This WACV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

No Annotations for Object Detection in Art through Stable Diffusion

Patrick Ramos[1] Nicolas Gonthier[2] Selina Khan[3] Yuta Nakashima[1] Noa Garcia[1]
[1]Osaka University [2]Univ Gustave Eiffel, ENSG, IGN, LASTIG, France [3]University of Amsterdam
{patrickramos@is., n-yuta@, noagarcia@}ids.osaka-u.ac.jp,
nicolas.gonthier@ign.fr, selinajasmin@gmail.com

Abstract

Object detection in art is a valuable tool for the digital humanities, as it allows for faster identification of objects in artistic and historical images compared to humans. However, annotating such images poses significant challenges due to the need for specialized domain expertise. We present NADA (no annotations for detection in art), a pipeline that leverages diffusion models' art-related knowledge for object detection in paintings without the need for full bounding box supervision. Our method, which supports both weakly-supervised and zero-shot scenarios and does not require any fine-tuning of its pretrained components, consists of a class proposer based on large vision-language models and a class-conditioned detector based on Stable Diffusion. NADA is evaluated on two artwork datasets, ArtDL 2.0 and IconArt, outperforming prior work in weakly-supervised detection, while being the first work for zero-shot object detection in art. Code is available at
https://github.com/patrick-john-ramos/nada

1. Introduction

Performance in object detection in paintings, which has applications such as art captioning [2,8,33], art visual question answering [13], art visual pattern discovery [51], musicological studies [22], or art exploration [36], lags behind traditional object detection in photographs [7, 27, 31, 45]. While traditional object detection enjoys success from large-scale annotated datasets such as MS-COCO [29] and OpenImages [28], these datasets are comprised mostly of natural images, limiting their use to other domains, *e.g.* art images. Paintings may contain objects that might not be of interest to standard detectors and usually portray them in a different style, documented as the cross-depiction problem [6, 17]. This domain gap can be addressed by training on datasets predominantly, if not completely, composed of non-natural images; however, annotating art images for object detection (*i.e.* with bounding box annotations) is time-consuming and requires domain expertise. For example, in Christian iconography, an annotator must be able to distinguish between St. Francis and St. Dominic. While these classes can be distinguished by their associated symbols as described in online iconography databases such as Iconclass[1], a deep familiarity with these relationships is still needed to annotate efficiently. As a result, existing object detection datasets in art [15,37,47,59,60] are much smaller than object detection datasets in natural images or only contain image-level annotations for training.

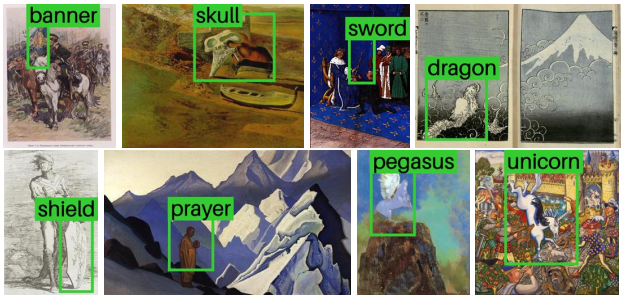


Figure 1. Art object detection *in the wild* with NADA's class-conditioned detector.

In response to these limitations, various methods have been proposed to minimize the supervision required for object detection in paintings, bypassing the need for fully annotated bounding boxes around objects of interest. A first step towards reducing supervised data is to tackle the task as *weakly-supervised* [15, 37], where object detectors are trained using only image-level labels rather than detailed object bounding boxes. Additionally, reducing annotations can be taken a step further with a *zero-shot* setting, where no annotations (neither bounding boxes nor class labels) are used. Due to the challenging nature of the zero-shot approach, it has not yet been explored in the art domain.

We address this gap by introducing NADA (no annotations for detection in art), an application for object detection in paintings that reduces the need for supervision and detects objects in both weakly-supervised and

[1] https://iconclass.org/

zero-shot settings. NADA, which leverages the inherent knowledge of art in computer vision models trained on vast amounts of data, consists of two modules: a *class proposer*, which, given an image of a painting and a list of potential classes, predicts the objects present in the image; and a *class-conditioned detector*, which locates the objects in the painting based on the predicted classes. The class proposer can be adapted according to the desired level of supervision. If image-level classes are available, (*i.e*., weakly-supervised setting), a lightweight classifier is trained to classify images from their CLIP [42] embeddings. In contrast, if no annotations are available at all (*i.e*., zero-shot setting), the class proposer relies on a vision-language model to predict the classes present in the image. The predicted classes are used by the class-conditioned detector, which leverages the generative capabilities of diffusion models [19, 40, 43, 48], particularly Stable Diffusion [48], to operate independently of the level of supervision. The classes are used to create an input prompt for regenerating the original image with the diffusion model. Given that diffusion models are trained on a large number of art images [49] (meaning they may be familiar with objects of interest to paintings) and have been shown to contain knowledge useful for style analysis in art [62] and segmenting objects in natural images [34, 58, 63], we extract and segment the cross-attention maps to generate object bounding boxes, effectively detecting the objects within the painting.

NADA is quantitatively evaluated on two art object detection datasets: ArtDL 2.0 [37] and IconArt [15]. In the weakly-supervised setting, NADA outperforms prior work on ArtDL 2.0 and stays competitive with other methods on IconArt. Meanwhile, NADA presents the first results for zero-shot object detection. Our ablation study isolates the influence of the class proposer by evaluating detection when labels are already known, boosting performance on both datasets and showing that the diffusion-based class-conditioned detector localizes objects in paintings effectively, but is reliant on accurate class proposals. Lastly, we showcase the applicability of NADA by detecting uncommon objects in standard object detector datasets, such as dragons or swords, *in the wild*, as shown in Fig. 1.

2. Related work

Object detection in art Localizing and recognizing objects in art presents some unique challenges compared to object detection in natural images [44, 45], primarily due to the interest in objects that are not common in natural images and the cross-depiction problem. While differences in style between natural images and paintings can contribute to the difficulty of object detection in art, previous work [1, 10, 14, 23, 59] have shown that transfer learning and domain adaptation techniques can perform reasonably well in bridging this gap. A survey on this topic is available in [4]. However, an important challenge arises when the classes to be detected, such as mythological creatures like *dragons* and *angels* or historical figures like *Napoleon Bonaparte*, are entirely different from those in natural image datasets, making transfer learning and domain adaptation techniques less effective. This problem is exacerbated by the cost of annotating bounding boxes for such novel classes. One approach is to leverage descriptions to address the knowledge gap [25], however this still requires painting descriptions. To address the difficulty of annotating data, other methods approach art object detection as a weakly supervised task.

In this setting, weakly supervised detectors are trained using only image labels, which indicate the classes present in the image but not their locations. This approach has been extensively studied for natural images, with methods based on end-to-end training of modified object detectors [5, 46, 50, 57]. In the domain of art, Gonthier *et al*. [15] treated the task as a multiple-instance-learning (MIL) problem by training a classifier on top of Faster R-CNN [45] bounding box features and objectness scores. This method was extended in [16] with a multi-layer model to boost the performance at minimal extra cost. Milani *et al*. [37] used pseudo-data by creating bounding boxes from class activation maps (CAMs) extracted from a ResNet-50 [18] fine-tuned on the target domain. At the extreme, [35] proposed a one-shot learning method using a modified co-attention and co-excitation framework [20] and data contextualization. Our approach, NADA, extends prior work which only goes as far as weak-supervision by also proposing a zero-shot method that does not require training on a target dataset.

Locating objects with diffusion models Diffusion models [19] are image generation models consisting of denoising auto-encoders that are often conditioned on text inputs [40, 43, 48]. Despite their main purpose being image generation, diffusion models have been leveraged for image segmentation [24, 34, 58, 61, 63] and object detection [12] in two main ways. The first way consists of generating synthetic training images, extracting attention maps from the diffusion model during generation, and converting these attention maps into pseudo-segmentation masks [34, 58, 61, 63]. The second approach attaches detection or segmentation modules directly to the internal representations of diffusion models [3, 12, 26, 28, 34] by obtaining noise corresponding to the input image through noising or diffusion inversion [54], denoising the noise, and extracting the intermediate representations to predict bounding boxes or segmentation masks using an encoder [3] or decoder head [26, 34], sometimes combined with text features [12, 28].

Of prior work, DiffusionSeg [34] is the most similar to our approach as it reports results on extracting segmentation masks from attention maps obtained from real images using diffusion inversion and without training on synthetic data. NADA differs from that study as we focus instead on object

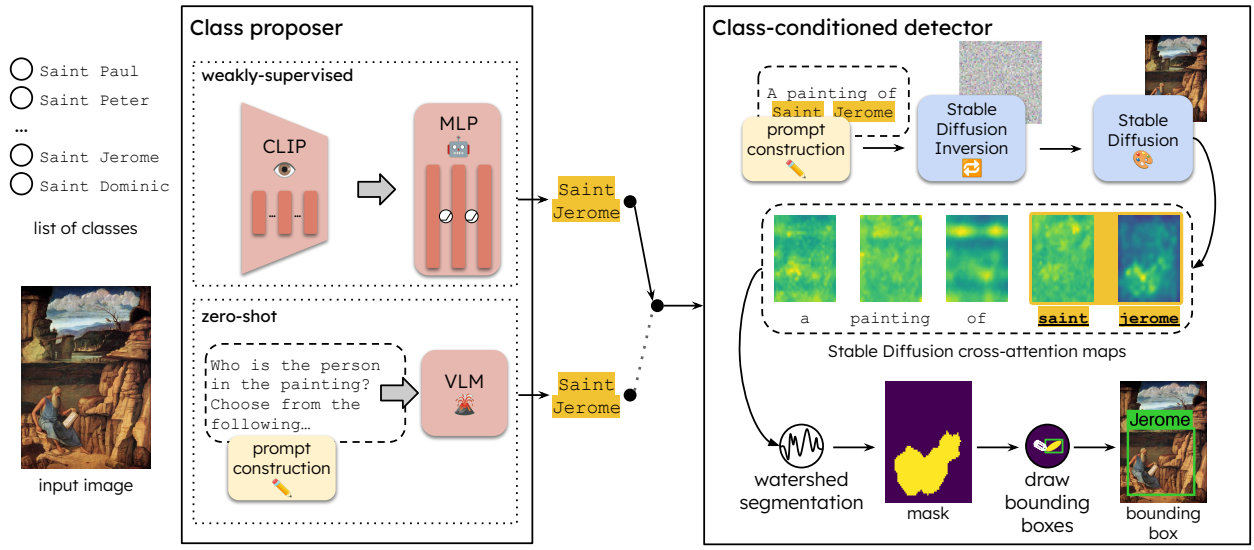


Figure 2. NADA consists of predicting classes from a painting with a class proposer and extracting bounding boxes for the predicted classes with a class-conditioned detector. The class proposer can operate in a weakly-supervised or a zero-shot setting. The class-conditioned detector leverages Stable Diffusion to extract bounding boxes by inverting and regenerating the painting conditioned on an input prompt. The cross-attention maps from the predicted class are aggregated and processed with watershed segmentation to find the bounding box.

detection, propose a simpler method of extracting bounding boxes, and specifically explore the suitability of leveraging Stable Diffusion's knowledge for art images.

3. Method

Figure 2 provides an overview of our method. Given an image i and a set $\mathcal{L} = \{l\}$ of classes l of possible objects[2], NADA predicts a set $\mathcal{B}$ of bounding boxes in i containing objects in $\mathcal{L}$. Following previous work [37], NADA is divided into a *class proposer* to predict a plausible set $\mathcal{L}' \subseteq \mathcal{L}$ of classes from i, and a *class-conditioned detector* to predict $\mathcal{B}$ from i and $\mathcal{L}'$. The class-conditioned detector leverages the art-related knowledge in Stable Diffusion to localize a given object in the image. Our class proposer module identifies $\mathcal{L}'$ without training with bounding box annotations, *i.e.* weakly-supervised or zero-shot.

3.1. Class proposer

We formulate the task of finding the set $\mathcal{L}'$ of class proposals that are likely to appear in i as a classification task. In the weakly-supervised version of our pipeline, we use class labels to train a simple classifier to predict which classes are in i. In the zero-shot version of our method, we task a frozen vision language model (VLM) to predict classes without any training.

Weakly-supervised class proposal (WSCP) We use a frozen CLIP image encoder followed by a multi-layer perceptron (MLP) to classify i. We formulate this as $\mathcal{L}' = \text{MLP}(\text{CLIP}(i))$. We train the MLP either with a single-label classification task using cross-entropy loss or with a multi-label classification task using binary cross-entropy loss. We leverage domain knowledge of the target art datasets to choose which task and loss to train the MLP with.

Zero-shot class proposal (ZSCP) We use a frozen VLM to classify i without any training. Given $\mathcal{L}$, we design a prompt q to ask the VLM to identify all $l \in \mathcal{L}$ in i. Class proposals are given by $\mathcal{L}' = P(\text{VLM}(i, q))$, where P is a simple text post-processing function.

3.2. Class-conditioned detector

This module takes i and each class label $l \in \mathcal{L}'$ to predict a set $\mathcal{B}_l$ of bounding boxes that contain an object of class l. To leverage art knowledge in Stable Diffusion, we obtain cross-attention maps from an input image i by performing a diffusion process (*i.e.*, Stable Diffusion inversion) followed by a reverse diffusion process (*i.e.*, Stable Diffusion), both guided by a prompt p containing the class label l. The reverse diffusion process provides a cross-attention map between each token in p and each patch in i, which identifies which patches in i are associated with l. Letting A_l denote the cross-attention map for l, we apply watershed segmentation [39,53] to A_l to find regions relevant to l. Then, bounding boxes are computed from each of the regions. Specific details for each of these processes are provided below.

Prompt construction Given a label $l \in \mathcal{L}'$, we construct a prompt p that describes the image and contains l. Dur-

[2]We use l to denote both the class label (*e.g.*, `mary`) and its textual representation (*e.g.*, "Mary") interchangeably depending on the context.

ing prompt construction, we modify labels to make them more concrete *e.g.* concretizing the label *nudity*, a state of existence, to the more perceivable *naked person*. We also generalize some labels depending on the scope of the domain, such as generalizing *Child Jesus* to the simpler concept *baby* if it is the only baby among the objects of interest. Note that label generalization is one area of improvement as there are cases where it may confuse classes *e.g.* generalizing *Child Jesus* to *baby* when *Child St. John the Baptist* is also in the data.

Stable Diffusion inversion Our method relies on the cross-attention maps of Stable Diffusion D as it produces the input image i with the prompt p. However, having a model designed to generate synthetic images output existing ones is less straightforward. To allow D to produce i, we first invert the image using null-text inversion [38], which generates noise n from an image-prompt pair that reproduces i when p is fed to D. We denote the inversion process as $n = N_D(i, p)$, where N_D is the inversion function. Note that the null-text inversion process is conducted over several steps, making it time-consuming and another area of improvement.

Stable Diffusion reconstruction With the noise obtained from inversion, we use Stable Diffusion to generate i. The reverse diffusion process is denoted as $i, \{A'_{jk}\}_{jk} = D(n, p)$, which produces the cross attention map A'_{jk} between p and i from the k-th cross-attention block at time step j of the reverse diffusion process ($k = 1, \ldots, K$, $j = 1, \ldots, J$). We discard the reconstructed i and keep only the attention maps $\{A'_{jk}\}_{jk}$.

Extracting image-text cross-attention maps The cross-attention map $A'_k \in \mathbb{R}^{T \times H \times W}$ encompasses the attention weights between each token in p and each patch in i, where T refers to the number of tokens in p, and H and W are the numbers of patches comprising the height and width of the attention map. $A_t \in \mathbb{R}^{H \times W}$ is the average of the cross-attention maps across all layers and time steps of the network corresponding to token $t \in p$, given by:

$$A_t = \frac{1}{JK} \sum_{j,k} A'_{jkt}, \tag{1}$$

where the summation is computed over $k = 1, \ldots, K$ and $j = 1, \ldots, J$, and $A'_{jkt} \in \mathbb{R}^{H \times W}$ is the map that contains the relevance of t to the image patches. As l may consist of multiple tokens (*e.g.*, $l =$ ``john the baptist'' may consist of `john`, `the`, and `baptist`), we again average all maps associated with l to obtain the attention map A_l for l, *i.e.*,

$$A_l = C(\frac{1}{|l|} \sum_{t \in l} A_t), \tag{2}$$

where $|l|$ denotes the number of tokens in l and C is a clamp function that clamps the map's values between 0 and 1.

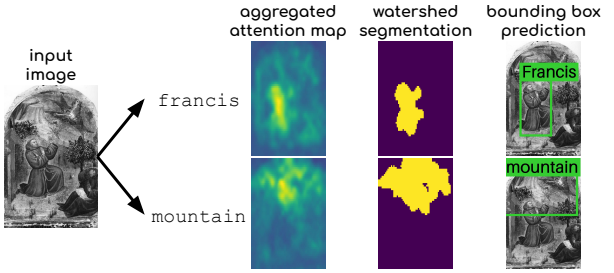


Figure 3. Bounding box extraction from attention maps.

Extracting bounding boxes As demonstrated by prior work [56], the attention map A_l gives a larger value to regions relevant to l. Based on this, we can extract regions by thresholding it. We use Otsu's method [41] to binarize the attention map and segment any region of interest with watershed segmentation [39, 53]. We identify the boxes enclosing the masked regions and take these as the predicted bounding boxes $\mathcal{B}_l$ and group them as $\mathcal{B} = \bigcup_l \mathcal{B}_l$. An example of the process of extracting bounding boxes from an input painting is shown in Fig. 3.

4. Experiments

Evaluation datasets We evaluate NADA on two standard object detection datasets in art: ArtDL 2.0 [37] and IconArt [15]. ArtDL 2.0 contains ten classes of Christian icons taken from the Iconclass database. The dataset comprises 21,673 images annotated with classes for training, along with 2,632 test images annotated with class labels only and 808 test images annotated with bounding boxes. Furthermore, 2,628 labeled images and 1,625 bounding-box annotated images are provided as validation sets. Similarly, IconArt focuses on seven classes of Christian iconography, with 1,421 images for classification training, 2,053 images for classification evaluation, 610 images for classification validation, and 1,480 images for detection evaluation. A summary of the evaluation datasets is provided in the Supplementary Material.

Implementation details In the weakly-supervised setting, we use a CLIP ViT-B/32[3] as the CLIP image encoder. We use two layers for the MLP for ArtDL 2.0 and three layers for IconArt. Both MLPs use a hidden size of 384 and ReLU activation. As most images in ArtDL 2.0 contain a single object, we train with single-label classification. Meanwhile, IconArt tends to have multiple classes in each image, so we use multi-label classification for it. All MLPs are trained for 100 epochs with AdamW [32] optimizer and a batch size of 512. Training hyperparameters per dataset are provided in the Supplementary Material.

[3] https://huggingface.co/openai/clip-vit-base-patch32

Table 1. Weakly-supervised classification results. All metrics are macro-averaged. Params indicates the number of trainable parameters.

Method	Params	ArtDL 2.0 [37]				IconArt [15]			
		P	R	F1	AP	P	R	F1	AP
Milani [37]	23.4M	72.7	69.8	**69.1**	**71.6**	71.7	61.9	65.6	73.1
MI-Max-HL [16]	3.7M	4.0	**85.0**	9.0	17.6	24.0	**97.0**	36.0	54.0
WSCP (ours)	0.4M	**78.1**	49.3	57.8	57.5	**80.5**	69.6	**74.1**	**80.7**

In the zero-shot setting, we use LLaVA-NeXT-34B[4] [30] as VLM with two types of prompts as input:

- **Choice**: We query the VLM to select which classes among $\mathcal{L}$ are present in the image. The text post-processing function P simply consists of extracting the predicted classes from the VLM text output.
- **Score**: We query the VLM to provide a confidence score $s_l \in [0, 1]$ for each $l \in \mathcal{L}$, with each score indicating the likelihood that a label l appears in the image. The text post-processing function P consists of extracting the labels and scores and thresholding the scores with predefined τ to identify a set $\mathcal{L}' = \{l \in \mathcal{L} | s_l > \tau\}$. We tune τ on the validation split of IconArt and set $\tau = 0.5$.

For the class-conditioned detector, we use Stable Diffusion 2[5] for inversion and reconstruction. We perform null-text inversion[6] [38] over 500 steps and reconstruct images for 50 steps. We consider two prompt construction methods for inversion and reconstruction:

- **Template**: We insert the class name into a pre-defined prompt template.
- **Caption**: We use the same VLM to describe the image with a caption that contains the class name. Captions that do not contain the class or contain the class at a position beyond the maximum input length of the diffusion model are prepended with a prompt template formatted with the class name.

For ArtDL 2.0, we use the Wikipedia[7] article titles corresponding to each class as labels. For IconArt, we change some of the class names as follows: *Saint Sebastien* to *person*, *child Jesus* to *baby*, and *nudity* to *naked person*. All the prompts can be found in the Supplementary Material.

4.1. Weakly-supervised evaluation

Baselines We compare our method against previous work on weakly-supervised object detection:

- PCL [55]: It uses an MIL network on top of projected CNN features to generate proposal scores and clusters, which are used to start an iterative refinement of an instance classifier. At each refinement step, the current instance classifier is guided by proposal clusters generated from the previous step.
- CASD [21]: It also has an MIL head and an iteratively refined instance classifier over image features, but features across input transformations and layers are aggregated to create comprehensive attention maps and are used to guide self-distillation of the detector.
- UWSOD [52]: Object locations are proposed with an anchor-based self-supervised object proposal generator. Both detection scores and boxes are progressively improved via a step-wise bounding-box fine-tuning process. A multi-rate resampling pyramid is used to combine multi-scale contextual information.
- CAM+PaS [37]: A ResNet-50 is fine-tuned on the target dataset and used to extract class-activation maps (CAMs) from images. Percentiles of the CAM values are used to threshold the CAMs and bounding boxes are drawn around the salient area.
- Milani [37]: CAM+PaS is used to create pseudo-ground-truth bounding boxes for a set of images and a Faster R-CNN is trained on them.
- MI-Max-HL [16]: A pretrained Faster R-CNN is used to extract proposal embeddings and objectness scores. Embeddings are processed by a fully connected layer and a MIL classifier before being multiplied by objectness scores. The highest-scoring proposals from the MIL classifier are taken as positive predictions during weakly supervised fine-tuning.

We do not re-implement the above baselines; instead, we report results as presented in previous works [16, 37, 55].

Classification results To evaluate classification accuracy, we report precision (P), recall (R), F1 score (F1), and classification average precision (AP) for each dataset in Tab. 1. Using only a simple MLP for training, our weakly-supervised class proposer (WSCP) achieves the highest P, F1 score and AP on IconArt. Moreover, it shows competitive performance compared to a more complex fully fine-tuned ResNet-50 (Milani) on the ArtDL 2.0 dataset, where it also obtains the best precision. In summary, our WSCP not only outperforms state-of-the-art models in terms of simplicity but also obtains competitive results, showcasing its effectiveness in weakly supervised object detection.

Object detection results Detection results are reported in Tab. 2 as AP_{50}, which measures the area under the

[4] https://huggingface.co/liuhaotian/llava-v1.6-34b
[5] https://huggingface.co/stabilityai/stable-diffusion-2-base
[6] https://github.com/google/prompt-to-prompt/
[7] https://www.wikipedia.org/

Table 2. Weakly-supervised object detection results as AP_{50}.

Method	Train detector?	ArtDL 2.0 [37]	IconArt [15]
PCL [55]	✓	24.8	5.9
CASD [21]	✓	13.5	4.5
UWSOD [52]	✓	7.6	6.2
CAM+PaS [37]	✓	40.3	3.2
Milani [37]	✓	41.5	**16.6**
MI-Max-HL [16]	×	8.2	14.5
NADA (with WSCP)	×	**45.8**	13.8

precision-recall curve for detections above a 0.5 intersection over union (IoU) threshold. For NADA, we report the result of the best prompt construction method per dataset. Results show that NADA achieves the highest performance on ArtDL 2.0 with an AP_{50} of 45.8. Meanwhile, on IconArt, NADA stays competitive with Milani and MI-Max-HL methods with only 0.7 and 2.8 AP_{50} points difference, respectively, while outperforming the remaining baselines by a higher score (7.6 AP_{50} points higher than the next best method). NADA achieves this while being one of only two evaluated methods that do not require training the detector. Note that we report more results in Tab. 2 than in Tab. 1 as some methods only reported detection and not classification scores.

Interestingly, superior AP classification accuracy does not necessarily translate to a better AP_{50} in object detection, as previously noted in [16]. This discrepancy suggests that while our Stable Diffusion-based method detects and localizes depicted classes more accurately, Faster R-CNN and its variants leverage stronger features than the off-the-shelf internal representations of Stable Diffusion. To investigate this, in Sec. 4.3, we isolate the influence of the class proposer and report the results of the class-conditioned detector when a perfect class proposal module is assumed.

4.2. Zero-shot evaluation

Baselines As there is no prior research on zero-shot object detection in art, and DiffusionSeg [34], which is the most closely related work leveraging Stable Diffusion for object segmentation, does not have publicly available code for reproduction, we compare our zero-shot NADA against two baseline class proposals methods: CLIP-based and InstructBLIP-based.

- CLIP: We use the standard zero-shot protocol in CLIP [42], where each test image is embedded with a pretrained CLIP image encoder and matched against a text embedded with a pretrained CLIP text encoder with the prompt ``A painting of [CLASS]''. Any [CLASS] with a cosine similarity greater than 0.28 is taken as a predicted class. This threshold is based on the CLIP filtering process of LAION-5B [49].
- InstructBLIP: We replace LLaVA with InstructBLIP-Vicuna-7B[8] [11]. We prompt InstructBLIP with each class individually using the query ``Is [CLASS] in the painting?''. An output containing ``yes'' is taken as a positive prediction and any other response is considered a negative prediction. We use one prompt per class as using our choice or score prompts tended to produce irrelevant results.

Note that whereas CLIP and InstructBLIP class proposals need to use a prompt for each class and image, our zero-shot class proposer (ZSCP) uses only one prompt per image. We use the same class-conditioned detector based on Stable Diffusion for baseline detection results.

Classification results Zero-shot classification results are shown in Tab. 3. We report the results of our ZSCP using the two types of prompts: choice and score. Results show that both choice and score prompts are not only the most efficient methods, requiring only one prompt per image compared to CLIP and InstructBLIP, which need a prompt per class and image, but also achieve the highest precision, F1 score and AP on both datasets. InstructBLIP provides the best recall, however, it also has the lowest precision on both datasets, indicating it is overpredicting classes. When comparing score to choice prompting methods, we observe inconsistent results. While the score prompt leads to more accurate predictions in terms of precision and AP for IconArt, the choice prompt leads to the best classification performance for ArtDL 2.0.

Object detection results Zero-shot detection results are reported in Tab. 4. For NADA, we report the best result among VLM prompts and prompt construction for each dataset. On ArtDL 2.0, following the classification results, the large gap between the ZSCP and the two baselines on F1 score and AP leads to NADA obtaining the highest detection performance with an AP_{50} of 21.8, surpassing InstructBLIP by 3.2 AP_{50} points. On the IconArt dataset, despite the classification results among all the models being closer in terms of F1 score and AP, NADA is able to achieve the highest performance with an AP_{50} of 15.8, which is 7.9 AP_{50} points above InstructBLIP. Notably, the zero-shot version of NADA even outperforms some weakly-supervised methods in Tab. 2 while requiring no annotations whatsoever. Although its performance may lag behind state-of-the-art fully-supervised methods, it does not require any training on the target dataset.

4.2.1 Qualitative analysis

We show attention map visualizations and bounding boxes predicted by NADA (with ZSCP) in Fig. 4. Stable Diffusion's knowledge of art images can indeed localize objects

[8] https://huggingface.co/Salesforce/instructblip-vicuna-7b

Table 3. Zero-shot classification results. Num. prompts indicate the number of prompts per image. All metrics are macro-averaged.

Class proposal	Num. prompts	ArtDL 2.0				IconArt			
		P	R	F1	AP	P	R	F1	AP
CLIP	Num. classes	30.2	27.7	15.2	14.6	65.8	55.1	49.9	48.5
InstructBLIP	Num. classes	27.3	**59.4**	32.5	20.3	61.2	**79.8**	65.2	52.8
ZSCP choice (ours)	1	**39.8**	41.4	**37.7**	**23.7**	62.6	80.2	**68.7**	55.8
ZSCP score (ours)	1	32.7	19.6	19.7	15.5	**84.9**	60.7	67.9	**63.0**

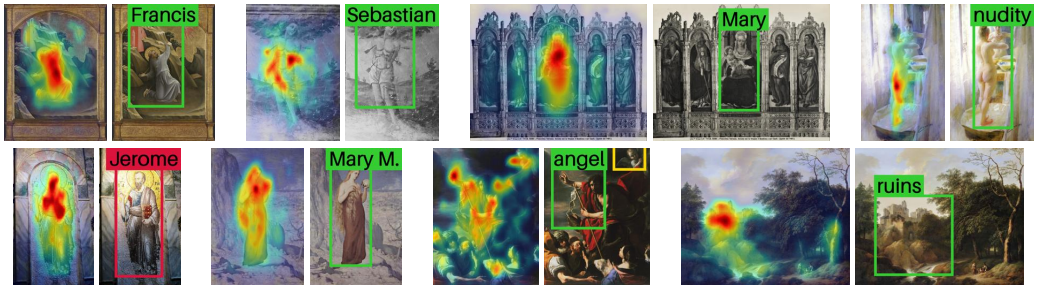


Figure 4. ArtDL 2.0 and IconArt test images overlaid with NADA (with ZSCP) attention maps and bounding boxes, shown in pairs. Redder areas indicate higher attention while bluer areas indicate lower attention. Correct model predictions are in green, incorrect model predictions are in red, and ground truth boxes when the predicted box has < 0.5 IoU with the ground truth are in yellow.

Table 4. Zero-shot object detection results as AP_{50}.

Class proposal	ArtDL 2.0 [37]	IconArt [15]
CLIP	13.3	6.8
InstructBLIP	18.6	7.9
NADA (with ZSCP)	**21.8**	**15.1**

in paintings, as the attention maps highlight the sought labels. The bounding boxes contain the salient regions of the attention map, showing that NADA can transform attention maps into meaningful bounding boxes. Even when there are multiple subjects, NADA is capable of detecting Mary among five people (top row, second from right). One can also see that Stable Diffusion knows about the iconographic attributes of some characters, such as the arrows of Saint Sebastian[9] (top row, second from left). NADA may fail when the class is incorrect, such as misclassifying Paul as Jerome (bottom row, leftmost), however it is still able to localize the subject. NADA may also correctly identify objects but incorrectly localize them, such as identifying the wrong person as an angel (bottom row, second from right).

4.3. Analysis and ablation studies

Upper bound object detection performance We measure the upper-bound detection performance of NADA when assuming a perfect class proposal module that always predicts the correct classes. This NADA configuration, referred to as *Oracle* allows us to discern the accuracy contribution of the class-conditioned detector and the adequacy of Stable Diffusion's cross-attention maps for art object detection. Given the correct labels, the Oracle substantially improves performance from 21.8 (zero-shot) and 45.8 (weakly-supervised) to 61.3 AP_{50} on the ArtDL 2.0 dataset. On the IconArt dataset, the Oracle boosts object detection results from 15.1 (zero-shot) and 13.8 (weakly-supervised) to 18.7 AP_{50}. This implies that a large part of the object detection performance is dependent on the accuracy of the class proposer. Within the same method, better class predictions lead to better object detection performance.

Impact of different thresholds We analyze the use of Otsu threshold in Figs. 5 and 6. Figure 5 shows the watershed segmentation mask prior to bounding box construction for thresholds ranging from 0.1 to 0.9 in intervals of 0.1. Lower thresholds lead to masks that are too large, while higher thresholds result in masks that are too small, with thresholds above 0.5 resulting in no mask being segmented. In this example, Otsu's method determined the optimal threshold value to be 0.33, which was not among any of the manually tested thresholds. Figure 6 shows AP_{50} performance on the validation detection split of ArtDL 2.0 for each of the thresholds. We find that the best-performing threshold is still the one determined via the Otsu's method.

Impact of prompting method Table 5 shows the impact of NADA's prompting method on object detection performance. We observe different behaviors between the two datasets, but consistent behaviors within. Caption prompts show consistently lower performance than tem-

[9]Saint Sebastian's association with arrows is a common representation in iconography and is discussed in https://en.wikipedia.org/wiki/Saint_Sebastian.

Prompt	ArtDL 2.0				IconArt			
	ZSCP choice	ZSCP score	WSCP	Oracle	ZSCP choice	ZSCP score	WSCP	Oracle
Template	21.8	13.8	45.8	61.3	7.8	12.1	11.7	15.2
Caption	20.2	12.6	42.5	58.0	9.9	15.1	13.8	18.7

Table 5. AP_{50} for different prompt construction methods across NADA systems with different class proposers.

Prompt	ArtDL 2.0		IconArt	
	single	multiple	single	multiple
Template	22.4	0.03	5.2	8.1
Caption	20.7	0.02	5.6	11.0

Table 6. AP_{50} for NADA (with ZSCP) on images with a *single* class and images with *multiple* classes.

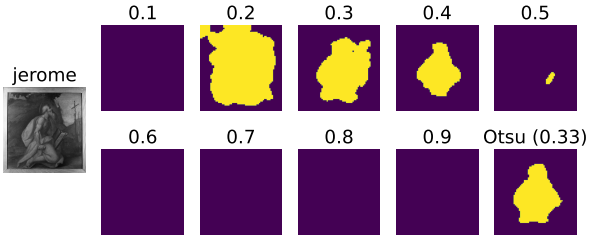


Figure 5. Mask prior to bounding box drawing for different thresholds, including Otsu's method.

plate prompts in ArtDL 2.0, but consistently outperform template prompts in IconArt. We initially believed this was due to ArtDL 2.0 and IconArt tending to have one and multiple classes per image respectively. However, upon checking how NADA (with ZSCP) performs on single and multi-object subsets of ArtDL 2.0 and IconArt in Tab. 6, we find that this is not the case. Regardless of the number of classes in an image, template prompt construction improves ArtDL 2.0 detection while caption prompt construction boosts IconArt detection. It should be noted however that templates outperform captions more on ArtDL 2.0 when there is only one label (+1.7 AP_{50}) than when there are multiple (+0.01 AP_{50}) while captions outperform templates more on IconArt when there are multiple labels (+2.9 AP_{50}) than where there is only one (+0.4 AP_{50}).

Object detection in the wild We use NADA's class-conditioned detector with caption prompt construction to detect objects in WikiArt images *in the wild*. Examples are shown in Fig. 1. These images contain subjects that are not typically considered in natural image object detectors and are even not among the classes in either ArtDL 2.0 or IconArt, while also covering a variety of styles including Renaissance, ukiyo-e, and surrealism. NADA is able to detect uncommon objects often portrayed in art such as banners and shields alongside mythological creatures such as dragons, Pegasus, and unicorns. NADA is also able to

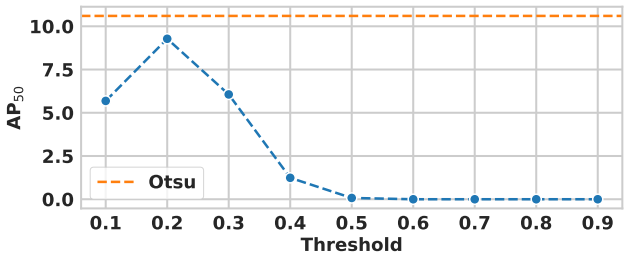


Figure 6. AP_{50} results on the ArtDL validation detection set for varying thresholds and when using Otsu's method.

understand artistic interpretations of these classes such as a surrealist rendering of melting skull. These results indicate that NADA is capable of bridging the domain gap in both classes and styles presented by art images. Furthermore, contrary to other methods performing object detection from text inputs, our approach does not rely on Google Images search of the objects of interest [9].

5. Conclusion

We introduced NADA, a method that applies diffusion models' knowledge of art to reduce the amount of supervision needed for object detection in paintings, specifically for weakly-supervised and zero-shot detection. Weakly-supervised NADA (with WSCP) competes closely with and outperforms other methods on art object detection, while zero-shot NADA (with ZSCP) is one of the first methods for zero-shot object detection in the domain of paintings. Detection performance improves when NADA's class proposer is always correct, demonstrating the importance of the class proposer to the whole pipeline. Prompting methods have varying effects on object detection depending on the target dataset. We use NADA's class-conditioned detector to detect objects in WikiArt images in the wild, demonstrating the detector's capacity to localize objects that are more commonly found in art images.

Acknowledgements This work is partly supported by JSPS KAKENHI No. JP23H00497 and JP22K12091, JST CREST Grant No. JPMJCR20D3, and JST FOREST Grant No. JPMJFR216O.

References

[1] Tasweer Ahmad and Maximilian Schich. Toward cross-domain object detection in artwork images using improved yolov5 and xgboosting. *IET Image Processing*, 17(8), 2023. 2

[2] Zechen Bai, Yuta Nakashima, and Noa Garcia. Explain me the painting: Multi-topic knowledgeable art description generation. In *ICCV*, 2021. 1

[3] Dmitry Baranchuk, Andrey Voynov, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. In *ICLR*, 2022. 2

[4] Siwar Bengamra, Olfa Mzoughi, André Bigand, and Ezzeddine Zagrouba. A comprehensive survey on object detection in visual art: taxonomy and challenge. *Multimedia Tools and Applications*, 83(5), 2024. 2

[5] Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *CVPR*, 2016. 2

[6] Hongping Cai, Qi Wu, and Peter Hall. Beyond photo-domain object recognition: Benchmarks for the cross-depiction problem. In *CVPRW*, 2015. 1

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*. Springer, 2020. 1

[8] Eva Cetinic. Iconographic image captioning for artworks. In *ICPRW*. Springer, 2021. 1

[9] Elliot J Crowley and Andrew Zisserman. In search of art. In *VISART Workshop at ECCV*. Springer, 2015. 8

[10] Elliot J Crowley and Andrew Zisserman. The art of detection. In *VISART Workshop at ECCV*. Springer, 2016. 2

[11] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*, 2023. 6

[12] Chengjian Feng, Yujie Zhong, Zequn Jie, Weidi Xie, and Lin Ma. Instagen: Enhancing object detection by training on synthetic dataset. In *CVPR*, 2024. 2

[13] Noa Garcia, Chentao Ye, Zihua Liu, Qingtao Hu, Mayu Otani, Chenhui Chu, Yuta Nakashima, and Teruko Mitamura. A dataset and baselines for visual question answering on art. In *VISART Workshop at ECCV*. Springer, 2020. 1

[14] Shiry Ginosar, Daniel Haas, Timothy Brown, and Jitendra Malik. Detecting people in cubist art. *AI Matters*, 1(3), 2015. 2

[15] Nicolas Gonthier, Yann Gousseau, Said Ladjal, and Olivier Bonfait. Weakly supervised object detection in artworks. In *VISART Workshop at ECCV*, 2018. 1, 2, 4, 5, 6, 7

[16] Nicolas Gonthier, Saïd Ladjal, and Yann Gousseau. Multiple instance learning on deep features for weakly supervised object detection with extreme domain shifts. *CVIU*, 214, 2022. 2, 5, 6

[17] Peter Hall, Hongping Cai, Qi Wu, and Tadeo Corradi. Cross-depiction problem: Recognition and synthesis of photographs and artwork. *Computational Visual Media*, 1, 2015. 1

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33, 2020. 2

[20] Ting-I Hsieh, Yi-Chen Lo, Hwann-Tzong Chen, and Tyng-Luh Liu. One-shot object detection with co-attention and co-excitation. *NeurIPS*, 32, 2019. 2

[21] Zeyi Huang, Yang Zou, BVK Kumar, and Dong Huang. Comprehensive attention self-distillation for weakly-supervised object detection. *NeurIPS*, 33, 2020. 5, 6

[22] Bekkouch Imad Eddine Ibrahim, Victoria Eyharabide, Valérie Le Page, and Frédéric Billiet. Few-shot object detection: Application to medieval musicological studies. *J. Imaging*, 8(2), 2022. 1

[23] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *CVPR*, 2018. 2

[24] Laurynas Karazija, Iro Laina, Andrea Vedaldi, and Christian Rupprecht. Diffusion models for zero-shot open-vocabulary segmentation. In *ECCV*, 2025. 2

[25] Selina Khan and Nanne van Noord. Context-infused visual grounding for art. In *VISART Workshop at ECCV*, 2024. 2

[26] Neehar Kondapaneni, Markus Marks, Manuel Knott, Rogério Guimaraes, and Pietro Perona. Text-image alignment for diffusion-based perception. In *CVPR*, 2024. 2

[27] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *CVPR*, 2022. 1

[28] Ziyi Li, Qinye Zhou, Xiaoyun Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Open-vocabulary object segmentation with diffusion models. In *ICCV*, 2023. 1, 2

[29] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*. Springer, 2014. 1

[30] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 36, 2024. 5

[31] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 1

[32] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *ICLR*, 2019. 4

[33] Yue Lu, Chao Guo, Xingyuan Dai, and Fei-Yue Wang. Data-efficient image captioning of fine art paintings via virtual-real semantic alignment training. *Neurocomputing*, 490, 2022. 1

[34] Chaofan Ma, Yuhuan Yang, Chen Ju, Fei Zhang, Jinxiang Liu, Yu Wang, Ya Zhang, and Yanfeng Wang. Diffusionseg: Adapting diffusion towards unsupervised object discovery. *arXiv preprint arXiv:2303.09813*, 2023. 2, 6

[35] Prathmesh Madhu, Anna Meyer, Mathias Zinnen, Lara Mührenberg, Dirk Suckow, Torsten Bendschus, Corinna Reinhardt, Peter Bell, Ute Verstegen, Ronak Kosti, et al.

One-shot object detection in heterogeneous artwork datasets. In *IPTA*. IEEE, 2022. 2

[36] Louie Meyer, Johanne Engel Aaen, Anitamalina Regitse Tranberg, Peter Kun, Matthias Freiberger, Sebastian Risi, and Anders Sundnes Løvlie. Algorithmic ways of seeing: Using object detection to facilitate art exploration. In *CHI*, 2024. 1

[37] Federico Milani, Nicolò Oreste Pinciroli Vago, and Piero Fraternali. Proposals generation for weakly supervised object detection in artwork images. *J. Imaging*, 8(8), 2022. 1, 2, 3, 4, 5, 6, 7

[38] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *CVPR*, 2023. 4, 5

[39] Peer Neubert and Peter Protzel. Compact watershed and preemptive slic: On improving trade-offs of superpixel segmentation algorithms. In *ICPR*. IEEE, 2014. 3, 4

[40] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *ICML*. PMLR, 2022. 2

[41] Nobuyuki Otsu et al. A threshold selection method from gray-level histograms. *Automatica*, 11(285-296), 1975. 4

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICLR*. PMLR, 2021. 2, 6

[43] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2), 2022. 2

[44] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, 2016. 2

[45] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS*, 28, 2015. 1, 2

[46] Zhongzheng Ren, Zhiding Yu, Xiaodong Yang, Ming-Yu Liu, Yong Jae Lee, Alexander G Schwing, and Jan Kautz. Instance-aware, context-focused, and memory-efficient weakly supervised object detection. In *CVPR*, 2020. 2

[47] Artem Reshetnikov, Maria-Cristina Marinescu, and Joaquim More Lopez. Deart: Dataset of european art. In *VISART Workshop at ECCV*. Springer, 2022. 1

[48] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2

[49] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *NeurIPS*, 35, 2022. 2, 6

[50] Jinhwan Seo, Wonho Bae, Danica J Sutherland, Junhyug Noh, and Daijin Kim. Object discovery via contrastive learning for weakly supervised object detection. In *ECCV*. Springer, 2022. 2

[51] Xi Shen, Alexei A Efros, and Mathieu Aubry. Discovering visual patterns in art collections with spatially-consistent feature learning. In *CVPR*, 2019. 1

[52] Yunhang Shen, Rongrong Ji, Zhiwei Chen, Yongjian Wu, and Feiyue Huang. Uwsod: Toward fully-supervised-level capacity weakly supervised object detection. *NeurIPS*, 33, 2020. 5, 6

[53] Pierre J Soille and Marc M Ansoult. Automated basin delineation from digital elevation models using mathematical morphology. *Signal Processing*, 20(2), 1990. 3, 4

[54] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *ICLR*, 2021. 2

[55] Peng Tang, Xinggang Wang, Song Bai, Wei Shen, Xiang Bai, Wenyu Liu, and Alan Yuille. Pcl: Proposal cluster learning for weakly supervised object detection. *IEEE TPAMI*, 42(1), 2018. 5, 6

[56] Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Türe. What the daam: Interpreting stable diffusion using cross attention. In *ACL*, 2023. 4

[57] Fang Wan, Chang Liu, Wei Ke, Xiangyang Ji, Jianbin Jiao, and Qixiang Ye. C-mil: Continuation multiple instance learning for weakly supervised object detection. In *CVPR*, 2019. 2

[58] Jinglong Wang, Xiawei Li, Jing Zhang, Qingyuan Xu, Qin Zhou, Qian Yu, Lu Sheng, and Dong Xu. Diffusion model is secretly a training-free open vocabulary semantic segmenter. *arXiv preprint arXiv:2309.02773*, 2023. 2

[59] Nicholas Westlake, Hongping Cai, and Peter Hall. Detecting people in artwork with cnns. In *VISART Workshop at ECCV*. Springer, 2016. 1, 2

[60] Qi Wu, Hongping Cai, and Peter Hall. Learning graphs to model visual objects across different depictive styles. In *ECCV*. Springer, 2014. 1

[61] Weijia Wu, Yuzhong Zhao, Mike Zheng Shou, Hong Zhou, and Chunhua Shen. Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. In *ICCV*, 2023. 2

[62] Yankun Wu, Yuta Nakashima, and Noa Garcia. Not only generative art: Stable diffusion for content-style disentanglement in art analysis. In *ICMR*, 2023. 2

[63] Ryota Yoshihashi, Yuya Otsuka, Tomohiro Tanaka, et al. Attention as annotation: Generating images and pseudo-masks for weakly supervised semantic segmentation with diffusion. *arXiv preprint arXiv:2309.01369*, 2023. 2