Journal of
Imaging

Article

Comparing CAM Algorithms for the Identification of Salient Image Features in Iconography Artwork Analysis

Nicolò Oreste Pinciroli Vago [1,2], Federico Milani [1,*], Piero Fraternali [1] and Ricardo da Silva Torres [2]

[1] Department of Electronics Information and Bioengineering, Politecnico di Milano, 20133 Milano, Italy; nicolooreste.pinciroli@mail.polimi.it (N.O.P.V.); piero.fraternali@polimi.it (P.F.)
[2] Department of ICT and Engineering, NTNU—Norwegian University of Science and Technology, 6009 Ålesund, Norway; nicoloop@stud.ntnu.no (N.O.P.V.); ricardo.torres@ntnu.no (R.d.S.T.)
* Correspondence: federico.milani@polimi.it

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 iconographic subjects in paintings and Convolutional Neural Networks 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 and if the architecture of a classifier trained on whole artwork images can be exploited to support the much harder task of object detection. 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 iconographic symbols well. Grad-CAM produces wider and more contiguous areas that cover large iconographic symbols better. The salient image areas computed by the CAM algorithms have been used to estimate object-level bounding boxes and a quantitative analysis shows that the boxes estimated with Grad-CAM reach 55% average IoU, 61% GT-known localization and 31% mAP. The obtained results are a step towards the computer-aided study of the variations of iconographic elements positioning and mutual relations in artworks and open the way to the automatic creation of bounding boxes for training detectors of iconographic symbols in Christian art images.

Keywords: convolutional neural network; class activation maps; explainability; iconography; artwork analysis

Citation: Pinciroli Vago, N.O.; Milani, F.; Fraternali, P.; da Silva Torres, R. Comparing CAM Algorithms for the Identification of Salient Image Features in Iconography Artwork Analysis. *J. Imaging* **2021**, *7*, 106. https://doi.org/10.3390/jimaging7070106

Academic Editors: Giovanna Castellano, Gennaro Vessio and Fabio Bellavia

Received: 15 May 2021
Accepted: 24 June 2021
Published: 29 June 2021

1. Introduction

Iconography is the discipline that concerns itself with the subject matter of artworks, as opposed to their form [1]. 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 [2] to the Byzantine [3] civilization. Iconography is a prominent topic of the art history studied through centuries [4–6]. The attribution of iconographic 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, iconographic 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 iconographic 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. The WikiPedia page on Christian Saint symbolism (https://en.wikipedia.org/wiki/Saint_symbolism—accessed on 15 May 2021) lists 257 characters with 791 attributes. This makes the manual attribution of iconographic 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.

A viable alternative relies on the use of semi-automatic computer-aided solutions supporting the expert annotator in the task of associating iconographic classes to art images. Computer Vision (CV) has already been used for artwork analysis tasks, such as genre identification [7], author identification [8], and even subject identification and localization [9]. The field of computer-aided iconographic analysis is more recent and addressed by few works [10,11]. Borrowing the standard CV terminology, the problem of computer-aided iconographic analysis can be further specialized into iconography classification, which tackles the association of iconographic 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 iconographic class appear.

Applying CV to the analysis of art iconographic 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 dataset in use today are very large and provided with huge numbers of annotations. Conversely, in the narrower art domain, image datasets 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 Convolutional Neural Networks (CNNs) for iconography classification [11], 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 [12–14], 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 image-wide annotations only. This result could reduce the effort of building training sets for the much harder task of iconography detection.

This paper addresses the following research questions:

- Are CAMs an effective tool for understanding how a CNN classifier recognizes the iconographic 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 paper can be summarized as follows:

- We apply four state-of-the-art class activation map algorithms (namely, CAM [15], Grad-CAM [16], Grad-CAM++ [17], and Smooth Grad-CAM++ [18]) to the CNN iconography classification model presented in [11], which exploits a backbone based on ResNet50 [19] trained on the ImageNet dataset [20] and refined on the ArtDL dataset (http://www.artdl.org—accessed on 15 May 2021) consisting of 42,479 images of artworks portraying Christian Saints divided into 10 classes. Note that, in order to avoid ambiguity, we refer to the specific algorithm as "CAM" and to the generic output as "class activation maps".

- For the quantitative evaluation of the different algorithms, a test dataset 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. We use the Intersection over Union (IoU) metrics to measure the agreement between the areas of the image highlighted by the algorithm and those annotated manually as ground truth. Furthermore, we analyze the class activation map area based on percentage of covered bounding boxes and percentage of covered area that does not contain any iconographic symbol.
- The comparison shows 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 [18] for natural images. Smooth Grad-CAM++ produces multiple disconnected image regions that identify small iconographic symbols quite precisely. Grad-CAM produces wider and more contiguous areas that cover well both large and small iconographic symbols. To the best of our knowledge, such a comparison has not been performed before in the context of artwork analysis.
- 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. A simple procedure is tested for selecting "good enough" class activation maps and for creating symbol bounding boxes automatically from them. The results of such a procedure are illustrated visually.
- We deepen the evaluation by measuring quantitatively the agreement between the ground-truth bounding boxes and the bounding boxes estimated from the class activation maps. The assessment shows that the whole Saint bounding boxes computed from the Grad-CAM class activation maps obtain 55% average IoU, 61% GT-known localization and 31% mAP. Such results obtained by a simple post-processing of the output of a general purpose CNN interpretability technique pave the way to the use of automatically computed bounding boxes for training weakly supervised object detectors in artwork images.

Figure 1 shows an example of the assessment performed in this paper. 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 the paper is organized as follows: Section 2 surveys related work; Section 3 describes the different CAM variants considered in our study; Section 4 describes the adopted evaluation protocol and the results of the quantitative and the qualitative analysis; finally, Section 5 draws the conclusions and outlines possible future work.

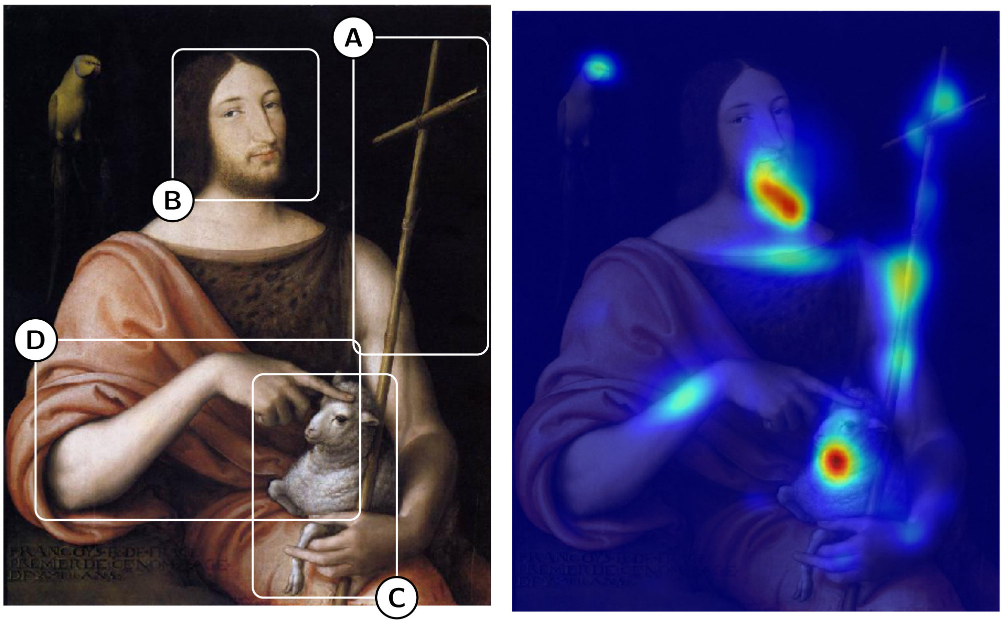

Figure 1. On the left: Saint John the Baptist image and iconographic symbols identified manually (e.g., cross (**A**), face (**B**), and lamb (**C**), and hand pointing at lamb (**D**)). On the right: the CAM heat map associated with classification results of a CNN-based solution.

2. Related Work

This section surveys the essential previous research in the fields of automated artwork analysis and CNN interpretability that are the foundations of our work.

2.1. *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 by means of Computer Vision and Deep Learning methods. Several datasets containing various types of artworks have been proposed to support such studies [10,11,21–26].

The performed analyses span several classification tasks and techniques: from style classification to artist identification, comprising also medium, school, and year classification [27–29]. 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 [30], and forgery detection [31].

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 [9–11,32–35]. Since there is 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 [11,35–38]. This approach is known as Transfer Learning and consists in 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.

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. In the literature, there are many techniques that aim at explaining the behavior of neural models [39,40]. 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 [15] have shown their effectiveness in highlighting the most discriminative areas of an image in several fields, ranging from medicine [41] to fault diagnostics [12]. The original formulation of CAMs has been subsequently improved. Selvaraju et al. [16] 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. [17] 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++ [18] applies Grad-CAM++ iteratively on the combination of the original image and a 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 [42].

This paper focuses on the comparison of different CAM algorithms on the task of iconography classification to determine which variant may be more suitable for weakly supervised studies. Since CAM algorithms are most often studied only for natural images, the aim of the work is also to address the research gap about the utility of CAMs for the art domain.

3. Class Activation Maps for Iconography Classification

This paper compares different CAM algorithms: Grad-CAM, Grad-CAM++, and Smooth Grad-CAM++. Their implementation is based on the mathematical definitions provided, respectively, by [15–18].

Figure 2 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 art works 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 [43].

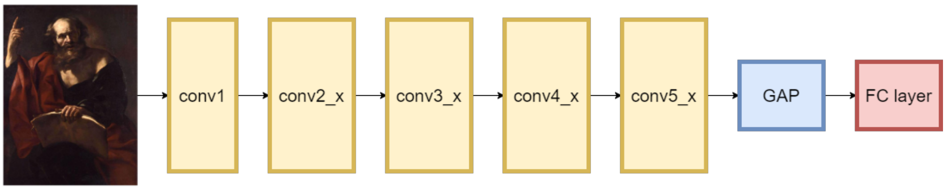

Figure 2. The ResNet50 architecture.

3.1. CAM

CAMs [15] are based on the use of GAP, which has been demonstrated to have remarkable localization abilities [44]. 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. The actual class activation map value $M_c(x, y)$ for a class c and a position x, y in the input image is expressed as follows:

$$M_c(x,y) = \sum_k w_k^c A_k(x,y) \tag{1}$$

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

Differently from the original approach, we compute the CAM output not only for the predominant class, but for all the classes. The ArtDL dataset 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 wrong classification.

3.2. *Grad-CAM*

Grad-CAM [16] is a variant of CAM which considers not only the weights but also the gradients flowing into the last convolution layer. In this way, also the layers preceding the last one 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:

$$\alpha_c^k = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{2}$$

where $\frac{1}{Z} \sum_i \sum_j$ denotes the global average pooling operation ($Z = i \cdot j$) and $\frac{\partial y^c}{\partial A_{ij}^k}$ denotes the back-propagation gradients. In the gradient expression, y^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:

$$M_{Grad-CAM}^c(x, y) = ReLU\left(\sum_k \alpha_c^k A^k(x, y)\right) \tag{3}$$

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.

3.3. *Grad-CAM++*

Grad-CAM++ [17] 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, with the purpose of covering a wider portion of the object. Given a class c with a score Y^c and the activation map A_{ij}^k calculated in the last convolutional layer, a parameter α_{ij}^{kc} can be defined as follows:

$$\alpha_{ij}^{kc} = \frac{\frac{\partial^2 Y^c}{(\partial A_{ij}^k)^2}}{2\frac{\partial^2 Y^c}{(\partial A_{ij}^k)^2} + \sum_a \sum_b A_{ab}^k \{\frac{\partial^3 Y^c}{(\partial A_{ij}^k)^3}\}} \tag{4}$$

The parameter w_k^c, which has the same role of α_k^c in Grad-CAM, is defined as:

$$w_k^c = \sum_i \sum_j \alpha_{ij}^{kc} ReLU\left(\frac{\partial Y^c}{\partial A_{ij}^k}\right) \tag{5}$$

which leads to

$$w_k^c = \sum_{i,j} [\frac{\frac{\partial^2 Y^c}{(\partial A_{ij}^k)^2}}{2\frac{\partial^2 Y^c}{(\partial A_{ij}^k)^2} + \sum_{a,b} A_{ab}^k \{\frac{\partial^3 Y^c}{(\partial A_{ij}^k)^3}\}}] ReLU\left(\frac{\partial Y^c}{\partial A_{ij}^k}\right) \tag{6}$$

As in the other CAMs, it holds that

$$M_{Grad-CAM++}^c(x, y) = ReLU\left(\sum_k w_k^c A^k(x, y)\right) \tag{7}$$

3.4. *Smooth Grad-CAM++*

Smooth Grad-CAM++ [18] 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 random Gaussian perturbations 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:

$$M_{c(x,y)}(z) = \frac{1}{n}\sum_{1}^{n} M^{GCpp}_{c(x,y)}(z + \mathcal{N}(0, \sigma^2)) \tag{8}$$

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 a random Gaussian noise.

4. Evaluation

The evaluation exploits the ArtDL dataset [11], an existing artwork collection annotated with image-level labels. The purpose of the evaluation is: (1) to understand whether the class activation maps are effective in localizing both the whole representation of an iconographic class and the distinct symbols that characterize it (the attributes associated with the classes present in the ArtDL dataset are illustrated in [45] and listed in [46]); (2) to compare CAMs algorithms in their ability to do so. To evaluate the localization ability of class activation maps, a subset of the images have been annotated with bounding boxes framing iconographic symbols associated with each Saint. Figure 3 illustrates the symbols in a painting of Saint Jerome. The bounding boxes are used for the quantitative assessment of class activation maps algorithms with the metrics described in Section 4.5. A qualitative analysis is reported in Section 4.6.

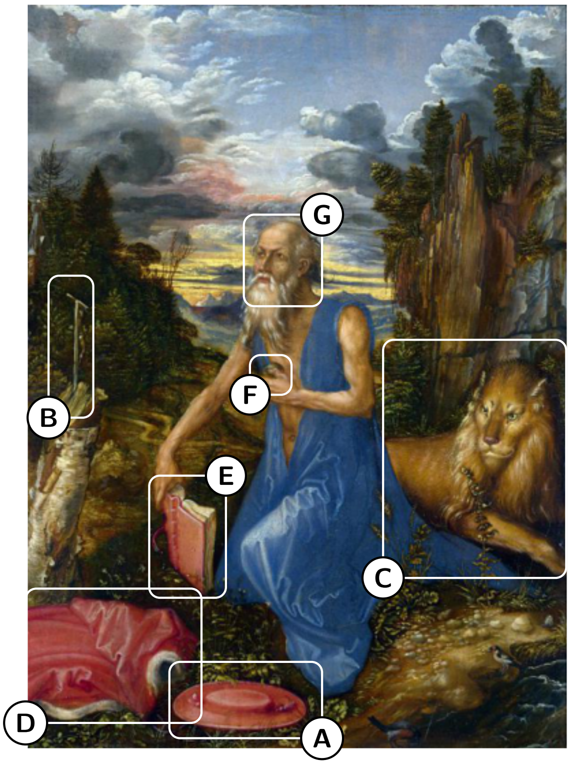

Figure 3. Saint Jerome—The cardinal's galero (**A**), the crucifix (**B**), the lion (**C**), the cardinal's vest (**D**), the book (**E**), the stone in the hand (**H**), and the face (**G**).

4.1. Dataset

The ArtDL dataset [11] comprises images of paintings that represent the Iconclass [47] 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. The representation of such classes in Christian art paintings exploit specific symbols, i.e., markers that hint at the identity of the portrayed character. Table 1 summarizes the symbols associated with the 10 Iconclass categories represented in the ArtDL dataset.

Table 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

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 the dataset with bounding boxes that localize the symbols listed in Table 1. Out of the whole dataset, 823 sample images were selected and manually annotated with bounding boxes that frame each symbol separately. A symbol can either be included completely within a single bounding box (e.g., Saint Jerome's lion) or be split into multiple bounding boxes (e.g., Saint Peter's bushy hair, which are usually divided in two parts separated by the forehead). We consider a symbol representation as 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 are annotated with multiple bounding boxes. When the same symbol relates to multiple saints (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 the symbol name and the Saint's name. 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. Table 2 summarizes the characteristics of the dataset used to compare the class activation maps algorithms.

Figure 4 shows the distribution of the bounding boxes within the images. Most images contain from 2 to 5 bounding boxes and a few images do not contain any annotation. The latter case occurs when the automatic classification of the ArtDL dataset is incorrect (e.g., for images in which a character named Mary was incorrectly associated with the Virgin Mary).

Table 2. 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

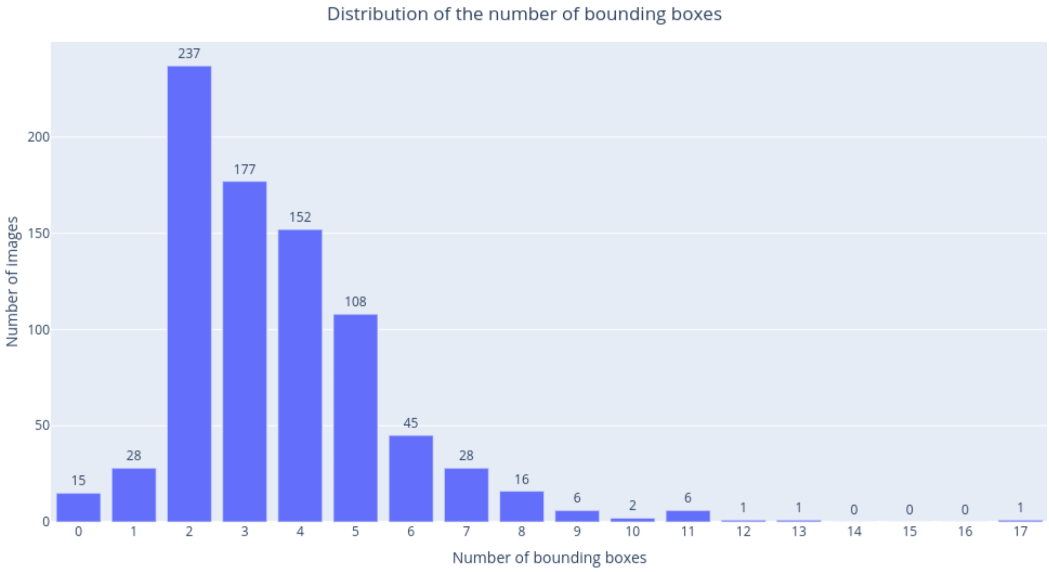

Figure 4. Bounding box distribution: most images contain from 2 to 5 bounding boxes (average = 3).

4.2. Class Activation Maps Generation

The class activation maps are generated by feeding the image to the ResNet50 model and applying the computation 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.

4.3. Choice of the Threshold Value

A class activation map contains values in the range from 0 to 1. Given a threshold t, it is possible to separate the class activation map into background (pixels with a value lower than t) and foreground (pixels with a value greater than t). The choice of the threshold value aims at making foreground areas concentrate on the Saints' figure and symbols. Figure 5 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 4.5 reports results obtained with multiple values uniformly distributed from 0 to 1 with a step of 0.05.

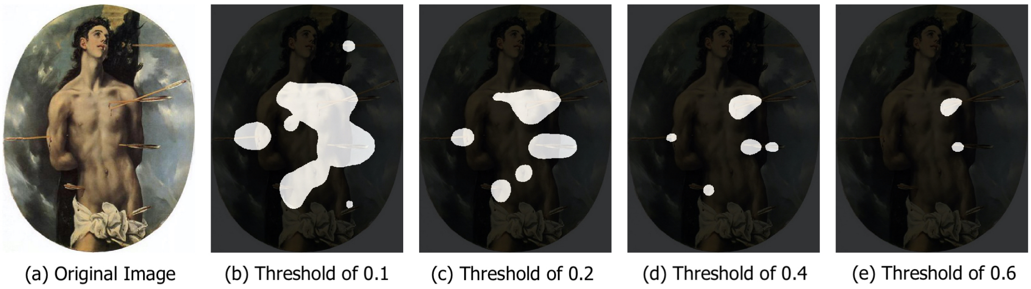

Figure 5. Analysis with different thresholds—black areas correspond to class activation map values below the specified threshold (background) while white pixels correspond to class activation map values greater or equal than the threshold (foreground). An increment in the threshold value results in smaller and more distinct areas. Original image (**a**), cam with threshold at 0.1 (**b**), cam with threshold at 0.2 (**c**), cam with threshold at 0.4 (**d**), cam with threshold at 0.6 (**e**).

4.4. Intersection Over Union Metrics

Intersection Over Union (IoU) is a standard metric used to compute the overlap between two different areas. It is defined as:

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

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. 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 of 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.

4.5. Quantitative Analysis

This section presents the results of comparing quantitatively 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 emphasize the boundary values reached by this algorithm. The number of samples is found to barely affect the results, whereas the standard deviation has more impact. To reduce the computational cost a lower number of samples is preferable.

Component IoU

This metric evaluates how well the class activation map focuses on the individual Saints' symbols. First, the class activation map 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 6 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 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 7 (San Sebastian's Martyrdom, Giovanni Maria Butteri, 1550–1559) 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.

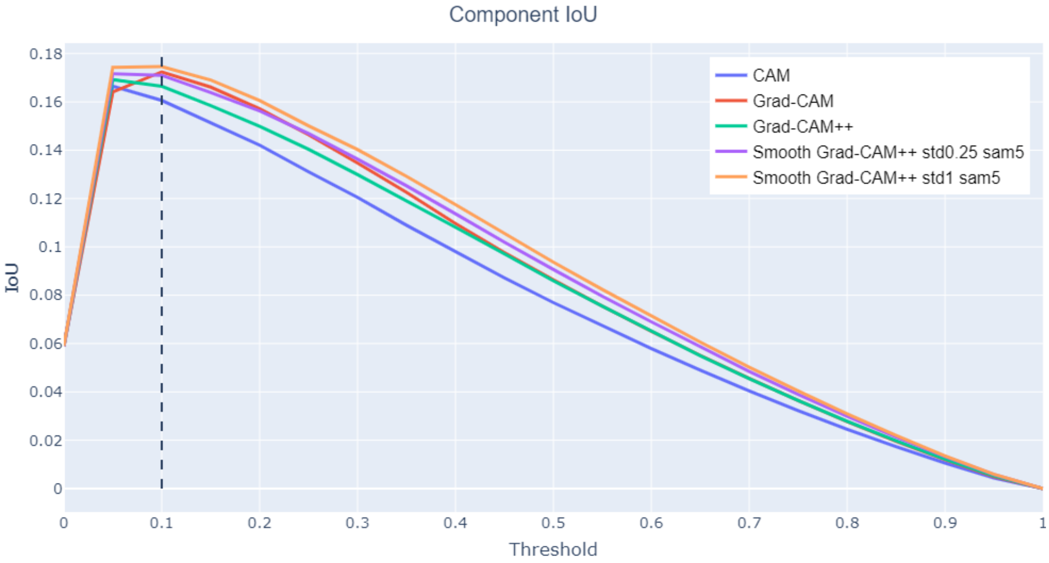

Figure 6. Component IoU at varying threshold levels.

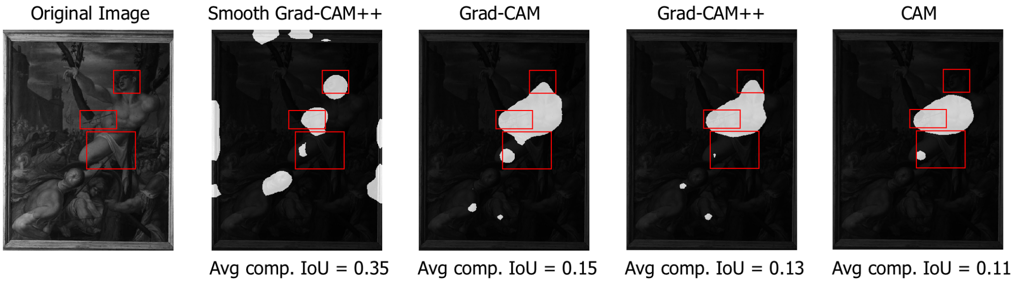

Figure 7. Different values of component IoU produced by different class activation map algorithms (Smooth Grad-CAM++ with $\sigma = 1$ and $s = 5$) at threshold $t = 0.1$. Ground-truth bounding boxes are shown in red.

Global IoU

An alternative metric 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 metric is calculated for all threshold values and assesses how the class activation map focuses on the whole representation of the Saint, favoring those class activation maps methods that generate wider and more connected areas rather than separated components. Figure 8 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 when the global metric is computed.

Figure 9 (Saint Jerome in the study, nd, 1604) compares the global IoU values produced on a sample image by different class activation map algorithms. For the same threshold value, Grad-CAM generates wider areas that cover more foreground pixels.

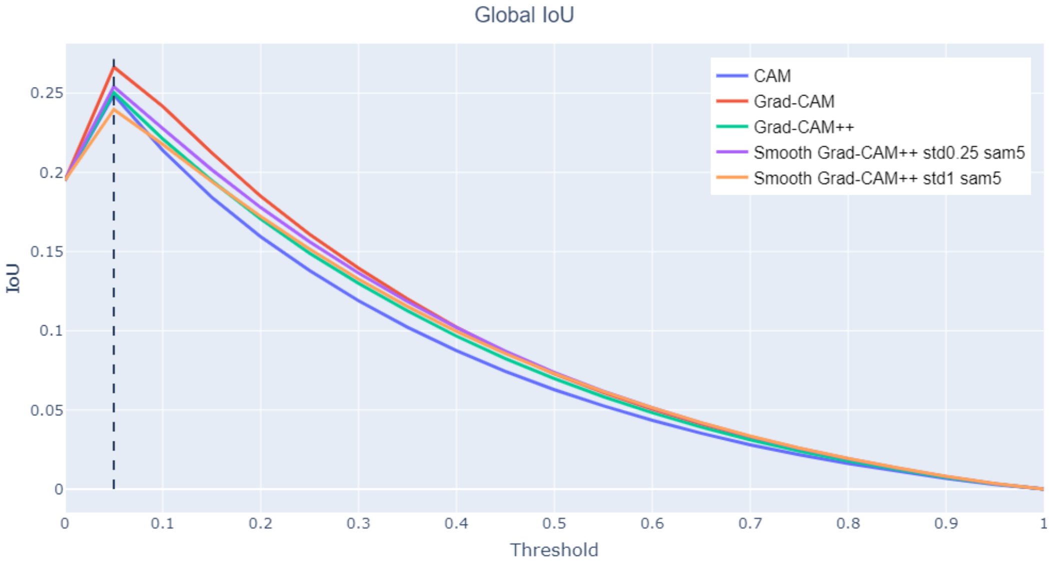

Figure 8. Global IoU at varying threshold levels.

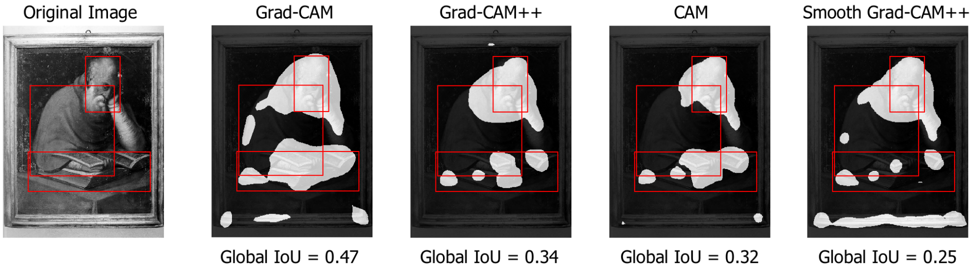

Figure 9. Different values of global IoU produced by different class activation map algorithms (Smooth Grad-CAM++ with $\sigma = 1$ and $s = 5$) at threshold $t = 0.05$. Manually annotated symbol bounding boxes are shown.

Bounding box coverage

When analyzing the class activation map algorithms, a factor to consider is also how many bounding boxes are covered by each class activation map. This metric alone is not enough to characterize the performance because a trivial class activation map that covers the entire image would have 100% coverage. However, coupled with the two previous metrics, it can give information about which method is able to generate class activation maps that can highlight a large fraction of the iconographic symbols that an expert would recognize. The bounding box coverage metric considers that a bounding box is covered by the class activation map only if their intersection is greater than or equal to 20% of the bounding box area. Figure 10 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 and less focused on the iconographic symbols with respect to the other approaches.

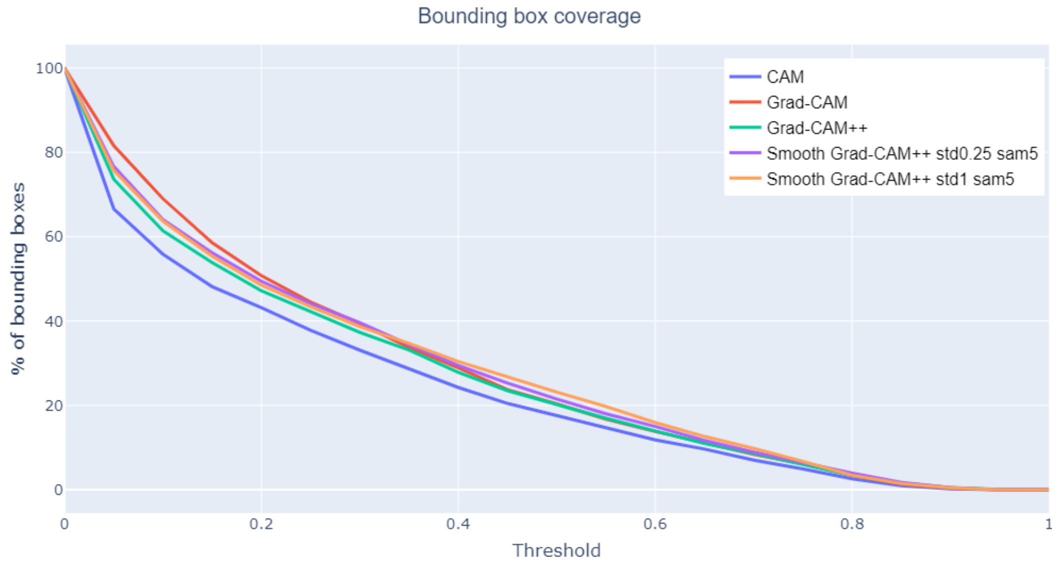

Figure 10. Bounding box coverage at varying threshold values.

Irrelevant attention

When evaluating the global IoU, a low value can occur for two reasons: (1) the two areas have a very small intersection or (2) the two areas overlap well but one is much larger than the other. Thus, an analysis on how much the class activation maps focus on irrelevant parts of the image helps characterizing low global IoU values. Irrelevant attention corresponds to the percentage of class activation map area outside any bounding box. Figure 11 shows that CAM has the less irrelevant attention, coherently with the previous results. Figure 12 (Madonna with Child and Infant St. John surrounded by Angels, Tiziano Vecellio, 1550) compares the irrelevant attention values produced on a sample image by different class activation map 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 Smooth Grad-CAM++ to focus on irrelevant areas can be seen also in Figures 7 and 9.

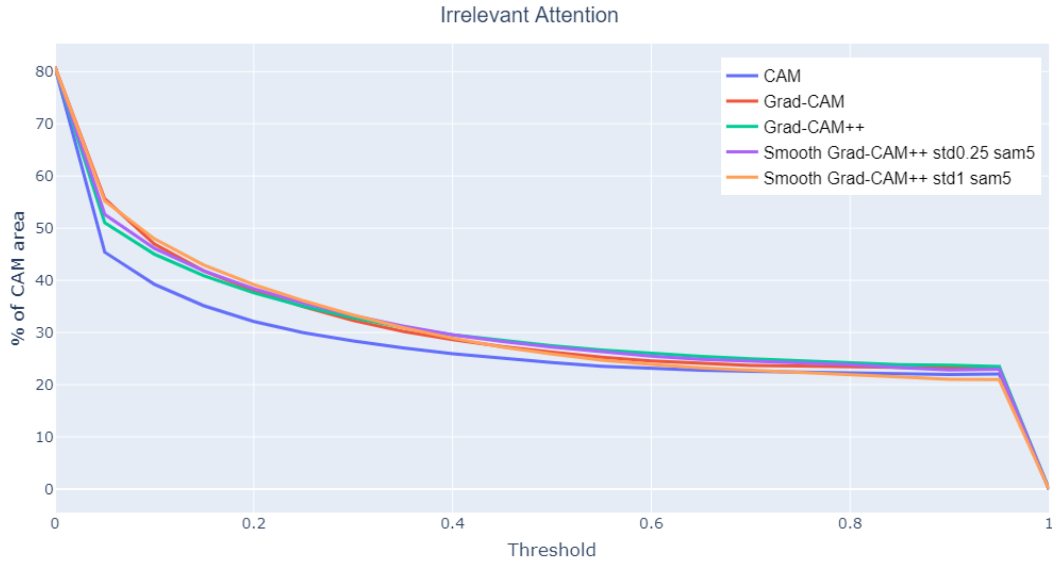

Figure 11. Irrelevant attention at varying threshold values.

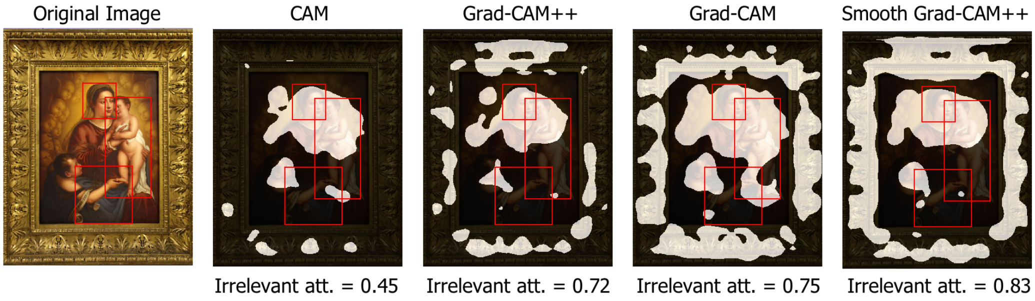

Figure 12. Different values of irrelevant attention produced by different class activation map algorithms (Smooth Grad-CAM++ with $\sigma = 1$ and $s = 5$) at threshold $t = 0.1$. Manually annotated symbol bounding boxes are reported.

4.6. Qualitative Analysis

This section presents a qualitative analysis of the results obtained by the different class activation map algorithms highlighting their capabilities and limitations. Each example shows the original image, the class activation maps generated by each algorithm (with background in black and foreground in white) and the ground-truth bounding boxes.

Positive examples

Figure 13 (Saint Jerome in his Study, Jan van Remmerswale, 1533) shows an example in which all the algorithms focus well on the iconographic symbols. The image contains seven symbols with different size, shape and position, which are all identified and separated by the class activation map algorithms. The irrelevant area on the top right corresponds to a piece of the cardinal's vest that has the same color and approximate shape of the cardinal's galero appearing in many paintings of Saint Jerome.

Figure 14 (St. Peter, Antonio Veneziano, 1369–1375) 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 map algorithms identify four out of the five symbols. The central ground-truth bounding box is not identified because it corresponds to a rather generic attribute (the bishop's vest), which is not evident in the drawing. Only CAM misses the book, which the other algorithms identify by focusing on the characteristic marks on the spine of the book or on the lock. The example of Figure 14 and many similar ones of black and white and poor quality images highlight the ability of class activation map algorithms to extract useful maps also when the image has low discriminative features.

A counterexample of the difficulty of detecting such generic attributes as the vest is illustrated in Figure 15 (Saint Dominic, Carlo Crivelli, 1472). The vest is identified thanks to a specific detail: the change of color typical of the black and white Dominican habit.

Negative examples

Class activation maps algorithms tend to fail consistently in two cases: when multiple symbols are too close or have a substantial overlap and when the representation of a symbol is rather generic and covers a wide area of the image. Figure 16 (Penitent St. Peter, Jusepe de Ribera, 1600–1649) illustrates a typical example: Saint Peter's bushy hair and beard are merged into a single region and the vest, which is a rather generic attribute, is missed completely or highlighted only through small irrelevant details.

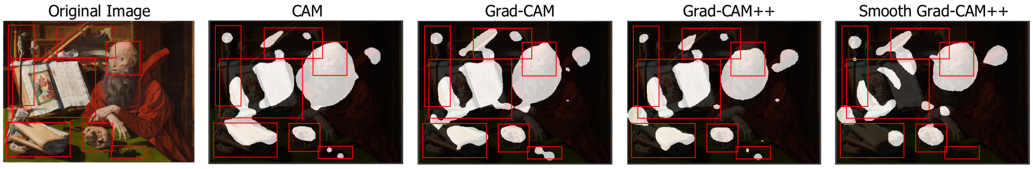

Figure 13. Class activation maps with seven recognized symbols associated with Saint Jerome.

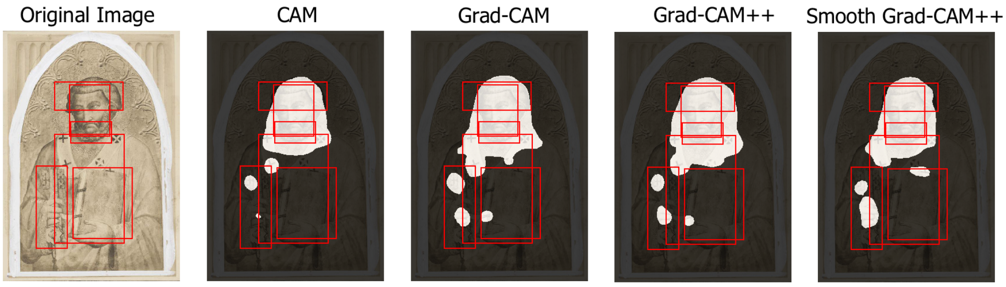

Figure 14. Class activation maps extracted from a drawing of Saint Peter. Four out of five symbols are identified despite their low visibility.

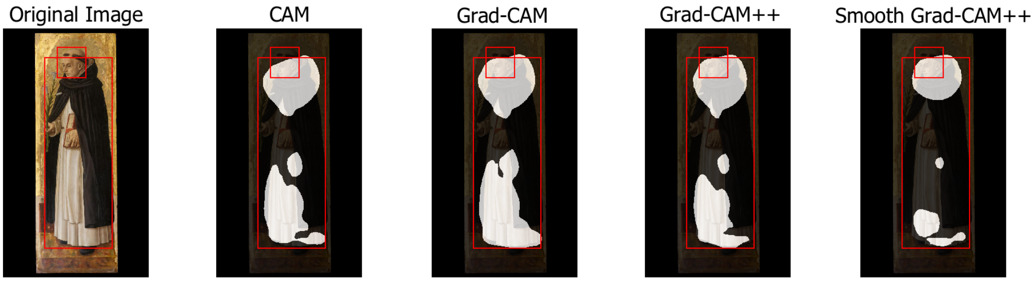

Figure 15. Class activation maps extracted from a paining of Saint Dominic. The rather generic vest attribute is identified by focusing on its double color.

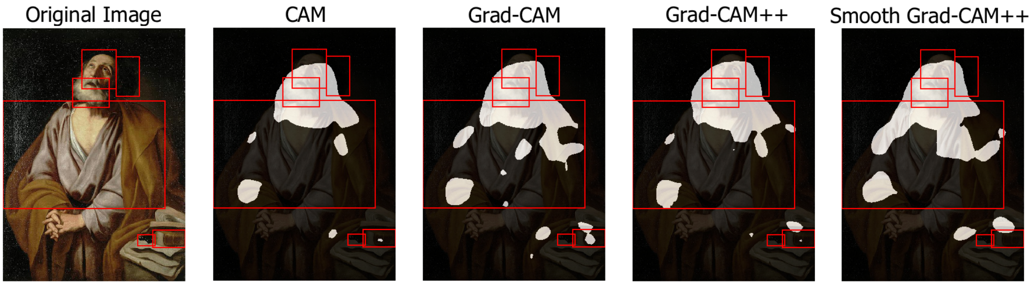

Figure 16. Class activation maps with merged symbols and missed generic attributes.

Relevant irrelevant regions

An interesting case occurs when the class activation map algorithms focus on an apparently irrelevant area, which instead contains a relevant iconographic attribute not present in the ground truth. Figure 17 illustrates three examples. The painting of Saint John the Baptist (a) (portrait of François I as St John the Baptist, Jean Clouet, 1518) contains an apparently 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 iconographic symbols used to annotate the images but appears in some of the paintings. The same happens with Saint Jerome (b) (Saint Jerome, Albrecht Durer, 1521), where the class activation map 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 Saint Jerome (c) (Landscape with St. Jerome, Simon Bening, 1515–1520), where the class activation map algorithms focus on the outdoor environment. This is a well-known symbol associated 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.

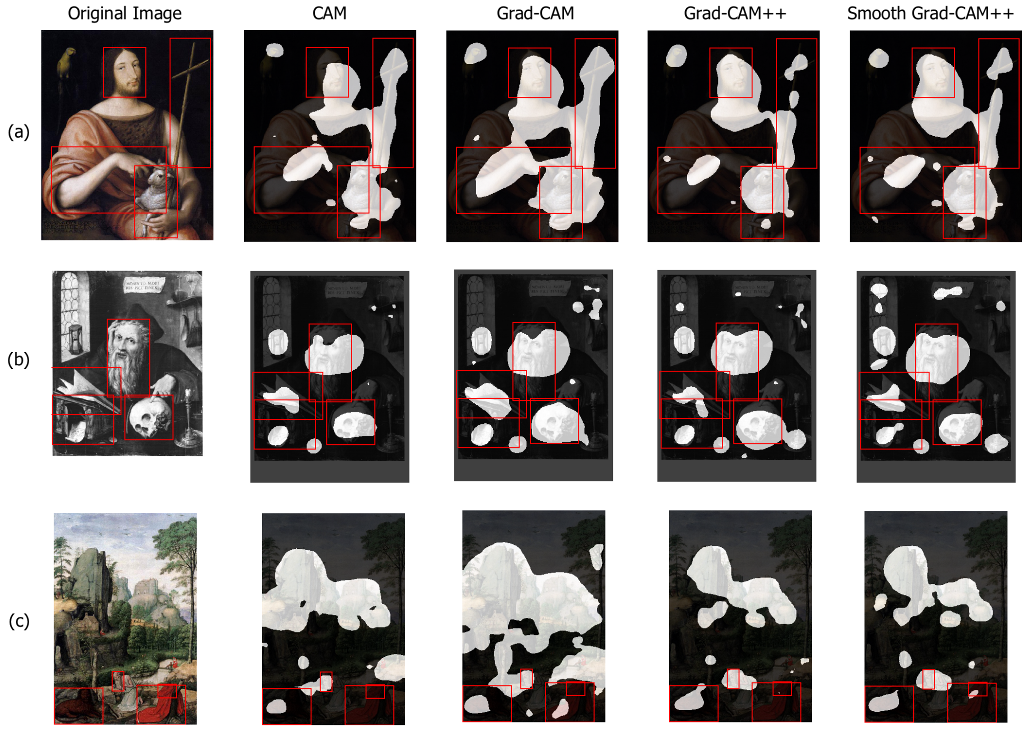

Figure 17. Class activation maps highlighting regions containing relevant iconographic attributes not present in the ground truth: a bird associated with Saint John the Baptist (**a**) an hourglass associated with Saint Jerome (**b**) and the wilderness where Saint Jerome retired (**c**).

Confusion with unknown co-occurring class

Figure 18 (Baptism of Christ, Pietro Perugino, 1510) presents an example in which all analyzed 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 dataset. 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 dataset were annotated with the Iconclass category for Jesus.

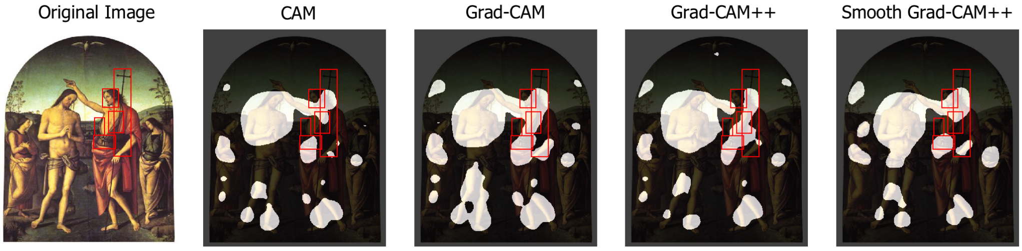

Figure 18. Class activation maps with confusion between Saint John the Baptist and Jesus Christ.

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 for 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 iconographic 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 of 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 [48]. To investigate the potential of the class activation maps to support 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 6) and for whole Saint bounding boxes global IoU greater than 0.24 at threshold 0.05 (see Figure 8);
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 12);
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.

Symbol bounding boxes

Figure 19 presents some examples of the computed symbol bounding boxes (green) compared with the ground-truth bounding boxes (red). The proposed procedure is able to generate boxes that in many cases correctly highlight and distinguish the most important iconographic 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 19).

Whole Saint bounding boxes

Figure 20 illustrates some examples of computed whole Saint bounding boxes (green) compared with the ground-truth boxes (red). The automatically generated bounding boxes localize almost entirely the Saint's figure and include only very small irrelevant areas.

Figures 19 and 20 show that the simple procedure for processing class activation map 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.

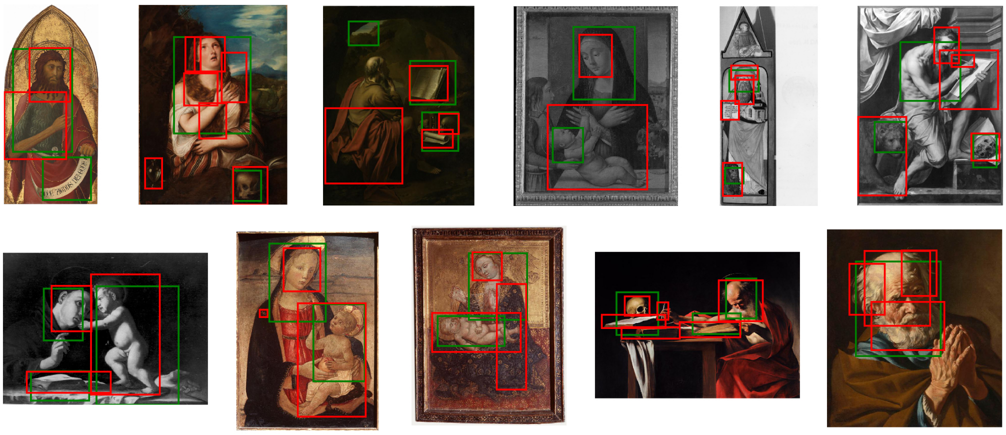

Figure 19. Examples of symbols bounding boxes generated from Grad-CAM (green) and manually annotated (red).

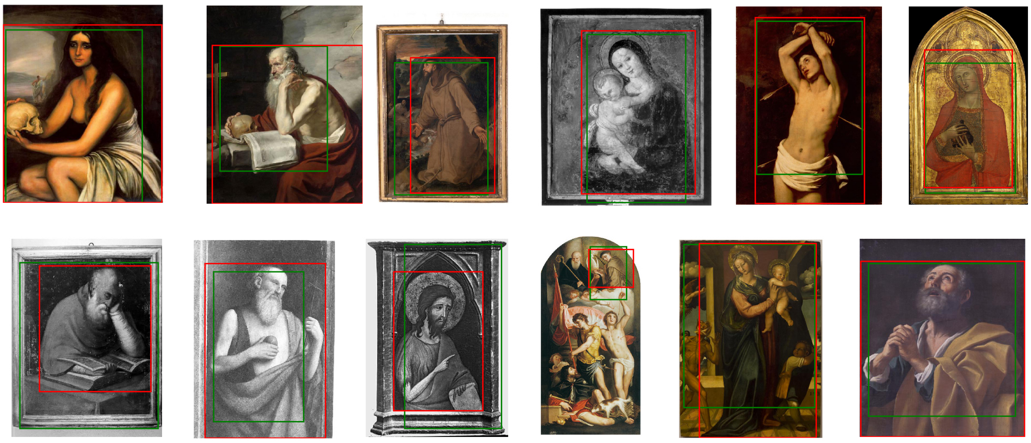

Figure 20. Examples of Saints bounding boxes generated from Grad-CAM (green) and manually annotated (red).

Quantitative evaluation of whole Saint bounding boxes

For the whole Saint case, each estimated bounding box can be labeled 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 ([49–51]) 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. Results are reported in Table 3: 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 21 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 shows the mAP values obtained with GradCAM on the ten ArtDL classes.

The whole Saint estimated bounding boxes appear to be 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 an object detection is usually evaluated with a minimal IoU threshold of 0.5 and 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, which is 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 datasets similar to ArtDL [10,52]. 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.

Table 3. 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 $\geq$ 0.5)
CAM	0.489	49.70	0.206
GradCAM	0.551	61.20	0.316
GradCAM++	0.529	59.88	0.292
Smooth-GradCAM++	0.544	61.18	0.307

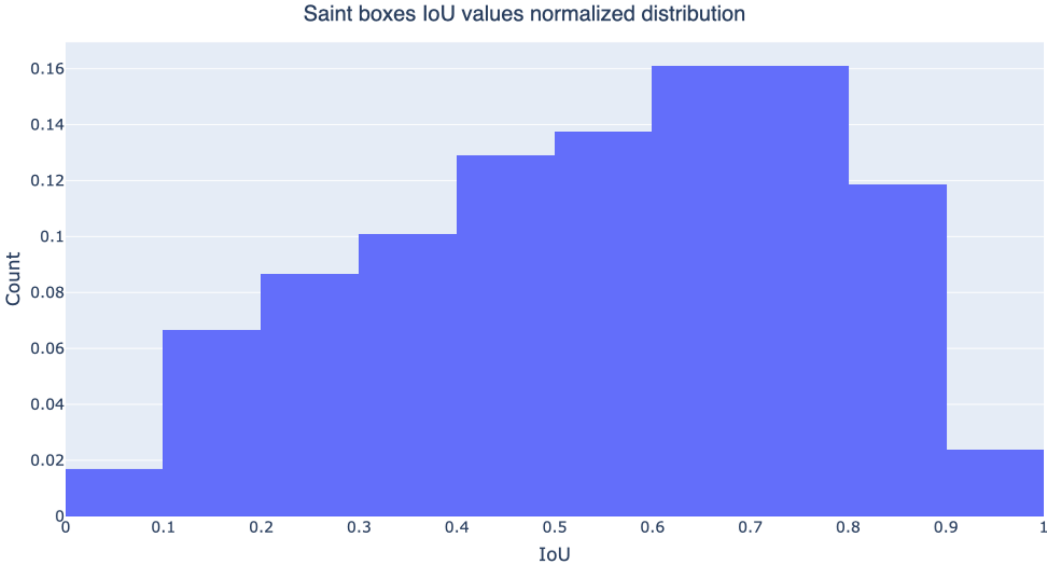

Figure 21. Normalized distribution of IoU values between whole-Saint Grad-CAM estimated bounding boxes and ground-truth bounding boxes.

Table 4. Mean Average Precision (mAP) values for each class of the ArtDL dataset. Bounding boxes are estimated with GradCAM.

Anthony	John	Paul	Francis	Magdalene	Jerome	Dominic	Virgin	Peter	Sebastian
0.076	0.289	0.173	0.33	0.616	0.228	0.142	0.442	0.399	0.468

5. Conclusions and Future Work

This work has presented a comparative study about the effectiveness of class activation maps as a tool for explaining of 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 metric 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 [18], in our use case Grad-CAM is the method of choice for deriving the bounding boxes from class activation maps necessary to train a weakly supervised detector.

Future work will concentrate on the comparison of other activation mapping techniques [50,51,53,54]. In particular, [50,51] 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 will be used to pursue the ultimate goal of our research, which is to use the output of class activation maps to create training datasets for weakly supervised iconographic symbol detection and segmentation. The implementation of an automated system for iconographic analysis of artworks could promote the development of educational applications for art history experts and students. Finally, another future research path consists in addressing more complex Iconclass categories involving complex scenes (e.g., the crucifixion, the nativity, the visitation of the magi, etc.) and in exploring the iconography of other cultures.

Author Contributions: Conceptualization, N.O.P.V., F.M., P.F. and R.d.S.T.; Methodology, N.O.P.V., F.M., P.F. and R.d.S.T.; Validation, N.O.P.V., F.M., P.F. and R.d.S.T.; Writing—original draft, N.O.P.V., F.M., P.F. and R.d.S.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: http://www.artdl.org (accessed on 29 June 2021).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Panofsky, E. *Studies in Iconology: Humanistic Themes in the Art of the Renaissance*; Routledge Taylor and Francis Group: New York, NY, USA, 1939; p. 262.
2. Proulx, D.A. *A sourcebook of Nasca Ceramic Iconography: Reading a Culture through Its Art*; University of Iowa Press: Iowa City, IA, USA, 2009.
3. Parani, M.G. *Reconstructing the Reality of Images: Byzantine Material Culture and Religious Iconography 11th-15th Centuries*; Brill: Leiden, The Netherlands, 2003; Volume 41.
4. Van Leeuwen, T.; Jewitt, C. *The Handbook of Visual Analysis*; Sage: Thousand Oaks, CA, USA, 2001; pp. 100–102.
5. King, J.N. *Tudor Royal Iconography: Literature and Art in an Age of Religious Crisis*; Princeton University Press: Princeton, NJ, USA, 1989.
6. Roberts, H.E. *Encyclopedia of Comparative Iconography: Themes Depicted in Works of Art*; Routledge: London, UK, 2013.
7. Zujovic, J.; Gandy, L.; Friedman, S.; Pardo, B.; Pappas, T.N. Classifying paintings by artistic genre: An analysis of features classifiers. In Proceedings of the 2009 IEEE International Workshop on Multimedia Signal Processing, Rio de Janeiro, Brazil, 5–7 October 2009; pp. 1–5.
8. Shamir, L.; Tarakhovsky, J.A. Computer Analysis of Art. *J. Comput. Cult. Herit.* **2012**, *5*. [CrossRef]
9. Cai, H.; Wu, Q.; Corradi, T.; Hall, P. The Cross-Depiction Problem: Computer Vision Algorithms for Recognising Objects in Artwork and in Photographs. *arXiv* **2015**, arXiv:1505.00110.

10. Gonthier, N.; Gousseau, Y.; Ladjal, S.; Bonfait, O. Weakly supervised object detection in artworks. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, Munich, Germany, 8–14 September 2018.
11. Milani, F.; Fraternali, P. A Data Set and a Convolutional Model for Iconography Classification in Paintings. *arXiv* **2020**, arXiv:2010.11697.
12. Sun, K.H.; Huh, H.; Tama, B.A.; Lee, S.Y.; Jung, J.H.; Lee, S. Vision-Based Fault Diagnostics Using Explainable Deep Learning With Class Activation Maps. *IEEE Access* **2020**, *8*, 129169–129179. [CrossRef]
13. Patro, B.; Lunayach, M.; Patel, S.; Namboodiri, V. U-CAM: Visual Explanation Using Uncertainty Based Class Activation Maps. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 7443–7452. [CrossRef]
14. Yang, S.; Kim, Y.; Kim, Y.; Kim, C. Combinational Class Activation Maps for Weakly Supervised Object Localization. In Proceedings of the 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), Snowmass Village, CO, USA, 1–5 March 2020; pp. 2930–2938. [CrossRef]
15. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning Deep Features for Discriminative Localization. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Vegas, NV, USA, 27–30 June 2016; pp. 2921–2929.
16. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 618–626.
17. Chattopadhay, A.; Sarkar, A.; Howlader, P.; Balasubramanian, V.N. Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018. [CrossRef]
18. Omeiza, D.; Speakman, S.; Cintas, C.; Weldermariam, K. Smooth grad-cam++: An enhanced inference level visualization technique for deep convolutional neural network models. *arXiv* **2019**, arXiv:1908.01224.
19. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. *arXiv* **2015**, arXiv:1512.03385.
20. Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; Li, F. ImageNet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), Miami, FL, USA, 20–25 June 2009; IEEE Computer Society: New York, NY, USA, 2009; pp. 248–255. [CrossRef]
21. Karayev, S.; Trentacoste, M.; Han, H.; Agarwala, A.; Darrell, T.; Hertzmann, A.; Winnemoeller, H. Recognizing image style. *arXiv* **2013**, arXiv:1311.3715.
22. Crowley, E.J.; Zisserman, A. *The State of the Art: Object Retrieval in Paintings Using Discriminative Regions*; British Machine Vision Association: Durham, UK, 2014.
23. Khan, F.S.; Beigpour, S.; Van de Weijer, J.; Felsberg, M. Painting-91: A large scale database for computational painting categorization. *Mach. Vis. Appl.* **2014**, *25*, 1385–1397. [CrossRef]
24. Strezoski, G.; Worring, M. Omniart: Multi-task deep learning for artistic data analysis. *arXiv* **2017**, arXiv:1708.00684.
25. Mao, H.; Cheung, M.; She, J. Deepart: Learning joint representations of visual arts. In Proceedings of the 25th ACM International Conference on Multimedia, Mountain View, CA, USA, 23–27 October 2017; pp. 1183–1191.
26. Bianco, S.; Mazzini, D.; Napoletano, P.; Schettini, R. Multitask painting categorization by deep multibranch neural network. *Expert Syst. Appl.* **2019**, *135*, 90–101. [CrossRef]
27. Castellano, G.; Vessio, G. Deep learning approaches to pattern extraction and recognition in paintings and drawings: An overview. In *Neural Computing and Applications*; Springer: New York, NY, USA, 2021; pp. 1–20.
28. Santos, I.; Castro, L.; Rodriguez-Fernandez, N.; Torrente-Patino, A.; Carballal, A. Artificial Neural Networks and Deep Learning in the Visual Arts: A review. In *Neural Computing and Applications*; Springer: New York, NY, USA, 2021; pp. 1–37.
29. Zhao, W.; Zhou, D.; Qiu, X.; Jiang, W. Compare the performance of the models in art classification. *PLoS ONE* **2021**, *16*, e0248414. [CrossRef]
30. Gao, Z.; Shan, M.; Li, Q. Adaptive sparse representation for analyzing artistic style of paintings. *J. Comput. Cult. Herit. (JOCCH)* **2015**, *8*, 1–15. [CrossRef]
31. Elgammal, A.; Kang, Y.; Den Leeuw, M. 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, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
32. Crowley, E.J.; Zisserman, A. Of gods and goats: Weakly supervised learning of figurative art. *Learning* **2013**, *8*, 14.
33. Shen, X.; Efros, A.A.; Aubry, M. Discovering Visual Patterns in Art Collections with Spatially-consistent Feature Learning. *arXiv* **2019**, arXiv:1903.02678.
34. Kadish, D.; Risi, S.; Løvlie, A.S. Improving Object Detection in Art Images Using Only Style Transfer. *arXiv* **2021**, arXiv:2102.06529.
35. Banar, N.; Daelemans, W.; Kestemont, M. *Multi-Modal Label Retrieval for the Visual Arts: The Case of Iconclass*; Scitepress: Setúbal, Portugal, 2021.
36. Gonthier, N.; Gousseau, Y.; Ladjal, S. An analysis of the transfer learning of convolutional neural networks for artistic images. *arXiv* **2020**, arXiv:2011.02727.

37. Cömert, C.; Özbayoğlu, M.; Kasnakoğlu, C. Painter Prediction from Artworks with Transfer Learning. In Proceedings of the IEEE 2021 7th International Conference on Mechatronics and Robotics Engineering (ICMRE), Budapest, Hungary, 3–5 February 2021; pp. 204–208.
38. Belhi, A.; Ahmed, H.O.; Alfaqheri, T.; Bouras, A.; Sadka, A.H.; Foufou, S. Study and Evaluation of Pre-trained CNN Networks for Cultural Heritage Image Classification. In *Data Analytics for Cultural Heritage: Current Trends and Concepts*; Springer: Cham, Switerland, 2021; p. 47.
39. Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; Pedreschi, D. A survey of methods for explaining black box models. *ACM Comput. Surv. (CSUR)* **2018**, *51*, 1–42. [CrossRef]
40. Buhrmester, V.; Münch, D.; Arens, M. Analysis of explainers of black box deep neural networks for computer vision: A survey. *arXiv* **2019**, arXiv:1911.12116.
41. Gupta, V.; Demirer, M.; Bigelow, M.; Yu, S.M.; Yu, J.S.; Prevedello, L.M.; White, R.D.; Erdal, B.S. Using Transfer Learning and Class Activation Maps Supporting Detection and Localization of Femoral Fractures on Anteroposterior Radiographs. In Proceedings of the 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), Iowa City, IA, USA, 3–7 April 2020; pp. 1526–1529.
42. Zhang, M.; Zhou, Y.; Zhao, J.; Man, Y.; Liu, B.; Yao, R. A survey of semi-and weakly supervised semantic segmentation of images. *Artif. Intell. Rev.* **2020**, 53, 4259–4288. . [CrossRef]
43. Lin, M.; Chen, Q.; Yan, S. Network in network. *arXiv* **2013**, arXiv:1312.4400.
44. Qiu, S. Global Weighted Average Pooling Bridges Pixel-level Localization and Image-level Classification. *arXiv* **2018**, arXiv:1809.08264.
45. Lanzi, F.; Lanzi, G. *Saints and Their Symbols: Recognizing Saints in Art and in Popular Images*; Liturgical Press: Collegeville, MN, USA, 2004; pp. 327–342.
46. Wikipedia: Saint Symbolism. https://en.wikipedia.org/wiki/Saint_symbolism (accessed on 24 April 2021).
47. Couprie, L.D. Iconclass: An iconographic classification system. *Art Libr. J.* **1983**, *8*, 32–49. [CrossRef]
48. Zhang, D.; Han, J.; Cheng, G.; Yang, M.H. Weakly Supervised Object Localization and Detection: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**. [CrossRef]
49. Singh, K.K.; Lee, Y.J. Hide-and-Seek: Forcing a Network to be Meticulous for Weakly-supervised Object and Action Localization. *arXiv* **2017**, arXiv:1704.04232.
50. Choe, J.; Shim, H. Attention-Based Dropout Layer for Weakly Supervised Object Localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition CVPR 2019, Long Beach, CA, USA, 16–20 June 2019; pp. 2219–2228. [CrossRef]
51. Bae, W.; Noh, J.; Kim, G. Rethinking Class Activation Mapping for Weakly Supervised Object Localization. In *Part XV, Proceedings of the Computer Vision - ECCV 2020—16th European Conference, Glasgow, UK, 23–28 August 2020*; Vedaldi, A., Bischof, H., Brox, T., Frahm, J., Eds.; Lecture Notes in Computer Science; Springer: New York, NY, USA, 2020; Volume 12360, pp. 618–634. [CrossRef]
52. Gonthier, N.; Ladjal, S.; Gousseau, Y. Multiple instance learning on deep features for weakly supervised object detection with extreme domain shifts. *arXiv* **2020**, arXiv:2008.01178.
53. Wang, H.; Du, M.; Yang, F.; Zhang, Z. Score-cam: Improved visual explanations via score-weighted class activation mapping. *arXiv* **2019**, arXiv:1910.01279.
54. Zhao, G.; Zhou, B.; Wang, K.; Jiang, R.; Xu, M. Respond-cam: Analyzing deep models for 3d imaging data by visualizations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: New York, NY, USA, 2018; pp. 485–492.