
VISION-LANGUAGE MODEL FOR ACCURATE CRATER DETECTION

Patrick Bauer

University of Technology of Troyes
Hochschule Darmstadt
patrick.bauer@utt.fr

Marius Schwinnig

GMV for European Space Agency
marius.schwinnig@ext.esa.int

Florian Renk

European Space Agency
florian.renk@esa.int

Andreas Weinmann

Technische Hochschule Würzburg-Schweinfurt
Hochschule Darmstadt
andreas.weinmann@thws.de

Hichem Snoussi

University of Technology of Troyes
hichem.snoussi@utt.fr

ABSTRACT

The European Space Agency (ESA), driven by its ambitions on planned lunar missions with the Argonaut lander, has a profound interest in reliable crater detection, since craters pose a risk to safe lunar landings. This task is usually addressed with automated crater detection algorithms (CDA) based on deep learning techniques. It is non-trivial due to the vast amount of craters of various sizes and shapes, as well as challenging conditions such as varying illumination and rugged terrain. Therefore, we propose a deep-learning CDA based on the OWLv2 model, which is built on a Vision Transformer, that has proven highly effective in various computer vision tasks. For fine-tuning, we utilize a manually labeled dataset from the IMPACT project, that provides crater annotations on high-resolution Lunar Reconnaissance Orbiter Camera Calibrated Data Record images. We insert trainable parameters using a parameter-efficient fine-tuning strategy with Low-Rank Adaptation, and optimize a combined loss function consisting of Complete Intersection over Union (CIoU) for localization and a contrastive loss for classification. We achieve satisfactory visual results, along with a maximum recall of 94.0% and a maximum precision of 73.1% on a test dataset from IMPACT. Our method achieves reliable crater detection across challenging lunar imaging conditions, paving the way for robust crater analysis in future lunar exploration.

Keywords Crater Detection · Vision Language Models · OWLv2 model

1 INTRODUCTION

Craters are among the most prominent features on the lunar surface, and their accurate detection is of critical importance to the European Space Agency (ESA) due to its direct impact on the success of planned lunar missions. The ESA’s *Terrae Novae 2030+* program [1], aligned with the exploration strategy *Explore2040* [2], outlines a detailed plan for future lunar exploration. They emphasize the importance of crater detection as a key component in achieving the goal of sending the first European astronaut to the lunar surface. ESA’s lunar lander program, Argonaut [3], will enable Europe to access the Moon as a fully European project. It consists of two main components: the Lunar Descent Element (LDE) and the Passenger. The LDE is responsible for both transporting and landing the Passenger on the lunar surface. Since even the smallest craters, only a few meters in diameter, can potentially cause a lander to tip over, their accurate detection conditions the success of the whole exploration mission. The main focus in recent crater detection research has been on detecting large craters [4]. In this work, we overcome this limitation by targeting craters of various sizes and shapes, also under varying sun incidence angles, resulting in craters that range from barely to highly shadowed.

According to Chaini et al. [5] crater detection algorithms (CDAs) can be divided into manual methods, where craters are labeled by humans, and in automatic crater detection approaches. Manual crater detection has the limitation of

being very time consuming and error-prone [6]. Further, according to Robbins et al. [7], manual crater count can vary by up to 40%. Hence, the focus of CDAs is on automatic crater detection, where traditional methods such as edge detection [8] have been used. In recent years, deep learning has become the primary focus in the field of CDAs, as the data provided by NASA's Lunar Reconnaissance Orbiter Camera (LROC) contains complex information, including varying crater shapes and sizes, diverse lighting conditions, and rugged terrain, where traditional approaches face significant limitations [5]. Several deep learning approaches have been proposed, e.g. Tewari et al. [9], that used a Mask R-CNN model with a ResNet-50 backbone and a feature pyramid network, trained on optical, elevation, and slope map inputs for crater detection. In a later work, Tewari et al. [10] proposed a novel crater shape retrieval system that first estimates crater boundaries using an unsupervised adaptive rim extraction algorithm on DEM data. Subsequently, they refine these boundaries through cascaded Mask R-CNNs in a semi-supervised manner, enabling highly accurate crater shape retrieval. Several researchers incorporated various versions of the You Only Look Once (YOLO) model as their CDA. Nan et al. [11] proposed YOLOv8-LCNET, an enhanced YOLOv8-based model for automatic lunar crater detection, which integrates a Partial Self-Attention (PSA) mechanism in the backbone and a Gather-and-Distribute (GD) module in the neck to improve multi-level feature information and detection accuracy. La Grassa et al. [12], Mu et al. [13] and Zhu et al. [14] also employed variants of the YOLO model and achieved good crater detection results. Zhang et al. [15] focused on small-scale crater detection using an anchor-free deep learning approach. Specifically, they applied CenterNet with transfer learning, detecting crater centers as peaks in a heat map and directly regressing their sizes, without relying on non-maximum suppression (NMS). Jia et al. [16] introduced the AE-TransUNet+, an improved hybrid Transformer network based on the TransUNet [17], which achieved high accuracy in detecting small craters. Leveraging pretrained foundation models and adapting them to specific downstream tasks has also proven promising [18]. For example, Giannakis et al. [19] applied SAM [20] for crater detection and achieved promising results, even without fine-tuning it on remote sensing images containing craters. For further existing CDA research, we refer to the review article of Chaini et al. [5].

In recent years, ViTs have played a vital role in the realm of computer vision. Tasks such as object detection, semantic segmentation, and image classification are increasingly addressed using ViTs [21]. The patch-based processing structure of ViTs enables them to capture long-range dependencies across an image, enabling them for tasks that demand a comprehensive understanding of complex visual content [22]. Through positional encoding and multi-head self-attention (MSA), ViTs have demonstrated strong performance on various vision benchmarks, often matching or outperforming convolutional neural networks (CNNs) in tasks such as object detection [23]. However, unlike CNNs, ViTs lack strong inductive biases and thus typically require extensive pre-training on large-scale datasets to generalize effectively. Extensive pre-training on large datasets enhances model performance and, as noted by Dosovitskiy et al. [24], can even surpass the benefits of inductive biases. This flexibility and scalability make ViTs a robust and versatile foundation for advanced methods, including feature extraction and task-specific adaptation. Many models based on ViTs have been proposed in recent years, such as the OWL-ViT [25] (Vision Transformer for Open-World Localization) and its successor, OWLv2 [26]. Previous work [27] demonstrated a promising few-shot crater detection approach using the OWLv2 model by customizing the similarity metric utilizing the high-dimensional image embeddings. In this work, we build on that approach by fine-tuning the OWLv2 model for the task of crater detection on lunar surface images. According to [18], pretraining foundation models on large-scale datasets and subsequently fine-tuning them for specific downstream tasks is a widely adopted paradigm in computer vision. To address the computational and memory challenges of full fine-tuning, several parameter-efficient fine-tuning (PEFT) techniques have been proposed in the literature, including Adapters [28], Prefix-Tuning [29], Prompt Tuning [30], BitFit [31], AdapterFusion [32], and IA³ [33]. These approaches enable efficient task-specific adaptation while keeping most of the base model parameters fixed. For our methodology, we adopt LoRA [34], that inserts low-rank trainable matrices into the MSA layers of frozen ViT weights and in the detection heads, allowing efficient fine-tuning by optimizing only a small number of additional parameters.

Contributions

Detecting craters on the lunar surface is an important research objective, and the ESA's ambitions for lunar exploration demand a reliable CDA that performs well across diverse lunar regions and imaging conditions. Building on the contributions in [27], which employed a few-shot crater detection strategy using the OWLv2 model and introduced a penalty term in the similarity score calculation, we propose a deep-learning based enhancement, summarized as follows:

- (i) We propose a method that fine-tunes the pre-trained OWLv2 model by introducing additional trainable parameters with LoRA, enabling efficient adaptation to the crater detection task. The model is trained by minimizing a combined loss of Complete Intersection over Union (CIoU) for localization and contrastive loss for classification.

- (ii) The model is trained and evaluated on a manually labeled dataset from the IMPACT project [35]. We observe strong performance in detecting craters of varying sizes and shapes, supported by both qualitative and quantitative evaluation.

2 METHODOLOGY

In this section, we present our method for fine-tuning the OWLv2 [26] model on a manually annotated dataset from the IMPACT [35] project for an novel accurate crater detection approach.

2.1 Vision Transformer in Computer Vision

With the introduction of Transformers [36] and the incorporation of attention modules, a new paradigm in natural language processing (NLP) emerged. Attention-based models enable the modeling of dependencies between any elements in the input or output sequence, regardless of their distance [36]. In this process, words or subwords are tokenized and embedded into a high-dimensional vector space. Positional encodings are added to the input embeddings to provide the model with information about the order of tokens [36]. Each layer applies multi-head self-attention by projecting the input tokens with weight matrices W_Q, W_K, W_V onto the queries Q , keys K , and values V . Attention is further the scaled dot-product

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \quad (1)$$

computed across multiple heads, where $\frac{1}{\sqrt{d_k}}$ denotes the scaling factor. The head outputs are then concatenated and projected with W_O . A position-wise feed-forward network follows, with residual connections and layer normalization around each sublayer, yielding the Transformer output tokens [36]. The original Transformer is structured as an encoder-decoder architecture. Several benchmark models such as BERT [37] and GPT-3 [38] have been developed, with GPT-3 trained on 45 TB of text data and containing approximately 175 billion parameters. Pre-trained models that are Transformer-based are capable of achieving state-of-the-art performances on various tasks [39]. As a consequence, the Transformer has become the standard architecture in NLP. Following the remarkable success of the Transformer in NLP, researchers began adapting the architecture to computer vision tasks such as image classification, object detection, and segmentation. Its first major breakthrough in image classification was achieved by the Vision Transformer (ViT) [24], which applies an encoder-only Transformer architecture directly to sequences of image patches. The idea is that an image is divided into fixed-size non-overlapping patches, and each patch, similarly to words or subwords in a standard Transformer, is treated as a token and embedded as a high-dimensional vector, with positional encodings added to retain spatial information [24]. In general, a class token is prepended to all tokens and serves as an overall representation of the image. Similar to large pre-trained Transformer models in NLP, several foundation models based on ViTs have emerged in computer vision, including CLIP [40] (Contrastive Language-Image Pretraining), SAM [20] (Segment Anything Model), and DINO [41] (Self-Distillation with No Labels). For example, CLIP is designed to align text and visual information in a shared embedding space and is, through extensive contrastive pre-training on large amounts of image-text data, capable of tasks such as zero-shot image classification. This means it is not limited to a fixed set of labels or categories seen during pre-training. Building on these developments, OWL-ViT [25] extends CLIP to the setting of object detection.

2.2 OWLv2: Vision Transformer for Open-World Localization

OWL-ViT [25] builds upon the CLIP [40] architecture by combining a ViT [24] for image encoding and a Transformer-based text encoder. These components are jointly pre-trained on 3.6 billion image-text pairs using contrastive learning to align visual and textual features in a shared embedding space. For object detection, OWL-ViT removes the token pooling layer, attaches classification and box regression heads to the Transformer output tokens, and fine-tunes the model on roughly 2 million object-level annotated images. OWLv2 [26] extends OWL-ViT by incorporating large-scale self-training: The open-vocabulary object detector OWL-ViT CLIP-L/14 [25] is used to generate pseudo-boxes over WebLI [42] (~ 10 billion image-text pairs), and models are self-trained on a ~ 2 billion image subset. A new objectness head improves efficiency by selecting tokens that are most likely to correspond to actual objects against the background, reducing unnecessary computations during detection.

OWLv2 processes the input images by first resizing them to a fixed image size of 960×960 pixels with bilinear interpolation. The image is then divided into 3600 non-overlapping patches of size 16×16 pixels. Each patch is linearly projected into a 768-dimensional vector space, and positional encodings are added to preserve spatial structure.

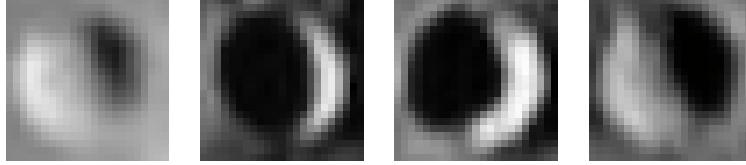


Figure 1: Examples of craters for the few-shot crater detection approach.

A special class token that captures global image information, is prepended to the sequence, resulting in a total of 3601 tokens. These tokens are then passed through the ViT encoder, which models local and global semantic relationships via MSA, analogous to token processing in NLP transformers. After encoding, the resulting embeddings capture local and global semantic relationships and are used for object detection. The class token plays a key role in aggregating global context. As noted by Minderer et al. [25], multiplying the class token with the patch embeddings can enhance object detection performance. Each of the 3600 patch embeddings is passed through three distinct prediction heads. The box head generates a four-dimensional vector representing the predicted bounding box coordinates. The objectness head produces a scalar score indicating the likelihood that an object is present in the corresponding bounding box or indicating background. The class head outputs a 512-dimensional image class embedding that encodes the visual features of the predicted object.

Simultaneously, the text encoder processes objects of interest provided as textual prompts (e.g., “dog”, “cat”, etc.). The output are 512-dimensional vectors representing the corresponding objects of interest. The core idea behind contrastive ViTs is that objects described by text should have a similar representation, based on cosine similarity, to the 512-dimensional image class embeddings derived from the visual features. Models like OWL-ViT and OWLv2 are capable of performing both zero-shot and few-shot object detection [25, 26]. Zero-shot object detection refers to the model’s ability to detect and localize objects from classes that were not necessarily seen during training. Instead of relying on visual examples for each class, the model utilizes semantic information provided through text prompts or class descriptions [43]. Few-shot object detection, on the other hand, addresses the challenge of recognizing novel classes using only a limited number of labeled examples. Hence, OWLv2 provides a framework for adapting zero- and few-shot object detection to the task of crater detection.

2.3 OWLv2 for Crater Detection

The naive approach of detecting craters using the zero-shot object detection approach using the pretrained OWLv2 model and prompting the word “crater” presents several challenges. First of all, it is not clear what the exact prompt should be, as alternatives like “craters”, “small craters”, “impact craters”, or even “circles” are all plausible, since craters appear circular when viewed from a nadir perspective [44]. Moreover, applying this approach revealed further limitations as the confidence scores indicating whether detected objects are actual craters were consistently low. This may be explained by the fact that the OWLv2 model was not specifically trained on remote sensing images of the Moon. While the word “crater” may occur in captions, it was likely not aligned with lunar craters as visual objects, which limits OWLv2’s ability to detect them reliably. In prior work [27], a few-shot approach was found more suitable by passing example craters (cf. Fig. 1) and their corresponding image class embeddings, then identifying similar objects, i.e. craters, based on cosine similarity. In this setup, only the ViT path of the OWLv2 model was used, while the text Transformer was not considered. The analysis of embeddings classified as craters revealed high similarity even between visually distinct instances, such as brightly illuminated craters and those almost entirely shadowed. This indicates a high intraclass similarity among the embeddings describing craters. However, the presence of similarly high interclass similarity, i.e., between craters and non-craters, motivated the introduction of a penalty mechanism to improve separation in the embedding space. Further, as noted by [45], the performance of frozen ViTs pretrained on natural images is often limited when directly applied to remote sensing tasks due to a significant domain gap, making further adaptation or fine-tuning typically necessary. Similarly, Luo et al. [46], in their adaptation of the Segment Anything Model (SAM) [20] to remote sensing, emphasized this domain gap and underscored the importance of domain-specific adaptation strategies. Thus, fine-tuning the OWLv2 model on lunar-specific remote sensing data is crucial to fully exploit its capability for detecting lunar craters.

In this approach, we utilize both the text Transformer and the ViT as shown in Fig. 2. The 512-dimensional embedding for the word “crater” serves as an anchor in a contrastive framework: crater image embeddings are pulled toward this anchor, while non-crater embeddings are pushed away, by maximizing the cosine similarity for craters and minimizing it for non-craters. The OWLv2 model processes 3600 patches per image, allowing detection of up to 3600 objects.

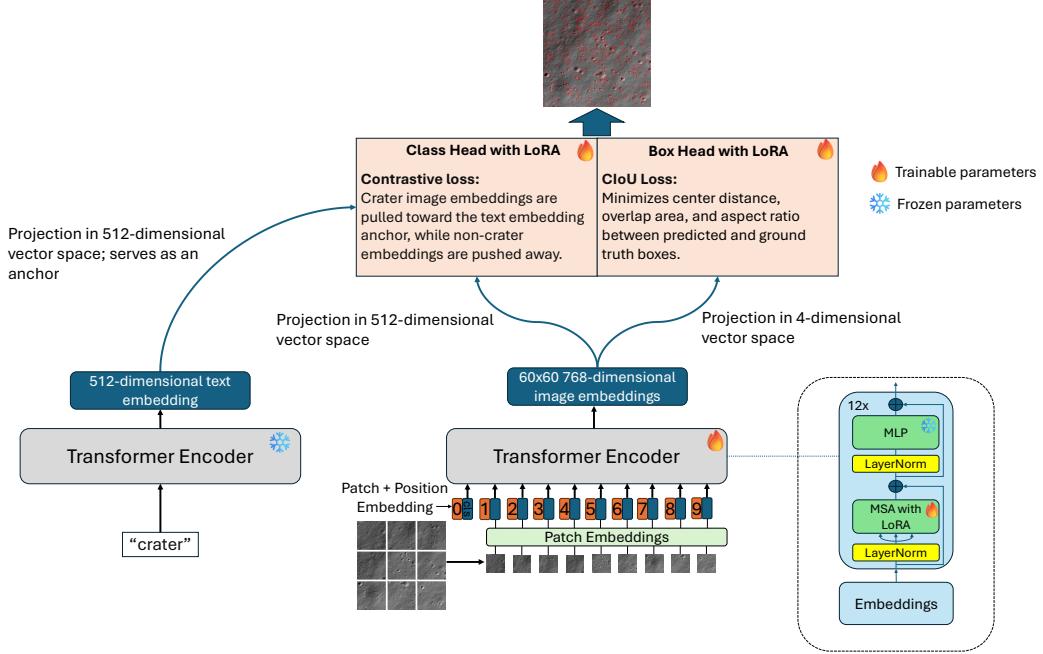


Figure 2: Schematic overview of the proposed crater detection method: Calibrated Data Record (CDR) images are processed by the OWLv2 Vision Transformer (ViT), where trainable LoRA parameters are inserted into the encoder. Simultaneously, the word “crater” is encoded by the frozen text Transformer and used as an anchor in the shared embedding space. The Box Head minimizes a CIoU-based loss between predicted boxes and ground truth, while the Class Head applies a contrastive loss to separate crater from non-crater embeddings utilizing the anchor vector, by adding LoRA parameters into both heads for efficient fine-tuning.

To assign predictions to ground truth (GT) craters, we use the Hungarian matching algorithm [47], which minimizes the cost function based on CIoU [48] (complete intersection over union). Let $y = \{y_i\}_{i=1}^M$ denote the set of GT craters for a given image, where M is the number of annotated craters, and let $\hat{y} = \{\hat{y}_j\}_{j=1}^N$ be the set of N model predictions, with $M \ll N$ (in OWLv2, $N = 3600$; median M_m of annotated craters in our GT dataset: $M_m = 200$). Following [49], we pad y with $(N - M)$ “no-object” entries so that both sets have equal cardinality and we denote the predictions by i (rows) and the padded GT boxes by j (columns). We compute a pairwise CIoU cost matrix $C \in \mathbb{R}^{N \times N}$,

$$C_{ij} = 1 - \text{CIoU}(\hat{b}_i, b_j), \quad (2)$$

where $\hat{b}_i \in [0, 1]^4$ and $b_j \in [0, 1]^4$ are predicted and GT boxes, respectively. We then obtain the alignment

$$\sigma^* = \underset{\sigma \in S_N}{\operatorname{argmin}} \sum_{i=1}^N C_{i,\sigma(i)}, \quad (3)$$

where S_N is the set of all permutations of N elements. The problem is solved efficiently with the Hungarian algorithm [47]. After matching, we compute the bounding box loss only for the M true GT and prediction pairs:

$$\mathcal{L}_{\text{box}} = \frac{1}{M} \sum_{i=1}^M [1 - \text{CIoU}(\hat{b}_{\sigma^*(i)}, b_i)], \quad (4)$$

where \mathcal{L}_{box} is the same CIoU-based regression loss used to define the matching cost.

Further, let $t \in \mathbb{R}^{512}$ denote the ℓ_2 -normalized text embedding of the word “crater”, and let $e_j \in \mathbb{R}^{512}$ denote the ℓ_2 -normalized image class embedding for prediction j . For each image, we mark the $\sigma^*(i)$ indices found by the Hungarian assignment of the matching result as positives and all other indices as negatives. We then apply a cosine-similarity contrastive loss that (i) pulls the positive image embeddings $e_{\sigma^*(i)}$ toward the anchor t and (ii) pushes the remaining negative embeddings toward orthogonality:

$$\mathcal{L}_{\text{cls}} = \frac{1}{P} \sum_{j \in \mathcal{P}} (1 - \cos(e_j, t)) + \frac{1}{N} \sum_{j \in \mathcal{N}} \max(0, \cos(e_j, t) - \tau),$$

with $y_j \in \{0, 1\}$ the binary label derived from σ^* , $\mathcal{P} = \{j \mid y_j = 1\}$ and $\mathcal{N} = \{j \mid y_j = 0\}$ denote the sets of positive and negative pairs with $P = |\mathcal{P}|$ and $N = |\mathcal{N}|$, respectively, and τ is a margin for the negatives (here, $\tau = 0.1$). That means, we let the model learn what a (visual) crater embedding $e_j \in \mathbb{R}^{512}$ should look like given the vector $t \in \mathbb{R}^{512}$ that corresponds to the textual description ‘‘crater’’. In total, we minimize the loss function

$$\mathcal{L}_{\text{total}} = \alpha_{\text{box}} \mathcal{L}_{\text{box}} + \alpha_{\text{cls}} \mathcal{L}_{\text{cls}}, \quad (5)$$

where we set $\alpha_{\text{box}} = 7.5$ and $\alpha_{\text{cls}} = 3.0$ to prioritize precise localization while still ensuring strong classification performance; these values were chosen empirically. After the training process, the crater detection pipeline follows the zero-shot object detection principle. By prompting the word ‘‘crater’’ through the text Transformer and passing out-of-distribution selected CDR images through the ViT, the model predicts only objects with 512-dimensional vectors similar to the vector representing ‘‘crater’’.

2.4 Injecting parameters with LoRA

As described in Subsection 2.1, the queries, keys, and values are computed by linearly projecting the input embeddings: $Q = W_Q x, K = W_K x, V = W_V x$. During fine-tuning, these weight matrices are frozen. Mathematically, for $m \in \{Q, K, V\}$, a pre-trained weight matrix $W_m \in \mathbb{R}^{d \times k}$ is updated by adding trainable low-rank matrices $A_m \in \mathbb{R}^{r \times k}, B_m \in \mathbb{R}^{d \times r}, r \ll \min(d, k)$. For an input x a corresponding step of the forward path yields

$$W_m x + \Delta W_m x = W_m x + B_m A_m x. \quad (6)$$

Note that during training, the pre-trained weight matrix W_m remains frozen and is not updated. In the image encoder, we apply LoRA weight matrices only to W_Q and W_V , as following the procedure in [34], and use a rank of $r = 8$. Traditionally, LoRA parameters are injected only into the multi-head self-attention (MSA) and feed-forward (MLP) layers of the Transformer. However, since our objective is to learn and represent crater-specific features within the high-dimensional embedding space, we also inject LoRA parameters into the class and box head instead of fully fine-tuning them, following the idea in [50], where LoRA parameters were added in the classification head. To enable this, we apply LoRA to both heads with a rank of $r = 16$. For the downstream task of crater detection with the OWLv2 model, we emphasize that fine-tuning the objectness head is unnecessary and would only add computational and memory overhead to the training process. Therefore, we remove the objectness head and focus solely on fine-tuning the box and classification heads. In doing so, only a small fraction (0.2%) of the OWLv2 model parameters are trainable.

3 EXPERIMENTAL SETUP

3.1 Dataset

NASA launched the Lunar Reconnaissance Orbiter (LRO) in June 2009 and since then it has been observing the Moon from an altitude of 50 km, providing various images from the Moon and its surface [51]. To do so, it has several instruments onboard such as the Lunar Reconnaissance Orbiter Camera (LROC). The LROC consists of two distinct camera systems, the Narrow Angle Cameras (NACs) and the Wide Angle Camera (WAC). The two NACs, the NAC-Left (NAC-L) and NAC-R (NAC-Right), were designed to capture high-resolution images down to ~ 0.5 m/pixel, and the WAC provides images at a scale of ~ 100 m/pixel [51]. The mission’s objectives include identifying potential landing sites, detecting surface hazards, and creating high-resolution maps of the lunar polar regions. For the task of detecting craters, we select images from the LROC Calibrated Data Record (CDR), that have undergone radiometric and geometric corrections, and downloaded them from the Planetary Data System archive (<https://pds.lroc.asu.edu/data/>).

For fine-tuning the OWLv2 model, we use the manually labeled dataset provided by the IMPACT project [35]. The project provides a game, where players are encouraged to annotate lunar craters as accurately as possible while building a base on the lunar surface. As a result, the project publishers, who plan to make the dataset publicly available in the near future (currently only available to ESA), have collected and verified a large number of crater annotations spanning various lunar regions. Example annotations are shown in Fig. 4. Crater annotations are provided as normalized circle coordinates (c_x, c_y, r) relative to the CDR image size of 2048×2048 pixels, with craters excluded that are smaller than 8 pixels in diameter. Each image is identified by its CDR image ID (e.g., M118673590LC), followed by the $(x_{\text{min}}, y_{\text{min}})$ coordinates of the upper left corner of image tile. The final number indicates the tile size, which is always 2048 (for example: M118673590LC_1508_36964_2048). For training and evaluation, we convert these circular coordinates to square bounding boxes, cut each image in 16 tiles of size 512×512 and use them as GT. This also motivates the use of LoRA in the box head: fully unfreezing the box head would lead the model to associate craters strictly with square shapes, which could hinder generalization to other celestial bodies or to specific emission angles, where craters can appear more elliptical and require rectangular shaped bounding boxes.

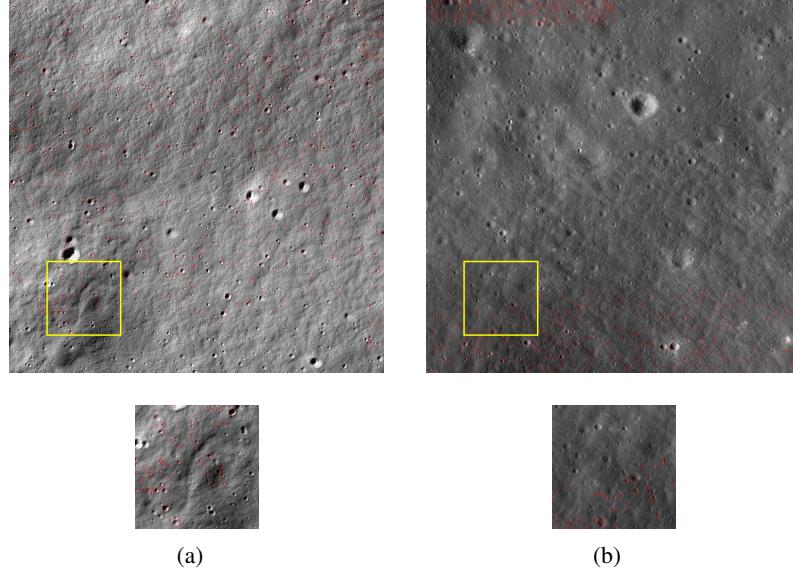


Figure 3: Examples of poorly annotated images. Each column: top: original 2048×2048 image with a yellow rectangle marking the zoomed region; bottom: cropped zoom corresponding to that rectangle. Red circles show annotated ground-truth craters.

3.2 Processed IMPACT Dataset

Since players are encouraged to label craters and, as previously noted, manual crater counts can vary widely, the developers introduced a metric to quantify annotation accuracy. Each crater receives an accuracy score based on the consistency of markings from multiple players. Regions incorrectly labeled as craters, where only a few players mark the same location, receive low accuracy scores. Hence, the raw data contains many false positive annotations, making data pre-processing essential to obtain a reliable dataset for training and validation. Mostly following the developers’ recommendations, we set an accuracy threshold of 0.55, slightly below the actually suggested value of 0.6, and remove noisy annotations that cover more than 85% black pixels to enhance dataset quality. This step is necessary because regions in full shadow frequently confused users, leading them to mark image noise. We consider any pixel with an intensity lower than a grayscale value of 30 (in the range $[0, 255]$) as black. Further visual inspection indicates that some images contain inconsistent labeling, with many craters missing. Therefore, we removed entire images that were of insufficient quality, cf. Fig. 3. As noted by Robbins et al. [52], the decision to label a feature as a crater depends on how conservative or liberal the annotator is. Since our goal is to support ESA’s lunar landing efforts, we prioritize the recall metric, aiming to minimize false negatives, and exclude images with missing annotations. Consequently, we adopt a more liberal approach to labeling features as craters. After preprocessing, we retain a total of 880 tiles of size 512×512 , containing a total of 178,812 crater annotations. We divide the dataset based on the following criteria. Since in some cases, multiple tiles are collected from the same image ID but from different regions, for instance, two distinct images may correspond to a single image ID (e.g. M118673590LC_1508_36964_2048 and M118673590LC_1508_41060_2048), we ensure that all images from the same image ID are placed in the same dataset split — train, validation, or test — to prevent data leakage. We target an 80 – 10 – 10 split for training, validation, and testing, respectively. We group all 2048×2048 tiles by their parent LROC image ID and randomly assign approximately 80% of image IDs to the training set, and approximately 10% each to validation and testing, ensuring that no CDR image ID is shared across splits. At the image level, the split results in 42/7/6 images for train, validation, and test, corresponding to approximately 76%, 13%, and 11%. The actual tile counts deviate slightly from the intended 80 – 10 – 10 split. We consider this modest drift acceptable given the strict grouping by image ID.

3.3 Data Augmentation and Implementation Details

To increase the size of the training set and to get a more diverse dataset we apply data augmentation strategies. We follow the idea presented in [53] by constructing five different augmentation policies. Each policy is selected randomly and within each policy, individual augmentation operations are applied with a fixed probability and a certain range of magnitude. We keep the exact same data augmentation policies as in [53], but replace the BBox_Only_TranslateY with a BBox_Only_Rotate technique to randomize the shadow orientation and discourage the model from associating

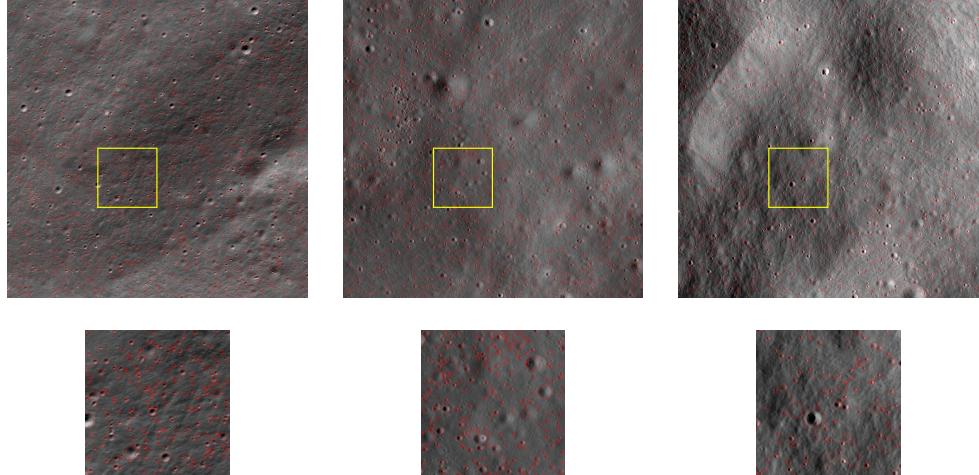


Figure 4: Each column: (top) Original 2048×2048 image with a yellow rectangle indicating the zoomed region. (bottom) Cropped and zoomed-in region corresponding to the yellow rectangle above. The red circles indicate the annotated ground truth craters.

Table 1: Data augmentation sub-policies and their parameters. Op: Operation, P: probability, RoM: Range of Magnitude. Following [53], each sub-policy consists of two operations with an associated probability and magnitude. We keep the continuous ranges provided by Albumentations [55], which directly reflect our implementation.

Sub-policy	Operation 1			Operation 2		
	Op	P	RoM	Op	P	RoM
Sub-policy 1	TranslateX	0.6	$[-0.4, 0.4]$	Equalize	0.8	—
Sub-policy 2	BBox_Only_Rotate	0.2	$[-180, 180]$	Scale	1.0	$[-0.3, 0.3]$
Sub-policy 3	ShearY	0.6	$[-10, 10]$	BBox_Only_Rotate	0.6	$[-180, 180]$
Sub-policy 4	Rotate	0.6	$[-30, 30]$	ColorJitter	1.0	br.=0.6, co.=0.5, sa.=0.5, hu.=0.1
Sub-policy 5	No operation	—	—	No operation	—	—

a specific shadow direction with the crater class. In addition, we replace `Cutout` with `Scale` to better capture craters at various sizes. An overview over the data augmentation strategy is provided in Table 1. For each image in the train dataset we apply one of the five policies and double the training set size. Training is performed for 100 epochs with a batch size of 4. We apply a weight decay of 1×10^{-3} and use the AdamW [54] optimizer. The learning rate is linearly warmed up from 0 to 1×10^{-4} during the first 10% of epochs, then decayed to zero over the remaining epochs following a cosine schedule. The experiments were conducted using an NVIDIA H100 NVL GPU with 94 GB of VRAM.

4 RESULTS

In this section, we evaluate our approach both qualitatively and quantitatively. The quantitative evaluation is based on recall and precision, defined as

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad (7)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad (8)$$

where TP denotes the number of true positives (actual craters), FN the number of false negatives (missed craters), and FP the number of false positives (incorrectly identified as craters). We consider a detection as a true positive (TP) if the IoU is greater than 0.30. In inference, we apply non-maximum suppression with a IoU threshold of 0.12 to remove duplicate boxes for the same crater.

4.1 Qualitative Evaluation

In Fig. 5, we observe good visual results in detecting craters across various lunar regions, that were not seen during training or validation. These images were selected as out-of-distribution samples from the CDR archive and the

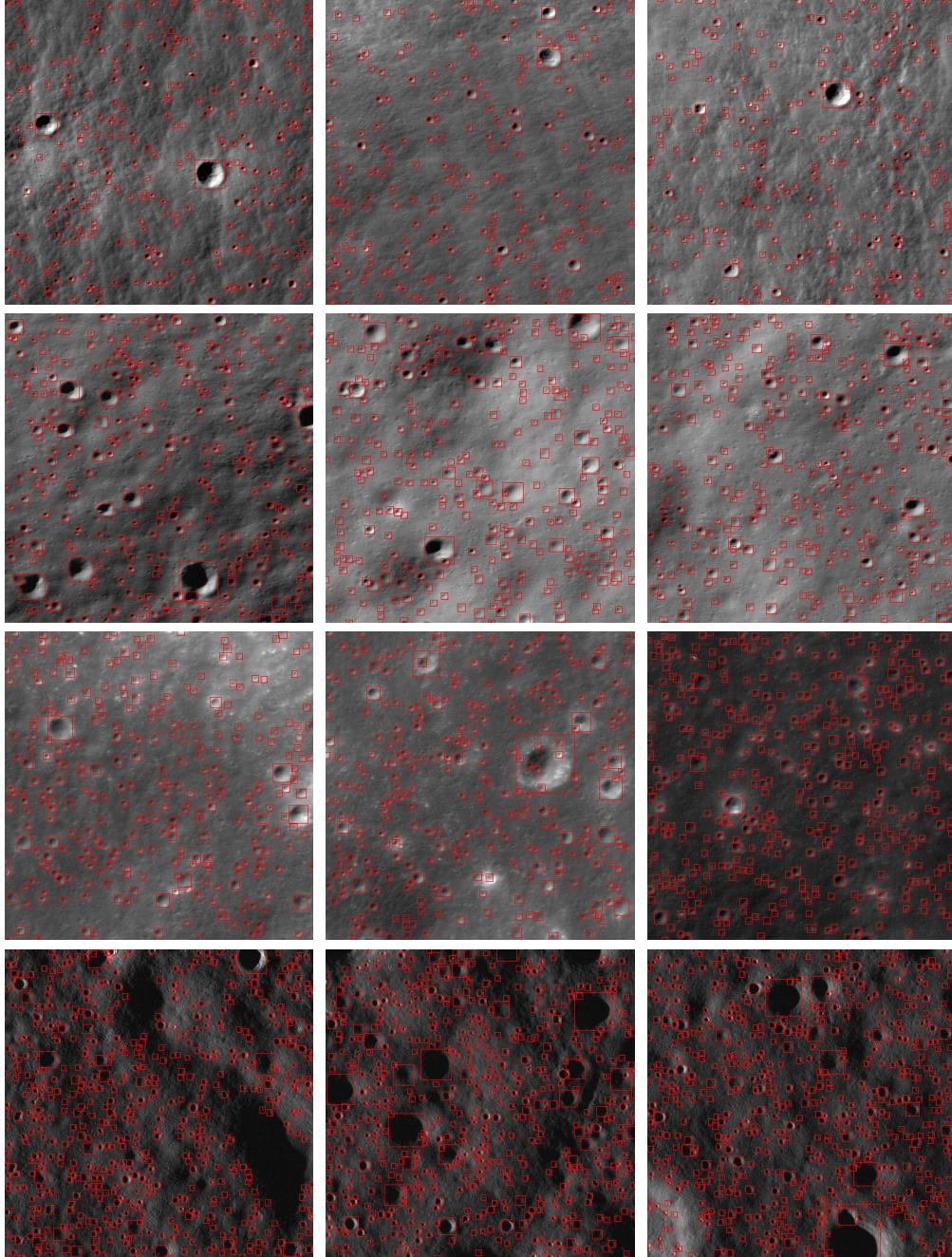


Figure 5: Qualitative results on CDR images unseen during both training and validation, selected randomly. Each row consists of three tiles of size 512×512 . First row: M1369716293R, second row: M1310840621L, third row: M1501897277L, fourth row: M1496678308R. Despite varying image conditions, such as incidence angles ranging from 40.2 to 85.1 degrees, an overall sufficient detection quality is achieved. In the fourth row, especially for larger craters, the bounding boxes often capture only the shadowed part of the crater, that also appear circular. Under these illumination conditions, however, even for human observers the crater contours are difficult to extract.

experiments were conducted on tiles of size 512×512 . For a robust evaluation, we selected images with different resolutions, ranging from $0.49 \text{ m}/\text{px}$ to $1.85 \text{ m}/\text{px}$, as well as varying illumination incidence angles ranging from 40.2 to 85.1 degrees, to demonstrate the model's applicability across diverse imaging conditions. Our results show that the method detects craters of various sizes and shapes and remains robust under varying image conditions. Considering the detections in Fig. 6, which showcases an image of the lunar south pole, where craters can appear more elliptical, our

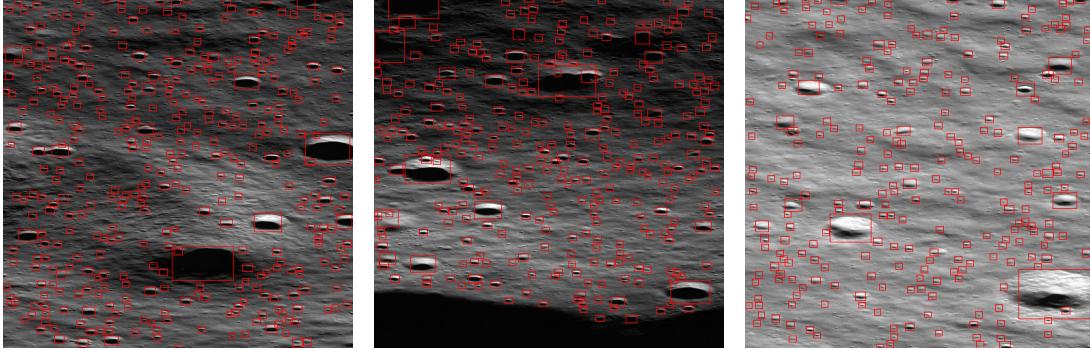


Figure 6: Qualitative results on CDR image M185603505L from the lunar south pole, unseen during both training and validation. Due to the emission angle, the craters appear more elliptical than circular yielding more rectangular shaped bounding boxes, a condition completely absent from training and validation. Nevertheless, during inference the craters are detected sufficiently well, highlighting the strong generalization capability of the fine-tuned OWLv2 model.

Table 2: Recall and precision of the fine-tuned OWLv2 model on the test dataset.

Image Name	Recall (%)	Precision (%)
M1184106708LC_1508_34916_2048	87.8	52.9
M1184106708LC_1508_47204_2048	85.5	29.3
M1250049562LC_1508_16484_2048	93.4	46.9
M1466265041LC_1508_100_2048	94.0	50.5
M1466265041LC_1508_12388_2048	82.2	73.1
M1466265041LC_1508_49252_2048	87.5	64.0

approach is not limited, as the craters are still detected accurately and represented with more rectangular bounding boxes. This demonstrates the model’s generalization capabilities. Further, we emphasize that our method does not produce many visually observable false positives. For nearly every detected crater, one could argue that it represents an actual crater, even if no corresponding GT annotation exists. Overall we note that the qualitative evaluation of our method yields that it is of sufficient quality. Nevertheless, we highlight the need for further methodological improvements, as in some cases, only the shadowed region of a crater is detected. This effect is even more evident in merely illuminated regions, where the shadowed areas appear circular, as shown in the last row of Fig. 5. Under such illumination conditions, it is difficult even for human observers to extract the exact crater boundaries.

4.2 Quantitative Evaluation

The quantitative evaluation is performed on the test dataset of 6 images and therefore 96 tiles of size 512×512 in total. In Table 2 we outline the corresponding evaluation results. The recall values range from 82.2% to 94.0% with an average of 88.4%, while the precision values range from 29.3% to 73.1% with an average of 52.8%. While the recall on each test image is relatively high, indicating a good detection rate of true positives, the precision is in comparison rather low. We emphasize that many detections not present in the GT, counted by definition as false positives, are in fact plausible crater detections. This is also highlighted in Fig. 7, where we evaluated our method on the test set with GT annotations. We observe a high consistency between our predictions (red boxes) and the GT annotations (yellow circles). Notably, in almost every case, the number of predictions exceeds the number of GT annotations. This can also be explained by the `scale` data augmentation technique, where a negative scale makes the GT craters smaller during training and causes the model to predict craters smaller than 8 pixels. This effect is also evident in Fig. 7, where most of the newly detected craters appear to be small. This is not a limitation of our method, since our goal is to detect craters across all sizes and not to exclude small ones. In our work, we aimed to construct a more liberal catalog out of the IMPACT annotations, to reduce the number of missed detections. Nevertheless, many craters were still absent from the GT annotations, partly due to the applied size constraint of 8 pixels in diameter and the accuracy threshold. As a result many detections counted as false positives are, in fact, likely true craters that were simply not labeled, a challenge also noted by Tewari et al. [10] and Silburt et al. [56].

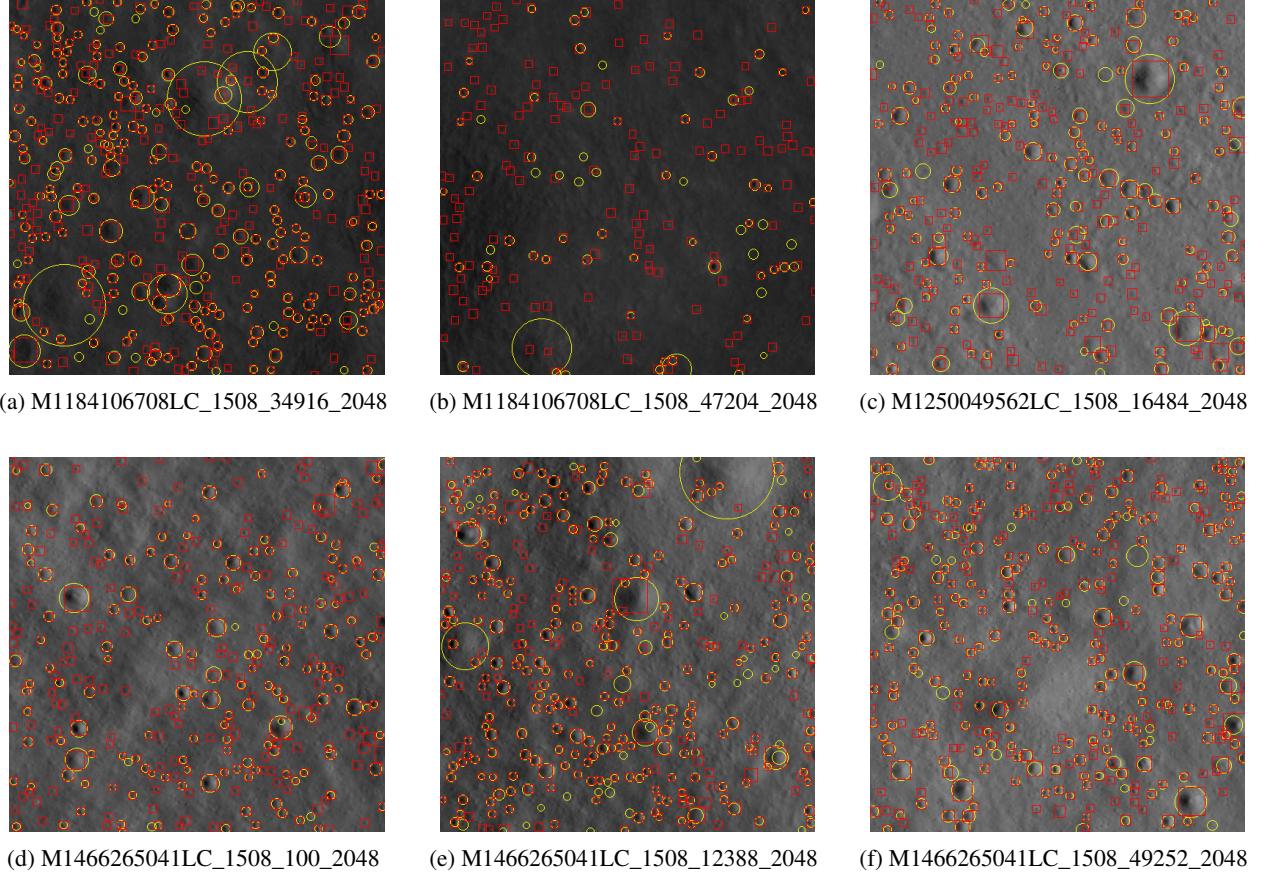


Figure 7: Qualitative results on one 512×512 tile from each of the six images from the IMPACT test dataset. Yellow circles indicate ground truth annotations, and red bounding boxes denote model predictions. The model predicts more craters than are annotated in the ground truth.

5 DISCUSSION AND FUTURE WORK

In this paper, we introduced a novel crater detection method based on fine-tuning the pre-trained OWLv2 model and an adapted loss function. This method aligns text and images in a shared embedding space using a text Transformer and a ViT, respectively. To do so, we inserted trainable parameters based on LoRA into the vision encoder as well as in classification and box head, respectively, while keeping the text Transformer frozen. We also removed the objectness head entirely, as it is of no use during inference and would add unnecessary memory and computational overhead. For training and evaluation, we utilized a manually labeled dataset from the IMPACT project. During training, we aligned the predicted bounding boxes with the GT annotations by minimizing a CIoU-based loss, while simultaneously aligning the corresponding class vector by maximizing the cosine similarity with the anchor vector obtained from the text Transformer encoding the word “crater”.

Our method is qualitatively evaluated on randomly selected CDR images and quantitatively on the test dataset from IMPACT. The qualitative evaluation shows that the method successfully detects craters of different sizes and shapes under varying illumination conditions. Also, the model generalizes well, as demonstrated by its ability to detect more elliptical craters, such as those on lunar south pole images. Visually, only a few false negatives occur, indicating that our model can be broadly applied to detect craters under various image conditions. This is supported by the quantitative evaluation, where recall values remain high, ranging from 82.2% to 94.0%. On the other hand, we observe notably lower precision values ranging from 29.3% to 73.1%. We emphasize that many of the false positives affecting the precision metric are in fact actual craters that were not labeled in the GT, either due to the accuracy score threshold or the minimum size constraint of 8 pixels. Nevertheless, we note that some detections are incorrectly classified as craters, especially in rugged lunar terrain.

Future work will focus on improving the detection of the smallest craters and on extending the application to other celestial bodies, such as Mars. Further we also aim to respect the circular morphology of craters and extract their shapes directly rather than relying on bounding boxes, which could involve the use of semantic segmentation techniques.

Acknowledgments

This research work is conducted via the Open Space Innovation Platform (<https://ideas.esa.int>) as a Co-Sponsored Research Agreement and carried out under the Discovery programme of, and funded by, the European Space Agency (contract number: 4000144734). The view expressed in this publication can in no way be taken to reflect the official opinion of the European Space Agency. This work is also partially funded by the Hochschule Darmstadt and the University of Technology of Troyes within the European University of Technology EUT+.

References

- [1] European Space Agency. *Terraen Novae: Europe's exploration vision*, 2025. URL: https://www.esa.int/Science_Exploration/Human_and_Robotic_Exploration/Exploration/Terraen_Novae_Europe_s_exploration_vision.
- [2] European Space Agency. *EXPLORE2040: The European Exploration Strategy*, 2024. URL: https://esamultimedia.esa.int/docs/HRE/Explore_2040.pdf.
- [3] European Space Agency. *Argonaut: Europe's lunar lander programme*, 2025. URL: https://www.esa.int/Science_Exploration/Human_and_Robotic_Exploration/Exploration/Argonaut_Europe_s_lunar_lander_programme.
- [4] Chen Zou, Jialong Lai, Yanshuang Liu, Feifei Cui, Yi Xu, and Le Qiao. Small lunar crater identification and age estimation in Chang'e-5 landing area based on improved Faster R-CNN. *Icarus*, 410:115909, 2024. doi:10.1016/j.icarus.2023.115909.
- [5] Chinmayee Chaini and Vijay Kumar Jha. A review on deep learning-based automated lunar crater detection. *Earth Science Informatics*, 17(5):3863–3898, July 2024. doi:10.1007/s12145-024-01396-2.
- [6] Atal Tewari, K Prateek, Amrita Singh, and Nitin Khanna. Deep Learning based Systems for Crater Detection: A Review, 2023. URL: <https://arxiv.org/abs/2310.07727>, arXiv:2310.07727.
- [7] Stuart J. Robbins, Irene Antonenko, Michelle R. Kirchoff, Clark R. Chapman, Caleb I. Fassett, Robert R. Herrick, Kelsi Singer, Michael Zanetti, Cory Lehan, Di Huang, and Pamela L. Gay. The variability of crater identification among expert and community crater analysts. *Icarus*, 234:109–131, 2014. doi:10.1016/j.icarus.2014.02.022.
- [8] Ebrahim Emami, George Bebis, Ara Nefian, and Terry Fong. *Automatic Crater Detection Using Convex Grouping and Convolutional Neural Networks*, page 213–224. Springer International Publishing, 2015. doi:10.1007/978-3-319-27863-6_20.
- [9] Atal Tewari, Vinay Verma, Pradeep Srivastava, Vikrant Jain, and Nitin Khanna. Automated Crater detection from Co-registered optical images, elevation maps and slope maps using deep learning. *Planetary and Space Science*, 218:105500, 2022. doi:10.1016/j.pss.2022.105500.
- [10] Atal Tewari, Vikrant Jain, and Nitin Khanna. Automatic crater shape retrieval using unsupervised and semi-supervised systems. *Icarus*, 408:115761, 2024. doi:10.1016/j.icarus.2023.115761.
- [11] Jing Nan, Yixin Wang, Kaichang Di, Bin Xie, Chenxu Zhao, Biao Wang, Shujuan Sun, Xiangjin Deng, Hong Zhang, and Ruiqing Sheng. YOLOv8-LCNET: An Improved YOLOv8 Automatic Crater Detection Algorithm and Application in the Chang'e-6 Landing Area. *Sensors*, 25(1), 2025. doi:10.3390/s25010243.
- [12] Riccardo La Grassa, Gabriele Cremonese, Ignazio Gallo, Cristina Re, and Elena Martellato. YOLOLens: A Deep Learning Model Based on Super-Resolution to Enhance the Crater Detection of the Planetary Surfaces. *Remote Sensing*, 15(5), 2023. doi:10.3390/rs15051171.
- [13] Lingli Mu, Lina Xian, Lihong Li, Gang Liu, Mi Chen, and Wei Zhang. YOLO-Crater Model for Small Crater Detection. *Remote Sensing*, 15(20), 2023. doi:10.3390/rs15205040.
- [14] Jionghao Zhu, Jiarui Liang, and Xiaolin Tian. Lunar Impact Crater Detection Based on Yolo V7 Using Multi-source Data. In *2023 IEEE International Conference on Control, Electronics and Computer Technology (ICCECT)*, pages 901–905, 2023. doi:10.1109/ICCECT57938.2023.10140508.

- [15] Shuowei Zhang, Peng Zhang, Juntao Yang, Zhizhong Kang, Zhen Cao, and Ze Yang. Automatic detection for small-scale lunar impact crater using deep learning. *Advances in Space Research*, 73(4):2175–2187, 2024. Synergistic Use of Remote Sensing Data and In-Situ Investigations to Reveal the Hidden Secrets of the Moon. doi:10.1016/j.asr.2023.05.041.
- [16] Yutong Jia, Zhijuan Su, Gang Wan, Lei Liu, and Jia Liu. AE-TransUNet+: An Enhanced Hybrid Transformer Network for Detection of Lunar South Small Craters in LRO NAC Images. *IEEE Geoscience and Remote Sensing Letters*, 20:1–5, 2023. doi:10.1109/LGRS.2023.3294500.
- [17] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, and Yuyin Zhou. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation, 2021. URL: <https://arxiv.org/abs/2102.04306>, arXiv:2102.04306.
- [18] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, and Sydney von Arx et al. On the Opportunities and Risks of Foundation Models, 2022. arXiv:2108.07258.
- [19] Iraklis Giannakis, Anshuman Bhardwaj, Lydia Sam, and Georgios Leontidis. A flexible deep learning crater detection scheme using Segment Anything Model (SAM). *Icarus*, 408:115797, 2024. doi:10.1016/j.icarus.2023.115797.
- [20] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment Anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, October 2023.
- [21] Xinyu Dong, Qi Wang, Hongyu Deng, Zhenguo Yang, Weijian Ruan, Wu Liu, Liang Lei, Xue Wu, and Youliang Tian. From Global to Hybrid: A Review of Supervised Deep Learning for 2-D Image Feature Representation. *IEEE Transactions on Artificial Intelligence*, 6(6):1540–1560, 2025. doi:10.1109/TAI.2025.3526138.
- [22] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in Vision: A Survey. *ACM Computing Surveys*, 54(10s):1–41, January 2022. doi:10.1145/3505244.
- [23] Shrishti Shah and Jitendra Tembhurne. Object detection using convolutional neural networks and transformer-based models: a review. *Journal of Electrical Systems and Information Technology*, 10(1), November 2023. doi:10.1186/s43067-023-00123-z.
- [24] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *CoRR*, abs/2010.11929, 2020. arXiv:2010.11929.
- [25] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. Simple Open-Vocabulary Object Detection. In *ECCV 2022*, page 728–755, Berlin, Heidelberg, 2022. Springer-Verlag. doi:10.1007/978-3-031-20080-9_42.
- [26] Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS ’23*, Red Hook, NY, USA, 2023. Curran Associates Inc.
- [27] Patrick Bauer, Marius Schwindinger, Florian Renk, Andreas Weinmann, and Hichem Snoussi. Small lunar crater detection using a few-shot object detection approach with Vision Transformer. In Jixin Ma, editor, *Sixth International Conference on Computer Vision and Information Technology (CVIT 2025)*, volume 13796, page 1379604. International Society for Optics and Photonics, SPIE, 2025. doi:10.1117/12.3077849.
- [28] Neil Houlsby, Andrea Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mohammad Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019.
- [29] Xiang Lisa Li and Percy Liang. Prefix-Tuning: Optimizing Continuous Prompts for Generation. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, 2021.
- [30] Brian Lester, Rami Al-Rfou, and Noah Constant. The Power of Scale for Parameter-Efficient Prompt Tuning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi:10.18653/v1/2021.e mnlp-main.243.

- [31] Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. BitFit: Simple Parameter-efficient Fine-tuning for Transformer-based Masked Language-models. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–9, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-short.1.
- [32] Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. AdapterFusion: Non-destructive task composition for transfer learning. In *Proceedings of The 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2021)*, Virtual Conference, 2021. Association for Computational Linguistics.
- [33] Xiang Lisa Liu, Aaron Tamkin, Xinyun Wu, Aakanksha Chowdhery, Sharan Narang, Aditi Raghunathan, Jared Kaplan, Dario Amodei, and Percy Liang. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *arXiv preprint arXiv:2205.05638*, 2022.
- [34] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models, 2021. URL: <https://arxiv.org/abs/2106.09685>, doi: 10.48550/ARXIV.2106.09685.
- [35] European Space Agency. Gaming to the Moon, 2024. URL: <https://blogs.esa.int/exploration/gaming-to-the-moon/>.
- [36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17*, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc.
- [37] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423.
- [38] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.
- [39] Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. A survey of transformers. *AI Open*, 3:111–132, 2022. doi: 10.1016/j.aiopen.2022.10.001.
- [40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021.
- [41] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé J’egou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9630–9640, 2021.
- [42] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. PaLI: A Jointly-Scaled Multilingual Language-Image Model, 2023. URL: <https://arxiv.org/abs/2209.06794>, arXiv:2209.06794.
- [43] Chufeng Tan, Xing Xu, and Fumin Shen. A Survey Of zero shot detection: Methods and applications. *Cognitive Robotics*, 1:159–167, 2021. doi: 10.1016/j.cogr.2021.08.001.
- [44] Peter H. Cadogan. Automated precision counting of very small craters at lunar landing sites. *Icarus*, 348:113822, 2020. doi: 10.1016/j.icarus.2020.113822.
- [45] Xingxing Weng, Chao Pang, and Gui-Song Xia. Vision-Language Modeling Meets Remote Sensing: Models, datasets, and perspectives. *IEEE Geoscience and Remote Sensing Magazine*, page 2–50, 2025. doi: 10.1109/mgs.2025.3572702.

- [46] Muying Luo, Tao Zhang, Shiqing Wei, and Shunping Ji. SAM-RSIS: Progressively Adapting SAM With Box Prompting to Remote Sensing Image Instance Segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024. doi:10.1109/TGRS.2024.3460085.
- [47] H. W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97, March 1955. doi:10.1002/nav.3800020109.
- [48] Zhaohui Zheng, Ping Wang, Dongwei Ren, Wei Liu, Rongguang Ye, Qinghua Hu, and Wangmeng Zuo. Enhancing Geometric Factors in Model Learning and Inference for Object Detection and Instance Segmentation. *IEEE Transactions on Cybernetics*, 52:8574–8586, 2020.
- [49] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-End Object Detection with Transformers. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I*, page 213–229, Berlin, Heidelberg, 2020. Springer-Verlag. doi:10.1007/978-3-030-58452-8_13.
- [50] Zheng Yuan, Jie Zhang, Shiguang Shan, and Xilin Chen. FullLoRA: Efficiently Boosting the Robustness of Pretrained Vision Transformers. *IEEE Transactions on Image Processing*, 34:4580–4590, 2024.
- [51] M. S. Robinson, S. M. Brylow, M. Tschimmel, D. Humm, S. J. Lawrence, P. C. Thomas, B. W. Denevi, E. Bowman-Cisneros, J. Zerr, M. A. Ravine, M. A. Caplinger, F. T. Ghaemi, J. A. Schaffner, M. C. Malin, P. Mahanti, A. Bartels, J. Anderson, T. N. Tran, E. M. Eliason, A. S. McEwen, E. Turtle, B. L. Jolliff, and H. Hiesinger. Lunar Reconnaissance Orbiter Camera (LROC) Instrument Overview. *Space Science Reviews*, 150(1–4):81–124, January 2010. doi:10.1007/s11214-010-9634-2.
- [52] Stuart J. Robbins. A New Global Database of Lunar Impact Craters > 1–2 km: 1. Crater Locations and Sizes, Comparisons With Published Databases, and Global Analysis. *Journal of Geophysical Research: Planets*, 124(4):871–892, April 2019. doi:10.1029/2018je005592.
- [53] Barret Zoph, Ekin D. Cubuk, Golnaz Ghiasi, Tsung-Yi Lin, Jonathon Shlens, and Quoc V. Le. *Learning Data Augmentation Strategies for Object Detection*, page 566–583. Springer International Publishing, 2020. doi:10.1007/978-3-030-58583-9_34.
- [54] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017.
- [55] Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A. Kalinin. Albumentations: Fast and Flexible Image Augmentations. *Information*, 11(2):125, February 2020. doi:10.3390/info11020125.
- [56] Ari Silburt, Mohamad Ali-Dib, Chenchong Zhu, Alan Jackson, Diana Valencia, Yevgeni Kissin, Daniel Tamayo, and Kristen Menou. Lunar crater identification via deep learning. *Icarus*, 317:27–38, 2019. doi:10.1016/j.icarus.2018.06.022.