].

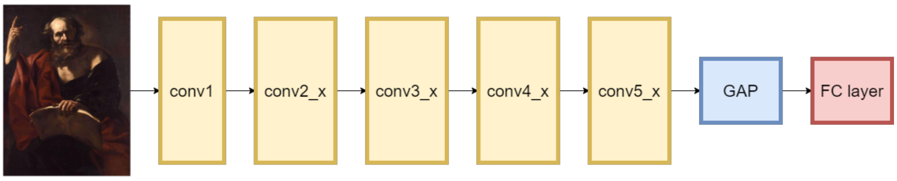


Figure 3.1: The ResNet50 architecture.

This chapter presents the four algorithms analyzed in this thesis. Section 3.1 introduces CAM; Section 3.2 introduces Grad-CAM; Sec-

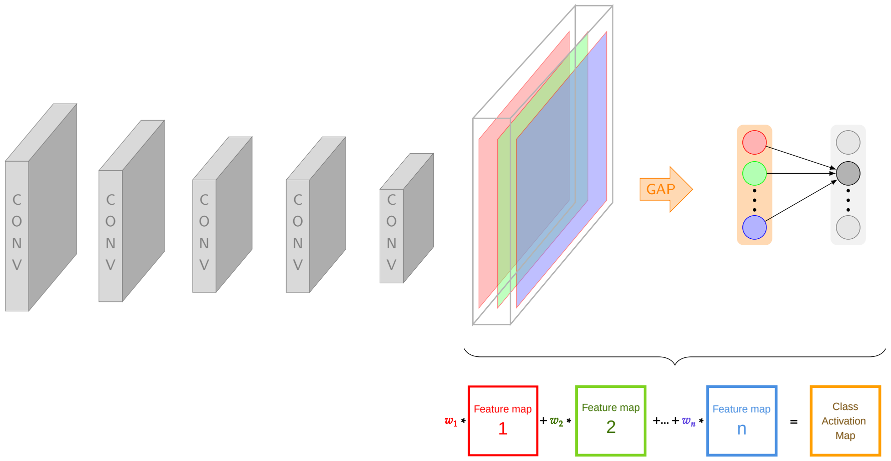


Figure 3.2: CAM structure – This figure presents an example of how CAM is calculated for an input image, suggesting that the most relevant areas are associated with higher values.

tion 3.3 introduces Grad-CAM++ and Section 3.4 introduces Smooth Grad-CAM++.

3.1 CAM

CAMs [88] are based on the use of GAP, which has been demonstrated to have remarkable localization abilities [60]. The GAP operation averages the feature maps of the last convolutional layer and feeds the obtained values to the final fully connected layer that performs the actual classification.

Class activation maps are generated by performing a weighted sum of the feature maps of the last convolutional layer for each class. Figure 3.2 presents CAM structure more precisely, and shows that the most prominent areas of the image correspond to higher values in the activation map.

Before introducing a compact definition of CAM, it is necessary to introduce the following quantities, defined on an input image[1]:

- c represents a class;
- (x, y) represents a spatial location in the input image (i.e., the position of a pixel);
- k is a unit in the last convolutional layer of the network;
- $A_k(x, y)$ represents the activation of unit k in the last convolutional layer at (x, y);
- $F_k = \sum_{x,y} A_k(x, y)$ is the result of the application of global average pooling;
- w_k^c is the weight associated with class c for unit k and indicates the importance of F^K for class c;
- $S^c = \sum_k w_k^c F_k$ is the input to softmax;
- $P^c = \frac{\exp(S^c)}{\sum_c \exp(S^c)}$ is the output of softmax for class c, assuming a bias term of zero;

Consequently, S^c can be formulated as:

$$S^c = \sum_k w_k^c \sum_{x,y} A_k(x, y) = \sum_{x,y} \sum_k w_k^c A_k(x, y) \tag{3.1}$$

[1] In the original paper, A_k is indicated as f_k. Here, A_k is chosen to keep a consistent notation across the analyzed methods.

The actual class activation map value $M^c(x, y)$ for a class c and a position x, y in the input image, then, is expressed as follows:

$$M^c(x, y) = \sum_k w_k^c A_k(x, y) \tag{3.2}$$

where $A_k(x, y)$ is the activation value of feature map k in the last convolutional layer at position (x, y), and w_k^c is the weight associated with feature map k and with class c.

By exploiting the definition of M^c, S^c can be rewritten as:

$$S^c = \sum_{x,y} M^c(x, y) \tag{3.3}$$

Intuitively, a high CAM value at position x, y is the result of an average high activation value of all the feature maps of the last convolution layer.

Differently from the original approach, we compute the CAM output not only for the predominant class but for all the classes. The ArtDL data set contains multi-class multi-label images and this formulation allows us to analyze which regions of the artwork are associated with which classes, also in the case of incorrect classification.

3.2 GRAD-CAM

Grad-CAM [65] is a variant of CAM which considers not only the weights but also the gradients flowing into the last convolution layer. In this way, the layers preceding the last one also contribute to the activation map. An advantage of using gradients is that Grad-CAM can be applied to any layer of the network. Still, the last one is especially relevant for the localization of the parts of the image that contribute most to the final prediction. Furthermore, the layer used as input for the prediction can be followed by any module and not only by a fully connected layer. Grad-CAM exploits the parameters α_k^c, which represents the *neuron importance weights* and are calculated as:

$$\alpha_k^c = \frac{1}{Z}\sum_{x,y} \frac{\partial S^c}{\partial A_k(x,y)} \tag{3.4}$$

where $\frac{1}{Z}\sum_{x,y}$ denotes the global average pooling operation ($Z = \sum_{x,y} 1$) and $\frac{\partial S^c}{\partial A_k(x,y)}$ denotes the back-propagation gradients. In the gradient expression, S^c is the score of the class c and A_k represents the k-th feature map. The Grad-CAM for a class c at position (x,y) is then given by:

$$M_{Grad-CAM}^c(x,y) = \text{ReLU}\left(\sum_k \alpha_k^c A_k(x,y)\right) \tag{3.5}$$

where the ReLU operator maps the negative values to zero. As in the case of CAM, we compute the output of Grad-CAM for all the classes under analysis.

It is possible to show that Grad-CAM is a generalization of CAM. Considering a set of feature maps A_k, indexed by the positions x and y, the score associated with a class c is defined as:

$$S^c = \sum_k w_k^c \cdot \frac{1}{Z}\sum_x \sum_y A_k(x,y) \tag{3.6}$$

Hence, the result of the application of GAP, F_k, is defined as:

$$F_k = \frac{1}{Z} \sum_x \sum_y A_k(x, y) \tag{3.7}$$

Considering now the gradient of the score S^c with respect to the feature map F_k, it results:

$$\frac{\partial S^c}{\partial F_k} = \frac{\frac{\partial S^c}{\partial A_k(x,y)}}{\frac{\partial F_k}{\partial A_k(x,y)}} = \frac{\partial S^c}{\partial A_k(x, y)} \cdot Z \tag{3.8}$$

Hence, the weight w_k^c is:

$$w_k^c = Z \cdot \frac{\partial S^c}{\partial A_k(x, y)} = \sum_x \sum_y \frac{\partial S^c}{\partial A_k(x, y)} \tag{3.9}$$

Hence, up to a constant term $\frac{1}{Z}$, normalized during the visualization, w_k^c is identical to α_k^c, and GradCAM is a strict generalization of CAM.

3.3 GRAD-CAM++

Grad-CAM++ [15] is a generalization of Grad-CAM aimed at better localizing multiple class instances and at capturing objects more completely. Differently from Grad-CAM, Grad-CAM++ applies a weighted average of the partial derivatives, to cover a wider portion of the object and of better detecting multiple occurrences of the same object. In particular, the authors propose to define the weights w_k^c by calculating the pixel-wise gradients weighted average, where α_k^c is the weighting factor:

$$w_k^c = \sum_{x,y} \alpha_k^c(x,y) \cdot \text{ReLU}\left(\frac{\partial S^c}{\partial A_k(x,y)}\right) \tag{3.10}$$

where $A_k(x,y)$ is the activation map calculated in the last convolutional layer as in the cases of Grad-CAM and Grad-CAM++.

In particular, the weighting factor is defined as:

$$\alpha_k^c(x,y) = \begin{cases} \frac{1}{\sum_{l,m} \frac{\partial S^c}{\partial A_k(l,m)}} & \text{if} \quad \frac{\partial S^c}{\partial A_k(x,y)} = 1 \\ 0 & \text{otherwise} \end{cases} \tag{3.11}$$

and highlights all the objects of a certain class giving them the same importance.

Given a class c, the score S^c is calculated based on α_k^c:

$$S^c = \sum_k \left[\sum_{a,b} \alpha_k^c(a,b) \cdot \text{ReLU}\left(\frac{\partial S^c}{\partial A_k(a,b)}\right)\right] \sum_{x,y} A_k(x,y) \tag{3.12}$$

where (a,b) and (x,y) are positions in A_k, over which they iterate.

Consequently, it holds that:

$$\frac{\partial^2 S^c}{(\partial A_k(x,y))^2} = 2\alpha_k^c(x,y) \frac{\partial^2 S^c}{(\partial A_k(x,y))^2} + \sum_{a,b} A_k(a,b) \left[\alpha_k^c(x,y) \frac{\partial^3 S^c}{(\partial A_k(x,y))^3}\right] \tag{3.13}$$

The parameter $\alpha_k^c(x,y)$ can be rewritten as follows:

$$\alpha_k^c(x,y) = \frac{\frac{\partial^2 S^c}{(\partial A_k(x,y))^2}}{2\frac{\partial^2 S^c}{(\partial A_k(x,y))^2} + [\sum_{a,b} A_k(a,b)] \frac{\partial^3 S^c}{(\partial A_k(x,y))^3}} \tag{3.14}$$

The parameter w_k^c, therefore, can be defined as:

$$w_k^c = \sum_x \sum_y \alpha_k^c(x,y) \mathrm{ReLU}\left(\frac{\partial S^c}{\partial A_k(x,y)}\right) \tag{3.15}$$

which leads to

$$w_k^c = \sum_{x,y} \left\{ \frac{\frac{\partial^2 S^c}{(\partial A_k(x,y))^2}}{2\frac{\partial^2 S^c}{(\partial A_k(x,y))^2} + \left[\sum_{a,b} A_k(a,b)\right] \frac{\partial^3 S^c}{(\partial A_k(x,y))^3}} \right\} \mathrm{ReLU}\left(\frac{\partial S^c}{\partial A_k(x,y)}\right) \tag{3.16}$$

As in the other CAMs, it holds that

$$M^c_{Grad-CAM++}(x,y) = \mathrm{ReLU}\left(\sum_k w_k^c A_k(x,y)\right) \tag{3.17}$$

3.4 SMOOTH GRAD-CAM++

Smooth Grad-CAM++ [52] is a variant of Grad-CAM++ that can focus on subsets of feature maps or of neurons for identifying anomalous activations. Smooth Grad-CAM++ applies a random Gaussian perturbation on the image z and exploits the visual sharpening of the class activation maps by averaging random samples taken from a feature map close to the input. The value of the activation map M^c in a position (x, y) is defined as:

$$M^c_{SGCpp}(x, y, z) = \frac{1}{n} \sum_{1}^{n} M^c_{GCpp}(x, y, z + \mathcal{N}(0, \sigma^2)) \tag{3.18}$$

where, for an image z, $M^c_{GCpp}(x, y, z) = M^c_{Grad-CAM++}(x, y)$. where n is the number of samples, $\mathcal{N}(0, \sigma^2)$ is the 0-mean Gaussian noise with standard deviation σ, and M^c_{SGCpp} is the activation map for the input $z + \mathcal{N}(0, \sigma^2)$. The final result is obtained by iterating the computation of Grad-CAM++ on inputs resulting from the overlap of the original image and random Gaussian noise.

Different from the previously analyzed approaches, Smooth Grad-CAM++ allows regulating the additional hyper-parameters σ and n. However, the original study did not focus on the comparison of different parameter settings. In this research, instead, different combinations are analyzed in the context of Christian artworks.

4

EVALUATION

This chapter addresses research activities related to evaluation of different CAM algorithms, adapting and extending the content presented in [57]. The evaluation exploits the ArtDL data set [47], an existing artwork collection annotated with image-level labels. The purpose of the evaluation is:

1. To understand whether the class activation maps effectively localize both the overall representation of an iconography class and the distinct symbols that characterize it.[1]

2. To compare CAM algorithms in their ability to do so. A subset of the images has been annotated with bounding boxes framing iconography symbols associated with each saint to evaluate the localization ability of class activation maps. Figure 4.1 illustrates the symbols in a painting of Saint Jerome.

This chapter is organized as follows: Section 4.1 presents the evaluation protocol adopted in this thesis, focusing on the data set and the experimental procedures; Section 4.2 presents the quantitative assessment of class activation map algorithms, while Section 4.3 concerns the qualitative analysis.

[1] The attributes associated with the classes present in the ArtDL data set are illustrated in [42] and listed in [79].

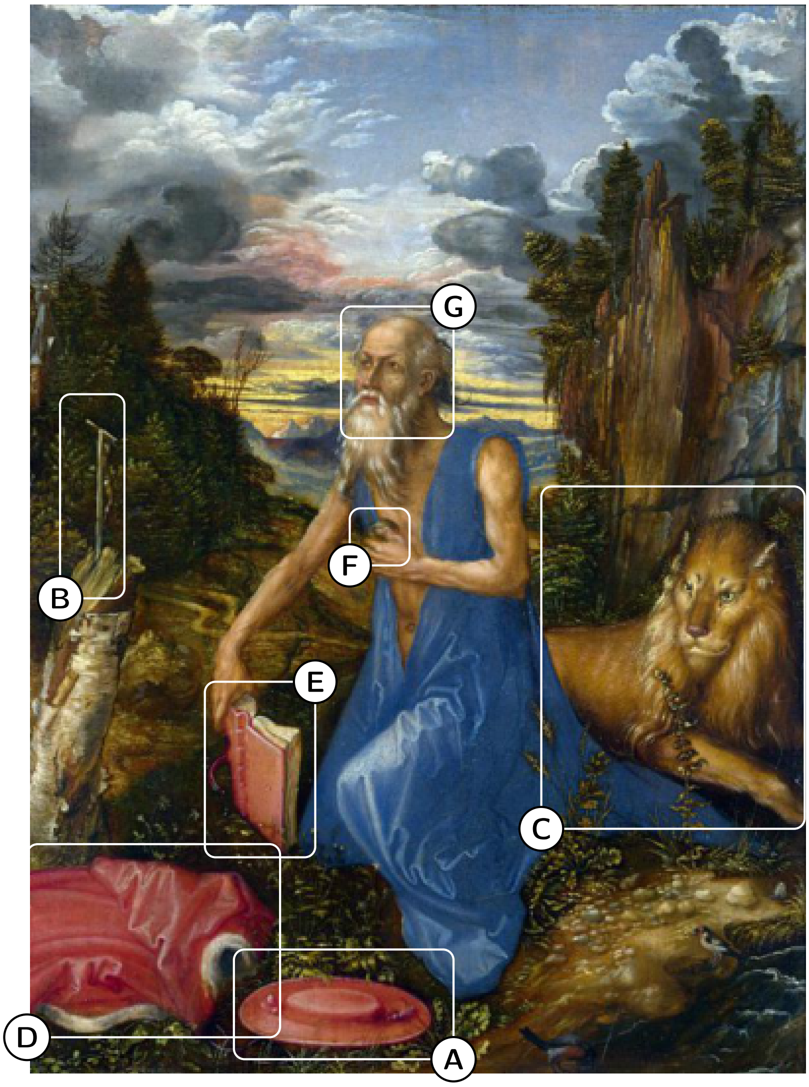


Figure 4.1: Saint Jerome – The cardinal's galero (A), the crucifix (B), the lion (C), the cardinal's vest (D), the book (E), the stone in the hand (H), and the face (G).

4.1 EVALUATION PROTOCOL

This section introduces the evaluation protocol adopted in this work. In particular, Section 4.1.1 presents the ArtDL data set, one of the contributions of this thesis, and Section 4.1.2 presents the experimental procedures used to evaluate the results, both quantitatively and qualitatively.

4.1.1 *Data set*

This section introduces the ArtDL data set and the definition of symbol-level and saint-level annotations. The ArtDL data set [47] comprises images of paintings that represent the Iconclass [20] categories of 10 Christian Saints: Saint Dominic, Saint Francis of Assisi, Saint Jerome, Saint John the Baptist, Saint Anthony of Padua, Saint Mary Magdalene, Saint Paul, Saint Peter, Saint Sebastian, and the Virgin Mary. In particular, out of the whole data set, 823 sample images were selected and manually annotated with bounding boxes that frame each symbol separately. The representation of such classes in Christian art paintings exploits specific *symbols*, i.e., markers that hint at the identity of the portrayed character. Table 4.1 presents the symbols associated with the 10 Iconclass categories represented in the ArtDL data set.

4.1.1.1 *The annotation process*

This section presents the annotation process of the test set in more detail, focusing first on the saints-level annotations and then on the symbol-level annotations, showing that this process is tedious, different from the automatic generation of bounding boxes. The annotations'

Table 4.1: Iconclass categories and symbols associated with them.

Iconclass category	Symbols
Anthony of Padua	Baby Jesus, bread, book, lily, face, cloth
Dominic	Rosary, star, dog with a torch, face, cloth
Francis of Assisi	Franciscan cloth, wolf, birds, fish, skull, stigmata, face, cloth
Jerome	Hermitage, lion, cardinal's galero, cardinal vest, cross, skull, book, writing material, stone in hand, face, cloth
John the Baptist	Lamb, head on platter, animal skin, pointing at Christ, pointing at lamb, cross, face, cloth
Mary Magdalene	Ointment jar, long hair, washing Christ's feet, skull, crucifix, red egg, face, cloth
Paul	Sword, book, scroll, horse, beard, balding head, face, cloth
Peter	Keys, boat, fish, rooster, pallium, papal vest, inverted cross, book, scroll, bushy beard, bushy hair, face, cloth
Sebastian	Arrows, crown, face, cloth
Virgin Mary	Baby Jesus, rose, lily, heart, seven swords, crown of stars, serpent, rosary, blue robe, sun and moon, face, cloth, crown

creation relied on ODIN [73], recently proposed by Torres et al. ODIN is a flexible tool, and allows creating annotations in a standard format, MS COCO, used in this research. ODIN gives the possibility of annotating images at different levels of detail, possibly relying on previous training outcomes. Additionally, this tool is compatible with *Jupyter notebook*, the environment used for developing this research's analyses, and allows the creation of annotations collaboratively.

SAINT-LEVEL ANNOTATION: This section introduces saint level annotations, focusing on the principal challenges in the annotation process.

The manual annotation of saints required a thorough analysis of artworks, focusing on the most prominent symbols associated with the saints, even for a limited subset of characters. The study of the artworks with the lowest quality (e.g., where the saints were exceedingly small compared to the overall artwork or when the picture's quality was low) presented additional challenges. In this situation,

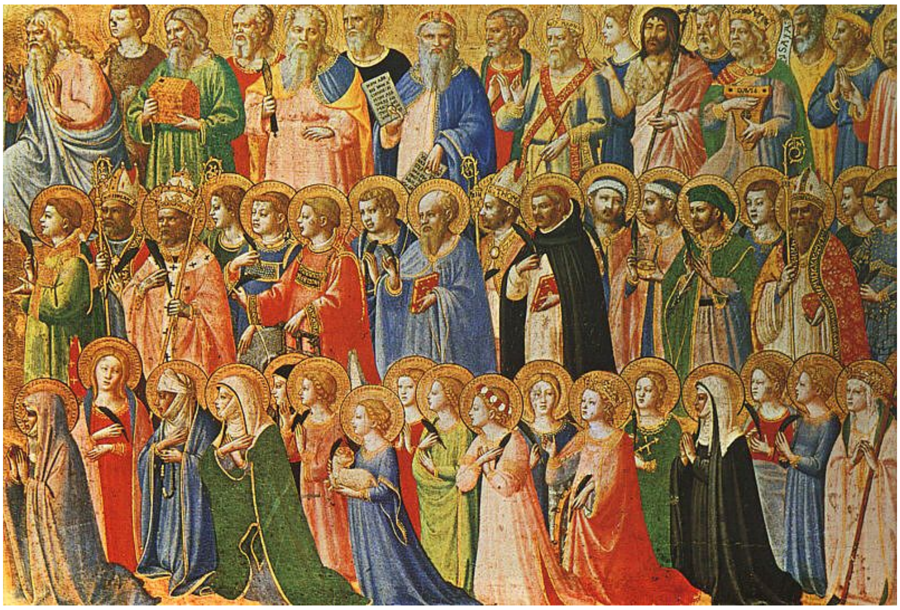


Figure 4.2: A multitude of saints – This artwork presents 50 characters, among whom there are Christian saints. Creating manual annotations according to the proposed workflow would require considerable human effort, time and expertise. Creating symbol-level annotations would require a considerable additional effort.

it was more laborious to identify the symbols and, consequently, to distinguish saints with similar characteristics (e.g., face or clothes).

Annotating an image consists of the following phases:

- Identification of (some) relevant symbols;
- Identification of symbols unique to a specific saint;
- Identification of the character associated with the symbols (an artwork may contain multiple saints, usually located close to their most relevant symbols);
- Creation of a bounding box surrounding the saint figure and possibly attached symbols (e.g., the ointment jar of Saint Mary Magdalene);

The proposed workflow allows creating fewer bounding boxes than the symbol-level annotations workflow but is tedious and time-

Table 4.2: Saint-level bounding boxes distribution – This table presents the number of saint-level bounding boxes associated with each saint. In particular, the most frequent saints are in general associated with a higher number of bounding boxes.

Saint	Saint bounding boxes
Anthony of Padua	26
Dominic	30
Francis of Assisi	85
Jerome	136
John the Baptist	82
Mary Magdalene	66
Paul	34
Peter	85
Sebastian	49
Virgin Mary	289

consuming. While this data set was limited to 10 saints, Christian artworks include hundreds of saints associated with symbols. A human annotator should understand, given a set of symbols, the associated saint. Figure 4.2[2] contains 50 figures, each associated with symbols, and is an example of this challenge. Table 4.2 presents the saint-level bounding boxes distribution.

SYMBOL-LEVEL ANNOTATIONS: This section introduces symbol-level annotations, accentuating why they are necessary to conduct accurate analyses and the related challenges.

The ArtDL images are associated with high-level annotations specifying which Iconclass categories appear in them (from a minimum of 1 to a maximum of 7). Whole-image labels are not sufficient to assess the different ways in which the class activation maps methods focus on the image content. For this purpose, it is necessary to annotate

[2] Predella of the Saint Domenico Altarpiece, Beato Angelico, 1423.

the data set with bounding boxes that localize the symbols listed in Table 4.1.

In particular, a symbol:

- Can be included completely within a single bounding box (e.g., Saint Jerome's lion);
- Can be split into multiple bounding boxes (e.g., Saint Peter's bushy hair, usually divided into two parts separated by the forehead).

A symbol representation is the union of all the bounding boxes annotated with the same symbol label. For instance, Saint Sebastian's arrows correspond to a unique symbol but consist of multiple bounding boxes annotations. When the same symbol relates to more than one saint (e.g., Baby Jesus may appear with both the Virgin Mary and St. Anthony of Padua), its presence is denoted with a label composed of the symbol name and the Saint's name. Hence, there will be multiple bounding boxes for the same symbol.

The creation of symbol-level annotations, for each image, relies on the following workflow:

- Identification of *all* the relevant symbols;
- Identification of the character associated with the symbols;
- For each symbol, identify its parts (e.g., Saint Peter's forehead separates his hair, so it usually needs two bounding boxes);
- For each part, create a bounding box associated with the symbol's label (which includes the saint name).

Creating this additional workflow is required since what discriminates a saint from the other is the difference in the symbols associated with them. It is expected, for this reason, that class activation maps focus chiefly on the relevant symbols regions (i.e., on the symbols specific to the saint under analysis). Saint-level bounding boxes, instead, are not able to suitably capture the reasons why an artwork was associated with a saint.

Then, a crucial phase consists of selecting the most frequent symbols across the entire data set (assuming that the symbols in the train and test set have an akin distribution). Infrequent symbols do not allow precise analyses because, presumably, they were not deemed relevant during the training.

A prominent Christian artwork particularly rich in symbols is *The Last Judgment* by Michelangelo Buonarroti, presented in Figure 4.3.[3] A variety of studies have focused on the analysis of this artwork and of its meaning [1, 6, 12], which underlines its complexity, the importance of using automated analyses to support experts and the difficulty of creating accurate annotations in the field of Art History.

COMPARISON BETWEEN SYMBOL AND WHOLE SAINT BOUNDING BOXES:

This section presents the importance of creating both symbol-level and whole Saint bounding boxes, which serve different purposes and can be used for different analyses.

Symbol bounding boxes and whole Saint bounding boxes have two different purposes:

- Symbol bounding boxes are used to evaluate the ability to locate characteristic symbols, rather than the whole Saint;
- Whole Saint bounding boxes are used to evaluate the ability to locate the entire body of the Saint.

[3] The Last Judgment, Michelangelo Buonarroti, 1537 - 1541.

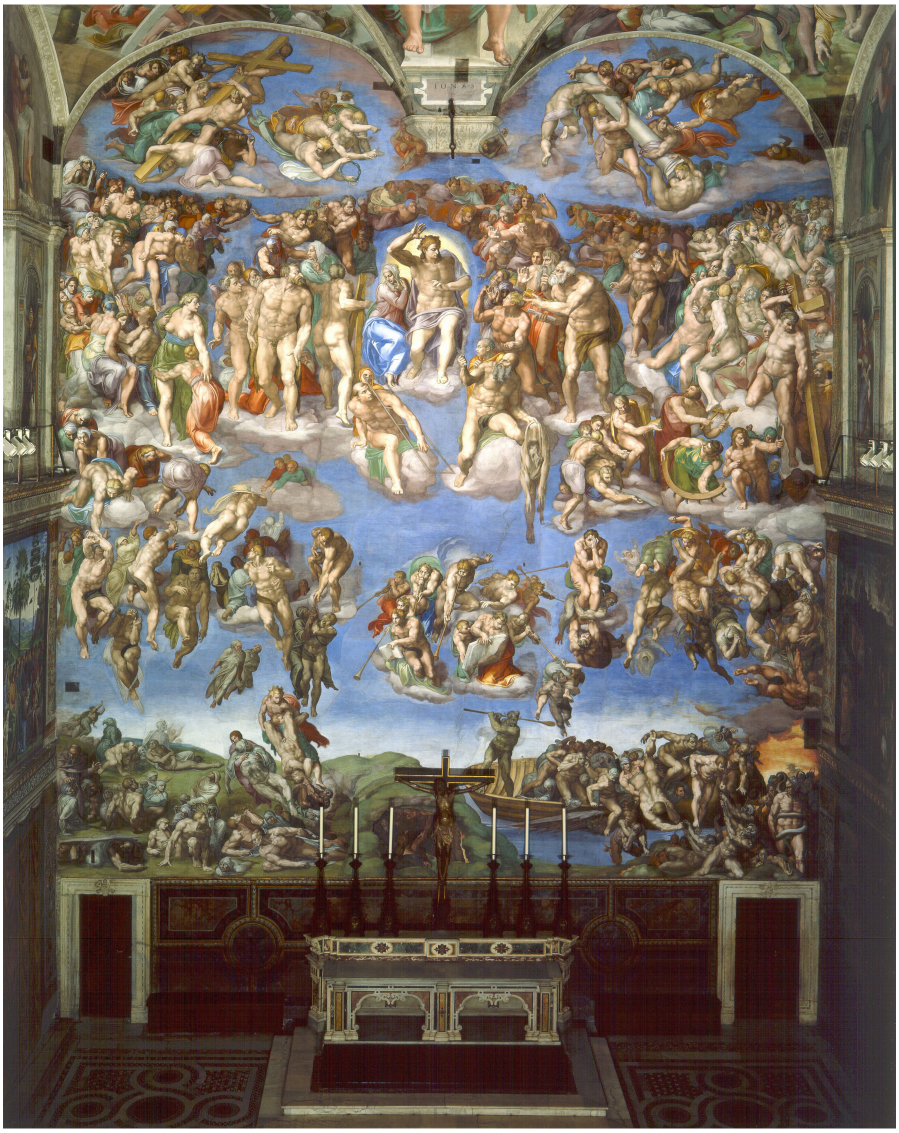


Figure 4.3: A multitude of symbols – This example shows a complex artwork, containing multiple characters, in part associated with symbols. A manual analysis would require expertise and considerable effort.

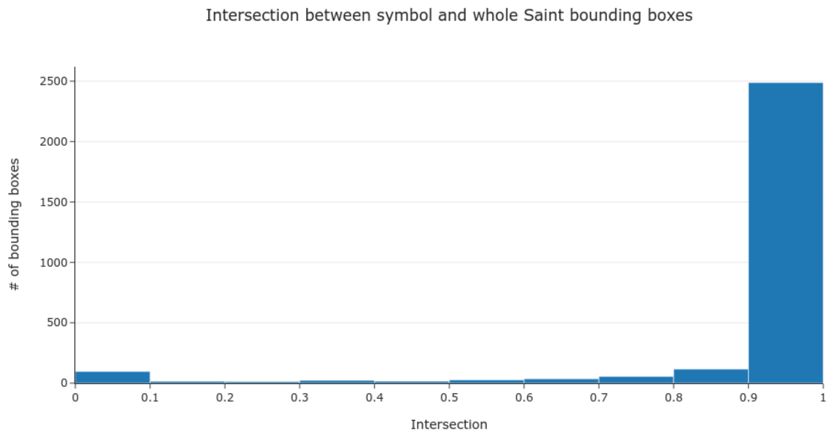


Figure 4.4: Distribution of the intersection values between the symbol bounding boxes and the corresponding whole Saint bounding boxes.

Symbol bounding boxes are more relevant from an iconographical point of view since in principle a Saint should be identified through the associated characteristic symbols. For this reason, the majority of the analyses in this research relied on symbol bounding boxes. The presence of two alternatives for annotations, then, is not the result of inaccurate annotations, but rather a choice aimed at evaluating results differently.

A prominent characteristic of the ArtDL dataset paintings is the closeness of symbols to the body of the Saint. On average, 92.66% of a symbol bounding box is contained inside the corresponding whole Saint bounding box. Less than 1.87% (54 out of 2887) of the symbol bounding boxes don't intersect the corresponding whole Saint bounding boxes, while 86.2% (2489 of 2887) intersect the whole Saint bounding box for at least 90% of their area inside. Figure 4.4 shows the distribution of the intersection values between the symbol bounding boxes and the corresponding whole Saint bounding boxes, emphasizing the closeness of the symbols to the corresponding Saint body.

Figure 4.5[4] present a typical example in which a symbol (here, Saint Jerome's galero) is detached from the Saint's body and does not cover a large proportion of the painting. Including the galero in the whole Saint bounding box would produce an exceedingly large bounding box, covering most of the painting and leading to biased analyses.

Saint Jerome is characterized by more symbols than the other saints, as shown in Table 4.1. For this reason, finding detached symbols is more common than in paintings depicting other Saints. However, other examples can be found. Figure 4.6[5] depicts Virgin Mary holding Jesus in her arms. He is attached to her body, and for this reason included inside the whole Saint bounding box (green), while the lily, a characteristic symbol of Virgin Mary, is detached from the body and covers a much smaller area with respect to it. Consistently with the other bounding boxes, the whole Saint bounding box does not cover the lily. By including it, the bounding box would extend for a much wider area, leading to biased analyses.

4.1.1.2 *Symbol filtering*

This section presents the procedures adopted for filtering symbols, based on the number of ground-truth annotations in the test set. While some symbols appear in the majority of the images of the corresponding Saint, others are absent or rarely present. For each Saint, only the symbols that appear in at least 5% of the paintings depicting the respective Saint are kept. This filter eliminates 23 of the 84 possible symbols associated with the 10 Iconclass categories and reduces the number of symbol bounding boxes from 2957 to 2887.

[4] The Penitent St Jerome, Filippino Lippi, ca. 1485.

[5] Madonna in trono con Sant'Antonio e San Benedetto, nd, ca. 1500-1540.

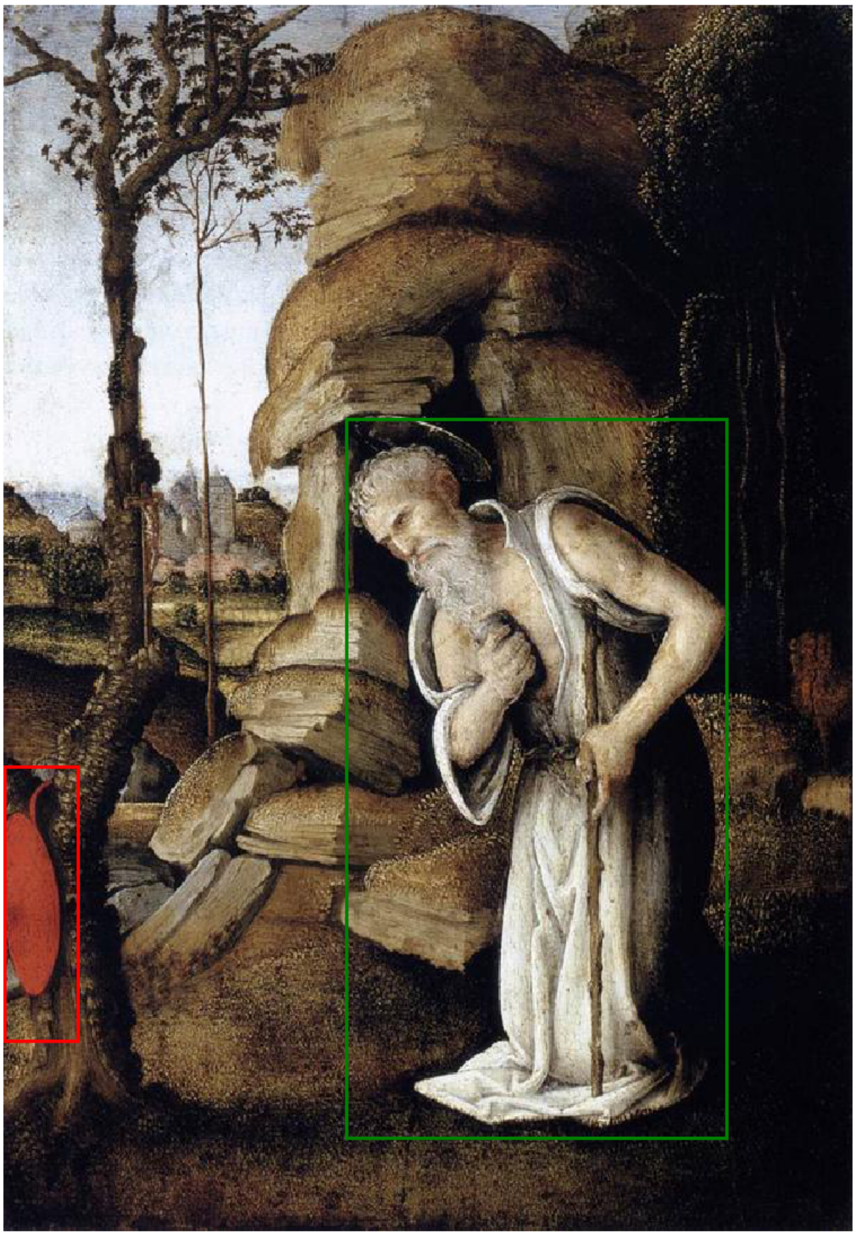


Figure 4.5: An example of distant symbol. In this painting, depicting Saint Jerome, his galero (red bounding box) is not attached to the saint's body. By including the galero in the whole Saint bounding box (green), it would have covered most of the painting surface, leading to biased analyses.

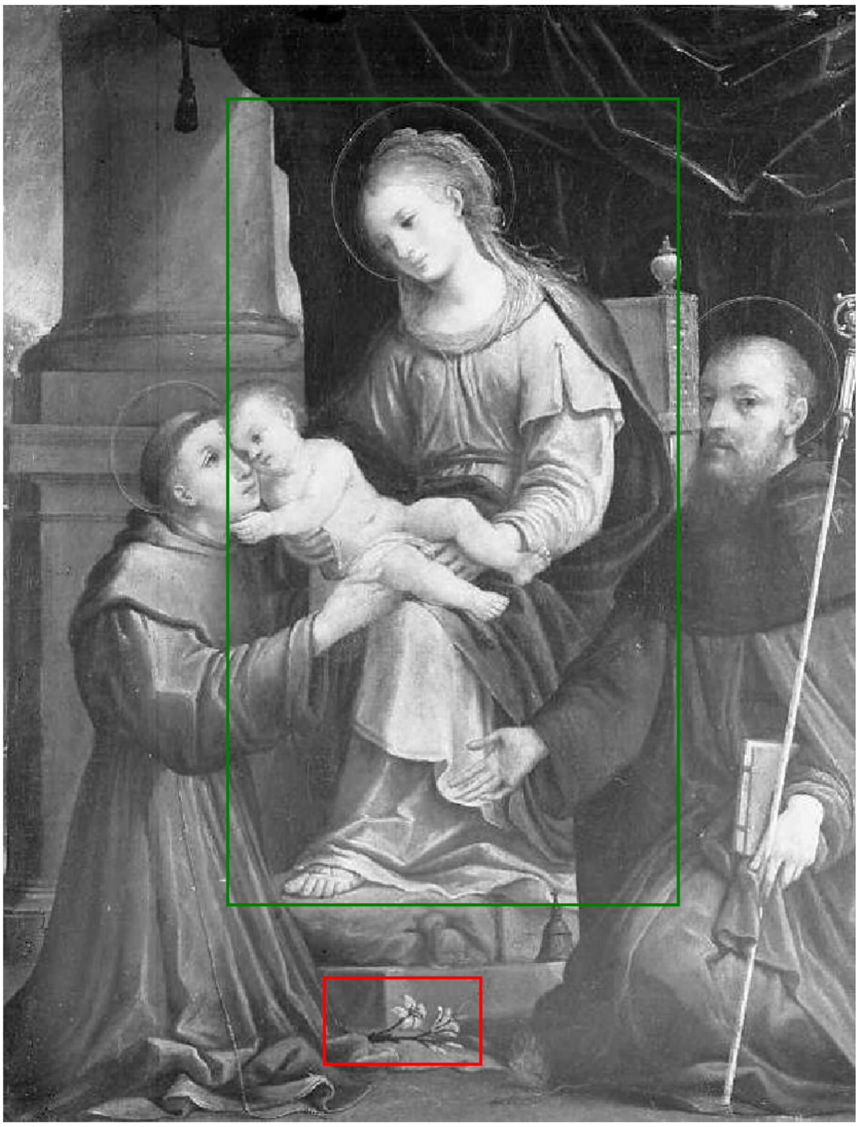


Figure 4.6: An example of symbol not attached to the body. In this painting, depicting Virgin Mary, the lily (red bounding box) is close to the saint (green bounding box), but covering it would have produced a much larger bounding box, leading to biased analyses.

Table 4.3: Symbol and bounding box distribution.

Iconclass category	Symbol classes	Symbol bounding boxes
Anthony of Padua	6	83
Dominic	4	59
Francis of Assisi	5	295
Jerome	11	434
John the Baptist	5	231
Mary Magdalene	5	283
Paul	6	132
Peter	9	408
Sebastian	3	267
Virgin Mary	7	695

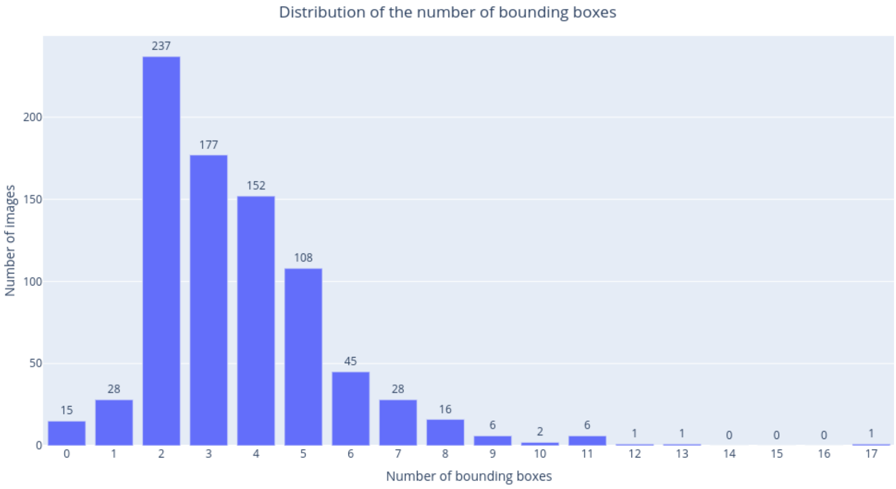


Figure 4.7: Bounding box distribution: most images contain from 2 to 5 bounding boxes (average = 3).

Table 4.3 presents the characteristics of the data set used to compare the class activation maps algorithms.

Figure 4.7 shows the distribution of the bounding boxes within the images. Most images contain from 2 to 5 bounding boxes. A few do not contain annotations. The latter case occurs when the automatic classification of the ArtDL data set is incorrect (e.g., for images in which a character named Mary was incorrectly associated with the Virgin Mary).

4.1.2 *Experimental procedures*

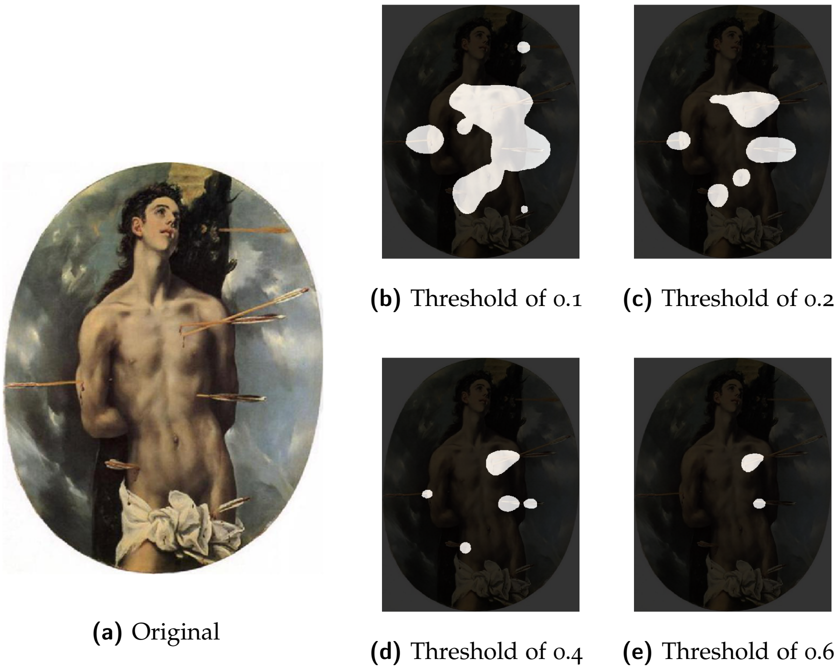
This section introduces the experimental procedures used in this thesis. Section 4.1.2.1 presents the generation of class activation maps; Section 4.1.2.2 introduces the selection of the threshold; Section 4.1.2.3 introduces the Intersection Over Union metrics, used for performing quantitative analyses and Section 4.1.2.4 outlines the most relevant analyses that can be performed.

4.1.2.1 *Generation of Class Activation Maps*

The class activation maps are generated by feeding the image to the ResNet50 model and applying the computations explained in Section 3. They have a size equal to $h \times w \times c$ where h and w are the height and width of the *conv5_x* layer and c is the number of classes. Since the output size (h, w) is smaller than the input size, due to the convolution operations performed by the ResNet architecture, each class activation map is upsampled with bilinear interpolation to match the input image size. Min-max scaling is applied to the upsampled class activation maps to normalize them in the $[0, 1]$ range.

4.1.2.2 *Choice of the threshold value*

A class activation map contains values in the range from 0 to 1. Given a threshold t, the class activation map can be separated into the background (pixels with a value lower than t) and the foreground (pixels with a value greater than or equal to t). The choice of the threshold value aims at making foreground areas concentrate on the Saints' figure and symbols. Figure 4.8 shows the impact of applying different threshold values to a class activation map. As the threshold value increases, the foreground areas (in white) become smaller and more distinct and the background pixels increase substantially at the cost of fragmenting the foreground areas and missing relevant symbols. To investigate the choice of the proper threshold, the quantitative evaluation of Section 4.2 reports results obtained with multiple values uniformly distributed from 0 to 1 with a step of 0.05.


(b) Threshold of 0.1 (c) Threshold of 0.2

(a) Original

(d) Threshold of 0.4 (e) Threshold of 0.6

Figure 4.8: Analysis with different thresholds – Black areas correspond to class activation map values below the specified threshold (background) while white pixels correspond to class activation map values greater or equal than the threshold (foreground). An increment in the threshold value results in smaller and more distinct areas.

4.1.2.3 *Intersection Over Union metrics*

IoU is a standard metrics used to compute the overlap between two different areas. It is defined as:

$$\text{IoU} = \frac{A_{\cap}}{A_{\cup}},$$

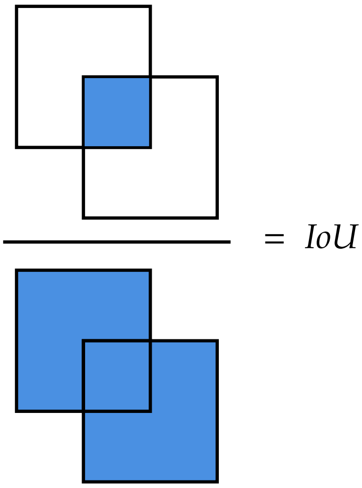


Figure 4.9: Intersection Over Union – This figure presents a graphical representation of the IoU metric, computed considering two rectangular areas. In general, the areas can have any shape.

where $A_{\cap}$ is the intersection between the two areas and $A_{\cup}$ is their union. IoU ranges between 0 and 1, with 0 meaning that the two areas are disjoint and 1 meaning that the two areas overlap and have equal dimensions (Figure 4.9). We use IoU to compare the foreground regions of the class activation maps with the ground truth bounding boxes. The computation of the class activation maps and the metrics does not depend on the number of Saints in the painting because every Iconclass category is associated with a different activation map independent of the others. All the reported results are valid regardless of the number of Saints. This metrics is employed in diverse quantitative analyses. Depending on the analysis, different definitions are employed to calculate the areas used in the computation of the union and the intersection.

4.1.2.4 *Relevant analyses*

This section presents the most relevant analyses that can be performed and explains why other analyses would be biased or impossible.

In this research, the evaluation relies on generated class activation maps (e.g., calculating the IoU between a class activation map and a GT bounding box) or on bounding boxes generated from the class activation maps (e.g., calculating the IoU between a generated bounding box and a GT bounding box, with the same granularity). Class activation maps have a coarser granularity and refer to the saint, since the network is trained on saints, rather than on symbols. As a consequence, they do not contain symbol-level information. Consequently, the generated bounding boxes also refer to the saint, and cannot contain symbol-level information. For this reason, any analysis relying on the confusion matrix (Table 4.4) at symbols level is not possible. On the other hand, coarse-grained analyses involving only the Saints are possible, and some of them were conducted by Milani and Fraternali [47].

The confusion matrix concepts are employed to define the Mean Average Precision (mAP). Given a GT bounding box, a generated bounding box and a threshold $t \in [0, 1]$, the IoU of the two bounding boxes can be computed. Then, it is possible to compute two quantities:

- False positive, when $\mathrm{IoU} < t$;
- True positive, when $\mathrm{IoU} \geqslant t$.

Table 4.4: The confusion matrix considering two classes: a positive and a negative class.

		Actual		
		Positive	Negative	
Predicted	Positive	True positive TP *(Correct)*	False positive FP *(Incorrect)*	Precision/Positive Predictive Value (PPV) $\frac{TP}{TP+FP}$
	Negative	False negative FN *(Incorrect)*	True negative TN *(Correct)*	Negative Predictive Value (NPV) $\frac{TN}{TN+FN}$
		Sensitivity/Recall Rate (RR) $\frac{TP}{TP+FN}$	Specificity Rate (SR) $\frac{TN}{TN+FP}$	

Consequently, it is possible to obtain the precision and the recall for a set of generated Saint bounding boxes with respect to the set of GT bounding boxes for the same images, using the definitions of Table 4.4. The mAP is the area under the precision-recall curve obtained by the combinations of precision and recall values for different IoU thresholds. Section 4.2.6.2 presents the mAP results for the ArtDL data set.

4.2 QUANTITATIVE ANALYSIS

This section presents the qualitative comparison results concerning the effectiveness of the class activation maps algorithms in the localization of iconography classes and their symbols.

Smooth Grad-CAM++ is the only method that requires hyper-parameters: the standard deviation σ and the number of samples s. To set the hyper-parameter values a grid-search was executed in the following space: $\sigma \in \{0.25, 0.5, 1\}$ and $s \in \{5, 10, 25\}$. Only the best and worst Smooth Grad-CAM++ configurations are reported, to highlight the boundary values reached by this algorithm. The results show that the number of samples barely affects the results, whereas the standard deviation has a higher impact. To reduce the computational cost, a lower number of samples is preferable.

This section is organized as follows: Section 4.2.1 presents the component IoU analysis, Section 4.2.2 presents the global IoU analysis, Section 4.2.3 presents the bounding box coverage analysis, Section 4.2.4 presents the symbols coverage analysis, Section 4.2.5 presents the irrelevant attention analysis and Section 4.2.6 presents the generation of bounding boxes from activation maps and quantitative analyses concerning them.

4.2.1 *Component IoU*

This metrics evaluates how well the class activation map focuses on the individual Saints' symbols. First, the class activation map's foreground area is divided into *connected components*, i.e., groups of pixels connected to each other. The IoU value is calculated between *each ground truth bounding box* and *the connected components that intersect it*. Then, the average IoU across all symbol classes is taken. This procedure is repeated for all threshold values. Figure 4.10 shows that the best results are obtained by Smooth Grad-CAM++ with a standard deviation $\sigma = 1$ and a number of samples $s = 5$. The reason for this is that Smooth Grad-CAM++ tends to produce smaller and more focused areas, which yield more connected components and better coverage of the distinct symbols. Grad-CAM tends to create larger and more connected areas. This increases the size of the union and such an increase is not compensated by an equivalent increase of the intersection, which motivates the lower IoU values. In all the

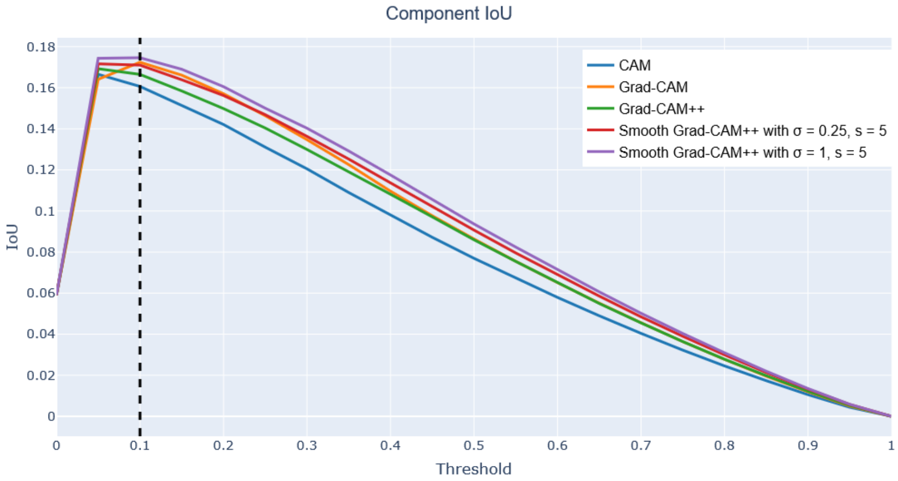


Figure 4.10: Component IoU at varying threshold levels.

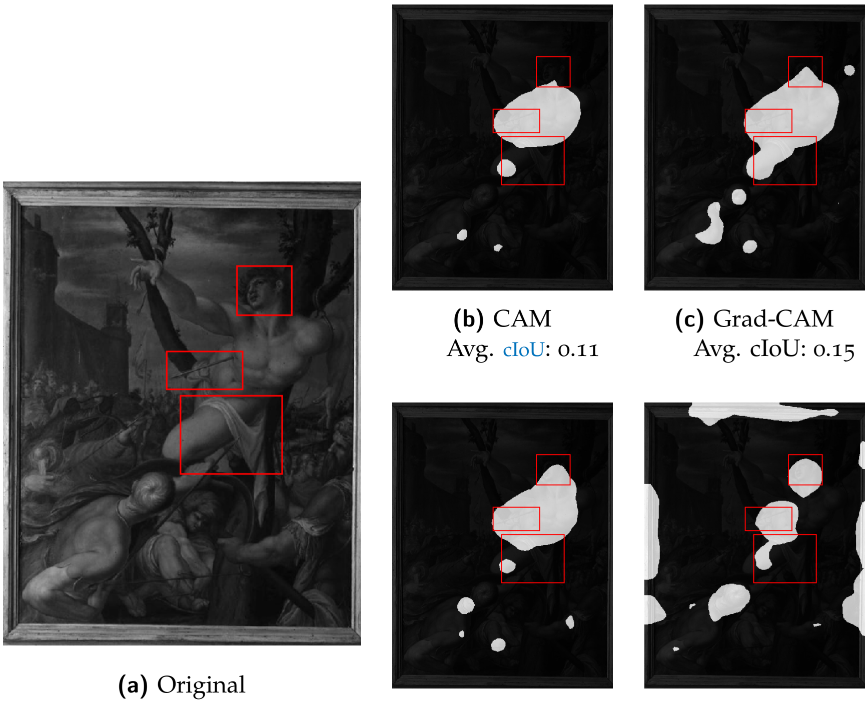
considered class activation maps variants, the component IoU peak is found for a threshold value $t \in \{0.05, 0.1\}$. Grad-CAM creates larger and more connected regions and thus a higher threshold is needed to obtain the same number of components as the other methods. This explains why the component IoU peak is found at a higher threshold. Figure 4.11[6] compares the component IoU values produced on a sample image by different class activation maps algorithms. For the same threshold value, Smooth Grad-CAM++ creates more and better-focused components.

Section A.1 further analyzes the hyper-parameters settings concerning Smooth Grad-CAM++.

4.2.2 *Global IoU*

An alternative metrics is the IoU between *the union of all the bounding boxes* in the image and *the entire foreground area* of the class activation map taken at a given threshold. This metrics is calculated for all threshold values and assesses how the class activation map focuses on

[6] San Sebastian's Martyrdom, Giovanni Maria Butteri, 1550-1559.



(a) Original

(b) CAM
Avg. cIoU: 0.11

(c) Grad-CAM
Avg. cIoU: 0.15

(d) Grad-CAM++
Avg. cIoU: 0.13

(e) Smooth GC++
Avg. cIoU: 0.35

Figure 4.11: Different values of component IoU produced by different class activation maps algorithms (Smooth Grad-CAM++ with $\sigma = 1$ and $s = 5$) at threshold $t = 0.1$. Ground truth bounding boxes are shown in red. Here, cIoU refers to the component IoU.

the whole representation of the Saint, favouring those class activation maps methods that generate wider and more connected areas rather than separated components. Figure 4.12 shows that Grad-CAM is significantly better than the other analyzed methods. As already observed, Grad-CAM tends to spread over the entire figure and covers better the Saint and the associated symbols. Due to the complementary role of the component and global IoU metrics, the method with the best component IoU (Smooth Grad-CAM++ with $\sigma = 1$ and $s = 5$) has the worst global IoU. Differently from the component IoU, the global IoU peak position on the x axis does not change across methods, because the influence of the number of components is less relevant

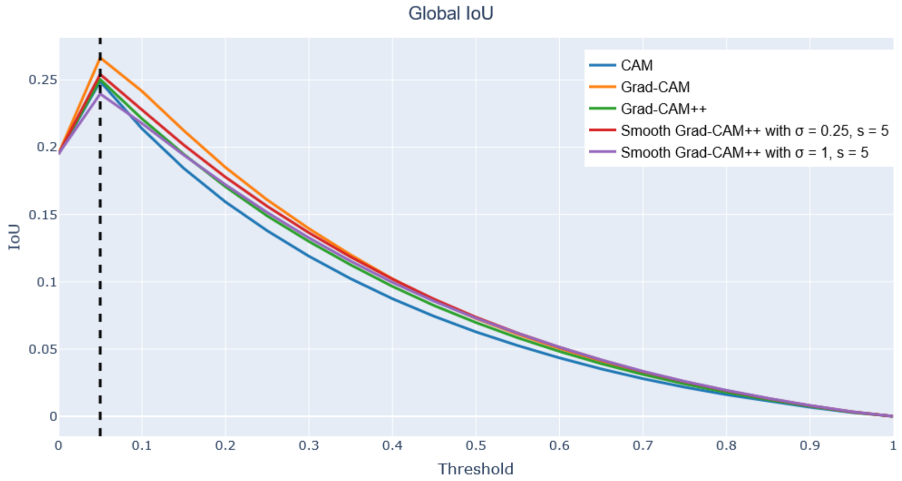


Figure 4.12: Global IoU at varying threshold levels.

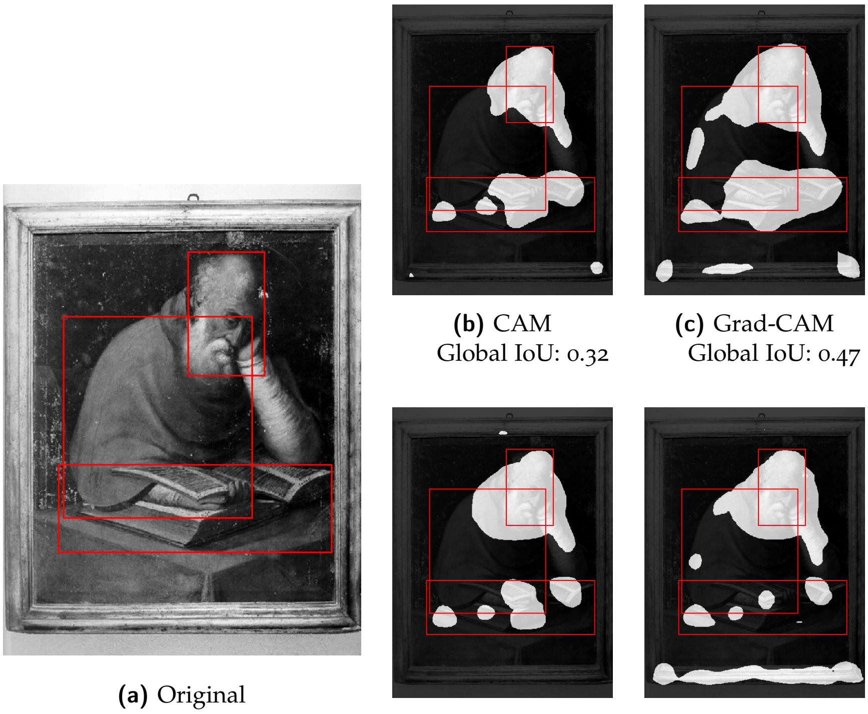
when the global metrics is computed. Figure 4.13[7] compares the global IoU values produced on a sample image by different class activation maps algorithms. For the same threshold value, Grad-CAM generates wider areas that cover more foreground pixels.

Section A.2 further analyzes the hyper-parameters settings concerning Smooth Grad-CAM++.

4.2.3 *Bounding box coverage*

When analyzing the class activation maps algorithms, a factor to consider is also how many bounding boxes are covered by each class activation map. This metrics alone is not enough to characterize the performance because a trivial class activation map covering the entire image would have 100% coverage. However, coupled with the two previous metrics, it can give information about which method can generate class activation maps that can highlight a considerable fraction of the iconographic symbols that an expert would recognize. The bounding box coverage metrics considers that a bounding box is

[7] Saint Jerome in the study, nd, 1604.



(a) Original

(b) CAM
Global IoU: 0.32

(c) Grad-CAM
Global IoU: 0.47

(d) Grad-CAM++
Global IoU: 0.34

(e) Smooth GC++
Global IoU: 0.25

Figure 4.13: Different values of global IoU produced by different class activation maps algorithms (Smooth Grad-CAM++ with $\sigma = 1$ and $s = 5$) at threshold $t = 0.05$. Manually annotated symbol bounding boxes are shown.

covered by the class activation map only if their intersection is greater than or equal to 20% of the bounding box area.

Figure 4.14 illustrates the functioning of this metrics by considering a bounding box covered by the class activation map (in green) and a bounding box that is not covered by the class activation map (in red).

Figure 4.15 presents the results: Grad-CAM and Smooth Grad-CAM++ intersect, on average, more bounding boxes than the other methods. This result confirms that Grad-CAM covers wider areas while focusing on the correct details at the same time. The worst method, CAM, performs poorly also in the two previous metrics. This indicates that it generates class activation maps that are smaller

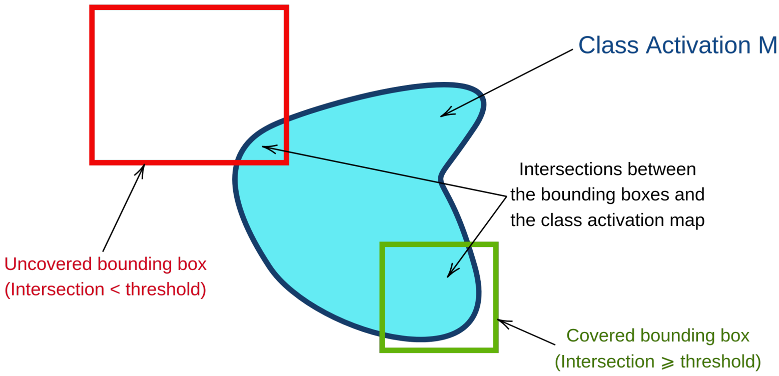


Figure 4.14: Bounding box coverage. This figure shows a symbol-level bounding box covered by the class activation map and a symbol-level bounding box not covered by the class activation map.

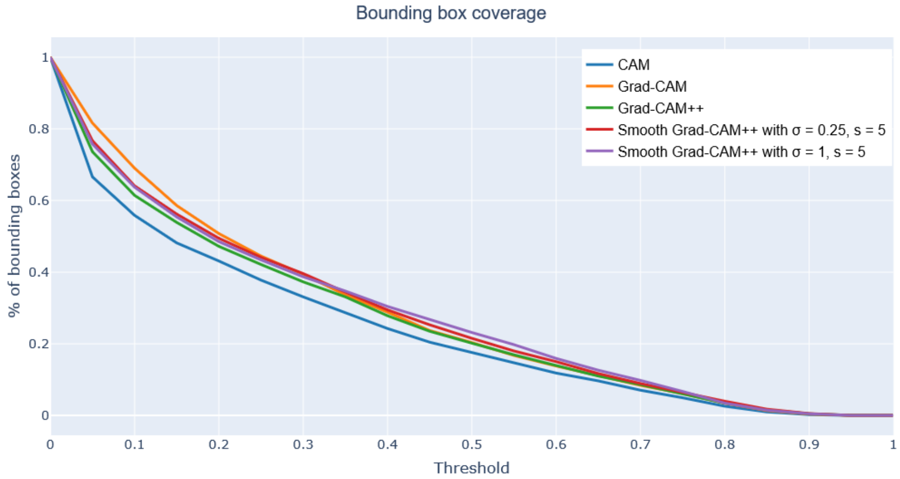


Figure 4.15: Bounding box coverage at varying threshold values.

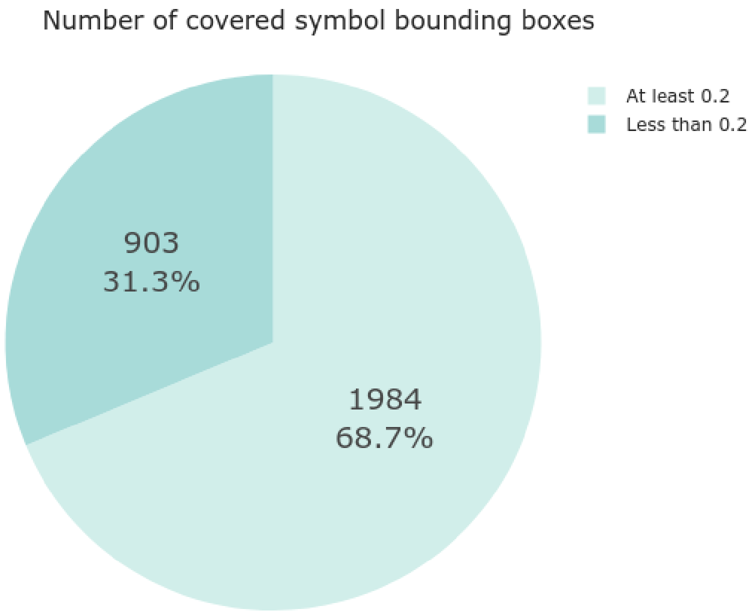


Figure 4.16: Bounding box coverage for $t = 0.1$, using Grad-CAM.

and less focused on the iconographic symbols compared to the other approaches.

Figure 4.16 shows the number and percentage of covered bounding boxes for a threshold $t = 0.1$ (i.e., the threshold maximizing the component IoU). A bounding box is considered covered when it is intersected by at least 20% of the corresponding class Grad-CAM.

Figure 4.17 shows six examples where the GT bounding boxes are entirely covered by the activation map as defined in the component IoU analysis (i.e., the portion of the class activation map above the threshold $t = 0.1$). Moreover, the selected examples show that the class activation map covers the most relevant areas of each painting, focusing on significant symbols (e.g., the arrows of Saint Sebastian). Figure 4.18, instead, shows six examples where more than 95% of each GT bounding box is covered by the activation map as defined in the component IoU analysis. As in Figure 4.17, the examples show that the class activation map covers the most relevant areas of each painting, focusing on significant symbols (e.g., the beard of Saint Peter). The bounding boxes are not entirely covered by class activation maps since they are close to the map boundaries. The class activation maps, indeed, tend to follow more closely the shape of the symbols. Figure 4.19 shows six examples where the GT bounding boxes are not sufficiently covered by the activation map as defined in the component IoU analysis. The selected examples show that the symbols not covered by the class activation map are either less discriminative (e.g., Jerome vest in the first painting is less discriminative than the lion) or are too thin or characterized by a low contrast (e.g., the crucifix of Saint John the Baptist and one of the arrows of Saint Sebastian).

Section A.3 further analyzes the hyper-parameters settings concerning Smooth Grad-CAM++.

4.2.4 *Symbols coverage*

When analyzing the class activation maps algorithms applied to iconography, it is interesting to evaluate also the number of covered *symbols*, rather than the number of covered bounding boxes. A symbol can spread over several bounding boxes (e.g., Saint Sebastian arrows), and the bounding boxes coverage analysis considered alone does not provide symbol-level information. As in the bounding box coverage analysis, a trivial class activation map covering the entire image would have 100% coverage. For this reason, this analysis is meaningful if coupled with the previous metrics. The symbol coverage metrics considers that a symbol is covered by the class activation map

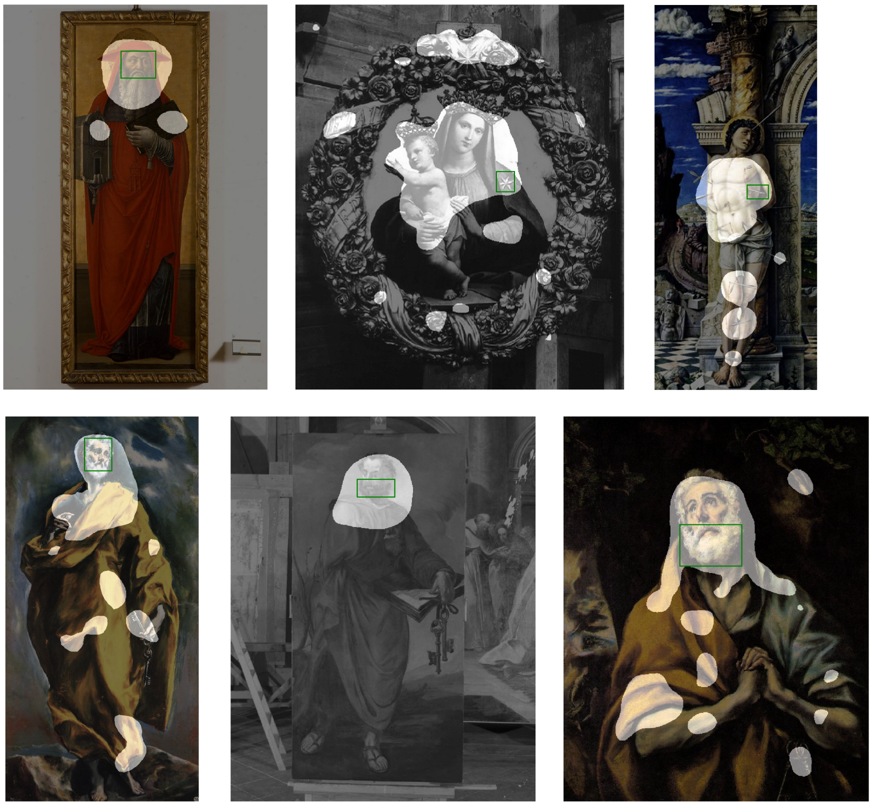


Figure 4.17: Examples of bounding boxes completely covered by the class activation map. The six examples presented in this figure illustrate the case of an intersection between a bounding box and the class activation map equal to 100%, using Grad-CAM.

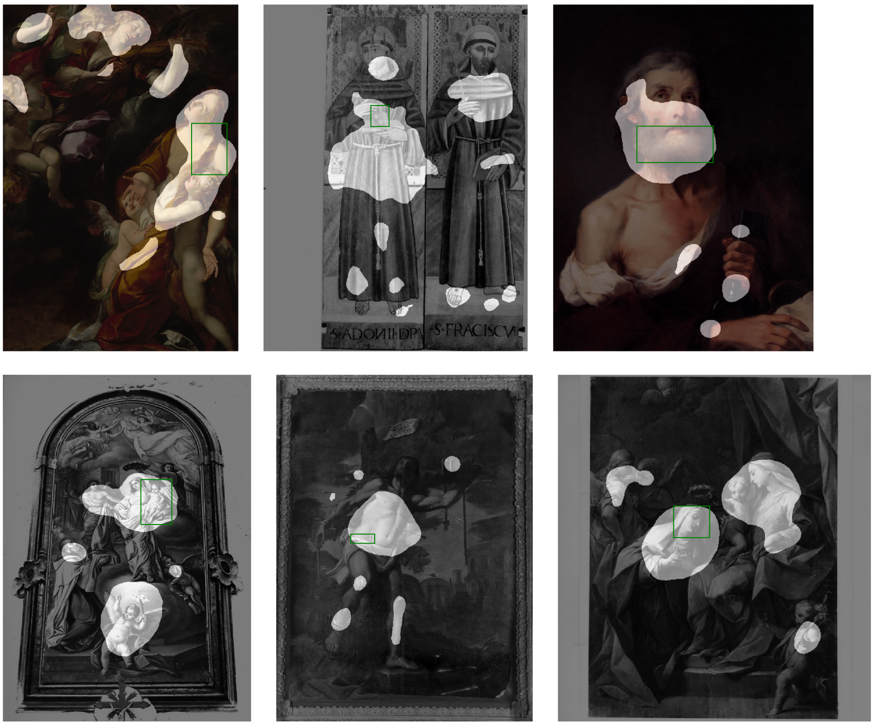


Figure 4.18: Examples of bounding boxes almost entirely covered by the class activation map. The six examples presented in this figure illustrate the case of an intersection between a bounding box and the class activation map ⩾ 95%, using Grad-CAM.

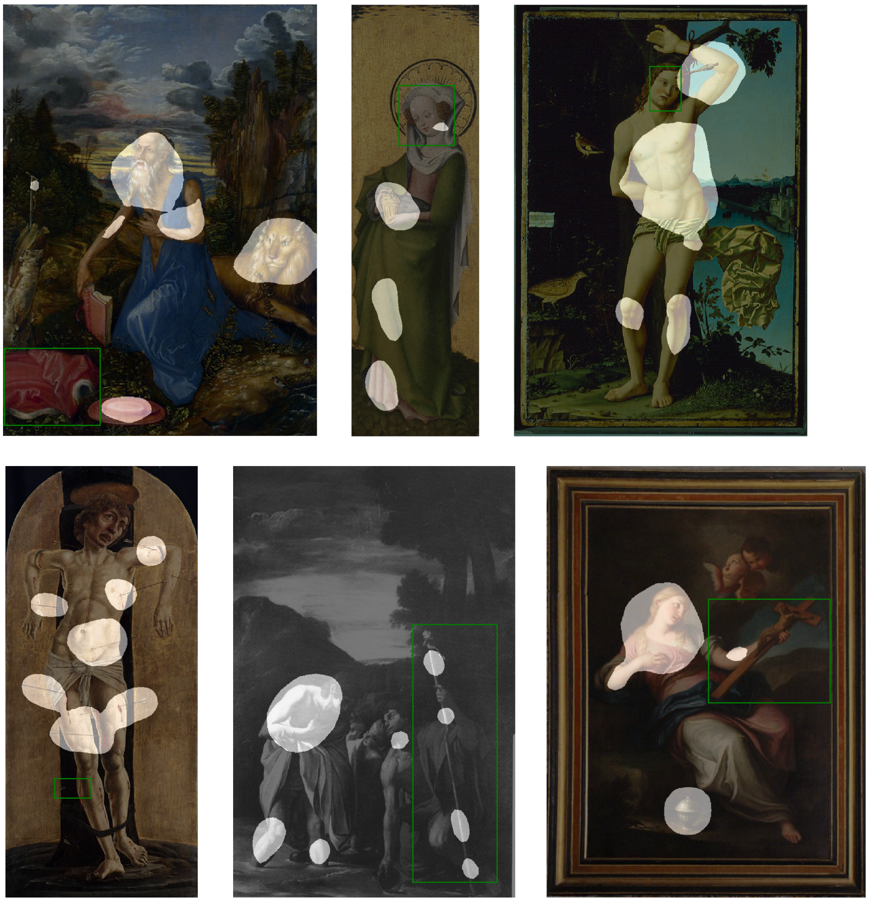


Figure 4.19: Examples of bounding boxes not sufficiently covered by the class activation map. The six examples presented in this figure illustrate the case of an intersection between a bounding box and the class activation map $\leqslant 10\%$, using Grad-CAM.

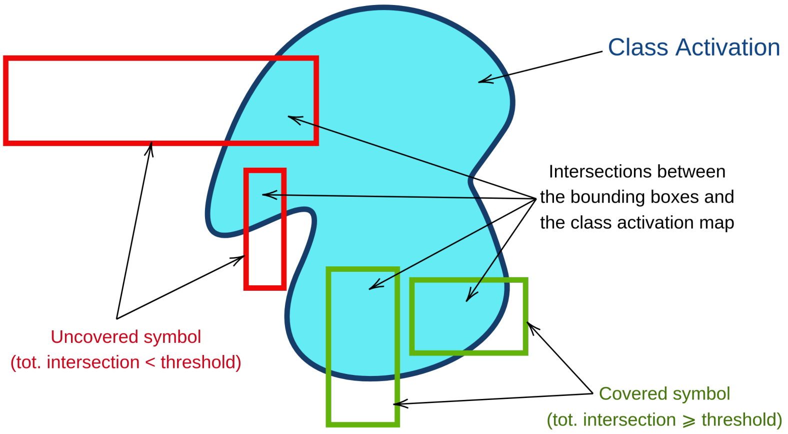

Class Activation Map
Intersections between the bounding boxes and the class activation map
Uncovered symbol
(tot. intersection < threshold)
Covered symbol
(tot. intersection ⩾ threshold)

Figure 4.20: Symbols coverage. This figure shows a symbol covered by the class activation map and a symbol not covered by the class activation map. The former is divided inside two green bounding boxes, while the latter is contained inside two red bounding boxes.

if the total intersection between the symbols bounding boxes and the class activation map is greater than or equal to 20% of the sum of the symbol bounding boxes areas:

$$\frac{\sum_i \mathit{Area}(\mathit{bbox}_i \cap \text{Grad-CAM})}{\sum_i \mathit{Area}(\mathit{bbox}_i)} \geqslant 20\% \tag{4.1}$$

Figure 4.20 illustrates the functioning of this metrics by considering a symbol covered by the class activation map (in green) and a symbol that is not covered by the class activation map (in red).

Figure 4.21 shows the number and percentage of covered symbols for a threshold $t = 0.1$ (i.e., the threshold maximizing the component IoU). It shows results comparable with the ones of the bounding box coverage, which suggests that, on average, symbols that spread in the image (i.e., symbols with multiple bounding boxes) are as likely as the other symbols to be highlighted by the class activation map.

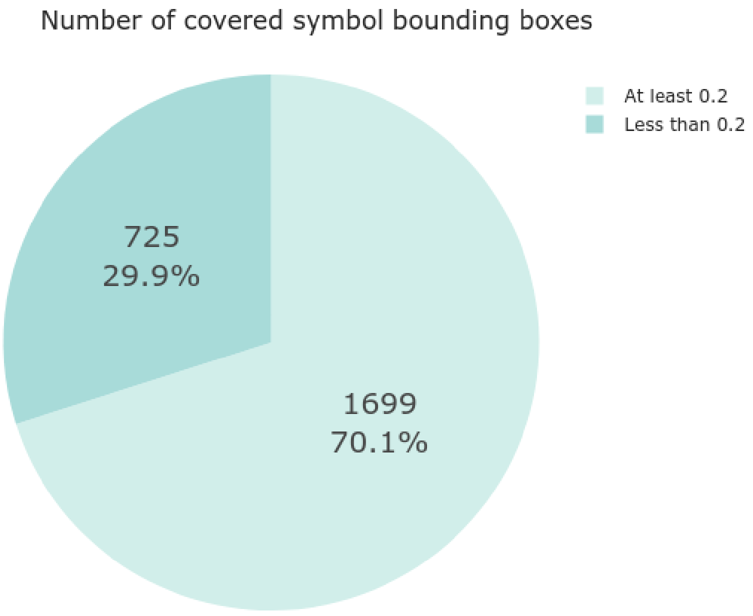


Figure 4.21: Symbols coverage for $t = 0.1$, using Grad-CAM.

Figure 4.22 shows four examples where GT symbols that spread through several bounding boxes are almost entirely covered by the activation map as defined in the component IoU analysis (i.e., the portion of the class activation map above the threshold $t = 0.1$). Moreover, the selected examples show that the class activation map covers relevant areas of the painting, focusing on significant symbols (e.g., the ointment jar of Mary Magdalene). Figure 4.23 shows two examples where GT symbols spreading through several bounding boxes are not sufficiently covered by the activation map as defined in the component IoU analysis. In the first example, the books are less discriminative than the lion (which instead characterizes only Saint Jerome), while in the second example, Saint Peter hair are not as bushy as in other paintings, making them less relevant than the beard.

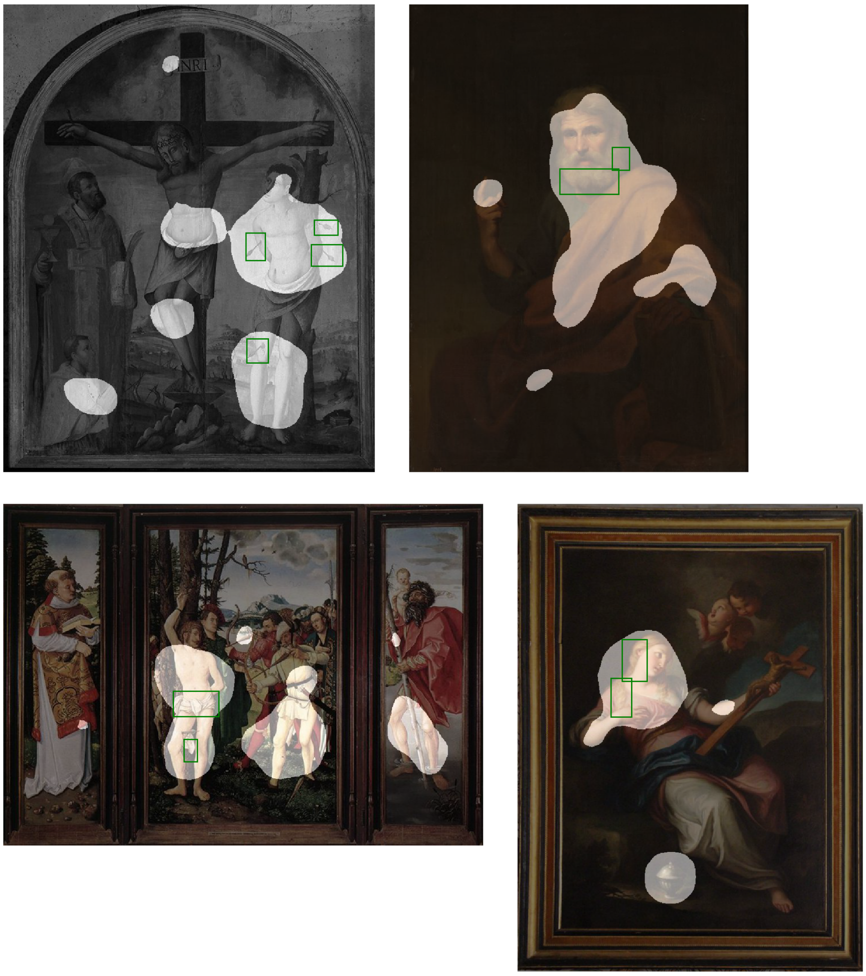


Figure 4.22: Examples of bounding boxes almost entirely covered by the class activation map. The six examples presented in this figure illustrate the case of an intersection between a symbol spreading through several bounding boxes and the class activation map $\geqslant 95\%$, using Grad-CAM.

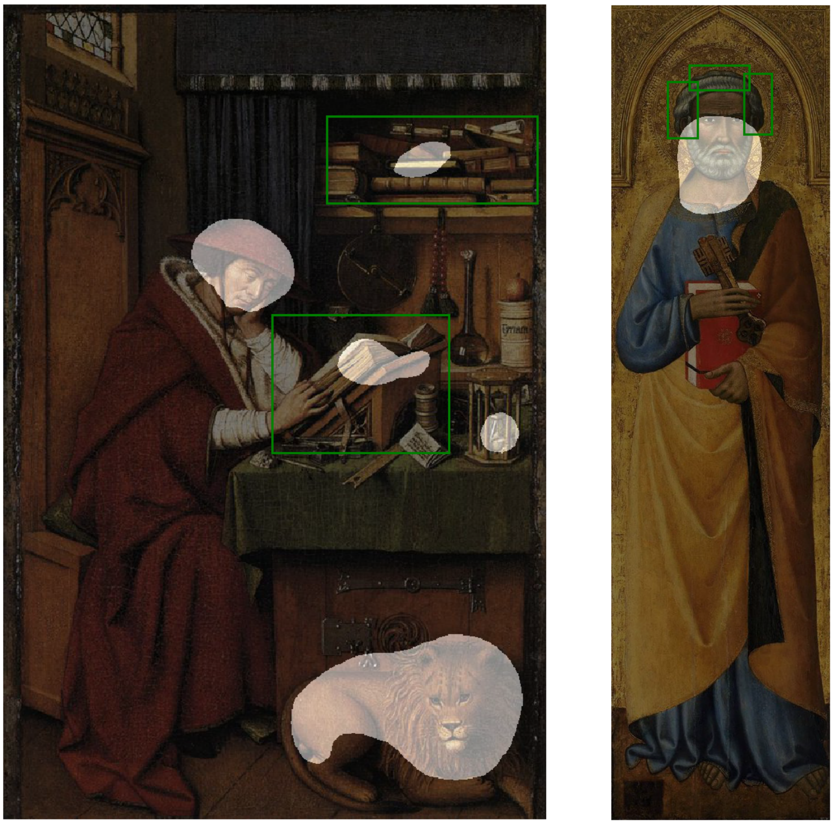


Figure 4.23: Examples of bounding boxes not sufficiently covered by the class activation map. The two examples presented in this figure illustrate the case of an intersection between a symbol spreading through several bounding boxes and the class activation map $\leqslant 5\%$, using Grad-CAM.

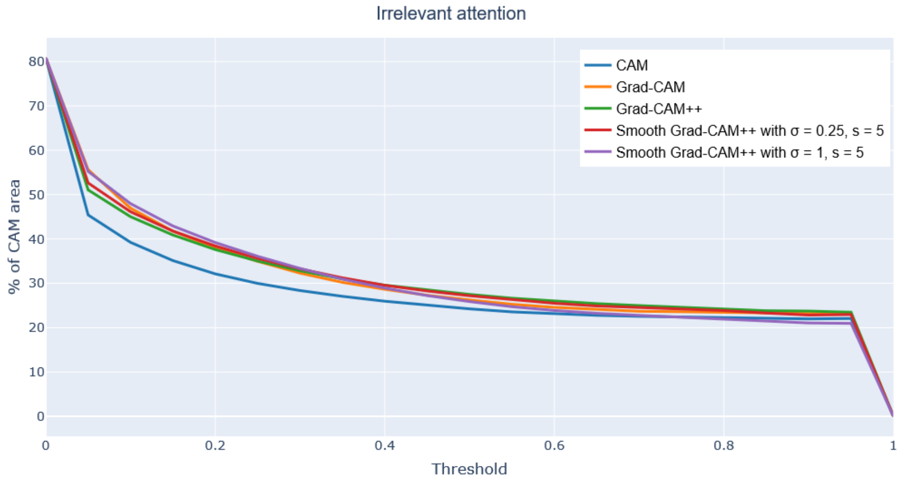


Figure 4.24: Irrelevant attention at varying threshold values.

4.2.5 *Irrelevant attention*

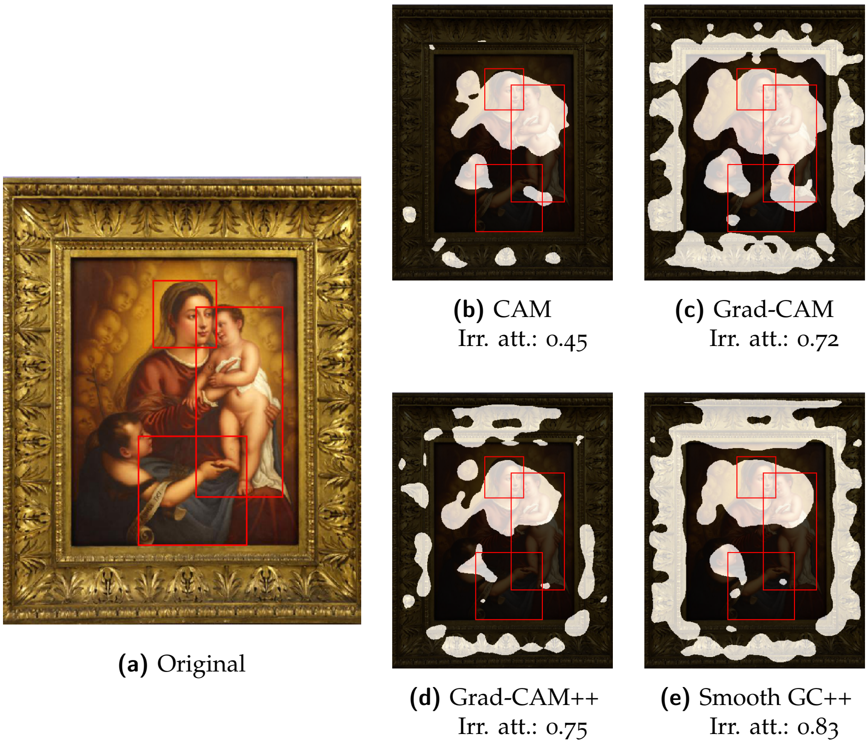
When evaluating the global IoU, a low value can occur for two reasons:

1. The two areas have a very small intersection;
2. The two areas overlap well, but one is much larger than the other.

Thus, an analysis of how much the class activation maps focus on irrelevant parts of the image helps to characterize low global IoU values. Irrelevant attention corresponds to the percentage of class activation map area outside any bounding box. Figure 4.24 shows that CAM has the lowest irrelevant attention, coherently with the previous results. Figure 4.25[8] compares the irrelevant attention values produced on a sample image by different class activation maps algorithms. For the same threshold value, CAM generates smaller irrelevant areas, whereas Grad-CAM and Smooth-Grad-CAM++ include more irrelevant regions corresponding to the painting frame. The tendency of

[8] Madonna with Child and Infant St. John surrounded by Angels, Tiziano Vecellio, 1550.

Smooth Grad-CAM++ to focus on irrelevant areas can be seen also in Figures 4.11 and 4.13.


(a) Original

(b) CAM
Irr. att.: 0.45

(c) Grad-CAM
Irr. att.: 0.72

(d) Grad-CAM++
Irr. att.: 0.75

(e) Smooth GC++
Irr. att.: 0.83

Figure 4.25: Different values of irrelevant attention produced by different class activation maps algorithms (Smooth Grad-CAM++ with $\sigma = 1$ and $s = 5$) at threshold $t = 0.1$. Manually annotated symbol bounding boxes are reported.

Section A.4 further analyzes the hyper-parameters settings concerning Smooth Grad-CAM++.

4.2.6 *Bounding boxes*

This section proposes a method for generating the bounding boxes from the class activation maps and an evaluation of the results comparing three metrics.

This section is organized as follows: Section 4.2.6.1 presents the generation of bounding boxes starting from class activation maps, and Section 4.2.6.2 compares the four class activation map algorithms considering three metrics.

4.2.6.1 *Bounding box generation*

The goal of the presented work is to compare the effectiveness of alternative class activation map algorithms in isolating the salient regions of artwork images that have the greatest impact on the attribution of a specific iconography class. The capacity of a class activation map algorithm to identify precisely the areas of an image that correspond to the whole Saint or to one of the iconography symbols that characterize him/her can help build a training set for the object detection task. The class activation map can be used as a replacement for the manual annotations necessary for creating a detection training set by computing the smallest bounding boxes that comprise the foreground area and using such automatically generated annotations for training an object detector. This approach is known as *weakly supervised* object detection and is an active research area [83].

To investigate the class activation maps' potential in supporting weakly supervised object detection, the region proposals obtained by drawing bounding boxes around the connected components of the class activation maps have been compared visually with the ground truth bounding boxes of the iconographic symbols. For completeness, we have also computed the bounding boxes surrounding all the foreground pixels and compared them with manually created bounding boxes surrounding the whole Saints. The candidate region proposals to use as automatic bounding boxes have been identified with the following heuristic procedure.

1. Collect the images on which all the four methods satisfy a minimum quality criterion: for symbol bounding boxes component IoU greater than 0.165 at threshold 0.1 (see Figure 4.10) and for whole Saint bounding boxes global IoU greater than 0.24 at threshold 0.05 (see Figure 4.12).

2. Compute the Grad-CAM class activation map of the selected images and apply the corresponding threshold: 0.1 for symbol bounding boxes and 0.05 for whole Saint bounding boxes.

3. Only for symbol boxes: split the class activation maps into connected components. Remove the components whose average activation value is less than half of the average activation value of all components. This step filters out all the foreground pixels with low activation that usually correspond to irrelevant areas (Figure 4.25).

4. For each Iconclass category, draw one bounding box surrounding each component (symbol bounding boxes) and one bounding box surrounding the entire class activation map (whole Saint bounding boxes).

In the procedure above, Grad-CAM is chosen to compute the candidate symbol and whole Saint bounding boxes because it has the highest value of the bounding box coverage metrics (together with Smooth Grad-CAM++) and covers wider areas at the same time focusing on the correct details.

4.2.6.2 *Whole Saint bounding boxes*

Figure 4.27 illustrates some examples of computed whole Saint bounding boxes (green) compared with the ground truth boxes (red). The

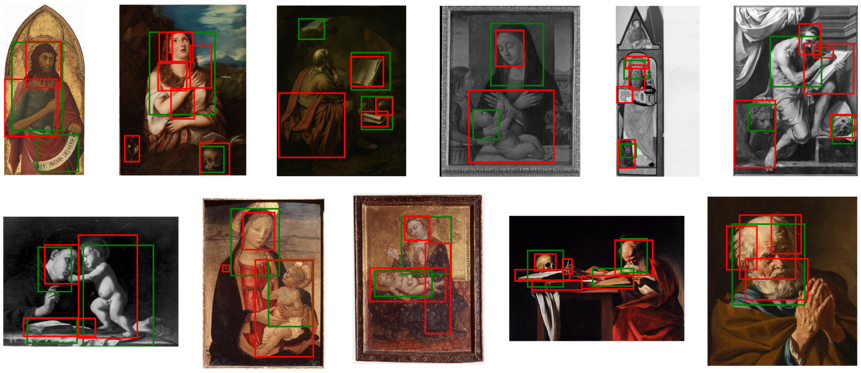


Figure 4.26: Examples of symbols bounding boxes generated from Grad-CAM (green) and manually annotated (red).

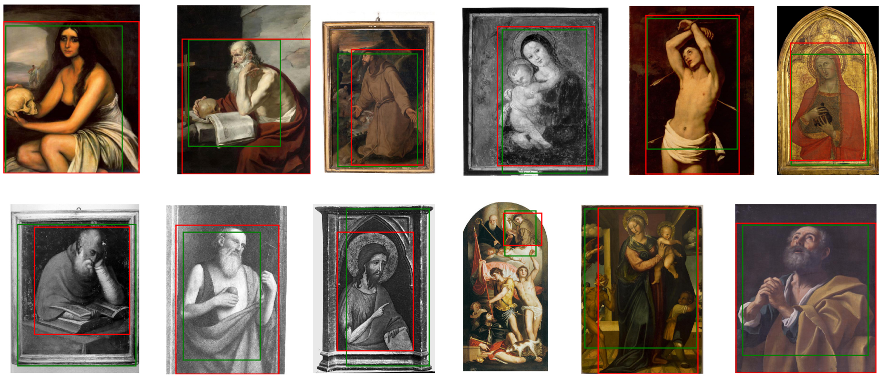


Figure 4.27: Examples of saints bounding boxes generated from Grad-CAM (green) and manually annotated (red).

automatically generated bounding boxes localize almost entirely the Saint's figure and include only very small irrelevant areas.

Figures 4.26 and 4.27 show that the simple procedure for processing class activation maps outputs is sufficient to generate good quality bounding boxes that can act as a proxy to the ground truth for training a fully supervised object detector.

For the whole Saint case, each estimated bounding box can be labelled with the iconography class of the corresponding Saint portrayed in the image. In this way, it is possible to quantify the coincidence between the bounding box of the ground truth and the bounding box computed from the class activation map.

For this purpose, three object detection metrics have been computed: the average IoU value between the GT and the estimated bounding boxes, Mean Average Precision and *GT-known Loc*. The latter is used in several works [4, 16, 68] to evaluate the localization accuracy of object detectors and is defined as the percentage of *correct* bounding boxes. A bounding box is considered correct only when the IoU between the GT box (for a specific class) and the estimated box (for the same class) is greater than 0.5.

Mean Average Precision cannot be used for symbol-level bounding boxes, since the ground-truth bounding boxes are annotated with classes (e.g., *Saint Sebastian - arrows*). On the other hand, the ones defined automatically cannot be associated with a class since the training process concerns only the saints and their classes. Similarly, *GT-known Loc* counts the number of correct bounding boxes with respect to ground-truth labels, but the network can extract only the saint-level labels.

Results are reported in Table 4.5. Grad-CAM confirms as the method with the best performances, Smooth-Grad-CAM++ yields similar results, and CAM is the worst performing method in all the computed metrics. Grad-CAM produces bounding boxes that, on average, have 0.55 IoU with the GT boxes and the *GT-Known Loc* metric shows that 61% of those boxes have an IoU value greater than 0.5.

Figure 4.28 presents the normalized distribution of IoU values for Grad-CAM. We can observe that ~ 83% of the generated boxes have an IoU value greater than 0.3 and that most values are in the range between 0.4 and 0.9, with ~ 12% having an IoU greater than 0.9. Table 4.6 shows the mAP values obtained with Grad-CAM on the ten ArtDL classes. It shows that in general common classes tend to have a higher mAP, consistently with other results.

The whole Saint estimated bounding boxes appear suitable for creating the pseudo ground truth for training an object detector with the weakly supervised approach. Two observations motivate the viability of Grad-CAM for this purpose. As in the GT-known Loc metrics, the goodness of object detection is usually evaluated with a minimal IoU threshold of 0.5. The boxes generated automatically with Grad-CAM obtain 0.55 IoU on average, which suggests that the automatically estimated bounding boxes have a quality similar to the bounding boxes produced by a *fully supervised* object detector, albeit inferior to the quality of the bounding boxes created by humans. Grad-CAM, designed to be an interpretability technique, can be used also to estimate bounding boxes that reach 31.6% mAP on cultural heritage data without any optimization. This finding compares well with the fact that methods designed and optimized specifically for weakly supervised object detection reach values around 14% on artworks data

Table 4.5: Average IoU, GT-Known accuracy and mAP values for the whole Saint bounding boxes estimated with the four analyzed class activation map techniques. The values are calculated with an activation threshold equal to 0.05.

Method	**Average IoU**	**GT-Known Loc (%)**	**mAP** (at IoU $\geqslant 0.5$)
CAM	0.489	49.70	0.206
Grad-CAM	0.551	61.20	0.316
Grad-CAM++	0.529	59.88	0.292
Smooth Grad-CAM++	0.544	61.18	0.307

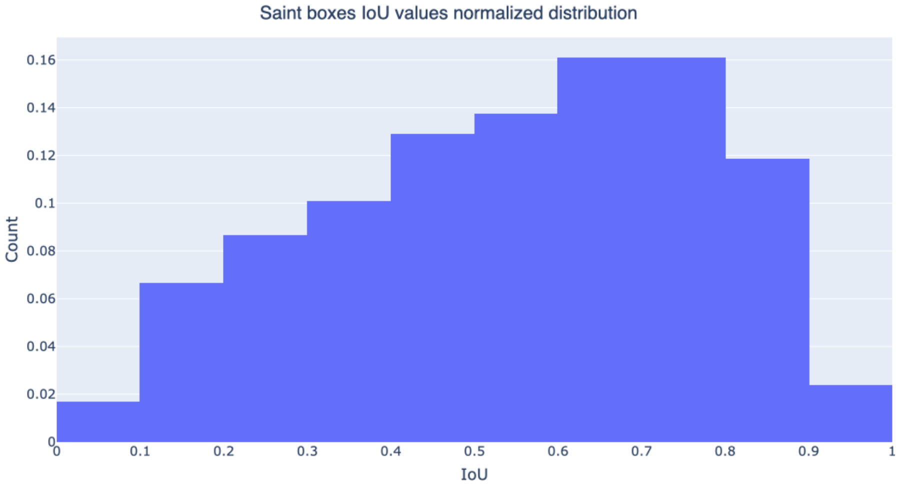


Figure 4.28: Normalized distribution of IoU values between whole-Saint Grad-CAM estimated bounding boxes and ground truth bounding boxes.

sets similar to ArtDL [28, 30]. For this reason, simple and generic techniques, such as Grad-CAM, which can localize multiple Saint instances and even multiple characteristic features, are a promising starting point for advancing weakly supervised object detection studies in the cultural heritage domain.

To conclude, this section has shown that Class Activation Map methods can be used effectively to create bounding boxes based on a simple heuristic procedure.

Table 4.6: Mean Average Precision (mAP) values for each class of the ArtDL data set. Bounding boxes are estimated with Grad-CAM.

Saint	Mean Average Precision
Anthony	0.076
John	0.289
Paul	0.173
Francis	0.330
Magdalene	0.616
Jerome	0.228
Dominic	0.142
Virgin	0.442
Peter	0.399
Sebastian	0.468

4.3 QUALITATIVE ANALYSIS

This section presents a qualitative analysis of the results obtained by the different class activation maps algorithms and highlights their capabilities and limitations. The examples concerning class activation maps, if not otherwise specified, show the original image, the class activation maps generated by each algorithm (with the background in black and foreground in white) and the ground truth bounding boxes. In this section, the generation of symbol-level bounding boxes serves the purpose of performing qualitative analyses.

This section is organized as follows: Section 4.3.1 presents qualitative results concerning the class activation maps, Section 4.3.2 presents qualitative results concerning the whole Saint bounding boxes generated from the class activation maps, and Section 4.3.3 presents qualitative results concerning the symbol bounding boxes generated from the class activation maps.

4.3.1 *Class Activation Maps*

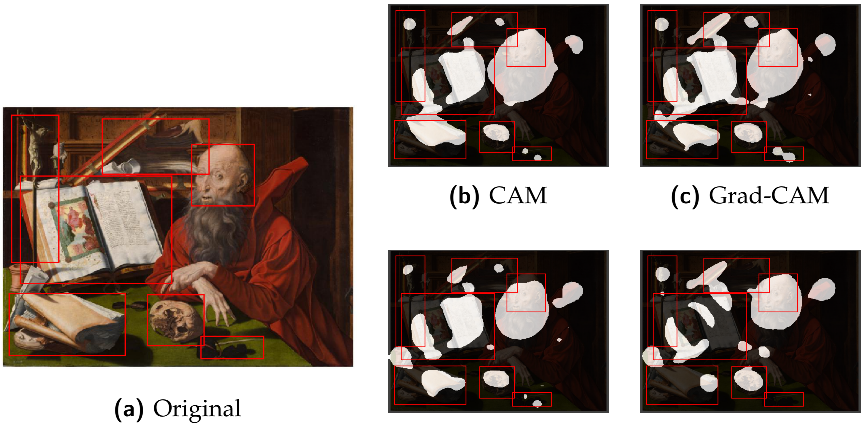
This section presents qualitative analyses of class activation maps and is organized as follows: Section 4.3.1.1 presents positive examples; Section 4.3.1.2 presents negative examples; Section 4.3.1.3 presents an example of sculpture with multiple instances; Section 4.3.1.4 presents some examples of symbols not indicated in the ground truth and found by the Class Activation Map algorithms; Section 4.3.1.5 presents the case of co-occurring classes.

4.3.1.1 *Positive examples*

Figure 4.29[9] shows an example in which all the algorithms focus well on the iconography symbols. The image contains seven symbols with different size, shape, and position, all identified and separated by the class activation map algorithms. The irrelevant area on the top right corresponds to a portion of the cardinal's vest that has the same colour and approximate shape of the cardinal's galero appearing in many paintings of Saint Jerome.

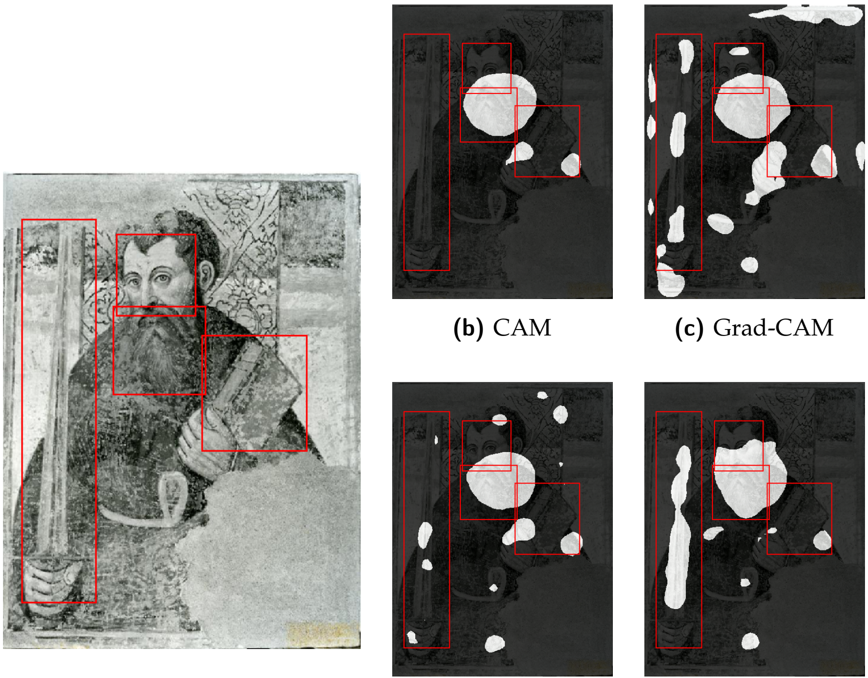
Figure 4.30[10] shows an example in which all the algorithms perform well on a painting in which the visibility of the symbols is very low. All class activation maps algorithms identify at least two out of the three symbols. Only CAM misses the sword, which the other algorithms identify by focusing on the hand holding it or on the sword blade. The example of Figure 4.30 and many similar ones of black and white and poor quality images highlight the ability of class activation maps algorithms to extract useful maps also when the image has low discriminative features.

[9] Saint Jerome in his Study, Jan van Remmerswale, 1533.
[10] St. Paul, nd, ca. 1510.



(a) Original (b) CAM (c) Grad-CAM (d) Grad-CAM++ (e) Smooth GC++

Figure 4.29: Class activation maps with seven recognized symbols associated with Saint Jerome.



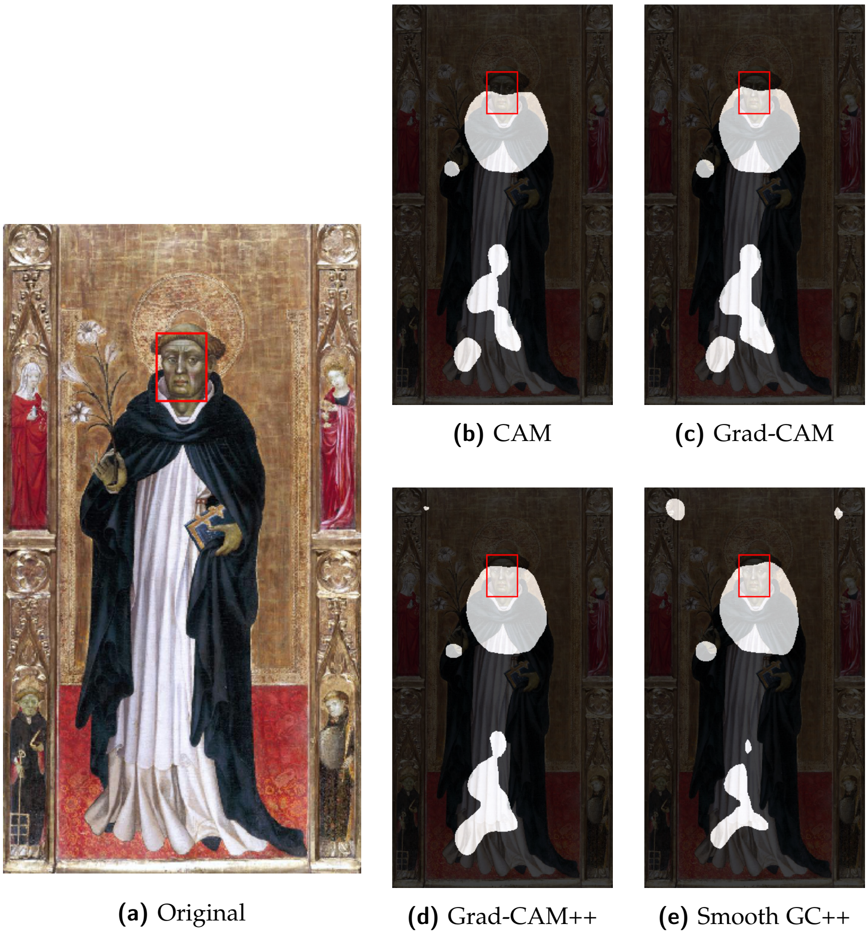
(a) Original (b) CAM (c) Grad-CAM (d) Grad-CAM++ (e) Smooth GC++

Figure 4.30: Class activation maps extracted from a drawing of Saint Paul. Four out of five symbols are identified despite their low visibility.

Figure 4.31[11] illustrates a counterexample of the difficulty of detecting such generic attributes as the vest. The vest is identified thanks to

[11] Saint Dominic Guzmán and Four Saints, Guerau Gener, 1405.

a specific detail: the change of colour typical of the black and white Dominican habit.


(a) Original

(b) CAM

(c) Grad-CAM

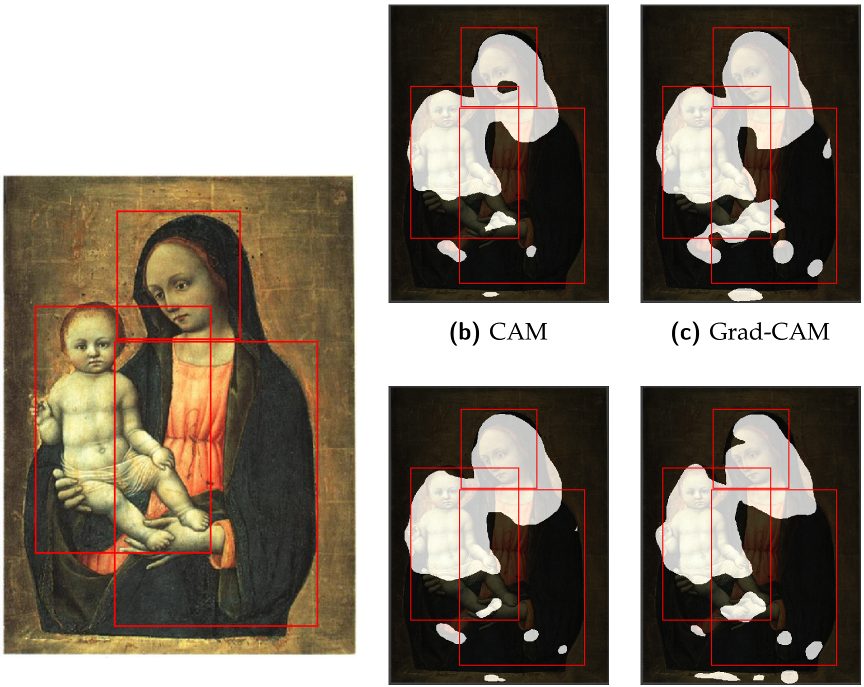
(d) Grad-CAM++

(e) Smooth GC++

Figure 4.31: Class activation maps extracted from a paining of Saint Dominic. The rather generic vest attribute is identified by focusing on its double color.

4.3.1.2 *Negative examples*

Class activation maps algorithms tend to fail consistently in two cases: when multiple symbols are too close or have substantial overlap and when the representation of a symbol is rather generic and covers a wide area of the image.



(a) Original

(b) CAM

(c) Grad-CAM

(d) Grad-CAM++

(e) Smooth GC++

Figure 4.32: Class activation maps with merged symbols and missed generic attributes.

Figure 4.32[12] illustrates a typical example: Virgin Mary's face and Baby Jesus are merged into a single region, while the vest, which is a rather generic attribute, is missed altogether or highlighted only through small irrelevant details.

4.3.1.3 *Multiple instances*

A few artworks contain multiple instances of the same saint (i.e., the same character is present multiple times in the artwork). In addition, some artworks contain scenes where the characters have similar poses and are associated with similar symbols. A notable example is presented in Figure 4.33,[13] which represents Saint Anne

[12] Madonna col Bambino, Antonio Vivarini, ca. 1441.

[13] This sculpture is located in the altarpiece of Kvernes stavkyrkje, in the Averøy Municipality, Møre og Romsdal, Norway. The original stave church was built during the first half of the 14th century, while there are no information on the author or realization period of the sculpture. Image source: `https://www.stavechurch.com/wp-content/uploads/2017/12/Maria-og-Anna.jpg`.

(on the left) and the adult Virgin Mary (on the right). In addition, Virgin Mary is depicted, as a child, also in the left panel, in the arms of Saint Anne.

Virgin Mary is recognized by all the class activation maps algorithms when depicted as an adult, in the right panel. Moreover, Grad-CAM, Grad-CAM++, and Smooth Grad-CAM++ highlight her as the daughter of Saint Anne, in the left panel. The model is confused by the presence of Saint Anne, which assumes the typical pose of Virgin Mary and that holds, in her right arm, a baby, which is likely Baby Jesus. Notably, Baby Jesus is recognized in both panels and with a similar class activation map intensity.

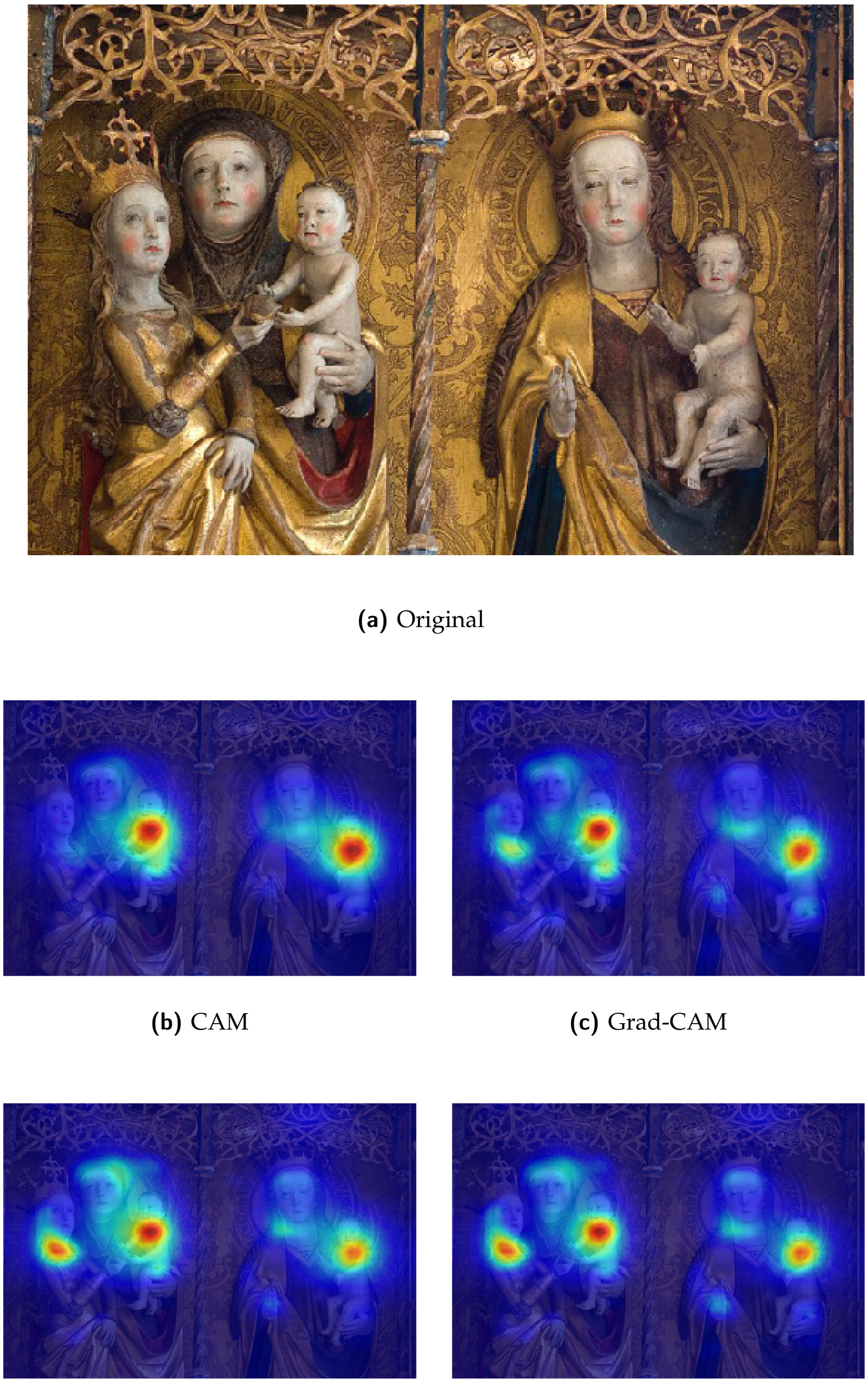
The choice of differentiating the most prominent regions of the artwork using gradients emphasizes the focus on the main characters and symbols more effectively than in single-instance examples.

4.3.1.4 *Relevant irrelevant regions*

An interesting case occurs when the class activation maps algorithms focus on a seemingly irrelevant area which, instead, contains a relevant iconography attribute not present in the ground truth. Figure 4.34 illustrates three examples. The painting of Saint John the Baptist (a)[14] contains a seemingly irrelevant area in the top left, which focuses on a bird. This is a less frequent attribute of the Saint that is not listed in the iconography symbols used to annotate the images but appears in some paintings. The same happens with Saint Jerome (b),[15] where the class activation maps algorithms highlight an hourglass, an infrequent symbol present only in a subset of the ArtDL images and not used in the annotation. Finally, another case occurs with the iconography of

[14] Portrait of François I as St John the Baptist, Jean Clouet, 1518.

[15] Saint Jerome, Albrecht Durer, 1521.



(a) Original

(b) CAM

(c) Grad-CAM

(d) Grad-CAM++

(e) Smooth Grad-CAM++

Figure 4.33: Class activation maps for multiple instances and similar poses and symbols.

Saint Jerome (c),[16] where the class activation maps algorithms focus on the outdoor environment. This is a well-known symbol associated

[16] Landscape with St. Jerome, Simon Bening, 1515-1520.

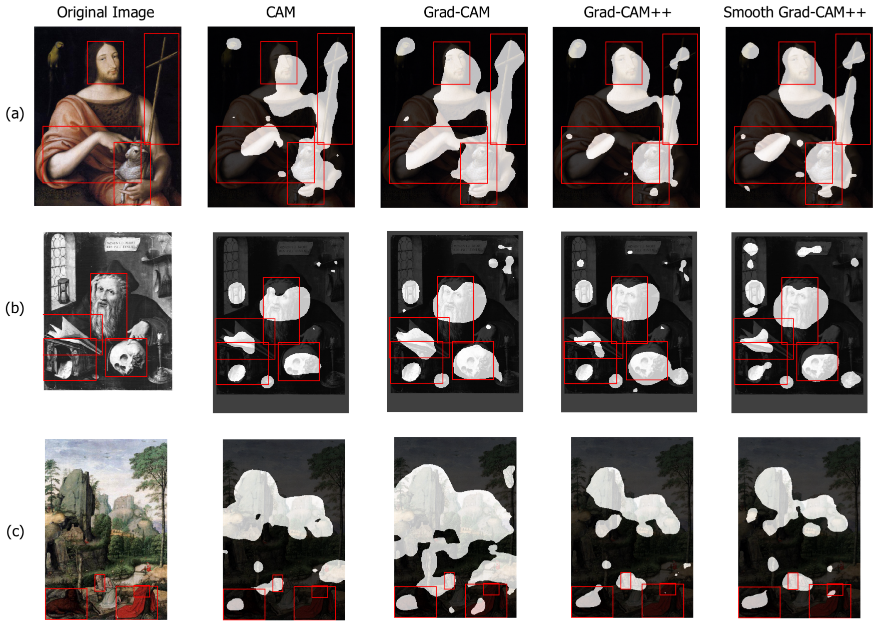


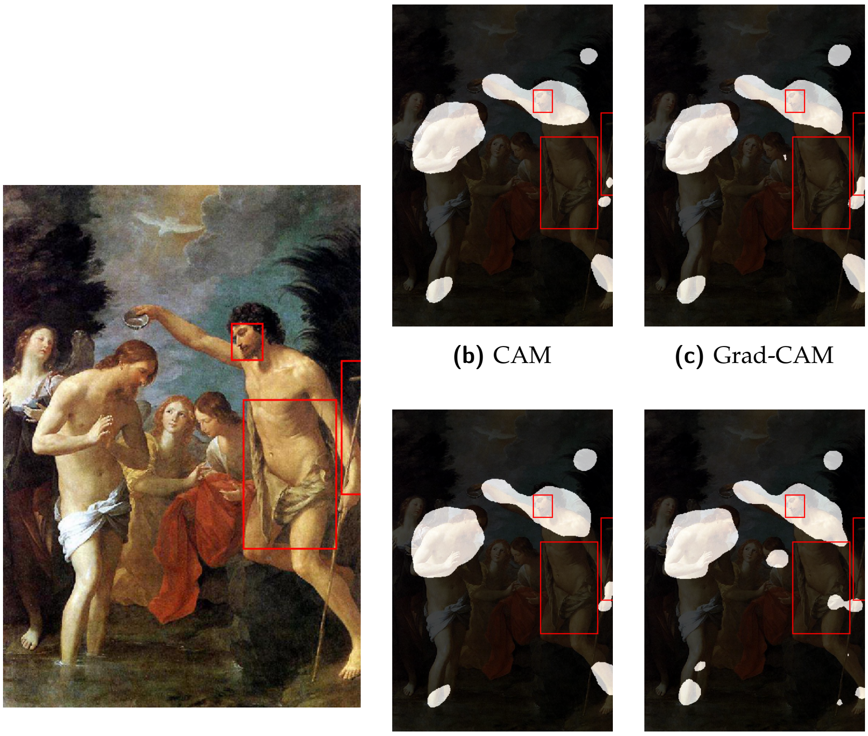
Figure 4.34: Class activation maps highlighting regions containing relevant iconography attributes not present in the ground truth: a bird associated with Saint John the Baptist (a) an hourglass associated with Saint Jerome (b) and the wilderness where Saint Jerome retired (c).

with the Saint, who retired in the wilderness, but one that is hard to annotate with bounding boxes and thus purposely excluded from the ground truth annotations.

4.3.1.5 *Confusion with unknown co-occurring class*

Figure 4.35[17] presents an example in which all analysed variants make confusion between Saint John the Baptist and Jesus Christ. The latter is an Iconclass category too, but not one represented in the ArtDL data set. Given the prevalence of paintings depicting Saint John the Baptist in the act of baptizing Christ over those where the Saint occurs alone, the CAM output highlights both the figures. This ambiguity would reduce if the data set were annotated with the Iconclass category for Jesus.

[17] Baptism of Christ, Guido Reni, ca. 1622-1623



(b) CAM

(c) Grad-CAM

(a) Original

(d) Grad-CAM++

(e) Smooth GC++

Figure 4.35: Class activation maps with confusion between Saint John the Baptist and Jesus Christ.

4.3.2 *Whole Saint bounding boxes*

This section presents qualitative examples of whole Saint bounding boxes generated from class activation maps and is organized as follows: Section 4.3.2.1 presents positive examples; Section 4.3.2.2 presents negative examples.

4.3.2.1 *Positive examples*

This section presents a deeper analysis of some of the examples presented in Figure 4.27.

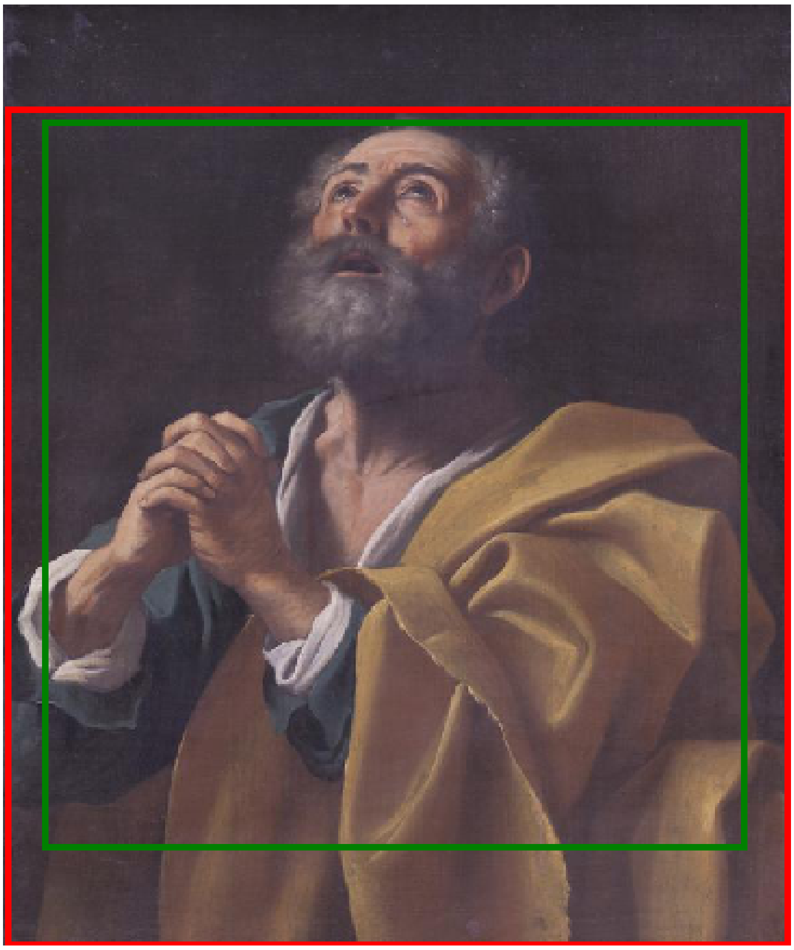


Figure 4.36: A positive example of saints bounding boxes generated from Grad-CAM (green) and manually annotated (red). This painting, depicting Saint Peter, shows good accordance between the generated bounding box and the GT bounding box. Only the peripheral parts of the saint are not included inside the generated bounding box.

SAINT PETER: Figure 4.36[18] presents a positive example in terms of the whole bounding box generation. In this case, the prominent features of the saint (e.g., the bushy hair and beard) are included in the bounding box, since they allow to discern Saint Peter from other saints. The saint's clothes are not as distinctive, so the bounding box does not include their lower part.

[18] San Pietro in preghiera, Sisto Badalocchio, ca. 1628-1629.

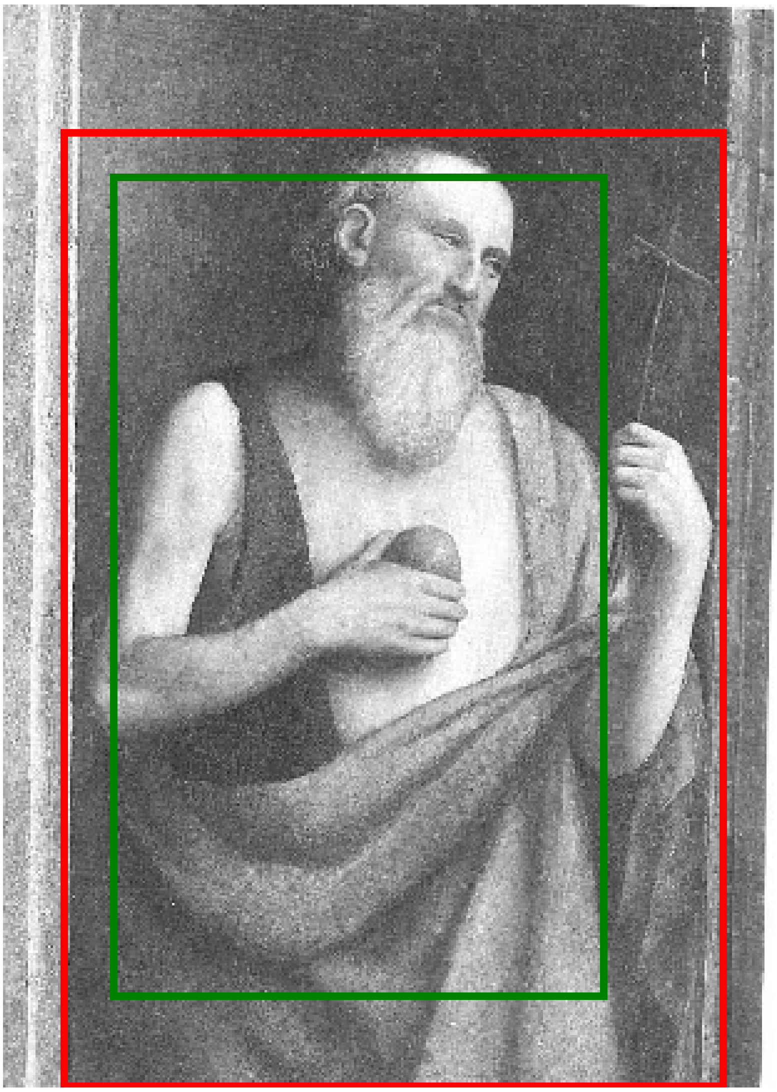


Figure 4.37: A positive example of saints bounding boxes generated from Grad-CAM (green) and manually annotated (red). This painting, depicting Saint Jerome, shows good accordance between the generated bounding box and the GT bounding box. Only the peripheral parts of the saint are not included inside the generated bounding box.

SAINT JEROME: Figure 4.37[19] presents a positive example in terms of the whole bounding box generation. In this case, the prominent features of the saint (e.g., the stone in his hand) are included in the bounding box since they allow to discern Saint Jerome from other saints. The saint's clothes are not as distinctive, so their lower part is excluded from the bounding box. Differently from Saint Peter, Saint Jerome face does not have peculiar characteristics. For this reason, the generated bounding box contains only part of it.

4.3.2.2 *Negative examples*

This section analyzes some negative examples in which the generated bounding boxes differ significantly from the GT bounding boxes.

VIRGIN MARY: Figure 4.38[20] presents a negative example in terms of the whole bounding box generation. In this case, the generated bounding box covers a wide area of the painting due to secondary characters in the lower part of the artwork. Such characters are not iconographical symbols, but it is not uncommon that Virgin Mary is depicted in a more complex scene. For example, this painting represents the Coronation of the Virgin, which has changed across the centuries and often contains saints and angels [63].

SAINT SEBASTIAN: Figure 4.39a[21] presents a negative example in terms of the whole bounding box generation. In this case, the generated bounding box covers a restricted area of the painting, focusing on the arrows plunged into Saint Sebastian legs and torso. The

[19] San Gerolamo, Bernardino Luini, ca. 1515-1520.

[20] Madonna del Rosario con San Francesco e Santa Chiara, Federico Fiori called Federico Barocci (attributed), ca. 1535.

[21] Saint Sebastian, Cosme Tura, ca. 1484.

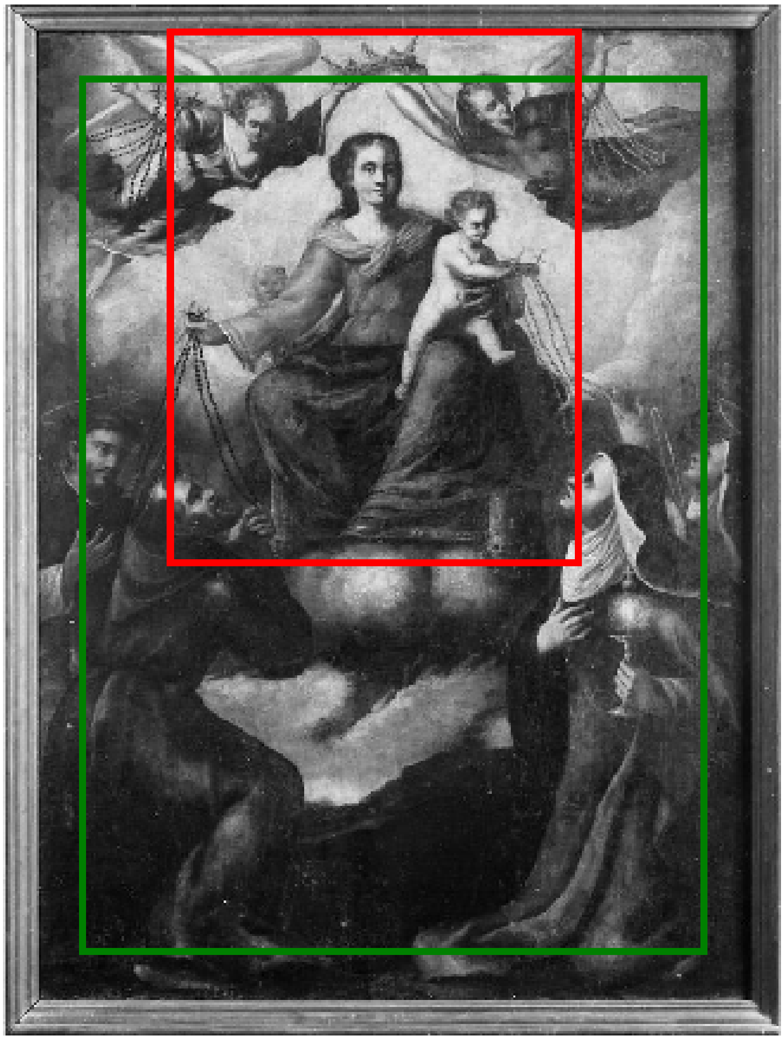
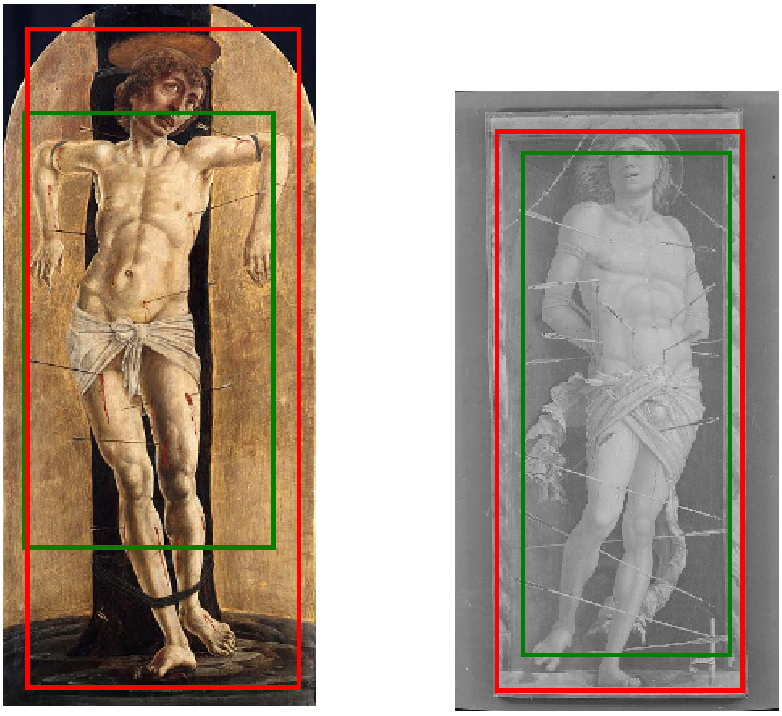


Figure 4.38: A negative example of saints bounding boxes generated from Grad-CAM (green) and manually annotated (red). This painting, depicting Virgin Mary, shows a generated bounding box that is much wider than the GT bounding box due to the presence of recurrent secondary characters.



(a)

(b)

Figure 4.39: Two examples of saints bounding boxes generated from Grad-CAM (green) and manually annotated (red). Those paintings, depicting Saint Sebastian, show a generated bounding box that is smaller than the GT bounding box due to the concentration of symbols below the head and above the feet.

head, instead, is not included in the bounding box because it is not a peculiar characteristic of this saint. Also Figure 4.39b[22] shows the same phenomenon. However, the bounding box is better than the other one because the arrows are longer and cause the bounding box to be wider. The bounding box not covering the entire face confirms that it is not a prominent feature.

4.3.3 *Symbol bounding boxes*

Figure 4.26 presents some examples of the computed symbol bounding boxes (green) compared with the ground truth bounding boxes (red). The proposed procedure can generate boxes that in many cases correctly highlight and distinguish the most important iconography symbols present in the images. When the symbols are grouped in a small area (e.g., the bushy hair and beard of Saint Peter), the procedure tends to generate one component that covers all of them, thus creating only one bounding box. Sometimes, elements in the image that have not been manually annotated in the ground truth are correctly detected (e.g., the scroll in the hand of Saint John the Baptist in the first painting of Figure 4.26).

This section is organized as follows: Section 4.3.3.1 presents positive examples; Section 4.3.3.2 presents negative examples.

4.3.3.1 *Positive examples*

This section analyzes some of the examples presented in Figure 4.26.

[22] San Sebastiano, Andrea Mantegna (attributed), ca. 1450-1499.

VIRGIN MARY: Figure 4.40[23] presents a positive example in terms of the whole bounding box generation. In this case, Baby Jesus, a prominent iconographical symbol of the saint, is characterized by a rather precise bounding box, including most of the body. The other generated bounding box, on the other hand, includes the face and part of the blue garment, another prominent symbol. Likely, the generated bounding boxes don't extend further because the class activation maps, once normalized, tend to give more importance to the most common symbols at the expense of less common ones.

SAINT JEROME: Figure 4.41[24] presents a positive example in terms of the whole bounding box generation. Almost all the symbols have been found correctly, except for the quill (part of the writing material symbol). However, close symbols have been merged, as observed for the skull, the closed book and the open book. This phenomenon occurs likely because of the low contrast between different objects and since class activation maps spread more widely than the output of an ideal segmentation, making the object boundaries more arduous to define.

4.3.3.2 *Negative examples*

This section analyzes some negative examples in terms of symbol-level bounding boxes, focusing on typical issues.

SAINT PETER: Figure 4.42[25] presents a negative example in terms of the whole bounding box generation. In this case, both the key and

[23] Madonna in trono con Bambino, Maestro Dei Polittici Crivelleschi, ca. 1482.

[24] San Girolamo in meditazione, Caravaggio, ca. 1605.

[25] Saint Peter, Benvenuto di Giovanni, 1479.

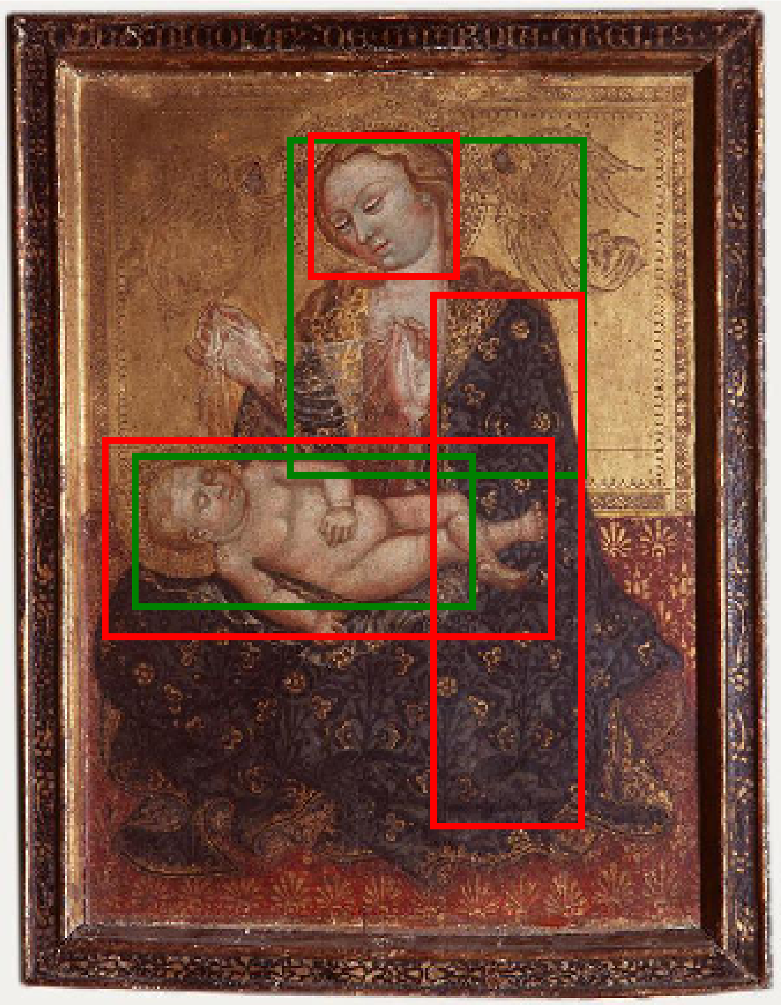


Figure 4.40: A positive example of symbols bounding boxes generated from Grad-CAM (green) and manually annotated (red). This painting, depicting Virgin Mary, shows good accordance between the generated bounding boxes and the GT bounding boxes, especially in the detection of Baby Jesus.

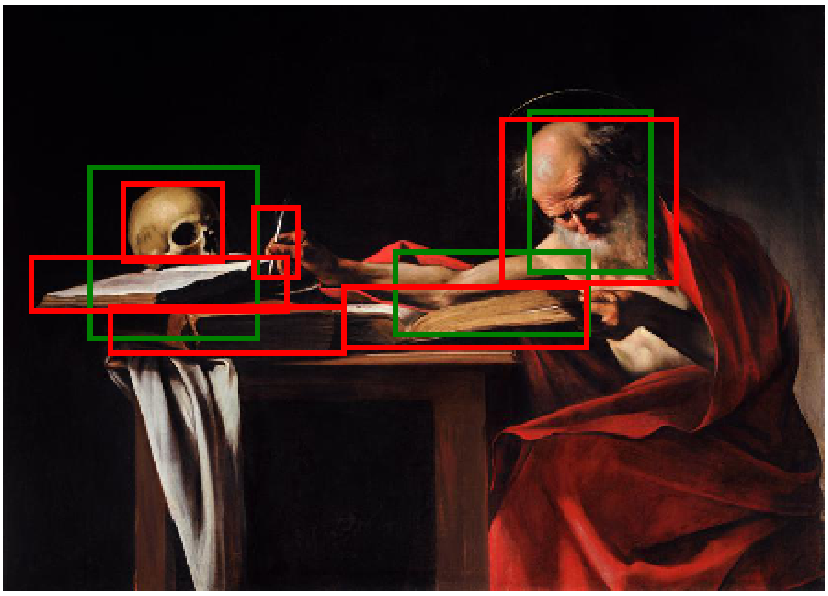


Figure 4.41: A positive example of symbols bounding boxes generated from Grad-CAM (green) and manually annotated (red). This painting, depicting Saint Jerome, shows good accordance between the generated bounding boxes and the GT bounding boxes, but some of the symbols are grouped together.

the book are not contained in any bounding box. This phenomenon is likely caused by the abundance of the bushy hair and bushy beard symbols across the data set, while the book characterizes many saints, and the keys appear only in some paintings containing Saint Peter. Those symbols, hence, are less decisive in terms of classification.

SAINT FRANCIS: Figure 4.43[26] presents a negative example in terms of the whole bounding box generation. In this case, none of the generated bounding boxes contains the skull and one of the stigmata. The gaze of the saint, oriented towards the sky, indeed, is peculiar of Saint Francis. Other saints usually observe the spectator (e.g., Saint Peter and Virgin Mary) or objects inside the painting (e.g., Saint Jerome looks at books or to the ground). This distinctive feature likely prevails

[26] Saint François d'Assise en oraison devant un crucifix, Francesco Albani, ca. 1640-1650.

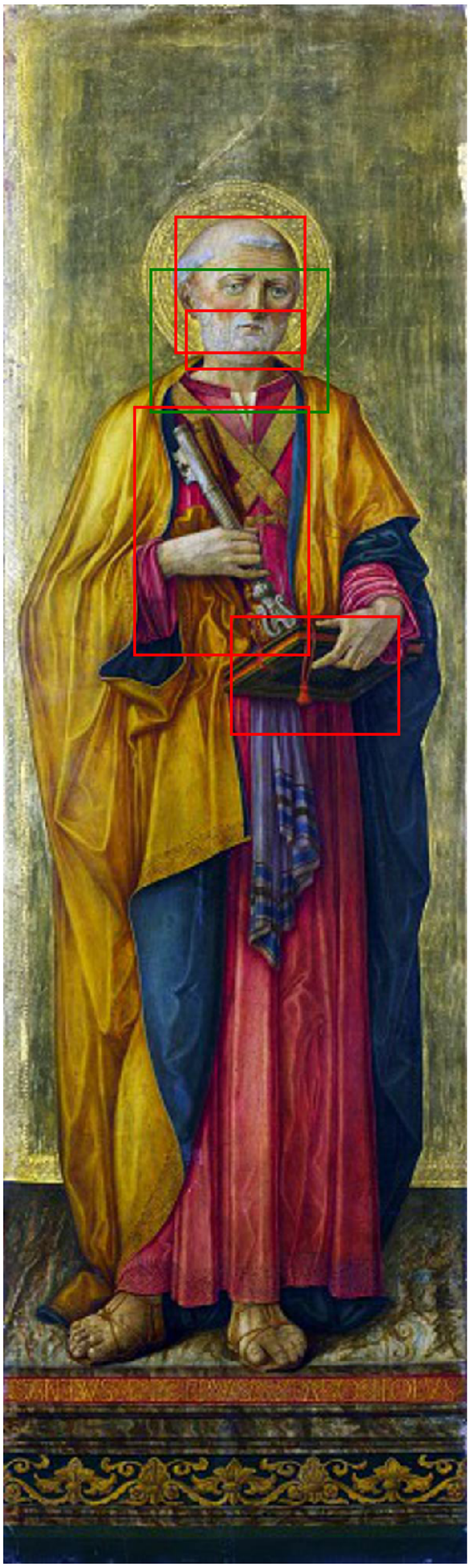


Figure 4.42: A negative example of symbols bounding boxes generated from Grad-CAM (green) and manually annotated (red). In this painting, depicting Saint Peter, the key and the book are not found automatically, although they are relevant symbols.

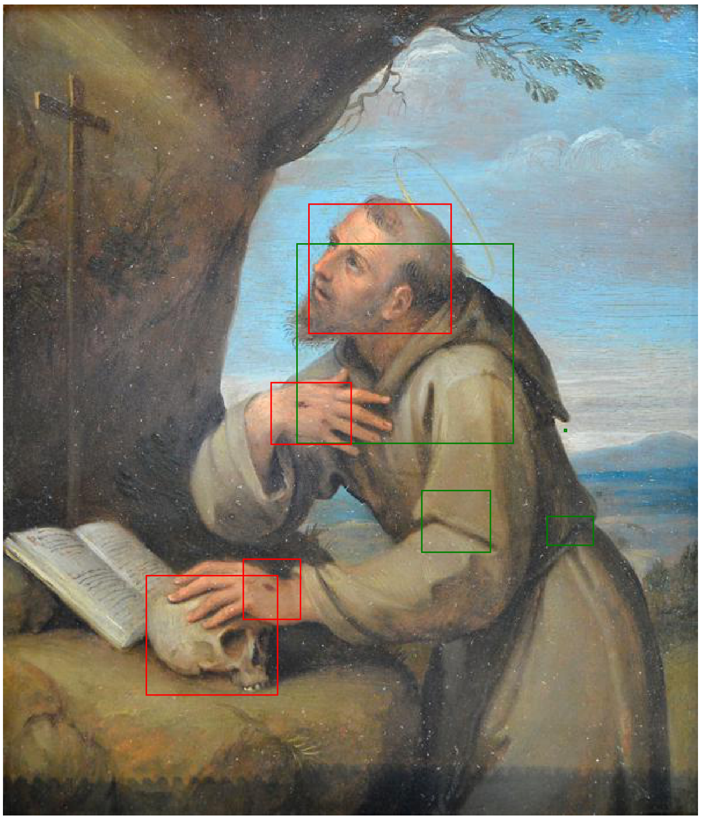


Figure 4.43: A negative example of symbols bounding boxes generated from Grad-CAM (green) and manually annotated (red). In this painting, depicting Saint Francis, one of the stigmata and the skull are not contained in any bounding box.

on small symbols far away from the face (one of the stigmata) and the ones characterizing multiple saints (e.g., the skull, a symbol of Saint Mary Magdalene). On the other hand, the creases on the saint cloth can be mistaken by parts of the characteristic cord he wears, partially hidden in this painting.

5

CONCLUSIONS AND FUTURE WORK

This chapter presents the conclusions drawn from this research and proposes future promising directions of research, in the context of a wider project, whose goal is to be able to explain the meaning of artworks automatically. Section 5.1 presents the contributions of the work and how it has contributed to answering the research questions identified in the Introduction. Section 5.2 outlines promising research directions, while Section 5.3 presents the context of this work, emphasizing the contribution of this research in the overall project.

5.1 CONTRIBUTIONS

This work has presented a comparative study about the effectiveness of class activation maps as a tool for explaining how a CNN-based classifier recognizes the Iconclass categories present in images portraying Christian Saints. The symbols relevant to the identification of the Saints were annotated with bounding boxes and the output of the class activation maps algorithms were compared to the ground truth using four metrics. The analysis shows that Grad-CAM achieves better results in terms of global IoU and covered bounding boxes and Smooth Grad-CAM++ scores best in the component IoU thanks to its precision in delineating individual small size symbols. The irrelevant attention metrics promotes the original CAM algorithm as the best approach, but the low component IoU and box coverage complement such an evaluation showing that CAM covers too small areas. While for natural images, Smooth Grad-CAM++ outperforms the other three algorithms [52], in our use case, Grad-CAM is the method of choice for deriving from class activation maps the bounding boxes necessary to train a weakly supervised detector.

This work has addressed the research questions presented in the Introduction:

- *Are CAMs an effective tool for understanding how a CNN classifier recognizes the iconography classes of a painting?*

  CAMs are effective for understanding how a CNN classifier recognizes the iconography classes of a painting since they highlight the most relevant areas of the painting and since from them it is possible to obtain bounding boxes surrounding the most relevant symbols.

- *Are there significant differences in the state-of-the-art CAM algorithms with respect to their ability to support the explanation of iconography classification by CNNs?*

  The differences among different CAM algorithms are not exceedingly significant, but this work shows that in general Grad-CAM yields better results both quantitatively and qualitatively.

- *Are the image areas highlighted by CAMs a good starting point for creating semi-automatically the bounding boxes necessary for training iconography detectors?*

  The image areas highlighted by CAMs are a good starting point for creating bounding boxes, but the extraction of better bounding boxes may rely either on new algorithms for extracting the bounding boxes or on other activation map algorithms since there is margin for improvement in terms, e.g., of the GT Loc metrics.

The contributions can be summarized as follows:

- Application of different state-of-the-art class activation map algorithms on 10 classes of the ArtDL data set;
- Creation of a test data set, comprising 823 annotated images annotated with 2957 bounding boxes surrounding specific iconographic symbols;
- Creation of an additional test data set, comprising 823 annotated images annotated with 882 bounding boxes surrounding specific saints;
- Quantitative analyses of different algorithms concerning the symbol bounding boxes;
- Quantitative analyses of different algorithms concerning the saint bounding boxes;
- Comparison of the results of the different class activation map algorithms with natural images;
- Qualitative evaluation aimed at identifying the strengths and weaknesses of the different class activation map algorithms.

5.2 FUTURE WORK

Future work will concentrate on the comparison of other activation mapping techniques [4, 16, 76, 86] and on devising precise bounding boxes surrounding the iconographical symbols. In particular, the studies presented in [16] and [4] are based on the re-training of the network, an approach quite different from the currently analyzed alternatives.

The results of the CAMs algorithms selection can lay the foundation to pursue the ultimate goal of our research, which is to use the output of class activation maps to create training data sets for weakly supervised iconography symbol detection and segmentation. An automated system for iconography analysis of artworks could promote educational applications development for Art History experts and students. Another future research path consists of addressing more complex Iconclass categories involving complex scenes (e.g., the crucifixion, the nativity, the visitation of the magi, etc.) and exploring the iconography of other cultures.

Among the proposed directions of research, the ability to create better bounding boxes starting from the class activation maps lays the foundations for thorough analyses of artworks. Several methods deal with the problem of fine-grained visual categorization (e.g., [27, 40, 84]), which is relevant because it focuses on confined areas of the images. In the context of artworks, such areas correspond to the relevant symbols associated with saints. Another approach consists of changing the network structure, to introduce additional components aiming at directing the focus on the most relevant areas of the painting. The more general methods trying to change the network layout are known as attention mechanisms [78].

The next sections analyze those alternatives more in detail. Section 5.2.1 presents promising fine-grained categorization methods, while Section 5.2.2 presents different recently proposed attention mechanisms.

5.2.1 *Fine-grained categorization*

This section introduces some promising fine-grained categorization approaches which, given an input image, can extract bounding boxes surrounding its most relevant parts.

Ge et al. [27] proposed to extract an initial segmentation from a class activation map. The initial segmentation is then passed through Mask R-CNN to obtain a probability map (i.e., a map that aims at covering the object more precisely than the class activation map) and, from this, a bounding box and segmentation. They propose the application of their method on two data sets containing natural images of animals and on one more generic data set. Their approach also requires the implementation of LSTM to combine the outputs of Mask R-CNN for generating the probability map. This method is promising in their results, but it would be necessary to verify whether their results are as promising when considering artworks and when dealing with multiple objects in the same image, since they show only single-instance classification examples.

Korsch et al. [40] proposed a different approach, based on both the initial prediction and the back-propagation of feature importance. This research considers two bird data sets, a car data set and a flowers data set. This research is interesting because it shows promising results using ResNet-50 as a backbone in the case of the cars data set. Their proposal, differently from Ge et al.'s pipeline, focuses on the computation of a sparse saliency map, from which they extract the bounding boxes of the parts of the image. Similarly to Ge et al., this research's examples contain a single object, and applying this approach to non-natural images containing multiple objects would also be an interesting research path.

Other methods rely on subsequent crops of the original image, to find the relevant part images. Zhang et al. [84] research shows that their method can achieve higher localization accuracy than other state-of-the-art methods without adding trainable parameters. Their proposal is based on the introduction of two novel modules, used for predicting the position of an object and for the attention part proposal. This approach is applied to images concerning specific fields (i.e., cars, birds, and aircraft), which does not allow to accurately estimate its performances on ArtDL. Moreover, similarly to the other presented approaches, all the examples focus on a single object per image.

5.2.2 *Attention mechanisms*

This section presents some recently proposed attention mechanisms. In general, their purpose is to generate activation maps, possibly changing the structure of the network. From activation maps, it is possible to extract bounding boxes, for example as shown in this research.

Bae et al. [3] proposed three changes to a CNN to improve the quality of the activation maps:

- Replacing GAP with Thresholded Average Pooling (TAP): since the score associated with a small relevant area in an image is low, CAMs compute a high corresponding weight to compensate it. For this reason, larger areas are considered relatively less important than smaller ones, even when they are associated with the same maximum score. TAP determines the score associated with a smaller area considering only the part of that area above a certain threshold;

- Introducing Negative Weight Clamping (NWC): since a class activation map is the weighted average of feature maps, such feature maps may also be associated with negative weights. The authors observed experimentally that not considering the feature maps associated with negative weights brought to better results, hence they neglect them;
- Introducing percentile: while the creation of bounding boxes traditionally relies on fixed thresholds, independent of the image (i.e., the class activation map always has values below a fixed threshold t outside the bounding box), this research proposes to create bounding boxes considering a threshold given by, e.g., the 90th percentile. This threshold depends on the distribution of the class activation map values.

This research is particularly promising, since it proposes three different methods, which can be applied separately, and their results can be studied on the ArtDL data set. In particular, the percentile method allows to compare results without re-training the network or re-computing the activation maps, hence it may be applied to the activation map methods presented in this thesis.

A completely different method, Attention Dropout Layer (ADL) was proposed by Choe and Shim [17]. The idea behind their method is to re-train the network, hiding the parts deemed most relevant at that moment, so that during the training the network can focus on other relevant parts. Given an input, the algorithm randomly chooses whether to hide parts of the image or using the entire image, for each iteration. This method may allow focusing on more symbols than the current ones, but in the case of multiple saints in an image, it may attribute symbols to the wrong saint, since they may be loosely related to him/her, and given more importance by ADL.

One possibility consists of combining different methods, comparing the results on the ArtDL data set. For instance, the percentile can be combined with Grad-CAM or with ADL. Such possibilities have not been attempted so far and may give promising results also in the analysis of natural images.

5.3 RESEARCH CONTEXT

The study described in this thesis fits in the context of a broader research project, whose final goal consists of devising a textual description of the artwork meaning, given the picture of an artwork as an input.

Figure 5.1 provides an overview of the main steps of the overall project, which includes class activation maps calculation. In particular, class activation mapping analyses rely on a previous data set preparation and a training procedure. The subsequent research steps consist of the creation of bounding boxes, given the activation maps, the extension of the proposed methodology to a more extensive data set, conduction of further analyses to improve the previous result, and, finally, the application in the field of digital humanities (e.g., historical and cultural studies or cultural heritage management operations). Each step in the presented workflow also includes the results' evaluation, which relies on different metrics.

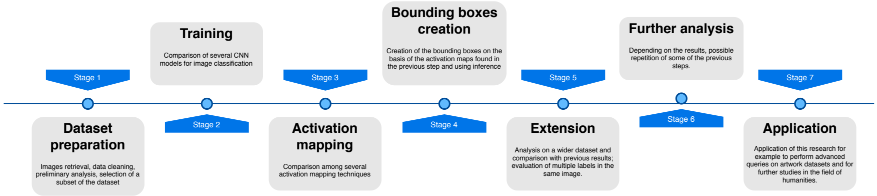

Stage 1
Dataset preparation
Images retrieval, data cleaning, preliminary analysis, selection of a subset of the dataset
Training
Comparison of several CNN models for image classification
Stage 2
Stage 3
Activation mapping
Comparison among several activation mapping techniques
Bounding boxes creation
Creation of the bounding boxes on the basis of the activation maps found in the previous step and using inference
Stage 4
Stage 5
Extension
Analysis on a wider dataset and comparison with previous results; evaluation of multiple labels in the same image.
Further analysis
Depending on the results, possible repetition of some of the previous steps.
Stage 6
Stage 7
Application
Application of this research for example to perform advanced queries on artwork datasets and for further studies in the field of humanities.

Figure 5.1: Overall project workflow – the research described in this thesis refers to the implementation of class activation maps (Step 3).

The study described in this thesis is meaningful since it performs quantitative and qualitative analyses of different class activation map algorithms., which are relevant because bounding boxes creation often relies on such algorithms. In particular, a quantitative comparison of class activation maps is necessary to determine which ones cover the correct area in different scenarios, given an activation threshold. Combining qualitative and quantitative comparisons, in turn, allows determining whether such class activation maps locate the artwork subject (i.e., a saint and the associated symbols). The coverage analysis would support the definition of the boundaries of a target object and, consequently, an appropriate bounding box surrounding it. Bounding boxes can be later used to re-train networks and be explored in the definition of more robust evaluation protocols to support comparisons with state-of-the-art approaches (e.g., the work of Milani and Fraternali [47]).

The main advantage in using the workflow proposed in Figure 5.1 consists of avoiding the manual creation of the bounding boxes for every image and later using the areas obtained by Class Activation Maps to train an object detection model based on automatically generated bounding boxes. In particular, to achieve the initial results, the overall research relies on a fine-tuning approach on Resnet50 [33], which exploits transfer learning, as proposed by Milani and Fraternali in [47]. In this case, high-level features, such as faces, objects, and animals are in the topmost layers of the pre-trained network without the need of completely retraining the proposed model.

A

APPENDIX

The appendix further analyzes the hyper-parameters of Smooth Grad-CAM++, showing quantitative results of the four metrics presented in Section 4.2. The appendix is organized as follows: Section A.1 presents the results for the component IoU metrics, Section A.2 presents the results for the global IoU metrics, Section A.3 presents the results for the bounding box coverage metrics, Section A.4 presents the results for the irrelevant attention metrics, and Section A.5 summarizes the main findings concerning Smooth Grad-CAM++.

A.1 COMPONENT IOU

This section presents the component IoU results for different values of the Smooth Grad-CAM++ hyper-parameters, as a deepening of the results presented in Section 4.2.

Figure A.1 presents an overview of the component IoU results for different thresholds. Its purpose is to emphasize the small differences introduced by changing the number of samples s and the standard deviation σ. As observed in the component IoU analysis concerning also CAM, GradCAM and GradCAM++, the best IoU values are found in the interval $t \in \{0.05, 0.15\}$. This phenomenon can be observed also in the Smooth Grad-CAM++ case. Figure A.2 focuses on the results obtained in this threshold range and shows that the best result is

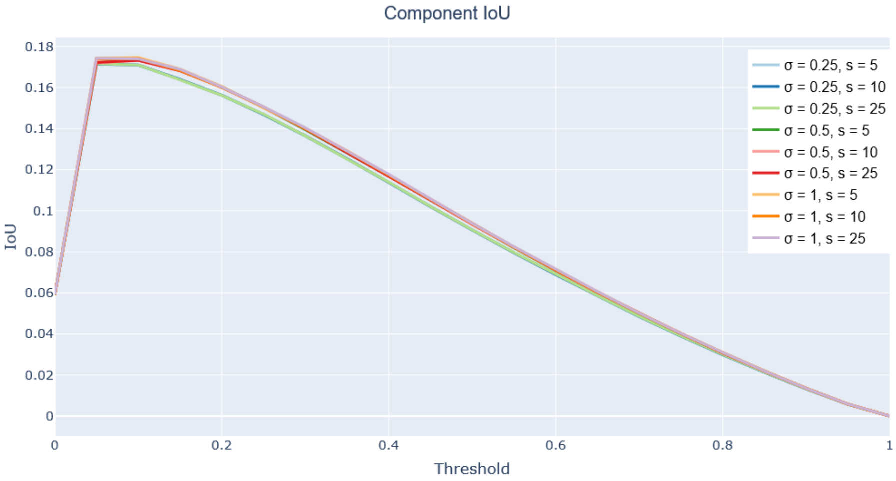


Figure A.1: Component IoU at varying threshold levels.

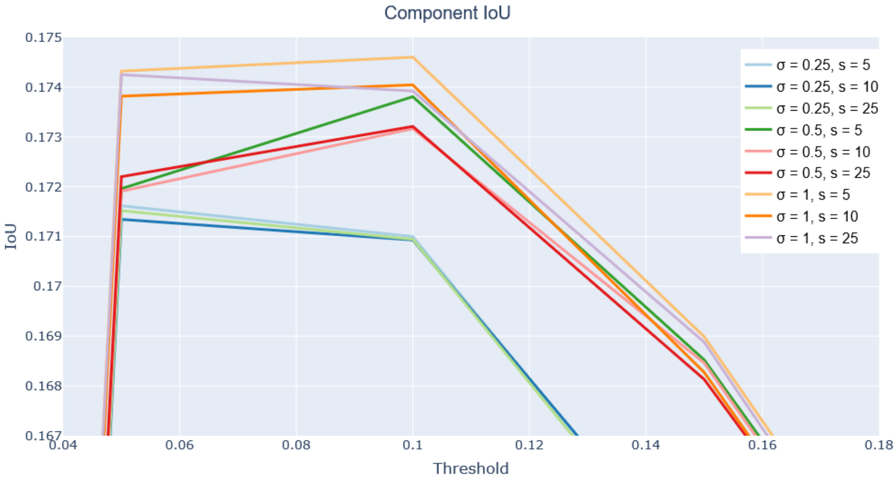


Figure A.2: Component IoU at varying threshold levels, with a focus on $t \in \{0.05, 0.15\}$.

obtained for $\sigma = 1, s = 5$, while the worst results are quantitatively close. In particular, the set of results with $\sigma = 0.25$ achieves the lowest IoU values. Since there are no significant differences for different values of s, it is preferable to choose the lowest number of samples, i.e. $s = 5$, to reduce the computational cost.

It is interesting to observe that, also for the other values of σ, a variation of the number of samples does not produce a significant change, while a variation of σ produces more substantial IoU variations. Figures A.3, A.4 and A.5 clearly show this phenomenon on the

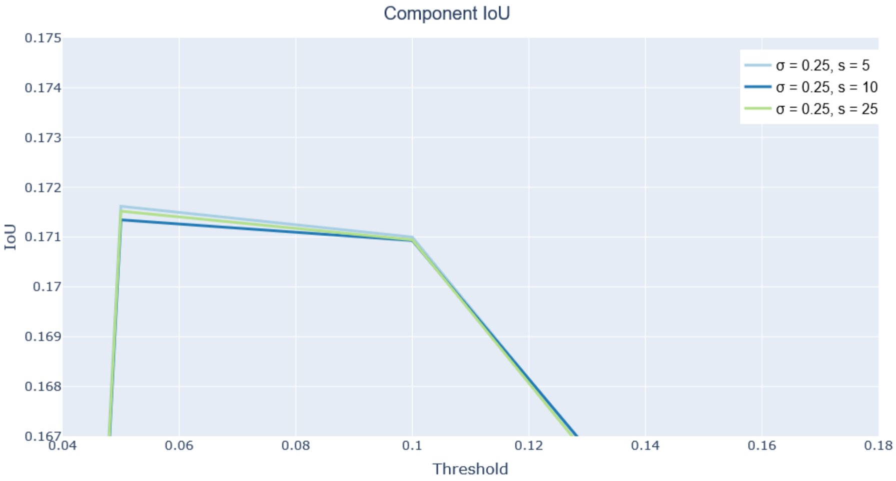


Figure A.3: Component IoU at varying threshold levels in the proximity of the peak for $\sigma = 0.25$.

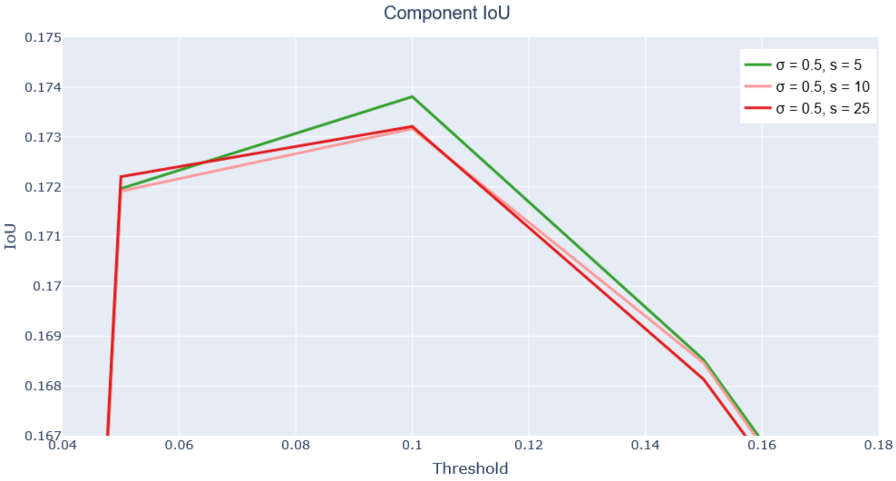


Figure A.4: Component IoU at varying threshold levels in the proximity of the peak for $\sigma = 0.5$.

component IoU peaks for $\sigma = 0.25$, $\sigma = 0.5$, and $\sigma = 1$ respectively. Since the purpose of the graphs is to highlight the differences observed for different values of σ, the scale of the y axis is much bigger than the one presented, for example, in Figure A.2. However, the maximum distance between peaks obtained for the same value of σ is less than 0.001. In each graph, the darker items in the legend are shown.

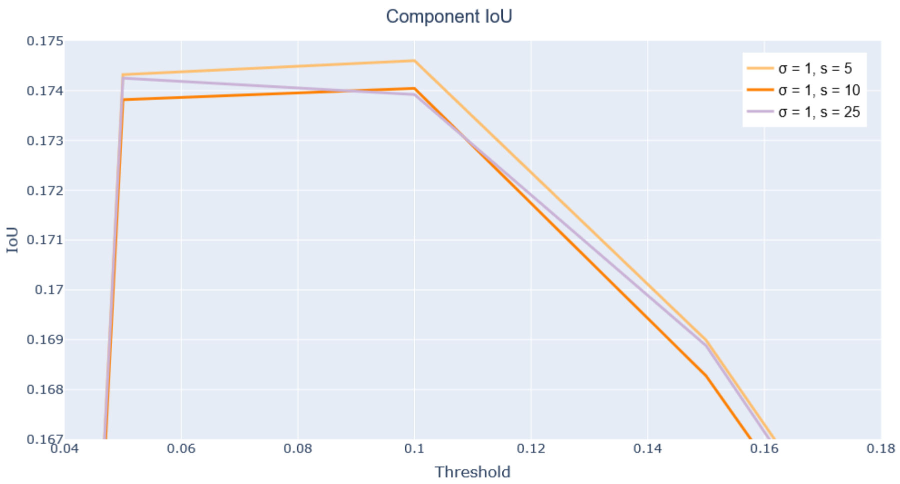


Figure A.5: Component IoU at varying threshold levels in the proximity of the peak for $\sigma = 1$.

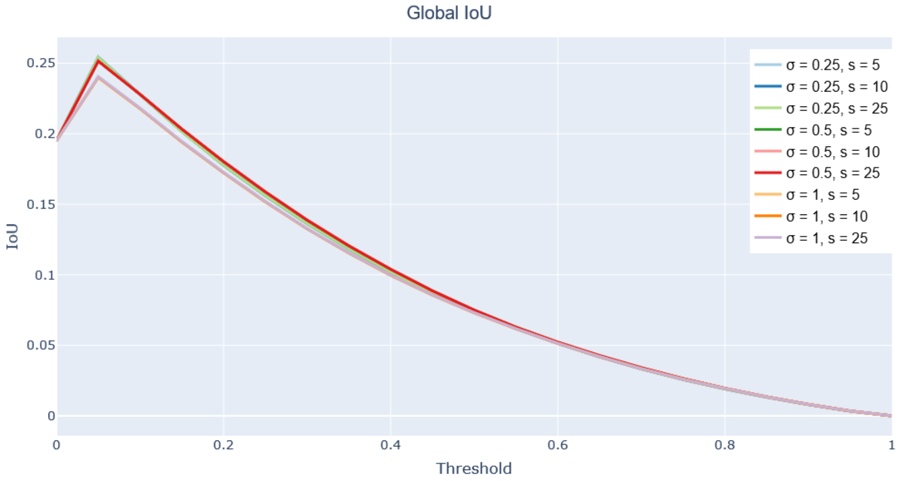


Figure A.6: Global IoU at varying threshold levels.

A.2 GLOBAL IOU

This section presents the global IoU results for different values of the Smooth Grad-CAM++ hyper-parameters, as a deepening of the results presented in Section 4.2.

Figure A.6 presents an overview of the global IoU results for different thresholds. Its purpose is to emphasize the small differences introduced by changing the number of samples s and the standard deviation σ. As observed in the global IoU analysis concerning also

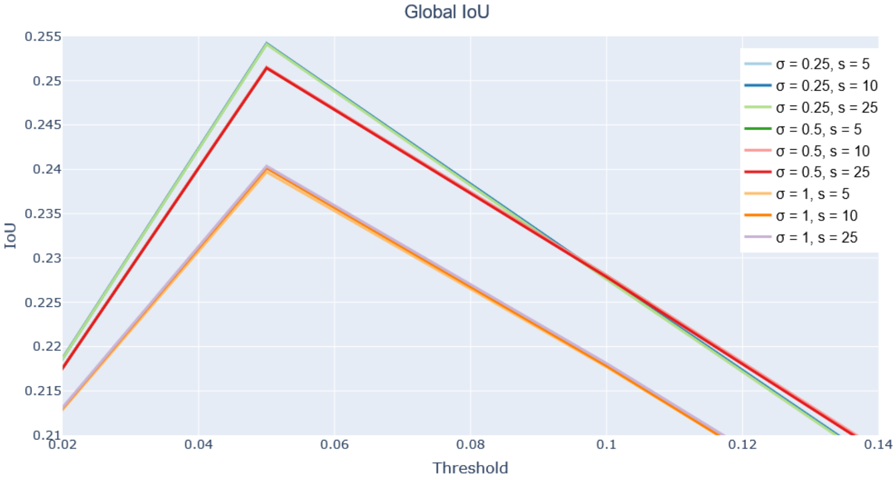


Figure A.7: Global IoU at varying threshold levels, with a focus around $t = 0.05\}$.

CAM, GradCAM and GradCAM++, the best IoU values are found in the for $t = 0.05$. This phenomenon can be observed also in the Smooth Grad-CAM++ case. Figure A.7 focuses on the results obtained for this threshold value and shows that the best result is obtained for $\sigma = 0.25, s = 5$, while the worst results are quantitatively close. In particular, the set of results with $\sigma = 1$ achieves the lowest IoU values. Since there are no significant differences for different values of s, it is preferable to choose the lowest number of samples, i.e. $s = 5$, to reduce the computational cost.

It is interesting to observe that, also for the other values of σ, a variation of the number of samples does not produce a significant change, while a variation of σ produces more substantial IoU variations. Figures A.8, A.9 and A.10 clearly show this phenomenon on the global IoU peaks for $\sigma = 0.25$, $\sigma = 0.5$, and $\sigma = 1$ respectively. Since the purpose of the graphs is to highlight the differences observed for different values of σ, the scale of the y axis is much bigger than the one presented, for example, in Figure A.7. However, the maximum

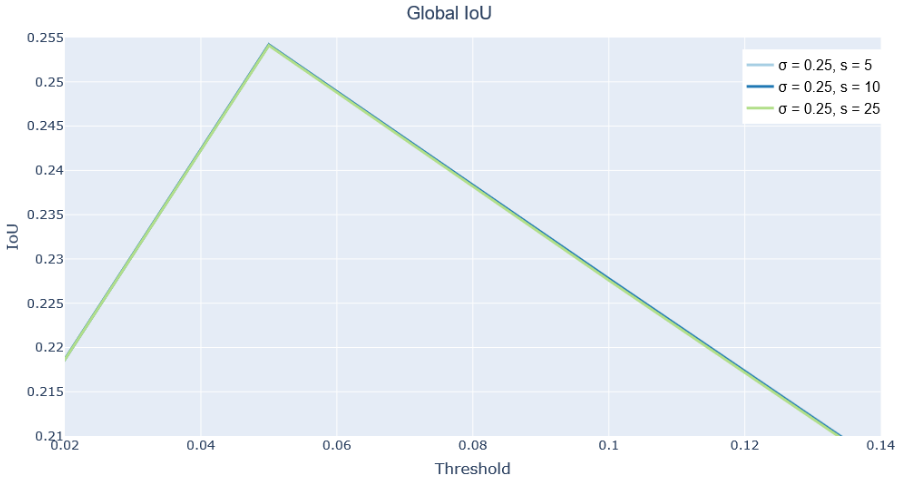


Figure A.8: Global IoU at varying threshold levels in the proximity of the peak for $\sigma = 0.25$.

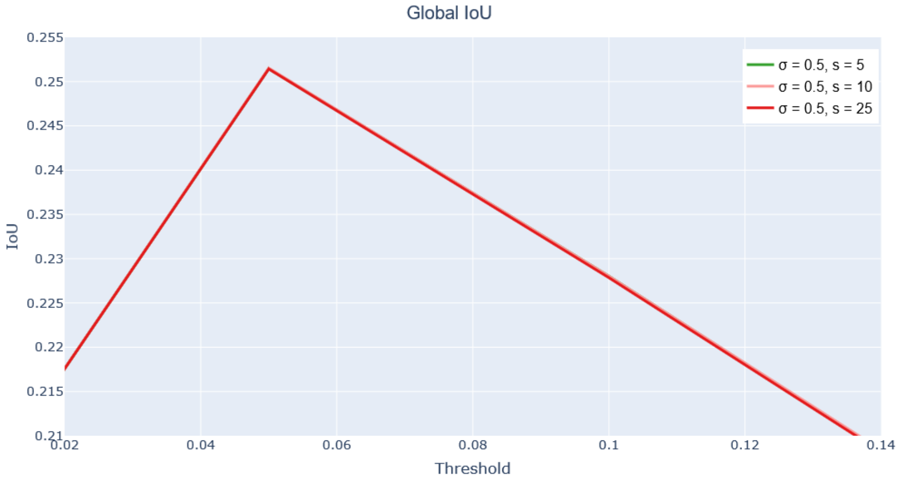


Figure A.9: Global IoU at varying threshold levels in the proximity of the peak for $\sigma = 0.5$.

distance between peaks obtained for the same value of σ is less than 0.005. In each graph, the darker items in the legend are shown.

Compared to the component IoU analysis, the best and the worst case are inverted. This phenomenon is observed because the role of component IoU and global IoU is complementary: while component IoU focuses on individual symbols, global IoU focuses on the whole representation of the saint. Since the main purpose of this research is finding relevant iconographical symbols, the best results for the component IoU are considered.

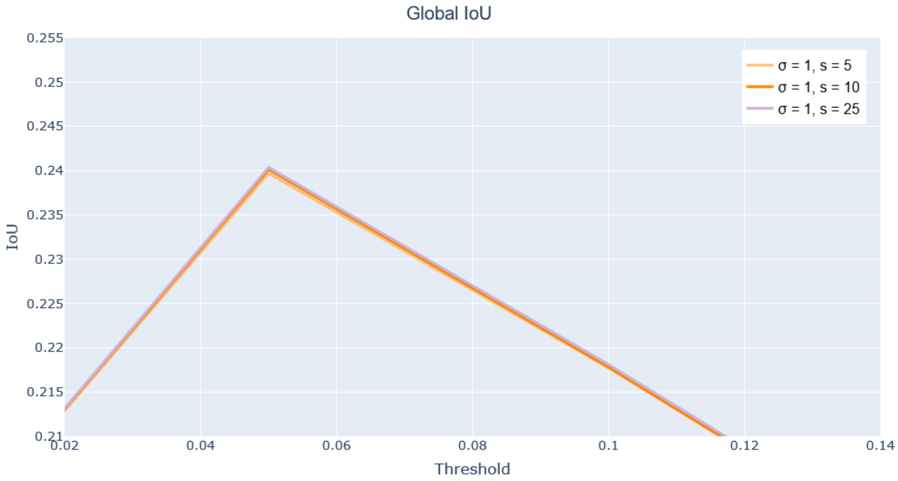


Figure A.10: Global IoU at varying threshold levels in the proximity of the peak for $\sigma = 1$.

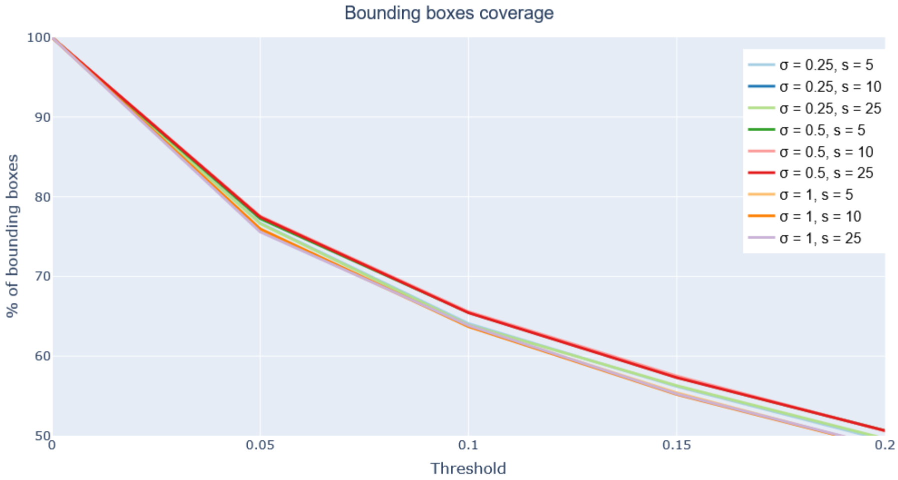


Figure A.11: Bounding box coverage at varying threshold levels.

A.3 BOUNDING BOX COVERAGE

This section presents the bounding box coverage results for different values of the Smooth Grad-CAM++ hyper-parameters, as a deepening of the results presented in Section 4.2.

Figure A.11 presents an overview of the bounding box coverage results for different thresholds. Its purpose is to emphasize the small differences introduced by changing the number of samples s and the standard deviation σ. We consider a threshold $t = 0.1$, since

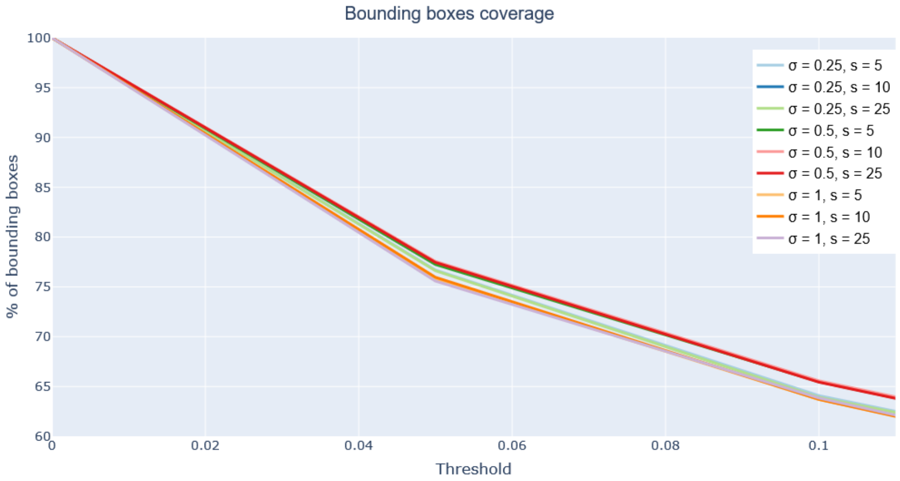


Figure A.12: Bounding box coverage at varying threshold levels, with a focus around $t = 0.1\}$.

this is the peak value obtained in the component IoU analysis. It is not meaningful, instead, to consider $t = 0$, since this would be a trivial case with the class activation map spreading on the entire image. Figure A.12 focuses on the results obtained for this threshold value and shows that the best result is obtained for $\sigma = 0.5$ (without relevant variations when the number of samples changes). The set of results with $\sigma = 1$, instead, achieves the lowest IoU values. Since there are no significant differences for different values of s, it is preferable to choose the lowest number of samples, i.e. $s = 5$, to reduce the computational cost.

It is interesting to observe that, also for the other values of σ, a variation of the number of samples does not produce a significant change, while a variation of σ produces more substantial IoU variations. Figures A.13, A.14 and A.15 clearly show this phenomenon on the bounding box coverage peaks for $\sigma = 0.25$, $\sigma = 0.5$, and $\sigma = 1$ respectively. Since the purpose of the graphs is to highlight the differences observed for different values of σ, the scale of the y axis is much bigger than the one presented, for example, in Figure A.12. However,

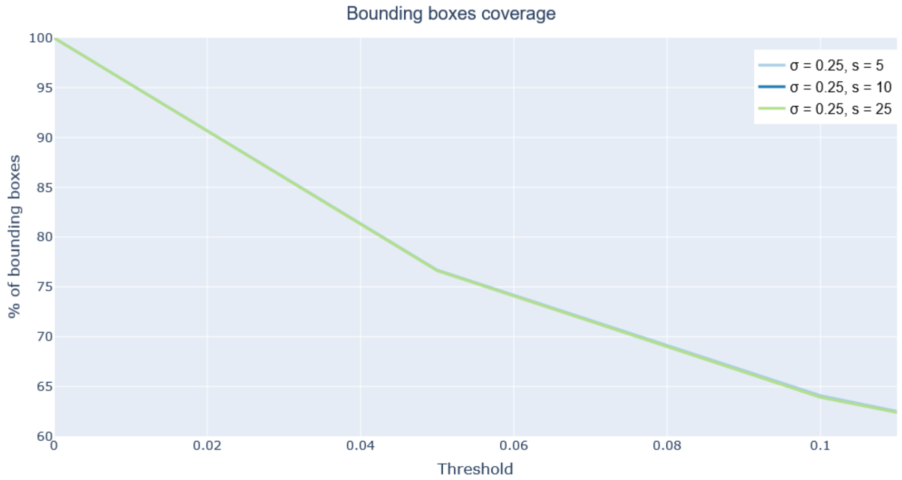


Figure A.13: Bounding box coverage at varying threshold levels in the proximity of the peak for $\sigma = 0.25$.

the maximum distance between peaks obtained for the same value of σ is less than 1%. In each graph, the darker items in the legend are shown.

Compared to the IoU analyses, the best and the worst case are different. The reason is that in this case, we are considering the *intersection*, rather than the *Intersection over Union*. Likely, intermediate values for σ (i.e. $\sigma = 0.5$) correspond to a greater probability of covering at least a part of each symbol, while low values correspond to more focused areas and high values to less focused areas. In the case of low σ values, the class activation maps values may be higher in the focused areas, but much lower in the surrounding areas. In the case of high σ, instead, the class activation maps may be lower than 0.1 in a greater part of such broader areas.

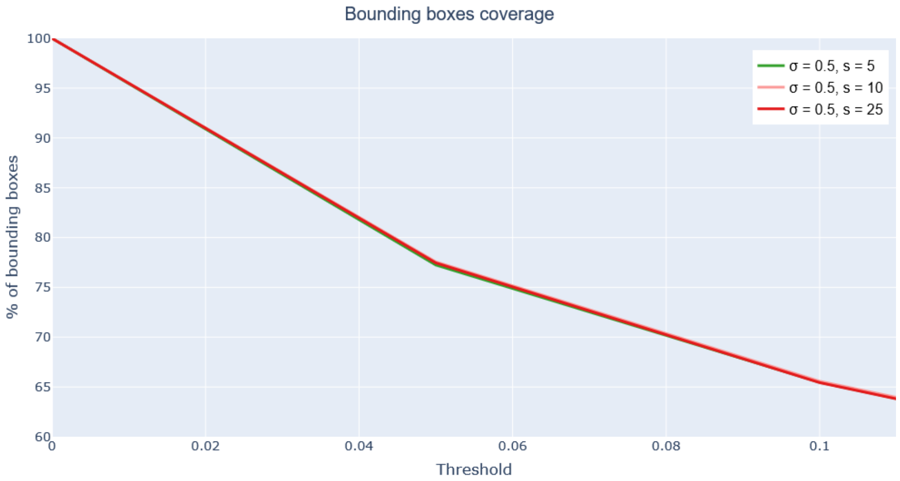


Figure A.14: Bounding box coverage at varying threshold levels in the proximity of the peak for $\sigma = 0.5$.

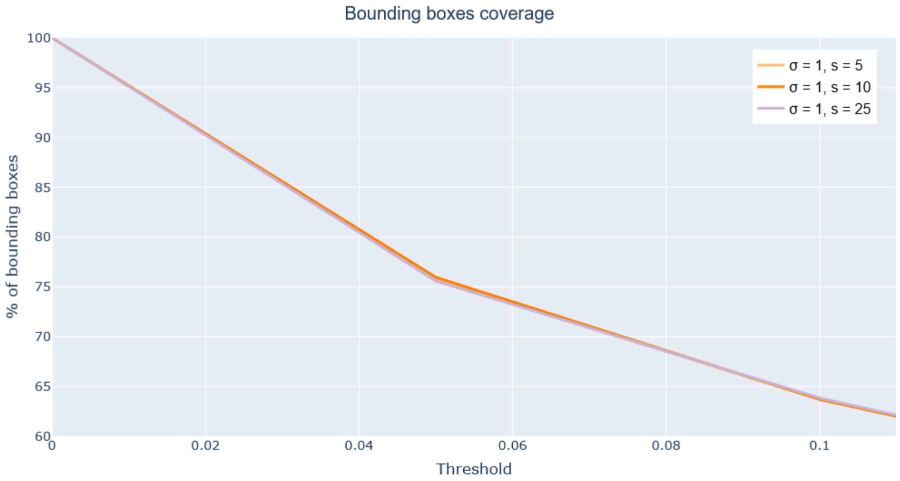


Figure A.15: Bounding box coverage at varying threshold levels in the proximity of the peak for $\sigma = 1$.

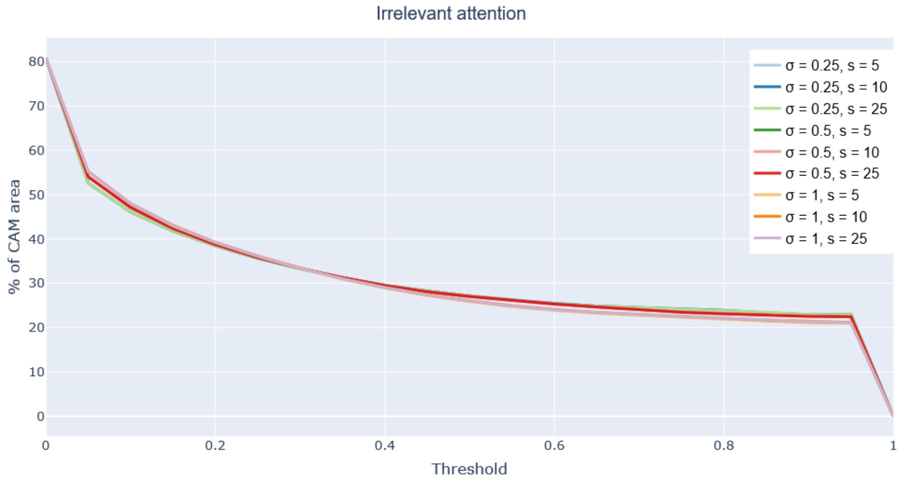


Figure A.16: Bounding box coverage at varying threshold levels.

A.4 IRRELEVANT ATTENTION

This section presents the irrelevant attention results for different values of the Smooth Grad-CAM++ hyper-parameters, as a deepening of the results presented in Section 4.2.

Figure A.16 presents an overview of the irrelevant attention results for different thresholds. Its purpose is to emphasize the small differences introduced by changing the number of samples s and the standard deviation σ. We consider a threshold $t = 0.05$, since the this analysis considers the area not included in *any* class activation map. It is not meaningful, instead, to consider $t = 0$, since this would be a trivial case with the class activation map spreading on the entire image (i.e., the irrelevant attention would not depend on the quality of the algorithm). Figure A.17 focuses on the results obtained for this threshold value and shows that the best result is obtained for $\sigma = 1$ (without relevant variations when the number of samples changes). The set of results with $\sigma = 0.25$, instead, achieves the lowest IoU values. Since there are no significant differences for different values of

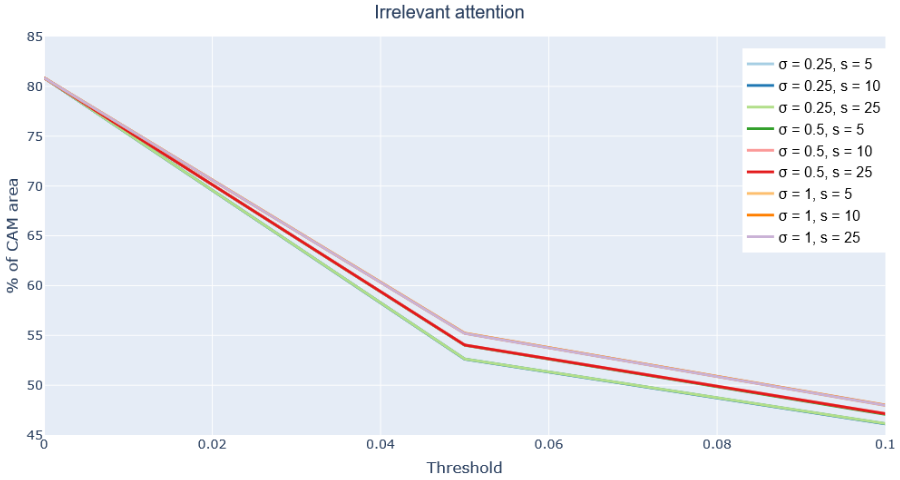


Figure A.17: Irrelevant attention at varying threshold levels, with a focus around $t = 0.05\}$.

s, it is preferable to choose the lowest number of samples, i.e. $s = 5$, to reduce the computational cost.

It is interesting to observe that, also for the other values of σ, a variation of the number of samples does not produce a significant change, while a variation of σ produces more substantial IoU variations. Figures A.18, A.19 and A.20 clearly show this phenomenon on the irrelevant attention peaks for $\sigma = 0.25$, $\sigma = 0.5$, and $\sigma = 1$ respectively. Since the purpose of the graphs is to highlight the differences observed for different values of σ, the scale of the y axis is much bigger than the one presented, for example, in Figure A.17. However, the maximum distance between peaks obtained for the same value of σ is less than 0.1%. In each graph, the darker items in the legend are shown.

Compared to the IoU analyses, the best and the worst cases are the same as the component IoU. This means that considering the Smooth Grad-CAM++ hyper-parameter combinations, $\sigma = 1$ produces more focused areas in correspondence of the symbols and less irrelevant areas outside of the symbols.

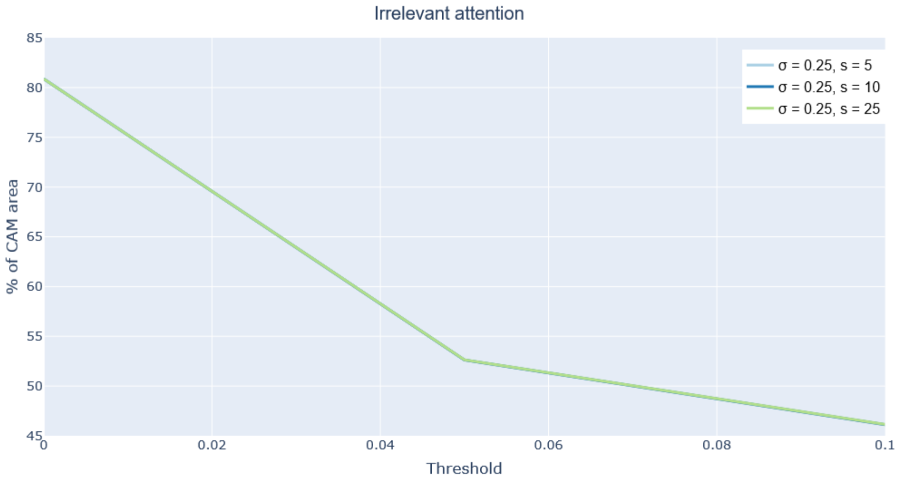


Figure A.18: Irrelevant attention at varying threshold levels in the proximity of the peak for $\sigma = 0.25$.

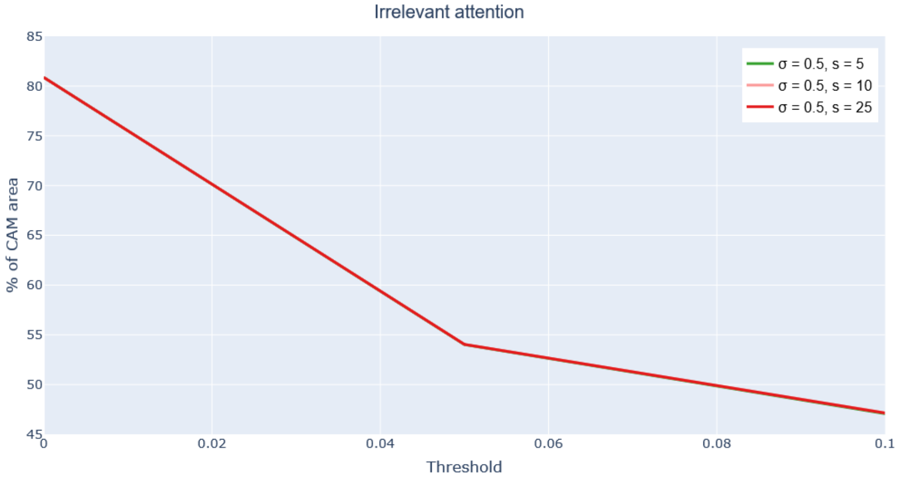


Figure A.19: Irrelevant attention at varying threshold levels in the proximity of the peak for $\sigma = 0.5$.

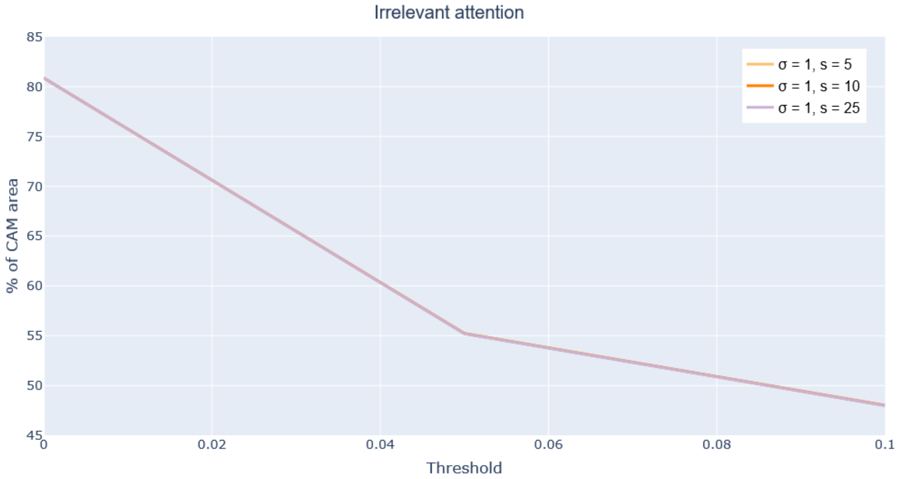


Figure A.20: Irrelevant attention at varying threshold levels in the proximity of the peak for $\sigma = 1$.

A.5 CONCLUSIONS ON SMOOTH GRAD-CAM++

The analyses presented in this appendix show three important characteristics of Smooth Grad-CAM++, when applied to the ArtDL dataset:

- A variation of the number of samples s, without a variation of the standard deviation σ does not produce significant differences, hence it is preferable to use $s = 5$ to keep the computational costs lower;
- A standard deviation $\sigma = 1$ gives more focused areas and less irrelevant areas;
- Changing σ produces variations that are more significant than the ones obtained by changing s, but they are still negligible with respect to the variations needed to overcome the results obtained using Grad-CAM.

BIBLIOGRAPHY

[1] Cristina Acidini Luchinat. “Da Michelangelo a Raffaello, due modelli opposti per il’Giudizio’nella cupola.” In: (1997) (cit. on p. 56).

[2] Mohamad Ali-Dib, Kristen Menou, Alan P Jackson, Chenchong Zhu, and Noah Hammond. “Automated crater shape retrieval using weakly-supervised deep learning.” In: *Icarus* 345 (2020), p. 113749 (cit. on pp. 12, 13).

[3] Wonho Bae, Junhyug Noh, and Gunhee Kim. “Rethinking class activation mapping for weakly supervised object localization.” In: *European Conference on Computer Vision*. Springer. 2020, pp. 618–634 (cit. on p. 116).

[4] Wonho Bae, Junhyug Noh, and Gunhee Kim. “Rethinking Class Activation Mapping for Weakly Supervised Object Localization.” In: *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XV*. Ed. by Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm. Vol. 12360. Lecture Notes in Computer Science. Springer, 2020, pp. 618–634 (cit. on pp. 88, 113, 114).

[5] Nikolay Banar, Walter Daelemans, and Mike Kestemont. “Multimodal Label Retrieval for the Visual Arts: The Case of Iconclass.” In: (2021) (cit. on p. 25).

[6] Bernadine Barnes. “Metaphorical Painting: Michelangelo, Dante, and the Last Judgment.” In: *The Art Bulletin* 77.1 (1995), pp. 65–81 (cit. on p. 56).

[7] Philip Bateman and Hans Georg Schaathun. “Image steganography and steganalysis.” In: *Department Of Computing, Faculty of Engineering and Physical Sciences, University of Surrey, Guildford, Surrey, United Kingdom, 4th August* (2008) (cit. on p. 19).

[8] Abdelhak Belhi, Hosameldin Osman Ahmed, Taha Alfaqheri, Abdelaziz Bouras, Abdul Hamid Sadka, and Sebti Foufou. “Study and Evaluation of Pre-trained CNN Networks for Cultural Heritage Image Classification.” In: *Data Analytics for Cultural Heritage: Current Trends and Concepts* (2021), p. 47 (cit. on p. 25).

[9] Simone Bianco, Davide Mazzini, Paolo Napoletano, and Raimondo Schettini. “Multitask painting categorization by deep multibranch neural network.” In: *Expert Systems with Applications* 135 (2019), pp. 90–101 (cit. on pp. 25, 28, 29).

[10] Mehdi Boroumand, Mo Chen, and Jessica Fridrich. “Deep Residual Network for Steganalysis of Digital Images.” In: *IEEE Transactions on Information Forensics and Security* 14.5 (2019), pp. 1181–1193 (cit. on p. 19).

[11] Vanessa Buhrmester, David Münch, and Michael Arens. “Analysis of explainers of black box deep neural networks for computer vision: A survey.” In: *arXiv preprint arXiv:1911.12116* (2019) (cit. on p. 23).

[12] Charles Burroughs. “The" Last Judgment" of Michelangelo: Pictorial Space, Sacred Topography, and the Social World.” In: *Artibus et historiae* (1995), pp. 55–89 (cit. on p. 56).

[13] Hongping Cai, Qi Wu, Tadeo Corradi, and Peter Hall. *The Cross-Depiction Problem: Computer Vision Algorithms for Recognising Objects in Artwork and in Photographs*. 2015. arXiv: 1505.00110 [cs.CV] (cit. on pp. 2, 25).

[14] Giovanna Castellano and Gennaro Vessio. "Deep learning approaches to pattern extraction and recognition in paintings and drawings: an overview." In: *Neural Computing and Applications* (2021), pp. 1–20 (cit. on p. 25).

[15] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. "Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks." In: *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)* (Mar. 2018) (cit. on pp. 4, 24, 39, 44).

[16] Junsuk Choe and Hyunjung Shim. "Attention-Based Dropout Layer for Weakly Supervised Object Localization." In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 2219–2228 (cit. on pp. 88, 113, 114).

[17] Junsuk Choe and Hyunjung Shim. "Attention-based dropout layer for weakly supervised object localization." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 2219–2228 (cit. on p. 117).

[18] Francois Chollet et al. *Deep learning with Python*. Vol. 361. Manning, New York, 2018 (cit. on pp. 10, 15).

[19] Ceren Cömert, Murat Özbayoğlu, and Coşku Kasnakoğlu. "Painter Prediction from Artworks with Transfer Learning." In: *2021 7th*

International Conference on Mechatronics and Robotics Engineering (ICMRE). IEEE. 2021, pp. 204–208 (cit. on p. 25).

[20] Leendert D Couprie. "Iconclass: an iconographic classification system." In: *Art Libraries Journal* 8.2 (1983), pp. 32–49 (cit. on p. 51).

[21] Elliot J Crowley and Andrew Zisserman. "Of gods and goats: Weakly supervised learning of figurative art." In: *learning* 8 (2013), p. 14 (cit. on pp. 25, 32).

[22] Elliot J Crowley and Andrew Zisserman. "The state of the art: Object retrieval in paintings using discriminative regions." In: (2014) (cit. on pp. 25, 31).

[23] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. "ImageNet: A large-scale hierarchical image database." In: *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*. IEEE Computer Society, 2009, pp. 248–255 (cit. on p. 4).

[24] Jiahua Dong, Yang Cong, Gan Sun, and Dongdong Hou. "Semantic-transferable weakly-supervised endoscopic lesions segmentation." In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 10712–10721 (cit. on p. 13).

[25] 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*. Vol. 32. 1. 2018 (cit. on p. 25).

[26] Zhi Gao, Mo Shan, and Qingquan Li. "Adaptive sparse representation for analyzing artistic style of paintings." In: *Journal*

on Computing and Cultural Heritage (JOCCH) 8.4 (2015), pp. 1–15 (cit. on p. 25).

[27] Weifeng Ge, Xiangru Lin, and Yizhou Yu. "Weakly supervised complementary parts models for fine-grained image classification from the bottom up." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 3034–3043 (cit. on pp. 114, 115).

[28] N. Gonthier, Y. Gousseau, S. Ladjal, and O. Bonfait. "Weakly Supervised Object Detection in Artworks." In: (2018), pp. 692–709 (cit. on pp. 2, 25, 29, 32, 34, 90).

[29] Nicolas Gonthier, Yann Gousseau, and Saïd Ladjal. "An analysis of the transfer learning of convolutional neural networks for artistic images." In: *arXiv preprint arXiv:2011.02727* (2020) (cit. on p. 25).

[30] Nicolas Gonthier, Saïd Ladjal, and Yann Gousseau. *Multiple instance learning on deep features for weakly supervised object detection with extreme domain shifts*. 2020. arXiv: 2008.01178 [cs.CV] (cit. on p. 90).

[31] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. "A survey of methods for explaining black box models." In: *ACM computing surveys (CSUR)* 51.5 (2018), pp. 1–42 (cit. on pp. 21, 23).

[32] V. Gupta, M. Demirer, M. Bigelow, S. M. Yu, J. S. Yu, L. M. Prevedello, R. D. White, and B. S. Erdal. "Using Transfer Learning and Class Activation Maps Supporting Detection and Localization of Femoral Fractures on Anteroposterior Radiographs."

In: *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. 2020, pp. 1526–1529 (cit. on p. 24).

[33] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. *Deep Residual Learning for Image Recognition*. 2015. arXiv: `1512.03385 [cs.CV]` (cit. on pp. 4, 18, 120).

[34] David Kadish, Sebastian Risi, and Anders Sundnes Løvlie. "Improving Object Detection in Art Images Using Only Style Transfer." In: *arXiv preprint arXiv:2102.06529* (2021) (cit. on p. 25).

[35] Fahdi Kanavati, Gouji Toyokawa, Seiya Momosaki, Michael Rambeau, Yuka Kozuma, Fumihiro Shoji, Koji Yamazaki, Sadanori Takeo, Osamu Iizuka, and Masayuki Tsuneki. "Weakly-supervised learning for lung carcinoma classification using deep learning." In: *Scientific reports* 10.1 (2020), pp. 1–11 (cit. on p. 12).

[36] Sergey Karayev, Matthew Trentacoste, Helen Han, Aseem Agarwala, Trevor Darrell, Aaron Hertzmann, and Holger Winnemoeller. "Recognizing image style." In: *arXiv preprint arXiv:1311.3715* (2013) (cit. on pp. 25, 28).

[37] Asifullah Khan, Anabia Sohail, Umme Zahoora, and Aqsa Saeed Qureshi. "A survey of the recent architectures of deep convolutional neural networks." In: *Artificial Intelligence Review* 53.8 (2020), pp. 5455–5516 (cit. on p. 18).

[38] Fahad Shahbaz Khan, Shida Beigpour, Joost Van de Weijer, and Michael Felsberg. "Painting-91: a large scale database for computational painting categorization." In: *Machine vision and applications* 25.6 (2014), pp. 1385–1397 (cit. on pp. 25, 29).

[39] John N King. *Tudor royal iconography: literature and art in an age of religious crisis*. Princeton University Press, Princeton, 1989 (cit. on p. 1).

[40] Dimitri Korsch, Paul Bodesheim, and Joachim Denzler. "Classification-specific parts for improving fine-grained visual categorization." In: *German Conference on Pattern Recognition*. Springer. 2019, pp. 62–75 (cit. on pp. 114, 115).

[41] Alex Krizhevsky, Geoffrey Hinton, et al. *Learning multiple layers of features from tiny images*. 2009 (cit. on p. 18).

[42] Fernando Lanzi and Gioia Lanzi. *Saints and their symbols: recognizing saints in art and in popular images*. Liturgical Press, Collegeville, 2004, pp. 327–342 (cit. on p. 49).

[43] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. "Enhanced deep residual networks for single image super-resolution." In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2017, pp. 136–144 (cit. on p. 19).

[44] Min Lin, Qiang Chen, and Shuicheng Yan. "Network in network." In: *arXiv preprint arXiv:1312.4400* (2013) (cit. on p. 39).

[45] Zhang Lu, Zhang Yu, Peng Yali, Liu Shigang, Wu Xiaojun, Lu Gang, and Rao Yuan. "Fast single image super-resolution via dilated residual networks." In: *IEEE Access* 7 (2018), pp. 109729–109738 (cit. on p. 19).

[46] Hui Mao, Ming Cheung, and James She. "Deepart: Learning joint representations of visual arts." In: *Proceedings of the 25th ACM international conference on Multimedia*. 2017, pp. 1183–1191 (cit. on pp. 25, 28, 29).

[47] Federico Milani and Piero Fraternali. *A Data Set and a Convolutional Model for Iconography Classification in Paintings*. 2020. arXiv: 2010.11697 [cs.CV] (cit. on pp. 2, 4, 25–27, 29, 30, 34, 35, 49, 51, 66, 119, 120).

[48] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. "DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks." In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 2574–2582 (cit. on pp. 23, 24).

[49] Pietro Morbidelli, Diego Carrera, Beatrice Rossi, Pasqualina Fragneto, and Giacomo Boracchi. "Augmented Grad-CAM: Heat-Maps Super Resolution Through Augmentation." In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2020, pp. 4067–4071 (cit. on p. 19).

[50] Huu-Giao Nguyen, Alessia Pica, Jan Hrbacek, Damien C Weber, Francesco La Rosa, Ann Schalenbourg, Raphael Sznitman, and Meritxell Bach Cuadra. "A novel segmentation framework for uveal melanoma in magnetic resonance imaging based on class activation maps." In: *International Conference on Medical Imaging with Deep Learning*. PMLR. 2019, pp. 370–379 (cit. on p. 24).

[51] Mark Nixon and Alberto Aguado. *Feature extraction and image processing for computer vision*. Academic press, Orlando, 2019, p. 1 (cit. on p. 19).

[52] Daniel Omeiza, Skyler Speakman, Celia Cintas, and Komminist Weldermariam. "Smooth grad-cam++: An enhanced inference level visualization technique for deep convolutional neural net-

work models." In: *arXiv preprint arXiv:1908.01224* (2019) (cit. on pp. 4, 5, 24, 39, 46, 112).

[53] Erwin Panofsky. *Studies in Iconology: Humanistic Themes in the Art of the Renaissance*. Ed. by New York Oxford University Press. 1939, p. 262 (cit. on p. 1).

[54] Maria G Parani. *Reconstructing the reality of images: Byzantine material culture and religious iconography 11th-15th centuries*. Vol. 41. Brill, Leiden, 2003 (cit. on p. 1).

[55] B. Patro, M. Lunayach, S. Patel, and V. Namboodiri. "U-CAM: Visual Explanation Using Uncertainty Based Class Activation Maps." In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019, pp. 7443–7452 (cit. on p. 3).

[56] Nicolò Oreste Pinciroli Vago, Ibrahim A. Hameed, and Michael Kachelriess. "Using Convolutional Neural Networks for the Helicity Classification of Magnetic Fields." In: *Proceedings of Science* Proceedings of the 37th International Cosmic Ray Conference -ICRC 2021- (2021) (cit. on p. 15).

[57] Nicolò Oreste Pinciroli Vago, Federico Milani, Piero Fraternali, and Ricardo da Silva Torres. "Comparing CAM Algorithms for the Identification of Salient Image Features in Iconography Artwork Analysis." In: *Journal of Imaging* 7.7 (2021). ISSN: 2313-433X (cit. on pp. 6, 49).

[58] Marcelo OR Prates, Pedro H Avelar, and Luis C Lamb. "Assessing gender bias in machine translation: a case study with google translate." In: *Neural Computing and Applications* (2019), pp. 1–19 (cit. on pp. 21–23).

[59] Donald A Proulx. *A sourcebook of Nasca ceramic iconography: Reading a culture through its art*. University of Iowa Press, Iowa City, 2009 (cit. on p. 1).

[60] Suo Qiu. “Global Weighted Average Pooling Bridges Pixel-level Localization and Image-level Classification.” In: *CoRR* abs/1809.08264 (2018). arXiv: 1809.08264 (cit. on p. 40).

[61] Sheng Ren, Deepak Kumar Jain, Kehua Guo, Tao Xu, and Tao Chi. “Towards efficient medical lesion image super-resolution based on deep residual networks.” In: *Signal Processing: Image Communication* 75 (2019), pp. 1–10 (cit. on p. 19).

[62] Helene E Roberts. *Encyclopedia of comparative iconography: themes depicted in works of art*. Routledge, London, 2013 (cit. on p. 1).

[63] José María Salvador González. “The Iconography of the Coronation of the Virgin in Late Medieval Italian Painting.” In: *Eikón / Imago* 2.1 (June 2013), pp. 1 –48 (cit. on p. 102).

[64] Iria Santos, Luz Castro, Nereida Rodriguez-Fernandez, Alvaro Torrente-Patino, and Adrian Carballal. “Artificial Neural Networks and Deep Learning in the Visual Arts: A review.” In: *Neural Computing and Applications* (2021), pp. 1–37 (cit. on p. 25).

[65] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization.” In: *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 618–626 (cit. on pp. 4, 24, 39, 42).

[66] Lior Shamir and Jane A. Tarakhovsky. “Computer Analysis of Art.” In: *Journal on Computing and Cultural Heritage* 5.2 (Aug. 2012). ISSN: 1556-4673 (cit. on p. 2).

[67] Xi Shen, Alexei A. Efros, and Mathieu Aubry. *Discovering Visual Patterns in Art Collections with Spatially-consistent Feature Learning*. 2019. arXiv: 1903.02678 [cs.CV] (cit. on p. 25).

[68] Krishna Kumar Singh and Yong Jae Lee. "Hide-and-Seek: Forcing a Network to be Meticulous for Weakly-supervised Object and Action Localization." In: *CoRR* abs/1704.04232 (2017). arXiv: 1704.04232 (cit. on p. 88).

[69] Gjorgji Strezoski and Marcel Worring. "Omniart: multi-task deep learning for artistic data analysis." In: *arXiv preprint arXiv:1708.00684* (2017) (cit. on pp. 25, 30).

[70] K. H. Sun, H. Huh, B. A. Tama, S. Y. Lee, J. H. Jung, and S. Lee. "Vision-Based Fault Diagnostics Using Explainable Deep Learning With Class Activation Maps." In: *IEEE Access* 8 (2020), pp. 129169–129179 (cit. on pp. 3, 24).

[71] Sudeepti Surapaneni, Sana Syed, and Logan Yoonhyuk Lee. "Exploring Themes and Bias in Art using Machine Learning Image Analysis." In: *2020 Systems and Information Engineering Design Symposium (SIEDS)*. IEEE. 2020, pp. 1–6 (cit. on pp. 24, 36, 37).

[72] Ying Tai, Jian Yang, and Xiaoming Liu. "Image super-resolution via deep recursive residual network." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 3147–3155 (cit. on p. 19).

[73] Rocio Nahime Torres, Piero Fraternali, and Jesus Romero. "ODIN: An Object Detection and Instance Segmentation Diagnosis Framework." In: *Computer Vision – ECCV 2020 Workshops*. Ed. by Adrien

Bartoli and Andrea Fusiello. Cham: Springer International Publishing, 2020, pp. 19–31. ISBN: 978-3-030-65414-6 (cit. on p. 52).

[74] Nicolò Oreste Pinciroli Vago, Ibrahim A. Hameed, and Michael Kachelriess. *Using Convolutional Neural Networks for the Helicity Classification of Magnetic Fields*. 2021. arXiv: 2106.06718 [astro-ph.HE] (cit. on p. 15).

[75] Theo Van Leeuwen and Carey Jewitt. *The handbook of visual analysis*. Sage, Thousand Oaks, 2001, pp. 100–102 (cit. on p. 1).

[76] Haofan Wang, Mengnan Du, Fan Yang, and Zijian Zhang. "Score-cam: Improved visual explanations via score-weighted class activation mapping." In: *arXiv preprint arXiv:1910.01279* (2019) (cit. on p. 113).

[77] Yu Wang, Fengqing Zhu, Carol J. Boushey, and Edward J. Delp. "Weakly supervised food image segmentation using class activation maps." In: *2017 IEEE International Conference on Image Processing (ICIP)*. 2017, pp. 1277–1281 (cit. on p. 24).

[78] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. "Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 12275–12284 (cit. on p. 114).

[79] *Wikipedia: Saint Symbolism*. https://en.wikipedia.org/wiki/Saint_symbolism. Accessed: 2021-04-24 (cit. on p. 49).

[80] Songtao Wu, Shenghua Zhong, and Yan Liu. "Deep residual learning for image steganalysis." In: *Multimedia tools and applications* 77.9 (2018), pp. 10437–10453 (cit. on p. 19).

[81] Heekyung Yang and Kyungha Min. “Classification of basic artistic media based on a deep convolutional approach.” In: *The Visual Computer* 36.3 (2020), pp. 559–578 (cit. on pp. 24, 35).

[82] S. Yang, Y. Kim, Y. Kim, and C. Kim. “Combinational Class Activation Maps for Weakly Supervised Object Localization.” In: *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2020, pp. 2930–2938 (cit. on p. 3).

[83] Dingwen Zhang, Junwei Han, Gong Cheng, and Ming-Hsuan Yang. “Weakly Supervised Object Localization and Detection: A Survey.” In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021) (cit. on p. 85).

[84] Fan Zhang, Meng Li, Guisheng Zhai, and Yizhao Liu. “Multi-branch and multi-scale attention learning for fine-grained visual categorization.” In: *International Conference on Multimedia Modeling*. Springer. 2021, pp. 136–147 (cit. on pp. 114, 116).

[85] Man Zhang, Yong Zhou, Jiaqi Zhao, Yiyun Man, Bing Liu, and Rui Yao. “A survey of semi-and weakly supervised semantic segmentation of images.” In: *Artificial Intelligence Review* (2019), pp. 1–30 (cit. on p. 24).

[86] Guannan Zhao, Bo Zhou, Kaiwen Wang, Rui Jiang, and Min Xu. “Respond-cam: Analyzing deep models for 3d imaging data by visualizations.” In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2018, pp. 485–492 (cit. on p. 113).

[87] Wentao Zhao, Dalin Zhou, Xinguo Qiu, and Wei Jiang. “Compare the performance of the models in art classification.” In: *Plos one* 16.3 (2021), e0248414 (cit. on p. 25).

[88] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. "Learning Deep Features for Discriminative Localization." In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 2921–2929 (cit. on pp. 4, 23, 39, 40).

[89] Zhi-Hua Zhou. "A brief introduction to weakly supervised learning." In: *National science review* 5.1 (2018), pp. 44–53 (cit. on p. 11).

[90] J. Zujovic, L. Gandy, S. Friedman, B. Pardo, and T. N. Pappas. "Classifying paintings by artistic genre: An analysis of features classifiers." In: *2009 IEEE International Workshop on Multimedia Signal Processing*. 2009, pp. 1–5 (cit. on p. 2).