

# Prompt me a Dataset: An investigation of text-image prompting for historical image dataset creation using foundation models

Hassan El-Hajj<sup>1,2</sup>[0000–0001–6931–7709] and Matteo Valleriani<sup>1,2,3,4</sup>[0000–0002–0406–7777]

- <sup>1</sup> Max Planck Institute for the History of Science, Boltzmannstr. 22, Berlin, 14195, Germany
- <sup>2</sup> BIFOLD – Berlin Institute for the Foundations of Learning and Data, Berlin, 10587, Germany
- <sup>3</sup> The Cohn Institute for the History and Philosophy of Science and Ideas, Faculty of Humanities, Tel-Aviv University, Tel-Aviv, 6997801, Israel
- <sup>4</sup> Institute of History and Philosophy of Science, Technology, and Literature, Faculty I, Technische Universität Berlin, Straße des 17. Juni 135, Berlin, 10623, Germany  
`{hhajj,valleriani}@mpiwg-berlin.mpg.de`

**Abstract.** In this paper, we present a pipeline for image extraction from historical documents using foundation models, and evaluate text-image prompts and their effectiveness on humanities datasets of varying levels of complexity. The motivation for this approach stems from the high interest of historians in visual elements printed alongside historical texts on the one hand, and from the relative lack of well-annotated datasets within the humanities when compared to other domains. We propose a sequential approach that relies on GroundDINO and Meta’s Segment-Anything-Model (SAM) to retrieve a significant portion of visual data from historical documents that can then be used for downstream development tasks and dataset creation, as well as evaluate the effect of different linguistic prompts on the resulting detections.

**Keywords:** SAM · GroundingDINO · Digital Humanities · Dataset Creation · Historical Documents · Text Prompts

## 1 Introduction

Technological advancements of the last decades have led to major digitization efforts focused on historical documents, such as the Google Book Search (GBS) and Open Content Alliance (OCA) [12]. This rapid growth of digitized historical documents has paved the way for computational historical document analysis, allowing researchers to comb through large number of documents and test hypotheses at scale.

With the advent of neural networks, new methods of image analysis and information extraction came to light. However, these methods are data intensive, and require a large amount of annotated and curated datasets in order to be

trained. The lacuna of such datasets pushed the digital humanities community towards collecting and publishing annotated and curated datasets to facilitate the training of state-of-the-art models. However, given the heterogeneous nature of historical data, and the high degree of inter –and intra– domain variability, such datasets often cover very specific historical topics and domains, with limited generalization possibilities.

In this paper, we propose a pipeline for information extraction from historical documents using image foundation models to support the work of historians. We discuss the current state of research for information extraction pipelines within the humanities in Section 2. In Section 3, we discuss our current pipeline as well as the current experience within the Max Planck Institute for the History of Science, evaluate it on three datasets in Section 4, and conclude with an overview of possible extensions to the proposed pipeline in Section 5.

## 2 Current State of the Research

Information Extraction (IE), including image (e.g., visual elements) extraction, from historical documents is playing an increasingly important role in formulating historical hypotheses [23], allowing researchers to tap into a large pool of information that would have been impossible to assemble without computational methods. Meanwhile, there have been many advances in text processing with regard to both printed and handwritten sources [10,7,9]. In this paper, we tackle the well-addressed technical problem of image extraction from historical documents while relying on foundation models and text prompts.

Current approaches to extract images from historical texts can be divided into two main groups: Segmentation and Object Detection approaches. Segmentation approaches often rely on FCN architectures such as U-Net [20] or Mask-RCNN [11] to generate masks of the desired image region. One of these approaches is the one proposed by [16] to extract images from a wide range of historical texts using a modified U-Net. Another similar approach is proposed by [7] to segment text lines in handwritten historical documents. Numerous further approaches have treated information extraction and, more specifically, image extraction as an object detection problem and tackled it with models such as EfficientDet [22], for instance in [8], where the authors extracted images from a corpus of Scottish Chapbooks. [4] used instead YOLOv5 [19] to extract different classes of visual elements from a large corpus of early modern books.

While the above-mentioned approaches are far from representing a comprehensive review of the current state of image extraction from historical documents, they highlight general trends within the community. Despite their differences, these approaches share a fundamental feature, namely that they were all trained on carefully annotated datasets. This is notable because, in contrast to other industry domains, *annotated* data within the humanities remains relatively scarce due to numerous reasons including a lack of expertise (compared to the difficulty of defining classes within heterogeneous data, as well as ambiguous data interpretations to name a few). Many of the approaches discussed above provide their

own datasets, such as the Synthetic *SynDoc* dataset presented in [16], the Chapbook dataset presented by [8], as well as the S-VED [5] presented by [4,5]. The amount of humanities and historical document datasets is continuously growing with numerous datasets covering different aspects of these fields [17]; this is also manifested by the growing number of datasets published on Hugging-Face’s BigLAM: Big-Science, Libraries, Archives and Museums group [1].

Despite the consistently growing number of datasets, the high level of heterogeneity of historical documents means that many of these datasets cover a small, often *niche-like*, group of target documents (e.g., Figure 1). This essentially means that the image extraction models – whether segmentation or object detection – often perform very well on in-domain data, but suffer from high performance degradation on out-of-domain data. This is clearly shown in the results presented in [4], where the performance of the YOLOv5 model trained on S-VED [5] (a dataset containing visual elements from early modern books on astronomy) degrades on out-of-domain datasets, such as Mandragore [2], a dataset consisting of diverse manuscripts from the Bibliothèque Nationale de France (BnF), or RASM [18], a dataset of historical Arabic manuscripts.



**Fig. 1.** Images of diverse types and styles. (Left to Right) A diagram from the S-VED dataset [4]. A colored image from the IlluHistDoc dataset [16]. An image from the Chapbook dataset [8]. A miniature from the HORAE dataset [3]

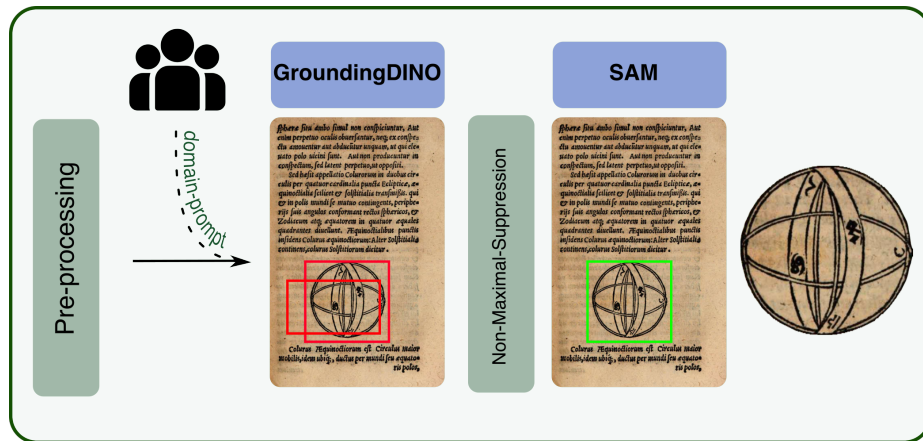
While these target-specific models are dependent on the presence of well-curated domain datasets, new foundation models are being developed, trained on immensely large datasets and able to perform **zero-shot** inference, which means that these models are able to perform well on out-of-domain images without requiring extra training.

### 3 Pipeline

While stand-alone models perform excellently on in-domain data, we aim to leverage *image* foundation models to help humanities researchers extract visual elements from their datasets as an end goal and, more importantly, quickly generate image datasets from broad data sources without requiring in-domain training.

This pipeline relies heavily on prompting these foundation models to achieve the best possible image extraction results without the use of domain-specific data. In this case, we chain a GroundingDINO [15] model with a Segment-Anything-Model (SAM) [13] to create a pipeline to generate visual element region masks from historical manuscripts (see Figure 2).

GroundingDINO is a model that relies on a Transformer-based end-to-end object detection DINO (DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection) [24], and fuses it with a Text-Encoder in order to detect objects based on human language input on open-domain data, achieving good zero-shot results on the COCO dataset [14]. GroundingDINO takes an image-text input pair, and returns a bounding box that corresponds to the image region that in turns semantically corresponds to its textual counterpart. These bounding boxes are then passed on as data prompts to SAM [13] in order to segment the desired semantically relevant object.



**Fig. 2.** Workflow from out-of-domain data entry on the left towards data extraction as bounding box with GroundingDINO [24] and as masks with SAM [13]

One obvious downside of such models is that despite the fact that they are trained on very large datasets (e.g., SA-1B dataset released by Meta contains 11 million images with 1 billion segmentation masks [13]), they often overlook humanities or historically oriented data, excluding data classes needed for manuscript and historical document information extraction. One of the major causes of the current status concerning data is the relatively low number of annotated manuscript and historical text data, as well as the difficulty in retrieving the domain knowledge required to annotate these images (e.g., think of the difference between different classes of images within the same manuscript or printed book). This situation makes it difficult to accommodate or create



such data using Mechanical Turk<sup>5</sup> workers with little to no historical domain knowledge. To circumvent these shortcomings, we propose utilizing targeted domain-aware prompts that can hone in on the desired objects, and fine-tuning GroundingDINO as part of future developments, as discussed in Section 5.

The proposed pipeline is composed of three blocks: A pre-processing block, an object detection block relying on Prompt engineering GroundingDINO, and a finer segmentation block relying on SAM. The pre-processing block resizes each image to a standard size of 1000x1000 px, and includes an autocontrast step with a 2nd and 98th percentile cutoff. These images are then passed on to the GroundingDINO module with engineered prompts. These prompts are designed in a way to inject domain knowledge while remaining general enough so that the Feature Enhancer block of the GroundingDINO model is able to fuse text and image features in an efficient way to return reliable results. Examples of these engineered prompts are shown in Section 4.

With the multiple prompt classes, multiple bounding box detections are expected. We thus add a Non-Maximal Suppression module that operates on the selected prompt group classes to ensure that each object is detected once. The cleaned results, i.e., the bounding boxes, are then passed on as box-prompts to the SAM block to return clean segmentation masks of the desired regions.

We acknowledge that this pipeline relies on two very large models and might not be efficient to run in production. However, we believe that this approach can drastically increase the amount of data at the disposal of humanities researchers, and allows them to create large datasets using language prompts. For production scenarios or domain-specific requirements, the proposed approach can be used for an initial data collection phase in preparation for the training of bespoke object detector or segmentation models.

## 4 Text-Image Prompt Evaluation

We conduct a preliminary evaluation of the pipeline above on subsets of the S-VED [5], Chapbook [8], and the HORAE datasets [3] and report the preliminary results below. These datasets are object detection datasets in historical documents dating from the 15–17th, 17–19th, and the 14th–16th centuries respectively. The S-VED dataset contains four semantically different classes, the most abundant being Content Illustrations which covers visual elements within the body of the text and intend to enrich it. Other classes include Initials, which represent often decorated letters (or drop cap) at the beginning of chapters and paragraphs, Decorations which represents small decorative elements on pages, and Printer’s Marks, which represent the emblem of the printer(s) who produced the book in question [4]. The Chapbook dataset consists of a single image class representing every image within a text page [8] while the pages of the HORAE dataset have the most detailed annotation scheme [3]. These cover Miniatures,

<sup>5</sup> Mechanical Turk is an Amazon based marketplace platform where organizations can hire workers, often for relatively low wage, to conduct some low-level work. This service is often used to annotate images and create large datasets.

which are illustrations embedded in the text, Decorations which are elements often placed along the page borders, as well as different types of Initials, such as simple initials (initials differing from the main body of the text in ink and size), decorated initials (initials with purely ornamental decoration style), and historiated initials (initials whose decoration depicts an iconographic element such as a scene or a character) [3].

Beyond the difference in classes, these three datasets represent different types of content which contain different styles of visual elements. The S-VED dataset derives from the Sphere Corpus<sup>6</sup>, which contains scientific books on geocentric astronomy used in pedagogical settings; the Chapbooks were booklets containing popular content ranging from literature, poems, religious texts, and riddles; and the HORAE dataset contains pages from the books of hours, which were a type of handwritten prayer book owned and widely circulated in the late middle ages. The difference between these types of primary sources is naturally reflected in the types of images they contain, with the S-VED containing a large number of orbital diagrams and geometric drawings, the Chapbooks dataset containing a wide range of daily life drawings featuring humans, animals, and in some cases abstract and stylized figures, and the HORAE containing a large amount of decorative elements placed around the textual area of the page, as well as a lot of religious illustrations (Figure 1).

In our attempt to evaluate the pipeline on the two models, we set the text and image thresholds to 0.35 within the GroundingDINO parameters and perform non-maximal suppression on the output boxes. We also cast all classes of the S-VED into a single visual element class in order to obtain a comparable result between the three chosen datasets. We evaluate the Average Precision (AP) [6] of different language prompts in order to examine their effect on the model’s ability to extract the needed information on such as out-of-domain data.

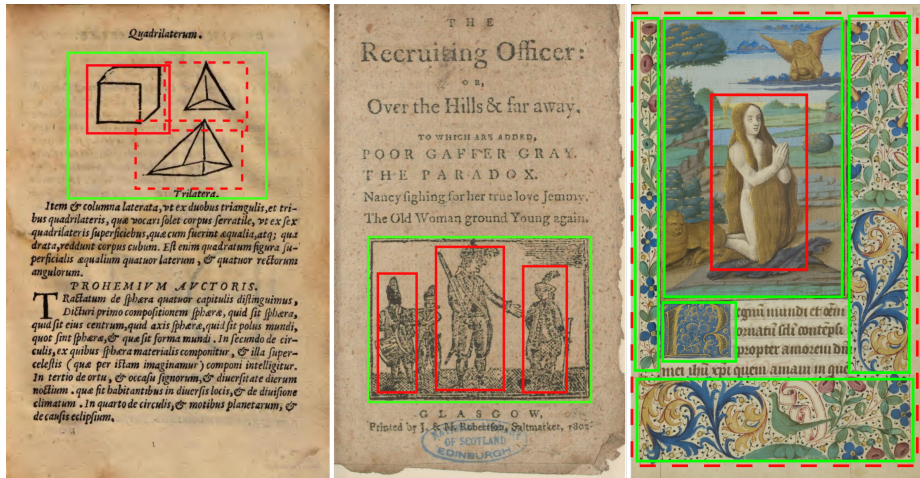
The first language prompt that we applied uses simple language prompts (i.e., single words) to try to extract the visual elements from both datasets. In this case, the prompt is constituted of the single word **{figure}**, which resulted in very good semantically meaningful results on the S-VED dataset, but appears to perform poorly on both the Chapbook and the HORAE dataset (see Table 1).

We investigate the detection and segmentation results from our pipeline in an effort to improve our prompts and retrieve a larger amount of visual elements. In the S-VED, the error sources were manifold. The first consists of missing small visual elements placed in the marginalia; the second concerns missing abstract geometric shapes that the model did not deem to be a fit for the given textual prompts. However, the highest contributor to the relatively modest AP score reported in Table 1 is the difference between the bounding boxes that our pipeline considered to be representing a figure, and the bounding boxes created by the annotators of the S-VED, which is highly abstract. A simple example is the presence of three different drawings on an S-VED representing semantically related topics, and thus annotated as a single image by the S-VED annotators. However,

<sup>6</sup> <https://sphaera.mpiwg-berlin.mpg.de/>

relying solely on the image and the text prompt, our pipeline returns multiple smaller bounding boxes with low Intersection-over-Union scores, leading to False Positive results (see Figure 3).

In the case of the Chapbook and HORAE datasets, the main cause of this error was a semantic mismatch between our prompt and the desired outcome. The suggested prompt of **{figure}** has led the GroundingDINO module to return bounding boxes of human figures within the visual elements, instead of the desired output of a figure in the literal sense (see Figure 3). Thus the low score was the result of the identification and segmentation of parts of the visual elements showing a human figure, diverging from the annotated ground truth data. This simple example proved interesting and highlights the multifaceted meaning that a single word prompt could have, and how it could affect the results (see Figure 3).



**Fig. 3.** (Left) A page from the S-VED dataset showing multiple detected regions. The solid red line represents the results obtained with the prompt **{figure}**, the dashed red lines represent the two extra boxes detected with the prompt **{figure - diagram - geometry - sketch}**, while the green box represent the ground truth data. (Center) A Chapbook page with multiple region predictions from a **{figure}** prompt in red, and a region prediction given a prompt **{image - square - rectangle - photo}**, in green corresponding to the ground truth. (Right) A HORAE page with ground truth bounding boxes in green showing a miniature, initial, and three decoration boxes around the page. The solid red box shows the prediction from a **{figure}** prompts, while the dashed red line shows the prediction from a **{floral - rectangle - flower - decorative - abstract}** prompt.

To inject more domain-knowledge into our prompts, we provide dataset-tailored prompts. For S-VED, we provide a textual prompt that better describes the content of the majority of its visual elements: **{figure - diagram - geom-**

**etry - sketch}**. In the Chapbook dataset, we focus on identifying the complete visual element, which is often square or rectangular in shape, thus the prompt in this case is **{image - square - rectangle - photo}**, while in the HORAE dataset, where we aim to detect miniatures, we provide a prompt that aims to describe their content based on our observations, which in this case is **{figure - lanscape - scene - square}**. We re-evaluate the pipeline with the aforementioned linguistic prompts, and notice a small increase in performance for the S-VED, largely due to the detection of some previously missed geometric shapes, and a large increase in performance on the Chapbook and HORAE datasets due to the fact that the new textual prompt aligns with the ground truth annotation scheme.

In order to better probe the limits of text-image pairings on a very specific dataset such as historical documents, we attempt to differentiate between the different classes of the S-VED and HORAE dataset. In the case of the S-VED, we focus our attention on differentiating between the Content Illustration and Initial classes, the most abundant classes in the S-VED. In the HORAE dataset we focus on differentiating between the Initial, Decoration, and Miniature classes. In the first case, We prompt the following in order to retrieve the S-VED Initials class **{dropcap - decorated letter - large letter}** and the following for the Content Illustration class **{figure - diagram - circle - planets}**. To differentiate between the HORAE classes, we utilize the same S-VED prompt for the Initials class, and use the following **{floral - rectangle - flower - decorative - abstract}** and **{scene - landscape - square}** for the Decoration and Miniatures classes respectively. However in the above cases, we see that we have possibly reached the limit of the pre-trained model’s text-image understanding, which is likely hindered by our efforts to differentiate the classes using very generalized terms. In the S-VED case, this is noticeable for example when an Initial such as “O” or “D” is classified as Content Illustration with high confidence due to its circular characteristic. In the HORAE examples, we encountered the same issues with the Initials class; but also faced some problems detecting the decorative elements according to the annotation scheme which divides the decorative elements according to their orientation in the manuscript pages. This meant that while our pipeline often recognizes decorative elements in the page, the detection box does not recognize the distinct decorative elements as per the annotation scheme, resulting in poor performance (see Figure 3 for a clear comparison between the annotation scheme and the detected areas). Such mishaps ultimately resulted in almost random class detections (AP scores of 0.1, 0.08, and 0.12 for the S-VED Initials, HORAE Initials, and Decorative elements respectively), and proved to be an inefficient avenue.

The results presented above are, despite their limitations, very promising, especially for researchers aiming to collect large image datasets from archival material at scale. It is clear that such models soon hit their limits when it comes to differentiating between image classes that might be of interest for historical research (e.g., Initials, Content Illustrations, and Decorations). However, the possibility of quickly collecting thousands of images from historical documents

Dataset	Prompt	AP
S-VED	{figure}	0.42
S-VED	{figure - diagram - geometry - sketch}	<b>0.51</b>
Chapbook	{figure}	0.19
Chapbook	{image - square - rectangle - photo}	<b>0.82</b>
HORAE	{figure}	0.15
HORAE	{figure - lanscape - scene - square }	<b>0.74</b>

**Table 1.** Average Precision score for object detection on a subset of S-VED, Chapbook, and HORAE datasets

using descriptive language remains enticing, and will increase the efficiency of data collection, which is often a major barrier to applying ML algorithms in the frame of historical research. In this case, the scholars could invest human power on fine-tuning the retrieved data and creating well-curated sub-classes, which can then power the training of an in-domain model.

#### 4.1 A note on the environment

As we are living in a climate-critical era, it is imperative that we take environmentally conscious choices when dealing with computational data at scale. In this case, we acknowledge that the use of both GroundingDINO and SAM comes at a high computational cost. Although these models have zero-shot capabilities, which means we do not need to spend energy on training them, a single inference across this pipeline takes ca. 40 times longer (on CPU) than a single inference using models such as YOLOv8. Thus, we highly recommend using such a pipeline for preliminary data collection followed by training a specific-domain model that can then perform inferences at scale.

## 5 Conclusion

In this paper, we explored the fast emerging field of multi-modal models and investigated its suitability for the digital humanist. The results of our investigation using the proposed pipeline show great potential from a technical aspect. We believe that this potential will lead to the generation of larger humanities datasets in the near future, but also to a larger interest and engagement from humanities scholars in computational approaches. This, we believe, is largely due to the linguistic interaction between the scholars and the machine, which is becoming one of the most human-computer interaction modes. This paper builds on the “multimodal turn in the Digital Humanities” [21]. This language interaction also forces us, as digital humanists, to reconsider object and class definitions, and reformulate them in a more computer-suited linguistic approach, which can often be very challenging, and often lead to new definitions and hypotheses.

The work on this pipeline is part of an ongoing infrastructure project at the Max Planck Institute for the History of Science that aims to collect large

amounts of visual content from heterogeneous historical documents. We used the pre-trained GroundingDINO as our object extractor in this paper; however, in the medium term we also plan to fine-tune this model on humanities-specific datasets in order to allow specific linguistic prompts to match the desired image region. In the long run, we plan to slowly build an application with a simple GUI around this pipeline to allow humanists with minimal computer science knowledge to extract such information from historical documents.

## Funding

This work was supported by the German Ministry for Education and Research as BIFOLD – Berlin Institute for the Foundations of Learning and Data (grant 01IS18037A) and the Max Plank Institute for the History of Science.

## Code availability

The code for this pipeline is available here: <https://github.com/hassanhajj910/prompt-me-a-dataset>.

## Acknowledgments

We would like to thank Lindy Divarci and Luis Melendrez Zehfuss for the English proofreading.

## References

1. Biglam: Bigscience libraries, archives and museums (2023), <https://huggingface.co/biglam>
2. Bibliothèque Nationale de France: Échantillon segmenté d’enluminures de mandragore (2019), <https://api.bnf.fr/mandragore-echantillon-segmente-2019>
3. Boillet, M., Bonhomme, M.L., Stutzmann, D., Kermorvant, C.: Horae: An annotated dataset of books of hours. In: Proceedings of the 5th International Workshop on Historical Document Imaging and Processing. p. 7–12. HIP ’19, Association for Computing Machinery, New York, NY, USA (2019). <https://doi.org/10.1145/3352631.3352633>, <https://doi.org/10.1145/3352631.3352633>
4. Büttner, J., Martinetz, J., El-Hajj, H., Valleriani, M.: Cordeep and the sacrobosco dataset: Detection of visual elements in historical documents. *Journal of Imaging* 8(10) (2022). <https://doi.org/10.3390/jimaging8100285>, <https://www.mdpi.com/2313-433X/8/10/285>
5. Büttner, J., Martinetz, J., El-Hajj, H., Valleriani, M.: Sacrobosco visual element dataset (s-ved) (Oct 2022). <https://doi.org/10.5281/zenodo.7142456>, <https://doi.org/10.5281/zenodo.7142456>
6. Cartucho, J., Ventura, R., Veloso, M.: Robust object recognition through symbiotic deep learning in mobile robots. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 2336–2341 (2018)

7. Droby, A., Kurar Barakat, B., Alaasam, R., Madi, B., Rabaev, I., El-Sana, J.: Text line extraction in historical documents using mask r-cnn. *Signals* **3**(3), 535–549 (2022). <https://doi.org/10.3390/signals3030032>, <https://www.mdpi.com/2624-6120/3/3/32>
8. Dutta, A., Bergel, G., Zisserman, A.: Visual analysis of chapbooks printed in scotland. In: The 6th International Workshop on Historical Document Imaging and Processing. p. 67–72. HIP '21, Association for Computing Machinery, New York, NY, USA (2021). <https://doi.org/10.1145/3476887.3476893>, <https://doi.org/10.1145/3476887.3476893>
9. Fischer, A., Liwicki, M., Ingold, R.: Handwritten Historical Document Analysis, Recognition, and Retrieval — State of the Art and Future Trends. *WORLD SCIENTIFIC* (2020). <https://doi.org/10.1142/11353>, <https://www.worldscientific.com/doi/abs/10.1142/11353>
10. Gaur, S., Sonkar, S., Roy, P.P.: Generation of synthetic training data for handwritten indic script recognition. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR). pp. 491–495 (2015). <https://doi.org/10.1109/ICDAR.2015.7333810>
11. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 2980–2988 (2017). <https://doi.org/10.1109/ICCV.2017.322>
12. Jones, E.: Large-scale book digitization in historical context: Outlines of a comparison. In: Proceedings of the 2011 IConference. p. 829–830. iConference '11, Association for Computing Machinery, New York, NY, USA (2011). <https://doi.org/10.1145/1940761.1940925>, <https://doi.org/10.1145/1940761.1940925>
13. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything (2023)
14. Lin, T., Maire, M., Belongie, S.J., Bourdev, L.D., Girshick, R.B., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. *CoRR* **abs/1405.0312** (2014), <http://arxiv.org/abs/1405.0312>
15. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., Zhang, L.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection (2023)
16. Monnier, T., Aubry, M.: docExtractor: An off-the-shelf historical document element extraction. In: ICFHR (2020)
17. Nikolaidou, K., Seuret, M., Mokayed, H., Liwicki, M.: A survey of historical document image datasets (2022). <https://doi.org/10.48550/arxiv.2203.08504>, <https://arxiv.org/abs/2203.08504>
18. Pattern Recognition & Image Analysis Research Lab: University of Salford, Manchester: Rasm2019 dataset (2019), <https://www.primaresearch.org/RASM2019/resources>
19. Redmon, J., Divvala, S.K., Girshick, R.B., Farhadi, A.: You only look once: Unified, real-time object detection. *CoRR* **abs/1506.02640** (2015), <http://arxiv.org/abs/1506.02640>
20. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. *CoRR* **abs/1505.04597** (2015), <http://arxiv.org/abs/1505.04597>
21. Smits, T., Wevers, M.: A multimodal turn in Digital Humanities. Using contrastive machine learning models to explore, enrich, and analyze digital visual historical collections. *Digital Scholarship in the Humanities* (03 2023). <https://doi.org/10.1093/llc/fqad008>, <https://doi.org/10.1093/llc/fqad008>, fqad008

22. Tan, M., Pang, R., Le, Q.V.: Efficientdet: Scalable and efficient object detection. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10778–10787. IEEE Computer Society, Los Alamitos, CA, USA (jun 2020). <https://doi.org/10.1109/CVPR42600.2020.01079>, <https://doi.ieeecomputersociety.org/10.1109/CVPR42600.2020.01079>
23. Valleriani, M., Vogl, M., el Hajj, H., Pham, K.: The network of early modern printers and its impact on the evolution of scientific knowledge: Automatic detection of awareness relationships. *Histories* **2**(4), 466–503 (2022). <https://doi.org/10.3390/histories2040033>, <https://www.mdpi.com/2409-9252/2/4/33>
24. Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L.M., Shum, H.Y.: Dino: Detr with improved denoising anchor boxes for end-to-end object detection (2022)