Weakly Supervised Object Detection in Artworks

Nicolas Gonthier[1(✉)], Yann Gousseau[1], Said Ladjal[1], and Olivier Bonfait[2]

[1] LTCI, Telecom ParisTech, Universite Paris-Saclay, 75013 Paris, France
{nicolas.gonthier,yann.gousseau,said.ladjal}@telecom-paristech.fr
[2] Universite de Bourgogne, UMR CNRS UB 5605, 21000 Dijon, France

Abstract. We propose a method for the weakly supervised detection of objects in paintings. At training time, only image-level annotations are needed. This, combined with the efficiency of our multiple-instance learning method, enables one to learn new classes on-the-fly from globally annotated databases, avoiding the tedious task of manually marking objects. We show on several databases that dropping the instance-level annotations only yields mild performance losses. We also introduce a new database, IconArt, on which we perform detection experiments on classes that could not be learned on photographs, such as Jesus Child or Saint Sebastian. To the best of our knowledge, these are the first experiments dealing with the automatic (and in our case weakly supervised) detection of iconographic elements in paintings. We believe that such a method is of great benefit for helping art historians to explore large digital databases.

Keywords: Weakly supervised detection · Transfer learning · Art analysis · Multiple instance learning

1 Introduction

Several recent works show that recycling analysis tools that have been developed for natural images (photographs) can yield surprisingly good results for analysing paintings or drawings. In particular, impressive classification results are obtained on painting databases by using convolutional neural networks (CNNs) designed for the classification of photographs [10,55]. These results occur in a general context were methods of transfer learning [14] (changing the task a model was trained for) and domain adaptation (changing the nature of the data a model was trained on) are increasingly applied. Classifying and analysing paintings is of course of great interest to art historians, and can help them to take full advantage of the massive artworks databases that are built worldwide.

More difficult than classification, and at the core of many recent computer vision works, the object detection task (classifying and localising an object) has been less studied in the case of paintings, although exciting results have been obtained, again using transfer techniques [11,28,52].

L. Leal-Taixé and S. Roth (Eds.): ECCV 2018 Workshops, LNCS 11130, pp. 692–709, 2019.
https://doi.org/10.1007/978-3-030-11012-3_53

Methods that detect objects in photographs have been developed thanks to massive image databases on which several classes (such as cats, people, cars) have been manually localised with bounding boxes. The PASCAL VOC [17] and MS COCO [34] datasets have been crucial in the development of detection methods and the recently introduced Google Open Image Dataset (2M images, 15M boxes for 600 classes) is expected to push further the limits of detection. Now, there is no such database (with localised objects) in the field of Art History, even though large databases are being build by many institutions or academic research teams, e.g. [16,38,39,43,44,53]. Some of these databases include image-level annotations, but none includes location annotations. Besides, manually annotating such large databases is tedious and must be performed each time a new category is searched for. Therefore, it is of great interest to develop *weakly supervised* detection methods, that can learn to detect objects using image-level annotations only. While this aspect has been thoroughly studied for natural images, only a few studies have been dedicated to the case of painting or drawings.

Moreover, these studies are mostly dedicated to the cross depiction problem: they learn to detect the same objects in photographs and in paintings, in particular man-made objects (cars, bottles ...) or animals. While these may be useful to art historians, it is obviously needed to detect more specific objects or attributes such as ruins or nudity, and characters of iconographic interest such as Mary, Jesus as a child or the crucifixion of Jesus, for instance. These last categories can hardly be directly inherited from photographic databases.

For these two reasons, the lack of location annotations and the specificity of the categories of interest, a general method allowing the weakly supervised detection on specific domains such as paintings would be of great interest to art historians and more generally to anyone needing some automatic tools to explore artistic databases. We propose some contributions in this direction:

- We introduce a new multiple-instance learning (MIL) technique that is simple and quick enough to deal with large databases,
- We demonstrate the utility of the proposed technique for object detection on weakly annotated databases, including photographs, drawings and paintings. These experiments are performed using image-level annotations only.
- We propose the first experiments dealing with the recognition and detection of iconographic elements that are specific to Art History, exhibiting both successful detections and some classes that are particularly challenging, especially in a weakly supervised context.

We believe that such a system, enabling one to detect new and unseen category with minimal supervision, is of great benefit for dealing efficiently with digital artwork databases. More precisely, iconographic detection results are useful for different and particularly active domains of humanities: Art History (to gather data relative to the iconography of recurrent characters, such as the Virgin Mary or San Sebastian, as well as to study the formal evolution of their representations), Semiology (to infer mutual configurations or relative dimensions of the iconographic elements), History of Ideas and Cultures (with category such as nudity, ruins), Material Culture Studies, etc.

In particular, being able to detect iconographic elements is of great importance for the study of spatial configurations, which are central to the reading of images and particularly timely given the increasing importance of Semiology. To fix ideas, we can give two examples of potential use. First, the order in which iconographic elements are encountered (e.g. Gabriel and Mary), when reading an image from left to right, has received much attention from art historians [20]. In the same spirit, recent studies [5] on the meaning of mirror images in early modern Italy could benefit from the detection of iconographic elements.

2 Related Work

Object Recognition and Detection in Artworks. Early works on cross-domain (or cross-depiction) image comparisons were mostly concerned with sketch retrieval, see e.g. [12]. Various local descriptors were then used for comparing and classifying images, such as part-based models [46] or mid-level discriminative patches [2,9]. In order to enhance the generalisation capacity of these approaches, it was proposed in [54] to model object through graphs of labels. More generally, it was shown in [25] that structured models are more prone to succeed in cross-domain recognition than appearance-based models.

Next, several works have tried to transfer the tremendous classification capacity of convolutional neural networks to perform cross-domain object recognition, in particular for paintings. In [10], it is shown that recycling CNNs directly for the task of recognising objects in paintings, without fine-tuning, yields surprisingly good results. Similar conclusions were also given in [55] for artistic drawings. In [32], a robust low rank parametrized CNN model is proposed to recognise common categories in an unseen domain (photo, painting, cartoon or sketch). In [53], a new annotated database is introduced, on which it is shown that fine-tuning improves recognition performances. Several works have also successfully adapted CNNs architectures to the problem of style recognition in artworks [3,31,36]. More generally, the use of CNNs opens the way to other artwork analysis tasks, such as visual links retrieval [45], scene classification [19], author classification [51] or possibly to generic artwork content representation [48].

The problem of *object detection* in paintings, that is, being able to both localise and recognise objects, has been less studied. In [11], it is shown that applying a pre-trained object detector (Faster R-CNN [42]) and then selecting the localisation with highest confidence can yield correct detections of PASCAL VOC classes. Other works attacked this difficult problem by restricting it to a single class. In [22], it is shown that deformable part model outperforms other approaches, including some CNNs, for the detection of people in cubist artworks. In [40], it is shown that the YOLO network trained on natural images can, to some extend, be used for people detection in cubism. In [52], it is proposed to perform people detection in a wide variety of artworks (through a newly introduced database) by fine-tuning a network in a supervised way. People can be detected with high accuracy even though the database has very large stylistic variations and includes paintings that strongly differs from photographs in the way they represent people.

Weakly supervised detection refers to the task of learning an object detector using limited annotations, usually image-level annotations only. Often, a set of detections (e.g. bounding boxes) is considered at image level, assuming we only know if at least one of the detection corresponds the category of interest. The corresponding statistical problem is referred to as multiple instance learning (MIL) [13]. A well-known solution to this problem through a generalisation of Support Vector Machine (SVM) has been proposed in [1]. Several approximations of the involved non-convex problem have been proposed, see e.g. [21] or the recent survey [6].

Recently, this problem has been attacked using classification and detection neural networks. In [47], it is proposed to learn a smooth version of an SVM on the features from R-CNN [23] and to focus on the initialisation phase which is crucial due to the non-convexity of the problem. In [41], it is proposed to learn to detect new specific classes by taking advantage of the knowledge of wider classes. In [4] a weakly supervised deep detection network is proposed based on Fast R-CNN [24]. Those works have been improved in [50] by adding a multi-stage classifier refinement. In [8] a multi-fold split of the training data is proposed to escape local optima. In [33], a two step strategy is proposed, first collecting good regions by a mask-out classification, then selecting the best positive region in each image by a MIL formulation and then fine-tuning a detector with those propositions as "ground truth" bounding boxes. In [15] a new pooling strategy is proposed to efficiently learn localisation of objects without doing bounding boxes regression.

Weakly supervised strategies for the cross domain problem have been much less studied. In [11], a relatively basic methodology is proposed, in which for each image the bounding box with highest (class agnostic) "objectness" score is classified. In [28], it is proposed to do mixed supervised object detection with cross-domain learning based on the SSD network [35]. Object detectors are learnt by using instance-level annotations on photographs and image-level annotations on a target domain (watercolor, cartoon, etc.). We will perform comparisons of our approach with these two methods in Sect. 4.

3 Weakly Supervised Detection by Transfer Learning

In this section, we propose our approach to the weakly supervised detection of visual category in paintings. In order to perform transfer learning, we first apply Faster R-CNN [42] (a detection network trained on photographs) which is used as a feature extractor, in the same way as in [11]. This results in a set of candidate bounding boxes. For a given visual category, the goal is then, using image-level annotations only, to decide which boxes correspond to this category. For this, we propose a new multiple-instance learning method, that will be detailed in Sect. 3.1. In contrast with classical approaches to the MIL problem such as [1] the proposed heuristic is very fast. This, combined with the fact that we do not need fine-tuning, permits a flexible on-the-fly learning of new category in a few minutes.

Figure 1 illustrates the situation we face at training time. For each image, we are given a set of bounding boxes which receive a label +1 (the visual category of interest is present at least once) or −1 (the category is not present in this image).

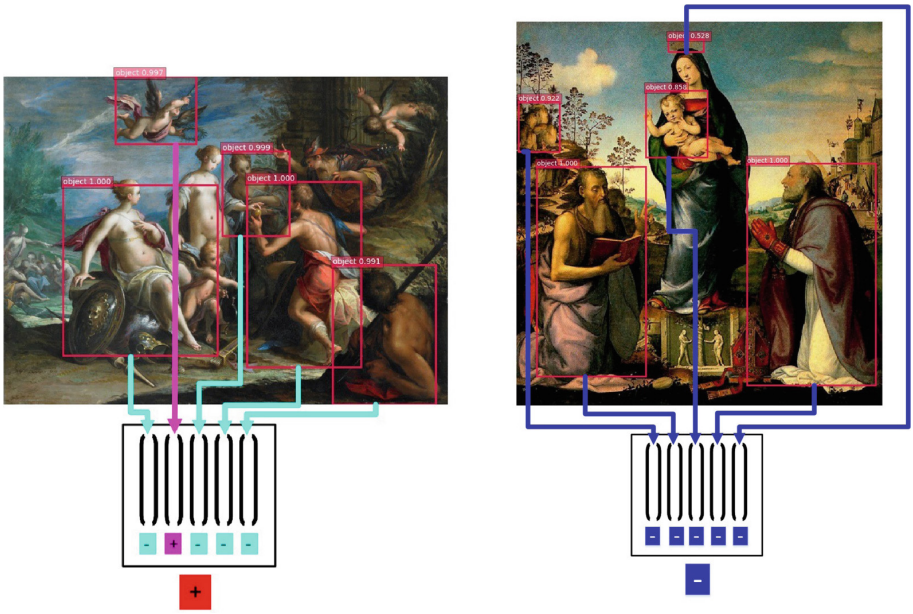

Fig. 1. Illustration of positive and negative sets of detections (bounding boxes) for the *angel* category.

3.1 Multiple Instance Learning

The usual way to perform MIL is through the resolution of a non-convex energy minimisation [1], although efficient convex relaxations have been proposed [29]. One disadvantage of these approaches is their heavy computational cost. In what follows, we propose a simple and fast heuristic to this problem.

For simplicity of the presentation, we assume only one visual category. Assume we have N images at hand, each of which contains K bounding boxes. Each image receives a label $y = +1$ when it is a positive example (the category is present) and $y = -1$ otherwise. We denote by n_1 the number of positive examples in the training set, and by n_{-1} the number of negative examples.

Images are indexed by i, the K regions provided by the object detector are indexed by k, the label of the i-th image is denoted by y_i and the high level semantic feature vector of size M associated to the k-th box in the i-th image is denoted $X_{i,k}$. We also assume that the detector provides a (class agnostic) "objectness" score for this box, denoted $s_{i,k}$.

We make the (strong) hypothesis that if $y_i = +1$, then there is at least one of the K regions in image i that contains an occurrence of the category. In a sense,

we assume that the region proposal part is robust enough to transfer detections from photography to the target domain.

Following this assumption, our problem boils down to the classic multiple-instance classification problem [13]: if for image i we have $y_i = +1$, then at least one of the boxes contains the category, whereas if $y_i = -1$ no box does. The goal is then to decide which boxes correspond to the category. Instead of the classical SVM generalisation proposed in [1] and based on an iterative procedure, we look for an hyperplan minimising the functional defined below. We look for $w \in \mathbf{R}^M$, $b \in \mathbf{R}$ achieving

$$min_{(w,b)}\mathcal{L}(w,b) \tag{1}$$

with

$$\phi(w,b) = \sum_{i=1}^{N} \frac{-y_i}{n_{y_i}} Tanh \left\{ \max_{k\in\{1..K\}} \left(w^T X_{i,k} + b\right) \right\} \tag{2}$$

and

$$\mathcal{L}(w,b) = \phi(w,b) + C * ||w||^2, \tag{3}$$

where C is a constant balancing the regularisation term. The intuition behind this formulation is that minimising $\mathcal{L}(w,b)$ amounts to seek a hyperplan separating the most positive element of each positive image from the least negative element of the negative image, sharing similar ideas as in MI-SVM [1] or Latent-SVM [18]. The $Tanh$ is here to mimic the SVM formulation in which only the worst margins count. We divide by n_{y_i} to account for unbalanced data. Indeed most example images are negative ones ($n_{-1} \gg n_1$)).

The main advantage of this formulation is that it can be realised by a simple gradient descent, therefore avoiding costly multiple SVM optimisation. If the dataset is too big to fit in the memory, we switch to a stochastic gradient descent by considering random batches in the training set.

As this problem is non-convex, we try several random initialisation and we select the couple w, b minimising the classification function $\phi(w,b)$. Although we did not explore this possibility it may be interesting to keep more than one vector to describe a class, since one iconographic element could have more that one specific feature, each stemming from a distinctive part.

In practice, we observed consistently better results when modifying slightly the above formulation by considering the (class-agnostic) "objectness" score associated to each box (as returned by Faster R-CNN). Therefore we modify function ϕ to

$$\phi^s(w,b) = \sum_{i=1}^{N} \frac{-y_i}{n_{y_i}} Tanh \left\{ \max_{k\in\{1..K\}} \left((s_{i,k} + \epsilon)\left(w^T X_{i,k} + b\right)\right) \right\} \tag{4}$$

with $\epsilon \geq 0$. The motivation behind this formulation is that the score $s_{i,k}$, roughly a probability that there is an object (of any category) in box k, provides a prioritisation between boxes.

Once the best couple $(w^\star, b^\star)$ has been found, we compute the following score, reflecting the meaningfulness of category association:

$$S(x) = Tanh\{(s(x) + \epsilon)\left(w^{\star T} x + b^\star\right)\} \tag{5}$$

At test time, each box with a positive score (5) (where $s(x)$ is the objectness score associated to x) is affected to the category. The approach is then straightforwardly extended to an arbitrary number of categories, by computing a couple $(w^\star, b^\star)$ per category. Observe, however, that this leads to non-comparable scores between categories. Among all boxes affected to each class, a non-maximal suppression (NMS) algorithm is then applied in order to avoid redundant detections. The resulting multiple instance learning method is called **MI-max**.

3.2 Implementation Details

Faster R-CNN. We use the detection network Faster R-CNN [42]. We only keep its region proposal part (RPN) and the features corresponding to each proposed region. In order to yield and efficient and flexible learning of new classes, we choose to avoid retraining or even fine-tuning. Faster R-CNN is a meta-network in which a pre-trained network is enclosed. The quality of features depends on the enclosed network and we compare several possibility in the experimental part.

Images are resized to 600 by 1000 before applying Faster R-CNN. We only keep the 300 boxes having best "objectness" scores (after a NMS phase), along with their high-level features[1]. An example of extracted boxes is shown in Fig. 2. About 5 images per second can be obtained on a standard GPU. This part can be performed offline since we don't fine-tune the network.

As mentioned in [30], residual network (ResNet) appears to be the best architecture for transfer learning by feature extractions among the different ImageNet models, and we therefore choose these networks for our Faster R-CNN versions. One of them (denoted RES-101-VOC07) is a 101 layers ResNet trained for the detection task on PASCAL VOC2007. The other one (denoted RES-152-COCO) is a 152 layers ResNet trained on MS COCO [34]. We will also compare our approach to the plain application of these networks for the detection tasks when possible, that is when they were trained on classes we want to detect. We refer to these approaches as FSD (fully supervised detection) in our experiments.

For implementation, we build on the Tensorflow[2] implementation of Faster R-CNN of Chen et al. [7][3].

MI-max. When a new class is to be learned, the user provides a set of weakly annotated images. The MI-max framework described above is then run to find a linear separator specific to the class. Note that both the database and the library of classifiers can be enriched very easily. Indeed, adding an image to the database only requires running it through the Faster R-CNN network and adding a new class only requires a MIL training.

For training the MI-max, we use a batch size of 1000 examples (for smaller sets, all features are loaded into the GPU), 300 iterations of gradient descent

[1] The layer $fc7$ of size $M = 2048$ in the ResNet case, often called 2048-D.
[2] https://www.tensorflow.org/.
[3] Code can be found on GitHub https://github.com/endernewton/tf-faster-rcnn.

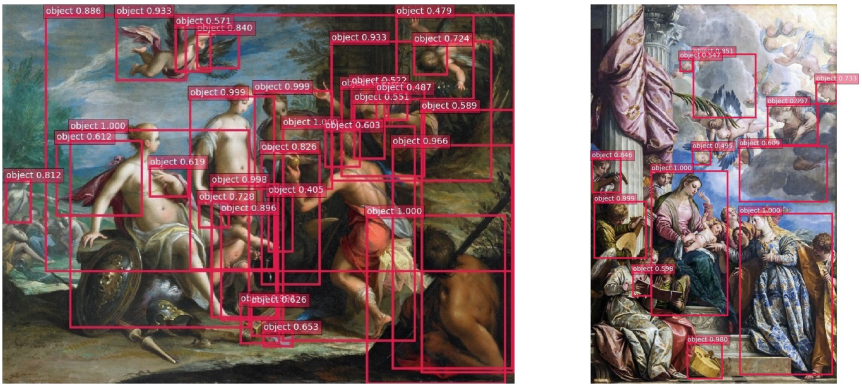

Fig. 2. Some of the regions of interest generated by the region proposal part (RPN) of Faster R-CNN.

with a learning rate of 0.01 and $\epsilon = 0.01$ (4). The whole process takes 750 s for 20 classes on PASCAL VOC07 trainval (5011 images) with 12 random start points per class, on a consumer GPU (GTX 1080Ti). Actually the random restarts are performed in parallel to take advantage of the presence of the features in the GPU memory since the transfer of data from central RAM to the GPU memory is a bottleneck for our method. The 20 classes can be learned in parallel.

For the experiments of Sect. 4.3, we also perform a grid search on the hyper-parameter C (3) by splitting the training set into training and validation sets. We learn several couples (w, b) for each possible value of C (different initialisation) and the one that minimises the loss (4) for each class is selected.

4 Experiments

In this section, we perform weakly supervised detection experiments on different databases, in order to illustrate different assets of our approach.

In all cases, and besides other comparisons, we compare our approach (MI-max) to the following baseline, which is actually the approach chosen for the detection experiments in [11] (except that we do not perform box expansion): the idea is to consider that the region with the best "objectness" score is the region corresponding to the label associated to the image (positive or negative). This baseline will be denoted as MAX. Linear-SVM classifier are learnt using those features per class in a one-vs-the-rest manner. The weight parameter that produces the highest AP (Average Precision) score is selected for each class by a cross validation method[4] and then a classifier is retrained with the best hyper-parameter on all the training data per class. This baseline requires to train several SVMs and is therefore costly.

[4] We use a 3-fold cross validation while [11] use constant training and validation set.

At test time, the labels and the bounding boxes are used to evaluate the performance of the methods in term of AP par class. The generated boxes are filtered by a NMS with an Intersection over Union (IoU) [17] threshold of 0.3 and a confidence threshold of 0.05 for all methods.

4.1 Experiments on PASCAL VOC

Before proceeding with the transfer learning and testing our method on paintings, we start with a sanity check experiment on PASCAL VOC2007 [17]. We compare our weakly supervised approach, MI-max, to the plain application of the fully supervised Faster R-CNN [42] and to the weakly supervised MAX procedure recalled above. We perform the comparison using two different architectures (for the three methods), RES-101-VOC07 and RES-512-COCO, as explained in the previous section.

Table 1. VOC 2007 test Average precision (%) Comparison of the Faster R-CNN detector (trained in a fully supervised manner: FSD) and our MI-max algorithm (trained in a weakly supervised manner) for two networks RES-101-VOC07 and RES-152-COCO.

Net	Method	aero	bicy	bird	boa	bot	bus	car	cat	cha	cow	dtab	dog	hors	mbik	pers	plnt	she	sofa	trai	tv	mean
RES-101-VOC07	FSD [26]	73.6	82.3	75.4	64.0	57.4	80.2	86.5	86.2	52.7	85.2	66.9	87.0	87.1	82.9	81.2	45.7	76.8	71.2	82.6	75.5	75.0
	MAX	20.8	47.0	26.1	20.2	8.3	41.1	44.9	60.1	31.7	54.8	46.4	42.9	62.2	58.7	20.9	21.6	37.6	16.7	42.0	19.8	36.2
	MI-max[a]	63.5	78.4	68.5	54.0	50.7	71.8	85.6	77.1	52.7	80.0	60.1	78.3	80.5	73.5	74.7	37.4	71.2	65.2	75.7	67.7	68.3 ± 0.2
RES-152-COCO	FSD [26]	91.0	90.4	88.3	61.2	77.7	92.2	82.2	93.2	67.0	89.4	65.8	88.0	92.0	89.5	88.5	56.9	85.1	81.0	89.8	85.2	82.7
	MAX [11]	58.8	64.7	52.4	8.6	20.8	55.2	66.8	76.1	19.4	66.3	6.7	59.7	56.4	43.3	15.5	18.3	80.3	7.6	71.8	32.6	44.1
	MI-max[a]	88.0	90.2	84.3	66.0	78.7	93.8	92.7	90.7	63.7	78.8	61.5	88.4	90.9	88.8	87.9	56.8	75.5	81.3	88.4	86.1	81.6 ± 0.3

[a]It is the average performance on 100 runs of our algorithm.

As shown in Table 1 our weakly supervised approach (only considering annotations at the image level[5]) yields performances that are only slightly below the ones of the fully supervised approach (using instance-level annotations). On the average, the loss is only 1.1% of mAP when using RES-512-COCO (for both methods). The baseline MAX procedure (used for transfer learning on paintings in [10]) yields notably inferior performances.

4.2 Detection Evaluation on Watercolor2k and People-Art Databases

We compare our approach with two recent methods performing object detection in artworks, one in a fully supervised way [52] for detecting people, the other using a (partly) weakly supervised method to detect several VOC classes on watercolor images [28]. For the learning stage, the first approach uses instance-level annotations on paintings, while the second one uses instance-level annotations on photographs and image-level annotations on paintings. In both cases, it

[5] However, observe that since we are relying on Faster R-CNN, our system uses a subpart trained using class agnostic bounding boxes.

is shown that using image-level annotations only (our approach, MI-max) only yields a light loss of performances.

Experiment 1: Watercolor2k. This database, introduced in [28], and available online[6], is a subset of watercolor artworks from the **BAM!** database [53] with instance-level annotations for 6 classes (bike, bird, dog, cat, car, person) that are included in the PASCAL VOC, in order to study cross-domain transfer learning. On this database, we compare our approach to the methods from [28] and from [4], to the baseline MAX discussed above, as well as to the classical MIL approach MI-SVM [1] (using a maximum of 50 iterations and no restarts).

In [28], a style transfer transformation (Cycle-GAN [56]) is applied to natural images with instance-level annotation. The images are transferred to the new modality (i.e. watercolor) in order to fine-tune a detector pre-trained on natural images. This detector is used to predict localisation of objects on watercolor images annotated at the image level. The detector is then fine-tuned on those images in a fully supervised manner. Bilen and Vedaldi [4] proposed a Weakly Supervised Deep Detection Network (WSDDN), which consists in transforming a pre-trained network by replacing its classification part by a two streams network (a region proposal stream and a classification one) combined with a weighted MIL pooling strategy.

Table 2. Watercolor2k (test set) Average precision (%). Comparison of the proposed MI-max method to alternative approaches.

Net	Method	bike	bird	car	cat	dog	person	Mean
VGG	WSDDN [4][a]	1.5	26.0	14.6	0.4	0.5	33.3	12.7
SSD	DT+PL [28][a]	76.5	54.9	46.0	37.4	38.5	72.3	54.3
RES-152-COCO	MAX [11]	74.0	34.5	26.8	17.8	21.5	21.0	32.6
	MI-SVM [1]	66.8	23.5	6.7	13.0	8.4	14.1	22.1
	MI-max [our][b]	85.2	48.2	49.2	31.0	30.0	57.0	50.1 ± 1.1

[a]The performance come from the original paper [28].
[b]Standard deviation computed on 100 runs of the algorithm.

From Table 2, one can see that our approach performs clearly better than the other ones using image-level annotations only ([4], MAX, MI-SVM). We also observe only a minor degradation of average performances (54.3% versus 48.9%) with respect to the method [28], which is retrained using style transfer and instance-level annotations on photographs.

Experiment 2: People-Art. This database, introduced in [52], is made of artistic images and bounding boxes for the single class *person*. This database is particularly challenging because of its high variability in styles and depiction techniques. The method introduced in [52] yields excellent detection performances on this database, but necessitates instance-level annotations for training.

[6] https://github.com/naoto0804/cross-domain-detection.

The authors rely on Fast R-CNN [24], of which they only keep the three first layers, before re-training the remaining of the network using manual location annotations on their database.

In Table 3, one can see that our approach MI-max yields detection results that are very close to the fully supervised results from [52], despite a much lighter training procedure. In particular, as already explained, our procedure can be trained directly on large, globally annotated database, for which manually entering instance-level annotations is tedious and time-costly.

Table 3. People-Art (test set) Average precision (%). Comparison of the proposed MI-max method to alternative approaches.

Net	Method	Person
Fast R-CNN (VGG16)	Fine tuned [52][a]	59
RES-152-COCO	MAX [11]	25.9
	MI-SVM [1]	13.3
RES-152-COCO	MI-max [our]	55.4 ± 0.7

[a]The performance come from the original paper.

4.3 Detection on IconArt Database

In this last experimental section, we investigate the ability of our approach to learn and detect new classes that are specific to the analysis of artworks, some of which cannot be learnt on photographs. Typical such examples include iconic characters in certain situations, such as Child Jesus, the crucifixion of Jesus, Saint Sebastian, etc. Although there has been a recent effort to increase open-access databases of artworks by academia and/or museums workforce [10,16, 31,36–38,44,48], they usually don't include systematic and reliable keywords. One exception is the database from the Rijkmuseum, with labels based on the IconClass classification system [27], but this database is mostly composed of prints, photographs and drawings. Moreover, these databases don't include the localisation of objects or characters.

In order to study the ability of our (and other) systems to detect iconographic elements, we gathered 5955 painting images from Wikicommons[7], ranging from the 11th to the 20th century, which are partially annotated by the Wikidata[8] contributors. We manually checked and completed image-level annotations for 7 classes. The dataset is split in training and test sets, as shown in Table 4. For a subset of the test set, and only for the purpose of performance evaluation, instance-level annotations have been added. The resulting database is called

[7] https://commons.wikimedia.org/wiki/Main_Page.
[8] https://www.wikidata.org/wiki/Wikidata:Main_Page.

IconArt[9]. Example images are shown in Fig. 3. To the best of our knowledge, the presented experiments are the first investigating the ability of modern detection tools to classify and detect such iconographic elements in paintings. Moreover, we investigate this aspect in a weakly supervised manner.

Table 4. Statistics of the IconArt database

Class	Angel	Child Jesus	Crucifixion	Mary	nudity	ruins	Saint Sebastian	None	Total
Train	600	755	86	1065	956	234	75	947	2978
Test for classification	627	750	107	1086	1007	264	82	924	2977
Test for detection	261	313	107	446	403	114	82	623	1480
Number of instances	1043	320	109	502	759	194	82		3009

Fig. 3. Example images from the IconArt database. Angel on the first line, Saint Sebastian on the second. We can see some of the challenges posed by this database: tiny objects, occlusions and large pose variability.

To fix ideas on the difficulty of dealing with iconographic elements, we start with a classification experiment. For this, we use the same classification approach as in [10], using InceptionResNetv2 [49] as a feature extractor[10]. We also perform classification-by-detection experiments, using the previously described MAX approach (as in [11]) and our approach, MI-max. In both cases, for each class, the score at the image level is the highest confidence detection score for this class on all the regions of the image. Results are displayed in Table 5. First, we observe that classification results are very variable depending on the class.

[9] The database is available online https://wsoda.telecom-paristech.fr/downloads/dataset/IconArt_v1.zip.

[10] Only the center of the image is provided to the network and extracted features are 1536-D.

Classes such as Jesus Child, Mary or crucifixion have relatively high classification scores. Others, such as Saint Sebastian, are only scarcely classified, probably due to a limited quantity of examples and a high variability of poses, scales and depiction styles. We can also observe that, as mentioned in [11], the classification by detection can provide better scores than global classification, possibly because of small objects, such as angels in our case. Observe that these classification scores can probably be increased using multi-scale learning (as in [51]), augmentation schemes and an ensemble of networks [11].

Table 5. IconArt classification test set *classification* average precision (%).

Net	Method	angel	JChild	crucifixion	Mary	nudity	ruins	StSeb	Mean
InceptionResNetv2 [49]		44.1	77.2	57.8	81.1	77.4	74.6	26.8	62.7
RES-152-COCO	MAX [11]	49.3	74.7	30.3	67.5	57.4	43.2	7.0	47.1
	MI-max [our]	57.4	60.7	79.9	70.4	65.3	45.9	17.0	56.7 ± 1.0
	MI-max-C [our]	61.0	68.9	80.2	71.4	66.3	51.7	14.8	59.2 ± 1.2

Next, we evaluate the detection performance of our method, first with a restrictive metric: AP per class with an IoU $\geqslant 0.5$ (as in all previous detection experiments in this paper), then with a less restrictive metric with IoU $\geqslant 0.1$. Results are displayed in Table 6. Results on this very demanding experiment are a mixed-bag. Some classes, such as crucifixion, and to a less extend nudity or Jesus Child are correctly detected. Others, such as angel, ruins or Saint Sebastian, hardly get it up to 15% detection scores, even when using the relaxed criterion IoU $\geqslant 0.1$. Beyond a relatively small number of examples and very strong scale and pose variations, there are further reasons for this:

- The high in-class depiction variability (for Saint Sebastian for instance)
- The many occlusions between several instances of a same class (angel)
- The fact that some parts of an object can be more discriminative than the whole object (nudity).

Illustrations of successes and failures are displayed, respectively on Figs. 4 and 5. On the negative examples, one can see that often a larger region than

Table 6. IconArt detection test set *detection* average precision (%). All methods based on RES-152-COCO.

Method	Metric	angel	JChild	crucifixion	Mary	nudity	ruins	StSeb	Mean
MAX [11]	AP IoU $\geqslant 0.5$	1.4	3.9	7.4	2.8	3.9	0.3	0.9	2.9
	AP IoU $\geqslant 0.1$	10.1	36.2	28.2	18.4	14.0	1.6	2.8	15.9
MI-max [our]	AP IoU $\geqslant 0.5$	0.3	0.9	37.3	3.8	21.2	0.5	10.9	10.7 ± 1.7
	AP IoU $\geqslant 0.1$	6.4	25.3	74.4	44.6	30.9	6.8	17.2	29.4 ± 1.7
MI-max-C [our]	AP IoU $\geqslant 0.5$	3.0	17.7	32.6	4.8	23.5	1.1	9.6	13.2 ± 3.1
	AP IoU $\geqslant 0.1$	12.3	41.2	74.4	46.3	31.2	13.6	16.1	33.6 ± 2.2

the element of interest is selected or that a whole group of instances is selected instead of a single one. Future work could focus on the use of several couples (w, b) instead of one to prevent those problems.

Fig. 4. Successful examples using our MI-max-C detection scheme. We only show boxes whose scores are over 0.75.

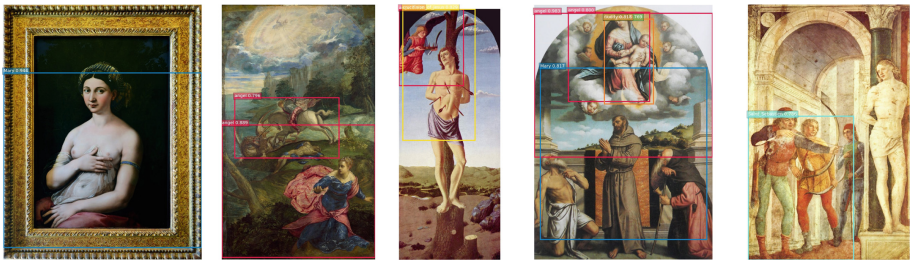

Fig. 5. Failure examples using our MI-max-C detection scheme. We only show boxes whose scores are over 0.75.

5 Conclusion

Results from this paper confirm that transfer learning is of great interest to analyse artworks databases. This was previously shown for classification and fully supervised detection schemes, and was here investigated in the case of weakly supervised detection. We believe that this framework is particularly suited to develop tools helping art historians, because it avoids tedious annotations and opens the way to learning on large datasets. We also show, in this context, experiments dealing with iconographic elements that are specific to Art History and cannot be learnt on natural images.

In future works, we plan to use localisation refinement methods, to further study how to avoid poor local optima in the optimisation procedure, to add

contextual information for little objects, and possibly to fine-tune the network (as in [15]) to learn better features on artworks. Another exciting direction is to investigate the potential of weakly supervised learning on large databases with image-level annotations, such as the ones from the Rijkmuseum [44] or the French Museum consortium [43].

Acknowledgements. This work is supported by the "IDI 2017" project funded by the IDEX Paris-Saclay, ANR-11-IDEX-0003-02.

References

1. Andrews, S., Tsochantaridis, I., Hofmann, T.: Support vector machines for multiple-instance learning. In: Advances in Neural Information Processing Systems, pp. 577–584 (2003)
2. Aubry, M., Russell, B.C., Sivic, J.: Painting-to-3D model alignment via discriminative visual elements. ACM Trans. Graph. (ToG) **33**(2), 14 (2014)
3. Bianco, S., Mazzini, D., Schettini, R.: Deep multibranch neural network for painting categorization. In: Battiato, S., Gallo, G., Schettini, R., Stanco, F. (eds.) ICIAP 2017. LNCS, vol. 10484, pp. 414–423. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68560-1_37
4. Bilen, H., Vedaldi, A.: Weakly supervised deep detection networks. In: IEEE Conference on Computer Vision and Pattern Recognition (2016)
5. de Bosio, S.: Master and judge: the mirror as dialogical device in Italian renaissance art theory. In: Zimmermann, M. (ed.) Dialogical Imaginations: Debating Aisthesis as Social Perception. Diaphanes (2017)
6. Carbonneau, M.A., Cheplygina, V., Granger, E., Gagnon, G.: Multiple instance learning: a survey of problem characteristics and applications. Pattern Recogn. **77**, 329–353 (2016). https://doi.org/10.1016/j.patcog.2017.10.009
7. Chen, X., Gupta, A.: An implementation of faster RCNN with study for region sampling. arXiv:1702.02138 [cs], February 2017
8. Cinbis, R.G., Verbeek, J., Schmid, C.: Weakly supervised object localization with multi-fold multiple instance learning. IEEE Trans. Pattern Anal. Mach. Intell. **39**(1), 189–203 (2016). https://doi.org/10.1109/TPAMI.2016.2535231
9. Crowley, E., Zisserman, A.: The state of the art: object retrieval in paintings using discriminative regions. In: BMVC (2014)
10. Crowley, E.J., Zisserman, A.: In search of art. In: Agapito, L., Bronstein, M.M., Rother, C. (eds.) ECCV 2014. LNCS, vol. 8925, pp. 54–70. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-16178-5_4
11. Crowley, E.J., Zisserman, A.: The art of detection. In: Hua, G., Jégou, H. (eds.) ECCV 2016. LNCS, vol. 9913, pp. 721–737. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46604-0_50
12. Del Bimbo, A., Pala, P.: Visual image retrieval by elastic matching of user sketches. IEEE Trans. Pattern Anal. Mach. Intell. **19**(2), 121–132 (1997)
13. Dietterich, T.G., Lathrop, R.H., Lozano-Pérez, T.: Solving the multiple instance problem with axis-parallel rectangles. Artif. Intell. **89**(1–2), 31–71 (1997)
14. Donahue, J., et al.: DeCAF: a deep convolutional activation feature for generic visual recognition. In: Xing, E.P., Jebara, T. (eds.) Proceedings of the 31st International Conference on Machine Learning. Proceedings of Machine Learning Research, PMLR, Bejing, China, vol. 32, pp. 647–655, 22–24 June 2014. http://proceedings.mlr.press/v32/donahue14.html

15. Durand, T., Mordan, T., Thome, N., Cord, M.: WILDCAT: weakly supervised learning of deep ConvNets for image classification, pointwise localization and segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017). IEEE, Honolulu, July 2017
16. Europeana: collections Europeana (2018). https://www.europeana.eu/portal/en
17. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL visual object classes challenge 2007 (VOC2007) results (2007). http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html
18. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Trans. Pattern Anal. Mach. Intell. **32**(9), 1627–1645 (2010)
19. Florea, C., Badea, M., Florea, L., Vertan, C.: Domain transfer for delving into deep networks capacity to de-abstract art. In: Sharma, P., Bianchi, F.M. (eds.) SCIA 2017. LNCS, vol. 10269, pp. 337–349. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-59126-1_28
20. Gasparro, D.: Dal lato dell'immagine: destra e sinistra nelle descrizioni di Bellori e altri. Ed. Belvedere (2008)
21. Gehler, P.V., Chapelle, O.: Deterministic annealing for multiple-instance learning. In: Artificial Intelligence and Statistics, pp. 123–130 (2007)
22. Ginosar, S., Haas, D., Brown, T., Malik, J.: Detecting people in cubist art. In: Agapito, L., Bronstein, M.M., Rother, C. (eds.) ECCV 2014. LNCS, vol. 8925, pp. 101–116. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-16178-5_7
23. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587, June 2014. https://doi.org/10.1109/CVPR.2014.81
24. Girshick, R.: Fast R-CNN. In: International Conference on Computer Vision (ICCV) (2015)
25. Hall, P., Cai, H., Wu, Q., Corradi, T.: Cross-depiction problem: recognition and synthesis of photographs and artwork. Comput. Vis. Media **1**(2), 91–103 (2015)
26. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
27. Iconclass: Home—Iconclass (2018). http://www.iconclass.nl/home
28. Inoue, N., Furuta, R., Yamasaki, T., Aizawa, K.: Cross-domain weakly-supervised object detection through progressive domain adaptation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018). IEEE (2018)
29. Joulin, A., Bach, F.: A convex relaxation for weakly supervised classifiers. arXiv preprint arXiv:1206.6413 (2012)
30. Kornblith, S., Shlens, J., Le, Q.V.: Do better ImageNet models transfer better? arXiv:1805.08974 [cs, stat], May 2018
31. Lecoutre, A., Negrevergne, B., Yger, F.: Recognizing art style automatically in painting with deep learning. In: ACML, pp. 1–17 (2017)
32. Li, D., Yang, Y., Song, Y.Z., Hospedales, T.M.: Deeper, broader and artier domain generalization. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 5543–5551, October 2017. https://doi.org/10.1109/ICCV.2017.591
33. Li, D., Huang, J.B., Li, Y., Wang, S., Yang, M.H.: Weakly supervised object localization with progressive domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3512–3520 (2016)

34. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
35. Liu, W., et al.: SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
36. Mao, H., Cheung, M., She, J.: DeepArt: learning joint representations of visual arts. In: Proceedings of the 2017 ACM on Multimedia Conference, pp. 1183–1191. ACM Press (2017). https://doi.org/10.1145/3123266.3123405
37. Mensink, T., Van Gemert, J.: The Rijksmuseum challenge: museum-centered visual recognition. In: Proceedings of International Conference on Multimedia Retrieval, p. 451. ACM (2014)
38. MET: image and data resources — the metropolitan museum of art (2018). https://www.metmuseum.org/about-the-met/policies-and-documents/image-resources
39. Pharos consortium: PHAROS: the international consortium of photo archives (2018). http://pharosartresearch.org/
40. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788 (2016)
41. Redmon, J., Farhadi, A.: YOLO9000: better, faster, stronger. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017). IEEE (2017)
42. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 28, pp. 91–99. Curran Associates, Inc. (2015). http://papers.nips.cc/paper/5638-faster-r-cnn-towards-real-time-object-detection-with-region-proposal-networks.pdf
43. Réunion des Musées Nationaux-Grand Palais: Images d'Art (2018). https://art.rmngp.fr/en
44. Rijksmuseum: online collection catalogue - research (2018). https://www.rijksmuseum.nl/en/research/online-collection-catalogue
45. Seguin, B., Striolo, C., diLenardo, I., Kaplan, F.: Visual link retrieval in a database of paintings. In: Hua, G., Jégou, H. (eds.) ECCV 2016. LNCS, vol. 9913, pp. 753–767. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46604-0_52
46. Shrivastava, A., Malisiewicz, T., Gupta, A., Efros, A.A.: Data-driven visual similarity for cross-domain image matching. ACM Trans. Graph. (ToG) **30**(6), 154 (2011)
47. Song, H.O., Girshick, R., Jegelka, S., Mairal, J., Harchaoui, Z., Darrell, T.: On learning to localize objects with minimal supervision. In: Xing, E.P., Jebara, T. (eds.) Proceedings of the 31st International Conference on Machine Learning. Proceedings of Machine Learning Research, PMLR, Bejing, China, pp. 1611–1619, No. 2, 22–24 June 2014, http://proceedings.mlr.press/v32/songb14.html
48. Strezoski, G., Worring, M.: OmniArt: multi-task deep learning for artistic data analysis. arXiv:1708.00684 [cs], August 2017
49. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: AAAI, p. 4 (2017)
50. Tang, P., Wang, X., Bai, X., Liu, W.: Multiple instance detection network with online instance classifier refinement. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3059–3067 (2017)

51. van Noord, N., Postma, E.: Learning scale-variant and scale-invariant features for deep image classification. Pattern Recogn. **61**, 583–592 (2017). https://doi.org/10.1016/j.patcog.2016.06.005
52. Westlake, N., Cai, H., Hall, P.: Detecting people in artwork with CNNs. In: ECCV Workshops (2016)
53. Wilber, M.J., Fang, C., Jin, H., Hertzmann, A., Collomosse, J., Belongie, S.: BAM! The behance artistic media dataset for recognition beyond photography. In: IEEE International Conference on Computer Vision (ICCV). IEEE (2017)
54. Wu, Q., Cai, H., Hall, P.: Learning graphs to model visual objects across different depictive styles. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8695, pp. 313–328. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10584-0_21
55. Yin, R., Monson, E., Honig, E., Daubechies, I., Maggioni, M.: Object recognition in art drawings: transfer of a neural network. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2299–2303. IEEE (2016)
56. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: 2017 IEEE International Conference on Computer Vision (ICCV) (2017)