# Object Detection/Image analysis version 1

Deliverable D2.3

Version DRAFT


NEGOTIATING OLFACTORY AND SENSORY EXPERIENCES IN CULTURAL HERITAGE PRACTICE AND RESEARCH

The Odeuropa project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004469. This document has been produced by the Odeuropa project. The content in this document represents the views of the authors, and the European Commission has no liability in respect of the content.

| | |
|---|---|
| **Grant Agreement No.** | 101004469 |
| **Project Acronym** | ODEUROPA |
| **Project full title** | Negotiating Olfactory and Sensory Experiences in Cultural Heritage Practice and Research |
| **Funding Scheme** | H2020-SC6-TRANSFORMATIONS-2020 |
| **Project website** | http://odeuropa.eu/ |
| **Project Coordinator** | Prof. Dr. Inger Leemans<br>KNAW Humanities Cluster<br>Email: inger.leemans@huc.knaw.nl |
| **Document Number** | Deliverable D2.3 |
| **Status & version** | FINAL |
| **Contractual date of delivery** | 30 June 2022 |
| **Date of delivery** | 30 July 2022 |
| **Type** | Report and Software |
| **Security (distribution level)** | Public |
| **Number of pages** | 27 |
| **WP contributing to the deliberable** | WP2 |
| **WP responsible** | WP2 |
| **EC Project Officer** | Hinano Spreafico |
| **Authors:** Mathias Zinnen,[1] Vincent Christlein[1] | |
| **Internal reviewers:** Lizzie Marx,[1] William Tullett[2] | |
| **Affiliations** (1) FAU , (2) ARU | |
| **Keywords:** Computer Vision, Object Detection, Software, Evaluation | |
| **Abstract:** This deliverable provides a summary of the activities and experiments we have conducted to automatically recognize olfactory references. We explain the techniques we apply to obtain domain adaptation and present the results of multiple object detection approaches on the dataset that has been used for the ODOR Challenge (D2.4, to be submitted in M20) and been presented in D2.2. In this context, the object detection accuracies for different smell-relevant objects are analyzed separately. Additionally, initial experiments with the aim of recognizing gestures in historical artworks are presented and an outlook about next steps in automatic smell reference recognition is provided. | |

# Table of Revisions

| Version | Date | Description and reason | By | Affected sections |
|---|---|---|---|---|
| 0.1 | June 2022 | Draft | Mathias Zinnen and Vincent Christlein | All |
| 0.2 | July 2022 | Revision after Review | Mathias Zinnen | |
| 1.0 | July 2022 | Final check and approval by project manager | Marieke van Erp | - |

## Executive Summary

This deliverable summarizes our efforts in the automatic recognition of olfactory references. Based on the taxonomy of olfactory references created in Deliverable 2.1, we focus on two types of olfactory references, i. e., Olfactory Objects, and Smell Gestures. We describe the experiments we have conducted to enable their automatic recognition, analyze the algorithms we have applied quantitatively. Furthermore, we provide an analysis of detection accuracy per object, and give qualitative examples of the recognition results.

Additionally, we explain the approaches we have taken to obtain domain adaptation, i. e., the ability to transfer knowledge obtained on large scale datasets of modern photographs to the domain of historical artworks we are interested in.

# Contents

Figure 1: *Still Life with Flowers*. Cornelia van der Mijn, 1762. Rijksmuseum Amsterdam.

# 1 Introduction

The visual branch of the Odeuropa project aims at investigating past scents of the past through their visual traces in art. One way of approaching visual smell references is by closely investigating specific artworks in their historical and cultural context. By providing a taxonomy of visual smell references (D2.1) and manually annotating a large number of images according to this scheme (D2.2), we enable this kind of smell research. Additionally and complementary, the Odeuropa project specifically aims at another, data-driven perspective on past smells in the history of European arts that can be called 'distant viewing' [Arnold and Tilton, 2019]. By looking at the visual arts at a larger scale we provide tools that enable smell researchers to investigate larger trends and structural developments that might remain invisible to an approach that focuses on qualitative inspections of singular works of arts: Research questions leveraging big data analysis might for example investigate historical shifts in the co-occurrences of specific flowers or fruits in still lifes as in Fig. 1 and thus analyze changes in smell-combinations. Or visual sources might be taken into account to analyze the presence of domestic animals in city scenes or fish markets in city scenes, inferring how urban smellscapes might have evolved over time.

While the existing meta-data for the images in digital collections can provide smell researchers hints to answer research questions in a quantitative manner, smell-related annotations related to the image content are mostly lacking. This motivates the implementation of learning systems that are capable to automatically annotate images with visible smell references at a larger scale than it is possible using human annotators.

In this deliverable, we report our experiments in the automatic extraction of smell references from historical artworks and describe possible uses of the results of automatic smell reference extraction.

In D2.1, we identified *olfactory objects*, *smell gestures*, *fragrant spaces*, and *olfactory iconographies* as four kinds of visual smell references. Here we are covering the detection of olfactory objects (Section 5) and recognition of smell gestures (Section 6), since these two domains have been the main focus of our efforts so far.

Recognition systems for all kinds of smell references have in common that they have to bridge

a domain gap between photographs, on which computer vision algorithms are usually trained, and historical artworks, on which we are applying them. In Section 4, we will thus present the experiments that we are conducting in order to overcome the gap between these domains.

Before diving into the technical details of automatic smell recognition, we give some examples of possible research questions for data-driven smell research that is based on smell references in Section 2 and how automatically extracted data can help in answering these questions.

# 2 Quantatitive Analysis of Visual Smell References

One key question for humanities researchers interested in smell and the past concerns how people react to smells in paintings. Whilst we cannot initially plot the reactions themselves, WP2 has begun to explore gestures of smelling in paintings through object-relation and pose-estimation. This will enable researchers to identify images where smelling is taking place, with the possibility of then exploring and categorising the reactions themselves using further levels of either computational or human interpretation. They are some of the research avenues that object detection will be able to develop. Combining these findings with the results of automatic object detection might then allow us to explore what circumstances and narratives are taking place where gestures of smelling are depicted, thereby helping us to discover more objects with olfactory significance that might then be annotated in images, or understand the types of emotions or facial responses that occur in images of smelling, thereby helping to compliment the emotion annotation being developed in WP3.

We started the project assuming that depictions of flower carry an olfactory dimension because they have been viewed as olfactory compositions as well as visual compositions. We started the project investigating whether depictions of flowers carry olfactory significance within the compositions. Using the results of a system that is able to automatically recognize flower species in a large corpus of still lifes and other flower compositions might help us to test this hypothesis by analyzing to what extent the flowers deemed to have olfactory significance in texts sourced by WP3 correspond to the trends in representing particular flowers in works of art. This research question actually guided us in the definition of distinguishable flower species where we analyzed 13 European perfumery texts ranging across the 1600 to 1925 period to identify flower species with high olfactory significance. These kinds of questions also enable researchers to validate researcher hypotheses across different modalities, i. e., text and images.

Of specific interest are olfactory artefacts, which are purposefully designed to carry odorants. These objects have a direct relation to the scents they carry and can serve as cues within texts and artworks to specific odorants. Two artefacts we have specifically focused on are scented gloves and pomanders which can both appear as olfactory accessories in portraiture: Pomanders are precious containers of aromatic substances that were used to drive away bad smells which were believed to cause disease. Gloves were often heavily scented to cover repulsive leather smells resulting from historical leather tanning processes. Both accessories are prominently displayed in historical portraiture (cf. Section 2). Among the many research possibilities that the olfactory objects present, tracing the relative frequency of their occurrences in portraits over time could lead to interesting insights into the tastes for olfactory accesories, the ways in which sitters are displayed, and what roles smells have played in this.

Providing the tools to quantitatively answer such questions or easily query large collections of visual data does not replace close reading and art-historical expertise in the humanities. But it has the potential to open up the field for new research questions and complement humanities research with another way of empirically grounding research hypotheses.

# 3 Evaluation Metrics

To evaluate the performance of the automatic smell reference recognition, we applied an evaluation metric for a quantitative evaluation, as well an inspection of the results for a qualitative evaluation.

(a) (b)

Figure 2: (a). Woman with a pomander. *Portrait of a Woman, probably Maria Schuurman*. Anonymous, c. 1599-1600. Rijksmuseum Amsterdam.
(b). Portrait of a woman with gloves. *Portrait of Johanna le Maire (c. 1601-60)*. Nicolaes Eliasz Pickenoy, c. 1622 - c. 1629. Rijksmuseum Amsterdam.

To evaluate object detection, we apply the COCO metric (mAP) [Lin et al., 2014], and pascal VOC [Everingham et al., 2010] metric ($\text{mAP}_{50}$), which are both standard evaluation metrics for object detection. Both metrics draw on Intersection over Union (IoU), which defines if a pair of predicted and ground truth bounding boxes are counted as correct prediction. IoU is defined as the ratio of intersection and union between the predicted and actual bounding box. A prediction is considered to be correct (True Positive, $TP$) if the IoU is greater than a predefined threshold value, and False Positive ($FP$) otherwise. For a set of prediction we can define the precision $P = \frac{TP}{TP+FP}$ and recall $R = \frac{TP}{TP+FN}$. For a given set of ground truth annotations, the ratio of precision and recall will vary depending on how many predictions are taken into account. The larger the set of considered predictions, the higher the recall will be and the lower the precision. Object detection metrics like COCO and pascal VOC compute the average precision $AP$ averaged over the precision values at multiple recall values $AP_r$ which gives the per-category $C$ average precision $AP(C)$. The mean average precision (mAP) is then defined as the mean over all category average precisions.

$$\frac{1}{|C|} \sum_{N} AP(C) \tag{1}$$

In pascal VOC, the IoU threshold is defined as 0.5 (IoU 0.5). For COCO evaluation, the thresholds range from 0.5 to 0.95 with a step size of 0.05 and the metrics is computed as the mean averaged over all threshold values (IoU 0.5:0.05:0.95).

Since smell-relevant objects are often relatively small, we additionally report the COCO mAP values for small ($\text{mAP}_S$), medium ($\text{mAP}_M$), and large ($\text{mAP}_L$) objects as defined by [Lin et al., 2014].

# 4 Domain Adaptation

Most state-of-the-art computer vision algorithms are trained and evaluated on large-scale datasets consisting of contemporary photographs such as ImageNet [Russakovsky et al., 2015], MS COCO [Lin et al., 2014], or OpenImages [Kuznetsova et al., 2020] which contain millions of annotated images. While this amount of available training data leads to impressive results on photographic data, it also induces a significant performance drop when applied to different imagery, such as paintings or drawings. The visual representation of objects differs significantly between artworks and photographs [Hall et al., 2015]. With realistic imagery, such as the bouquets of xviith-century Dutch still lifes, the mismatch between historic paintings and modern computer vision datasets is comparatively small. In these cases, many computer vision algorithms can directly be applied without much modification: the close observation of the flowers are integral to their depictions. However, as the level of abstraction increases, off-the-shelf algorithms struggle more and more to transfer their vision capabilities trained on photographic material. Apart from this *domain gap* between photographs and artistic imagery, there is also a *content mismatch* between classification categories present in modern datasets and historical olfactory references, cause by historical diachrony on the one hand [Marinescu et al., 2020], and the particularity of some smell-relevant objects and gestures on the other [Zinnen, 2021, Ehrich et al., 2021].

Techniques to transfer knowledge from one domain to a different one are called *domain adaptation* [Farahani et al., 2021]. Most network architectures in computer vision can be divided into a network *head* and a *backbone* where the backbone is extracting compressed feature representations (or *embeddings*) of the high resolution pixel space of the input images and the network head is performing a classification, detection or segmentation task based on these feature vectors. Taking this perspective of a network as the composition of a feature-extracting backbone and network head, domain adaptation techniques can be described as the attempt to obtain feature embeddings that are either robust to style variations or adapted to a specific target domain.

In order to leverage the potential of large scale photographic datasets, we experiment with three different techniques of domain adaptation, i. e., *transfer learning*, *style transfer*, and *self-supervision*.

## 4.1 Transfer Learning

Transfer learning is a training strategy where machine learning algorithms are pre-trained in one domain and then fine-tuned in another, greatly decreasing the amount of required training data in the target domain ([Pan and Yang, 2009, Zhuang et al., 2020]). We conducted transfer learning experiments to improve the recognition of smell-related objects in historical artworks. In these experiments, we defined a subset of our complete set of annotations consisting of 1,126 images with 10,818 annotations in 29 categories (cf. Section 3.1: Initial Experiments Dataset in D2.2 and [Zinnen et al., 2022]).

A common transfer learning procedure is to use detection backbones that have been pre-trained on ImageNet and fine-tune them for object detection [Zhuang et al., 2020]. We expand this strategy by an additional pre-training step, where we train an ImageNet pre-trained object detection network [Ren et al., 2015] using different datasets. Finally, we fine-tune the resulting model using our olfactory artworks dataset (Fig. 3).

For pre-training, we use three different datasets, deviating to varying degreees from our olfactory artworks dataset in terms of categories and style (Table 1): a) Same Categories, Different Styles - A subset of OpenImages (OI) containing only odor objects results in a complete category match (Fig. 4); however, since OpenImages contains only photographs, there is a considerable style difference. b) Different Categories, Same Styles - We apply two object detection datasets from the art domain, which are more similar in terms of style but contain different object categories, namely IconArt (IA) [Gonthier et al., 2018] and PeopleArt (PA) [Westlake et al., 2016].

To ensure a fair comparison between the different pre-training datasets, we reduce each of the datasets to the same size, train three models, and select the best according to a fixed validation

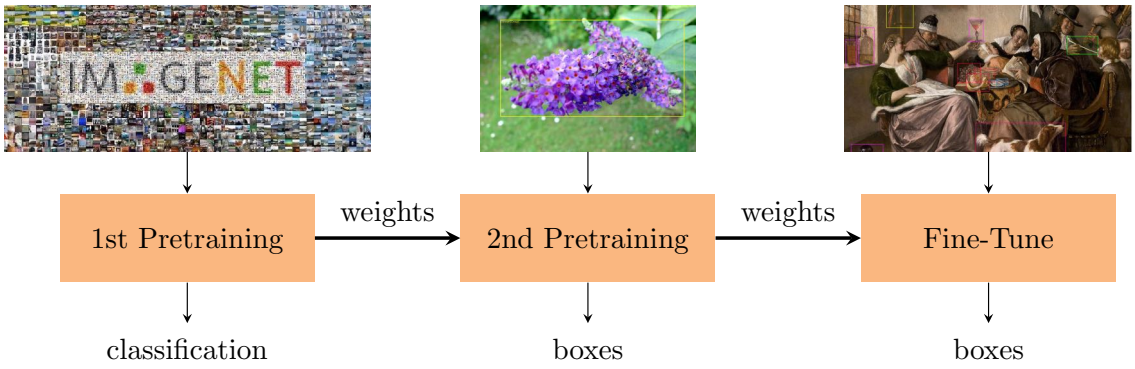



Figure 3: Transfer learning training strategy illustration. We start with a backbone pre-trained on ImageNet for classification, use this model to train an object detection system using different datasets. Finally, the object detection model is fine-tuned on the olfactory artworks dataset.

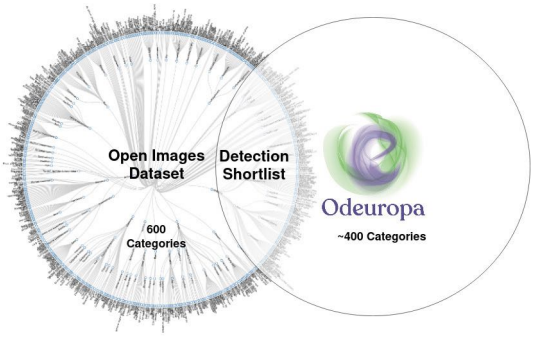



Figure 4: Category overlap between Odeuropa & OpenImages categories

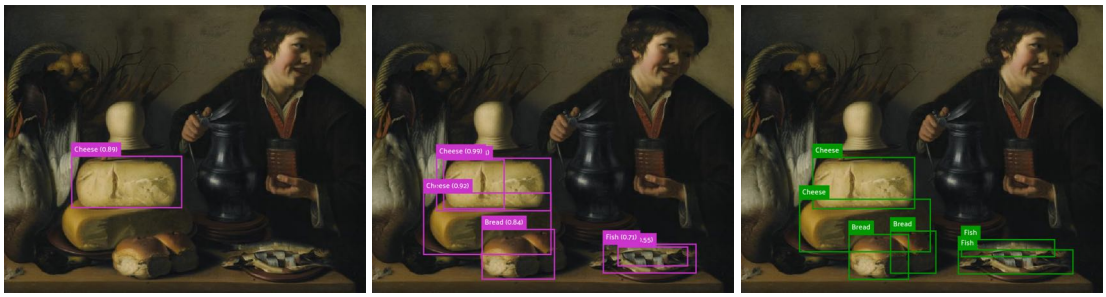
set for each dataset. The resulting models are then fine-tuned on the training set of the olfactory artworks dataset and evaluated on a separate test set. We observe a performance increase for all used pre-training datasets, with a 6.5%/3,4% boost in mAP 50 and COCO mAP, respectively, for the best performing pre-training scheme, which was achieved using the OI dataset. The exemplary object predictions in Fig. 5 show that adding an additional pre-training stage can increase the number of recognized objects.

While adding the additional pre-training stage considerably improved the performance on the 29 categories subset, our experiments with the full ODOR dataset (cf. D2.2) did not show a comparable improvement yet. We are currently investigating the reasons for this and trying to adapt our training strategy to be applicable on the full dataset as well.

## 4.2 Style Transfer

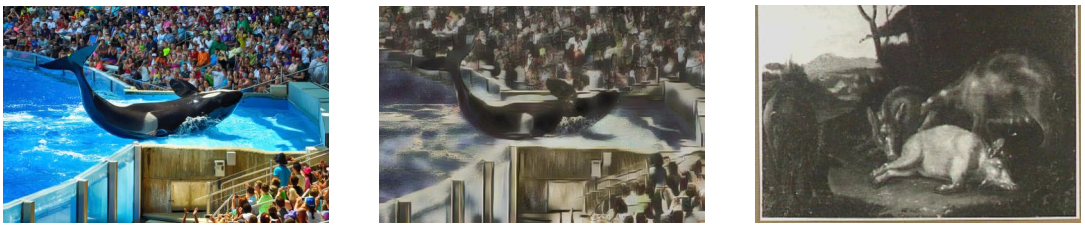
In Style Transfer, labelled photographic datasets are artificially transferred to the style of the target domain before training a recognition system, e. g., by mimicking the style of a specific artist or period. There are multiple ways of achieving this style adaptation, including Generative Adversarial Networks (GANs) [Goodfellow et al., 2014] such as CycleGAN [Zhu et al., 2017], Adaptive Instance Localization (AdaIN) [Huang and Belongie, 2017], or generative image-to-image translation [Isola et al., 2017]. We are using style transfer in multiple experiments:

(1) In CycleGAN, consistency between an image and it stylized counterpart is ensured by additionally learning an inverse function that generates a reconstruction of the original image. The consistency between original and reconstructed images is enforced by a loss over



(a) No pretraining. (b) PA pretraining. (c) Ground truth.

Figure 5: Exemplary object predictions for a detection model without intermediate training (a), with PeopleArt pretraining (b), and ground truth bounding boxes (c). Painting: *Boy holding a pewter tankard, by a still life of a duck, cheeses, bread and a herring.* 1625 – 1674. Gerard van Honthorst. RKD Digital Collection (https://rkd.nl/explore/images/287165). Public Domain.



(a) (b) (c)

Figure 6: Using AdaIN to transfer the style of a training image from OpenImages (a) to the style of one of the artworks from our dataset (c, detail only) generating the stylized training sample (b).

Table 1: An overview of domain & category similarity of the experiment datasets to our olfactory artworks

| Dataset | domain similarity | category similarity | # categories |
|---|---|---|---|
| OpenImages | low | complete match | 29 |
| IconArt | high | medium | 10 |
| PeopleArt | medium | low | 1 |

Table 2: Evaluation of object detection performance. The best performing model pre-trained with OI achieves an improvement of 6.5% pascal VOC mAP, and 3.4% COCO mAP over the baseline method without intermediate training. We report the evaluation for each pre-training dataset, averaged over five models, fine-tuned for 50 epochs on our olfactory artworks datasets. Best evaluation results are highlighted in bold. The merge of two datasets $D_1$ and $D_2$ is written as $D_1 \cup D_2$.

| Pretraining Dataset | $mAP_{50}$ | mAP |
|---|---|---|
| None (Baseline) | 16.8%($\pm$1.3%) | 8.4%($\pm$0.4%) |
| OI | **23.3%($\pm$0.5%)** | **11.8%($\pm$0.4%)** |
| IA | 22.6%($\pm$1.2%) | 10.9%($\pm$0.9%) |
| PA | 21.9%($\pm$0.4%) | 10.5%($\pm$0.2%) |

pixel-wise image differences which we replace by a modified consistency loss that evaluates the semantic similarity between original and reconstructed image on an object level;

(2) As an extension to the aforementioned transfer-learning experiments, we created stylized versions of IconArt, OpenImages, and PeopleArt pretraining datasets using AdaIN, similar to the approach of [Madhu et al., 2020] take to estimate poses in ancient vases. Figure 6 gives one example from the stylized OpenImages dataset. Unfortunately, we did not observe an increase of detection accuracy for the downstream task;

(3) As part of a contrastive learning model architecture, that will be described below, we generate stylized objects from ImageNet and OmniArt [Strezoski and Worring, 2018] using AdaIN and paste them onto random background images. Learning to localize these object crops should result in image embeddings that are robust to different styles like photographs, artworks, or prints.

## 4.3 Self-Supervision

Self-supervised learning is a technique where learning algorithms are trained without supervision, typically by masking parts of the training data. The learning algorithm learns properties of the data distribution by filling in the artificial gaps as accurate as possible. In the context of computer vision, self-supervision is often applied to learn feature representations of visual data that are robust to style variations. The method is particularly well suited for artistic domain adaptation tasks since with OmniArt [Strezoski and Worring, 2018] and various museum collections there are large amounts of digitized artistic imagery available but labels for large-scale supervised learning are still lacking. To use the potential of this training strategy, we apply and modify Instance Localization for Self-Supervised Detection Pretraining (InsLoc) [Yang et al., 2021], which is specifically tailored to generate representations suited for object detection. The main idea behind InsLoc is to randomly crop patches from foreground objects and paste them onto background images at random locations and scale. The network is then trained to generate matching representations for crops on different background images that have been pasted at different locations and scales. Training a computer vision backbone with this procedure leads to feature representations that are equivariant to shifts

in scale and location, a property that is essential to object detection, but unnecessary for image classification. We extend and improve the crop and paste technique applied by InsLoc to account for the stylistic properties of the artistic domain.

### 4.4 Summary

With *transfer learning*, *style transfer*, and *self-supervision* we are exploring the possibilities of three different techniques of domain adaptation. Using transfer learning alone already gave us a considerable increase of object detection accuracy on the dedicated experiment dataset. We are currently investigating why this performance increase was not observable on the full ODOR challenge dataset. In our current experiments, we are additionally combining multiple techniques, e. g., by leveraging style transfer to improve a contrastive learning technique, or by using stylized datasets in a transfer learning setup.

## 5 Object Detection

Olfactory objects are objects that either carry a strong smell themselves or act as indication to the presence of smells (cf. D2.1). From a technical perspective, the recognition of olfactory objects is more straightforward than the other three types of olfactory references since it has been a research area in computer vision for more than 20 years and there are various well established techniques that show impressive performance on photographic benchmark datasets [Zou et al., 2019]. While the question of olfactory relevance of singular objects remains debatable in many cases, the presence and position of objects with specific olfactory functions in images might serve as the basis for answering more complex research questions as described in Section 2. This research potential of olfactory objects in conjunction with the relative technical feasibility of their detection led the image analysis team of the Odeuropa project to lay their main focus on object detection.

### 5.1 Algorithms

Object detection algorithms can broadly be classified into one-stage approaches, two-stage approaches, [Jiao et al., 2019] and, more recently, transformer-based algorithms [Carion et al., 2020]. In two-stage algorithms, candidate object regions are proposed in a first step by an object proposal network and those regions are then refined and classified in a second step. One stage algorithms, on the other hand, operate on a predefined grid on which candidate objects are simultaneously predicted and classified in different aspect ratios and sizes. Originally, one-stage algorithms were considered to have a higher inference speed whereas two-stage algorithms were assumed to have a higher accuracy [Jiao et al., 2019]. However, as the following results will show, this assumption does not necessarily hold for modern architectures. Transformer-based approaches can both be implemented as one-stage and two-stage architectures. They are characterized by the replacement of at least part of the Convolutional Neural Network (CNN) network with transformer modules [Liu et al., 2021a]. We list them as a separate category because they lately outperform convolution-based models on nearly every object detection benchmark. For each of these object detection paradigms, we trained one representative state-of-the-art method and compare the performance on the ODOR Challenge dataset in Table 3.

(1) *One Stage: PPYOLO*
PPYOLO [Long et al., 2020] is a recent version of the classical one-stage object detection algorithm YOLO [Redmon et al., 2016] which combines a modernized architecture with multiple tricks to tweak object detection performance. We applied PPYOLOE [Xu et al., 2022a], a further improved version of the algorithm which can be trained using the PaddlePaddle [Authors, 2019] framework. For comparison we trained PPYOLOE with three different backbones that differ in the number of parameters and thus in their hardware requirements

and inference speed. Quantitative results of the three models on the ODOR Challenge dataset are reported in rows 1–3 of Table 3.

(2) *Two Stage: Faster R-CNN*
Although being the oldest of the three applied object detection algorithms, Faster R-CNN [Ren et al., 2015] is still widely used. We apply a more recent Faster R-CNN with Feature Pyramid Networks (FPN) [Lin et al., 2017a], where the feature extracting backbone is complemented with lateral connections between layers that allow capturing semantic features at multiple scales. We used the icevision[1] and MMDetection [Chen et al., 2019] frameworks to train and evaluate the Faster R-CNN models. When training with icevision, we achieved a slightly improved performance for a model that has been pretrained on PeopleArt as described above in Section 4 (Faster $\text{RCNN}_{\text{PA}}$). Detection performance of the Faster R-CNN models is reported in rows 4–5 of Table 3.

(3) *Transformer-based: SWIN*
Transformers are a network architectures capable of modeling both long term and short term dependencies within structured data that have become the default architecture in natural language processing (NLP) and have shown impressive results in language understanding and generation [Lin et al., 2021], [Brown et al., 2020]. Recently, transformers backbones have been adopted in computer vision and outperform their convolutional counterparts in most vision benchmarks [Liu et al., 2021a]. We apply Shifted Window Transformer (SWIN) [Liu et al., 2021b], a vision transformer backbone that performs particularly well on dense vision tasks such as object detection and segmentation. The main idea behind SWIN is to decompose the input image into a set of image windows on which the network constituents called transformer blocks can operate under feasible hardware requirements. In subsequent stages the local windows are then gradually merged to generate a hierarchical representation of the input features. Cross window-connections between local windows are modeled by shifting the windows for the respective page at each stage of the network computation, hence the name *Shifted* Windows Transformer.

Replacing the convolutional backbones with SWIN backbones, we trained a one stage detection architecture called Retinanet [Lin et al., 2017b] ($\text{Retinanet}_{\text{SWIN}}$) and a Faster RCNN ($\text{Faster RCNN}_{\text{SWIN}}$) and report the results in line 7–9 of Table 3.

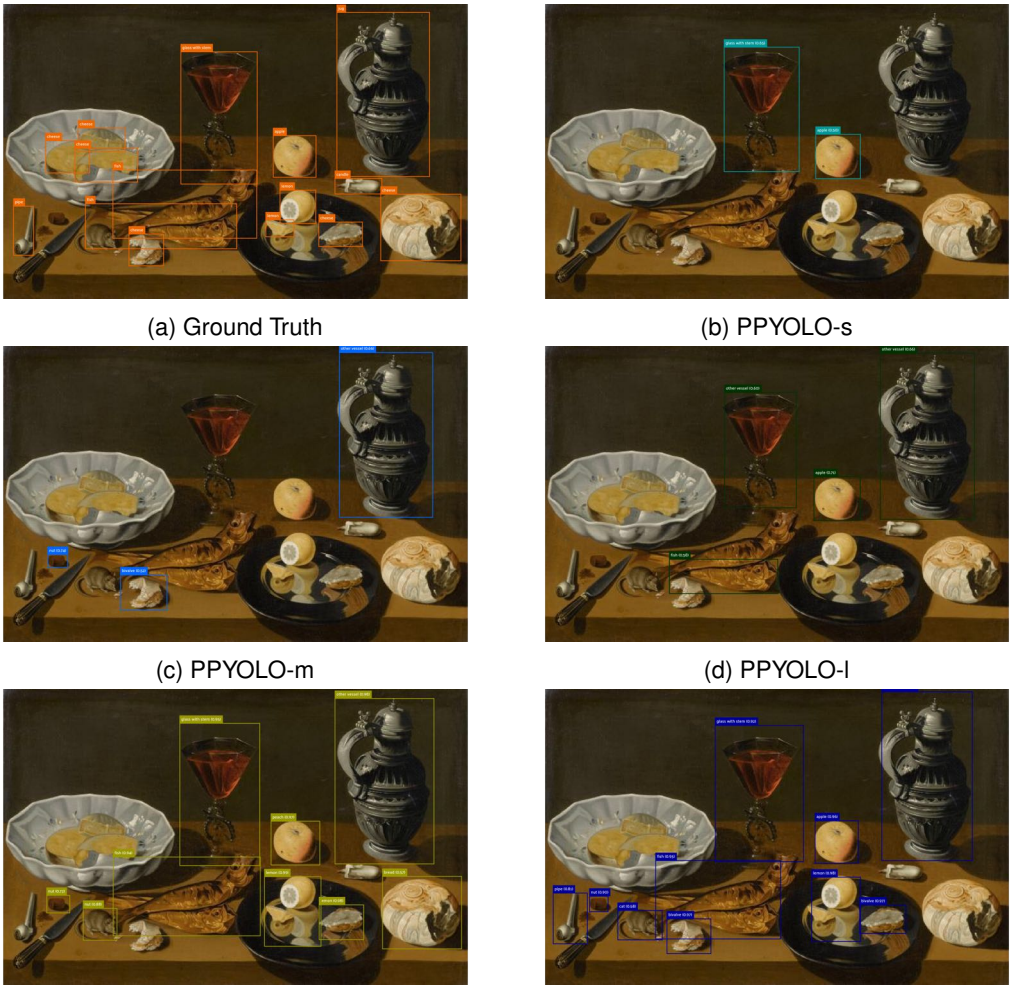
## 5.2 Summary

We observe that the Faster RCNN with with the SWIN backbone outperforms all other models by a large margin of 1.3% COCO mAP, especially on the detection of small objects. The performance of the one-stage detection algorithm PPYOLO lies between that of the Faster RCNN models with a convolutional backbone, and that of the Faster RCNN with the transformer backbone. Interestingly, plugging in a SWIN transformer in the one-stage detection algorithm Retinanet [Lin et al., 2017b] lead to a significant drop in accuracy. We are currently working on combining the better performing PPYOLO architecture with a SWIN backbone to check whether this gives us better results. In terms of object size, we observe that the faster RCNNs with convolutional backbones outperform the other models when detecting large objects but work much worse for the detection of small objects.

Figure 7 displays a still life with a variety of objects from our datasets and exemplary predictions for each of the detection algorithms we applied.

In the following, we will discuss the detection capabilities with respect to specific categories based on the evaluation of the best-performing $\text{FRCNN}_{\text{SWIN}}$ algorithm (cf. Table 4 for an overview, see also Table 6 in the appendix for the full picture). A positive surprise was the good performance on pomanders (51.6% mAP) and gloves (32.3 %), which both have a high olfactory relevance. Figure 8 shows three examples of pomander and gloves recognition. The results we observed for

[1] https://airctic.com/



(a) Ground Truth (b) PPYOLO-s

(c) PPYOLO-m (d) PPYOLO-l

(e) $\text{FRCNN}_{\text{PA}}$ (f) $\text{FRCNN}_{\text{SWIN}}$

Figure 7: Qualitative comparison of predictions for the different detection algorithms. While we see a gradual increase of detected objects with the increased model size in the one-stage detectors, both two-stage detectors recognize most of the annotated objects. The only model that captures the highly smell-relevant pipe object is the Faster RCNN with SWIN backbone.
Image credits: *Still life with herring, cheese, wine and a mouse*. Anonymous (France). 1650-1949. Oil on canvas. RKD - Netherlands Institute for Art History, RKDimages (226217).


(a) (b) (c)

Figure 8: Examples of successful pomander recognition. In the left and middle image the pomanders are detected correctly while the right image features a false positive next to the correct prediction. The left artwork additionally has a succesfull detection of gloves.
Image credits (left to right):
(a) *Portrait of an unknown woman*. Attributed to Pieter Soutman. c. 1625 – 1630. Oil on panel. RKD - Netherlands Institute for Art History, RKDimages (11254).
(b) *Portrait of an 18-year olf woman*. Attributed to Pieter Pourbous. 1574. Oil on panel. RKD - Netherlands Institute for Art History, RKDimages (280945).
(c) *Portrait of a woman, probably Maria Schurman*. Anonymous. 1599 – 1600. Oil on panel. RKD - Netherlands Institute for Art History, RKDimages (33713).

Table 3: Comparison of detection performance on the ODOR test set. We report COCO mAP, the weaker pascal VOC metric ($mAP_{50}$), and average precision for small, medium and large objects.

| | mAP | $mAP_{50}$ | $mAP_S$ | $mAP_M$ | $mAP_L$ |
|---|---|---|---|---|---|
| PPYOLOE-s | 6.9 % | 13.5 % | 3.5 % | 8.1 % | 16.0 % |
| PPYOLOE-m | 9.5 % | 18.0 % | 4.4 % | 10.4 % | 20.2 % |
| PPYOLOE-l | 9.9 % | 18.7 % | 5.0 % | 11.5 % | 19.9 % |
| Faster $RCNN_{PA}$ | 6.6 % | 13.9 % | 1.1 % | 6.6 % | 21.7 % |
| Faster $RCNN_N$ | 6.5 % | 13.7 % | 0.7 % | 6.6 % | 23.1% |
| Faster $RCNN_{SWIN}$ | 11.2% | 23.4 % | 5.8 % | 12.2% | 20.9% |
| $Retinanet_{SWIN}$ | 4.6 % | 9.2 % | 2.4 % | 5.4 % | 8.4 % |

Table 4: Class-wise COCO mAP for a selection of categories, reported for the best-performing $FRCNN_{SWIN}$ model.

| Category | mAP |
|---|---|
| pomander | 51.6% |
| gloves | 32.3 % |
| fish | 14.2 % |
| censer | 11.9 % |
| fruits (avg.) | 7.3 % |
| drinking vessel (avg.) | 5.1% |
| fire | 2.5% |
| flowers (avg.) | 2.3 % |
| smoke | 1.5% |

the detection of fish (14.2% mAP) and censers (11.9 % mAP) match roughly the average precision over all classes (11.2% mAP). Figure 9 displays possible reasons for the decrease performance compared to higher accuracy categories as pomanders and gloves:

Difficult as expected was the detection of flowers (average of 2.3% mAP) and fruits (avg. 7.3 % mAP) with their fine-grained, and often visually similar subcategories.

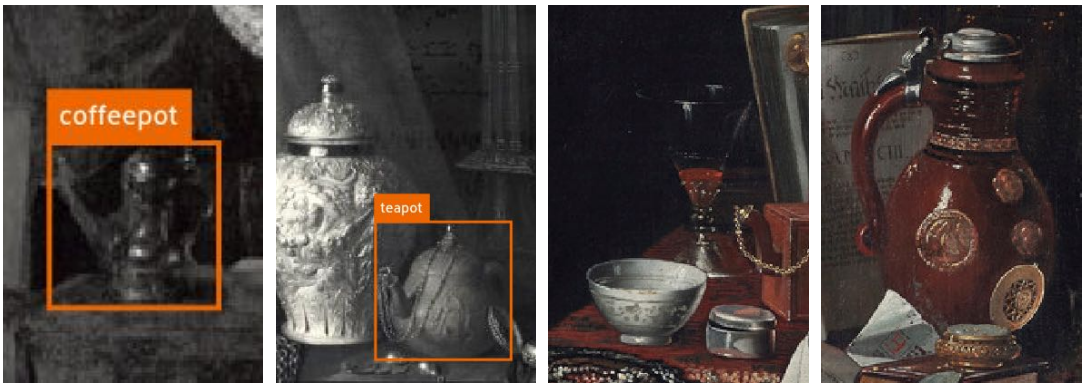
Various kinds of drinks like coffee, tea, beer or wine can have a very particular smell and thus can be of high olfactory significance. However, since it can be very hard to visually recognize the liquids in drinking vessels, we refrained from directly detecting the liquids and focused on recognizing specialized drinking vessels that may serve as a cue for their content. Figure 10 shows some examples of different drinking vessels that we identified. While this improved our ability to recognize drinks in artworks, detecting drinking vessels still remains challenging: Consider the example of tea-, and coffeepots between which the differentiation is challenging even for humans and requires a good amount of background knowledge. Figure 11 exemplifies the difficulties with drinking vessel detection. All shown predictions misinterpret the teapot as a jug, a drinking vessel that is highly represented in our training data.

# 6 Gesture Detection

For investigations into past conceptualizations and evaluations of the olfaction, smell gestures play an outstanding role. More than objects, they provide a direct gateway to cultural dimensions of smell: Olfactory gestures like covering the nose in reaction to a smell are overt manifestations of historic conceptualisations of smell. The olfactory objects that provoke the gestures will also be of interest for object detection. However, their recognition is considerably more challenging than


(a) Fish (b) Fish

Figure 9: Two detections of fish. In picture (a), the fish is detected, but the overlap between ground truth bounding box and detected bounding box is relatively low. Picture (b) has two overlapping fish which partially occlude each other. The model is only able to detect one of the occluding fishes.
Image credits:
(a) *Still life with herring and a teapot on a marble ledge*. Jacob van Walscapelle. c. 1675. Oil on canvas. RKD - Netherlands Institute for Art History, RKDimages (201923).
(b) *Boy holding a pewter tankard, by a still life of a duck, cheeses, bread and a herring*. Circle of Gerard van Honthorst. 1625 – 1674. Oil on canvas. RKD - Netherlands Institute for Art History, RKDimages (287165).




(a) Coffeepot (b) Teapot (c) Cup and glass with stem (d) Jug

Figure 10: Exemples of drinking vessels from our dataset. Details from (left to right):
(a) *Interior with a a company smoking, drinking and playing cards.*. Quiring van Brekelenkam. c.1661. Oil on panel. RKD - Netherlands Institute for Art History, RKDimages (248041).
(b) *Vanitas still life with ornamental vessels and a globe*. Pieter van Roestraeten. After c.1665. Oil on canvas. RKD - Netherlands Institute for Art History, RKDimages (283016).
(c & d) *Vanitas still life with books, documents and other objects*. Pseudo-Roestraten. Late 17th century. Oil on canvas. RKD - Netherlands Institute for Art History, RKDimages (238595).


(a) Ground truth (b) $\text{FRCNN}_{\text{ResNet}}$ (c) PPYOLO-L (d) PPYOLO-M (e) $\text{FRCNN}_{\text{SWIN}}$

Figure 11: Detections of a candle and a drinking vessel by three of our detection systems. The faster RCNN with convolutional backbone, and both PPYOLO models misinterpret the teapot as a jug while the candle is correctly detected by both PPYOLO models, and the fire is detected by the Faster RCNN. The faster RCNN with SWIN backbone detects the candle and classifies the teapot as 'other vessel'. The remaining models did not produce any detections.
Image credit: Detail from *Stil life of books, letters, a watch and other objects on a table draped with a carpet*. Pseudo-Roestraten. 1675 – 1725. Oil on panel. RKD - Netherlands Institute for Art History, RKDimages (186464).

Figure 12: Example of an alleged smelling gesture. The hand bringing the fruits to the nose could be interpreted as sniffing or as a mere coincidence. If it is interpreted as a sniffing gesture, the exact localization remains debatable. Print: Der Geruch. Johann Elias Ridinger. 1717/1760. Herzog Anton Ulrich-Museum, Braunschweig. CC BY-NC-ND 4.0

those of olfactory objects for three reasons:

(1) Unlike objects, gestures do not have clear boundaries. Their localization in an images is very subjective and different people might consider different visual elements part of a gesture. Fig. 12 shows an example of a gesture that has been annotated as sniffing. Whether the hand that brings an object to the nose to sniff belongs to the gesture or not is one example of an element of subjectivity in gesture localization;

(2) Not only the localization but also the identification of smell gestures can be subjective. What might be considered a smell gesture by one person, could be perceived as an arbitrary movement by another. The interpretation of a body pose as a specific gesture requires considerably more context information than the recognition of an object (cf. Fig. 12);

(3) It is difficult to find enough depictions of smell gestures to train and evaluate recognition systems because smells only play a minor role in existing museum collection metadata [Ehrich et al., 2021]. This restriction lead us to start our experiments with two smell gestures only, i. e., *sniffing* and *holding the nose*. For all the other smell gestures that are of interest, we simply could not find enough training and evaluation samples to reasonably start an automatic recognition effort.

As a naive solution, we experimented with detecting olfactory gestures in as similary way as to detecting objects and annotated them with bounding boxes. The results are reported in Table 5. As expected due to the lack of clear spatial boundaries of smell gestures, the results are not good. To improve on this, we are currently experimenting with multiple approaches:

**Pose Estimation** If we assume that smelling gestures are associated with characteristic body postures like an arm that is bent towards the nose, we can use estimations of body poses to predict whether a depicted person is performing a smell gesture or not. To obtain quantifiable estimations of body poses, there are plenty of *pose estimation* algorithms available. These algorithms predict a number of keypoints that belong to specific body parts and in conjunction, define a depicted person's pose. Typical keypoints localize the position of joints, and facial features. We apply a ViTPose [Xu et al., 2022b], a state-of-the-art pose estimation algorithm

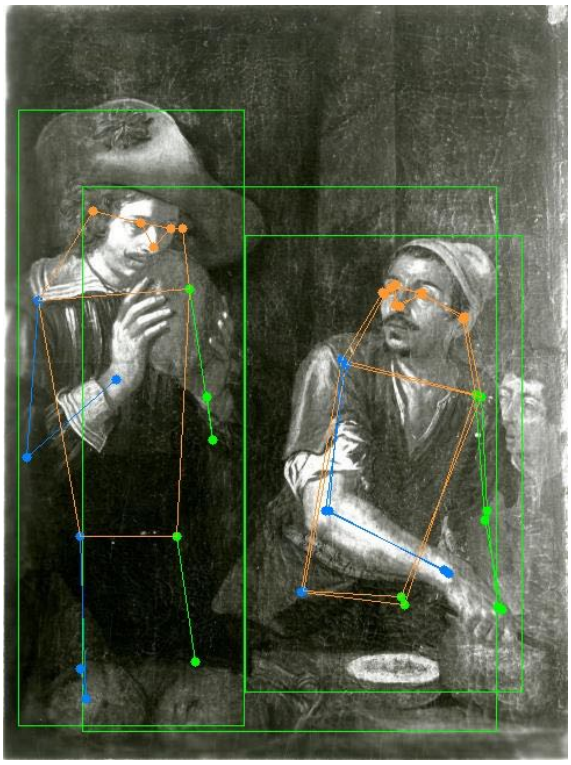


Figure 13: Person detection boxes and pose estimation keypoints on a person sniffing on a watermelon. Keypoints and boxes generated by Azhar Hussian using ViTPose [Xu et al., 2022b]. Image credit: *The sense of smell (one of the five senses)*. Joannes van Houbracken. 1615–1665. Oil on canvas. RKD - Netherlands Institute for Art History, RKDimages (244412).

Table 5: Initial results for the detection of smell gestures via pose estimation.

| Gesture | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| No Gesture | 74 % | 76% | 75% | 70 |
| Holding the Nose | 44 % | 42% | 43 % | 26 |
| Sniffing | 75% | 71% | 73% | 17 |

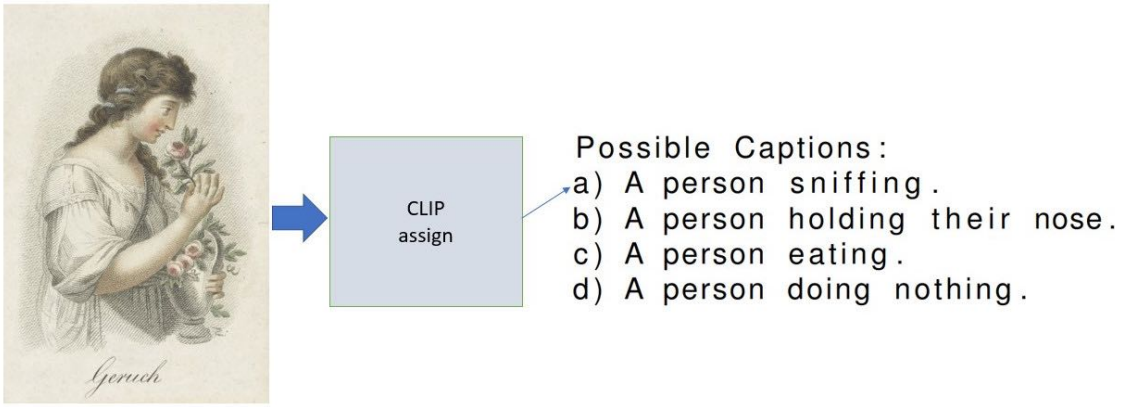



Figure 14: Proposed application of CLIP to classify smelling gestures

that produces accurate results, even on artworks. The keypoints we extract will then be fed to a classification algorithm, possibly in conjunction with other features like the position of extracted objects such as flowers to determine whether a depicted person is performing a smell gesture. Table 5 lists initial results for the classification of Sniffing and Holding the Nose in cropped persons from our collected artworks.

**Multimodal detection** Gestures can be considered more 'semantic' than objects since instead of solely relying on visual features, their recognition often requires an understanding of the narrative of an artwork. This consideration suggests the application of multi-modal image understanding models that aim at a textual description of the semantic content of an image instead of the exact localization of specific objects. One model that has been extremely successful in this regard is CLIP [Radford et al., 2021] which has been pre-trained to predict the correct matching between images and captions on a massive dataset of 400 Million image-text pairs. We can leverage CLIPs language and visual understanding capabilities by generating a set of optional captions for the model to choose from as illustrated in Fig. 14.

# 7 Conclusions & Outlook

Automatic recognition of visual smell references can be a key requisite to enable empirically backed research approaches that complement classical historical and art historical methodology. However, their recognition bears challenges due to the visual properties of their artistic representation and the particularity of some of the smell-related phenomena. To overcome these challenges, we have experimented with multiple different techniques of domain adaptation that have proven their capabilities to bridge the gap between large-scale photographic datasets with modern categories and the historical artworks we are interested in. In a prior deliverable (D2.1) we have created a taxonomy of visual smell references that consists of olfactory objects, smell gestures, fragrant spaces, and olfactory iconography.

Up to now, the bulk of our efforts has been directed on the detection of olfactory objects, which

then might serve as a basis for the recognition of more complex types of smell references. In olfactory object detection, we have experimented with multiple different detection paradigms of which we found two-stage object detection backed with a transformer backbone to be the most effective. In our detection models, we observe a large variety of detection accuracy depending on the object of interest. Fortunately, some objects with a high olfactory significance like pomanders and gloves exhibit a good detection accuracy. Other smell-relevant objects like specific flowers or smoke on the other hand can not yet be reliably detected and distinguished. By continuously increasing the size and quality of our fine-tuning dataset, and by further improving our detection algorithms we will work on increasing the performance on these difficult categories.

Apart from object detection, we have also conducted initial experiments with the recognition of smell gestures, specifically the gestures of *sniffing* and *holding the nose*. Initial results for the recognition based on body pose models are encouraging and motivate us to further work in this direction.

Future lines of research to improve of the automatic smell reference recognition are, apart from continuing the already experiments, the implementation of a system to automatically recognize fragrant spaces on image-level, and the development of multi-modal models that enable us to additionally leverage textual descriptions that often exist in the image meta-data.

# References

[Arnold and Tilton, 2019] Arnold, T. and Tilton, L. (2019). Distant viewing: analyzing large visual corpora. *Digital Scholarship in the Humanities*, 34(Supplement_1):i3–i16.

[Authors, 2019] Authors, P. (2019). Paddledetection, object detection and instance segmentation toolkit based on paddlepaddle. https://github.com/PaddlePaddle/PaddleDetection.

[Brown et al., 2020] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

[Carion et al., 2020] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. (2020). End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer.

[Chen et al., 2019] Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., Zhang, Z., Cheng, D., Zhu, C., Cheng, T., Zhao, Q., Li, B., Lu, X., Zhu, R., Wu, Y., Dai, J., Wang, J., Shi, J., Ouyang, W., Loy, C. C., and Lin, D. (2019). MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*.

[Ehrich et al., 2021] Ehrich, S. C., Verbeek, C., Zinnen, M., Marx, L., Bembibre, C., and Leemans, I. (2021). Nose first: Towards an olfactory gaze for digital art history. *First International Workshop on Multisensory Data and Knowledge*. Online accessed, December 09, 2021.

[Everingham et al., 2010] Everingham, M., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338.

[Farahani et al., 2021] Farahani, A., Voghoei, S., Rasheed, K., and Arabnia, H. R. (2021). A brief review of domain adaptation. *Advances in Data Science and Information Engineering*, pages 877–894.

[Gonthier et al., 2018] Gonthier, N., Gousseau, Y., Ladjal, S., and Bonfait, O. (2018). Weakly supervised object detection in artworks. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0.

[Goodfellow et al., 2014] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27.

[Hall et al., 2015] Hall, P., Cai, H., Wu, Q., and Corradi, T. (2015). Cross-depiction problem: Recognition and synthesis of photographs and artwork. *Computational Visual Media*, 1(2):91–103.

[Huang and Belongie, 2017] Huang, X. and Belongie, S. (2017). Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510.

[Isola et al., 2017] Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134.

[Jiao et al., 2019] Jiao, L., Zhang, F., Liu, F., Yang, S., Li, L., Feng, Z., and Qu, R. (2019). A survey of deep learning-based object detection. *IEEE access*, 7:128837–128868.

[Kuznetsova et al., 2020] Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al. (2020). The open images dataset v4. *International Journal of Computer Vision*, 128(7):1956–1981.

[Lin et al., 2021] Lin, T., Wang, Y., Liu, X., and Qiu, X. (2021). A survey of transformers. *arXiv preprint arXiv:2106.04554*.

[Lin et al., 2017a] Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. (2017a). Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125.

[Lin et al., 2017b] Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017b). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

[Lin et al., 2014] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

[Liu et al., 2021a] Liu, Y., Zhang, Y., Wang, Y., Hou, F., Yuan, J., Tian, J., Zhang, Y., Shi, Z., Fan, J., and He, Z. (2021a). A survey of visual transformers. *arXiv preprint arXiv:2111.06091*.

[Liu et al., 2021b] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021b). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022.

[Long et al., 2020] Long, X., Deng, K., Wang, G., Zhang, Y., Dang, Q., Gao, Y., Shen, H., Ren, J., Han, S., Ding, E., et al. (2020). Pp-yolo: An effective and efficient implementation of object detector. *arXiv preprint arXiv:2007.12099*.

[Madhu et al., 2020] Madhu, P., Villar-Corrales, A., Kosti, R., Bendschus, T., Reinhardt, C., Bell, P., Maier, A., and Christlein, V. (2020). Enhancing human pose estimation in ancient vase paintings via perceptually-grounded style transfer learning. *arXiv preprint arXiv:2012.05616*.

[Marinescu et al., 2020] Marinescu, M.-C., Reshetnikov, A., and López, J. M. (2020). Improving object detection in paintings based on time contexts. In *2020 International Conference on Data Mining Workshops (ICDMW)*, pages 926–932. IEEE.

[Pan and Yang, 2009] Pan, S. J. and Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

[Radford et al., 2021] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

[Redmon et al., 2016] Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788.

[Ren et al., 2015] Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 201.

[Russakovsky et al., 2015] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252.

[Strezoski and Worring, 2018] Strezoski, G. and Worring, M. (2018). Omniart: a large-scale artistic benchmark. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 14(4):1–21.

[Westlake et al., 2016] Westlake, N., Cai, H., and Hall, P. (2016). Detecting people in artwork with cnns. In *European Conference on Computer Vision*, pages 825–841. Springer.

[Xu et al., 2022a] Xu, S., Wang, X., Lv, W., Chang, Q., Cui, C., Deng, K., Wang, G., Dang, Q., Wei, S., Du, Y., et al. (2022a). Pp-yoloe: An evolved version of yolo. *arXiv preprint arXiv:2203.16250*.

[Xu et al., 2022b] Xu, Y., Zhang, J., Zhang, Q., and Tao, D. (2022b). Vitpose: Simple vision transformer baselines for human pose estimation. *arXiv preprint arXiv:2204.12484*.

[Yang et al., 2021] Yang, C., Wu, Z., Zhou, B., and Lin, S. (2021). Instance localization for self-supervised detection pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3987–3996.

[Zhu et al., 2017] Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232.

[Zhuang et al., 2020] Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., and He, Q. (2020). A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76.

[Zinnen, 2021] Zinnen, M. (2021). How to see smells: Extracting olfactory references from artworks. In *Companion Proceedings of the Web Conference 2021*, pages 725–726.

[Zinnen et al., 2022] Zinnen, M., Madhu, P., Bell, P., Maier, A., and Christlein, V. (2022). Transfer learning for olfactory object detection. In *Digital Humanities Conference*.

[Zou et al., 2019] Zou, Z., Shi, Z., Guo, Y., and Ye, J. (2019). Object detection in 20 years: A survey. *arXiv preprint arXiv:1905.05055*.

# A Object Detection Accuracies

| Category | mAP |
|---|---|
| anemone | 0.7% |
| carnation | 7.0% |
| columbine | 1.2% |
| cornflower | 3.5% |
| daffodil | 0.2% |
| geranium | 1.0% |
| heliotrope | 0.0% |
| hyacinth | 0.0% |
| iris | 1.9% |
| jasmine | 0.6% |
| lavender | 0.0% |
| lilac | 0.0% |
| lily | 0.4% |
| lily of the valley | 0.2% |
| neroli | 0.0% |
| petunia | 0.0% |
| poppy | 2.1% |
| rose | 11.3% |
| tulip | 15.1% |
| violet | 1.8% |
| other flower | 0.8% |
| other fruit | 0.6% |
| apple | 15.4% |
| cherry | 4.3% |
| peach | 16.1% |
| currant | 0.0% |
| fig | 2.6% |
| grapes | 4.1% |
| lemon | 14.1% |
| melon | 12.4% |
| pear | 14.4% |
| plum | 2.8% |
| strawberry | 0.9% |
| artichoke | 6.2% |
| carrot | 0.0% |
| garlic | 3.6% |
| mushroom | 12.4% |
| olive | 1.6% |
| onion | 3.3% |
| pumpkin | 4.8% |
| other vessel | 1.7% |
| glass with stem | 17.0% |
| glass without stem | 10.2% |
| jug | 16.7% |
| cup | 0.0% |
| chalice | 0.0% |
| wine bottle | 2.4% |
| carafe | 0.0% |
| coffeepot | 2.9% |

| | |
|---|---|
| teapot | 0.0% |
| other vertebrate | 0.0% |
| animal carcass | 11.5% |
| bird | 9.8% |
| cat | 10.0% |
| cow | 32.5% |
| dog | 31.6% |
| donkey | 15.6% |
| fish | 14.2% |
| goat | 20.3% |
| horse | 16.7% |
| pig | 16.9% |
| sheep | 27.7% |
| whale | 31.9% |
| other invertebrate | 3.0% |
| bivalve | 23.2% |
| butterfly | 34.7% |
| caterpillar | 4.4% |
| fly | 2.8% |
| lobster | 31.2% |
| prawn | 11.8% |
| bug | 0.0% |
| bracelet | 13.7% |
| pomander | 51.6% |
| ring | 6.4% |
| ashtray | 24.7% |
| bread | 29.4% |
| candle | 1.9% |
| censer | 11.9% |
| cheese | 27.7% |
| fire | 2.5% |
| gloves | 32.3% |
| meat | 6.6% |
| nut | 11.5% |
| pipe | 19.3% |
| smoke | 1.6% |

Table 6: Per-Class COCO mAP metrics for all classes of the dataset.