CHAI: Craters in Historical Aerial Images

Marvin Burges
Computer Vision Lab
TU Wien, Austria
mburges@cvl.tuwien.ac.at

Sebastian Zambanini
Computer Vision Lab
TU Wien, Austria
zamba@cvl.tuwien.ac.at

Philipp Pirker
Luftbilddatenbank Dr. Carls GmbH
Austria
pirker@luftbilddatenbank-gmbh.at

Abstract

In this paper we highlight the importance of historical aerial images in better understanding past events and their impact on their surroundings. More specifically, we are interested in studying bomb craters from World War II in Central Europe. We note the scarcity of publicly accessible datasets that provide labeled bomb craters and subsequently introduce a novel, domain-expert-annotated dataset comprised of 99 historical aerial images of Austria and Germany. We divide said data into training, validation, and test sets, and conduct training and evaluation using different object detectors - both general-purpose and specifically designed for remote sensing applications. This dataset thus serves as a benchmark for developing and evaluating (several) algorithms dedicated to the automated detection and analysis of bomb craters in historical aerial images. We underscore the uniqueness of this dataset as the first publicly available resource containing annotated bomb craters, thereby offering researchers a valueable and novel opportunity for future exploration. Lastly, we investigate possibilities for extending and enriching our data to enhance future studies, particularly within the context of preliminary risk estimation for unexploded bombs.

1. Introduction

Historical images, including aerial ones, offer valuable information that can provide insights into past events and their impact on the environment. In particular, aerial images captured during World War II can help researchers and explosive ordnance disposal services understand the extent of air strikes, damage to infrastructure and the natural landscape, as well as the potential dangers in construction projects [21]. The legacy of these air strikes during World War II is still present today, as numerous unexploded bombs are uncovered yearly in (Central) Europe [37]. Examining aerial images from surveillance flights during World War II makes it possible to make preliminary risk assessments based on the presence of bomb craters.

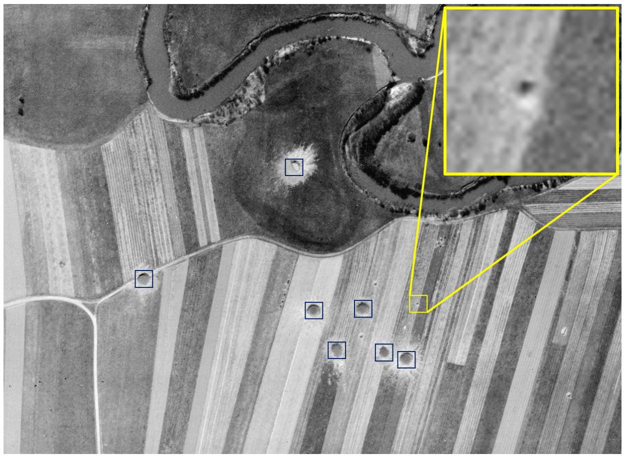


Figure 1. Seven bomb craters in a historical aerial image marked in dark blue and one unexploded bomb marked in yellow.

A practical example can be observed in Figure 1, displaying seven bomb craters. From this, an expert can deduce two key points. First, they can create an ”explosive ordnance map” featuring markings for each crater and outlining a 50m safety perimeter around each one. Construction endeavors within this radius typically demand added safety measures. Second, considering the historical context that bombers during that era commonly carried either 8 or 16 bombs, the expert can reasonably infer the presence of at least one unexploded bomb within the presented image, which was later found and diffused by the local authorities. However, annotating each crater manually can be a difficult and time-consuming process. Here, a (semi-)automated method, trained for this particular purpose, could thus greatly aid in expediting this process. Additionally, it could also be adapted for other regions around the world, such as historical aerial images from Cambodia following the Civil War (1967-1975) or the Vietnam War (1955-1975) [19, 27]

However, creating automated detection methods for identifying bomb craters in aerial images is challenging due to the absence of suitable and publicly available datasets. While datasets are available for Martian and moon craters [11, 34, 43], the same cannot be said for historical aerial images. Currently, to the author's best knowledge, only one dataset is available for historical aerial photos: the EuroSDR TIME dataset [15], consisting of 941 unlabeled images from Europe. However, of these 941 images, only 20 contain any craters and none feature any (expert) annotation.

This limitation makes it difficult for researchers to develop and test algorithms for the automated detection of bomb craters in aerial images. In rural areas, LIDAR image analysis can also be used for this purpose, as the morphological and morphometric characteristics of the crater are still visible today [36]. However, using historical aerial images is the only feasible option in urban areas where such characteristics have been lost owing to construction and development.

To overcome this limitation, this study introduces a historical aerial dataset of 99 aerial images taken in Austria and Germany during World War II, containing manually annotated bomb craters. Our industry partner provided these images and labels originating from their completed projects. The original images were sourced from national archives.

To promote replicability across experiments, we have made the raw images available in their original size, complete with crater annotations, Region Of Interest (ROI) delineations, and estimated Ground Sample Distance (GSD). Additionally, we have created a dataset from the raw images to facilitate ease of use in various experiments. The images in our dataset have been divided into predefined training, validation, and test sets. Furthermore, we have pre-tiled the images into image patches measuring 960x960 pixels, with a 20% overlap between adjacent patches.

We trained 15 object detectors on said training data and the resulting performances demonstrate the challenging nature of the given task, underscoring the importance of the presented dataset. Overall, the historical aerial dataset represents a valuable asset for researchers seeking to explore the consequences of air strikes during World War II on the environment. Moreover, the dataset offers the opportunity for domain adaptation, enabling the application of these algorithms to modern satellite images to detect craters.

While this dataset is primarily designed for bomb crater detection tasks, it can also be used for related tasks. One example would be its utilization for environmental impact studies that investigate long-term consequences of wartime activities [27]. Another use case could be educational or cultural heritage applications, where students or researchers can visualize the scale and impact of historical events.

The main contributions of this paper are twofold. First, the creation of a distinctive dataset featuring historical aerial images from Austria and Germany between 1943 and 1945, with a specific focus on identifying bomb craters. Second, the training and evaluation of 15 object detectors on said dataset, reveals its challenging nature, underscoring the importance of the presented dataset.

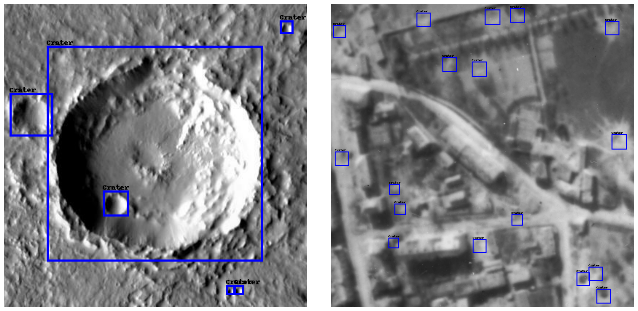
The remainder of this paper is structured as follows. First, in Section 2, the related work on datasets and object detection in historical aerial images and associated domains is presented, followed by the dataset proposed in this paper in Section 3. In Section 4 the experiments and their results are presented and in Section 5 the impact and the limitations of the dataset are discussed.

2. Related Work

Datasets. While large datasets are nothing new to the vision community, they typically focus on common objects in common areas using common cameras, like the COCO dataset [29] or the PASCAL VOC dataset [13]. Similarly, in remote sensing, there are large datasets available for object detection, like XView [24], the FMOW dataset [9], the DOTA dataset [39] or the DIOR dataset [25], which contain common classes like passenger vehicle, truck or building. This is likely because this imagery is comparatively easy to collect and annotate. In contrast, however, only a limited amount of training data is available for remotely sensed historical images, as this data can almost only be exclusively obtained from national archives. Furthermore, these images cannot be annotated by laypersons and require an expert, as the objects to find usually require special domain knowledge. Additionally, interpreting historical aerial images can be challenging due to factors like image quality, resolution, obstructions or the lack of context. Another factor, in the case of craters, is that precise localizations are required for the "explosive ordnance maps" and laypersons may lack the scientific background to reliably identify and measure in historical aerial images.

Currently, to our knowledge, only one historical aerial dataset is available containing craters, the EuroSDR TIME benchmark [15]. However, the primary focus of this dataset is to offer many historical aerial image blocks to test 2D and 3D image processing algorithms, specifically for aerial triangulation, georeferencing, generating digital surface models, and orthoimage production. Of the 941 images in the dataset, only 20 are captured during World War II, while all other images are from 1950 or later. Furthermore, none of the 20 images have crater annotations and are not usable for training a supervised machine learning approach. They could, however, be used for qualitative analysis or unsupervised training.

An alternative to bomb craters on Earth could be impact craters on Mars or Earth's moon. There have been multiple datasets proposed, like the Mars crater segmentation dataset presented by DeLatte *et al.* [12], the DeepCraters dataset by Yang and Guan [40], the Mars-Lunar crater dataset by Zhou [43] or various other datasets [11, 34]. However, lunar and



(a) Moon: Meteorite craters.

(b) Earth: Bomb craters.

Figure 2. Comparison of meteorite and bomb craters.

Martian craters vary significantly in size, from less than 30 meters to more than 1.938 meters, and are also considerably easier to detect, as there is no vegetation or human made structures [17]. Additionally, Burges *et al.* [4] showed that training with Moon images does not improve performance. Examples of a moon and a historical earth image are presented in Figure 2.

Object Detection in Historical Aerial Images. There is significant interest in historical aerial data. For instance, Brenner *et al.* [2] introduced a fully automatic system for detecting craters in historical aerial imagery. However, their evaluation revealed that human expertise is still necessary due to the low performance of their approach for fully automated deployment. Unfortunately, no dataset was released by Brenner *et al.* for reproducibility. Barone [1] presented a system that combined ground penetrating radar, a non-destructive technique, with historical aerial images to map potential unexploded bombs at archaeological sites. This method aimed to minimize damage to archaeological heritage and reduce risks during excavations. Similarly, Clermont *et al.* [10] developed a technique using convolutional neural networks and blob detection to detect bomb craters in aerial wartime images. However, Clermont *et al.* did not release any dataset for reproducibility. In another study, Kruse *et al.* [22] proposed using ellipses as object models to represent bomb craters. They employed a probabilistic algorithm based on marked point processes to determine the most likely object configuration. In a subsequent study, Kruse *et al.* [20] compared circles and ellipses as object models in a stochastic approach for automatically detecting bomb craters in aerial wartime images from World War II. Their method incorporated a term requiring homogeneous gray values within the object. In a follow-up paper, Kruse *et al.* [21] further optimized their approach and compared results obtained from single images with those from multiple photos of the same region. Utilizing multiple images improved their approach's F1-score from 39% to 67%. In their latest paper [23], Kruse *et al.* evaluated their approach on a larger dataset of 74 images. They compared it with a state-of-the-art convolutional neural network trained on a subset of their dataset. They found that the CNN outperformed their approach, but only with sufficient training data, highlighting the need for a large public dataset. However, no dataset was released by Kruse *et al.* for reproducibility. Taking a different approach, Waga *et al.* [37] employed digital elevation models to detect and evaluate morphometric parameters of craters in the Kedzierzyn-Kozle region. In contrast to the abovementioned methods, Geiger *et al.* [19] proposed a domain adaptation method for detecting bomb craters in aerial images. They highlighted the expensive and time-consuming nature of the current manual process and suggested that deep learning could offer a promising solution. However, they acknowledged the lack of a large labeled dataset for this task. To overcome this challenge, they proposed leveraging labeled moon surface images and a novel domain adaptation method to generate synthetic data by combining moon surface images with authentic aerial images captured during surveillance flights after allied air raids in World War II, using a CycleGAN. Unfortunately, no dataset was released by Geiger *et al.* for reproducibility.

Although Brenner *et al.* and Clermont *et al.* utilize deep learning techniques in their respective approaches, comparing them proves to be challenging due to the lack of publicly available datasets required for effective training. Conversely, Kruse *et al.* propose a non-deep learning-based method that eliminates the need for large datasets. However, it is essential to acknowledge that this approach falls short compared to the performance of the learned techniques with access to much training data. Moreover, Geiger *et al.* stress the importance of incorporating high-quality, real-world images in the training data and highlight the complementary role of synthetic data in the training process. Nevertheless, relying solely on synthetic data cannot fully replace the necessity for authentic, high-quality, real-world images.

Unfortunately, due to the absence of publicly available data and code from existing approaches, a direct and precise comparison is challenging. As a solution, we release our unique dataset to the public, aiming to facilitate future research by providing a standardized benchmark for evaluating algorithms in the context of historical aerial image analysis and crater detection. With this dataset, researchers will be able to develop, implement, and compare their approaches effectively.

3. CHAI Dataset

Data. Our proposed CHAI dataset consists of 99 images from 1943 to 1945. We obtained the images with the annotations from finished projects of our industry partner, covering both urban and rural areas. An overview of the different locations, the number of images per location, as well as

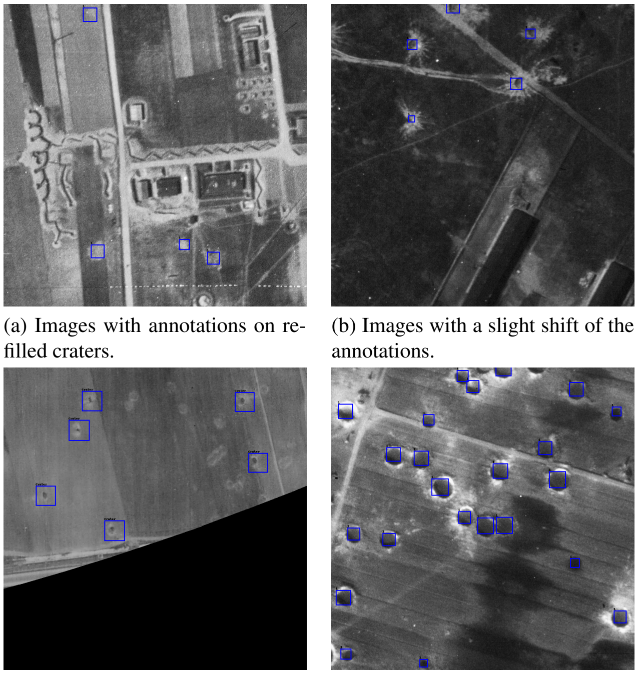
Location	Images	Split	GSD			Craters
			mean	min.	max.	
DE, Various	15	Test	0.27	0.17	0.38	2,923
AT, Graz Area	50	Train	0.21	0.16	0.31	15,315
AT, Linz	3	Val	0.22	0.19	0.24	138
AT, Vienna Area	31	Val	0.23	0.16	0.64	1,242
Combined	99	-	0.23	0.17	0.39	19,618

Table 1. General statistics of the images contained in the dataset. DE refers to Germany and AT for Austria. Craters refers to the annotations extracted from the industry partners projects, includes duplicate craters due to the overlap of images from the same flight and is not adjusted for the removed inconsistent annotations. Detailed information can be found in the supplementary material.

the approximate GSD and the total bombload per project, is presented in Table 1. Historical aerial images have been selected for these locations, and experts have mapped the bomb crater. In their workflow, the experts first survey all available photos that cover the ROI and then select a subset for the georeferencing and annotation. This subset consists of high-quality images with sufficient resolution and clarity and discards any image with, for example, excessive cloud coverage or bad lighting conditions. By carefully curating this subset of high-quality photos, they can streamline the subsequent georeferencing and annotation processes, minimizing potential errors and optimizing the overall efficiency of their workflow. Only in edge cases do they also annotate images of lower quality, but only if no higher quality image is available for the defined ROI. During the annotation process, multiple images from a similar time period and area may be accessible. However, due to cost-effectiveness, only one image is typically referenced and annotated, while the additional photos are usually reserved for addressing edge cases that may require additional information. This results in a dataset consisting of images with reasonable image quality and accurate annotations in the world coordinates, but the overlap of the referenced and annotated images is minimal.

Annotator Expertise and Quality Control Measures. Our industry partner possesses 12 years of experience and has completed approximately 2,000 projects in collaboration with various institutions in Central Europe. The task of annotation was undertaken by their team of experts, comprising historians and remote sensing specialists with a deep understanding of Central European history and expertise in analyzing aerial imagery. These annotators utilize specialized software designed for accurately pinpointing the exact locations of each bomb crater.

Each image undergoes annotation by a single annotator, and subsequently, the annotations are subjected to review by another expert. In instances where intricate situations arise, additional specialists are consulted. Moreover, additional images of the same geographical area are employed to evaluate challenging cases where the identification of craters is not straightforward. These additional images also aid in determining the specific type of bomb used in an attack, considering its distinctive blast radius, or in verifying whether an unexploded bomb that was initially detected has detonated in a subsequent image.



(a) Images with annotations on refilled craters.

(b) Images with a slight shift of the annotations.

(c) Loose annotations around a crater

(d) Too tight annotations around a crater.

Figure 3. Challenges from the annotation process.

Inconsistent Annotations. We encountered difficulties with the bounding boxes stemming from the annotation process and the inherent variability among human annotators. The annotation style of our industry partner aims to obtain a ”explosive ordnance map” efficiently, with minimal individual image labeling. Annotations relate to multiple georeferenced images, not just one. This leads to two main issues. First, with overlapping images or images from the same region but at a different time, it is unclear in which one a crater was annotated. To address this, we assumed that craters visible in earlier images also appear in later ones. While this is mostly true, some craters are re-filled. After manually investigating the whole dataset, we found that around 5% of annotations lacked visible craters, an example of this is presented in Figure 3a. As a result, we manually removed any annotations, under the supervision of our industry partner, that did not contain any visible crater residue, while keeping any annotations where traces of re-filled craters are visible.

On the other hand, georeferencing historical aerial images is not trivial. Even when done by experts, there might be slight inaccuracies in a small subset of the data, especially in rural areas, where reference points between historical aerial images and modern-day satellite images are hard to find. This, combined with the 1-to-many relationship, can lead to a slight shift in the annotations in these images. Figure 3b shows an example of this. A further challenge is the inherent variability of human annotators. Multiple experts have annotated the projects, and everyone might have a slightly different way of annotating a crater. For example, some might include ejecta rays (Figure 3c), while others annotate the crater too tightly (Figure 3d). We found, that of the 99 images in the dataset, six exhibit bounding box shifts in parts of the image, two contain loose annotations and three overly tight ones. We manually adjusted all inconsistent bounding boxes for the derived datasets.

Out of the 19,618 annotations for craters, we eliminated 497 because they lacked any indication of the presence of a crater. Additionally, we adjusted the size and position of 2,621 bounding boxes. The bounding boxes were removed or modified by the authors under the supervision of an industry expert. Both annotation sets will be released, with further details about the manual correction of the bounding boxes in the supplementary material.

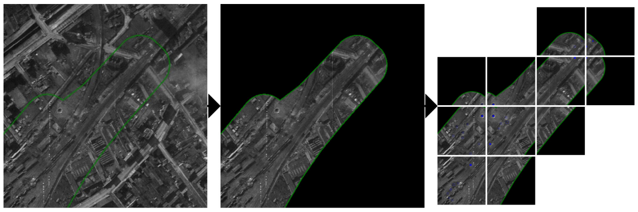


Figure 4. The patch extraction technique used to extract patches from the historical aerial images involves several steps. First, the area outside of the Region of Interest is made black to eliminate any unannotated craters that might be present. Next, the image is divided into overlapping patches, and patches located outside of the region of interest are subsequently eliminated.

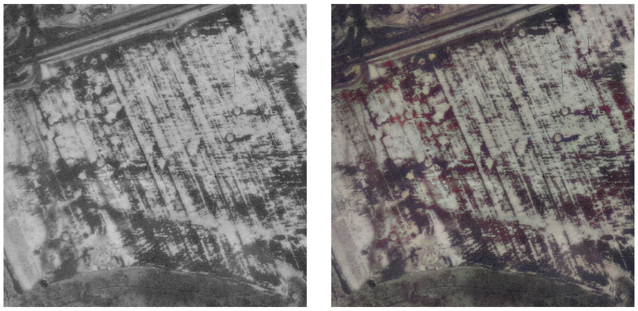
Dataset Extraction. The analysis is only performed within the defined ROI. Areas outside of the ROI are blackened, as these can contain unannotated craters. We also resized all images to the same, minimal GSD found in the data (0.16m per pixel), as we found that this does improve performance, compared to training on unnormalized images. In total, the provided projects contain 19,121 bomb craters in image coordinates (after the manual correction). We split all the images, using a custom script, into 960 × 960 patches with an overlap of 20%, a size typical in remotely sensed datasets. An overview of this pipeline is presented in Figure 4. We decided to derive two datasets from the data. In the first, we kept all patches inside of the ROI, and in the second, we additionally removed all patches that do not contain at least one crater annotation. A complete list of the available information for each image is presented in the supplementary material. The number of images and annotations for each dataset variant and split are presented in Table 2. Several randomly selected patches are visualized with bounding boxes in the supplementary material.

	Full dataset			Light dataset		
	Train	Val	Test	Train	Val	Test
Craters	22,798	1,933	4,458	22,798	1,933	4,458
Patches	12,145	1,173	2,161	3,748	461	990

Table 2. Patches and annotations per dataset split for the full dataset and light dataset. The total number of crater annotations increased from 19,121 to 29,189, due to the 20% overlap.

It is crucial to highlight that, while images within one location may overlap, images between location have no overlap. We therefore chose, that the greater area of Graz will be utilized as training dataset, the remaining locations in Austria (Vienna and Linz) will serve as the validation set and images obtained from German cities will be used for testing (a map is presented in the supplementary material). The data and the derived datasets will be available for download via Zenodo [3].

Coloring the Dataset. In addition, we conducted an experiment involving the application of deep learning for colorizing grayscale images, utilizing the Hyper-U-Net method proposed by Farella *et al.* [14]. We employed the model provided on their GitHub page along with the pre-trained weights provided by them, to color our **light** dataset. An example of an original and a colorized image are presented in Figure 5.



(a) Grey scale image. (b) Colored image.

Figure 5. Grey scale image before and after the colorization using Hyper-U-Net.

4. Dataset Evaluation

In Section 4.1, we first explain the experimental setup of how we evaluated the dataset on 15 detectors. These methods encompass a wide spectrum, including renowned detectors such as Faster-RCNN [33], cutting-edge approaches like DINO [42], and specialized methods like oriented RepPoints [26] developed specifically for remote sensing. Followed by the results on the light dataset, full dataset and the colored light dataset in Sections 4.2, 4.3 and 4.4, respectively. Finally, in Section 4.5 we present three images predicted by the best-performing model DINO, highlighting the potential challenges of our dataset.

4.1. Experimental Setup

Models. The detailed information regarding the selected models for this analysis is presented in Table 3. Our selection comprises object detectors that have demonstrated state-of-the-art performance on widely used datasets such as COCO [29]. These models include DETR [5] and its variants [30, 32, 45], Dino [42], Sparse-RCNN [35], YoloX [18], and YoloF [7]. Additionally, we considered well-established object detectors like SSD [31], RetinaNet [28], and Faster-RCNN [33]. We also incorporated object detectors specifically designed for remotely sensed images or detecting small objects. QueryDet [41] is an object detector specifically designed for the detection of (accelerating) small objects, while oriented RepPoints is specifically designed for aerial object detection. It is worth noting that oriented RepPoints predicts Oriented Bounding Boxes (OBB), instead of the commonly used Horizontal Bounding Boxes (HBB). While this is not advantageous for crater detection, we still chose oriented RepPoints due to their State-Of-The-Art performance on DOTA, a dataset that contains several small object classes.

Model	Year	Backbone	Prop.	Param.	Pret.
Faster-RCNN [33]	2015	ResNet50	HBB	42M	COCO
SSD [31]	2016	SSDVGG	HBB	24M	COCO
RetinaNet [28]	2017	ResNet50	HBB	36M	COCO
DETR [5]	2020	ResNet50	HBB	42M	COCO
Cond. DETR [32]	2021	ResNet50	HBB	43M	COCO
Def. DETR [45]	2021	ResNet50	HBB	40M	COCO
Sparse-RCNN [35]	2021	ResNet50	HBB	106M	COCO
YoloX-l [18]	2021	CSPDarknet	HBB	54M	COCO
YoloF [7]	2021	ResNet50	HBB	42M	COCO
TOOD [16]	2021	ResNet50	HBB	32M	COCO
DDOD [8]	2021	ResNet50	HBB	32M	COCO
DAB-DETR [30]	2022	ResNet50	HBB	44M	COCO
DINO [42]	2023	Swin-l	HBB	85M	COCO
QueryDet [41]	2022	ResNet50	HBB	39M	COCO
Ori. RepPoints [26]	2022	ResNet50	OBB	37M	DOTA

Table 3. Summary of object detection models. Models with Horizontal Bounding Box (HBB) proposals trained in mmdetection, with the exception of QueryDet trained using Detectron2 and Ori. RepPoints trained in mmrotate. All models have been pretrained on either the COCO or the DOTA dataset.

Size	Minimum	Maximum
Small	0	32
Medium	33	48
Large	49	∞

Table 4. Average diameter of the crater, and correspondingly the height and width of the bounding box, in pixels.

Metrics. For the experiments performed in this section, we are using the CocoAPI [29], with a modified definition for small, medium, and large bounding boxes, presented in Table 4. The evaluation metrics used include Average Precision (AP) for small (AP$_s$), medium (AP$_m$), and large (AP$_l$) craters, as well as AP with Intersection-Over-Union (IOU) thresholds of 25%, 50%, and 75%. The AP$_s$, AP$_m$, AP$_l$ metric are calculated for $IoU = 0.25 : 0.95$ instead of the common $IoU = 0.5 : 0.95$. The exception is oriented RepPoints, which is trained using the DOTA dataset format, here we only report AP$_{25}$, AP$_{50}$ and AP$_{75}$. For this, we converted the dataset from the COCO format to the DOTA format, utilizing eight $(x_1, y_1, ..., x_4, y_4)$ bounding box coordinates instead of the usual four ($XYXY$ or $XYHW$) coordinates. It is important to note that we did not rotate the bounding boxes during this conversion process.

Data Augmentation and Training. All models were trained using a consistent data augmentation pipeline. This pipeline involves random cropping, resizing images, random rotation up to 90 degrees, and random flipping, followed by normalization. The mmdetection [6], mmrotate [44] or the Detectron2 [38] frameworks were employed for training, and the configurations and details about the models and augmentations used can be found in our GitHub repository. Fine-tuning was performed using DOTA pretrained weights for Oriented RepPoints and COCO pretrained weights for other models, as this approach led to faster convergence. Specific details about the training of each model is available in the supplementary material.

4.2. Light Dataset

In Table 5, we provide a comprehensive overview of the results obtained from each model on the light dataset. Among the models, DINO achieves the highest overall performance in all tested metrics. Condi. DETR performed second best in the AP$_{25}$ metric, while YoloX-l performed second best in the AP$_s$ metric. The overall performance of TOOD, DDOD and DAB-DETR was close to the performance of DINO, with TOOD performing the second best in the AP$_m$ and AP$_l$ metric, DDOD second best in the AP$_{75}$ metric and DAB-DETR second best in the AP$_m$ and AP$_{50}$ metric. The discrepancy regarding Ori. RepPoints can be attributed to the inherent challenges associated with learning from OBBs, which introduce complexities that are not necessarily required for accurately detecting craters. It is also surprising to observe that the results for QueryDet are comparatively poor, considering that it is based on RetinaNet, which generally exhibits better performance. One potential reason for this behavior could be its specific focus on detecting rapidly moving small objects. Nevertheless, additional testing is essential to provide a more definitive explanation.

In conclusion, the evaluation results from Table 5 demonstrate the varying performance levels of different object detection models when applied to crater detection. While some models excel in specific metrics, the DINO detector emerges as the best across all evaluation criteria.

	AP$_s$	AP$_m$	AP$_l$	AP$_{25}$	AP$_{50}$	AP$_{75}$
Faster-RCNN	0.417	0.312	0.265	0.741	0.660	0.170
SSD	0.420	0.327	0.269	0.699	0.631	0.226
RetinaNet	0.440	0.338	0.293	0.800	0.731	0.188
DETR	0.430	0.357	0.307	0.778	0.702	0.191
Condi. DETR	0.425	0.359	0.324	0.804	0.722	0.167
Def. DETR	0.410	0.334	0.301	0.775	0.711	0.153
Sparse-RCNN	0.382	0.250	0.241	0.726	0.648	0.116
YoloX-l	0.456	0.318	0.272	0.769	0.619	0.229
YoloF	0.444	0.336	0.309	0.794	0.719	0.194
TOOD	0.440	0.364	0.328	0.774	0.714	0.231
DDOD	0.434	0.343	0.304	0.775	0.710	0.233
DAB-DETR	0.442	0.364	0.326	0.800	0.739	0.202
DINO	**0.474**	**0.393**	**0.333**	**0.828**	**0.759**	**0.273**
QueryDet	0.008	0.166	0.257	0.442	0.352	0.075
Ori. RepPoints	-	-	-	0.225	0.177	0.059

Table 5. Results: Trained object detection models on different average precision (AP) metrics, averaged over five runs. The best result **bold** and second best underlined. All models have been trained and tested on the light dataset. Extended table with STD in the supplementary material.

4.3. Full Dataset

We also evaluated the full dataset, which contains all patches within the ROI, and compared it to just training on the light dataset, where each patch extracted contains at least one annotation. This is a more realistic experiment, as the selected ROI rarely contains craters everywhere. For this experiment, we used DINO, as it showed the best performance in the previous experiment, and trained it on each dataset, additionally, we pre-trained DINO on the light dataset and then finetuned it on the full dataset. We evaluated all DINO weights obtained on the **full** test set. The results are presented in Table 6. It is evident that when DINO is trained on the light dataset, its performance on the full test set surpasses that achieved by training solely on the full dataset or the combination of both datasets, across all metrics. As a result, we conclude that additional empty patches in the training data do not enhance the model's training, and that training exclusively on the light dataset proves to be sufficient. An additional benefit of exclusively training on the light dataset is the time saved. Training on the light dataset takes approximately 48 minutes per epoch on average, with on average 10 epochs required, whereas using the full dataset necessitates around 204 minutes per epoch on average, with an average requirement of 4 epochs, leading to a reduction by approximately 59% in training time required (Light: 480; Full: 816 minutes). However, when comparing DINO in Table 5 and 6 one can see that testing on the light dataset overestimates the models performance by about 14% on the AP$_{50}$ metric, while the AP$_{75}$ does not change significantly. A table of all models trained on the light dataset, but tested on the full, can be found in the supplementary material.

	AP$_s$	AP$_m$	AP$_l$	AP$_{50}$	AP$_{75}$
Light	**0.484**	**0.408**	**0.341**	$\mathbf{0.659 \pm .025}$	$\mathbf{0.274 \pm .034}$
Full	0.458	0.374	0.302	$0.653 \pm .015$	$0.216 \pm .036$
FT-1	0.468	0.392	0.320	$0.622 \pm .030$	$0.225 \pm .007$
FT-3	0.450	0.372	0.294	$0.618 \pm .031$	$0.183 \pm .049$
FT-6	0.452	0.366	0.275	$0.595 \pm .038$	$0.182 \pm .029$

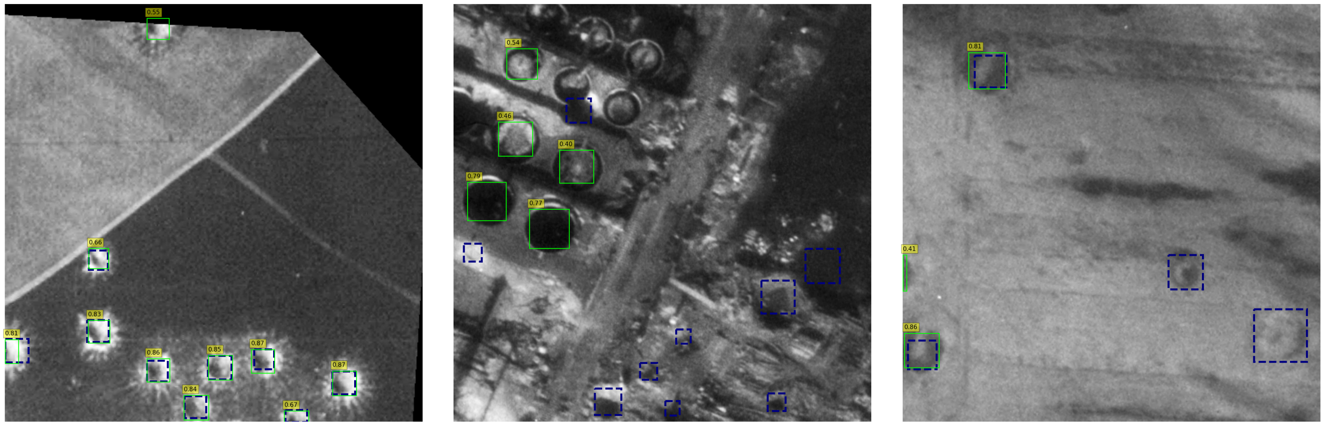
Table 6. Results: DINO, trained on the full, and light dataset, and pre-trained on the light and then FineTuned (FT) for 1, 3, 6 epochs on the full dataset. Best result in **black**, values averaged over five runs. Extended table with STD in the supplementary material.

4.4. Colored Light Dataset

We then trained DINO on the resulting dataset and compared it to its performance on the grayscale version. The result can be seen in Table 7. It is visible that the colorized images produced using the Hyper-U-Net method only minimally affect the performance of the DINO model, with the exception of the AP$_{75}$, where the colorization improved performance by about 30%. These results show, that colorizing greyscale images can improve performance, however, at the cost of a reduced inference speed. DINO alone took about 24ms per image, while colorizing the image took about 100ms per image, measured on an Nvidia RTX 3090.

	AP$_s$	AP$_m$	AP$_l$	AP$_{50}$	AP$_{75}$
Grey	0.474	0.393	**0.333**	$\mathbf{0.759 \pm .015}$	$0.273 \pm .044$
RGB	**0.490**	**0.396**	0.318	$0.741 \pm .012$	$\mathbf{0.356 \pm .063}$

Table 7. Results: DINO, trained on the grayscale and the RGB dataset. Best result in **black**, averaged over five runs. Extended table with STD in the supplementary material.



(a) Rural area.

(b) Urban area.

(c) New domain.

Figure 6. Predictions of the DINO model on the test set with a fixed threshold of 0.4.

4.5. Visual Analysis

We present three predictions on the test set in Figure 6, all predictions are obtained from DINO with a fixed threshold of 0.4. This threshold was established through an iterative process of trial and error on the validation set. Figure 6a depicts an urban example and it is visible that all craters have been found correctly. In Figure 6b, a complex scene of a bombed factory is depicted. In this scenario, DINO struggles to identify any craters correctly. Instead, it erroneously identifies storage tanks as craters. An evident shift in domain is illustrated in Figure 6c, where the craters are only faintly visible due to a fresh layer of snow covering them. Here, DINO accurately detects only two out of four craters while mistakenly flagging a crater at the image periphery.

5. Discussion

We have introduced a novel dataset comprising historical aerial images taken in Austria and Germany between 1943 and 1945, with the primary objective of identifying craters. In total, 99 images were carefully selected, and domain experts diligently annotated 19,618 bomb craters. This dataset is specifically designed to propel the advancement of object detection algorithms tailored for analyzing historical aerial imagery. While computer vision has seen the emergence of extensive datasets, we discovered a remarkable absence of a large-scale dataset for this particular task.

It is important to also acknowledge the inherent limitations of this dataset, particularly regarding significant domain gaps, as seen in the image with snow coverage or in urban areas. We are aware that the generalizability of the trained algorithms may be compromised under such circumstances. Nevertheless, we anticipate that the release of this dataset will encourage other researchers to contribute to the research community by openly sharing their code and presenting their findings using our dataset. This collaborative approach will undoubtedly enhance overall comprehension and progress in the field.

Furthermore, it is vital to acknowledge that every instance of a missed crater detection within the dataset represents a potential unexploded bomb during construction activities. This underscores the significance of employing responsible and explainable approaches in the development of detection algorithms. Addressing these aspects and devising methods to interpret and elucidate the algorithm's decision-making process represent promising avenues for future research endeavors.

Acknowledgments

This work was supported by the Austrian Research Promotion Agency (FFG) under project grant 880883. Acquisition of historical aerial imagery: Luftbilddatenbank Dr. Carls GmbH; Sources of historical aerial imagery: National Archives and Records Administration (Washington, D.C.) and Historic Environment Scotland (Edinburgh).

References

[1] Pier Matteo Barone. Bombed archaeology: Towards a precise identification and a safe management of wwii's dangerous unexploded bombs. *Heritage*, 2(4):2704–2711, 2019. 3

[2] Simon Brenner, Sebastian Zambanini, and Robert Sablatnig. Detection of bomb craters in wwii aerial images. In *OAGM Workshop 2018: Medical Image Analysis*, pages 94–97, 05 2018. 3

[3] Marvin Burges, Sebastian Zambanini, and Philipp Pirker. Craters in historical aerial images (chai) dataset. 10.5281/zenodo.10068633, 2023. 5

[4] Marvin Burges, Sebastian Zambanini, and Robert Sablatnig. Exploring learning-based approaches for bomb crater detection in historical aerial images. In *OAGM Workshop 2022:*

Digitalization for Smart Farming and Forestry, pages 60–66, 09 2022. 3

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, volume 12346 of *Lecture Notes in Computer Science*, pages 213–229. Springer, 2020. 6

[6] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. Mmdetection: Open mmlab detection toolbox and benchmark. *CoRR*, abs/1906.07155:1–13, 2019. 6

[7] Qiang Chen, Yingming Wang, Tong Yang, Xiangyu Zhang, Jian Cheng, and Jian Sun. You only look one-level feature. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 13039–13048. Computer Vision Foundation / IEEE, 2021. 6

[8] Zehui Chen, Chenhongyi Yang, Qiaofei Li, Feng Zhao, Zheng-Jun Zha, and Feng Wu. Disentangle your dense object detector. In Heng Tao Shen, Yueting Zhuang, John R. Smith, Yang Yang, Pablo César, Florian Metze, and Balakrishnan Prabhakaran, editors, *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, pages 4939–4948. Acm, 2021. 6

[9] Gordon A. Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 6172–6180. Computer Vision Foundation / IEEE Computer Society, 2018. 2

[10] Dominic Clermont, Christian Kruse, Franz Rottensteiner, and Christian Heipke. Supervised detection of bomb craters in historical aerial images using convolutional neural networks. *ISPRS - International Society for Photogrammetry and Remote Sensing*, Xlii-2/w16:67–74, 2019. 3

[11] Annotation Group Crater. Crater dataset. `https://universe.roboflow.com/crater-zqpjg/crater-vrqmn`, June 2022. visited on 2023-08-21. 2

[12] Danielle DeLatte, Sarah T. Crites, Nicholas Guttenberg, Elizabeth J. Tasker, and Takehisa Yairi. Mars crater segmentation dataset, May 2022. Accessed on 2023-06-27. 2

[13] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010. 2

[14] Elisa M. Farella, Salim Malek, and Fabio Remondino. Colorizing the past: Deep learning for the automatic colorization of historical aerial images. *Journal of Imaging*, 8(10):269, 2022. 5

[15] Elisa M. Farella, Luca Morelli, Fabio Remondino, Jon P. Mills, Norbert Haala, and Joep Crompvoets. The eurosdr time benchmark for historical aerial images. In *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, volume 43, pages 1175–1182, 2022. 2

[16] Chengjian Feng, Yujie Zhong, Yu Gao, Matthew R. Scott, and Weilin Huang. TOOD: task-aligned one-stage object detection. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 3490–3499. Ieee, 2021. 6

[17] Alistair Francis, Jonathan Brown, Thomas Cameron, Reuben Crawford Clarke, Romilly Dodd, Jennifer Hurdle, Matthew Neave, Jasmine Nowakowska, Viran Patel, Arianne Puttock, Oliver Redmond, Aaron Ruban, Damien Ruban, Meg Savage, Wiggert Vermeer, Alice Whelan, Panagiotis Sidiropoulos, and Jan-Peter Muller. A multi-annotator survey of sub-km craters on mars. *Data*, 5(3):70, 2020. 3

[18] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. YOLOX: exceeding YOLO series in 2021. *CoRR*, abs/2107.08430:1–7, 2021. 6

[19] Marco Geiger, Dominik Martin, and Niklas Kühl. Deep domain adaptation for detecting bomb craters in aerial images. In Tung X. Bui, editor, *56th Hawaii International Conference on System Sciences, HICSS 2023, Maui, Hawaii, USA, January 3-6, 2023*, pages 825–834. ScholarSpace, 2023. 1, 3

[20] Christian Kruse, Franz Rottensteiner, and Christian Heipke. Marked point processes for the automatic detection of bomb craters in aerial wartime images. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Xlii-2/w13:51–60, 2019. 3

[21] Christian Kruse, Franz Rottensteiner, and Christian Heipke. Using redundant information from multiple aerial images for the detection of bomb craters based on marked point processes. *ISPRS - International Society for Photogrammetry and Remote Sensing*, V-2-2020:861–870, 2020. 1, 3

[22] Christian Kruse, Franz Rottensteiner, Thorsten Hoberg, Marcel Ziems, Julia Rebke, and Christian Heipke. Generating Impact Maps from Automatically Detected Bomb Craters in Aerial Wartime Images Using Marked Point Processes. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, Iv3:127–134, Apr. 2018. 3

[23] Christian Kruse, Dennis Wittich, Franz Rottensteiner, and Christian Heipke. Generating impact maps from bomb craters automatically detected in aerial wartime images using marked point processes. *ISPRS Open Journal of Photogrammetry and Remote Sensing*, 5:100017, 2022. 3

[24] Darius Lam, Richard Kuzma, Kevin McGee, Samuel Dooley, Michael Laielli, Matthew Klaric, Yaroslav Bulatov, and Brendan McCord. xview: Objects in context in overhead imagery. *CoRR*, abs/1802.07856:1–16, 2018. 2

[25] Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*, 159:296–307, 2020. 2

[26] Wentong Li, Yijie Chen, Kaixuan Hu, and Jianke Zhu. Oriented reppoints for aerial object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition,*

CVPR 2022, New Orleans, LA, USA, June 18-24, 2022, pages 1819–1828. Ieee, 2022. 6

[27] Erin Lin, Rongjun Qin, Jared Edgerton, and Deren Kong. Crater detection from commercial satellite imagery to estimate unexploded ordnance in cambodian agricultural land. *Plos One*, 15(3):1–22, 03 2020. 1, 2

[28] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2999–3007. IEEE Computer Society, 2017. 6

[29] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer, 2014. 2, 6

[30] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. DAB-DETR: dynamic anchor boxes are better queries for DETR. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. 6

[31] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: single shot multibox detector. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I*, volume 9905 of *Lecture Notes in Computer Science*, pages 21–37. Springer, 2016. 6

[32] Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang. Conditional DETR for fast training convergence. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 3631–3640. Ieee, 2021. 6

[33] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 91–99, 2015. 6

[34] Ari Silburt, Mohamad Ali-Dib, Chenchong Zhu, Alan Jackson, and Kristen Menou. Deepmoon supplemental materials, Jan. 2018. 2

[35] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, and Ping Luo. Sparse R-CNN: end-to-end object detection with learnable proposals. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 14454–14463. Computer Vision Foundation / IEEE, 2021. 6

[36] Jan M. Waga and Maria Fajer. The heritage of the second world war: Bombing in the forests and wetlands of the koźle basin. *Antiquity*, 95(380):417–434, 2021. 2

[37] Jan M. Waga, Bartłomiej Szypuła, and Maria Fajer. The archaeology of unexploded world war ii bomb sites in the koźle basin, southern poland. *International Journal of Historical Archaeology*, 27:688 – 713, 2022. 1, 3

[38] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019. 6

[39] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge J. Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. DOTA: A large-scale dataset for object detection in aerial images. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 3974–3983. Computer Vision Foundation / IEEE Computer Society, 2018. 2

[40] Chen Yang and Renchu Guan. Deep craters. `https://figshare.com/articles/dataset/CE_DeepCraters/12768539/1`, 8 2020. visited on 2023-08-21. 2

[41] Chenhongyi Yang, Zehao Huang, and Naiyan Wang. Querydet: Cascaded sparse query for accelerating high-resolution small object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 13658–13667. Ieee, 2022. 6

[42] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. *CoRR*, abs/2203.03605:1–7, 2022. 6

[43] Lincoln Zhou. Mars - lunar crater dataset. `https://universe.roboflow.com/lincoln-zhou/mars-lunar-crater`, Oct. 2022. visited on 2023-08-21. 2

[44] Yue Zhou, Xue Yang, Gefan Zhang, Jiabao Wang, Yanyi Liu, Liping Hou, Xue Jiang, Xingzhao Liu, Junchi Yan, Chengqi Lyu, Wenwei Zhang, and Kai Chen. Mmrotate: A rotated object detection benchmark using pytorch. In João Magalhães, Alberto Del Bimbo, Shin'ichi Satoh, Nicu Sebe, Xavier Alameda-Pineda, Qin Jin, Vincent Oria, and Laura Toni, editors, *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, pages 7331–7334. Acm, 2022. 6

[45] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: deformable transformers for end-to-end object detection. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 6