University of West Bohemia
Faculty of Applied Sciences
Department of Cybernetics

Master's thesis

Detection of Archaeological Sites Using Remote Sensing

Plzeň 2025 Pavel Balda

ZÁPADOČESKÁ UNIVERZITA V PLZNI
Fakulta aplikovaných věd
Akademický rok: 2024/2025

ZADÁNÍ DIPLOMOVÉ PRÁCE

(projektu, uměleckého díla, uměleckého výkonu)

Jméno a příjmení:	**Bc. Pavel BALDA**
Osobní číslo:	**A23N0047P**
Studijní program:	**N0714A150011 Kybernetika a řídicí technika**
Specializace:	**Umělá inteligence a automatizace**
Téma práce:	**Detekce archeologických nalezišť pomocí dálkového průzkumu Země**
Zadávající katedra:	**Katedra kybernetiky**

Zásady pro vypracování

1. Seznamte se s přístupy pro detekci archeologických nalezišť z obrazových dat.
2. Zvolte vhodný dataset a jeho data předzpracujte.
3. Implementujte vybraný algoritmus pro detekci archeologických nalezišť.
4. Zvolený algoritmus natrénujte a vyhodnoťte.
5. Zhodnoťte dosažené výsledky pomocí zvolených metrik.

Rozsah diplomové práce: **40-50 stránek A4**
Rozsah grafických prací:
Forma zpracování diplomové práce: **tištěná/elektronická**
Jazyk zpracování: **Angličtina**

Seznam doporučené literatury:

Seznam doporučené literatury bude dodán vedoucím DP.

Vedoucí diplomové práce: **Ing. Ivan Gruber, Ph.D.**
Výzkumný program 1

Datum zadání diplomové práce: **1. října 2024**
Termín odevzdání diplomové práce: **19. května 2025**

Západočeská univerzita v Plzni
Fakulta aplikovaných věd
L.S.

Doc. Ing. Miloš Železný, Ph.D.
děkan

Doc. Dr. Ing. Vlasta Radová
vedoucí katedry

V Plzni dne 1. října 2024

Declaration

I hereby declare that this master's thesis is completely my own work and that I used only the cited sources.

Plzeň, 19th May 2025

Pavel Balda

Abstract

This thesis explores deep learning approaches for detecting archaeological cropmarks in aerial imagery. To address the challenge of limited annotated data, a specialized dataset from the Czech Republic was curated and extended with newly identified locations. The research compares convolutional neural network models and develops synthetic data generation methods, including procedurally generated cropmark patterns and neural generation using fine-tuned diffusion models. Experimental results demonstrate that pre-training on procedurally generated synthetic data followed by fine-tuning on real cropmark images yields optimal performance, achieving an F1 score of 0.673 on the test set. Validation on real-world orthophotos confirms practical applicability, with the model successfully identifying 75% of known archaeological test sites across multiple years of imagery.

Abstrakt

Tato práce zkoumá přístupy hlubokého učení pro detekci archeologických nalezišť pomocí vegetačních příznaků na leteckých snímcích. Pro řešení problému omezeného množství anotovaných dat byl sestaven obrázkový dataset lokalit v České republice. Ten byl následně rozšířen o nově identifikované lokality. Výzkum porovnává natrénované modely konvolučních neuronových sítí a vyvíjí metody generování syntetických dat - procedurálně generovaných vzorů vegetačních příznaků a neurálně generovaných obrázků pomocí difúzních modelů. Experimentální výsledky ukazují, že předtrénování na procedurálně generovaných syntetických datech s následným dotrénováním na reálnýchsnímcích vede k optimálnímu výkonu a dosahuje skóre F1 0,673 na testovací sadě. Validace na reálných ortofotosnímcích potvrzuje praktickou použitelnost - model úspěšně identifikoval 75 % archeologických testovacích lokalit napříč několika lety snímkování.

Contents

1 Introduction

1.1 Motivation and significance of the topic

Archaeology investigates past human societies by analyzing their material remains, including structures, tools, and various artifacts. These remnants, collectively referred to as material culture, provide crucial insights into the ways of life of previous civilizations [1].

Archaeology benefits society by authenticating historical artifacts and reconstructing credible narratives of the past. It provides tangible connections to history, fostering awareness and appreciation. Through systematic research, archaeology dispels myths, enhances critical thinking, and offers insights into cultural and technological evolution. Engaging the public through museums and media strengthens historical literacy and helps draw lessons from past societies. Ultimately, archaeology preserves our shared heritage, deepens cultural identity, and informs our understanding of the present [2].

One of the primary topics of interest in archaeology are archaeological sites. These are locations where material historical heritage has been preserved. Investigating these sites using archaeological survey methods can provide valuable insights into past settlements and contribute to a deeper understanding of historical knowledge.

The targeted discovery of archaeological sites is a challenging task in many aspects. Archaeological excavations are often preceded by a multi-step process aimed at maximizing the efficiency of subsequent invasive investigations [3]. In some cases, archaeological sites are discovered by chance, for example, during expansion of the infrastructure, declining water level, or agricultural activities. However, modern approaches to archaeology increasingly rely on non-invasive site detection methods, particularly leveraging the rapidly advancing techniques of remote sensing [4].

Remote sensing has been employed since the 19th century [5]. For archaeological research purposes, it has been used for more than a century [6] (as evidenced by the 1906 image of Stonehenge in figure 1.1). In recent years, we have witnessed rapid advancements in computing technology and machine learning methods, which enable the processing of image data faster than ever before. This progress has opened new opportunities for applications in numerous fields, including archaeology.

Figure 1.1: An image of Stonehenge taken from a military balloon in 1906, likely the earliest surviving aerial photograph of an archaeological site in England from National Monuments Record for England. [6]

1.2 Objectives of the thesis

The primary objective of this thesis is to explore and evaluate machine learning methods for the identification of archaeological cropmarks in high-resolution aerial imagery. The work aims to build an end-to-end pipeline capable of identifying relevant locations in orthophotos of the Czech Republic.

To this end, the following specific goals were pursued:

- Collect, clean, and structure a geographically and temporally diverse dataset of annotated aerial images suitable for supervised learning.
- Design and evaluate a neural classification model for binary cropmark detection, including multiple training scenarios and hyperparameter settings.
- Investigate the impact of various data augmentation and domain transfer techniques to improve generalization to real-world data.
- Validate and compare different experimental setups using consistent metrics (F1 score, precision, recall), and evaluate models not only at the tile level but also at the level of real archaeological sites across multiple years.
- Reflect on the challenges and limitations of current approaches and propose viable directions for future work, including potential improvements to dataset quality, evaluation protocols, and model architectures.

1.3 Structure of the Thesis

The thesis is divided into seven chapters. Chapter 1 introduces the motivation, objectives, and archaeological context. Chapter 2 outlines the theoretical background in computer vision and neural networks relevant to the task and Chapter 3 provides an overview of related work.

Chapter 4 describes the raw data and its pre-processing and image acquisition. Chapter 5 details the classification models, loss functions, metrics, and synthetic data pipelines.

Chapter 6 presents the experimental results across various training strategies and also evaluates model performance on real aerial data in automated task. Chapter 7 concludes the thesis' outputs.

Some visual outputs and image-based results are provided in Appendix B.

1.4 Technology

In addition to the technologies discussed in Sections 2 and 5, the author utilized the following tools for programming and data processing: the Python programming language [7], and frameworks including PyTorch [8], Torchvision [9], Torchmetrics [10], Scikit-learn [11], OpenCV [12], and the Diffusers framework [13]. The Albumentations library [14] was used for data augmentation. Google Colaboratory [15] and its run environments (T4, L4, A100) were used to train several models.

For monitoring and tracking experiments, the tools Weights & Biases [16] and tqdm [17] were employed. Visualization of results and creation of figures was performed using Matplotlib [18]. For programming assistance, translation, information retrieval, and writing support, the author made use of publicly available large language model assistants, specifically ChatGPT [19] and Claude.ai [20].

For version control and code management, the Git versioning system [21] and Github environment [22] were used throughout the project. A partial representative repository of this thesis is (or will be) publicly available on GitHub `https://github.com/pavbal/MT-Detection-of-Archaeological-Sites-Using-Remote-Sensing`.

2 Theoretical Background

This chapter provides an overview of key concepts and methods in remote sensing, archaeology, and computer vision and machine learning that underpin the approaches used in this thesis.

2.1 Remote sensing

Remote sensing refers to the collection and analysis of data on objects or phenomena without direct physical contact. This broad concept includes applications in various fields, such as medical imaging (e.g., X-rays, MRI) and environmental monitoring [23].

For the purpose of this study, the term remote sensing will be understood as the remote sensing of the Earth's surface using imagery acquired by aircraft or satellites. The measured variables will be considered as electromagnetic radiation at various wavelengths, most commonly visible light. However, there are also remote archaeological sensing methods that utilize the detection of reflected mechanical waves (e.g., sonar technology [24]).

Remote sensing methods (or sensors) are generally classified into *active* and *passive* categories. Passive methods rely on naturally emitted or reflected radiation from the Earth's surface across various spectra (e.g., orthophotos, infrared radiation). In contrast, active methods utilize an artificial source of radiation, such as radio waves, laser beams, or ultrasound, to illuminate the target and measure the reflected signal. [25].

In the context of this study, remote sensing represents a fundamental method for non-invasive surface investigation. With rapidly advancing technology, we are witnessing significant developments in the necessary tools and an increasing prominence of remote sensing across various disciplines, including geography [26], ecology and nature conservation [27][28], geology and mineral exploration [29], agriculture and precision farming [30] and military purposes [31]. Since the inception of this scientific field, remote sensing has also played a crucial role in archaeology.

2.1.1 Remote Sensing in Archaeology

Remote sensing in archaeology utilizes two types of data: oblique aerial imagery and vertical aerial imagery. Each of these image types possesses distinct qualitative properties. [4]

Oblique imagery captures the Earth's surface from a perspective other than the vertical, providing a more detailed representation of terrain relief and elevation differences. This perspective also allows for the identification of features on prismatic and overhanging structures. The acquisition of oblique images requires careful selection of the aircraft's flight altitude, the angle between the sensor and the Earth's surface, and the direction of capture. If an oblique image is to be georeferenced (e. g., for use in GIS), rectification is necessary. [4]

Due to the specific requirements of oblique imagery, archaeologists typically capture these images themselves [32].

Vertical (metric) imagery captures the Earth's surface from a nadir perspective (top view). As a result, elevation information is lost in the raw data. Satellite imagery also experiences distortion, necessitating correction to align with the coordinate system. However, many sources provide pre-processed, georeferenced data. These vertical images are particularly suitable for measuring object dimensions and vectorization without the need for additional preprocessing [4].

2.1.2 Aerial LiDAR Scanning of Landscapes

LiDAR (Light Detection and Ranging) is a method used to obtain a georeferenced elevation map of terrain and anthropogenic structures. It is an active remote sensing technique that determines spatial positions based on the specific properties of reflected laser beams. [32]

LiDAR is one of the most effective methods for surveying cultural and historical landscapes. It enables the identification and description of anthropogenic structures preserved on the Earth's surface, such as burial mounds, ditches, embankments, abandoned villages and field systems, charcoal kilns, and ancient roads. However, since laser beams do not penetrate solid materials, LiDAR cannot be used to detect fully buried structures. [32]

The raw data obtained from LiDAR scanning take the form of a point cloud and are typically affected by noise resulting from multiple reflections of laser beams from various obstacles. The laser beam width can reach several centimeters on the Earth's surface, contributing to inaccuracies. Vegetation, with its high variability, is a particularly significant source of noise. Furthermore, LiDAR data are influenced by atmospheric conditions and seasonal changes. [32]

Therefore, substantial preprocessing of the raw data is required. Points in the point cloud are classified into categories such as terrain, vegetation, buildings, and overhead structures. Generally, the last recorded laser reflection corresponds to the terrain. Auto-

mated filters, such as low-pass filters, and interpolation algorithms (e.g., irregular triangulation) are used for processing. The result is a *digital surface model (DSM)*, which includes buildings and infrastructure, and a *digital terrain model (DTM)*, which represents the bare Earth's surface. [32]

Both DSM and DTM can be transformed into standard digital image data, making them suitable for further processing using machine learning methods.

2.1.3 Spectral Bands Beyond RGB

In remote sensing, it is possible to capture spectral bands beyond the visible light spectrum. Objects may exhibit significant variation in non-visible spectral regions. Many objects that appear visually indistinguishable to the human eye can possess distinct properties in invisible radiation bands, allowing them to be identified and differentiated based on these characteristics. A compelling example of this can be found in the classification of forest vegetation species using multispectral data [33] [34].

In the domain of archaeological site detection, vegetation indices are frequently employed. One of the most widely used indices is the Normalized Difference Vegetation Index (NDVI). NDVI serves as a metric for assessing vegetation health and density based on sensor data, and it is derived from two spectral bands: red and near-infrared (NIR). [25]

2.2 Archaeology and archaeological site

Archaeology is defined as the study of the human past through material traces that have survived [35]. These traces can be found in various locations and under diverse conditions across the world.

These specific places, where past human activity has been preserved (*archaeological sites*) are considered key to understanding historical contexts. Consequently, their identification is one of the primary objectives of archaeological research.

2.2.1 Cropmarks

A significant method for identifying archaeological sites involves detecting so-called vegetation (crop)marks in aerial imagery using remote sensing. These marks appear most commonly on arable land, although they can also be observed in pastures and mowed meadows [36]. They provide valuable information about buried structures beneath the surface [36]. The existence of cropmarks has been known since the 16th century, but

they were not systematically described until the first half of the 20th century [36]. Cropmarks are caused by local changes in biomass production in the crop and can therefore be classified into two fundamental categories: negative and positive, based on their relative change of biomass production.

Positive cropmarks emerge due to the presence of past subsurface (sunken) structures, such as pits, ditches or graves, which could now retain water and nutrients [36]. In these areas, the soil layer is typically thicker, leading to the formation of visible patterns on agricultural land characterized by taller, darker, and / or denser vegetation caused by increased biomass production [37]. A color difference can also be observed, resulting from local variations in soil composition and nutrient availability, which affect the pigmentation of plant tissues. [37]

In contrast, *negative cropmarks* arise from the remnants of above-ground structures, such as walls or roads, buried beneath the soil surface [36]. These locations usually have a thinner soil layer than the surrounding area, and therefore less space for root growth and less available nutrients, resulting in visible patterns typically formed by shorter, lighter, and / or sparser vegetation [37].

Due to the nature of cropmarks, oblique aerial photography plays a significant role in aerial archaeology.

The nature of cropmark formation is illustrated in Figure 2.1.

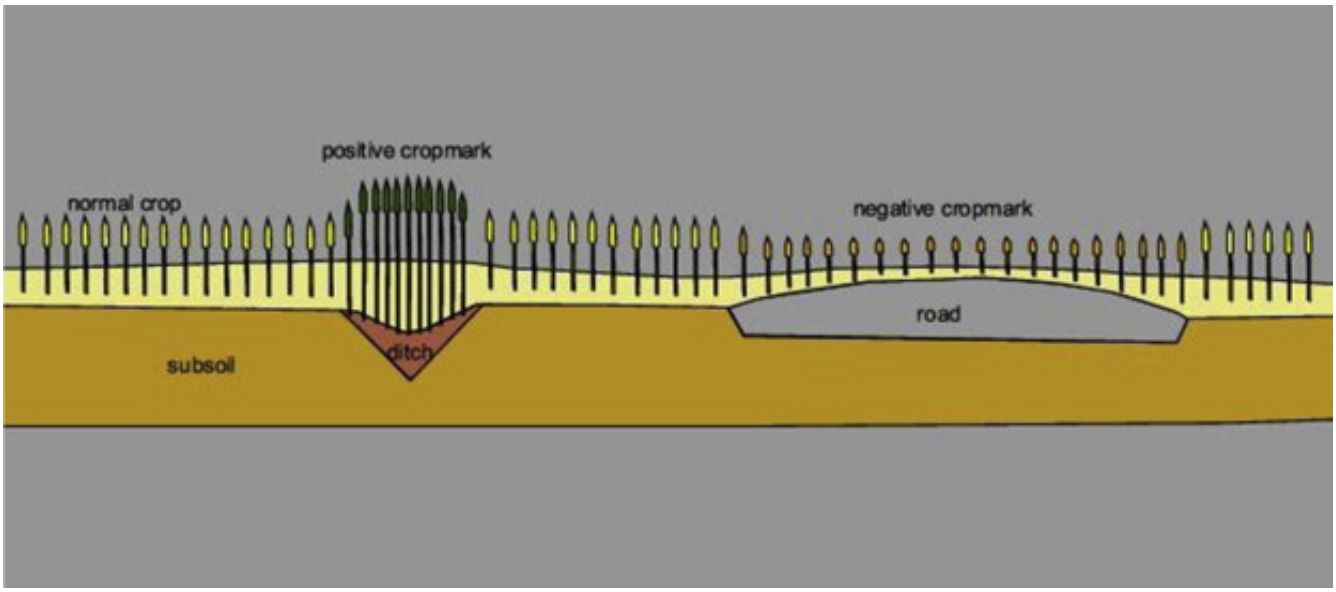

Figure 2.1: A diagram illustrating the nature of positive and negative cropmarks [38].

An important technology for the identification of cropmarks is imaging in spectral bands beyond the visible spectrum enabling more accurate detection of qualitative differences in vegetation (see Section 2.1.3).

Challenges in Cropmark Identification

A key challenge in identifying circular cropmarks is the presence of various phenomena that produce similar visual effects. One such phenomenon is the *fairy ring*, which results from the concentric growth of fungi or fungal infections in grass [39]. However, due to their natural growth pattern, Fairy rings tend to expand annually, which can aid in distinguishing them from true archaeological features. An example of fairy ring growth is shown in Figure 2.2.

Other similar formations may arise from tractor tracks, irregular fertilization, or circular irrigation, as illustrated in Figure 2.3. These phenomena can introduce type I errors (false positives) in the automated detection of archaeological sites, potentially leading to misclassification.

Examples of cropmarks resulting from subsurface human-made structures are shown in Figure 2.4. These cropmarks are captured in orthophotographic maps available on the Mapy.cz portal.

Figure 2.2: The image illustrates the temporal development of a circular fungal infection in grass (commonly referred to as a Fairy Ring) at coordinates 50.4118400N, 14.1883908E. This formation bears a striking resemblance to patterns typically associated with archaeological sites, posing a potential challenge in automated detection. The varying visibility of the feature across different years is most likely due to differences in the timing of aerial imaging, affecting the detectability of vegetation anomalies. Location: Nížebohy, Czech Republic. Source: Mapy.cz.

2.3 Computer vision in remote sensing and aerial archaeology

Since 2012, when the revolutionary approach to computer vision AlexNet was introduced [40], both fundamental and advanced neural network architectures have found significant applications in remote sensing tasks. In the domain of aerial data, most standard computer vision tasks (e.g., classification [41], segmentation [42], detection [43]) are

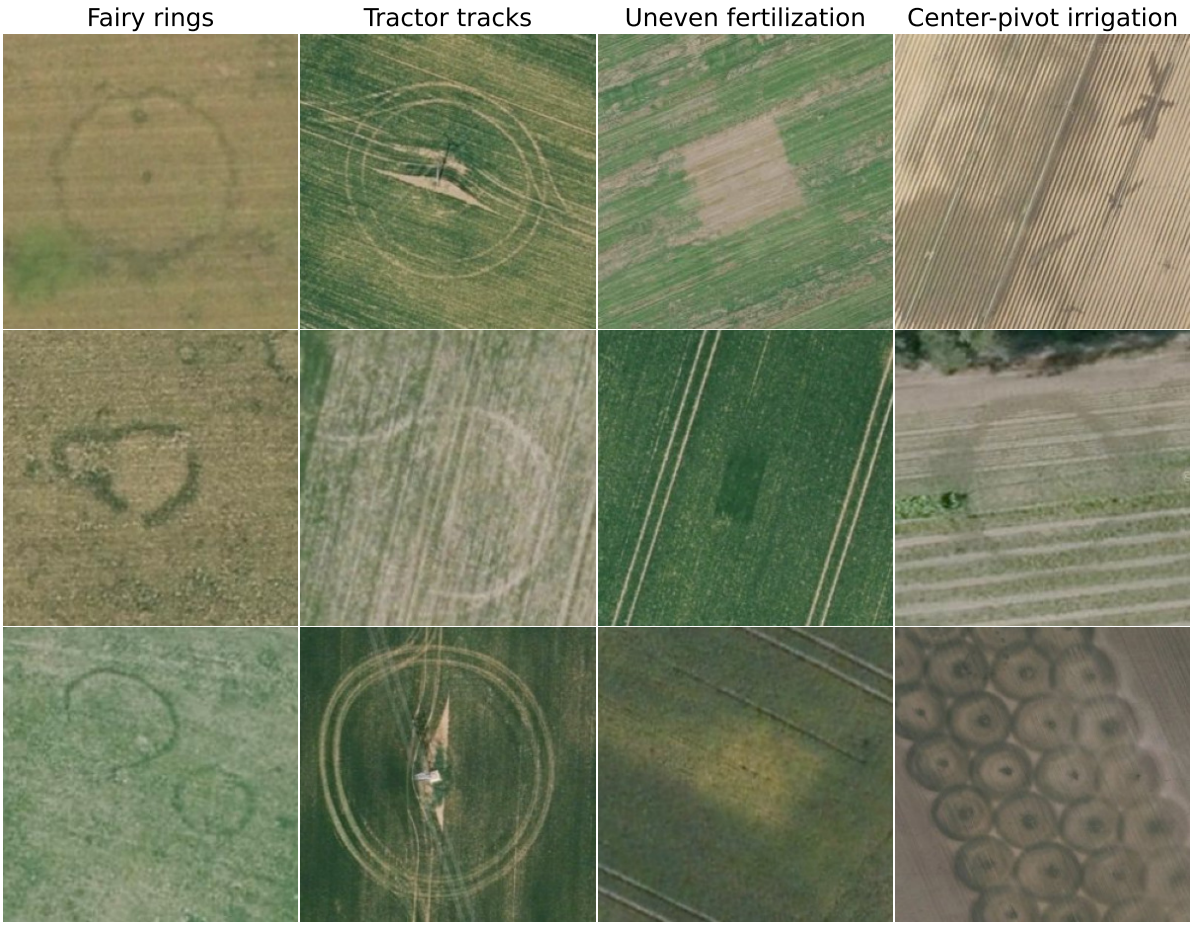

Figure 2.3: Illustration of a selected set of phenomena that resemble the targeted features but originate from different causes and are not relevant to our analysis. These phenomena can contribute to Type I errors (false positives). From left to right: circular fungal infection in grass (Fairy Rings), tractor tracks (e.g., around impassable objects), uneven fertilization (e.g., by agrochemical sprayers), and patterns formed by circular irrigation systems. Source: Mapy.cz.

applied, along with more advanced techniques such as synthetic data generation [44], change detection [45], and predictive modelling [46].

These methods have naturally extended to the field of archaeology, where they are primarily used for the identification and mapping of archaeological sites. Given the vast availability of aerial data and the labor-intensive nature of manually locating archaeological sites, the most valuable application in aerial archaeological surveying is the task of automated *detection*.

The greatest significance of detection methods for archaeologists lies in the automatic pre-selection of locations suspected to contain archaeological sites. This process reduces the amount of data that experts (archaeologists) need to manually review in order to identify new sites.

Given this context, it is crucial to minimize *Type II errors* (false negatives), as it

is preferable to have more false positives rather than risk missing actual archaeological sites. However, it is also necessary to ensure that the resulting dataset does not exhibit an excessively high *Type I error* rate (false positives). Essentially, this presents the classic trade-off between *recall* and *precision*, where achieving the optimal balance is critical for practical application.

Recent research demonstrates that AI-based approaches - particularly convolutional neural networks (CNNs), LiDAR, and ground-penetrating radar (GPR) - can significantly improve this balance. [47]

2.3.1 Detection

The task of *detection* generally involves determining the location of a specific object. In the context of computer vision, this task is understood as identifying the object of interest within an image using a bounding box. Detection often extends the classical image classification problem, which is inherently included in detection tasks (at a minimum, classification is represented by a binary distinction between objects of interest and other objects).

The detection effect can, in certain cases, be achieved through a series of classifications by dividing an image $I[x, y]$ into smaller subimages (tiles) $T_{ij}(a, b)$ of dimensions (a, b) and coordinates $[i, j]$ in the original image. Depending on the nature of the task, these tiles can be classified into a single class ω_{ij} when detecting a specific phenomenon (or into multiple classes that are strictly disjoint within each tile). Each tile can also be classified into multiple classes based on the presence of target objects. Thus, each tile provides information about the classes present within it. In combination with the tile coordinates, this approach yields an approximate location of the detected objects in the original image $I[x, y]$, where the dimensions of each tile form a standardized bounding box for each class contained within the tile. For a tile $T_{ij}(a, b)$, the bounding box has dimensions (a, b) and the top left coordinate at $I[ai, bj]$.

In remote sensing tasks, this classification method is often referred to as *scene classification* and has significant applications in detecting compact objects such as aircraft, swimming pools, sports fields, and vehicles [41][48]. This detection method can be further refined by performing an additional detection task on tiles classified as positive for a certain class. This approach can significantly increase the speed of the algorithm and generally reduces the demand for training data. For improved results, tile overlap can be introduced.

For unbounded objects covering large areas (e.g., forests, fields, water bodies, urban areas), an analogous approach can be used for coarse *semantic segmentation* (one of the

available datasets can be found in [49]). The original image $I[x, y]$ is divided into tiles. Each tile (assumed to be square) $T_{ij}(a)$ of size a pixels is classified into a single class ω_{ij}. This process results in a matrix $\Omega[i, j] = \omega_{ij}$, which represents a *semantic map* with a resolution reduced by a factor of a compared to the original image. This technique enables, for example, the segmentation of lower-resolution satellite images using scene classification from higher-resolution aerial imagery.

2.3.2 Generative Models for Synthetic Data Generation

In computer vision tasks with limited or imbalanced datasets, synthetic data generation has emerged as a powerful tool for improving generalization. Traditional augmentation techniques (e.g., rotation, scaling, color jitter) are often insufficient when dealing with rare patterns or underrepresented classes. Recent research has shifted towards neural generative models capable of synthesizing entirely new samples that mimic the distribution of real data.

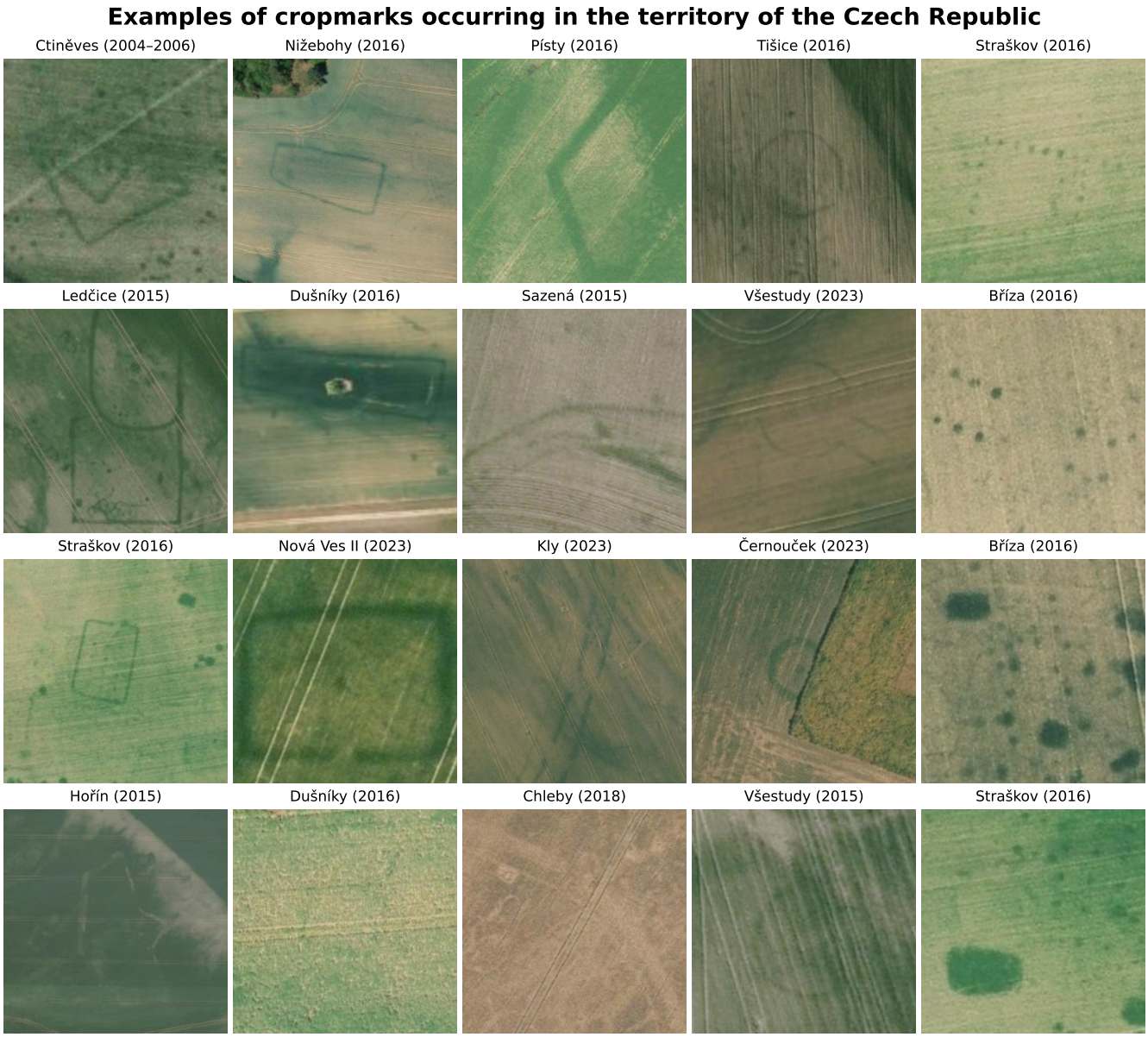

Figure 2.4: Examples of vertical aerial images with cropmarks in Czech Republic. Each image is labeled with the name of the present-day municipality and the year of image acquisition as provided by the Mapy.cz portal. The cropmarks are intuitively grouped by the author into columns based on morphological similarity. Same images were later used as a part of training and test datasets. Data source: Mapy.cz.

3 Related work

The concept of automatic detection of archaeological sites emerged even before the advent of advanced neural network capabilities. Early approaches included detecting circular structures through enhancement techniques followed by template matching [50] or automated detection of subsurface and pit structures using LiDAR images, developed by the same author [51]. In the preceding decades, research has focused primarily on the (automated) extraction of archaeological features (e.g., [52]), driven by advances in remote sensing technologies and image resolution.

Recent research has focused on identifying the optimal spectral space for the detection of features within the range of captured electromagnetic bands. This research has employed *decision tree* algorithms to identify optimal wavelengths and threshold values for classifying spectral signatures into 'healthy' versus 'stressed' vegetation categories [53]. This methodology facilitated the identification of novel combinations of spectral bands - such as those near 570 nm - that are particularly sensitive to vegetation stress induced by subsurface archaeological features.

In the past decade, fueled by the rapid development of computational power and deep learning, there has been a significant increase in the application of machine learning in archaeology. This trend is supported by the growing number of scientific publications on the topic, as illustrated in Figure 3.1 [54].

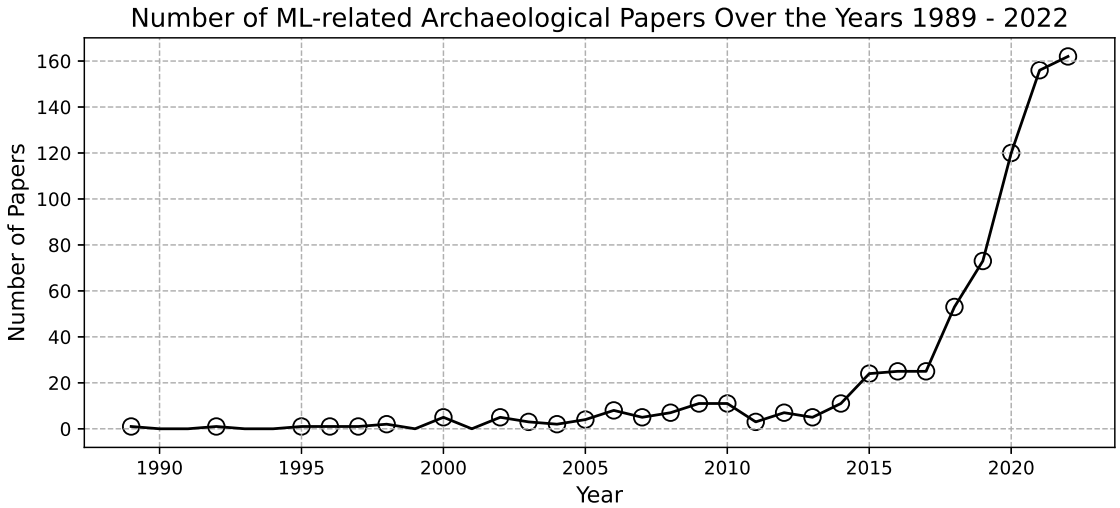

Figure 3.1: Number of Archaeological Scientific Publications Related to Machine Learning from 1989 to 2022, sourced from [54]. The circles in the graph indicate that at least one paper was published in the corresponding year.

Archaeologists have increasingly utilized aerial and satellite imagery processed through convolutional neural networks (CNNs) and other models to enable automated site detection, significantly enhancing the efficiency and scalability of archaeological surveys [25].

These algorithms can identify characteristic pattern anomalies in the image data, such as cropmarks in fields or terrain relics in LiDAR models, which serve as indicators of subsurface structures. Using machine learning techniques, these methods improve the efficiency of archaeological surveys by detecting subtle visual cues that might otherwise be overlooked in manual analysis.

Furthermore, contemporary deep learning models have demonstrated the capacity to analyze aerial and satellite imagery and to automatically detect areas that are likely to contain archaeological features based on subtle visual anomalies present in the data [25].

Recent work ([55]) explores the applicability of deep learning networks for archaeological site detection using high-resolution satellite imagery. The study evaluates two fully convolutional architectures: SegNet and a custom-built 8-layer network (SimpleNet), both trained on manually labeled Google Earth images from Peru. Despite the limited size of the training dataset (500–2000 images), the models, especially SimpleNet, achieved promising results in identifying archaeological features, particularly linear structures such as canals and walls. Although some false positives occurred due to visual similarities with natural formations, the networks still proved useful for large-scale landscape analysis. The authors conclude that even with minimal resources, deep learning can support archaeological prospecting, and advocate open sharing of labeled datasets to accelerate progress in the field. [55]

Another recent work [56] demonstrated the feasibility of detecting circular archaeological structures (ringforts) in Ireland using classical machine learning techniques applied to aerial imagery. Their approach, which combined Histogram of Oriented Gradients (HOG) with Support Vector Machine (SVM), achieved high classification accuracy and supports the general premise that morphologically distinct cropmarks can be successfully identified in orthophotographic data. Although their method differs from the neural architectures employed in this thesis, the study reinforces the value of aerial image analysis for archaeological prospection and highlights the effectiveness of focusing on geometrically well-defined features such as enclosures.

A comprehensive survey [57] outlines recent advances in the classification of Indian monuments using machine learning techniques, with a strong emphasis on deep convolutional neural networks (DCNNs) and their ability to learn architectural styles from images. The paper highlights the transition from classical handcrafted features to learned representations, the use of both global and local descriptors, and the increasing role of transfer learning and mobile-friendly models. Importantly, it includes cases using aerial

and drone imagery, demonstrating the feasibility of automated heritage classification from remotely sensed data. This aligns closely with the goals of this thesis.

3.1 Drone deployment

The deployment of drones equipped with multispectral sensors enables the capture of imagery at moments when such spectral indicators are most pronounced. In particular, during the investigation of the ancient city of Veii in Italy, the combined use of unmanned aerial systems and vegetation index analysis proved to be more effective in revealing small-scale archaeological structures than conventional RGB imagery [58].

3.2 LiDAR

A current trend in archaeological remote sensing is the shift toward semantic segmentation of LiDAR data: rather than simply outputting the coordinates of a detected site, algorithms now predict its semantic map directly. In one study, two deep learning architectures, U-Net [59] and Mask R-CNN [60], were applied to LiDAR imagery of a Maya archaeological site in Guatemala. The U-Net model successfully identified areas of ancient construction activity and even the foundations of buildings with an accuracy of approximately 60–66%, outperforming Mask R-CNN [61]. This demonstrates that deep learning can effectively detect settlement remnants in tropical rainforests, despite dense vegetation and erosion.

Further advancement is illustrated by a study conducted in Spain. By combining LiDAR data with Sentinel-2 multispectral imagery and employing a hybrid algorithm, researchers discovered over 10,000 previously unknown burial mounds in northwestern Iberia [62]. This method achieved a detection rate of 89.5% and precision of 0.97, providing a reliable alternative to manual mapping. Such approaches enable automated scanning of tens of thousands of square kilometers, dramatically increasing the number of sites detected and enhancing the scale of archaeological exploration.

4 Data

In order to effectively apply machine learning techniques to the detection of archaeological sites, a suitable dataset must first be identified and curated. This chapter describes the sources, characteristics, acquisition, and preprocessing steps of the data used in this thesis. The limitations of available data, such as resolution, seasonal variability, and annotation accuracy, are discussed in the context of their impact on downstream model performance.

The dataset used in this thesis consists of annotated aerial images capturing vegetation marks (cropmarks) associated with subsurface archaeological features on the territory of the Czech Republic. However, the data provided are in the form of *geographical coordinates*, specifically latitude and longitude. Therefore, the coordinates provided includes only locations where cropmarks indicating potential archaeological features have been observed. The collection of negative samples was furnished by the author of this thesis.

4.1 Limitations of the Dataset

The dataset used in this thesis was initially limited to a small set of coordinates representing locations where cropmarks had previously been identified. Building upon this foundation, all subsequent data processing steps, ranging from the acquisition of aerial imagery from public sources, through the construction of an image-based dataset, the classification and typology of cropmark structures, to the selection of negative samples and the identification of additional candidate sites, were carried out by the author of this thesis. This involved extensive manual inspection and annotation of orthophotos, as well as refinement of regional categorization for subsequent analysis. The following section describes the origin and characteristics of the original dataset from which this process began.

4.2 Data source

The data used in this study were kindly provided by Professor prof. PhDr. Martin Gojda, CSc., DSc., a leading expert in aerial archaeology at the University of West Bohemia. Professor Gojda has been systematically studying the occurrence of cropmarks in the Czech landscape for several decades and has published extensively on the

topic (e.g., [4], [5], [32], [37], [63]). His research has significantly contributed to the development of non-invasive archaeological survey methods in Central Europe, with a particular emphasis on the use of aerial imagery and remote sensing technologies for the detection and interpretation of buried cultural features.

4.3 Data Extension and Author Contributions

While the original dataset provided a solid foundation for initiating the detection pipeline, it was insufficient in terms of both volume and geographic variety to robustly train and evaluate machine learning models. In order to address these limitations, a systematic effort was undertaken to extend the dataset with additional annotated sites. Examples of cropmarks discovered by the author are shown in Appendix B, Figure B.1. These cropmarks may have been discovered earlier, but the author identified them independently through their own investigation.

The newly added locations were identified through an author's manual survey of high-resolution aerial imagery available on the Mapy.cz portal. The selection strategy focused on regions with a rich history of human settlement, particularly those near known archaeological clusters. For each candidate region, aerial images from multiple years and different seasons were examined, including the most recent from years 2022 and 2023. In many cases, these recent captures proved to be invaluable, offering markedly enhanced contrast and clarity of vegetation marks.

The process relied initially on published morphological typologies and visual patterns as presented in the referenced literature, particularly the comparative graphical table summarizing the identified cropmark shapes in the Czech territory (Table A.1). These references provided a conceptual framework for identifying promising patterns. Subsequently, the annotation decisions were informed by the author's own observational experience, developed iteratively throughout the data curation process.

It is important to acknowledge that the author is not a trained aerial archaeologist. As a result, despite rigorous cross-checking, the newly identified sites may contain a certain degree of noise, particularly in the form of false positives (Type I errors). These occur when naturally occurring patterns or agricultural artifacts are mistakenly interpreted as archaeological features. While this uncertainty is not unexpected in semi-automated or manual data expansion efforts, it may introduce a bias that affects the model's precision and overall performance.

The final coordinate dataset, which includes both the original sites and the newly identified locations, is visualized in Figure 4.1. The merged dataset not only increases the robustness of the learning task but also approximates a more realistic and heterogeneous

set of conditions for evaluating archaeological site detection models.

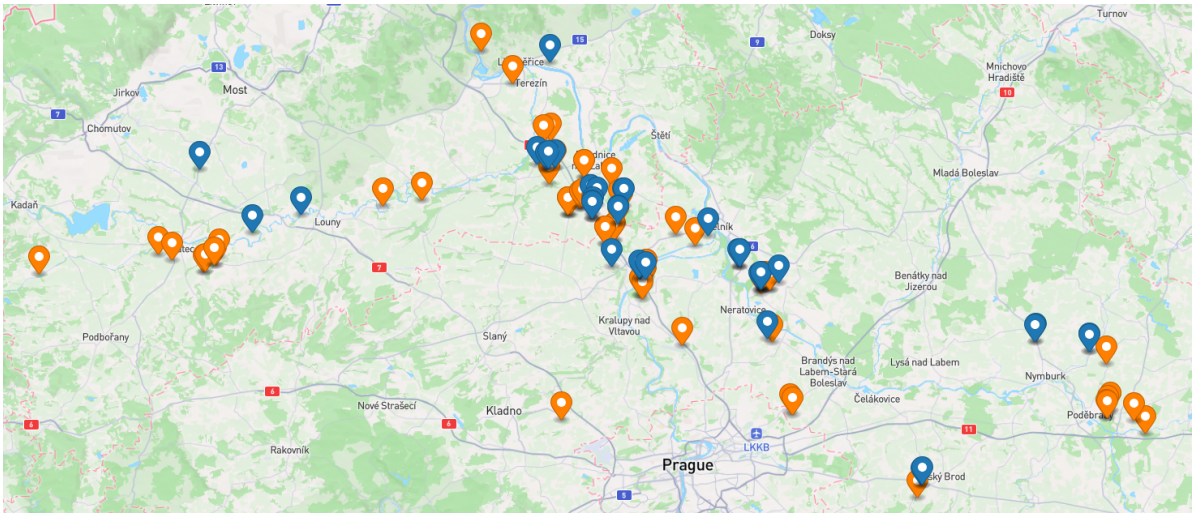

Figure 4.1: Map showing the location of data representing cropmarks within the Czech Republic. The original data (blue) were extended by the author (orange) through the identification of new sites and the subdivision of existing ones due to their large spatial extent. Generated using https://geojson.io.

4.4 Data Characteristics

4.4.1 Geographical Location

The dataset used in this thesis is geographically concentrated in Central and Northern Bohemia, with the highest density of samples located in the Elbe River lowlands (Polabí). Particularly well-represented are the regions of Žatec (Ohře River Basin), Poděbrady, Litoměřice, Veltrusy, Neratovice, Mělník, and Roudnice nad Labem. These areas form a continuous band of fertile agricultural land, where cropmarks appear with higher frequency due to favorable soil composition, stable land use, and long-term cultivation. A map showing the spatial distribution of these locations is provided in Figure 4.1, while representative examples of cropmark patterns are shown in Figure 2.4.

The original dataset, provided by Professor Gojda, comprised approximately six main regions (divided into 24 locations) containing aerial imagery samples with visible archaeological vegetation marks. To enhance the spatial precision and analytical utility of the data, several of the larger regions were subdivided into smaller, more homogeneous zones. Furthermore, the author identified and added approximately seven new regions (divided into several locations) based on the visual detection of cropmarks in publicly available orthophotos described in 4.3.

Despite these efforts to expand the dataset, its geographic distribution remains uneven. This imbalance is not accidental but reflects the historical settlement patterns of the Czech landscape. Consequently, the dataset is naturally biased toward areas with a higher probability of revealing archaeological features, limiting its geographic generalizability but reinforcing its domain-specific relevance.

4.4.2 Visual Characteristics and Typology of Cropmarks

Based on the dataset used in this study, several distinct feature types can be identified:

Circular features typically represent prehistoric enclosures, burial mounds, or ritual circles. These appear as ring-shaped vegetation marks of varying scale.

Rectangular features often indicate the presence of buried foundations. These rectilinear patterns may appear as complete or partial enclosures with defined corners and straight or slightly convex edges.

Linear (border) features may represent paths, roads, field boundaries, or defensive ditches. These appear as elongated vegetation marks of various thickness.

Round borders are circular enclosures with particularly pronounced boundaries, often indicating defensive structures (walls, ditches).

Dotted grids or **patterns** suggest systematic arrangements of postholes, foundation pillars, or organized settlement layouts. These appear as regular arrays of small vegetation marks.

Patches and pits represent discrete subsurface features such as storage pits, waste deposits, graves or other sunken structures. These typically appear in clusters.

Examples of individual cropmark feature types can be found in Figure 2.4.

4.5 Data Splitting Strategy

Given the geospatial nature and typology of the dataset and its intended use for training and validating machine learning models, particular care was taken in the process of dividing the data into training and test sets. This section outlines the reasoning behind the chosen data splitting methodology and describes the steps taken to ensure statistical integrity and geographical relevance.

4.5.1 Spatially-Aware Splitting

The coordinate dataset consists exclusively of positive sample locations where vegetation marks indicative of subsurface archaeological features have been detected. These

samples exhibit strong spatial clustering, both in terms of geographic proximity and morphological similarity. A naive random split of individual locations could result in training and test sets containing near-identical samples, potentially leading to data leakage and overestimation of model performance.

To mitigate these risks, a spatially-aware strategy was employed, based on unsupervised clustering using the K-Means [64] algorithm applied to geographical coordinates (latitude and longitude). This approach allows for the grouping of geographically proximate samples into distinct clusters that can then be used as the basis for the split.

4.5.2 Grid Search for Data Division

Given the exploratory nature of this research and the constraints posed by class balance and morphological diversity, a grid search was conducted to identify suitable clustering configurations. The following parameters were varied:

- Number of clusters.
- Random state of the KMeans algorithm.
- Random seed used for selecting test clusters.
- Number of clusters included in the test set, constrained within a fractional range of total clusters (approximately 16–26%).

For each combination, a split was considered valid only if it met all of the following criteria based on the cropmark feature types (feature types are described in Section 4.4.2):

1. The test set contained between 4 and 8 structures of type *circular* and also *rectangular* (A single clustered location may contain multiple structures of the same type, making it essential to adequately subdivide less numerous or spatially sparse objects to ensure balanced representation).
2. All feature types were represented at least once in the test set
3. The test set never contains more than 40% of the locations for any given type of feature

This filtering process ensured that the test set remained both representative and challenging, while the training set could retain sufficient variability to support model generalization.

4.5.3 Final Split and Visualization

One configuration was selected based on semantic balance and spatial separation. The final split consists of *49 training* and *12 test* locations. The number of unique archaeological sites, defined as distinct combinations of location and feature type, is 77 in the training set and 20 in the test set.

Figures 4.2 and 4.3 illustrate the spatial and semantic characteristics of the resulting datasets.

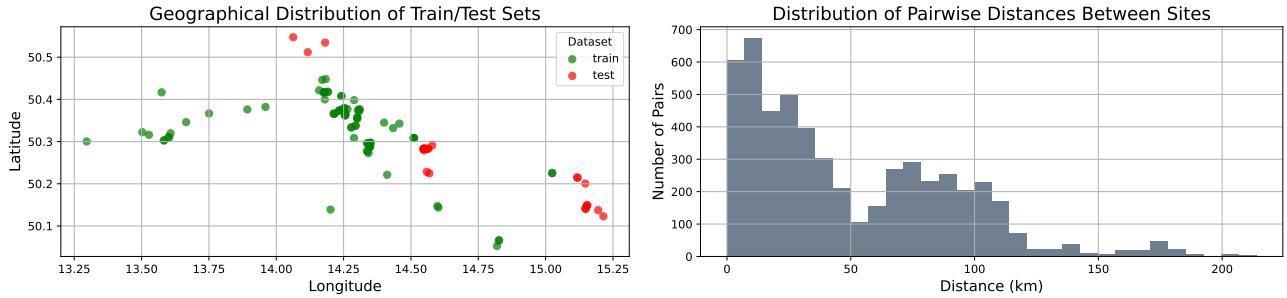

Figure 4.2: Spatial properties of the dataset. (Left) Geographical distribution of training and test locations. (Right) Histogram of pairwise distances between sites, in kilometers This highlighting the spatial clustering of the data. Most sites are located within a relatively narrow distance range.

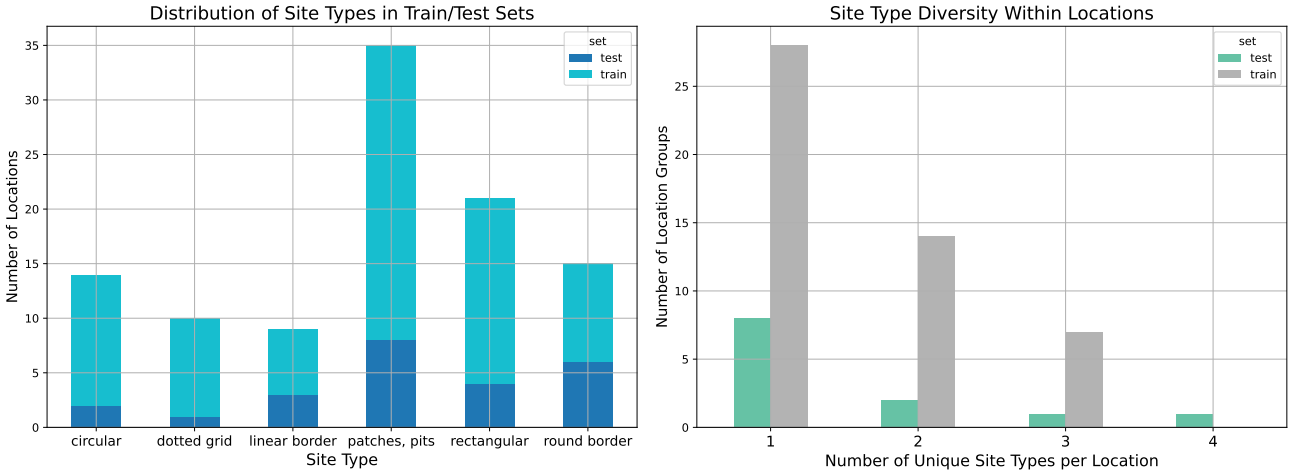

Figure 4.3: Semantic composition of the dataset. (Left) Distribution of site feature types in training and test sets. (Right) Number of location groups by the number of unique site feature types contained within them. This visualization helps assess whether training and test sets reflect comparable morphological variability.

4.6 Image Acquisition and Preprocessing

Since the default data were composed solely of geographic coordinates, the corresponding image data had to be obtained separately. This process was carried out by the author using two complementary strategies. The part of the images (majority of positive class) were manually cropped from the publicly accessible orthophoto viewer available at Mapy.cz. The second strategy was to obtain images through automated export from the official geoportal operated by the Czech State Administration of Land Surveying and Cadastre (ČÚZK).

The goal of the image acquisition step was to capture clear representations of the annotated features while preserving relevant spatial context. As a result of the manual (crop) workflow, the resulting image dimensions vary significantly. Overview of the resolution variance of the positive class is demonstrated later in this chapter in Figures 4.4 and 4.5. In some cases, deliberately larger image crops were used in order to preserve broader spatial context, which may later facilitate data augmentation techniques such as random cropping, rotation or zooming.

All manually collected images were cropped with care to ensure that subsequent random cropping during augmentation would still include at least part of the archaeological feature. This precaution was taken to avoid training the model on effectively negative samples labeled as positive.

Some images (both positive and negative) originating from the Mapy.cz platform contain visible watermarks. Although the presence of watermarks may introduce a potential bias in model training (e.g., through spurious feature learning or dataset leakage), their inclusion was in many cases deemed necessary in order to preserve the clarity and legibility of cropmark patterns.

4.6.1 Cropmark Test Set

To ensure consistency and comparability across all model evaluations, a single test dataset was established and used uniformly throughout the thesis. This section describes the construction, composition, and rationale behind the *cropmark test set*, which serves as the final benchmark for all trained models.

The process of obtaining positive-class images is described in Section 4.6; however, it is still necessary to establish a clear methodology for selecting negative samples.

Negative Sample Construction

Approximately half of the negative samples (circa 170 samples) in the test set were selected from areas geographically close to annotated positive samples. The aim was to test the model's ability to correctly distinguish subtle differences within similar environmental and agricultural contexts.

An additional 25% of the negative samples from the test set were selected from locations geographically distant from any of the annotated positive sites. While some of these regions may still theoretically possess documented historical settlement, the intent was to evaluate the model's ability to generalize its predictions beyond the spatial context of the training data. These distant negatives serve as a check against overfitting to regional characteristics, ensuring that the model can handle input from previously unseen or geographically distinct areas.

The remaining quarter of negative samples was purposefully curated to include visual phenomena that can be mistaken for archaeological cropmarks. These include, but are not limited to, *fairy rings*, *tractor tracks*, *uneven fertilization*, and *center-pivot irrigation*. The typology presented in Figure 2.3 illustrates several such misleading patterns. It should be noted that potential false positives of this kind also appear often within the earlier two segments of the negative sample set. However, this final 25% was specifically assembled to contain clearly non-archaeological, yet structurally ambiguous, patterns placed in randomly selected locations.

Final Construction of the Cropmark Dataset

After constructing the negative samples, we can now establish the image test dataset, referred to as the *cropmark test set*.

The sample images in the *cropmark test set* are present in a positive to negative ratio of 2:7. It consists of 95 images for positive class and 336 images for negative class. This imbalance is intended to approximate the uneven class distribution typically encountered in real-world scenarios.

To clarify, all images in test set were manually curated from the Mapy.cz portal as described in Section 4.6.

4.6.2 Cropmark Dataset

Negative Sample Construction

It is necessary to construct a robust set of train set negative samples. Two complementary strategies were used to build the negative portion of the training set (both

mentioned in Section 4.6).

The first half of the negative samples was obtained through the same method described for constructing the *cropmark test set*. Specifically, the selected images were manually curated from the Mapy.cz portal and included areas with no visible cropmarks. These regions were sampled across diverse spatial and agricultural contexts to ensure generalizability. The proportional distribution of contextual types was roughly preserved: approximately half of the negatives originated from locations near known archaeological sites, while the remaining samples were selected from more distant areas or were chosen for their visual resemblance to cropmarks (e.g., fairy rings, tractor tracks, or irrigation patterns) with a similar intention as in *cropmark test set*. This ensured that the training and test sets remained comparable not only in terms of semantic diversity but also spatial representativeness.

The second half of the negative training samples was acquired through automated image selection described in Section 4.6. For each candidate location - randomly distributed within the Czech Republic and at least 20 km distant from any test site - a historical aerial image was retrieved for a randomly selected year between 2006 and 2022. Each image then underwent a two-stage filtering process. In the first stage, a neural network classifier (see Section 5.2.1) was used to verify whether the image contained a valid aerial view, thereby eliminating empty, blank, or otherwise irrelevant tiles. In the second stage, only those images depicting open fields or grassy areas were retained, as determined by a second neural model (described in Section 5.2.2). This ensured that the final selection of negatives included semantically consistent terrain while avoiding regions that could visually resemble archaeological features.

All automatically obtained negative samples in *cropmark train set* were subsequently manually reviewed to ensure their validity.

Final Construction of the Cropmark Dataset

After defining the negative class and test dataset, the complete *cropmark dataset* was assembled. It consists of three exclusive subsets: the *cropmark train set*, the *cropmark validation set*, and the *cropmark test set*. Test set construction is described in previous section.

The *cropmark train set* has a positive-to-negative ratio of approximately 2:5 (437 positive and 1039 negative), compared to 2:7 (95 positive and 336 negative) in the *cropmark test set*. This milder imbalance was intentionally designed to ensure sufficient representation of the cropmark examples (rare and often subtle) during training. The goal was to support better recall without overwhelming the model with negative samples while still preserving exposure to diverse negative contexts.

To mitigate the risk of the model learning to associate all imagery from the geoportal with the negative class, a small subset (approximately 15%) of positive training examples was also obtained from the same platform. These images are included in reported statistics.

The *cropmark validation set* was manually curated by the author as a subset of the original training data. Care was taken to ensure that it did not contain any images depicting the same cropmark structures as those present in the remaining training set. In addition, this validation set was constructed to include samples from regions geographically close to the training areas. This approach allowed the model to retain access to full coverage of annotated locations during training while still providing an informative and independent validation benchmark. The resulting *cropmark validation set* contains 40 images for positive class and 100 images for negative class (maintaining 2:5 ratio). All models in this thesis trained on *cropmark train set* were validated using *cropmark validation set*.

Figure 4.4 provides an overview of the dimensions and brightness nature of the positive images in the *cropmark train set*.

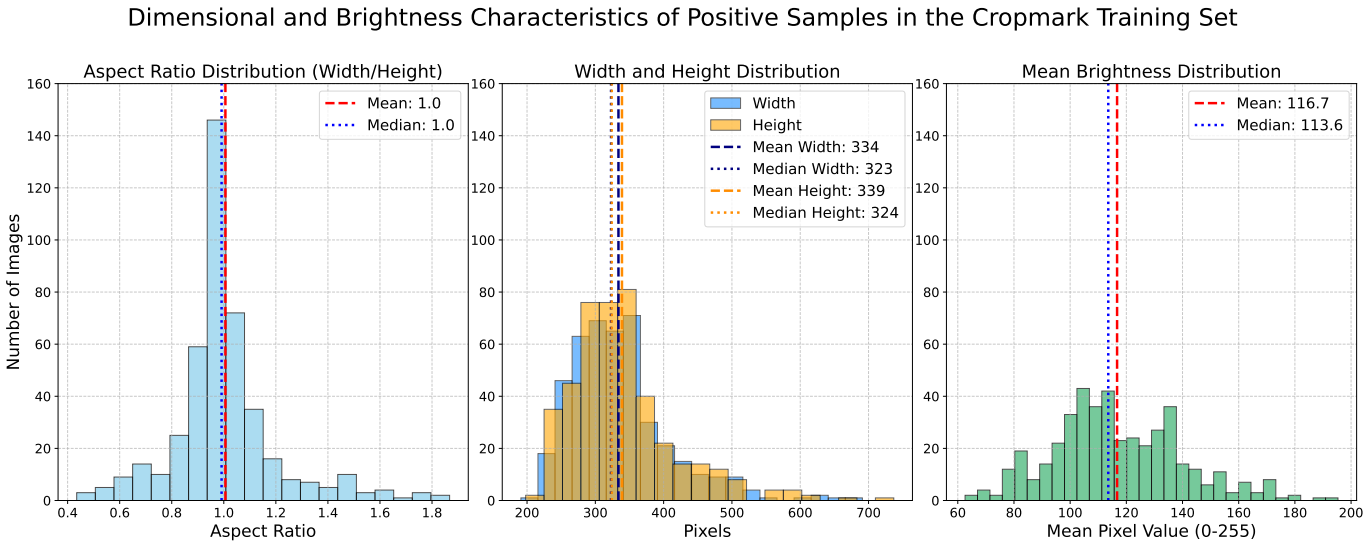

Figure 4.4: The figure summarizes the dimensional and brightness characteristics of positive samples of *cropmark train set*. The aspect ratios are mostly close to 1, indicating near-square crops. The distribution of image sizes shows a clear positive skew, with most samples being relatively small. The brightness distribution is multimodal, with a dominant peak around mid-range pixel values.

4.6.3 Extended Cropmark Dataset

To further enhance the robustness and generalizability of the detection models, an extended training dataset, referred to as the *extended cropmark train set*, was created. This set was constructed by approximately doubling the original number of training samples (in terms of image count) while preserving class balance (2:5). There are 896 images for the positive class and 2236 images for the negative class. The new samples consist of British cropmarks (sourced from the Colfein database [65]). Most of the added samples represent enclosures and defended enclosures of various geometric forms, but other types of archaeological features are also present.

Figure 4.5 provides an overview of the dimensions and brightness nature of the positive images in the *extended cropmark train set*.

The *extended cropmark validation set* was derived as a superset of the original validation set. In addition to all images from the *cropmark validation set*, it includes additional samples selected from the extended training data, following similar principles - namely, avoiding direct cropmark duplication while maintaining geographic coherence with the training regions. The total size of the *extended cropmark validation set* is 80 images for positive class and 200 images for negative class (maintaining 2:5 ratio). All models trained on *extended cropmark train set* were validated using *extended cropmark validation set*.

While the majority of the newly added images maintain a vertical or near-vertical perspective, a noticeable portion contains oblique views. Although the presence of tilted images introduces additional variability, it was considered acceptable, given that the primary goal was to expand the training dataset. Moreover, this variability may strengthen the model's ability to generalize, making them more robust to different imaging angles.

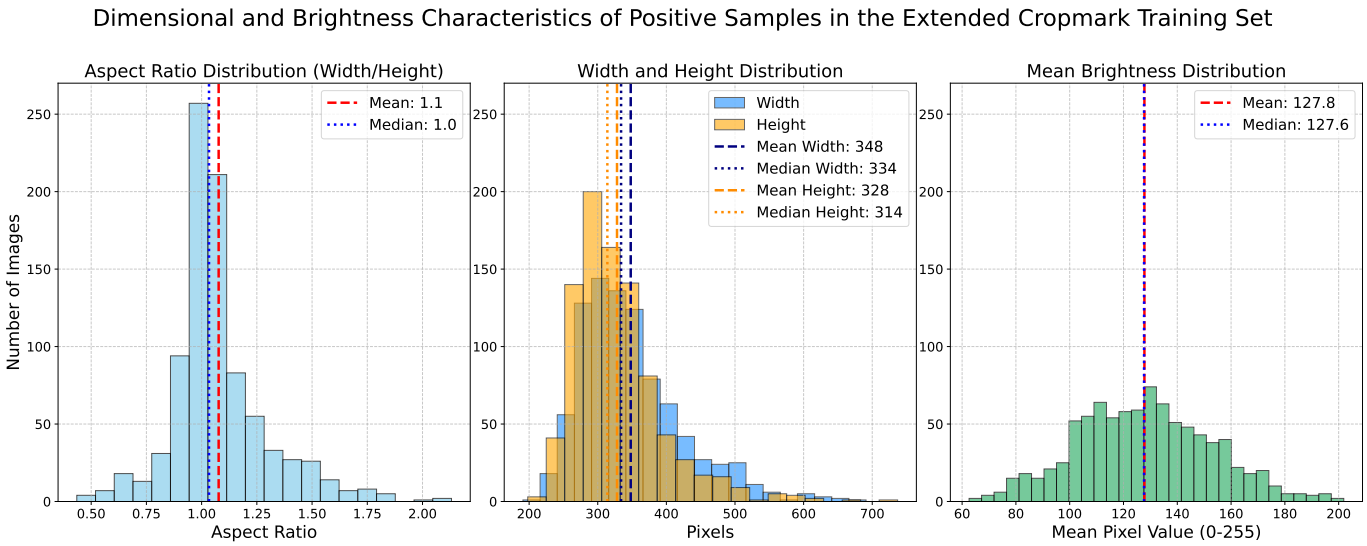

Figure 4.5: The figure summarizes the dimensional and brightness characteristics of the extended positive training samples in *extended cropmark train set*. Compared to the original dataset, the brightness distribution has become unimodal and closely approximates a normal distribution, with the mean value nearly identical to the median. The distribution of image dimensions appears sharper, with a more concentrated clustering around typical sizes. Apart from these changes, the fundamental structure and nature of the data distribution remain consistent with the *cropmark train set*.

4.6.4 Auxiliary Datasets

For the purposes of the remaining models in the final pipeline, dedicated train and test datasets were constructed to train the auxiliary neural models.

Aerial Dataset

The *aerial dataset* was created for training a future model within the pipeline. It consists of a *aerial test set*, *aerail train set* and *aerial validation set*.

Aerial test set is used later in this thesis to assess the performance of the binary classifier distinguishing between aerial and non-aerial images (see 5.2.1). It consists of 380 annotated images with a balanced 50:50 class distribution. The aerial (positive) samples were mostly derived from the *cropmark test set* originally assembled for cropmark classification. The non-aerial (negative) samples were selected mainly from the DomainNet dataset [66].

The *aerial train set* used for aerial versus non-aerial scene classification comprised approximately 40,000 labeled images, evenly balanced between the two classes. In addition, a separate validation set of around 4,000 images was used to monitor model performance during training.

Aerial images in *aerial train set* were sourced from Geoportal (ČÚZK), Mapbox [67], and MLRSNet [68], ensuring coverage of diverse regions and acquisition conditions. Non-aerial samples were drawn primarily from the COCO [69] and DomainNet [66] datasets, representing indoor, urban, and other irrelevant scenes. The *aerial validation set* included held-out examples from DomainNet and OpenImages [70] (for the non-aerial class), as well as previously unseen aerial scenes from the same sources as the training data.

Agricultural Dataset

A second set, called the *agricultural dataset* (divided into *agricultural test set*, *agricultural train set* and *agricultural validation set*), was constructed to support the binary classification task that distinguishes between aerial imagery relevant to agriculture and irrelevant.

Agricultural test set contains 190 images with a balanced 50:50 class distribution. The positive class comprises cropmark samples from the *cropmark test set*, representing typical agricultural landscapes. The negative class consists of aerial images from the Czech Republic depicting forests, urban areas, water bodies, and other land cover categories unsuitable for cropmark detection. The negative class was obtained through manual cropping from the Mapy.cz portal, ensuring the same visual style and acquisition conditions as the positive class, in order to avoid distributional bias. Later in the thesis (Section 5.2.2), this test set serves to evaluate the model's ability to filter out scenes where archaeological features are unlikely to occur, thus streamlining subsequent analysis.

Agricultural train set (approximately 8,000 labeled images and 50:50 class distribution) was constructed from multiple publicly available land cover sources. The positive class, representing agricultural land, was composed of samples extracted from the MultiScene [71], LandCover.ai [72], and MLRSNet [68] datasets. These were filtered to include only categories corresponding to cultivated fields and related terrain types. The negative class included non-agricultural scenes such as forests, urban areas, and water bodies, selected from RSSCN7 [73], UCMerced [74], and additional non-agricultural categories within MLRSNet.

Agricultural validation set (around 2,000 samples, 50:50 class distribution) was then constructed to mirror the training structure. Agricultural samples were obtained from Mapbox, RSSCN7, and UCMerced, while the non-agricultural class was validated using representative samples from PatternNet [75] and additional scenes from RSSCN7 and UCMerced.

4.7 Use of Lidar

Although LiDAR has demonstrated considerable potential in the identification of anthropogenic structures, the quality of publicly available LiDAR image datasets in the Czech Republic presents significant limitations for the task at hand. Although larger features, such as fortification lines or expansive earthworks, may still produce recognizable elevation signatures (as shown in Figure 4.7), these are the exception rather than the rule. The majority of archaeological remains detectable through remote sensing consist of small-scale structures, for which the current available LiDAR resolution is insufficient (as shown in Figure 4.6). These subtle terrain anomalies often vanish in the noise of coarse-resolution data, making their reliable detection challenging or even infeasible.

Combining LiDAR and RGB data could in theory improve detection performance by integrating structural and spectral information. However, this approach was not pursued due to practical limitations. High-resolution LiDAR data from ČÚZK (shown in Figures 4.7, 4.6) cannot be downloaded automatically, making large-scale alignment infeasible. Lower-resolution data from other sources are technically accessible, but their quality is insufficient for detecting small-scale archaeological features targeted in this study.

Given these constraints, LiDAR data were not employed in this thesis. Instead, the focus was placed on high-resolution aerial imagery, which offers more consistent and interpretable visual cues for the types of archaeological features targeted in this thesis.

Figure 4.6: Comparison of Vegetation Marks from 2010 and 2017 with the Digital Terrain Model (DTM) of the Czech Republic (2023) (Enhanced Contrast) for the Same Area (Černouček, 50.3571572N, 14.3009792E). In this case, it would be extremely difficult to distinguish the pattern using the DTM, most likely due to the small spatial size of the object and the insufficient resolution of the elevation data. Source: ČÚZK [2010, 2017, 2023].

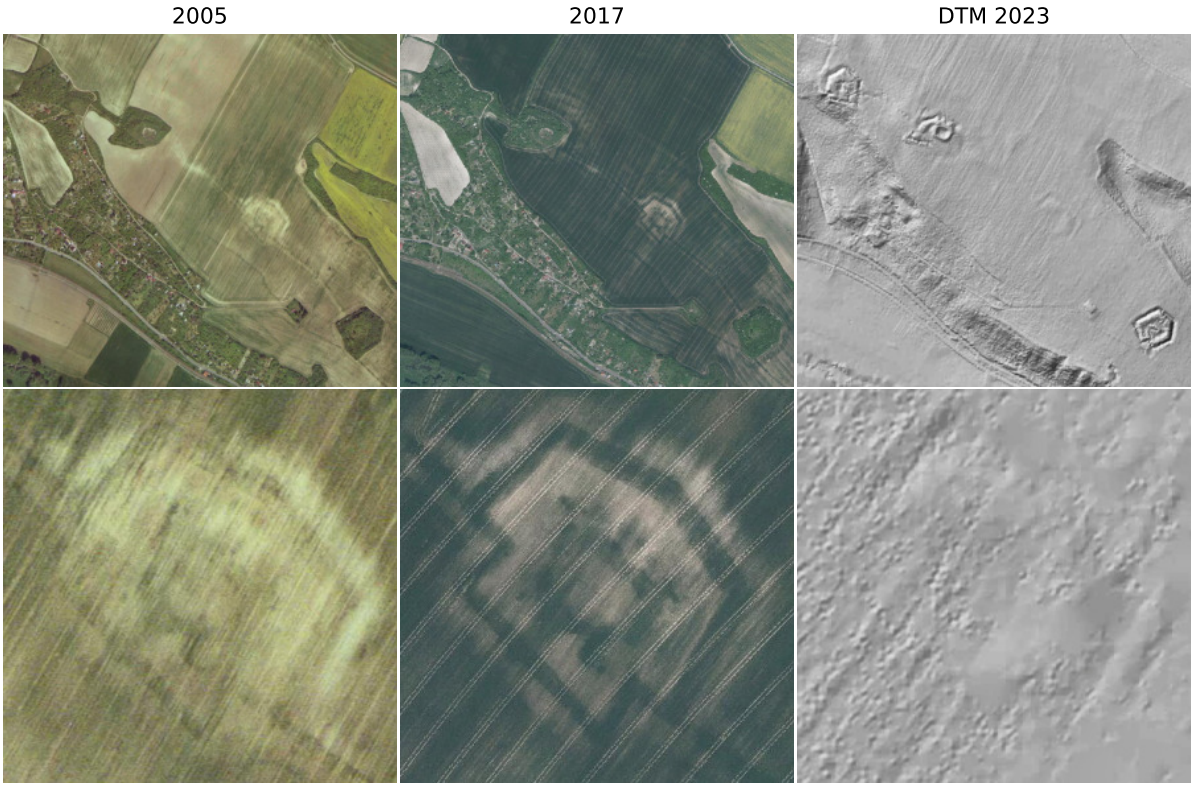

Figure 4.7: Comparison of Vegetation Marks from 2005 and 2017 with the Digital Terrain Model (DTM) of the Czech Republic (2023) (Enhanced Contrast) for the Same Area (Třeboutice, near Litoměřice, 50.5345525N, 14.1815553E). In this case, the remnants of the structure (fortified bridgehead of Terezín, see more detailed case study in [32]) can be analyzed not only through vegetation marks but also using DTM. Below: Zoomed-in view of the area of interest. Source: ČÚZK [2005, 2007, 2023].

5 Methods

This chapter describes the methodological foundation of the thesis. It introduces the evaluation metrics, training strategy and technology, initial augmentation, and synthetic data generation techniques used throughout the Experiments 6.

Given the nature of the dataset task, this thesis approaches the detection of archaeological cropmarks as a scene classification problem (described in Section 2.3.1).

5.1 Metrics

5.1.1 Evaluation Metrics

To evaluate the performance of binary classifiers developed in this thesis, several standard metrics were used. The primary focus is placed on three core metrics: **precision**, **recall**, and their harmonic mean, the **F1 score**. These metrics are particularly well suited for imbalanced classification tasks, where the number of positive examples is typically much smaller than the number of negatives.

Let the following quantities be defined:

- **True Positives (TP)**: the number of correctly predicted positive samples,
- **False Positives (FP)**: the number of negative samples incorrectly predicted as positive,
- **False Negatives (FN)**: the number of positive samples incorrectly predicted as negative,
- **True Negatives (TN)**: the number of correctly predicted negative samples.

Precision measures the proportion of true positive predictions among all samples predicted as positive. It reflects the model's ability to avoid false alarms:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5.1}$$

Recall (also known as sensitivity) quantifies the proportion of correctly identified positive samples out of all actual positives. It expresses the model's ability to detect relevant instances:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5.2}$$

F1 Score is the harmonic mean of precision and recall. It serves as a balanced metric that considers both false positives and false negatives.

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5.3}$$

In all classification experiments presented in this work, F1 score is used as the primary selection criterion for choosing the best-performing models, unless stated otherwise. Precision and recall are reported alongside to provide a more complete understanding of model behavior, particularly in relation to false positives and false negatives.

While **accuracy** is a commonly used metric in classification tasks, it is not well suited for the detection of rare phenomena such as archaeological cropmarks. Accuracy is defined as the ratio of correctly classified samples (both positives and negatives) to the total number of samples:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.4}$$

In highly imbalanced datasets, where negative samples dominate, a classifier may achieve high accuracy simply by predicting the majority class, while failing to detect most or all positive cases. As such, accuracy can be misleading in scenarios where detecting the minority class is of primary importance - as is the case in this thesis. For this reason, accuracy is reported only as a supplementary metric, while recall and F1 score remain the central focus.

Some of the evaluation metrics discussed above, particularly in the context of classification performance, have already been mentioned in Section 2.3.

5.1.2 Training Loss Functions

To train the classification models presented in this thesis, several loss functions were employed, each selected to reflect the nature of the task and the model architecture.

Binary Cross-Entropy Loss (BCE). For binary classification tasks, such as distinguishing between images with and without cropmarks, the most commonly used loss function is Binary Cross-Entropy (BCE) loss. It is defined as:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \tag{5.5}$$

where $y_i \in \{0, 1\}$ is the ground truth label, and $\hat{y}_i \in (0, 1)$ is the predicted probability for the positive class. The loss is averaged over a batch of N samples, where N denotes the number of examples in the current mini-batch. BCE loss penalizes incorrect predictions more strongly the more confident they are, encouraging the model to make accurate and calibrated probability estimates.

Diffusion Model Loss (Latent Prediction). In the case of generative models - specifically Stable Diffusion fine-tuned with LoRA - the training loss is fundamentally different. Instead of comparing predicted labels with ground truth, the model learns to denoise latent representations corrupted by Gaussian noise. The objective is to predict the noise vector ϵ added to the latent variable z, given the noised latent z_t and the timestep t. The corresponding loss is typically a simple Mean Squared Error (MSE) between the predicted and actual noise vectors:

$$\mathcal{L}_{\text{diff}} = E_{z,\epsilon,t}\left[\|\epsilon - \hat{\epsilon}_\theta(z_t, t, c)\|^2\right] \tag{5.6}$$

where $\hat{\epsilon}_\theta$ is the noise predicted by the model, and c is the conditioning input (in this case, a text prompt). This loss encourages the model to learn the conditional distribution of plausible images matching the prompt, and serves as the foundation for generative pre-training in Stable Diffusion models.

This denoising-based objective is used during LoRA fine-tuning as well, where the model is adapted to domain-specific features (e.g., cropmarks) using a relatively small number of training images and carefully engineered prompts.

5.2 Auxiliary Tasks

In addition to the main classification task, two auxiliary tasks were designed to support data preprocessing. Resulting models also play a role in the potential final classification pipeline.

5.2.1 Binary Classification of Aerial vs. Non-Aerial Imagery

For the task of distinguishing aerial photographs from non-aerial images, a convolutional neural network (CNN) based on the ResNet18 architecture [76] was used - a well-established model known for its simplicity and efficiency.

The model was initialized with pre-trained weights from ImageNet [77], which provides a general visual representation suitable for transfer learning. The final fully-connected

layer was replaced with a new dense layer with two outputs, corresponding to the binary classification task.

This architecture was selected as it provides sufficient model capacity for capturing domain-relevant image features, while allowing for efficient training and hyperparameter experimentation. For our purposes, ResNet18 offers a balanced trade-off between complexity and training speed, facilitating flexible exploration of different learning configurations.

The following augmentations were used to improve generalization:

- RandomResizedCrop(224) – simulates varying scale and framing.
- RandomHorizontalFlip – handles left-right symmetry.
- RandomRotation(15) – adds viewpoint variation.
- ColorJitter – introduces brightness and contrast noise.
- GaussianBlur – mimics blur and low quality.
- JPEG Compression – simulates compression artifacts.
- ImageNet normalization.

This model is intended to serve as a safeguard in the final detection pipeline, where it can filter out irrelevant out-of-domain images that may occur due to retrieval errors or inconsistencies in third-party datasets.

5.2.2 Binary Classification of Agricultural vs. Non-Agricultural Landscape

The second classification task focused on distinguishing between agriculturally relevant areas (such as fields, grasslands, and meadows) and non-agricultural land covers (including forests, urban settlements, water bodies, and others). As in the previous task, we employed a ResNet18 convolutional neural network with ImageNet [77] pre-trained weights, modified for binary output:

These extensive image augmentations were applied to the training set as the basic setup:

- Resize to 256×256 pixels.
- Random crop to 224×224.

- Random horizontal flips.
- Random rotations up to $\pm 15°$.
- Brightness and contrast jittering.
- Random Gaussian blur.
- Simulated JPEG compression artifacts.
- ImageNet normalization.

5.3 Main Task - Cropmark Identification

5.3.1 Basic Classification

Following the preliminary experiments, a baseline classification model was trained to distinguish aerial orthophoto tiles containing cropmarks from those without. Two lightweight convolutional neural network architectures were evaluated: ResNet18 [76] and MobileNetV2 [78].

Weights and Parameters

Both models (ResNet18, MobileNetV2) were pre-trained on ImageNet and modified for binary classification by replacing the final fully-connected layer with a single output neuron. Binary Cross-Entropy with Logits Loss [8] was used as the loss function, and training was performed using both the SGD optimizer [79] and the Adam [80] or AdamW [81] optimizer.

Extensive data augmentation was applied to the training images as a basic setup: [14]. The applied augmentations included:

- Random rotations of ± 90 degrees.
- Horizontal and vertical flips.
- Random zooms up to 30%.
- Maximal-square center crop.
- Random crops to a fixed size of 224×224 pixels.
- Brightness and contrast jittering.

- Hue and saturation perturbations.
- ImageNet normalization.

5.3.2 Synthetic Data Generation for Dataset Augmentation

Although the dataset was extended, it became evident during the course of this work that the limited number of available positive examples could pose a significant limitation for training robust models. To mitigate this constraint and improve the generalizability of the networks, the generation of synthetic training data was subsequently explored.

Synthetic data generation has emerged as an increasingly valuable technique in domains where real annotated samples are scarce or costly to obtain. In the context of archaeological remote sensing, generating plausible cropmark patterns offers the potential to expand the training dataset manifold while preserving the critical characteristics required for model learning.

Two approaches to synthetic data generation were considered.

Diffusion Models

Recent advances in diffusion-based image synthesis, notably Stable Diffusion[82], provide an alternative pathway. Fine-tuning diffusion models using a small set of domain-specific cropmark images could allow text-to-image generation of realistic scenes. Moreover, the ControlNet framework[83] permits conditioning the generation process on structural inputs such as edge maps or semantic masks, thus offering fine-grained control over the appearance and placement of synthetic cropmarks.

Recent advances in diffusion-based image synthesis, particularly the SDXL 1.0 model [84], provide a powerful foundation for generating realistic domain-specific imagery. However, full fine-tuning of such large models could be computationally demanding and often impractical for limited datasets. To address this, this thesis employs Low-Rank Adaptation (LoRA) [85], a parameter-efficient fine-tuning method that introduces low-rank updates into selected layers of the pre-trained network. LoRA enables targeted domain adaptation - in this case, learning the visual appearance of archaeological cropmarks - using a small set of annotated samples without modifying the entire base model.

Classical Augmentation and Procedural Simulation

Beyond deep generative models, simpler methods such as procedural overlay of extracted cropmark textures onto real aerial backgrounds were considered. Techniques including

random placement, rotation, brightness variation, and blending could simulate the appearance of cropmarks across a wide range of environmental conditions. Additionally, synthetic masks (watermarks) emulating typical archaeological shapes (circular enclosures, linear ditches, etc.) could be algorithmically generated and inserted into aerial images.

6 Experiments

This chapter presents a series of experiments that aim to evaluate the effectiveness of various training strategies, model architectures, and synthetic data sources for the task of cropmark classification in aerial imagery.

The chapter also introduces the overall workflow and adapts its structure in response to findings obtained throughout the experimental process.

6.1 Auxiliary tasks

6.1.1 Binary Classification: Aerial vs. Non-Aerial

The first experiment was focused on binary classification between aerial and non-aerial imagery. The final model is intended to be used in the final pipeline as a safeguard. For more information see section 5.2.1.

During training (*aerial train set*), a comprehensive hyperparameter sweep was performed. The best-performing model achieved the highest validation (*aerial validation set*) accuracy in epoch 20 with the following configuration.

- **Optimizer:** SGD with momentum 0.9.
- **Learning Rate:** 0.00067.
- **Batch Size:** 16.
- **Weight Decay:** 8.47×10^{-6}.

The resulting classifier demonstrated robust separation between the two domains and provided a reliable filtering mechanism for downstream tasks.

Results on Aerial Test Set

The binary classifier achieved an overall accuracy of 95.00% on the *aerial test set* (Section 4.6.4). The confusion matrix indicates a high level of performance, with 183 out of 190 aerial images correctly identified (recall = 0.96) and a precision of 0.94 for the aerial class. The F1 score for aerial images was 0.95.

To better understand the limitations of the model, Figures 6.2 and 6.1 present samples of incorrectly classified images. Figure 6.2 shows non-aerial images that were mistakenly classified as aerial, potentially introducing noise into downstream tasks.

Figure 6.1: This figure displays all images from the *aerial test set* that were incorrectly classified as non-aerial by the binary classifier (based on the ResNet18 architecture). Notably, none of the misclassified samples depict agricultural fields, indicating that (despite the suboptimal performance) the model is unlikely to eliminate data of archaeological interest in the downstream detection pipeline.

Figure 6.2: This figure presents a selection of images that were incorrectly classified as aerial by the chosen model (based on the ResNet18 architecture). It can be observed that the misclassified samples predominantly feature natural patterns and textures. Although such misclassifications are undesirable, the pipeline is not expected to encounter similar input data, and thus these errors are unlikely to impact practical performance.

6.1.2 Binary Classification: Agricultural vs. Non-Agricultural

The agricultural classifier (see 5.2.2) was trained (*agricultural train set*) using the same architecture and training strategy as the previous model.

To identify optimal training settings, a hyperparameter sweep was performed. The configuration that yielded the best validation performance (*agricultural validation set*) in epoch 23 was as follows:

- **Optimizer:** SGD with momentum 0.9.
- **Learning Rate:** 0.00043.
- **Batch Size:** 32.
- **Weight Decay:** 1.02×10^{-5}.

The best models were saved based on validation accuracy and validation loss for future integration into the full archaeological site detection pipeline. Their primary

role is to act as a secondary filter - rejecting inputs that, while aerial in origin, depict irrelevant landscapes for the domain of archaeological interest.

Results on Agricultural Test Set

The classifier was evaluated on a dedicated *agricultural test set* (Section 4.6.4). The model achieved an overall accuracy of 94.74%, with a recall of 0.99 and an F1 score of 0.95 for the positive class (field). The precision for this class reached 0.91, indicating a few non-field images were incorrectly accepted.

Figures 6.4 and 6.3 present all classification errors. These include field images incorrectly rejected and non-field images incorrectly accepted, providing insight into the model's decision boundaries and remaining ambiguities in visually similar land categories.

Figure 6.3: This figure presents a selection of images that were incorrectly classified as agricultural by the chosen model (based on the ResNet18 architecture). It can be observed that the model has difficulty correctly classifying water bodies and cleared forest areas.

Figure 6.4: This figure displays a single image of cropmark from *cropmark test set* that was incorrectly classified as non-agricultural by the binary classifier (based on the ResNet18 architecture).

6.2 Main Task - Cropmark Identification

6.2.1 Basic Classification

A series of experiments was conducted to establish a baseline classification model for cropmark identification, following the methodology described in 5.3.1. A hyperparameter sweep was performed. The best configuration on the validation set employed:

- ResNet18 model architecture.
- Batch size of 32.
- Initial learning rate of 8.8×10^{-4}.
- Step-based learning rate scheduler.
- Weight decay of 3.9×10^{-5}.
- AdamW optimizer.

Training (*cropmark train set*) was carried out for 30 epochs using the AdamW optimizer. Data random augmentations are listed in Section 5.3.1.

The model achieved a validation F1 score of 0.85 and a recall of 0.85. These results are particularly notable given the small size, noise, and morphological variability of the dataset. Data augmentation strategies, especially random flipping, brightness and contrast adjustment, and hue-saturation perturbations, were found to significantly improve generalization.

The evolution of validation metrics (*cropmark validation set*) during the training of the ResNet18 model for basic cropmark classification is illustrated in Figure 6.5. Notably, around epoch 5, the model achieved perfect precision on the validation set, at the expense of a sharp drop in recall due to an increase in false negatives. After approximately 10–15 epochs, the trend of validation metrics stabilizes and begins to oscillate without a clear upward or downward trajectory. The model checkpoint after 17 epochs was selected for final evaluation, because it reached its highest validation F1 score. As shown in Figure 6.6, the lowest validation loss throughout entire 30-epoch training also occurred around this point, supporting the selection.

The training and validation loss curves are presented in Figure 6.6. A noticeable peak in the validation loss is observed at epoch 5, corresponding to the sharp deviation seen in the validation metrics. This moment reflects a sudden increase in false negatives, as previously discussed. However, after this fluctuation, the model quickly returned to its prior trend, and the overall training dynamics remained stable throughout the subsequent epochs.

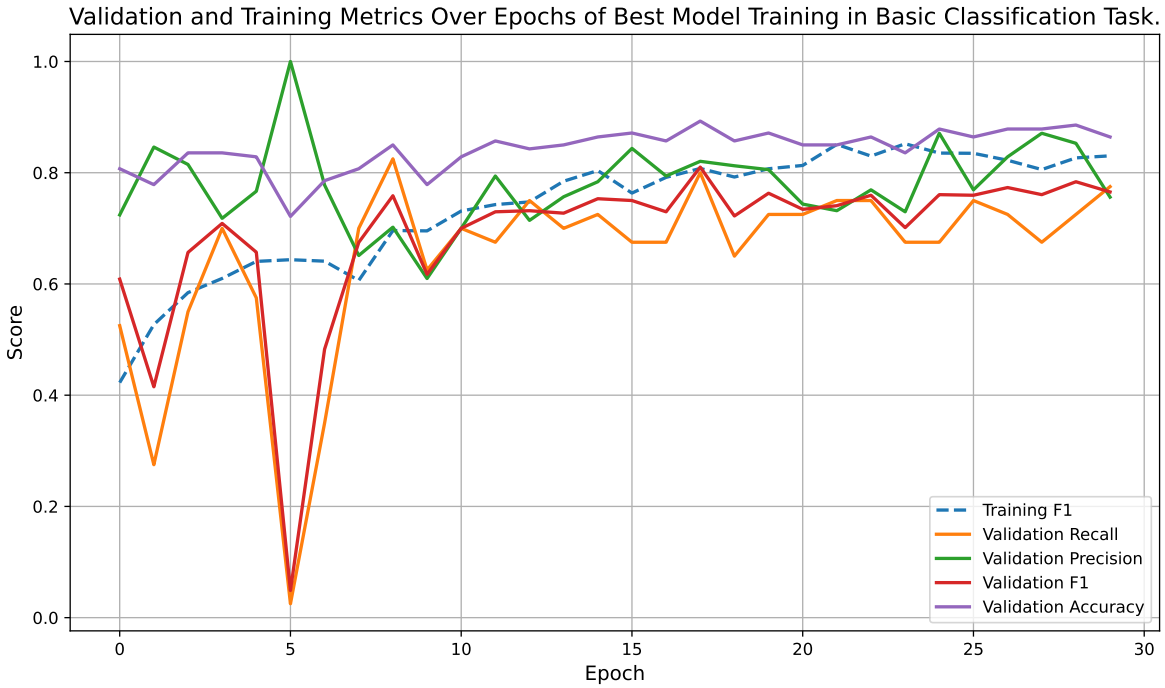

Figure 6.5: Validation metrics and Training set F1 score during training of the model (ResNet18) with hyperparameters from the best sweep in the basic classification task. Notably, at epoch 5, the model achieved perfect precision (no false positives) on the validation set, at the cost of a sharp increase in false negatives and a corresponding decrease in recall and F1 score consequently.

Results on Cropmark Test Set

The model achieving the best validation F1 score was selected after 17 epochs and evaluated on the *cropmark test set*. The classification report and confusion matrix are shown in Tables 6.1 and 6.2.

Table 6.1: Classification report on the *cropmark test set* (ResNet18, cropmark training dataset).

Class	Precision	Recall	F1-score	Support
Negative (0)	0.8773	0.8512	0.8640	336
Positive (1)	0.5238	0.5789	0.5500	95
Accuracy		0.7912		
Macro Avg	0.7006	0.7151	0.7070	431
Weighted Avg	0.7994	0.7912	0.7948	431

The model achieved an overall accuracy of 79.1% on the *cropmark test set*. It demonstrated relatively strong performance for the negative class (non-cropmark areas) with a precision of 0.88 and recall of 0.85. For the positive cropmark class, the model achieved a precision of 0.52 and recall of 0.57, resulting in an F1-score of 0.55.

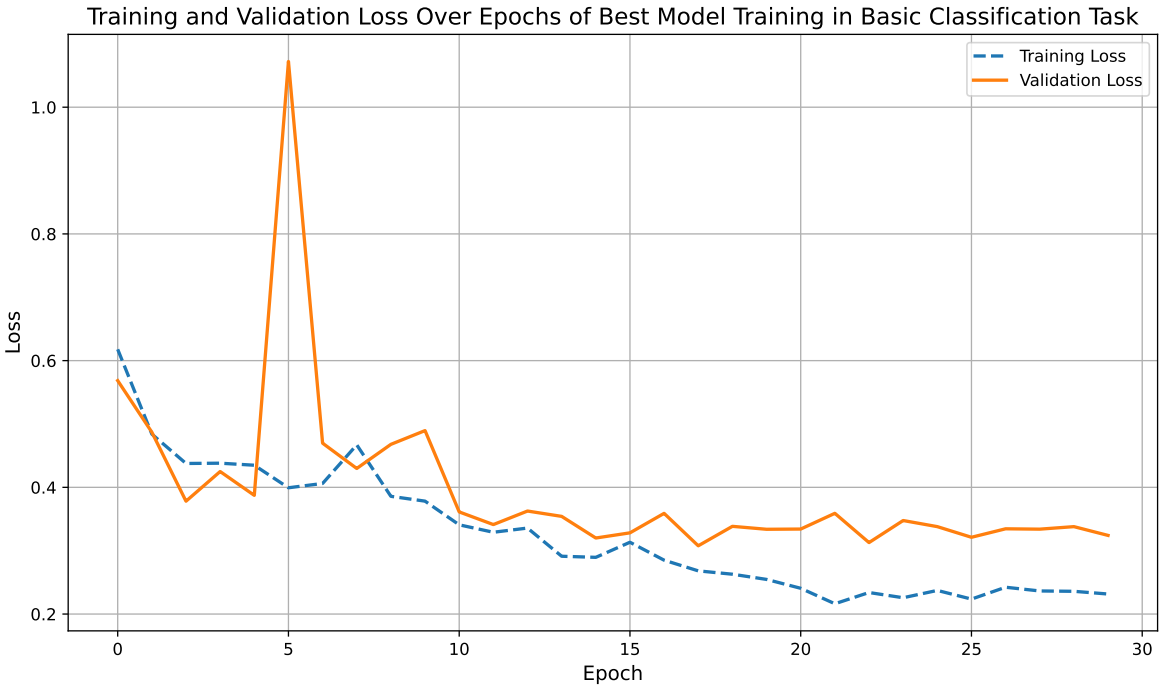

Figure 6.6: Validation and Training loss during training of the model (ResNet18) with hyperparameters from the best sweep in the basic classification task. The model reached the minimum validation loss at epoch 17. A noticeable deviation occurs at epoch 5, correlating with a sharp decline in several other monitored metrics in figure 6.5.

Table 6.2: Confusion matrix for *cropmark test set* predictions (ResNet18, cropmark training dataset).

	Predicted Negative	Predicted Positive
True Negative	286	50
True Positive	40	55

Examples of classification errors are illustrated in Figure 6.7. A rich representative sample of false positives and false negatives with detailed captions is provided in Appendix B, Figures B.3 and B.2. The false positives indicate that the model sometimes confuses fairy rings, unusual soil or vegetation patterns, reclamation systems, or water flow artifacts across fields with archaeological cropmarks. In contrast, the false negatives show that the model struggles to detect faint, blurry structures and cropmarks captured at a larger spatial scale, such as fortification boundaries (e.g., Třeboutice near Litoměřice and Chleby fortifications). This limitation is likely due to the lower representation of large-scale features in the training dataset, where smaller and more localized objects predominate. The model also encounters difficulties in identifying cropmarks with a brownish background, which are less visually distinct.

An additional factor potentially contributing to the suboptimal performance was

identified in the data augmentation process. During training, square central crops were applied, resulting in the consistent use of only the central portion of each image. This disproportionately affected rectangular images, leading to the loss of peripheral contextual information. As shown in Figure 4.4, a significant portion of the training set consists of non-square images. Consequently, the model may not have fully exploited the available spatial variability. To address this issue, subsequent model versions incorporated custom augmentations specifically designed to preserve and utilize the full spatial extent of the training samples.

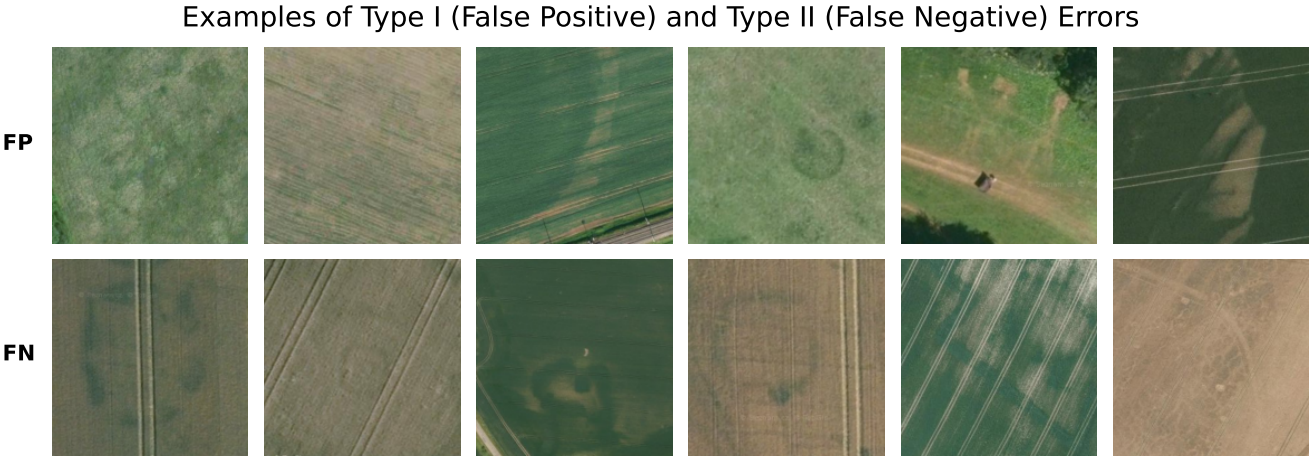

Figure 6.7: Examples of classification errors made by the model (ResNet18, *cropmark train set*) on the *cropmark test set* in binary classification task. View figures B.3 and B.2 for more advanced analysis.

Correction of Augmentation Strategy and New Results

To address the identified limitations in spatial cropping and over-reliance on central regions of the input images, the augmentation strategy was revised. A rotation of up to ± 15 degrees was introduced with a probability of 0.5, using the *border reflect* mode to avoid the introduction of empty (black) regions. All other hyperparameters, including learning rate, scheduler, batch size, and optimizer, were kept unchanged.

Multiple models were trained under this new setup. Their performance was evaluated based on validation recall, F1 score, and training F1 score, as illustrated in Figure 6.8. The model checkpoint from epoch 21 was selected for evaluation, as it yielded the highest validation F1 score with only a minimal decrease in recall.

The classification results for this final model are presented in Table 6.3, and the corresponding confusion matrix is shown in Table 6.4.

Table 6.3: Classification report on the *cropmark test set* (ResNet18, cropmark training dataset), corrected.

Class	Precision	Recall	F1-score	Support
Negative (0)	0.9000	0.8571	0.8780	336
Positive (1)	0.5676	0.6632	0.6117	95
Accuracy		0.8144		
Macro Avg	0.7338	0.7602	0.7448	431
Weighted Avg	0.8267	0.8144	0.8193	431

Table 6.4: Confusion matrix for *cropmark test set* predictions (ResNet18).

	Predicted Negative	Predicted Positive
True Negative	288	48
True Positive	32	63

Compared to the original baseline, the corrected model demonstrated consistent improvement across all key evaluation metrics. Precision for the positive class increased from 0.52 to 0.57, while recall rose from 0.58 to 0.66, leading to an F1-score improvement from 0.55 to 0.61. The overall accuracy also increased from 79.1% to 81.4%. The confusion matrix confirms a reduction in both false positives and false negatives. These results suggest that the revised augmentation strategy could successfully mitigated the shortcomings of the initial model.

Although performance on the positive class remains a challenge, the results demonstrate meaningful progress. The noticeable difference in models' performance highlights the significant impact of data transformations during training. This underlines not only the importance of careful preprocessing but also motivates further experimentation with augmentation strategies introduced in next section.

Augmentation Sweep

Following the baseline experiments, it became evident that although the model learned and achieved reasonable accuracy and recall, its performance on cropmarks in certain domains was significantly suboptimal. To address this limitation, a targeted sweep focused on data augmentation probabilities was performed using the *cropmark train set*.

During this sweep, the core hyperparameters such as the optimizer (AdamW), learning rate (8.8×10^{-4}), weigth decay (3.9×10^{-5}), batch size (32), scheduler (step-based)

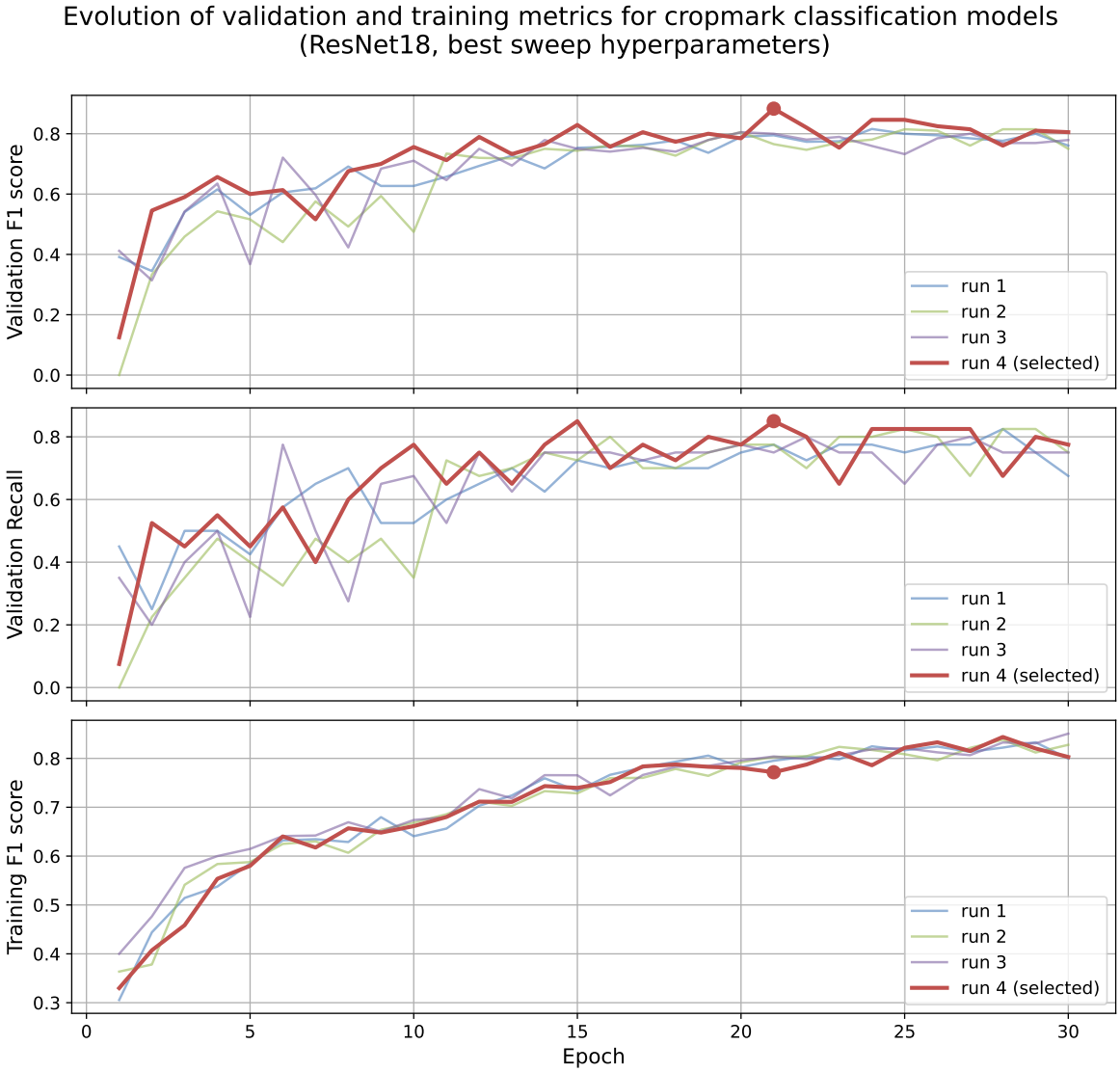

Figure 6.8: Evolution of validation and training metrics during training of four models with identical hyperparameters. Validation F1 score, validation recall, and training F1 score across 30 training epochs are displayed. A red marker indicates the epoch at which model achieved best performance. All models exhibit a steady increase in training F1 score, and their final performance levels are similar, suggesting stable learning behavior across runs.

and model architecture (ResNet18) were kept constant. The sweep systematically varied the probabilities of applying specific augmentations, including random rotation, flipping, scaling, brightness-contrast adjustment, and hue-saturation shifts. An enhancement of grid distortion was newly incorporated into the sweep.

Based on the results of the augmentation sweep (particularly the correlations ob-

served between individual augmentations and validation recall and the F1 score) and the analysis of misclassifications in the test set, as well as the realization of suboptimal augmentation design and underutilization of available data, a new augmentation strategy was designed for subsequent training on the *extended cropmark train set*. The applied augmentations are illustrated in Figure 6.9 and are considered as standard in the remaining parts of the thesis if not stated otherwise.

In addition to standard augmentations, two custom transformations were introduced. *MaxCenterCropResize* crops each image to the largest centered square and resizes it to 256×256 pixels (applied to 50% images), while *ConditionalResize* adjusts images with a shorter side below 224 pixels to a sufficient size while maintaining aspect ratio. Both augmentations were specifically implemented for purposes of this thesis to better handle the variability of the extended dataset.

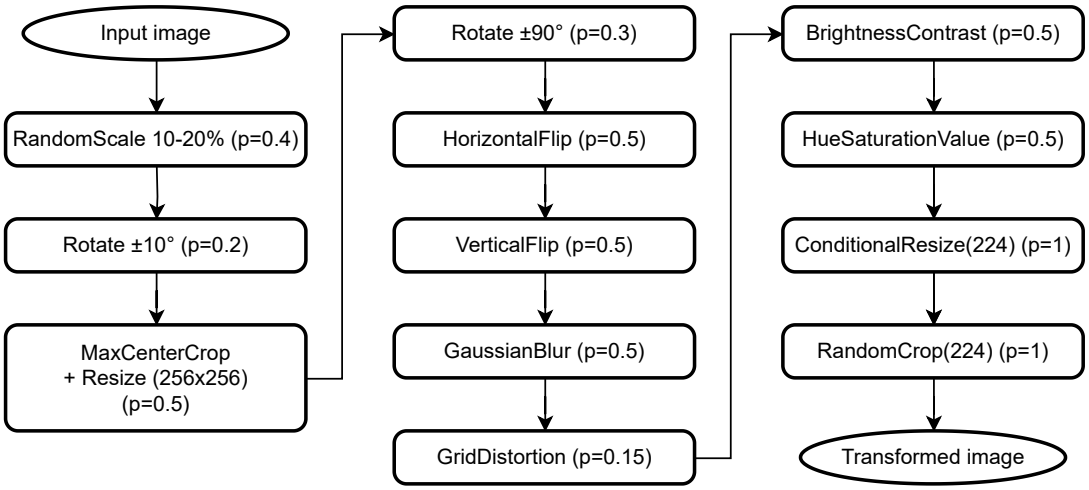

Figure 6.9: Diagram illustrating the sequence of augmentations applied to input images during model (ResNet18) training on cropmark extended training dataset. These transformations were selected and parameterized based on the results of the augmentation sweep and the observed deficiencies during training. The scheme shows individual transformation steps, including geometric operations and colour adjustments. Conditional resize augmentation applies only on the images with insufficient dimensions to a following random crop. The probabilities associated with individual augmentations have changed during future experiments.

Hyperparameter sweep (Extended Cropmark Train Set)

The next logical step toward improving performance is the incorporation of the *extended cropmark training set*.

A fresh hyperparameter sweep was performed, and the best configuration was selected, featuring a learning rate of 4.904×10^{-5}, a cosine-based scheduler, a batch size of

32, a training duration of 50 epochs, and a weight decay of 3.734×10^{-5}.

Results (Extended Cropmark Train Set)

The model achieving the best validation F1 score during training on the *extended cropmark validation set* was evaluated on the *cropmark test set*. The corresponding classification report is presented in Table 6.5, and the associated confusion matrix is shown in Table 6.6.

Compared to the baseline model trained on the original dataset, this configuration achieved a noticeable improvement in recall for the positive class, increasing the proportion of correctly identified cropmark images. However, this gain was accompanied by a significant decrease in precision, reflecting a higher number of false positives.

Table 6.5: Classification report on the *cropmark test set* for the ResNet18 model trained on the *extended cropmark train set* (best validation F1 score).

Class	Precision	Recall	F1-score	Support
Negative (0)	0.9011	0.7321	0.8079	336
Positive (1)	0.4304	0.7158	0.5375	95
Accuracy		0.7285		
Macro Avg	0.6657	0.7240	0.6727	431
Weighted Avg	0.7973	0.7285	0.7483	431

Table 6.6: Confusion matrix for predictions on the *cropmark test set* using the ResNet18 model trained on the *extended cropmark train set*.

	Predicted Negative	Predicted Positive
True Negative	246	90
True Positive	27	68

This result suggests that the model, now exposed to a wider variety of cropmark types and shapes during training, has improved its ability to detect true positives. The increased variability in the training data likely enabled the model to generalize better and recognize cropmark features that were underrepresented in the original dataset. At the same time, however, this diversity may have made the model more permissive, leading it to incorrectly classify visually similar structures as cropmarks. This effect is particularly relevant given the deliberately challenging composition of the test set, which includes many non-archaeological patterns specifically chosen to resemble archaeological cropmarks.

Another possible factor is the composition of the negative training data. In the extended dataset, a substantial portion of negative samples was obtained automatically from the Geoportal.cz, whereas the test set negatives primarily originate from the manually curated Mapy.cz imagery. This domain mismatch may have weakened the model's ability to distinguish genuine cropmarks from misleading patterns within the test set.

Additional Fine-tune

The performance of the model trained on the *extended cropmark train set* proved to be less satisfactory than initially expected. One possible explanation is that the model overadapted to the newly introduced data, which - although valuable - differ from the original test set in terms of visual and contextual characteristics. Nevertheless, it is reasonable to expect that a larger and more diverse dataset should, in principle, support improved generalization. For this reason, the best-performing model from the extended training was subsequently fine-tuned on the original cropmark dataset to restore alignment with the characteristics of the evaluation domain. The learning rate of 1×10^{-5} was chosen.

The following results correspond to the ResNet18 model that was first trained on the *extended cropmark train set* and subsequently fine-tuned on the original *cropmark train set*. The model selected for evaluation was the one that achieved the highest validation F1 score during fine-tuning. Evaluation was performed on the *cropmark test set*. Table 6.7 summarizes the class-wise metrics. The corresponding confusion matrix is provided in Table 6.8.

Table 6.7: Classification report on the *cropmark test set* for the ResNet18 model trained on the *extended cropmark train set* (best validation recall) and subsequently fine-tuned on the original *cropmark train set* (best validation F1 score).

Class	Precision	Recall	F1-score	Support
Negative (0)	0.8959	0.7173	0.7967	336
Positive (1)	0.4136	0.7053	0.5214	95
Accuracy		0.7146		
Macro Avg	0.6547	0.7113	0.6590	431
Weighted Avg	0.7896	0.7146	0.7360	431

This experiment, however, did not yield the desired results. Both evaluation metrics - *precision* and *recall* - decreased compared to the previous model. This outcome may be due to random variation in model selection, but it could also be influenced by the overfitting of feature extractors caused by a higher number of training epochs.

Table 6.8: Confusion matrix for predictions on the *cropmark test set* using the ResNet18 model trained on the *extended cropmark train set* and fine-tuned on the original *cropmark train set*.

	Predicted Negative	Predicted Positive
True Negative	241	95
True Positive	28	67

6.3 Synthetic Data Generation - procedular

Despite the new augmentation technique and extension of the dataset using additional real-world imagery, the resulting evaluation metrics remained unsatisfactory. The performance of the classifier varied during process, but did not improve significantly (with the first augmentation fix as an exception). This could be attributed to domain mismatch, suboptimal annotation quality, or insufficient quantity or variability of the data.

6.3.1 Procedural Synthetic Data Generation

As a response, the decision was made to generate synthetic training data. A procedurally heuristic approach was adopted. Artificial cropmark masks were generated and blended with real aerial imagery of agricultural land. This method enabled the controlled simulation of cropmark-like features while maintaining photorealistic texture and plausible geographic context.

Data acquisition

The underlying aerial images were obtained by automatically downloading archival orthophotos from the ČÚZK geoportal (same way as desrcibed in Section 4.6). Since many of the available images were found to be unsuitable, due to image non-existence, or the wrong land cover category, a multi-stage filtering process was designed to ensure only high-quality agricultural scenes were retained. This process is illustrated in Figure 6.10. Previously trained Aerial and Agricultural classifiers were used (Sections 5.2.1 and 5.2.2).

All returned images were resized or cropped to a fixed resolution of 350×350 pixels. This size corresponds to output format of the synthetic mask generators, which always produce centered cropmark structures. Since the neural network models used later in this work operate on 224×224 inputs, applying random cropping to the larger 350×350 images ensures that most of the resulting training samples will include at least a portion of the mask in positive examples, as it is shown on the Figure 6.11. At the same time,

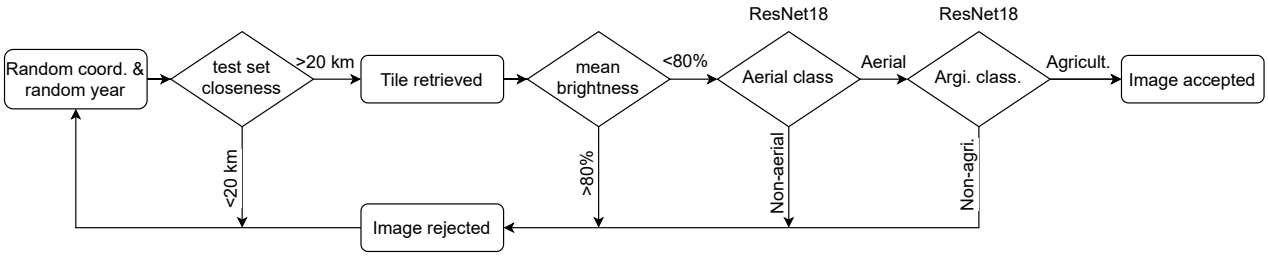

Figure 6.10: Overview of the filtering pipeline for selecting suitable archival orthophotos. The process includes random coordinate sampling, orthophoto retrieval, image brightness checks, aerial binary classification (ResNet18) and agricultural binary classification (ResNet18).

this approach introduces translational variation, effectively simulating positional shifts of the target feature while maintaining its central placement in the original image.

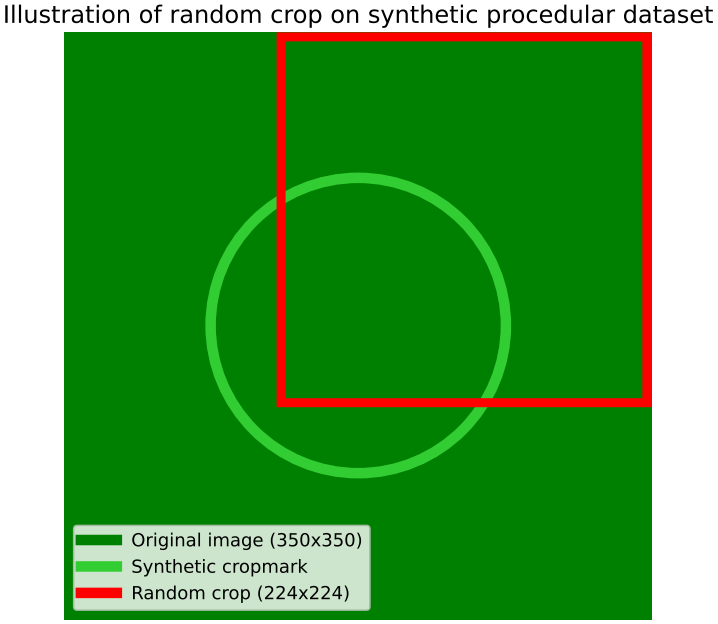

Figure 6.11: Diagram of an extreme case of Random Crop from an Image with a Centered Synthetic Mask. This cropping strategy simulates partial visibility of the target structure while ensuring that positive samples retain a characteristic portion of the mask essential for learning.

It is acknowledged that a small number of naturally occurring cropmarks may still be present in the downloaded imagery, despite no explicit effort being made to remove them. However, given the low prevalence of visible cropmarks across the general orthophoto archive, the likelihood of such false positives is considered minimal. Furthermore, the impact of any such samples is expected to be negligible compared to the synthetic overlays introduced in subsequent steps. Notably, the dataset constructed through this process is not meant to be used for final model training, but rather for the purpose of pre-training, where the focus lies in learning generalized low-level shape features and

spatial patterns.

Definition of Morphological Types

To determine which cropmark shapes should be synthetically generated, the entire training dataset was manually reviewed (described in Section 4.4.2). Several recurring morphological types were identified, with some further subdivided into subtypes to capture their internal variability. This typology, described in more detail in the following sections, was informed both by visual inspection and the reference table A.1.

For each type, a set of specific generating functions was implemented. These functions procedurally generate grayscale masks representing cropmark structures, which are later composited into aerial backgrounds. In addition, each main type was assigned a weight determining its relative frequency in the resulting dataset. Within each type, subtypes (i.e., generation functions) were also assigned individual probabilities. Table 6.9 summarizes the full configuration used for synthetic data generation.

Table 6.9: Summary of synthetic cropmark mask types and their generation probabilities.

Type	**Weight (Relative Frequency)**
`circular`	0.20
`rectangular`	0.25
`border`	0.20
`grid`	0.17
`patch (filled)`	0.18

Each of the defined morphological types corresponds to visual patterns commonly found in aerial images of archaeological sites. Examples of the generated mask types can be found in Figure B.4, which illustrates their visual diversity and resemblance to real cropmarks.

Shared Principles of Synthetic Cropmark Generation

All procedural functions used to generate synthetic cropmark masks follow a set of shared design principles, aimed at maximizing visual realism and morphological variability. The shared elements are summatized in Table 6.10. These principles ensure that generated structures resemble archaeological cropmarks not only in shape, but also in their irregularities, spatial distribution, and interaction with the image background. The most important shared aspects are summarized below.

Scale Adaptivity. Each generating function accepts the parameter `m_px` (meters per pixel, input image atribute), which controls scale-dependent properties such as object size, edge deformation, blur strength, and noise magnitude. In approximately 70% of cases, a randomized scale factor based on `m_px` is applied to introduce variability and to simulate cropmarks under different spatial resolutions.

Deformation, Waviness, and Imperfection Modeling. Edges of geometric primitives (circles, ellipses, rectangles, etc.) are perturbed using sinusoidal modulations or randomized displacements. This simulates natural imperfections and imprecise imaging conditions.

Blur and Fade Effects. To blend synthetic structures with background textures and to mimic natural softening, all generated shapes undergo Gaussian blurring. The kernel size and number of blur repetitions are again influenced by `m_px` and random factors. In many cases, cosine-based fade functions are used to gradually reduce the intensity of strokes toward their edges. Since each mask is normalized to values between 0 and 1 during the preprocessing phase, the blur effect not only introduces blurring but also influences the thickness of lines and the size of point-like structures. Therefore, the blur effect is inherently dependent on the scale of the image.

Controlled Noise Injection. All generators apply spatially variable brightness or intensity noise, often using a bicubic-resampled Gaussian map, constrained to the foreground mask of the shape. This simulates vegetation variability and environmental inconsistencies.

Rotation and Spatial Offsets. Most generators include randomized rotation within a defined angular range (typically 90 or 180 interval), and many allow for slight translation, misalignment, or variable placement of repeated structures. These operations increase dataset diversity and reduce geometric bias.

Table 6.10: Overview of key shared elements across synthetic cropmark generators.

Aspect	**Description**
`m_px` scaling	Adjusts size, deformation, blur, and noise strength based on spatial resolution.
Deformation and waviness	Adds irregularities to edges and contours for realism.
Gaussian blur	Integrates shapes into background, softens and widens edges.
Cosine fade	Reduces intensity locally.
Masked noise injection	Introduces brightness variation only within cropmark region.
Rotation and offset	Increases variety and reduces directional bias.

These shared design choices serve as the foundation upon which individual mask types build additional constraints and characteristics. Specific implementations for each morphological group are described in the following sections.

Mask types

The following section provides an overview of the individual mask types used in procedural generation, along with the specific strategies applied to each. These types reflect morphological patterns commonly observed in aerial imagery of archaeological sites, such as enclosures, ditches, and posthole arrangements. Examples of procedurally generated masks representing each type - including their internal variability - are illustrated in Appendix B, Figure B.4.

A random example of one mask per type is shown in Figure 6.12.

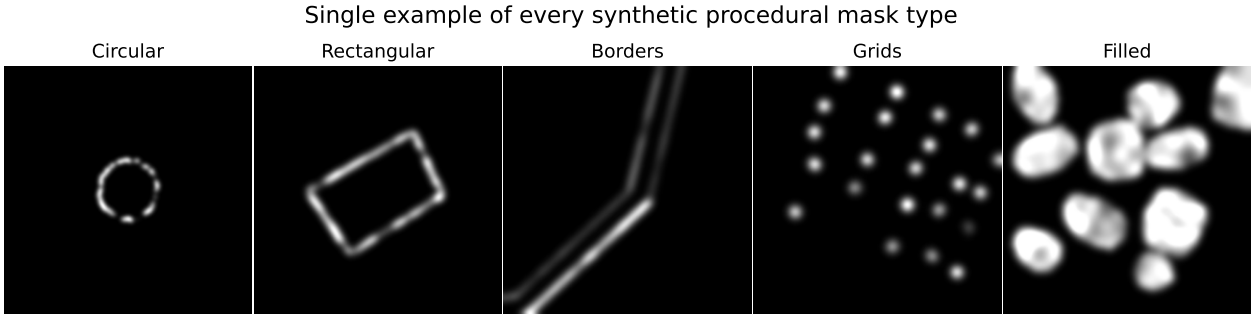

Figure 6.12: Example of one randomly generated procedural mask for each of the five defined types. For a more comprehensive overview, multiple samples per type are shown in Figure B.4.

Circular mask type. The Circular type includes elliptical and circular shapes with varying eccentricity. The deformation is achieved by modulating the ellipse radius with randomized offsets, while the axis ratio is sampled from a defined range to control elongation. Rotation is applied to the entire structure after generation, and the intensity along the contour varies using a cosine-based fading function combined with random jitter. Infrequently, parameters such as the fade factor or intensity variation are adjusted more aggressively, resulting in sharper or more eroded visual effects. This variability allows the generator to approximate a wide spectrum of circular archaeological features, from ring ditches to rounded enclosures.

Rectangular mask type. The rectangular type comprises three distinct subtypes-standard rectangles, rounded rectangles, and trapezoids — each introducing structural diversity while preserving the overall quadrilateral morphology. Cosine-based fade effect could be also applied. Rectangles are generated with randomized dimensions and subtle corner displacements to avoid perfect symmetry. Rounded rectangles extend this by varying the corner radius individually for each corner, resulting in softer shapes that resembling real structures. Trapezoids simulate tapering foundations or irregular plans, with additional stochastic removal of one side. All variants incorporate individual deformation patterns and rotation. The trapezoid generator also introduces variable top-to-base ratios and optional side omission, further increasing shape heterogeneity.

Border mask type. The borders type encompasses arc-based and linear (broken line) structures designed to mimic boundaries, ditches, and other enclosure-like features. A defining property of these generators is the use of elliptical or circular arcs with adjustable angular spans, flattening, and positional shifts. Functions create paired, spatially offset curves, often with wave-based modulation to imitate constructed boundaries, or emphasize continuity and smooth curvature, with randomized angular lengths and rotation. In some cases, side omission or irregular spacing is introduced to simulate real imperfections. These generators emphasize spatial flow and directional structure, distinguishing them from the more compact and enclosed mask types.

Grid mask type. The grid type includes both dotted rectangles and irregular point grids. These shapes simulate posthole patterns or foundation remains. Grid generators create arrays of circular elements with variable spacing, rotation, and missing points. Dotted rectangle function uses four-sided outlines filled with equidistant dots, with probabilistic omission of corners or points to simulate incomplete structures. The The grid function variant builds free-form grids with randomized dimensions and spacing.

Both approaches introduce controlled irregularities through deformation, dropout, and intensity variation, yielding highly structured yet naturally inconsistent patterns.

Filled mask type. The filled type includes compact, non-outlined structures resembling pits, ditches, or filled sunken features. It consists of three variants: randomly deformed filled patches, multi-patch compositions, and clusters of wavy ellipses (small patches). First two functions create one or more irregularly rounded shapes with bulging sides and randomized positioning near the image center. Overlap is prevented algorithmically. The last function distributes multiple distorted ellipses around a central point using a circular layout with significant probabilistic perturbation (regularization factor). Together, these masks introduce dense, high-contrast regions with organic variability and high morphological expressiveness, well-suited to simulating clustered archaeological features.

Integration of Synthetic Cropmarks into Real Imagery

Each mask type is associated with two to four distinct generation functions, which differ in both their parameterization and resulting shape (e. g., a broken line border, a rounded border, or a wavy border). Each of these functions is assigned a specific probability of use within the corresponding mask type.

Each real aerial image was combined with one (85%), two (10%) or three (5%) synthetic masks of randomly selected types. For each mask, a type was chosen at random, and one of the corresponding generation functions was applied to the image. This process was repeated accordingly for the specified number of masks. This reflects real-world scenarios where multiple cropmark types may appear within a single scene. In cases where a single mask type was selected, there is a 5% chance it could be applied repeatedly with a new random seed and with a random shift and spatial changes. This phenomenon also reflects real-world scenarios. Examples of multiple masks applied to a single image are shown in Figure B.6.

The final composite image was formed by overlaying the masks on the input image using either darkening (80% of cases - positive cropmarks) or brightening (20% - negative cropmarks), controlled by an opacity factor α. In 40% of the cases α was fixed ($\alpha = 0.1$), which the author considered to be the most realistic looking value, where synthetic structures are still observable even under challenging circumstances (e.g., dark imagery, grainy textures). In other cases, α was sampled uniformly from the interval $[0.1, 0.24]$ to ensure the variability in the distinctiveness of the mask.

Each mask was scaled according to the spatial resolution parameter m_{px}, which influenced size, deformation, and blur intensity. In case a selected mask function

failed to generate a valid output (e.g., due to extreme parameterization), an alternative function from the same mask morphology type was chosen.

Synthetic Dataset Construction

To pre-train a model capable of recognizing cropmark-like features, a synthetic dataset was created using procedurally generated masks. A total of approximately 25,000 orthophotos were obtained from the Czech Geoportal, covering a wide range of spatial resolutions ($m\_px \in \{0.1, 0.15, 0.2, 0.25, 0.3, 0.4, 0.6, 1.0, 1.8\}$), with roughly 3,000 samples per scale. The dataset was split into 93% training and 7% validation subsets.

To construct positive and negative samples, a structured sampling strategy was applied. Specifically, 25% of the images were used exclusively as positives by injecting synthetic cropmark masks, 25% were used only as negatives (unaltered), and the remaining 50% were duplicated across both classes - once with a mask and once without. This setup ensured that many images shared identical backgrounds across classes, differing only in the presence of cropmark-like features. The aim was to direct the model's attention toward the presence or absence of cropmarks, rather than irrelevant contextual cues.

This sampling strategy resulted in a synthetic dataset containing approximately 33,000 training samples and 2,300 validation samples. Importantly, the classes in this dataset are balanced (1:1 ratio), unlike the composition of the previous dataset *cropmark train set* (2:5 ratio). The real data validation set was introduced and curated using positive samples from the *extended cropmark train set* and negative samples from the original, non-extended *cropmark train set* - approaching a 1:1 class distribution in the real validation set.

The resulting synthetic cropmarks are, in many cases, highly convincing and difficult to distinguish from real ones. However, in some samples, their synthetic nature is clearly apparent - for example, patterns may extend into roads, forests, or other structures where cropmarks do not naturally occur. Nevertheless, the primary objective of the dataset is to increase the number of training samples and to enable the model to better generalize the concept of a cropmark, learning to distinguish between images with and without them.

Although some samples operate slightly outside the intended domain due to unsuitable background content, this is assumed to have minimal negative impact relative to the potential benefits of dataset augmentation. Examples of synthetic images generated by this method are presented in Appendix B, Figure B.5.

A comparison of synthetic and real cropmarks is shown in Figure B.7. The presence of structures resembling real data suggests that a procedurally generated dataset has

the potential to improve the performance of future models.

Pretraining Experiments on Synthetic Procedural Data

To evaluate the effectiveness of the synthetic dataset, a simple ablation study was conducted using a ResNet18 model pretrained on ImageNet. The models were trained on 30 epochs with hyperparameters obtained from last hyperparameter sweep on *extended cropmark train set*.

The model was trained four times using fixed random seed, each time with a different learning rate: 1×10^{-5}, 4×10^{-5}, 7×10^{-5}, and 1×10^{-4}. Performance was evaluated on two validation sets: the 7% synthetic validation subset and a real-data set based on the *extended cropmark train set*. This dual-validation setup was designed to assess both in-domain stability and out-of-domain generalization on the real data.

Figure 6.13 displays validation F1 scores and recall values over time. As expected, metrics on the synthetic validation set exhibit smoother trends, while those on real data fluctuate more significantly. This reflects the greater variability and noise inherent in real-world samples.

The best generalization to real validation data (*extended cropmark train set*) was achieved by the model with learning rate $4 \cdot 10^{-5}$, reaching an F1 score of 0.66 (and best overall accuracy 80.6%) on the real validation set in epoch 27. The following classification report (Table 6.11) and confusion matrix (Table 6.12) summarize its performance on *cropmark test set*:

Table 6.11: Classification report - *cropmark test set* (trained on procedural dataset, real validation set best F1 model, ResNet18, lr = 4×10^{-5}, epoch 27).

Class	Precision	Recall	F1-score	Support
Negative	0.843	0.911	0.876	336
Positive	0.559	0.400	0.466	95
Accuracy		0.798		431
Macro Avg	0.701	0.655	0.671	431
Weighted Avg	0.780	0.798	0.785	431

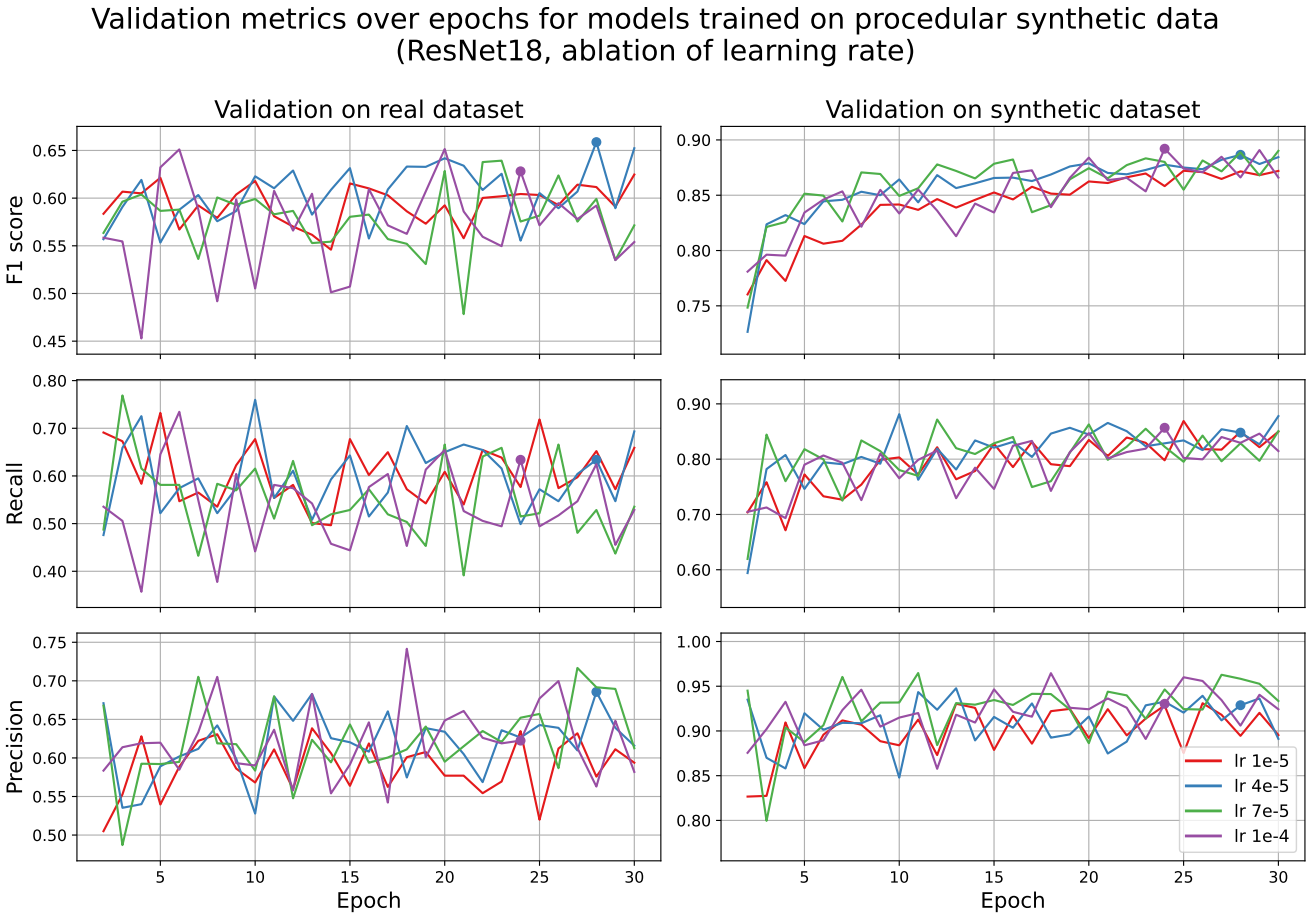

Figure 6.13: Validation performance across learning rates on synthetic procedural cropmark data. Each row shows a validation metric (F1 score, recall, precision) over 30 epochs, with evaluation on real data (left) and synthetic data (right). All models use ResNet18 and differ only in learning rate (color-coded). Highlighted points mark best-performing epochs of chosen models (best achieved F1 score). The best result on the real validation set was achieved by the model with a learning rate of 4×10^{-5} in epoch 27, while on the synthetic set, the best performance was obtained by the model with a learning rate of 10^{-4} in epoch 23.

Table 6.12: Confusion matrix - *cropmark test set* (trained on procedural dataset, real validation set best F1 model, ResNet18, lr $= 4 \times 10^{-5}$, epoch 27).

	Predicted Negative	Predicted Positive
True Negative	306	30
True Positive	57	38

The best result on the synthetic validation set was obtained using the highest learning rate ($1 \cdot 10^{-4}$), reaching an F1 score of 0.89 in epoch 23. This model was subsequently

evaluated on the *cropmark test set* (Tables 6.13 and 6.14) and surprisingly performed better than the previous model despite having never seen real data during training (directly or indirectly).

Table 6.13: Classification report - *cropmark test set* (trained on procedural dataset, synthetic validation set best F1 model, ResNet18, lr = 1×10^{-4}, epoch 23).

Class	Precision	Recall	F1-score	Support
Negative	0.884	0.774	0.825	336
Positive	0.445	0.642	0.526	95
Accuracy		0.745		431
Macro Avg	0.665	0.708	0.676	431
Weighted Avg	0.788	0.745	0.759	431

Table 6.14: Confusion matrix - *cropmark test set* (trained on procedural dataset, synthetic validation set best F1 model, ResNet18, lr = 1×10^{-4}, epoch 23).

	Predicted Negative	Predicted Positive
True Negative	260	76
True Positive	34	61

Discussion

Despite the expectation that validation performance on real data would better predict final test performance, the results suggest otherwise. The best-performing model on the *cropmark test set* was not the one that performed best on the real validation set but the one that achieved the best F1 score on the synthetic validation subset. This discrepancy may stem from a domain mismatch: the positive samples in the *extended cropmark train set* used for validation may differ strongly from the final test set due to the inclusion of new data. This shift potentially undermines its reliability as a proxy for test performance. Alternatively, the result could be explained by stochastic variance in the training process.

Nonetheless, the relative success of the synthetic-validation-optimized model indicates that high synthetic validation F1 score may serve as a better proxy for generalization in some scenarios.

These results provided a strong motivation for further experimentation, including finetuning of pretrained models on real datasets..

Cropmark Train Set Fine-tune

A logical next step was to fine-tune best-performing model using real annotated cropmark data to adapt the network to the real-world domain by leveraging its pre-learned representations of cropmark-like features from synthetic imagery.

As the starting point, we selected the best-performing model on the synthetic validation set - specifically, the ResNet18 model trained with a learning rate of 1×10^{-4}, which reached an F1 score of 0.89 on synthetic validation set and 0.526 on the *cropmark test set* in the Section 6.3.1. The same augmentation pipeline as illustrated in Figure 6.9 was used during fine-tuning.

The model was fine-tuned on the original (non-extended) *cropmark train set* for 15 epochs using AdamW optimizer with a learning rate of 5×10^{-5} and a weight decay of 1×10^{-5}. The best checkpoint, selected based on validation F1 score, was achieved after 9 epochs, where the model reached an F1 score of 0.795 and Recall of 0.825 on the *cropmark validation set*.

The selected checkpoint was subsequently evaluated on the *cropmark test set*. The resulting classification metrics are presented in Table 6.15 and the corresponding confusion matrix in Table 6.16.

Table 6.15: Classification report - *cropmark test set* (fine-tuned on *cropmark train set*, ResNet18, lr = 5×10^{-5}, epoch 9).

Class	Precision	Recall	F1-score	Support
Negative (0)	0.912	0.896	0.904	336
Positive (1)	0.653	0.695	0.673	95
Accuracy		0.852		
Macro Avg	0.783	0.795	0.789	431
Weighted Avg	0.855	0.852	0.853	431

Table 6.16: Confusion matrix - *cropmark test set* (fine-tuned on *cropmark train set*, ResNet18, lr = 5×10^{-5}, epoch 9).

	Predicted Negative	Predicted Positive
True Negative	301	35
True Positive	29	66

Compared to the previous best models, this fine-tuned variant exhibited a substantial performance boost. The resulting F1 score of 0.673 marks the highest observed so

far on the *cropmark test set*. Although the recall of 0.695 is slightly lower than the 0.716 achieved by the model trained on the extended dataset (this observed difference corresponds to two positive images in the test set, see more in Table 6.5), this fine-tuned model exhibits a significantly higher F1 score (by nearly 0.14, compared to the circa 0.02 recall difference), which justifies considering it the best-performing model up to this point.

These results validate the proposed two-step strategy: pre-training on procedurally generated data to learn general cropmark representations, followed by finetuning on domain-specific real-world imagery. In the subsequent section, this approach is further extended to incorporate the larger extended dataset during fine-tuning, aiming to push performance limits even further.

Extended Cropmark Train Set Finetune

Following the success of fine-tuning on the original *cropmark train set*, it was natural to explore whether further improvements could be achieved using the larger *Extended cropmark train set*. Based on results from Section 6.2.1, one might intuitively expect an increase in recall and a corresponding decrease in precision, reflecting the trade-offs observed in earlier experiments when the training data were expanded.

This hypothesis was tested by fine-tuning the model (initialized from the best synthetic-pretrained checkpoint) using the same training setup as in previous experiment. The best result (best F1 score on *extended cropmark validation set*) was obtained after 5 epochs of training.

Table 6.17: Classification report - *cropmark test set* (fine-tuned on *extended cropmark train set*, ResNet18, lr $= 5 \times 10^{-5}$, epoch 5).

Class	Precision	Recall	F1-score	Support
Negative (0)	0.936	0.658	0.773	336
Positive (1)	0.410	0.842	0.552	95
Accuracy		0.698		431
Macro Avg	0.673	0.750	0.662	431
Weighted Avg	0.820	0.698	0.724	431

Table 6.18: Confusion matrix - *cropmark test set* (fine-tuned on *extended cropmark train set*).

	Predicted Negative	Predicted Positive
True Negative	221	115
True Positive	15	80

The outcome confirmed the expectations: On the *cropmark test set*, precision for the positive class dropped to 0.410, while the recall reached 0.842 - the highest value achieved so far. Although the F1 score of 0.552 exceeds that of most models evaluated previously, it remains well below the 0.673 achieved by fine-tuning on the original, smaller dataset (Table 6.15). This suggests that while nearly three-quarters of the actual cropmarks were successfully identified, the model also misclassified more than a third of negative samples as positive.

This behavior highlights a recurring challenge. While high recall is desirable in many archaeological applications (to minimize missed sites), excessive false positives may limit the practical usability of the model, especially given that cropmarks are exceedingly rare in real landscapes. The *cropmark test set* was intentionally designed to include difficult negative cases that resemble cropmarks, which means that model performance may appear worse than it would in operational deployment. Nevertheless, if applied over large geographic regions, a high false positive rate could result in an overwhelming volume of false detections.

Training on Combined Dataset

To further evaluate the impact of synthetic data, additional experiments were conducted where the training set consisted of a combination of the *extended cropmark train set* and the *synthetic procedural train set*. Validation was performed on the two sets: the *synthetic validation set* and the real *extended cropmark validation set*.

The model was trained using the same configuration as the best-performing setup described in Section 6.2.1, including learning rate, batch size and optimizer. The checkpoint with the highest F1 score on the *extended cropmark validation set* was selected for evaluation on the final cropmark test set.

However, the model exhibited noticeably worse validation performance during training, and the final results confirmed this trend. The classification report is summarized in Table 6.19, with the corresponding confusion matrix in Table 6.20.

Table 6.19: Classification report - *cropmark test set* (trained on combined - real and procedural - dataset, best real F1 model).

Class	Precision	Recall	F1-score	Support
Negative (0)	0.767	0.500	0.605	336
Positive (1)	0.208	0.463	0.287	95
Accuracy		0.492		431
Macro Avg	0.487	0.482	0.446	431
Weighted Avg	0.644	0.492	0.535	431

Table 6.20: Confusion matrix - *cropmark test set* (trained on combined real and procedural data).

	Predicted Negative	Predicted Positive
True Negative	168	168
True Positive	51	44

The results were unsatisfactory. While recall was below average, precision reached the lowest value across all evaluated models. This indicates a high rate of false positives with only limited true positive detection. Given the poor performance and in comparison with the substantially better results achieved through the pre-train & finetune strategy, this direct combination approach was deemed ineffective and was not further developed.

6.3.2 Neural Synthetic Data Generation

Following the success of procedurally generated data, it was logical to explore complementary approaches based on neural generative models.

Image Generation Approach 1 - SDXL

As an initial attempt at neural image synthesis, a Stable Diffusion XL 1.0 (SDXL) model was fine-tuned directly on the *extended cropmark train set*. The training process employed standard optimization settings: a learning rate of 1×10^{-5}, batch size of 4, and 30 training epochs with gradient accumulation over 4 steps. Input images were resized to 512×512 pixels and augmented with random horizontal flips ($p = 0.5$), slight rotations ($\pm 10°$), and mild brightness and contrast jitter. All images were normalized to a mean and standard deviation of 0.5 in each RGB channel.

The training initially focused only on the positive class, using three descriptive prompts targeting typical cropmark structures. However, subsequent experiments at-

tempted a two-class fine-tuning strategy, using six analogous prompts for each class to better reflect binary classification needs.

Despite these efforts, the results were unsatisfactory. The generated samples appeared overly homogeneous, blurred, and lacked distinctive features. Although some faint structures loosely resembling cropmarks were present, they were visually insignificant and did not convey the desired characteristics. Representative examples of such outputs are shown in Figure 6.14.

The failure to adapt SDXL effectively can likely be attributed to the small size and low resolution of the training data. Given the scale and complexity of the SDXL architecture, successful fine-tuning typically requires large, high-quality datasets - conditions that were not met in this case. As a result, a more efficient and data-friendly alternative was pursued: Low-Rank Adaptation (LoRA), which allows fine-tuning only selected submodules of the diffusion model without modifying the entire network.

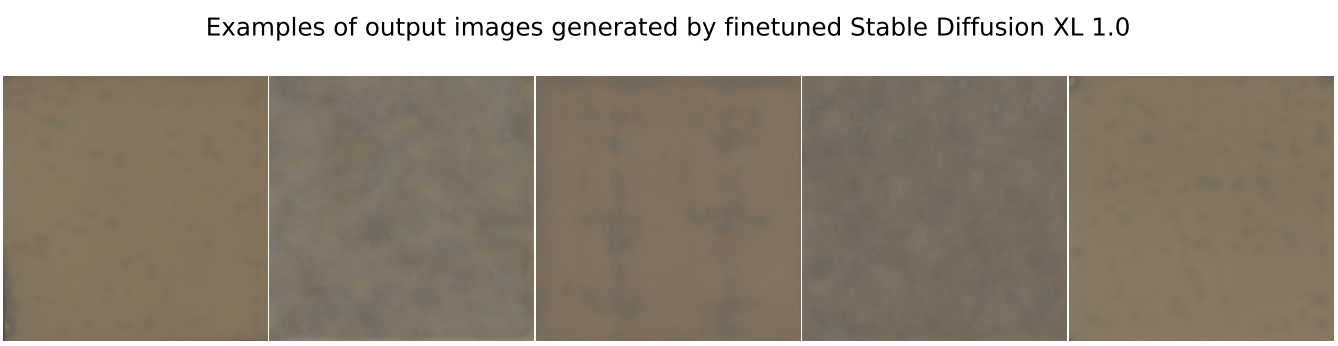

Figure 6.14: Validation output of Stable Diffusion XL 1.0 after full-model fine-tuning on the positive class from the *extended cropmark train set*. Fine-tuning the entire model with limited data proved ineffective. The outputs are overly homogeneous, blurred, and lack sufficiently distinct target cropmark structures. However, certain visual artifacts resembling patterns from real-world data can still be observed.

Image Generation Approach 2 - LoRA

As a second approach, LoRA fine-tuning was chosen for the reasons outlined above. To ensure comparability with the procedural dataset, the same morphological cropmark categories were retained. Real training images (*extended cropmark train set*) were divided into these categories, and a separate SDXL model was trained for each of this types using LoRA fine-tuning. All training was performed in Google Colab using the repository by *jhj0517* (available at `https://github.com/jhj0517`). For each type, the base training prompt is listed in Table 6.21. The generation phase then used 12 varied prompts per type, designed with the assistance of Claude.ai [20]. These prompts

altered conditions such as scale, number and size of cropmarks, visibility, background texture, and color variation, increasing diversity in the generated dataset. For instance, all prompts used to generate *rectangular type* images are listed in Table 6.22.

Table 6.21: Training prompts used during SDXL LoRA fine-tuning for individual cropmark types.

Type	**Training Prompt**
`negative`	Aerial orthophoto of a regular agricultural field with no visible archaeological features.
`grids`	Aerial orthophoto of a field with a posthole doted structure (remains of pillar holes, slightly darker and regularly placed dots in crop) of archaeological origin.
`borders`	Aerial orthophoto of a field with a linear (curved, angled, or parallel) archaeological cropmark in it.
`circular`	Aerial orthophoto of a field with distinctive rectangular archeological cropmark in it.
`rectangular`	Aerial orthophoto of a field with distinctive circular archeological cropmark in it.
`filled`	Aerial orthophoto of a field with an archaeological remains of pits and sunken structures — multiple clustered filled compact darker dots and patches of greener vegetation.

Table 6.22: Generation prompts used for synthetic data creation of the `rectangular` cropmark type.

Prompt
Aerial orthophoto of a field with a distinctive rectangular archaeological cropmark clearly visible in the center.
Aerial orthophoto of a green field with a small, high-contrast rectangular cropmark in sharp edges.
Aerial orthophoto of a golden-yellow mature field with a thick-lined, well-defined rectangular cropmark.
Aerial view of a field showing a weathered but still readable rectangular cropmark with slightly irregular borders.
Aerial orthophoto showing two intersecting rectangular cropmarks on a dry soil background.
High-resolution aerial image of a narrow rectangular cropmark on a grassy field with mild distortions.
Aerial image of a field with a large rectangular ring-shaped cropmark with bold outlines on brown soil.
Orthophoto of a greenish field with a pale but geometrically regular rectangular cropmark, partially surrounded by noise.
Aerial orthophoto of a field with two small parallel rectangular cropmarks close to each other.
Top-down aerial image of a distinct rectangular archaeological structure with clear right angles.
Aerial orthophoto of a broken rectangular cropmark with one faded side, but still structurally visible.
Satellite-style orthophoto of a large field showing a long, thin rectangular cropmark with sharp corners.

The training was conducted with the following hyperparameters: learning rate 5×10^{-5} (cosine scheduler), mixed precision set to *bf16*, batch size 1, *gradient accumulation steps* of 4, checkpointing every 5 steps, a total of 100 training steps. These settings were selected to balance training stability, memory efficiency, and runtime constraints and costs of the Colab environment (L4, A100).

To increase diversity and realism, the final dataset incorporates samples generated from model checkpoints saved at various training steps. This approach captures the stylistic evolution of the model over time: in the early stages of training, cropmarks tend to appear overly artificial or even cartoon-like, while the background textures are

relatively detailed and varied. As training progresses, the cropmarks become more realistic and better integrated into the imagery, whereas the backgrounds gradually lose complexity and tend toward homogeneity. This progression reflects a well-documented behavior of diffusion models under limited-step fine-tuning. Representative examples of generated images at different training stages are shown in Figure 6.15.

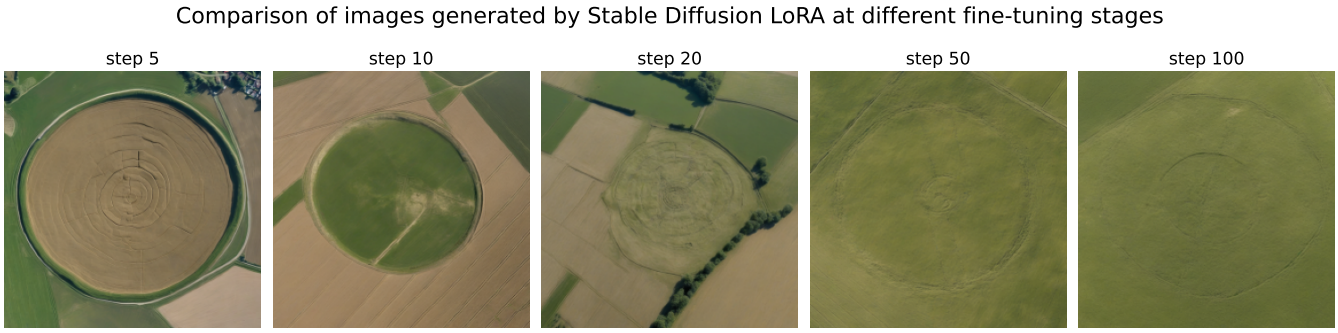

Figure 6.15: Examples of circular cropmark images generated at different stages of LoRA Stable Diffusion 1.0 fine-tuning. A gradual increase in realism can be observed, accompanied by a decrease in contrast and a reduction texture distinctivness.

Synthetic Neural Dataset The neural generative model (SDXL) produces high-resolution images at 1024×1024 pixels, which is unnecessarily large for the downstream classification task. To reduce computational requirements and introduce controlled variability, a preprocessing pipeline was applied to rescale and crop these images prior to their use in training.

Three augmentation strategies were employed (one or more for each image):

- **Direct resizing** to 400×400 pixels using Lanczos resampling.
- **Center cropping** a 512×512 region from the middle of the image followed by resizing.
- **Random cropping** a 512×512 region from a random location followed by resizing.

The cropping size of 512×512 was selected to ensure that even after downscaling, relevant cropmark features remain visible and spatially discernible. Since the cropmarks were generated (not intentionally) to be centrally positioned within the SDXL outputs, both center and random crops are likely to retain relevant structures. Random cropping introduces additional spatial variation, which enhances the model's robustness to feature location and reduces overfitting to positional priors.

During preprocessing, each image was transformed using one of the three strategies with fixed probabilities (60% resize, 20% center crop, 20% random crop). Additionally,

in 30% of cases, all remaining transformations were also applied, resulting in multiple augmented versions per original image.

The resulting dataset consists of approximately 8,500 negative samples and 7,000 positive samples, each divided into training and validation subsets. The class balance and diversity of spatial configurations aim to support stable training and mitigate overfitting, while reflecting the inherent structure of the LoRA-generated data.

Representative examples of the generated training images are provided in Appendix B, Figures B.8 and B.9. Figure B.8 shows a selection of positive samples categorized by type, while Figure B.9 illustrates typical outputs for the negative class.

A visual inspection reveals several consistent trends across the generated data. Notably, the model tends to overproduce circular features, even in cases where other morphologies were intended. This behavior is likely a residual effect of pretraining data, where circular structures and enclosures are frequent. As a consequence, the distinction between types such as *grids* and *filled* is often poorly expressed. The model also often struggles to produce convincing examples of the *rectangular* type, which are prone to distortion or lack geometric clarity. *Borders* type tends to be rendered more successfully, particularly in cases where curved or rounded borders are expected.

The visual quality of the synthetic cropmarks varies: some examples appear artificial or stylized, while others reflect real-world cropmarks with surprising realism. This heterogeneity reflects both the strengths and limitations of the LoRA-based generative approach.

Neural Synthetic Pre-training

Following the successful training on procedurally generated data, we investigated whether a model pre-trained on synthetically generated images from a neural diffusion model (Stable Diffusion XL fine-tuned via LoRA) could be used as a viable initialization for downstream cropmark identification. The motivation for this approach was twofold: first, to explore the utility of higher-fidelity synthetic data, and second, to evaluate whether feature extractors trained on these data could contribute to downstream performance even if the data generation was not strictly controlled.

A ResNet18 model (initialized with ImageNet weights) was therefore trained on the neural synthetic dataset. The training configuration (hyperparameters and augmentations) matched that of the procedural pre-training described in Section 6.3.1. Validation was carried out using two sets: (1) a randomly selected 10% subset of the neural synthetic dataset and (2) the same real validation set (subset of *extended cropmark train set* maintaining a 1:1 ratio, i.e., all psoitive samples and circa 40% of negative samples are present).

Due to the method of generating the neural synthetic data, there is no guaranteed correspondence in the background between positive and negative samples. As a result, the model may learn artifacts of the generation process rather than the desired morphological characteristics of cropmarks.

As illustrated in Figure 6.16, the model rapidly improved its performance on the synthetic validation set but failed to achieve meaningful gains on the real validation set, which could confirm previous concerns. This suggests that the model may overfit to features specific to the generated images (possibly due to homogeneity of backgrounds or other synthetic artifacts) without learning generalizable cropmark-relevant representations.

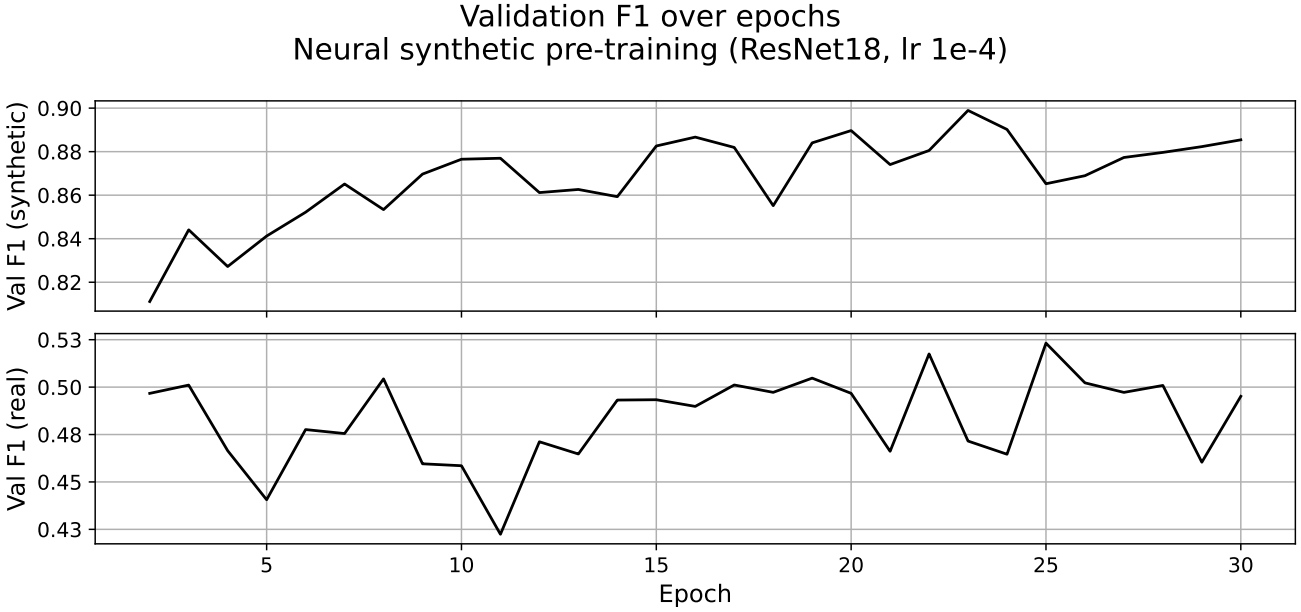

Figure 6.16: F1 score progression on both validation sets (synthetic and real) during pretraining of the ResNet18 model on procedurally generated synthetic data. The learning rate was set to 1×10^{-4}. A rising trend in F1 score can be observed on the synthetic validation set, while the real validation set exhibits a more oscillatory behavior.

The best-performing model on the synthetic validation set was subsequently evaluated on the final *cropmark test set*. However, its performance was worse than that of a baseline classifier making decisions solely based on class priors in the dataset (F1 score of 0.155). For this reason, the corresponding results are omitted from the main evaluation tables.

Despite its failure as a standalone classifier, the pre-trained model may still possess useful internal representations, such as filters for high-level shapes or spatial arrangements. Therefore, in the following section, we explore whether this model can be

effectively fine-tuned on real data to improve its performance in a real-world context.

Model Fine-tuning (Neural Synthetic Dataset)

Despite the poor performance observed during pre-training on neurally synthesized data alone, a follow-up experiment was conducted to assess whether fine-tuning such a model on the *cropmark train set* could still lead to improvement. The best validation F1 score was achieved after epoch 11 (out of 15), and the corresponding model was evaluated on the *cropmark test set* (Tables 6.23 and 6.24).

Table 6.23: Classification report - *cropmark test set* (neural synthetic pre-training with fine-tuning on *cropmark train set*).

Class	Precision	Recall	F1-score	Support
Negative (0)	0.824	0.875	0.848	336
Positive (1)	0.432	0.337	0.379	95
Accuracy		0.756		431
Macro Avg	0.628	0.606	0.614	431
Weighted Avg	0.737	0.756	0.745	431

Table 6.24: Confusion matrix - *cropmark test set* (neural pre-training with *cropmark train set* fine-tune).

	Predicted Negative	Predicted Positive
True Negative	294	42
True Positive	63	32

The resulting F1 score for the positive class (0.379) is notably lower than that of the best model trained on procedurally generated data followed by real-data fine-tuning (F1 = 0.67). In addition, recall dropped to 0.337, the lowest among all finetuned tested models, while precision reached only 0.432. Although the outcome could partially reflect random variation or noise in the validation process, there is no evidence suggesting that this model could outperform the top-performing baseline. These results reinforce the conclusion that procedural synthetic data offer a more suitable pre-training foundation for this task than neural synthetic data alone.

The model initially pretrained on neurally synthesized data was also fine-tuned on the *extended cropmark train set*. Compared to fine-tuning on the original (non-extended) training set, this model achieved slightly better results on the *cropmark test set* (Tables 6.25 and 6.26).

Table 6.25: Classification report - *cropmark test set* (neural synthetic pre-training with fine-tuning on *extended cropmark train set*).

Class	Precision	Recall	F1-score	Support
Negative (0)	0.865	0.741	0.798	336
Positive (1)	0.392	0.589	0.471	95
Accuracy		0.708		431
Macro Avg	0.628	0.665	0.634	431
Weighted Avg	0.760	0.708	0.726	431

Table 6.26: Confusion matrix - *cropmark test set* (neural pre-training with fine-tune on *extended cropmark train set*).

	Predicted Negative	Predicted Positive
True Negative	249	87
True Positive	39	56

Once again, the characteristic effect of training on the extended cropmark dataset was observed: a decrease in precision (0.392) accompanied by an increase in recall (0.589), compared to the same model fine-tuned on the non-extended dataset. This trend is consistent with earlier findings (see Section 6.2.1).

However, the absolute performance of this approach still lags behind the best-performing model - trained on procedurally synthesized data and fine-tuned on real cropmark data - across all key metrics. This confirms the earlier conclusion that procedurally generated data remain more effective as a pretraining domain in this task.

Double Synthetic Pre-train and Fine-tune

As one of the final experimental setup, a multi-stage training pipeline was tested. The model was first pre-trained (learning rate 1×10^{-4}) on neurally synthesized data, selecting the checkpoint with the best F1 score on the synthetic validation set. This model was then further pre-trained (learning rate 1×10^{-4}) on procedurally generated synthetic data, again choosing the best checkpoint based on synthetic validation F1. Finally, the resulting model was fine-tuned (learning rate 5×10^{-5}) on the real *cropmark train set*, and the version with the best F1 score on the real validation set (epoch 12) was selected. The final model was evaluated on the standard *cropmark test set* (Tables 6.27 and 6.28).

Table 6.27: Classification report - *cropmark test set* (double synthetic pre-training followed by real finetuning).

Class	Precision	Recall	F1-score	Support
Negative (0)	0.895	0.866	0.880	336
Positive (1)	0.575	0.642	0.607	95
Accuracy		0.817		431
Macro Avg	0.735	0.754	0.744	431
Weighted Avg	0.825	0.817	0.820	431

Table 6.28: Confusion matrix - *cropmark test set* (double synthetic pre-training with finetune).

	Predicted Negative	Predicted Positive
True Negative	291	45
True Positive	34	61

While the model performs reasonably well, achieving an F1 score of 0.607 for the positive class and an overall accuracy of 81.7%, it is still outperformed across all key metrics by the model trained solely on procedurally generated data followed by fine-tuning on the real *cropmark train set*. This suggests that the added complexity of double synthetic pretraining does not necessarily translate into better performance and may introduce domain inconsistencies. On the other hand, this may represent a fluctuation reflecting how well the model happened to align with the preferences of the *cropmark test set*.

In the next logical experiment, the model was pre-trained first on neurally synthesized data (the same pretrained model was chosen), then on procedurally generated synthetic data, and finally fine-tuned on the *extended cropmark train set*. The evaluation on the *cropmark test set* yielded the following results:

Table 6.29: Classification report - *cropmark test set* (neural + procedural pre-training, finetune on *extended cropmark train set*).

Class	Precision	Recall	F1-score	Support
Negative (0)	0.908	0.705	0.794	336
Positive (1)	0.418	0.747	0.536	95
Accuracy		0.715		431
Macro Avg	0.663	0.726	0.665	431
Weighted Avg	0.800	0.715	0.737	431

Table 6.30: Confusion matrix - *cropmark test set* (neural + procedural pre-training, finetune on *extended cropmark train set*).

	Predicted Negative	Predicted Positive
True Negative	237	99
True Positive	24	71

We once again observe the now familiar effect of using the *extended cropmark train set*: a noticeable increase in recall (0.747) comes at the cost of a substantial drop in precision (0.418). Although the recall exceeds that of the best-performing model, the degradation in precision is significant enough that the model trained on the procedural dataset and fine-tuned on the original *cropmark train set* remains the overall best performer.

6.4 Evaluation

6.4.1 Cropmark Test Set Results Summary

Over the course of the thesis, a total of thirteen (plus one omitted) distinct models were trained and evaluated on the *cropmark test set*. These models represent a wide range of training strategies, including baseline training on real data, pre-training on synthetic datasets (both procedural and neural), fine-tuning, as well as combinations thereof. A complete overview of their performance is provided in Table 6.31.

Among all evaluated models, the best-performing configuration was the one pre-trained on the synthetic procedural train set and subsequently fine-tuned on the original *cropmark train set*. This model was selected from four training runs with varying learning rates, and was chosen based on the best F1 score achieved on the synthetic validation set. On the held-out *cropmark test set*, this model reached the highest F1 score overall, while also achieving the best precision and accuracy across all models.

Although some alternative configurations yielded higher recall - particularly those involving training on the extended dataset - these gains were consistently offset by a more substantial drop in precision, ultimately leading to a lower F1 score. For this reason, the mentioned model is considered the most balanced and effective among all tested configurations and is further analysed in the following section.

Table 6.31: Summary of all models evaluated in this work (precision, recall and F1 score) on the *cropmark test set*. The table includes the applied initial learning rate (LR) in fine-tune phase, along with the datasets used for pre-training and fine-tuning. The *pre-train* column indicates the dataset used to initialize the model, while *fine-tune* denotes the dataset used in the final training phase. For clarity, the following abbreviations are used: CTrS - *cropmark train set*, ECTrS - *extended cropmark train set*, SPTrS - *synthetic procedural train set*, and SNTrS - *synthetic neural train set*. Additional details on the models can be found in the Experiments section.

model	**note**	**pre-train**	**fine-tune**	**LR**	**Prec.**	**Rec.**	**F1**	**Acc.**
model 1	augm. v1	ImageNet	CTrS	8.8e-4	0.523	0.579	0.550	0.795
model 2	augm. v2	ImageNet	CTrS	8.8e-4	0.568	0.663	0.612	0.814
model 3		ImageNet	ECTrS	4.9e-5	0.430	0.716	0.538	0.729
model 4		model 3	CTrS	1e-5	0.414	0.705	0.521	0.715
model 5	best real. val	ImageNet	SPTrS	4e-5	0.559	0.400	0.466	0.798
model 6	best syn. val	ImageNet	SPTrS	1e-4	0.445	0.642	0.526	0.745
model 7	best perform	model 6	CTrS	5e-5	**0.653**	0.695	**0.673**	**0.852**
model 8		model 6	ECTrS	5e-5	0.410	**0.842**	0.552	0.698
model 9	combined dataset	ImageNet	SPTrS & ECTrS	5e-5	0.208	0.463	0.287	0.492
model 10		ImageNet, SNTrS	CTrS	5e-5	0.432	0.337	0.379	0.756
model 11		ImageNet, SNTrS	ECTrS	5e-5	0.392	0.589	0.471	0.708
model 12		ImageNet, SNTrS, SPTrS	CTrS	5e-5	0.575	0.642	0.607	0.817
model 13		ImageNet, SNTrS, SPTrS	ECTrS	5e-5	0.418	0.747	0.536	0.715

6.4.2 Real Application Results

To evaluate the best model performance on real obtainable and aerial imagery for automation process, a set of testing tiles was acquired from open orthophoto data provided by the Czech Office for Surveying, Mapping and Cadastre (ČÚZK) portal. For each test location, a 7×7 grid of tiles was extracted around a central coordinate, covering an area of approximately 78.4×78.4 meters. Each tile was 224×224 pixels in size, corresponding to a ground resolution of 0.3 m/px, which was chosen as the target scale because it enhances both small and large-scale cropmark features and corresponds to the median resolution used in the synthetic procedural dataset.

To mitigate potential bias introduced by manually selected site centers, each location was subject to a small random shift - uniformly sampled within one-third of the tile width - in both latitude and longitude. This approach ensured more robust and reproducible evaluation. Tiles were extracted for all 12 positive test locations and a subset of 29 negative ones. The negative set was downsampled from a larger pool to maintain a similar class distribution to the original test set.

The dataset used for this evaluation was curated from orthophotos acquired over multiple years. The selection of years was made deliberately by the author, who excluded certain time frames due to unsuitable vegetation or seasonal conditions. Since aerial surveying in the Czech Republic is conducted in spatially staggered cycles, the available years differ across regions - some localities are represented by different time spans than others.

For the purpose of this evaluation experiment, a simplified version of the processing pipeline, as illustrated in Figure 6.10, was employed. Specifically, an agricultural classifier was used to filter out irrelevant scenes and a brightness-based thresholding step was applied to filter out blank images of regions not captured in the given year by the geoportal imagery. Aerial/non-aerial classifier was deemed unnecessary in this case.

Each location was assessed in its available years and a location was considered to be recognized as positive if at least one tile in the central layers (layer 0 or 1, i.e., the 3×3 region centered on the location) was classified as containing cropmarks. The resulting per-location statistics - including the name, number of observed and years, number of positive tiles in the center, the proportion of non-agricultural tiles removed (filtered out by a pre-trained agricultural classifier), and the final ratio of positive tiles - are summarized in Tables 6.32 and 6.33.

Importantly, the tile generation methodology and selection criteria were defined parametrically and with respect to data grouping before any model inference was made, ensuring that the process adhered to the proper validation principles.

Table 6.32: Per-location summary of evaluation results for the positive class — center detections, positive tile ratio, and domain filtering.

Location	**Center Pos. Y.**	**Total Y.**	**Pos. Tile Ratio**	**Non-Field Ratio**
Velké Žernoseky	2	5	0.01	0.21
Terezín	0	5	0.00	0.02
Litoměřice	1	5	0.04	0.01
Tišice	3	5	0.03	0.06
Přívory (1)	3	5	0.06	0.17
Přívory (2)	2	5	0.02	0.04
Kostelec	1	5	0.02	0.02
Chleby	0	5	0.02	0.01
Netřebice	0	5	0.00	0.02
Poděbrady	2	5	0.05	0.08
Opolany	2	5	0.03	0.24
Libice	2	5	0.07	0.03

Table 6.33: Per-location summary of evaluation results for the negative class — center detections, positive tile ratio, and domain filtering.

Location	Center Pos. Y.	Total Y.	Pos. Tile Ratio	Non-Field Ratio
Terezín	0	5	0.01	0.00
Tašov	1	5	0.02	0.17
Litoměřice	1	5	0.02	0.02
Středenice	0	5	0.00	0.01
Přívory	1	5	0.02	0.00
Kostelec n. L.	1	5	0.03	0.00
Benátky n. J.	2	5	0.12	0.00
Předměřice	0	5	0.04	0.00
Chleby	0	5	0.00	0.00
Poděbrady	2	5	0.03	0.14
Opolany	0	5	0.01	0.11
Domáslovice	1	5	0.00	0.01
Radovesnice	0	6	0.01	0.20
Kundratice	0	6	0.02	0.08
Kladruby n. L.	0	5	0.00	0.04
Přelouč	0	5	0.01	0.10
Habrkovice	0	5	0.00	0.27
Číčovice	0	5	0.00	0.15
Koleč	0	5	0.00	0.07
Plzeň (1)	0	5	0.01	0.00
Brod u Stříbra	0	5	0.01	0.00
Plzeň (2)	2	5	0.18	0.00
Libice	2	5	0.02	0.06
Kochánky	2	5	0.10	0.00
Třebel	0	5	0.00	0.32
Stebno	2	5	0.20	0.07
Libouchec	1	5	0.00	0.03
Stadice	2	5	0.02	0.07
Rýdeč	0	2	0.00	0.89

Although the nature of this evaluation differs from the main experiments - focusing on aggregate location-level behavior rather than independent image classification - it is still meaningful to compute the same performance metrics as those used throughout the thesis. This consistency allows for easier comparison and situates the results within

the broader context of the classification task. To that end, the confusion matrix and derived classification metrics for this real-world evaluation are shown in Table 6.35 and Table 6.34, respectively. It should be emphasized, however, that this evaluation scheme represents only one of many possible methodologies and may not be fully optimal in terms of sensitivity, robustness, or practical deployment alignment.

Table 6.34: Classification report - real data test experiment (best model from procedural pre-training followed by fine-tuning on real data).

Class	Precision	Recall	F1-score	Support
Negative (0)	0.850	0.586	0.694	29
Positive (1)	0.429	0.750	0.545	12
Accuracy		0.634		41
Macro Avg	0.639	0.668	0.620	41

Table 6.35: Confusion matrix - real data test experiment (best procedural fine-tuned model).

	Predicted Negative	Predicted Positive
True Negative	17	12
True Positive	3	9

The model achieved a recall of 0.750 and a precision of 0.429 for the positive class (Table 6.34).The lower precision reflects a higher rate of false positives. However, these values should be interpreted with caution, as the evaluation procedure in this case differs substantially from the main experiments and is therefore not directly comparable.

These results are in line with earlier observations: models often detect cropmarks well (realtively high recall) but at the cost of precision. Compared to the best model evaluated on the *cropmark test set* (0.67 F1 score for positive class), the current result is lower, but still meaningful given the different evaluation setup and smaller sample size.

Discussion

The relatively high recall observed in this evaluation may be partially explained by the structure of the experiment itself. In many cases, cropmarks appear in clusters, or multiple morphological types co-occur within a single locality. Since the evaluation considers a location as positive if at least one central tile is predicted as such in any observed year, even a single strong cropmark instance can determine the result. This differs from previous experiments, where each image was evaluated independently, without

temporal or spatial aggregation. As a result, the observed recall may appear higher and less directly comparable.

A possible refinement would be to require a positive center prediction in at least two years (rather than one) for a location to be classified as positive. This thresholding approach is justifiable, as it helps reduce the impact of year-specific anomalies or isolated false positives. In this adjusted setup, the model's metrics would shift approximately to: precision 0.54, recall 0.58, and F1 score 0.56. While recall decreases, the notable gain in precision indicates a more conservative and reliable performance.

Such stricter criteria may be better suited for real-world deployment, where false positives can overwhelm users and degrade overall trust in the system. Increasing decision robustness through multi-year evidence would reduce noise in candidate results and provide users with more actionable outputs. Additional parameter adjustments - including classifier threshold tuning or spatial aggregation logic - remain possible and may further improve the evaluation fidelity. The evaluation method itself - based on tile-level predictions aggregated per location - could also be improved or adapted.

As shown in Table 6.33, some predicted positive tiles appeared within the neighbourhood of the negative class, typically in a small percentage of cases. This statistic is aggregated across years - if a tile is predicted as positive in at least one year, it is considered positive for the purposes of this analysis. There is still potential for further refinement by adjusting model parameters as stated above. A more detailed breakdown by year and frequency of occurrence could enable the development of a relatively effective model suitable for deployment in practical archaeological field surveys.

6.4.3 Best-performing Models Analysis

To conclude the experimental part of this work, we now perform an analysis of the best-performing model on the *cropmark test set*, which may provide valuable insights. Such an analysis is conducted at this stage to avoid compromising the integrity and objectivity of earlier test evaluations.

To further investigate the behavior of the best-performing model (procedural pretraining followed by fine-tuning on the *cropmark train set*), a threshold sensitivity analysis was conducted. Instead of fixing the classification threshold at the default value of 0.5, various thresholds were applied to the model's sigmoid outputs on the test set, and key evaluation metrics - precision, recall, F1 score, and accuracy - were computed for each setting.

The results of this analysis are visualized in Figure 6.17. It can be observed that the model exhibits robust performance under small deviations from the standard threshold.

Slightly increasing the threshold leads to a modest improvement in overall F1 score, suggesting that the default threshold may be slightly suboptimal. However, considering that the model was not tuned on the test set, this behavior remains acceptable and indicates good generalization.

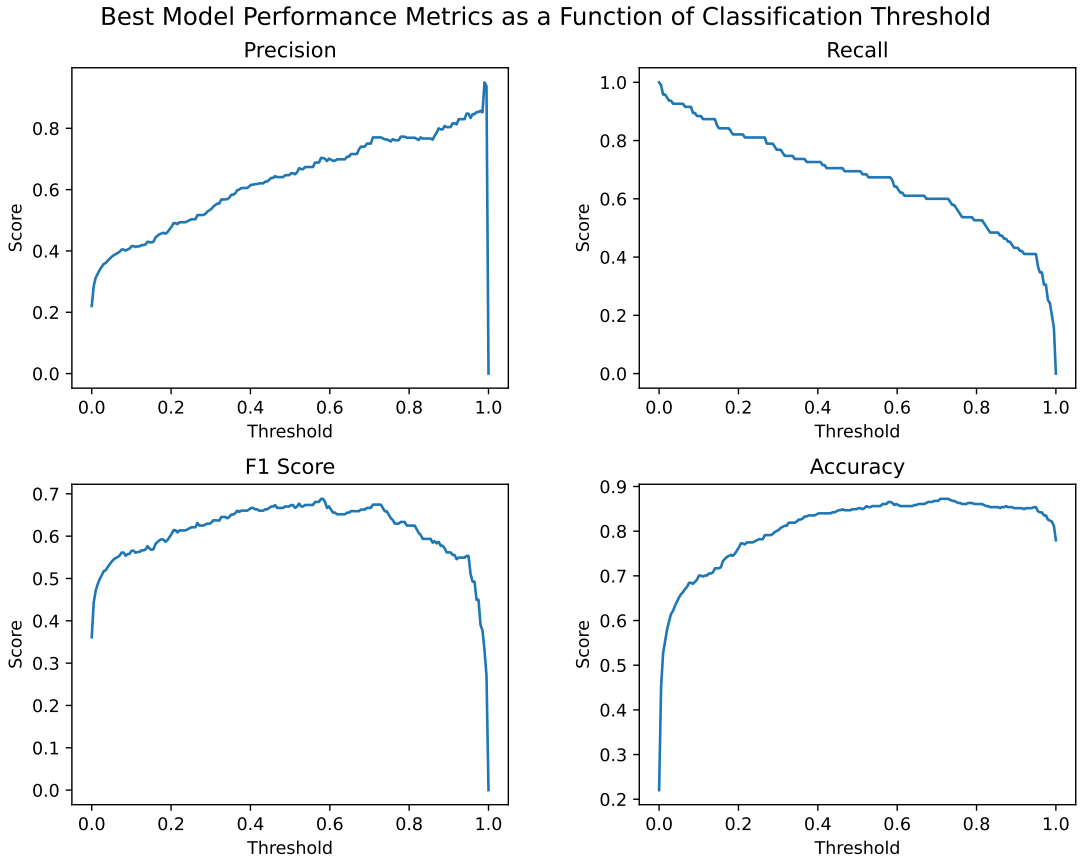

Figure 6.17: Test set metric curves on the *cropmark test set* as a function of the classification threshold for the best-performing model (Model 7). The figure shows that the model's performance, particularly in terms of F1 score, remains robust across small variations in the decision threshold.

As expected, the trade-off between precision and recall is evident: increasing the threshold improves precision at the cost of recall, while decreasing it has the opposite effect. This confirms that the threshold can be adjusted depending on the desired operating point - prioritizing either sensitivity (recall) or specificity (precision) depending on the practical use case. In the case of small threshold adjustments, this prioritization can be implemented without a significant impact on the F1 score.

However, it is important to note that any adjustment of the classification threshold should ideally be performed on validation data, rather than on the test set. This analysis was not intended to optimize model performance, but rather to provide an informative insight into the model's behavior on unseen dat and discussion topic.

To gain further insight into the model's decision boundaries, Figure 6.18 displays the ten most uncertain predictions made by the best-performing model - specifically, the samples whose predicted probabilities are closest to the default threshold of 0.5. This subset includes both positive and negative examples and highlights cases where the model exhibited the least confidence in its classification.

Among the negative-class samples, we can observe instances such as a former scout camp site, circular trails, and features that may resemble cropmarks but are likely caused by subsurface moisture variation or water flow.

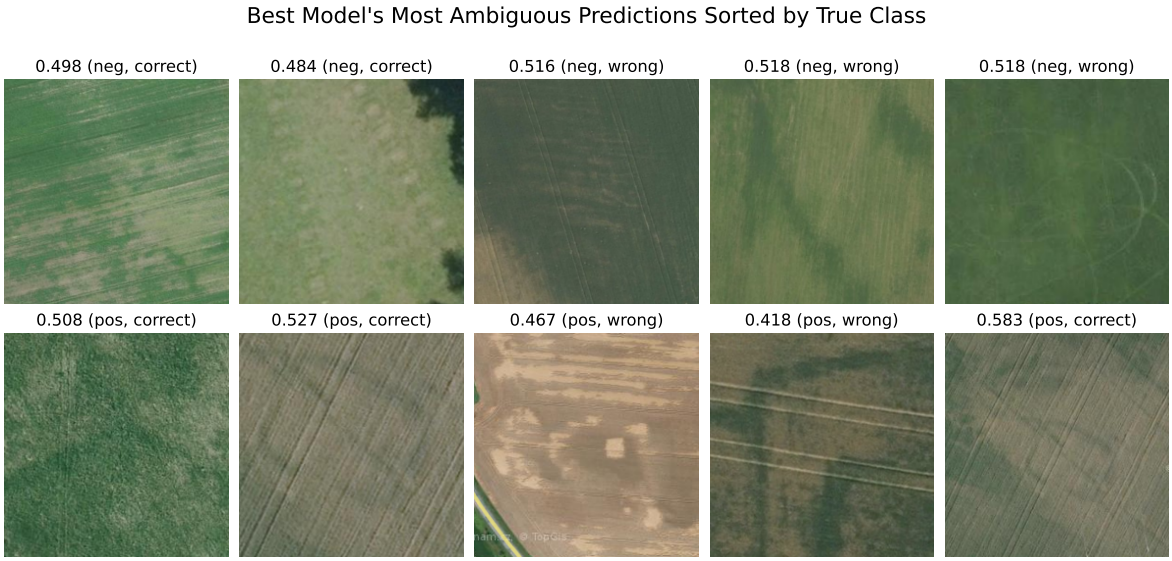

Figure 6.18: Ten most uncertain test samples (5 from each class) according to predictions by the best-performing model (Model 7). Each image is annotated with the predicted score, true class (pos/neg), and whether the classification was correct at the default threshold of 0.5.

Figure 6.19 shows the distribution of predicted probabilities for the misclassified samples of the best-performing model. Notably, a substantial portion of these incorrect predictions lies near the extremes of the probability scale, indicating that the model was highly confident in its (incorrect) decisions.

This observation may reflect two contributing factors. First, it suggests that certain challenging samples exhibit visual patterns that strongly resemble the opposite class, leading the model to make confident but incorrect inferences. Second, it may point to limitations in the training data - for example, mislabeled or borderline examples - or to inherent ambiguities in the visual features associated with cropmarks and their mimics. In either case, the result highlights the importance of refining the training data and incorporating mechanisms for uncertainty calibration in future work.

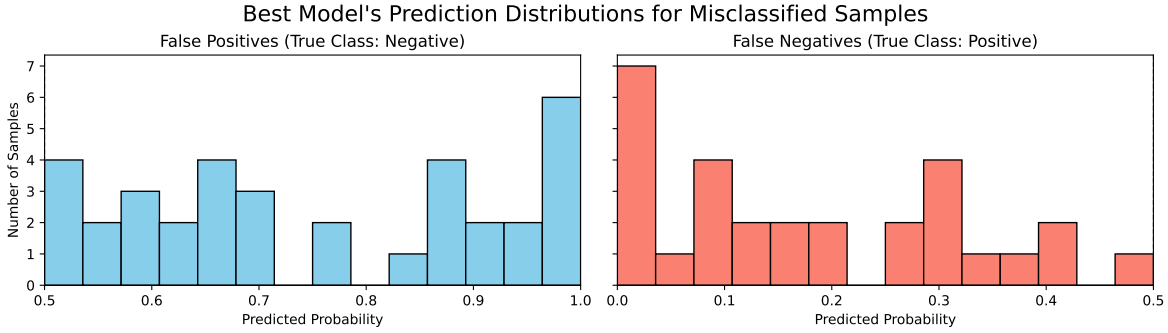

Figure 6.19:

6.4.4 Limitations and Potential Improvements

While the experiments in this thesis demonstrate the feasibility of cropmark detection using neural networks, several limitations emerged throughout the process. This section summarizes key challenges and outlines possible directions for future work.

Domain Shift and Generalization.

A consistent issue was the mismatch between validation and test performance, particularly when using the *extended cropmark train set*. This indicates a domain shift likely caused by regional, seasonal, or stylistic differences in the imagery. Incorporating domain adaptation techniques or measuring domain similarity explicitly could improve robustness.

Limited Real Training Data.

Despite efforts to expand the dataset, the number of annotated real positives remained limited. While synthetic data alleviated this to some extent, further improvement depends on access to high-quality annotations. In the presence of richer data, segmentation-based approaches (e.g., mask prediction or heatmap regression) could be employed instead of binary classification, allowing more detailed supervision.

Precision–Recall Trade-off.

Many models achieved high recall but low precision, especially those trained on synthetic or extended datasets. This poses a challenge in large-scale applications, where false positives may overwhelm the results. Threshold tuning, calibrated scoring, or postprocessing heuristics could be used to mitigate this.

Synthetic Data Quality.

Procedural data provided useful structure, while neural synthesis added visual realism. However, both have drawbacks: procedural masks lack fine texture variation, and diffusion-based outputs tend to become overly homogeneous with increasing training steps and has controlability issues. Refinement of prompts, hybrid data blending, domain-aware constraints or another approaches may lead to further gains.

Model Capacity.

The models in this work used ResNet18 for its simplicity and efficiency. Future work could explore deeper architectures or transformers to better capture spatial context and large-scale cropmark structures.

Evaluation Dataset Design.

The test set included deliberately ambiguous cases to challenge generalization. While this is valuable, it may overrepresent edge cases. A stratified evaluation protocol, combining easy and hard examples, would provide a more balanced performance overview.

Temporal Data and Sequence Modeling.

Many false positives, such as fairy rings or seasonal patterns, change significantly between years. If well-aligned imagery from multiple years were acquired, temporal modeling (e.g., sequence classification or multi-year concatenation) could improve robustness. By leveraging temporal consistency, the model could learn to discount ephemeral or unstable features. It would be valuable to establish a collaboration with the Mapy.cz portal, which could potentially provide aligned and high-quality data to support further research.

Multimodal and Contextual Information.

The study relied solely on RGB aerial imagery. Incorporating other modalities (e.g., multispectral, LiDAR, historical maps) or contextual data (e.g., known settlement regions, soil types) could help disambiguate difficult cases and reduce false detections.

Toward Multi-class or Type-specific Classification

An alternative approach worth exploring is to treat cropmark identification as a multi-class problem, where each morphological type (e.g., circular, rectangular, grids, filled)

is considered a separate class. Another option is to train individual binary classifiers for each type, which could improve precision by reducing confusion with unrelated patterns.

Procedurally generated synthetic data - already organized by type - could serve as a strong base for such models and be further fine-tuned on real examples. While not tested in this thesis, this strategy has clear potential for enhancing interpretability and performance and merits further investigation.

7 Conclusion

In this thesis author explored the application of deep learning techniques to the challenging task of detecting archaeological cropmarks in aerial imagery. By leveraging computer vision and remote sensing approaches, the work aimed to assist archaeologists in the labor-intensive process of identifying potential archaeological sites across large geographical areas.

The research began with a comprehensive analysis of the domain, including the theoretical foundations of remote sensing in archaeology and the characteristics of cropmark patterns that indicate subsurface archaeological features. Building upon this understanding, a novel dataset was constructed by combining existing annotations with newly identified cropmark locations discovered through manual exploration of orthophotos.

To address the inherent data scarcity in this specialized domain, several innovative approaches to synthetic data generation were developed and evaluated. Procedural generation techniques proved particularly effective, creating morphologically diverse cropmark patterns that closely resembled real archaeological features. Neural generation using Stable Diffusion with LoRA fine-tuning produced visually appealing results but yielded less satisfactory performance when used for model pre-training.

The experimental results demonstrated that a two-stage training strategy - pre-training on procedurally generated synthetic data followed by fine-tuning on real cropmark images - significantly outperformed both direct training on real data alone and other synthetic data approaches. The best-performing model achieved an F1 score of 0.673 on the cropmark test set, with a precision of 0.653 and recall of 0.695.

Validation on real-world orthophotos confirmed the practical applicability of the approach, with the model successfully identifying 75% of known archaeological sites across multiple years of imagery. While false positives remain a challenge, the results suggest that temporal aggregation and threshold adjustments could further enhance performance in practical deployment.

Several limitations and future directions were identified. The domain shift between different data sources and the inherent visual similarity between cropmarks and certain natural or agricultural phenomena present ongoing challenges. Future improvements could include incorporating temporal data to distinguish between stable archaeological features and transient patterns, exploring segmentation-based approaches for more precise localization, and integrating aligned information such as LiDAR or multispectral imagery.

In conclusion, this research demonstrates that deep learning, particularly when combined with domain-specific synthetic data generation, offers a promising pathway for enhancing archaeological prospection through remote sensing. By reducing the manual effort required for site detection, such techniques can enable archaeologists to survey larger areas more efficiently, potentially leading to the discovery of previously unknown archaeological sites and contributing to our understanding of cultural heritage.

A Annotation Materials

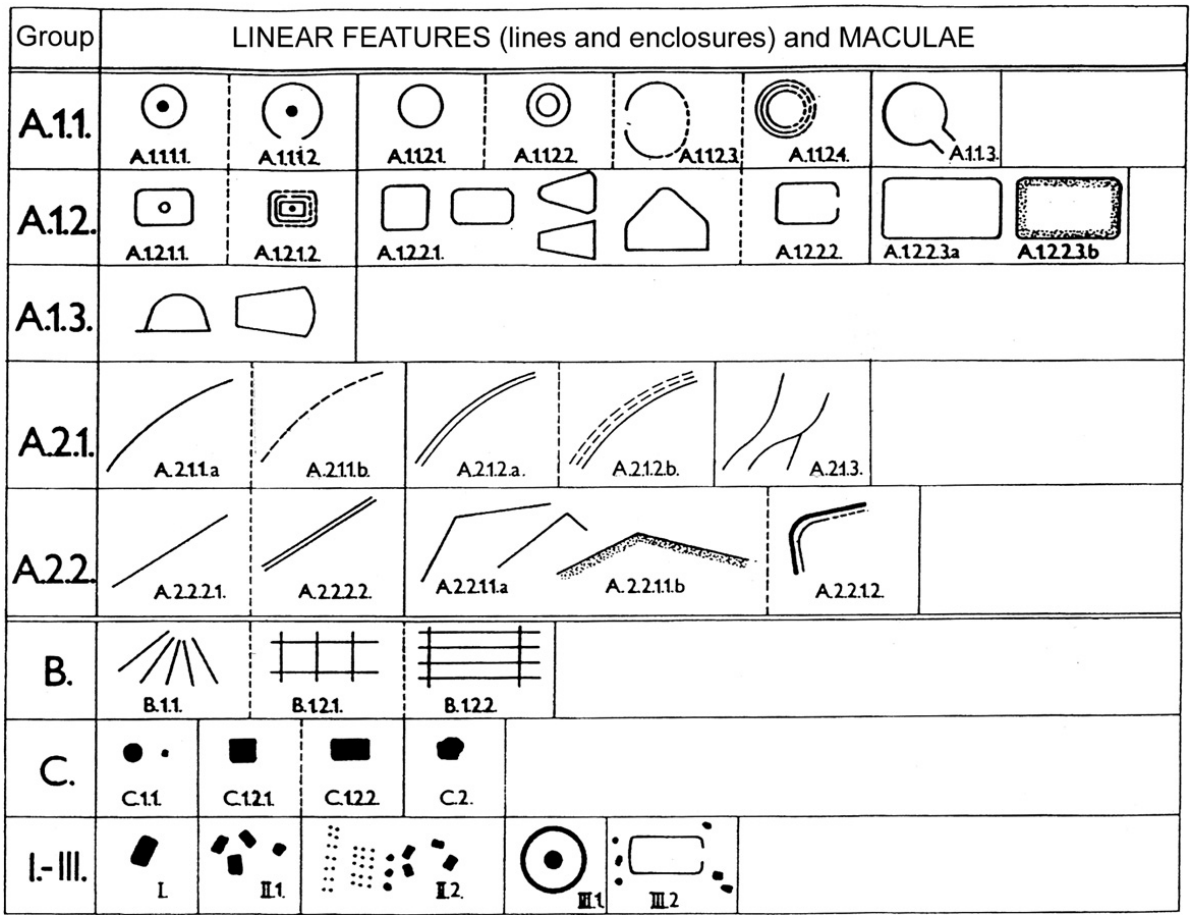

Figure A.1: "The morphology of buried features (in non-equivalent scale) as identified by air survey in Bohemia since 1992" [63][37]. The table relates to the time period 1992–2004.

B Results Visualization and Image Examples

Figure B.1: Examples of cropmarks identified by the author during their own survey. Each cropmark is accompanied by the name of the nearest village, its latitude and longitude coordinates, and an indicator of its relative proximity to a previously annotated location. The label *near* indicates that the cropmark is close to an annotated site but was not directly annotated, whereas *far* denotes a completely distinct location. The author does not claim to be the first to observe or map these cropmarks (although no prior records were found) and does not guarantee their archaeological origin. Image source: Mapy.cz.

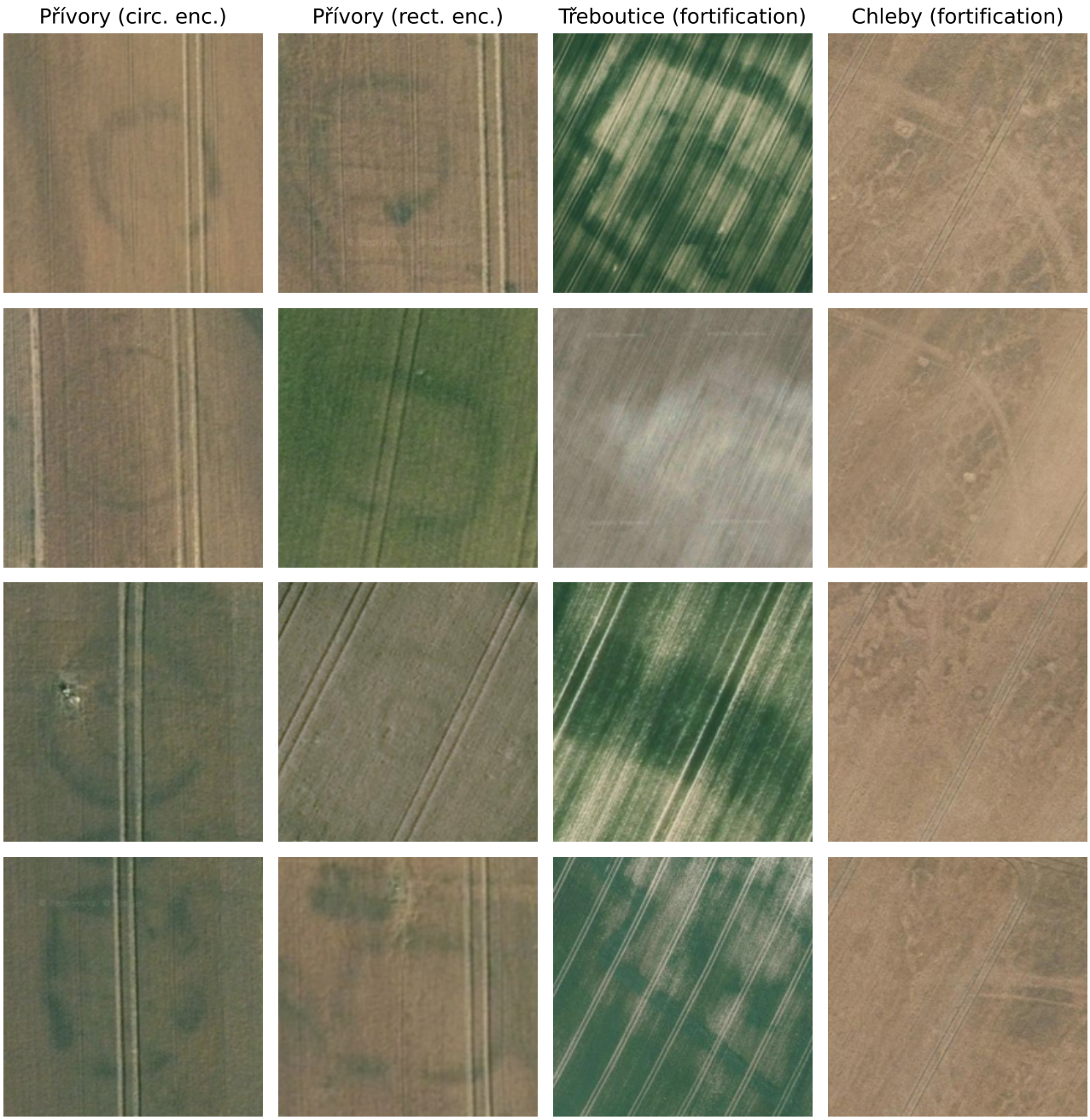

Figure B.2: Representative examples of false negative predictions made by the ResNet18 model trained on the *cropmark train set*. Each column highlights a specific type of missed archaeological feature. The first two columns (Circular and Rectangular Enclosures – Přívory) show enclosures the model failed to detect. The third column illustrates fortified structures near Terezín, while the fourth shows a large circular fortification near Chleby, highlighting the model's difficulties with extensive patterns. The examples demonstrate that the model struggles with large-scale features and cropmarks on brownish backgrounds typical of mature crops, emphasizing the need for further dataset refinement and model adaptation.

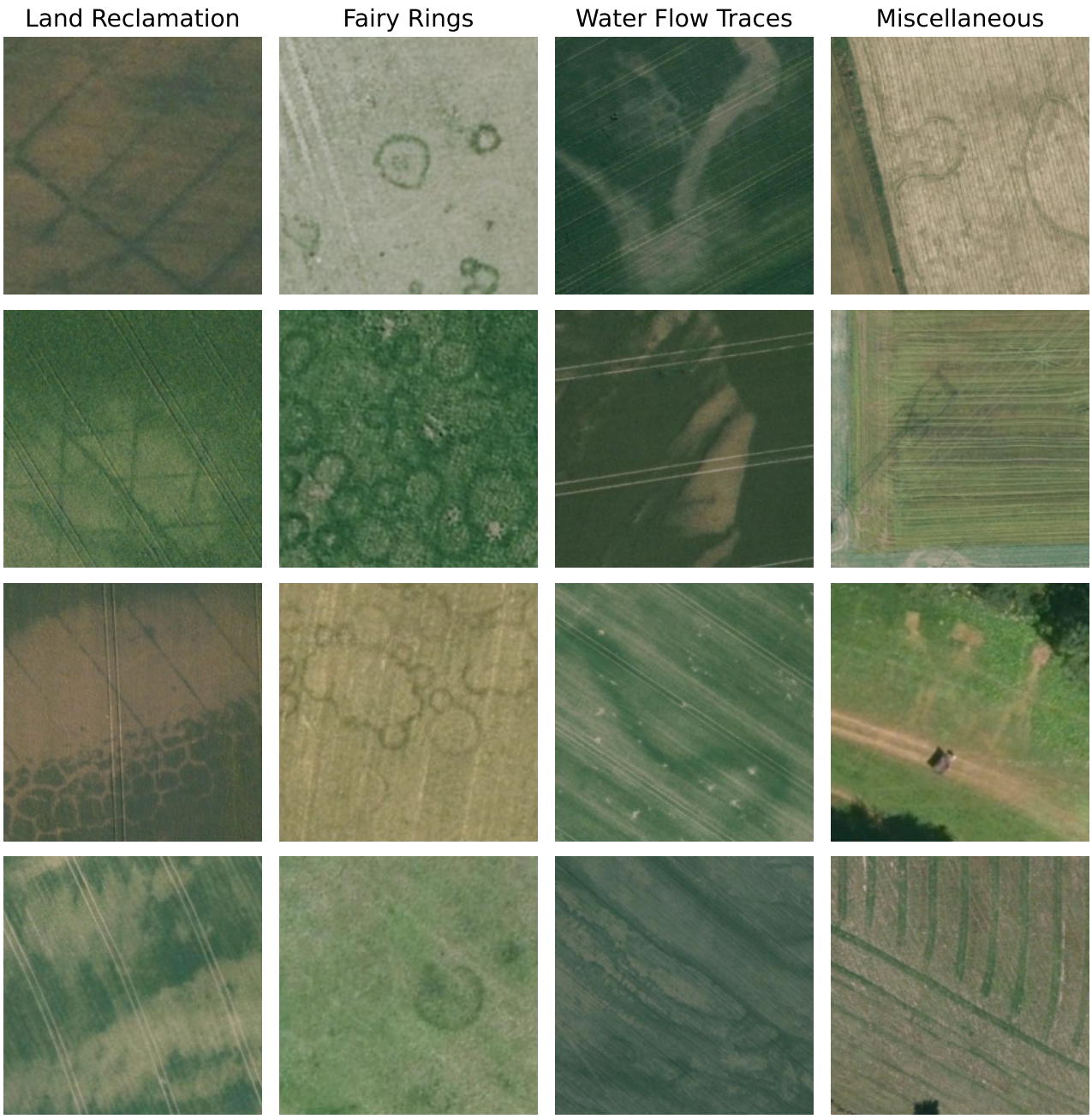

Figure B.3: Representative examples of false positive predictions produced by the ResNet18 model trained on the *cropmark train set*. The figure is organized into four thematic columns based on the visual characteristics of the errors. The first column (Land Reclamation) shows grid-like patterns from agricultural interventions. The second (Fairy Rings) presents natural fungal growths. The third (Water Flow Traces) includes surface patterns from water movement. The final column (Miscellaneous) features (from top to bottom) tractor tracks, a utility pole shadow, remnants of a scout camp, and an atypical field pattern. These examples illustrate the model's difficulties in distinguishing archaeological features from visually similar non-archaeological phenomena.

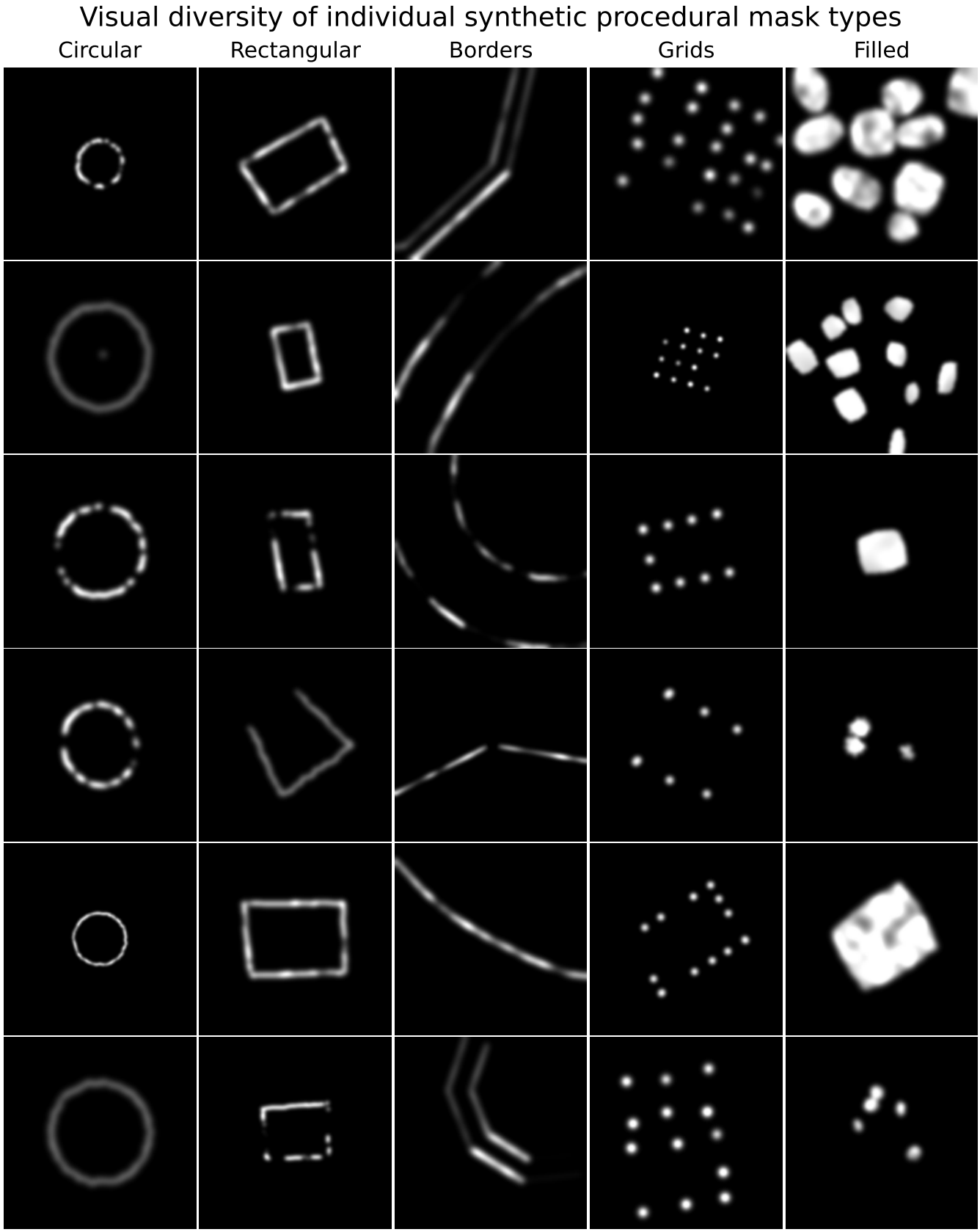

Figure B.4: Examples of randomly generated masks that are subsequently inserted (at varying intensities and polarity) into real aerial images. This process transforms negative samples into synthetic positive samples for the purpose of classifier training. The masks exhibit irregular fading effects, blurring, and distortion, all of which are applied to increase the realism and variability of the resulting synthetic samples.

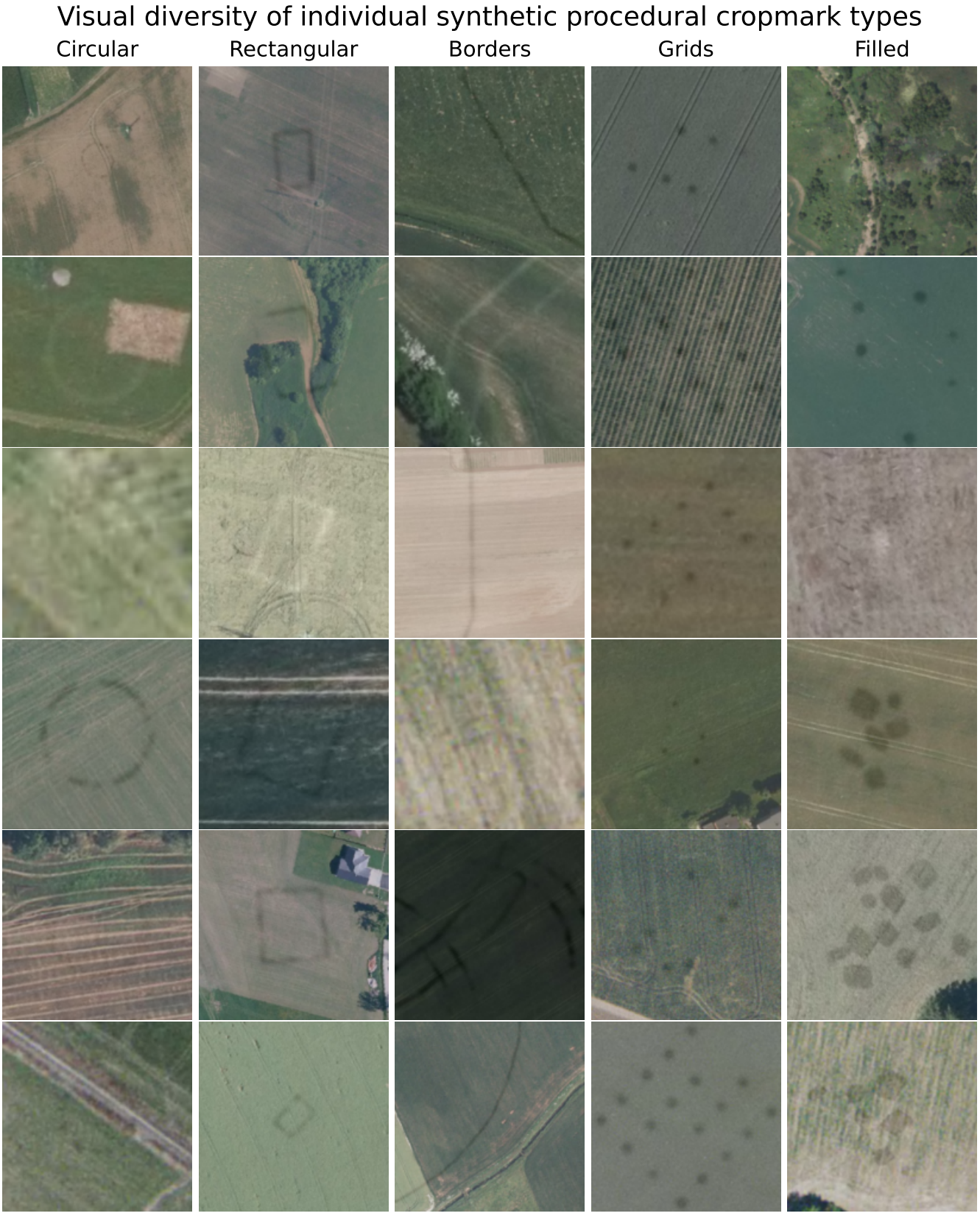

Figure B.5: The figure shows random examples of procedurally generated synthetic cropmarks of five types. Some samples do not fully correspond to real-world cropmark conditions, for instance, they appear in unsuitable environments such as forested areas or roadways. In some cases, particularly with the "Filled" and "Grids" type, cropmarks are less discernible due to environmental noise or the combination of random parameters. These occurrences are minor; importantly, even when visually indistinct, the synthetic marks are still present and can potentially be recognized by a neural network. The "Filled" type corresponds to the "pits, patches" category discussed in the data analysis section.

Examples of images with multiple generated synthetic shapes (masks)

Figure B.6: The figure shows selected examples from the synthetic procedural training set, featuring images with two or three applied masks. While this method of generation can occasionally resemble real-world scenarios, a large portion of such images does not accurately reflect reality (e. g., mixing positive and negative cropmark essence). These composite patterns may (like an augmentation technique) enhance the model's generalization ability, especially in cases involving geometrically complex structures that are not explicitly considered during synthetic generation due to their atypical nature.

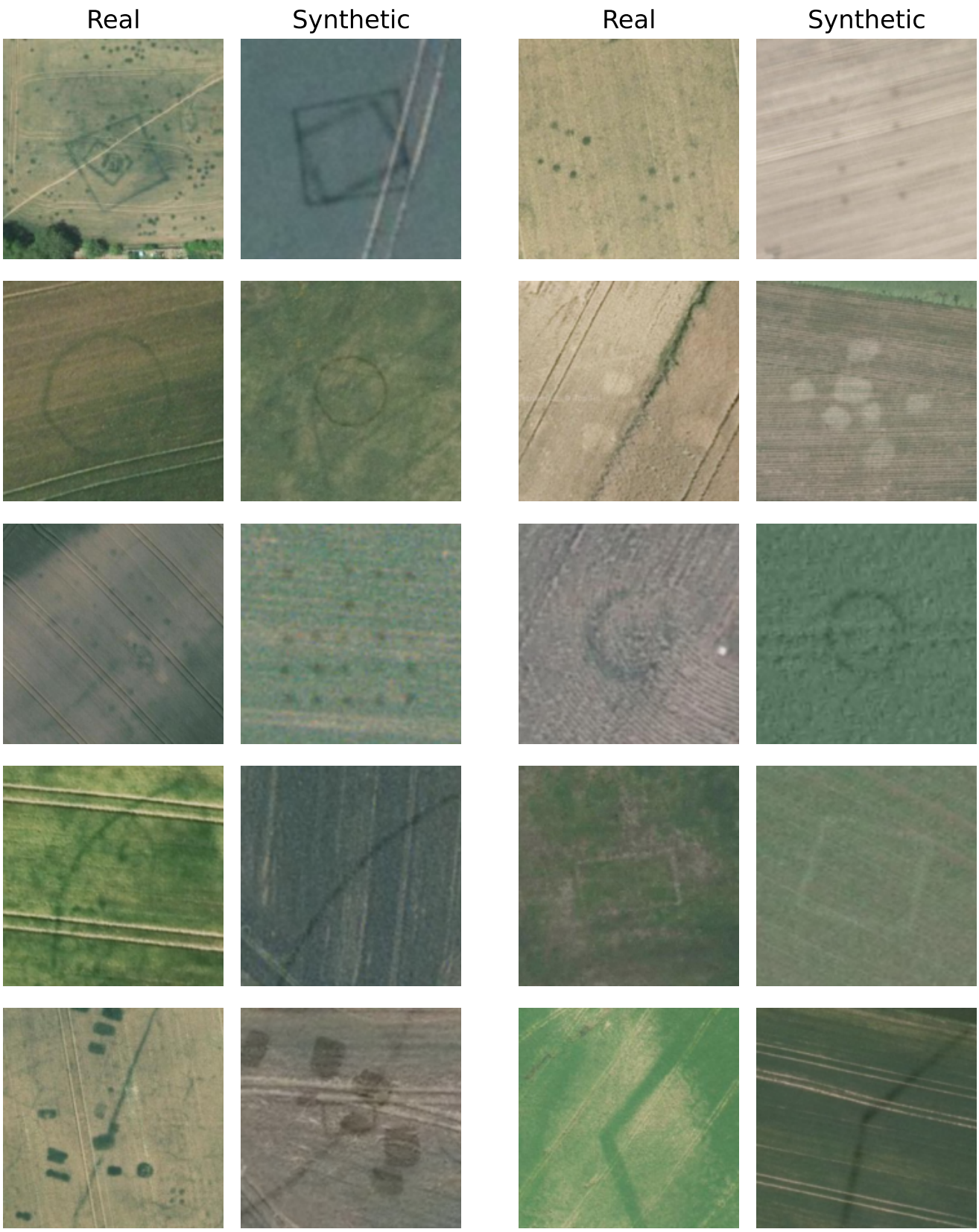

Figure B.7: Comparison of 10 real cropmark examples with their synthetic counterparts from the training and validation portions of the synthetic dataset. Some synthetic structures closely resemble real-world patterns. A model pre-trained on this synthetic data and subsequently fine-tuned on real data generally outperformed a model trained solely on real samples.

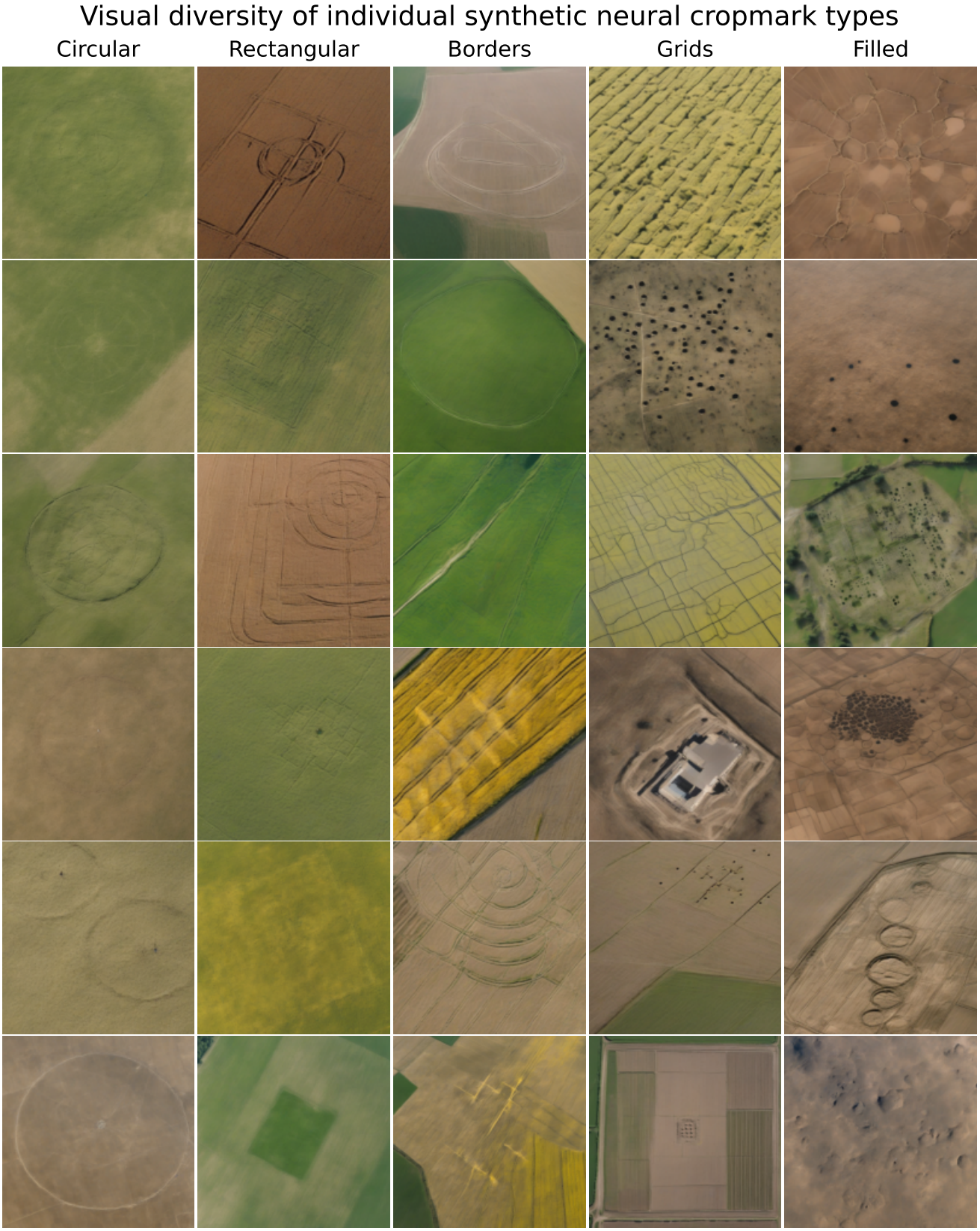

Figure B.8: Examples of synthetic data generated using Stable Diffusion XL 1.0 fine-tuned via LoRA on the *extended cropmark train set*. Each column represents one type of positive class. While some cropmarks appear hyperrealistic and closely resemble real-world samples, others exhibit unnatural characteristics. The models also tend to favor circular structures, while struggling to accurately replicate more complex types such as *grids* and *filled* (pits, patches). The model tends to generate images from a greater distance and often produces oblique aerial views.

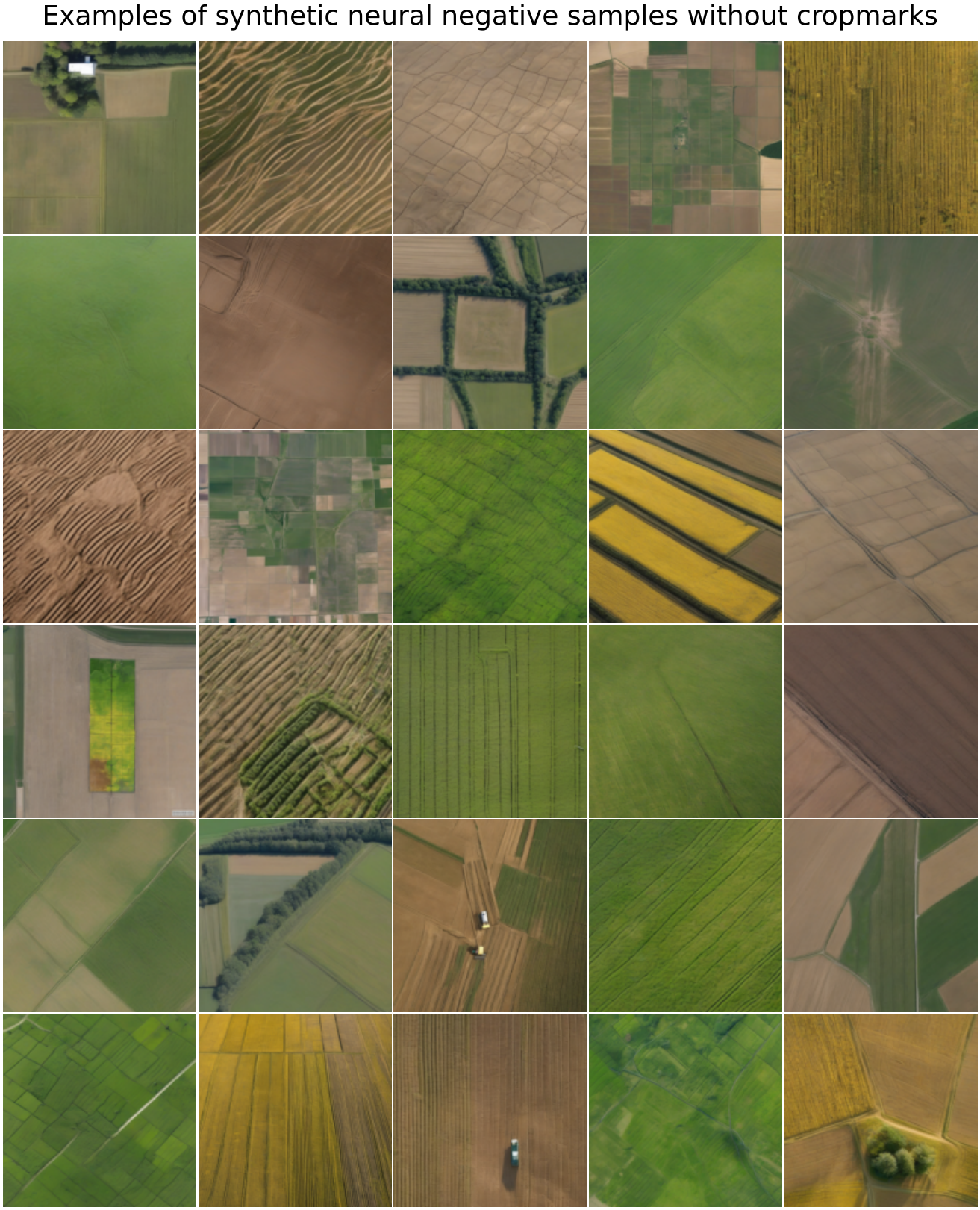

Figure B.9: Examples of synthetically generated images of the negative class from the synthetic training dataset—normal fields produced by Stable Diffusion XL fine-tuned via LoRA on a small subset of the *extended cropmark train set*. The model tends to generate images from a greater distance and often produces oblique aerial views, rather than strictly vertical perspectives.

Bibliography

[1] C. Renfrew and P. Bahn, *Archaeology: Theories, Methods and Practice*, 4th ed. London: Thames & Hudson, 2004.

[2] W. D. Lipe, "Public benefits of archaeological research," in *Public Benefits of Archaeology*. University Press of Florida, 2002, pp. 20 – 28.

[3] A. W. Paskey and A. B. Cisneros, *Digging into Archaeology: A Brief OER Introduction to Archaeology with Activities*. Academic Senate for California Community Colleges, 2020, open educational resource, licensed under Attribution-NonCommercial. [Online]. Available: https://mds.marshall.edu/oa-textbooks/822

[4] M. Gojda, "Dálkový průzkum a jeho proměny v oblasti detekce a mapování archeologického dědictví," *Studia archaeologica Brunensia*, pp. 5–28, 01 2021.

[5] M. Gojda, D. Novák, M. Kuna, P. Vavřín, and J. Bíšková, *Metodika zpracování a evidence dat leteckého průzkumu v archeologii*. Praha: Archeologický ústav AV ČR, 2022.

[6] R. H. Bewley, "Aerial survey for archaeology," *The Photogrammetric Record*, vol. 18, no. 104, pp. 273–292, 2003. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1046/j.0031-868X.2003.00023.x

[7] G. Van Rossum and F. L. Drake Jr, *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam, 1995.

[8] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," 2019. [Online]. Available: https://arxiv.org/abs/1912.01703

[9] TorchVision maintainers and contributors, "Torchvision: Pytorch's computer vision library," 2016, accessed: 2025-05-13. [Online]. Available: https://github.com/pytorch/vision

[10] N. S. Detlefsen, J. Borovec, J. Schock, A. Harsh, T. Koker, L. D. Liello, D. Stancl, C. Quan, M. Grechkin, and W. Falcon, "Torchmetrics - measuring reproducibility in pytorch," *Journal of Open Source Software*, vol. 7, no. 70, p. 4101, 2022, https://github.com/Lightning-AI/torchmetrics. [Online]. Available: https://www.pytorchlightning.ai

[11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[12] Itseez, "Open source computer vision library," https://github.com/itseez/opencv, 2015.

[13] P. von Platen, S. Patil, A. Lozhkov, P. Cuenca, N. Lambert, K. Rasul, M. Davaadorj, D. Nair, S. Paul, S. Liu, W. Berman, Y. Xu, and T. Wolf, "Diffusers: State-of-the-art diffusion models," 2022. [Online]. Available: https://github.com/huggingface/diffusers

[14] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: Fast and flexible image augmentations," *Information*, vol. 11, no. 2, 2020. [Online]. Available: https://www.mdpi.com/2078-2489/11/2/125

[15] G. Research, "Colaboratory: Google's free jupyter notebook environment," https://colab.research.google.com/, 2023, accessed: 2025-05-16.

[16] L. Biewald, "Experiment tracking with weights and biases," 2020, software available from wandb.com. [Online]. Available: https://www.wandb.com/

[17] C. da Costa-Luis, S. K. Larroque, K. Altendorf, H. Mary, richardsheridan, M. Korobov, N. Yorav-Raphael, I. Ivanov, M. Bargull, N. Rodrigues, Shawn, M. Dektyarev, M. Górny, mjstevens777, M. D. Pagel, M. Zugnoni, JC, CrazyPython, C. Newey, A. Lee, pgajdos, Todd, S. Malmgren, redbug312, O. Desh, N. Nechaev, M. Boyle, M. Nordlund, MapleCCC, and J. McCracken, "tqdm: A fast, extensible progress bar for python and cli," Nov. 2024. [Online]. Available: https://doi.org/10.5281/zenodo.14231923

[18] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007.

[19] OpenAI, "Chatgpt (may 13 version)," https://chat.openai.com, 2024, accessed: 2025-05-13.

[20] Claude AI, "Claude AI Natural Language Generation Service," https://www.anthropic.com, February 2024, anthropic.

[21] S. Chacon and B. Straub, *Pro git*. Apress, 2014.

[22] github, "Github," 2020. [Online]. Available: https://github.com/

[23] S. Khorram, F. H. Koch, C. F. Van der Wiele, and S. A. Nelson, *Remote sensing*. Springer Science & Business Media, 2012.

[24] D. M. Hanumant Singh, Jonathan Adams and B. Foley, "Imaging underwater for archaeology," *Journal of Field Archaeology*, vol. 27, no. 3, pp. 319–328, 2000. [Online]. Available: https://doi.org/10.1179/jfa.2000.27.3.319

[25] A. Argyrou and A. Agapiou, "A review of artificial intelligence and remote sensing for archaeological research," *Remote Sensing*, vol. 14, no. 23, 2022. [Online]. Available: https://www.mdpi.com/2072-4292/14/23/6000

[26] D. A. Quattrochi, S. J. Walsh, J. R. Jensen, and M. K. Ridd, "Remote sensing and its relationship to geography," *Geography in America at the Dawn of the 21st Century*, p. 376, 2003.

[27] N. Pettorelli, W. F. Laurance, T. G. O'Brien, M. Wegmann, H. Nagendra, and W. Turner, "Satellite remote sensing for applied ecologists: opportunities and challenges," *Journal of Applied Ecology*, vol. 51, no. 4, pp. 839–848, 2014.

[28] R. A. Rose, D. Byler, J. R. Eastman, E. Fleishman, G. Geller, S. Goetz, L. Guild, H. Hamilton, M. Hansen, R. Headley *et al.*, "Ten ways remote sensing can contribute to conservation," *Conservation Biology*, vol. 29, no. 2, pp. 350–359, 2015.

[29] E. Bedini, "The use of hyperspectral remote sensing for mineral exploration: A review," *Journal of Hyperspectral Remote Sensing*, vol. 7, no. 4, pp. 189–211, 2017.

[30] R. P. Sishodia, R. L. Ray, and S. K. Singh, "Applications of remote sensing in precision agriculture: A review," *Remote sensing*, vol. 12, no. 19, p. 3136, 2020.

[31] S. Majumdar, "The role of remote sensing and gis in military strategy to prevent terror attacks," *Intelligent Data Analytics for Terror Threat Prediction: Architectures, Methodologies, Techniques and Applications*, pp. 79–94, 2021.

[32] M. Gojda, J. John, R. Brejcha *et al.*, *Archeologie a letecké laserové skenování krajiny: Archaeology and Airborne Laser Scanning of the Landscape*, 1st ed. Plzeň: Katedra archeologie, Západočeská univerzita v Plzni, 2013, obsahuje bibliografii, ilustrováno (některé obrázky barevné), mapy.

[33] S. G. Yel and E. Tunc Gormus, "Exploiting hyperspectral and multispectral images in the detection of tree species: A review," *Frontiers in Remote Sensing*, vol. 4, 2023. [Online]. Available: https://www.frontiersin.org/journals/remote-sensing/articles/10.3389/frsen.2023.1136289

[34] M. P. Ferreira, M. Zortea, D. C. Zanotta, Y. E. Shimabukuro, and C. R. de Souza Filho, "Mapping tree species in tropical seasonal semi-deciduous forests with hyperspectral and multispectral data," *Remote Sensing of Environment*, vol. 179, pp. 66–78, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0034425716301134

[35] P. Bahn, "1introduction," in *Archaeology: A Very Short Introduction*. Oxford University Press, 08 2012. [Online]. Available: https://doi.org/10.1093/actrade/9780199657438.003.0011

[36] Z. Czajlik, M. Árvai, J. Mészáros, B. Nagy, L. Rupnik, and L. Pásztor, "Cropmarks in aerial archaeology: New lessons from an old story," *Remote Sensing*, vol. 13, no. 6, 2021. [Online]. Available: https://www.mdpi.com/2072-4292/13/6/1126

[37] M. Gojda and M. Hejcman, "Cropmarks in main field crops enable the identification of a wide spectrum of buried features on archaeological sites in central europe," *Journal of Archaeological Science*, vol. 39, no. 6, pp. 1655–1664, 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0305440312000350

[38] S. A. Aqdus, W. S. Hanson, and J. Drummond, "The potential of hyperspectral and multi-spectral imagery to enhance archaeological cropmark detection: a comparative study," *Journal of Archaeological Science*, vol. 39, no. 7, pp. 1915–1924, 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0305440312000465

[39] R. Jackson, "Fairy rings," *Canadian Medical Association Journal*, vol. 105, no. 7, pp. 703–704, October 9 1971.

[40] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., vol. 25. Curran Associates, Inc., 2012. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

[41] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proceedings of the IEEE*, vol. 105, no. 10, pp. 1865–1883, 2017.

[42] J. Wang, Z. Zheng, A. Ma, X. Lu, and Y. Zhong, "Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation," in *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, J. Vanschoren and S. Yeung, Eds., vol. 1, 2021. [Online]. Available: https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/4e732ced3463d06de0ca9a15b6153677-Paper-round2.pdf

[43] Z. Zhang, L. Zhang, Y. Wang, P. Feng, and R. He, "Shiprsimagenet: A large-scale fine-grained dataset for ship detection in high-resolution optical remote sensing images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 8458–8472, 2021.

[44] J. Zhu and T. Kelly, "Seamless satellite-image synthesis," 2021. [Online]. Available: https://arxiv.org/abs/2111.03384

[45] Z. L. Qiqi Zhu, Xi Guo and D. Li, "A review of multi-class change detection for satellite remote sensing imagery," *Geo-spatial Information Science*, vol. 27, no. 1, pp. 1–15, 2024. [Online]. Available: https://doi.org/10.1080/10095020.2022.2128902

[46] S. Demir, M. Dedeoğlu, and L. Başayiğit, "Yield prediction models of organic oil rose farming with agricultural unmanned aerial vehicles (uavs) images and machine learnaing algorithms," *Remote Sensing Applications: Society and Environment*, vol. 33, p. 101131, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352938523002136

[47] S. Vats and S. Mehta, "Revolutionizing archaeological discoveries: The role of artificial intelligence and machine learning in site analysis," in *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 2024, pp. 1–5.

[48] G.-S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, and X. Lu, "Aid: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 7, p. 3965–3981, Jul. 2017. [Online]. Available: http://dx.doi.org/10.1109/TGRS.2017.2685945

[49] Q. Zou, L. Ni, T. Zhang, and Q. Wang, "Deep learning based feature selection for remote sensing scene classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 11, pp. 2321–2325, 2015.

[50] D. Trier, S. Larsen, and R. Solberg, "Automatic detection of circular structures in high-resolution satellite images of agricultural land," *Archaeological Prospection*, vol. 16, no. 1, pp. 1–15, 2009. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/arp.339

[51] D. Trier and L. H. Pilø, "Automatic detection of pit structures in airborne laser scanning data," *Archaeological Prospection*, vol. 19, no. 2, pp. 103–121, 2012. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/arp.1421

[52] J. M. Lemmens, R. Verwaal *et al.*, "Automated archaeological feature extraction from digital aerial photographs," 1993.

[53] A. Agapiou and E. Gravanis, "A machine-learning-assisted classification algorithm for the detection of archaeological proxies (cropmarks) based on reflectance signatures," *Remote Sensing*, vol. 16, no. 10, 2024. [Online]. Available: https://www.mdpi.com/2072-4292/16/10/1705

[54] M. Bellat and T. Scholten, "Automated features detection in archaeology: Standardisation in the area of big data," in *CAA51st Across the Horizon.* Auckland, New Zealand: Computer Applications and Quantitative Methods in Archaeology (CAA)

and University of Auckland, Apr. 2024. [Online]. Available: https://hal.science/hal-04612232

[55] A. Karamitrou, F. Sturt, P. Bogiatzis, and D. Beresford-Jones, "Towards the use of artificial intelligence deep learning networks for detection of archaeological sites," *Surface Topography: Metrology and Properties*, vol. 10, no. 4, p. 044001, oct 2022. [Online]. Available: https://dx.doi.org/10.1088/2051-672X/ac9492

[56] K. Phelan and D. Riordan, "Detection of ringforts from aerial photography using machine learning," in *2020 31st Irish Signals and Systems Conference (ISSC)*, 2020, pp. 1–6.

[57] A. J. Paul, S. Ghose, K. Aggarwal, N. Nethaji, S. Pal, and A. Dutta Purkayastha, "Machine learning advances aiding recognition and classification of indian monuments and landmarks," in *2021 IEEE 8th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, 2021, pp. 1–8.

[58] F. Materazzi and M. Pacifici, "Archaeological crop marks detection through drone multispectral remote sensing and vegetation indices: A new approach tested on the italian pre-roman city of veii," *Journal of Archaeological Science: Reports*, vol. 41, p. 103235, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352409X21004478

[59] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *arXiv preprint arXiv:1505.04597*, 2015, conditionally accepted at MICCAI 2015. [Online]. Available: https://doi.org/10.48550/arXiv.1505.04597

[60] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988.

[61] M. Bundzel, M. Jaščur, M. Kováč, T. Lieskovský, P. Sinčák, and T. Tkáčik, "Semantic segmentation of airborne lidar data in maya archaeology," *Remote Sensing*, vol. 12, no. 22, 2020. [Online]. Available: https://www.mdpi.com/2072-4292/12/22/3685

[62] I. Berganzo-Besga, H. A. Orengo, F. Lumbreras, M. Carrero-Pazos, J. Fonte, and B. Vilas-Estévez, "Hybrid msrm-based deep learning and multitemporal sentinel 2-based machine learning algorithm detects near 10k archaeological tumuli in north-western iberia," *Remote Sensing*, vol. 13, no. 20, 2021. [Online]. Available: https://www.mdpi.com/2072-4292/13/20/4181

[63] M. Gojda, Ed., *Ancient Landscape, Settlement Dynamics and Non-Destructive Archaeology: Czech Research Project 1997-2002 = Dávnověká krajina a sídla ve světle nedestruktivní archeologie : český výzkumný projekt 1997-2002.* Praha: Academia, 2004.

[64] S. Lloyd, "Least squares quantization in pcm," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[65] Royal Commission on the Ancient and Historical Monuments of Wales, "Coflein: Online database for the national monuments record of wales," https://coflein.gov.uk/, 2025, accessed: 2025-05-09.

[66] X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang, "Moment matching for multi-source domain adaptation," 2019. [Online]. Available: https://arxiv.org/abs/1812.01754

[67] Mapbox, "Mapbox," https://www.mapbox.com/, n.d., accessed: 2025-04-24.

[68] X. Qi, P. Zhu, Y. Wang, L. Zhang, J. Peng, M. Wu, J. Chen, X. Zhao, N. Zang, and P. T. Mathiopoulos, "Mlrsnet: A multi-label high spatial resolution remote sensing dataset for semantic scene understanding," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 169, pp. 337–350, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271620302677

[69] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft coco: Common objects in context," 2015. [Online]. Available: https://arxiv.org/abs/1405.0312

[70] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov, T. Duerig, and V. Ferrari, "The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale," *International Journal of Computer Vision*, vol. 128, no. 7, p. 1956–1981, Mar. 2020. [Online]. Available: http://dx.doi.org/10.1007/s11263-020-01316-z

[71] Y. Hua, L. Mou, P. Jin, and X. X. Zhu, "Multiscene: A large-scale dataset and benchmark for multiscene recognition in single aerial images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, p. 1–13, 2022. [Online]. Available: http://dx.doi.org/10.1109/TGRS.2021.3110314

[72] A. Boguszewski, D. Batorski, N. Ziemba-Jankowska, T. Dziedzic, and A. Zambrzycka, "Landcover.ai: Dataset for automatic mapping of buildings, woodlands, water and roads from aerial imagery," 2022. [Online]. Available: https://arxiv.org/abs/2005.02264

[73] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proceedings of the IEEE*, vol. 105, no. 10, pp. 1865–1883, 2017.

[74] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM GIS)*. ACM, 2010.

[75] W. Zhou, S. Newsam, C. Li, and Z. Shao, "Patternnet: A benchmark dataset for performance evaluation of remote sensing image retrieval," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 145, p. 197–209, Nov. 2018. [Online]. Available: http://dx.doi.org/10.1016/j.isprsjprs.2018.01.004

[76] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015. [Online]. Available: https://arxiv.org/abs/1512.03385

[77] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[78] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," 2019. [Online]. Available: https://arxiv.org/abs/1801.04381

[79] H. E. Robbins, "A stochastic approximation method," *Annals of Mathematical Statistics*, vol. 22, pp. 400–407, 1951. [Online]. Available: https://api.semanticscholar.org/CorpusID:16945044

[80] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017. [Online]. Available: https://arxiv.org/abs/1412.6980

[81] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2019. [Online]. Available: https://arxiv.org/abs/1711.05101

[82] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," 2022. [Online]. Available: https://arxiv.org/abs/2112.10752

[83] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," 2023. [Online]. Available: https://arxiv.org/abs/2302.05543

[84] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, "Sdxl: Improving latent diffusion models for high-resolution image synthesis," 2023. [Online]. Available: https://arxiv.org/abs/2307.01952

[85] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," 2021. [Online]. Available: https://arxiv.org/abs/2106.09685