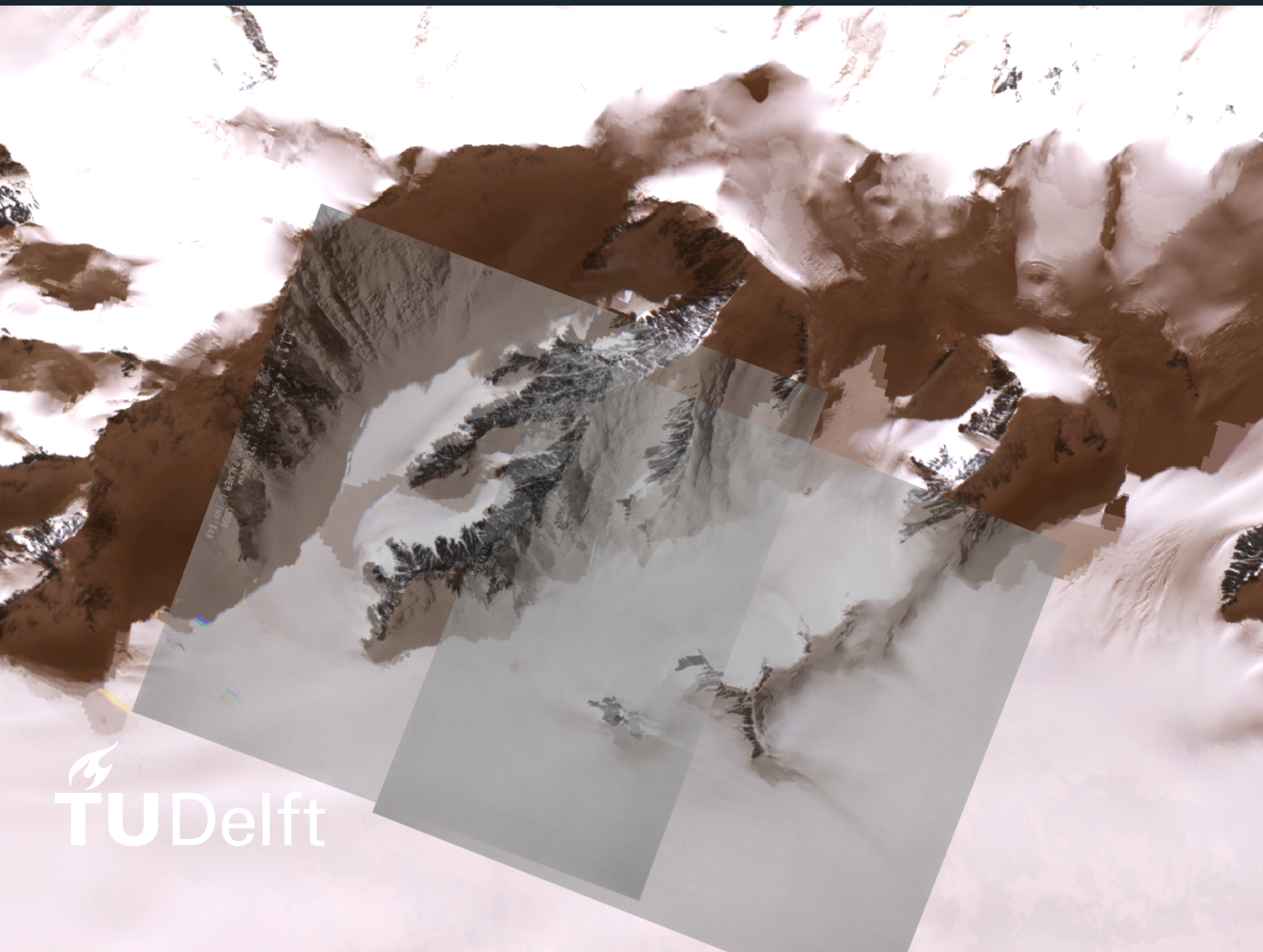
Deep learning-based Image similarity estimation for geo-localization of Historical Aerial imagery

Yushan Liu



TU Delft

Deep learning-based Image similarity estimation for geo-localization of Historical Aerial imagery

Master Thesis

by

Yushan Liu

Thesis committee:

F. (Felix) Dahle	TU Delft
Dr. R.C. (Roderik) Lindenbergh	TU Delft
Dr. Ir. B. (Bert) Wouters	TU Delft
Place:	Faculty of Civil Engineering and Geosciences, Delft
Project Duration:	April, 2023 - February, 2024
Student number:	5525829

An electronic version of this thesis is available at http://repository.tudelft.nl/.

Faculty Of Civil Engineering and Geosciences · Delft University of Technology

Abstract

Historical aerial imagery serves as a valuable data source for observing Antarctica, facilitating an extended temporal scale of observation and enabling comparisons to deepen understanding of glacier dynamics. However, many historical aerial datasets, including the Antarctica Single Frames dataset utilized in this study, lack geo-referencing and orientation metadata essential for spatial analysis. One method of geo-referencing these historical images involves image matching to establish Ground Control Points (GCPs). This study focuses on the prerequisite for image matching: ensuring alignment between unreferenced historical images and already geo-referenced images in terms of scene and approximate resolution, a process termed 'geo-localization' herein.

Geo-localization is achieved by comparing the historical image with positions within a predefined geo-referenced Area of Interest (AoI). Two predefined remote sensing datasets are used: Sentinel-2 and Quantarctica Rock Outcrop Mask, from which AoIs are generated. Positions within the AoI exhibiting the highest similarity to the historical image are likely to correspond to the same ground area, thus providing the location of the historical imagery.

This similarity assessment employs two Siamese Networks: SigNet and ResNet-50. SigNet, originally designed for signature verification tasks, consists of four convolutional layers. In contrast, ResNet-50, initially developed for image classification purposes, is characterized by its deep architecture comprising approximately 50 convolutional layers, as suggested by its name. In this study, these two models are initially pre-trained on cross-domain datasets and subsequently adaptively trained with task-specific datasets created in this study. The adaptive training datasets comprise triplets of similar and dissimilar images pre-processed using methods devised in this study. An evaluation methodology based on confidence level is developed to assess the model and workflow performance, which is then applied to 51 test historical image samples.

Overall, the results indicate that the ResNet-50 based network outperforms SigNet, achieving a 95.5% average confidence level. However, the method does not meet the initial expectation of directly providing the location of the historical image within the AoI. Instead, it identifies potential locations. Nevertheless, this outcome is valuable as it streamlines the search process for subsequent image matching steps. For instance, a 95.5% average confidence level for the ResNet-50 based network correlates with an approximate 95.5% reduction in processing time for geo-referencing when integrated with image matching in subsequent steps.

Contents

List of Figures

List of Tables

1 Introduction

Antarctica's vast icy landscapes have intrigued scientists, explorers, and the public for generations. Monitoring Antarctica is crucial because it plays a unique role in global climate systems and environmental dynamics. Despite its remoteness, studies about Antarctica is crucial for understanding climate change, sea-level rise, biodiversity, and global environmental dynamics.

Over the past few decades, there has been an increasing number of glacier monitoring programs in Antarctica. Notable examples include the Antarctic Seismic Data Library System (SDLS) [1] and Operation IceBridge [2]. However, the programs and the models only extend from the early 1960s to present, which covers the extent of the modern satellite era [3]. These programs and models give a better understanding of different types of changes in Antarctica. In this study, the aim is to expand the temporal scale of Antarctica observation by geo-localizing historical imagery known as Antarctica Single Frames. Once the locations of these historical images are determined, they can be geo-referenced and compared with modern satellite imagery for comparison.

1.1. Problem description

While historical aerial imagery holds great potential for extending the temporal scale of Antarctica's observations, the absence of orientation metadata often presents challenges in accurately geo-referencing these historical images. The dataset utilized in this study is known as Antarctic Single Frames (1946-2000), comprising historical aerial imagery captured using the TMA photography camera system. Therefore, in this study, these historical images are referred to as TMA images. A more detailed introduction to this dataset can be found in chapter 3. In the Antarctic Single Frames dataset, accompanying shapefiles contain supplemental information, including approximate camera height and camera center details.

In a previous study [4], we were able to compute 'rough' footprints for the historical aerial imagery. This is possible thanks to the fact that altimeters indicate the height of the camera on the images. The height was extracted by reading the altimeter with Computer Vision (CV) methods and combined with the approximate values of camera center coordinates obtained from additional shapefiles in Antarctica Single Frames. However, it is important to note that these footprints are only a rough indication of the real location. In most cases, they cover a much larger area than the actual footprint, providing only a general area of interest. In Fig. 1.1, the middle image depicts the rough footprint with a 10 km buffer added in all directions, while the top-right image displays the actual footprint. The size of the 'rough' footprints can vary from 5 km $\times$5 km to 30 km $\times$30 km or more, depending on the height of the camera. To allow for some errors in the 'rough' footprints, a buffer of 10 km is added in each direction, resulting in a final area of interest ranging from 25 km $\times$25 km to 50 km $\times$50 km or even larger when the height reading is larger than usual. In contrast, the size of the real footprints typically ranges from 3 km $\times$3 km to 10 km $\times$10 km, which is significantly smaller than the area of interest.

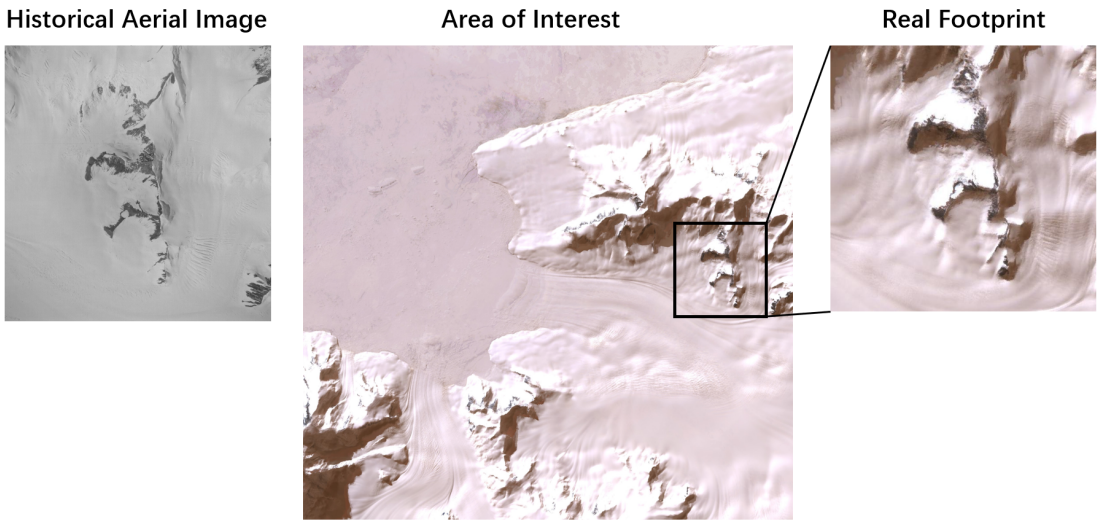


Figure 1.1: Area of interest (AoI) and real footprint of a historical aerial image.

Before utilizing the historical imagery, geo-referencing is a prerequisite to establish their spatial context. Adequate and appropriately distributed Ground Control Points (GCPs) are crucial for this process, yet they are currently absent in the historical images. The creation of GCPs can be supported by image matching, a technique that aligns the historical, un-referenced images with already geo-referenced counterparts, such as satellite images. Several efficient image matching methods are available for this purpose [5] [6] [7]. However, successful image matching necessitates that both images – the geo-referenced and the un-referenced – depict the same scene and exhibit roughly similar resolutions. This process is referred to as 'geo-localization' in this study.

Under the conditions where the aforementioned criteria are not met, meaning when the same scene as depicted in the un-referenced imagery cannot be pinpointed within the already geo-referenced Area of Interest (AoI), image matching needs to be applied at every position on the AoI to find the match. Typically, the searching strategy for image matching involves starting from the center of the AoI and then spiraling outward while continuously attempting to find a match. However, such an image matching searching strategy is highly time-consuming and inefficient.

Therefore, the primary objective in this study is to meet the prerequisite for image matching, which is also referred to as 'geo-localization' in this study: ensuring that the un-referenced historical images align with the already geo-referenced images in terms of scene and approximate resolution.

To achieve this alignment, deep learning-based image similarity estimation algorithms are explored. Deep learning is a type of machine learning that employs artificial neural networks to learn from data and make predictions. Using deep learning image similarity algorithms, attempts are made to locate the same ground structure or scene on the AoI for a historical image by comparing the similarity level between the historical image and each position on the AoI image.

1.2. Objectives and Research Questions

1.2.1. Research Objectives

The main objective of this study is to develop an algorithm to find the rough footprints of historical aerial images by matching these historical imagery to modern remote sensing datasets.

Specifically, the objectives are as follows:

- Develop suitable architectures for the deep neural network of image similarity estimation.
- Establish a well-defined and structured workflow for this project.
- Create a suitable training dataset for the proposed SigNet model and ResNet-50 based model.

1.2.2. Research Questions

Considering the aforementioned objectives, the main research question for this project is summarized as: How can deep learning-based image similarity algorithms be used to locate historical aerial imagery on modern remote sensing datasets?

To answer this question, the following sub-questions will be addressed:

1. How to establish a structured workflow for the project.
2. How to create a suitable training dataset?
3. How to assess image similarity using deep neural networks?
4. Based on the model predictions, how to locate the rough footprint?
5. How can the relative positions between historical images on the same flight lines be utilized in the algorithm?
6. How to evaluate the performance of the algorithm?

1.3. Thesis Outline

Chapter 2 of this thesis presents the conceptual framework and definitions related to the topic of the study, including a literature review on geo-referencing. In chapter 3, the three datasets utilized in this study are introduced, and the study area is specified. Chapter 4 details the methodology developed to address the research questions, encompassing the workflow, data pre-processing steps, introduction of the networks, and evaluation methods. Chapter 5 presents the results from both the SigNet and ResNet-50 based networks, along with a discussion and comparison of these results. Finally, chapter 6 concludes the study by addressing the research questions and providing recommendations for future work.

2

Conceptual Framework and Definitions

In this study, the utilization of historical aerial imagery, particularly the Antarctic Single Frames (1946-2000), poses challenges in practical application due to geo-referencing complexities. These challenges primarily stem from the absence of Ground Control Points (GCPs) within the images, exacerbated by the lack of onboard Global Navigation Satellite System (GNSS) technology or incomplete GNSS data on the capturing aerial platforms. Additionally, image quality issues and the intricate landscape further complicate traditional geo-referencing methods reliant on GCPs and direct GNSS data. To navigate these complexities, this study leverages geo-localization and image processing techniques, using image matching to establish surrogate GCPs. This chapter introduces the employed techniques.

2.1. Geo-referencing

Geo-referencing involves aligning geographic data to a known coordinate system to facilitate analysis, query, and visualization alongside other geographic data sources [8]. While various remote sensing (RS) data sources, such as satellite images, maps, point clouds, and aerial data, undergo geo-referencing, the process often is dependent on ground control points (GCPs) [9] [10].

In some applications, GCPs are manually created by visual comparison, for instance, geo-referencing Building Information Models (BIM) in GIS environments [11], and geo-referencing old maps to modern GIS [11]. However, this method is time-consuming and labor-intensive, making it impractical for many remote sensing images that often cover large areas.

In this study, we focus on geo-referencing aerial imagery. The traditional approach of geo-referencing such images relies heavily on existing GCPs [12]. However, aerial imaging often encounters scenarios where there's a scarcity of existing GCPs. Furthermore, manual selection of GCPs is impractical, particularly in inaccessible or hazardous areas, or when visually identifying similar features becomes difficult.

Presently, unmanned aerial vehicles (UAVs) with onboard global navigation satellite system-real-time kinematic (GNSS RTK) receiver are increasingly used. The precise knowledge of the camera's position during image acquisition, suggested with centimeter-level accuracy, has been proposed as a potential replacement for the need of Ground Control Points (GCPs) in geo-referencing photogrammetric models [13]. This approach eliminates the reliance on GCPs, offering advantages in terms of cost-effectiveness and simplification of measurements.

Other methods use a small number of GCPs [14], or even a single one [15] in combination with the camera position provided by GNSS RTK, significantly improving accuracy. Additionally, using oblique images is another promising alternative. Studies such as [16], [17], and [18] have demonstrated that incorporating oblique images can lead to improved accuracy, attributed to their ability to offer enhanced perspective, improved depth perception, and increased coverage.

Additionally, various image-based geo-referencing methods have been proposed in the literature. In [19], an automatic framework is presented for transforming the local coordinates of a Building Information Models (BIM) model into its real-world geographic coordinates. The authors of [20] introduce an autonomous image geo-referencing algorithm based on a template image matching approach, specifically designed for LAPAN-A3/IPB multispectral images. Furthermore, [21] proposes an automatic geometric correction system based on contour-matching techniques capable of geo-referencing satellite images with high

accuracy. In general, the task of image matching-based geo-referencing can be divided into several different methods, intensity-based [22], geo-localization template matching [23], and local feature-based. Among these, the local feature-based method received the most attention in the field of computer vision [24] [25], photogrammetry and remote sensing [26] [27] [28].

Deep learning methods have been explored for feature extraction in image matching. In [29], Where-CNN (Convolutional Neural Network) is proposed to find matches between street view and aerial view imagery. In [30], the comparison of multiple different image sources is achieved by applying a Siamese Network. A Siamese network is a type of neural network that employs two identical sub-networks to process two different inputs and learns to measure similarity or dissimilarity between them. While CNN-based geo-localization has been extensively applied in street-level image matching [31], our case focuses on matching aerial imagery with satellite imagery. Despite sharing similar underlying theories, our dataset, the Antarctica Single Frames, presents unique challenges. The images predominantly feature snow and rock formations, which are relatively irregular and uniform compared to urban landscapes. Furthermore, the Antarctica Single Frames exhibit a temporal disparity of approximately 50 years, with many images captured decades before the satellite imagery utilized in this study. These temporal and environmental differences present challenges in feature recognition and can complicate the matching and georeferencing processes.

2.2. Geo-localization

In computer vision, the task of coarsely estimating the place where a photo was taken based on a set of previously visited locations is called Visual (Image) Geo-localization (VG) or Visual Place Recognition (VPR) and it is addressed using image matching and retrieval methods on a database of images of known locations [32]. Geo-localization is a fundamental task in computer vision that enables the integration of visual information with geospatial data and maps, bridging the gap between visual perception and geospatial analysis.

Traditional image-based geo-localization is normally done in the context where both the query and geo-tagged reference images in the database are taken from the ground view [33]. The ”query” image refers to the image for which the location needs to be determined, while the ”geo-tagged reference” images are images in a database with known geographic coordinates, serving as a reference for location comparison or matching. From the last decade, ground-to-aerial image geo-localization has become a increasingly popular approach [33] [34] [35] [36], where ground view photos are matched to aerial imagery to locate the ground view photos.

2.3. Image Similarity Estimation

Image similarity quantifies the extent of resemblance between two images using a similarity metric or distance function. Its applications span diverse domains, including Content-Based Image Retrieval (CBIR) [37], Image Classification and Categorization [38], Face Recognition [39], and Image Recommendation, among others. The core of image similarity algorithms involves two key steps: feature extraction and the selection of an appropriate similarity metric [40]. These algorithms extract distinct visual features from images and subsequently compare them based on various visual attributes such as color, texture, shape, and other pertinent factors such as patterns, spatial layouts, gradients, edges, structural elements, and even higher-level semantics. The choice of feature extraction method and similarity metric is tailored to the specific application requirements and image characteristics.

Feature extraction is the pivotal process of deriving meaningful and representative visual features from images to assess their similarity. Methods include hand-crafted approaches like Histogram of Oriented Gradients (HOG) [41] and Scale-Invariant Feature Transform (SIFT) [42], which extract low-level or mid-level features capturing specific aspects of image content. These handcrafted features are engineered to represent certain visual characteristics effectively.

A concept similar to feature extraction is image embedding. In this study, we consider image embedding as one of the methods within feature extraction and thus refer to it collectively as feature extraction. Image embedding involves learning a mapping from images to a high-dimensional feature space using deep neural networks, often convolutional neural networks (CNNs). These networks are trained to extract abstract and semantically rich representations directly from the raw image data. The goal of image embedding is to learn representations that encode rich semantic information about the images, making them suitable for

various downstream tasks such as classification, retrieval, and similarity assessment. Unlike handcrafted features in traditional feature extraction methods, image embedding learns representations directly from the data, capturing more abstract and semantic information. [43]

Once features or embeddings are extracted, a distance metric or similarity score gauges the likeness between two images based on their feature representations. The choice of similarity metric is guided by the unique requirements of the task and the characteristics of the data. Commonly employed similarity metrics encompass Euclidean distance [44], Cosine similarity [45], Jaccard similarity [46], and Hamming distance [47].

2.3.1. Deep learning-based Image similarity estimation

Deep learning-based image similarity estimation networks harness the capabilities of deep neural networks to extract meaningful feature representations from raw image data, forming the cornerstone for quantifying image similarity. These networks, encompassing architectures like Siamese Networks and Triplet Networks, operate by leveraging sophisticated convolutional neural networks (CNNs) to automatically discern intricate visual patterns such as edges, textures, and shapes, making them ideal for various image analysis tasks ([40] [48]).

In Siamese Networks, a pair of identical CNNs processes each image to generate feature vectors capturing their visual essence. These vectors are then compared using distance metrics like Euclidean distance or cosine similarity, enabling the computation of a similarity score reflecting image resemblance. The network is trained using a contrastive loss function, which penalizes pairs of dissimilar images that are incorrectly judged to be similar and vice versa. This training process optimizes the network parameters to effectively learn discriminative features for image similarity assessment.

Alternatively, Triplet Networks employ three CNNs: one for the anchor image, one for a positive image resembling the anchor, and one for a negative image dissimilar to the anchor [49]. The objective of the network is to learn feature representations that minimize the distance between the anchor and positive image while maximizing the distance between the anchor and negative image. This framework facilitates the learning of embeddings capable of effectively discerning image similarities and dissimilarities, thereby advancing image analysis tasks. Triplet Networks are trained using a triplet loss function, which encourages the network to push the positive image closer to the anchor while simultaneously pushing the negative image further away.

Furthermore, recent advancements in deep learning-based similarity estimation networks have led to the development of more sophisticated architectures, such as deep metric learning networks [49] and attention-based models [50]. These models enhance the representation learning process by focusing on informative regions of the images and incorporating semantic information to improve the accuracy of similarity estimation. Additionally, the integration of domain-specific knowledge and data augmentation techniques further enhances the robustness and generalization capabilities of these networks, making them applicable to a wide range of image analysis tasks across various domains.

Overall, deep learning-based similarity estimation networks provide a robust and versatile approach for extracting and comparing meaningful visual features, thereby contributing to advancements in various fields reliant on image analysis. Further research in this area holds promise for addressing complex challenges and unlocking new opportunities in image understanding, retrieval, and interpretation.

3

Data Discription and Study Area

This chapter commences by presenting the historical aerial imagery dataset, the Sentinel-2 dataset, and the Quantarctica Rock Outcrop Mask. Subsequently, an introduction to the study area is provided.

3.1. Data Discription

3.1.1. TMA historical aerial imagery

Antarctic Single Frames (1946-2000) is a collection of trimetrogon aerial (TMA) photographs over the Antarctic from the US military between 1946 and 2000. The TMA photography camera system captures static frames from left-oblique, nadir (directly down), and right-oblique perspectives. The TMA images are taken along multiple flight lines with the oblique cameras pointed at a depression angle of 30°. Each of the cameras has an angular field of view of 60°, which provides a 180° horizon to horizon coverage when the images are placed side-by-side. There are black-and-white, natural color, and color infrared images included in this collection, however, the majority of the photographs are black-and-white, which is why in this thesis we only look into the black-and-white photographs. In Fig. 3.1 are four example photographs from the collection. These four photographs are on the same flight line and were taken at consecutive positions, which is why the coastline in the four photographs gradually moves with respect to the last one. The ground sampling distance (GSD) slightly varies based on the specific camera height of each image. However, generally, the GSD of the TMA images remains at a sub-meter level.

Figure 3.1: Example images of Antarctica single frame collection.

In the dataset there are 2143 flight lines, and over 330,000 single-frame aerial photographs, covering

most of the glaciers in Antarctica, including most areas in the eastern part of Antarctica, and some of the coastlines in the western part. The flight lines are depicted in Fig. 3.2.

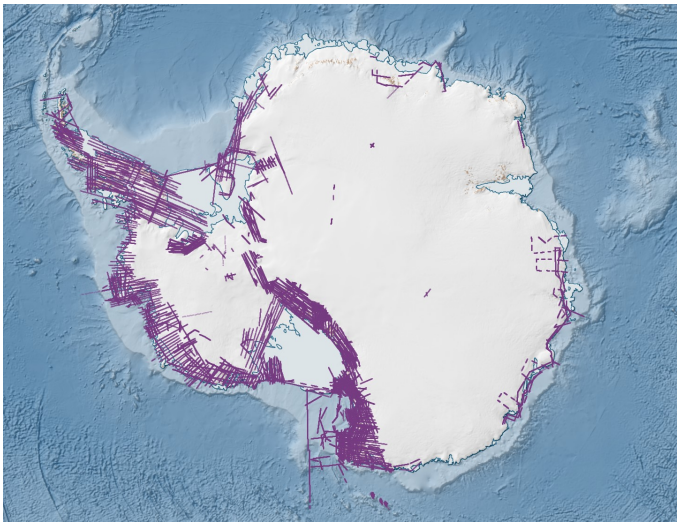


Figure 3.2: Flight lines in of Antarctica single frame collection.

3.1.2. Sentinel-2

The Sentinel-2 mission, launched by the European Space Agency (ESA) as a vital component of the Copernicus program, plays a pivotal role in facilitating a comprehensive assessment of Earth's environmental dynamics. This mission comprises two identical satellites, Sentinel-2A and Sentinel-2B, which operate in tandem as they orbit the Earth. This satellite constellation offers several advantages, including frequent revisit times and global coverage. Together, these satellites capture the entirety of the Earth's land surface every five days, ensuring extensive and recurrent data acquisition. The high-resolution multispectral imagery from Sentinel-2 enables intricate observations of various Earth surface elements, including forests, agricultural areas, urban landscapes, and aquatic ecosystems.

Sentinel-2 imagery is an accessible tool for monitoring the Antarctic Peninsula. Its frequent revisits, coupled with its high spatial resolution of up to 10 meters, make it ideal for examining diverse phenomena. This encompasses changes in ice coverage, glacial movement, sea ice patterns, and alterations within both terrestrial and aquatic ecosystems.

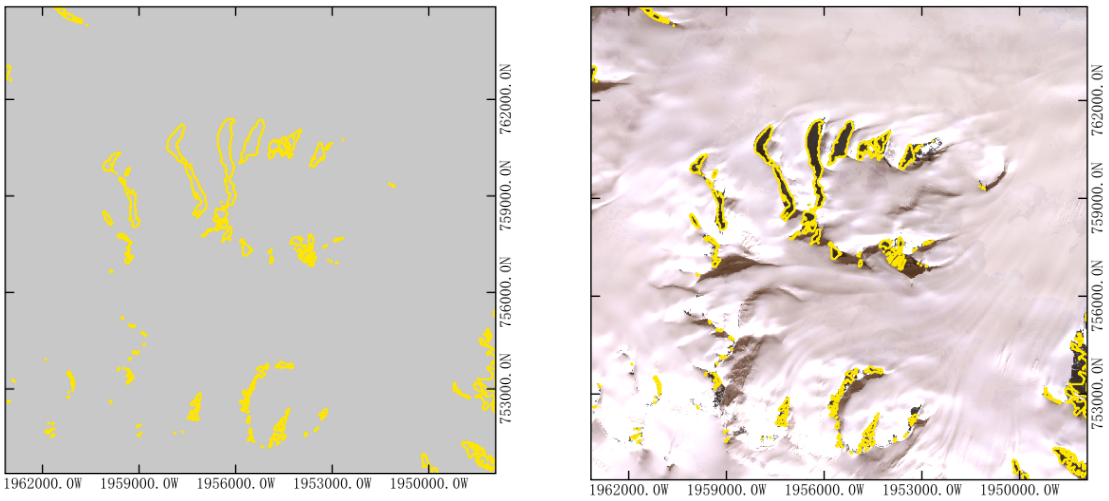
Sentinel-2 satellites capture images across 16 spectral bands, ranging from 0.43 µm to 2.30 µm, providing detailed information about the Earth's surface. This study focuses on utilizing bands 2, 3, and 4, corresponding to the blue, green, and red wavelengths. By assigning distinct colors to each band, Sentinel-2 images can be visualized in true color mode. However, for the purposes of this study, the images are converted into grayscale. Despite the conversion, the grayscale images retain information from all three bands, offering valuable insights into the monitored area.

3.1.3. Quantarctica Rock Outcrop Mask

Quantarctica gathers Antarctic geographical datasets intended for research, education, operations, and management purposes. It enables the exploration, importation, visualization, and sharing of Antarctic data. This compilation includes community-contributed, peer-reviewed data spanning ten scientific themes, complemented by a professionally-designed basemap [51].

Quantarctica provides a diverse package of datasets for exploring Antarctica's geology. Among these datasets, ADD Rock outcrop (Landsat8) employs an automated methodology on Landsat 8 imagery to differentiate rock formations from snow, clouds, and sea in Antarctica [52]. This dataset provides a vector layer delineating rocks as polygon geometries, as shown in Fig. 3.3, achieving a reported classification accuracy of 74 ± 9% with 1SD error [52]. While not explicitly specified in the original paper, this study observes an approximate spatial resolution of 10 meters for rock classification, consistent with the resolution of the Sentinel-2 imagery employed in this research. In this study, the vector layer was initially converted

to a raster layer and subsequently saved as a black-and-white TIFF image, with black representing rocks and white denoting snow.



(a) Example of ADD Rock Outcrop (Landsat 8) Dataset with yellow outline.

(b) Example of ADD Rock Outcrop (Landsat 8) Dataset with yellow outline overlaid on Sentinel-2 image.

Figure 3.3: Example: ADD Rock Outcrop (Landsat 8) Dataset and overlaid on Sentinel-2 Imagery.

3.2. Study Area

The study area is situated within the Antarctic Peninsula, indicated by the zoomed-out section at the bottom left in Fig. 3.4. The TMA-archive contains more than 100 flight lines over the Peninsula, providing a total of over 15,000 images that are suitable for training and testing. Besides, the Antarctic Peninsula is a mountainous region characterized by numerous glaciers, rock structures, ice shelves, and steep-sided valleys, which make ground structures more visible on aerial and satellite imagery compared to other regions. This is essential for image similarity estimation, since image similarity algorithms rely on ground structures to extract features for training.

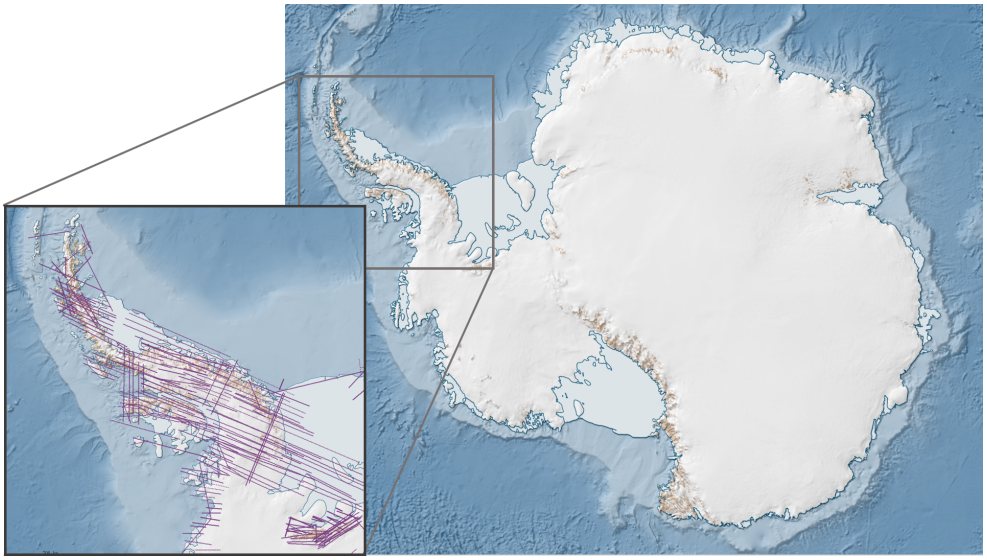


Figure 3.4: Study area: Antarctic Peninsula.

4

Methodology

This chapter outlines the methodology employed in this research. The study leverages two deep learning networks, and their workflow charts are depicted in Fig. 4.1 and Fig. 4.2, respectively. The primary distinction in the workflow of these two networks lies in the processing of input TMA images and the Sentinel-2 area of interest. Chapter 4.1 delves into the methods employed for handling the input TMA images and the Sentinel-2 area of interest (AoI). Subsequently, chapters 4.2 and 4.3 introduce the two networks, namely SigNet and the ResNet50-based network.

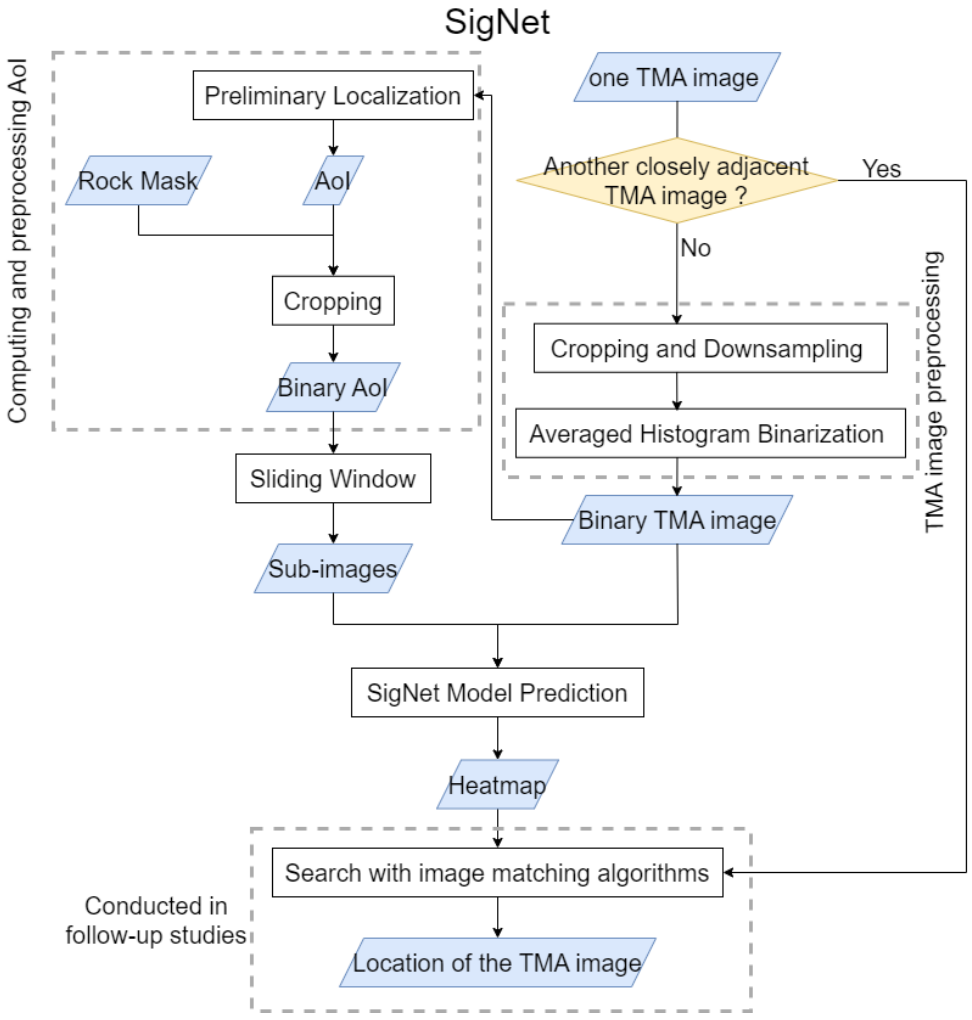


Figure 4.1: Workflow of the project using SigNet.

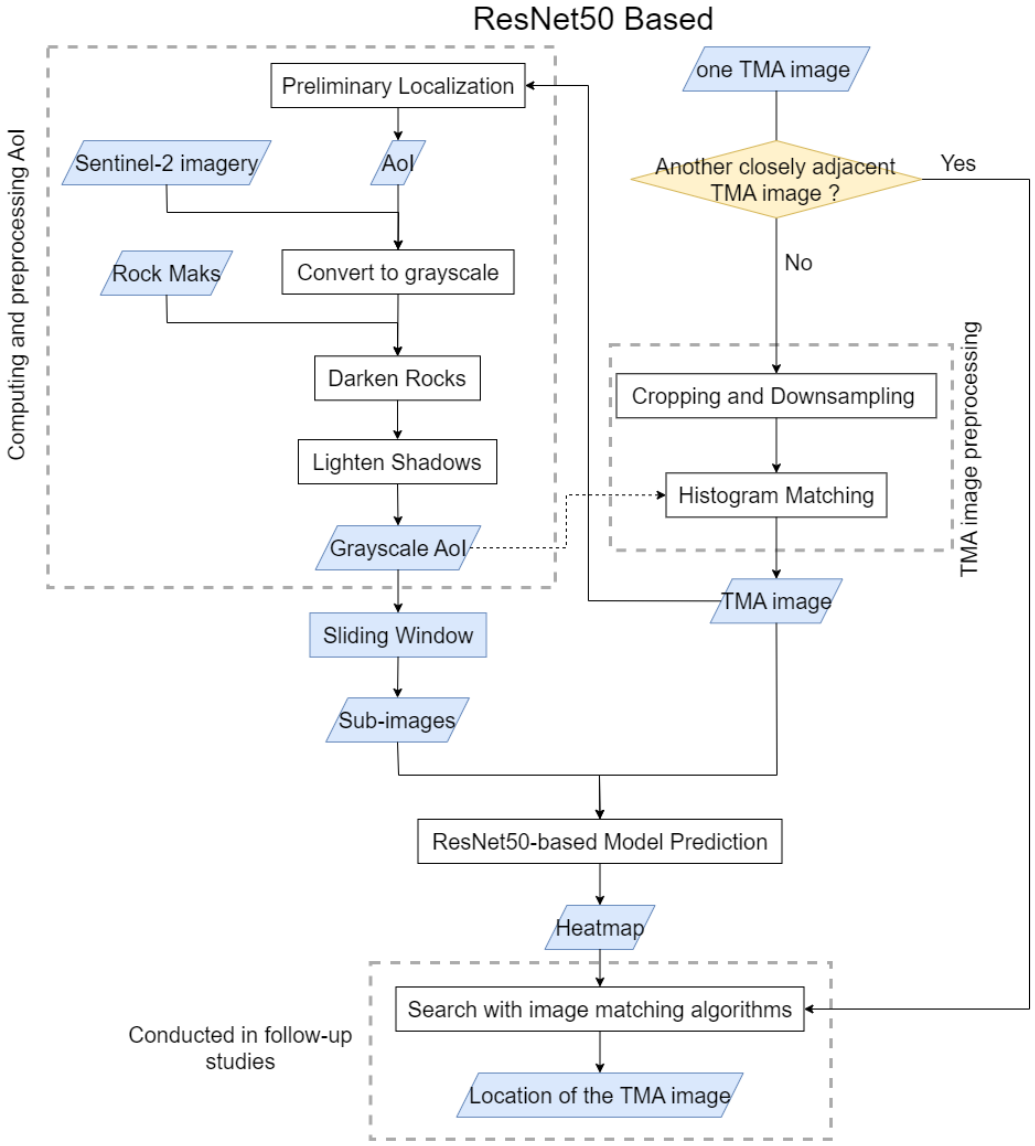


Figure 4.2: Workflow of the project using ResNet-50.

The overall workflow remains consistent when utilizing both networks. For each input TMA image, first it is checked if another TMA image exists on the same flight line within a distance of five or fewer shooting points. If such an image is identified, we leverage the positional relationships between the TMA images on the same flight line to deduce the location of the current TMA image. If not, the following steps are required: initially, image processing is conducted following the steps outlined in the workflow charts. Subsequently, preliminary localization is performed using a method developed in a prior study [4] to compute the coordinates of the Area of Interest (AoI) specific to the TMA image. Depending on the selected geo-referenced data source to serve as the reference image for geo-localization, the AoI is cropped from the data source. Image processing steps are then applied to obtain the AoI image.

Following this, a sliding window method is employed across the AoI image to generate sub-images, which are later compared with the TMA image. Fig. 4.3 illustrates how AoI images are transformed into sub-images using the sliding window method. Subsequently, each sub-image cropped from the AoI image, along with the pre-processed TMA image, is used as an image pair for the image similarity estimation model to predict the similarity scores between the TMA image and each sub-image. These similarity scores collectively form a heatmap.

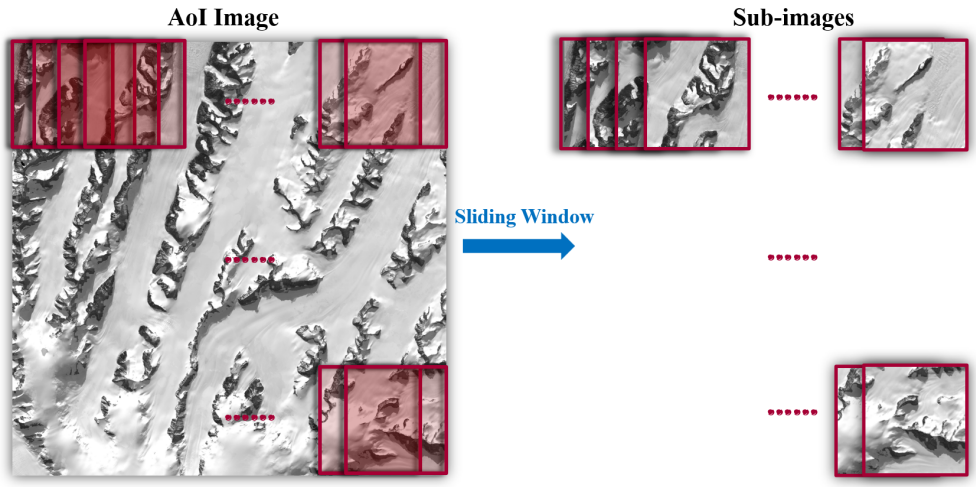


Figure 4.3: Generating sub-images from the grayscale Sentinel-2 Area of Interest (AoI) image using the sliding window method

Lastly, the heatmap is evaluated by traversing from pixels or sub-images with the highest similarity scores to the lowest. While this study does not explicitly detail the process of determining whether a pixel or sub-image is a match with the TMA image, the concept involves applying an image matching algorithm. This algorithm determines the authenticity of a match for each pixel or sub-image searched. Thus, the more confident or accurate the geo-localization method proposed in this study is in predicting the approximate location, the faster the subsequent image matching process can identify the precise location.

4.1. Image Triplet Generation

This study evaluates two networks for image similarity estimation: SigNet and a ResNet50-based network. SigNet operates on binary TMA images alongside corresponding binary rock mask images as Areas of Interest (AoI), while the ResNet50-based network processes grayscale TMA images paired with their respective grayscale Sentinel-2 images as AoI.

Both networks take image triplets as input for training. An image triplet includes an anchor, a positive, and a negative image. The anchor image serves as the reference against which other images are compared, establishing the baseline for determining similarity. A positive image is one that is similar to the anchor image. During training, the network aims to minimize the distance between the anchor and positive images in the feature space. Conversely, a negative image is one that is dissimilar to the anchor image. The network learns to maximize the distance between the anchor and negative images in the feature space.

Image triplets are put into the network for similarity estimation. Given Antarctica's landscape, dominated primarily by snow and rocks without other distinct ground features, rocks stand out as the most significant element. Accordingly, the primary objective during image pre-processing is to enhance the visibility of rocks within the image pairs and align their shades to make them more similar.

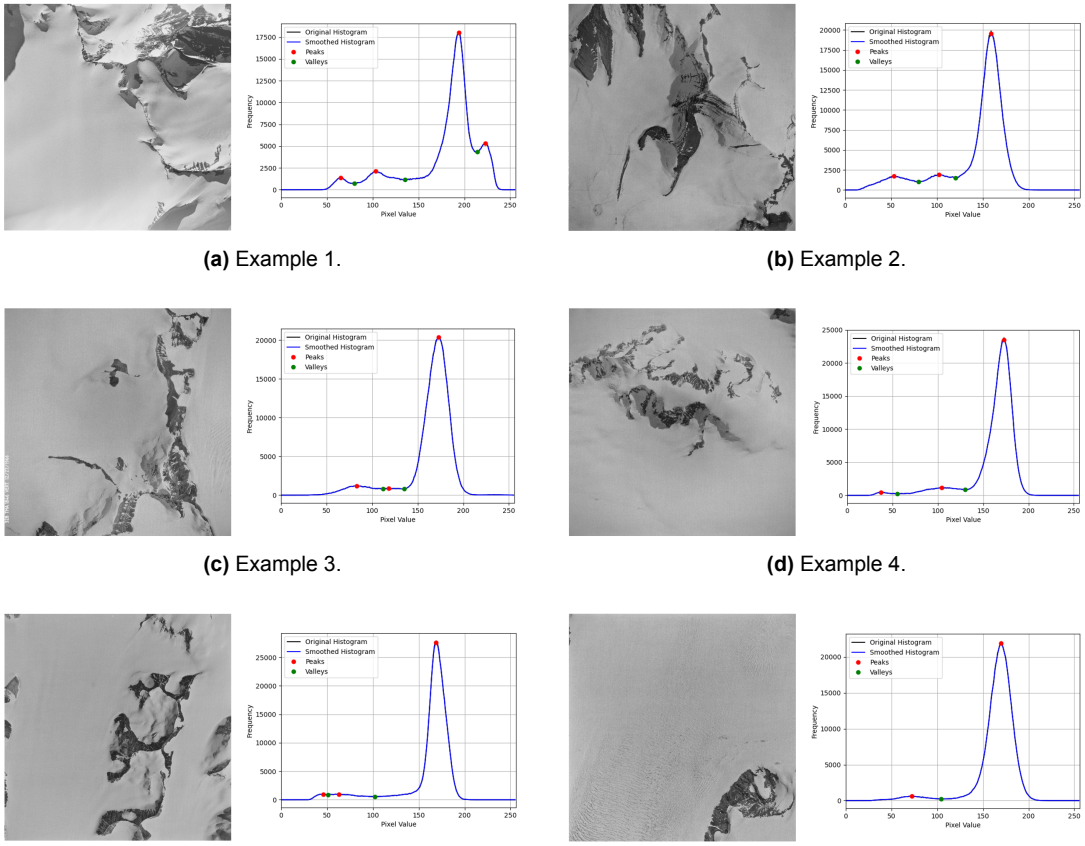
4.1.1. Binarizing TMA imagery for SigNet application

SigNet, one of the two networks utilized in this study, requires binary image triplets for similarity estimation. Hence, the challenge arises in binarizing the TMA images and the computed Quantarctica Rock Outcrop Area of Interest (AoI) image before feeding them into the network for prediction. The AoI image, extracted from the cropped Quantarctica Rock Outcrop (hereafter referred to as rock mask), is already binary, thus shifting the focus to binarizing the TMA images. The primary objective is to render the rock pixels black and other areas white, facilitating the identification of rocks between the TMA images and the rock mask.

Averaged Histogram Binarization

The binarization approach adopted for TMA image binarization in this study is termed 'averaged histogram binarization.' This method capitalizes on the characteristic histograms of TMA images. Upon inspecting these histograms, a discernible pattern emerges. As shown in Fig. 4.4, typically, a dominant peak

emerges in the range of pixel values between 150 and 200, signifying snow in the TMA images, as snow predominantly covers these areas. Additionally, two smaller peaks usually precede this dominant high peak. The first, found on the left, lies between 50-70 and represents rock pixels, being the darkest entities in the TMA images. The second smaller peak, closer to the highest peak, denotes shadows. Some TMA images exhibit only one small peak, as seen in Fig. 4.4f, possibly due to minimal shadow presence.



(a) Example 1. **(b)** Example 2.

(c) Example 3. **(d)** Example 4.

(e) Example 5. **(f)** Example 6.

Figure 4.4: Six examples of grayscale TMA image and histograms of their gray values.

Understanding these histogram patterns forms the basis for the averaged histogram binarization technique. Leveraging the fact that the first small peak on the left denotes rock pixels, the method sets the threshold at the valley between the first and second small peaks to differentiate between rock and shadows during binarization. To enhance the method's applicability across various images, the average histogram of 48 TMA images is calculated, and the threshold is set at the first valley. As illustrated in Fig. 4.5, the pixel value at this valley (79 in our case) serves as the threshold for subsequent TMA image binarization.

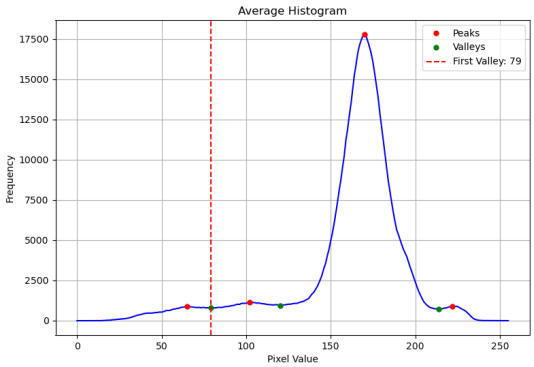
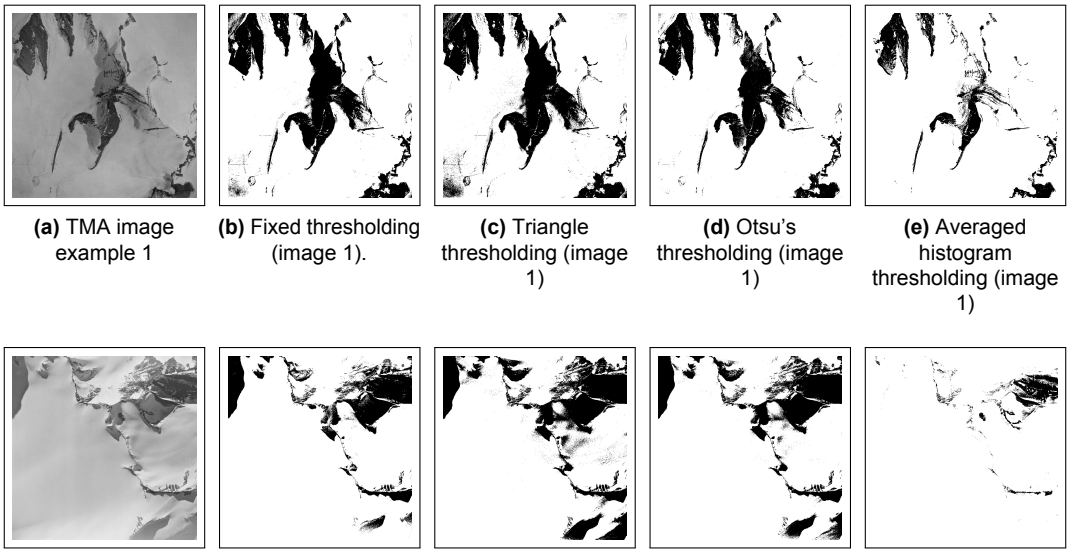


Figure 4.5: Averaged histogram over 48 TMA images.

The efficacy of averaged histogram binarization in TMA image processing surpasses other tested thresholding methods. Fig. 4.6 displays the binarization outcomes using different methods for two TMA images: fixed thresholding, triangle thresholding, Otsu's thresholding, and the averaged histogram thresholding employed in this study. Notably, the first three methods often render shadows as black, obscuring the shape of bare rocks. In contrast, the averaged histogram thresholding preserves the visibility of rock shapes by preventing shadow pixels from turning black.



(a) TMA image example 1 | **(b)** Fixed thresholding (image 1). | **(c)** Triangle thresholding (image 1) | **(d)** Otsu's thresholding (image 1) | **(e)** Averaged histogram thresholding (image 1)

(f) TMA image example 2 | **(g)** Fixed thresholding (image 2). | **(h)** Triangle thresholding (image 2) | **(i)** Otsu's thresholding (image 2) | **(j)** Averaged histogram thresholding (image 2)

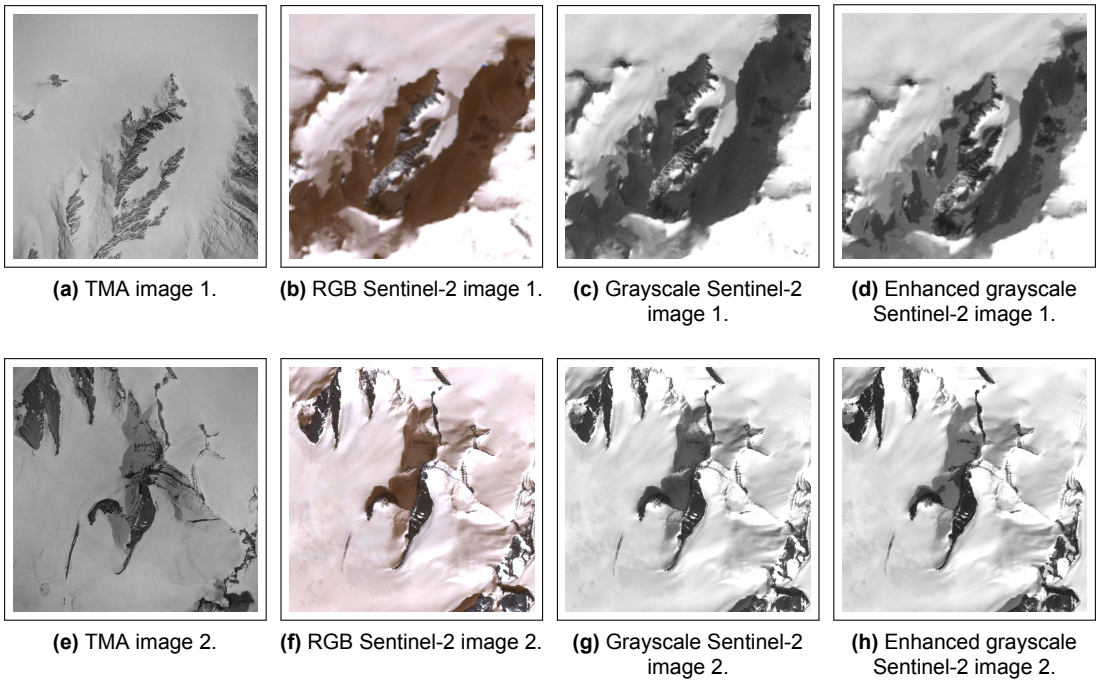
Figure 4.6: Comparison of different thresholding methods for binarizing TMA images.

4.1.2. Enhancing Grayscale Images for ResNet-50 based Network

The ResNet-50 based network utilized in this study is primarily intended for processing RGB image pairs to estimate similarity. Nevertheless, the TMA images employed in this analysis are grayscale. To ensure consistency and alignment between AoI and TMA images, it is needed to convert the RGB Sentinel-2 images to grayscale. This conversion aims to synchronize the tones and characteristics of the Sentinel-2 images with those of the TMA images, thereby facilitating a more precise evaluation of their similarity.

Sentinel-2 AoI
Converting RGB Sentinel-2 images into grayscale poses a challenge, as evident when comparing Fig. 4.7a to Fig. 4.7b and Fig. 4.7c. Even after converting to grayscale (Fig. 4.7c), significant differences persist between the TMA and corresponding grayscale Sentinel-2 images. The discrepancy mainly arises from the considerable darkness in the shadows of Sentinel-2 images, making it difficult to discern rocks from shadows. Conversely, TMA images typically exhibit fewer shadows.

The rock mask is employed in the image pre-processing phase. As both rock and shadow pixels appear quite dark in grayscale Sentinel-2 images, identifying rock pixels using the rock mask enables distinguishing between rock and shadow areas. By reducing pixel values for shadowed regions and increasing values for rock pixels, we diminish the darkness of shadowed areas while intensifying rock areas. This process enhances the distinction between rock and shadow pixels, making grayscale Sentinel-2 images resemble TMA images more closely, notably by mitigating the excessive darkness in shadowed regions.



(a) TMA image 1. **(b)** RGB Sentinel-2 image 1. **(c)** Grayscale Sentinel-2 image 1. **(d)** Enhanced grayscale Sentinel-2 image 1.

(e) TMA image 2. **(f)** RGB Sentinel-2 image 2. **(g)** Grayscale Sentinel-2 image 2. **(h)** Enhanced grayscale Sentinel-2 image 2.

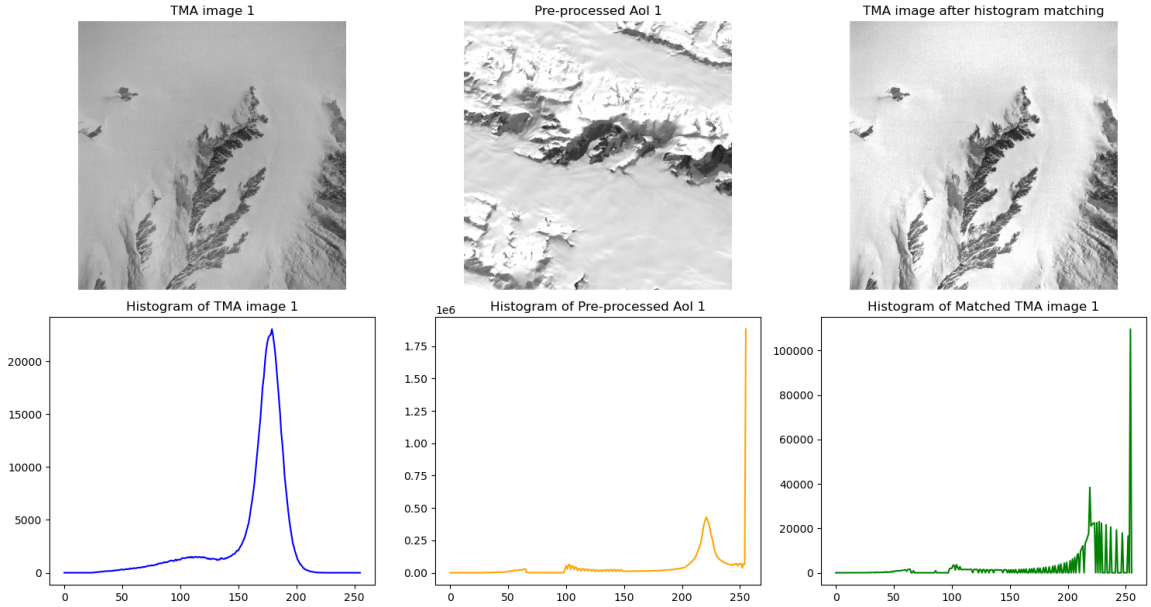
Figure 4.7: Comparison of TMA images and corresponding Sentinel-2 images at the same scene.

Examples of pre-processed grayscale Sentinel-2 images are displayed in Fig. 4.7d and Fig. 4.7h. Compared to merely converting RGB Sentinel-2 images to grayscale (Fig. 4.7c and Fig. 4.7g), shadows in these examples are less dark, although still darker than those in the TMA images. Additionally, rock pixels are emphasized. In summary, the contrast between rock and shadow pixels is heightened, making the grayscale Sentinel-2 images more analogous to the corresponding TMA images, thereby aiding in similarity estimation.

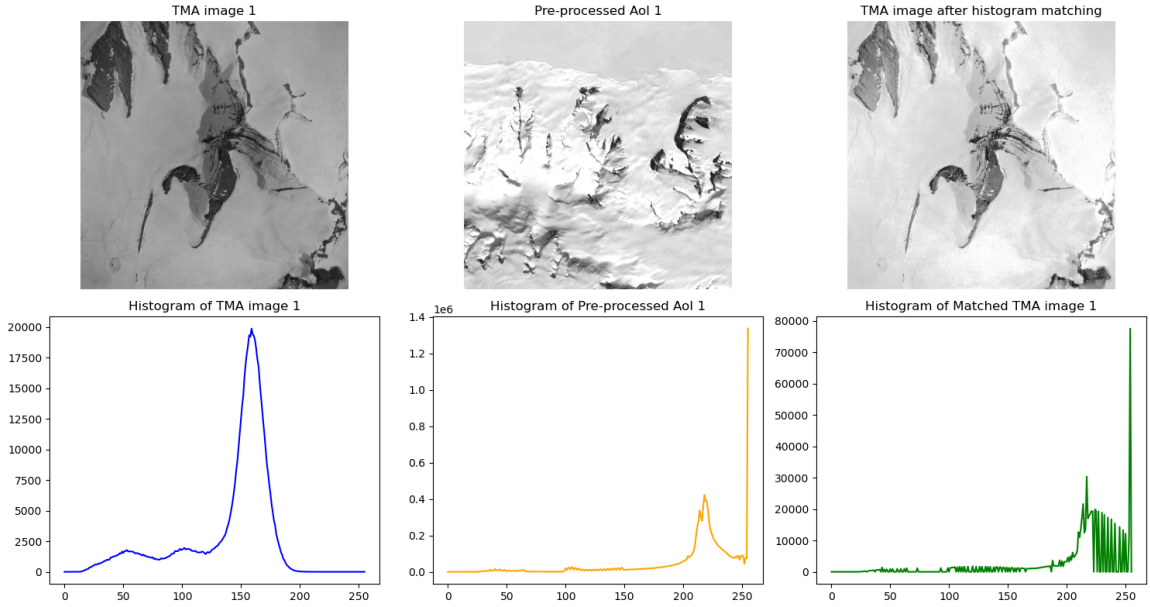
TMA images
After pre-processing the grayscale Sentinel-2 images, the resulting images depict shadows that are less dark and more discernible from the rocks. However, upon comparison between Fig. 4.7a and Fig. 4.7d, and also Fig. 4.7e and Fig. 4.7h, a noticeable discrepancy in shading persists between the TMA images and the pre-processed Sentinel-2 images. Specifically, the pre-processed Sentinel-2 images appear considerably brighter than the TMA images. Upon inspecting the histograms, it is evident that the Sentinel-2 images are saturated, indicating an excess of white pixels (pixel value 255) in the image, notably showcased in the second histograms (yellow) in Fig. 4.8.

Therefore, to align the TMA images and Sentinel-2 images more closely, we employed histogram matching for the TMA images to match the corresponding Sentinel-2 images. Fig. 4.8 showcases the histogram

matching results for two images. Notably, the TMA images are not matched precisely to the corresponding enhanced Sentinel-2 images at the same scene. Due to the nature of computing an Area of Interest (AoI) for each TMA image before network prediction, the matching involves the TMA image and the AoI image. However, this distinction does not diminish its efficacy, as the shading of the AoIs should ideally be consistent with the actual scene.



(a) Histogram matching of TMA image 2.



(b) Histogram matching of TMA image 2.

Figure 4.8: Two examples of histogram matching of TMA images.

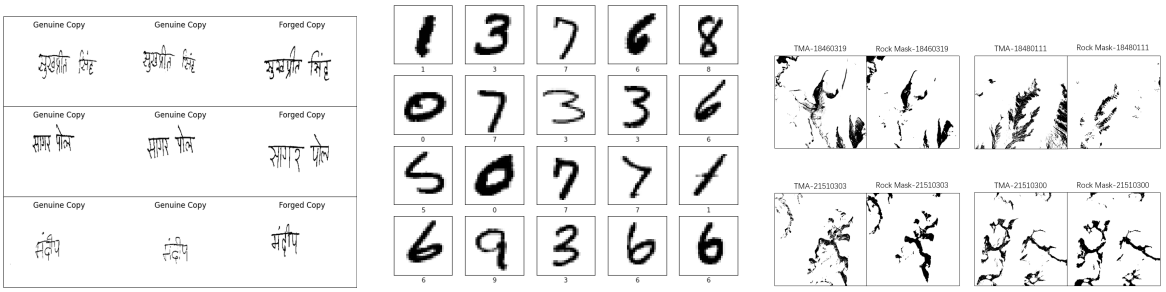
As shown in Fig. 4.8, after histogram matching, the TMA images appear considerably brighter. The histograms post-matching indicate saturation in the TMA images, similar to the pre-processed Sentinel-2 images. The post-matching TMA images show markedly improved similarity in shading to the Sentinel-2 AoI images.

4.2. SigNet Siamese Network

SigNet, a convolutional Siamese network proposed in [53], is specifically designed to address the offline writer-independent signature verification problem. This problem involves verifying the authenticity of handwritten signatures without relying on information about the specific individuals who wrote them. In essence, it seeks to authenticate signatures across various writers under conditions where signatures are not captured in real-time or with prior knowledge of the signers. In the original paper, training images were resized to 155×220, ensuring uniform image sizes for batch training in neural networks. In this study, a resizing to 155×220 is also employed, as the model is pre-trained on images of such dimensions.

In this study, SigNet was selected as an image similarity estimation model due to its adoption of a classic network architecture. It stands as a representative example of 'simple' Convolutional Neural Networks (CNNs) notable for its resemblance to early classic CNN architectures such as Lenet [54] and AlexNet [55], particularly bearing similarities to AlexNet. As a matter of fact, the SigNet architecture is inspired by Alexnet in the first place [53]. These simpler CNN architectures have been extensively employed across diverse datasets and various research studies [56] [57] [58].

Moreover, the selection of SigNet is influenced not only by its representation of simple CNNs but also by its original design for signature verification. As shown in Fig. 4.9, signatures primarily consist of black and white lines, similar to the rock structures observed in the context of this study. Although rock structures may exhibit greater complexity compared to signatures, exploring the application of the straightforward SigNet network to this problem is a promising avenue worth investigating.



(a) Hindi Signatures dataset - Each individual has a genuine signature, a forged signature, forming triplets.

(b) MNIST Handwritten Digits Dataset.

(c) Rock Dataset - four pairs of images generated for the SigNet model within this study.

Figure 4.9: Comparison of TMA and Rock dataset and other signature datasets.

In this study, transfer learning, or adaptive learning was implemented within the SigNet network architecture. This technique involves harnessing a model previously trained on a particular task or dataset to address a related, distinct task or dataset, thereby accelerate learning or enhancing performance. Unlike many other studies relying on existing pre-trained models, the absence of an established pre-trained SigNet model necessitated the development of a custom pre-trained model. Initially, the model underwent training on a Hindi signature dataset [59], acquiring unique features. Subsequently, leveraging these learned weights, the entire model underwent retraining on the TMA-Rock Triplet dataset created in this study. This approach facilitated the update of parameters across all layers using the new dataset. Consequently, the knowledge of distinctive features learned from Hindi signatures was effectively transferred to the TMA-Rock Triplets. This adaptation through transfer learning proved pivotal in our study, given the constraint of a limited dataset size. By reusing the features from the pre-trained model, the model adeptly adjusted these general features to the specific intricacies of the new task, enabling more efficient learning from limited data.

4.2.1. Network Architecture

A Siamese neural network constitutes a class of architectures featuring twin identical sub-networks. These parallel CNNs are characterized by matching configurations, parameters, and shared weights. Updates to parameters across the network are mirrored in both sub-networks. The convergence of these sub-networks occurs at a loss function atop the network, which computes a similarity metric, often involving the Euclidean Distance between feature representations from each side of the Siamese network. Two widely used loss

functions in Siamese networks are contrastive loss [60] and triplet loss [61]. In the context of SigNet, the contrastive loss function is employed and defined as follows:

$$L(s_1 s_2, y) = \frac{1}{2} \times (y \times d(s_1, s_2)^2 + (1 - y) \times max(0, m - d(s_1, s_2))^2) \tag{4.1}$$

Here, $d(s_1, s_2)$ denotes the distance between the embeddings of samples s_1 and s_2 in the embedding space. y = 1 signifies similar pairs (positive pairs), while y = 0 denote dissimilar pairs (negative pairs); In our case, the margin hyperparameter (m) is set to one. Unlike conventional approaches that assign binary similarity labels to pairs, a Siamese network endeavors to bring output feature vectors closer for input pairs labeled as similar and pushes the feature vectors apart for dissimilar input pairs. Each branch of the Siamese network functions as an embedding function, transforming input images into a space where pairs deemed similar are closer than dissimilar pairs. The network branches are interconnected by a layer that computes the Euclidean distance between two points in the embedded space. To obtain the similarity score between image pairs, the distance is normalized between 0 and 1, representing the final similarity score where 1 indicates exact similarity and 0 signifies no similarity.

The CNN layer parameters employed in SigNet are detailed in Table 4.1. For convolutional and pooling layers, the filter sizes are denoted as N×H×W, where N represents the number of filters, H is the filter height, and W denotes the filter width. 'Stride' denotes the spacing between filter applications during convolution and pooling operations, while 'pad' indicates the width of added borders to the input. Rectified Linear Units (ReLU) serve as the activation function across all convolutional and fully connected layers. To enhance the generalization of learned features, Batch Normalization is applied using the parameters outlined in Table 4.1. Additionally, Dropout is implemented with rates of 0.3 and 0.5 for the last two pooling layers and the initial fully connected layer, respectively.

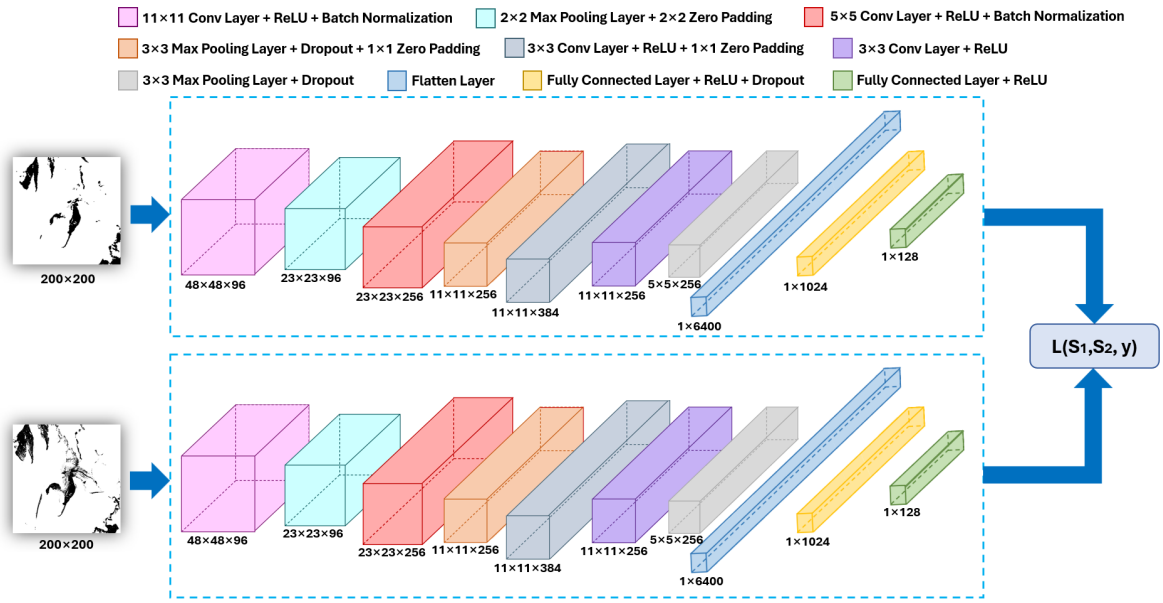


Figure 4.10: SigNet architecture.

As depicted in Fig. 4.10, the initial convolutional layers filter the 155×220 input image utilizing 96 kernels of size 11×11 with a stride of 1 pixel. Subsequently, the second convolutional layer processes the (normalized and pooled) output from the first layer using 256 kernels sized 5×5. The third and fourth convolutional layers are directly connected without intermediate pooling or normalization layers. The third layer employs 384 kernels sized 3×3, linked to the (normalized, pooled, and dropout) output of the second convolutional layer. The fourth convolutional layer incorporates 256 kernels of size 3×3. This architecture enables the neural network to learn fewer lower-level features with smaller receptive fields and more features for higher-level or abstract representations. The initial fully connected layer comprises 1024 neurons, while

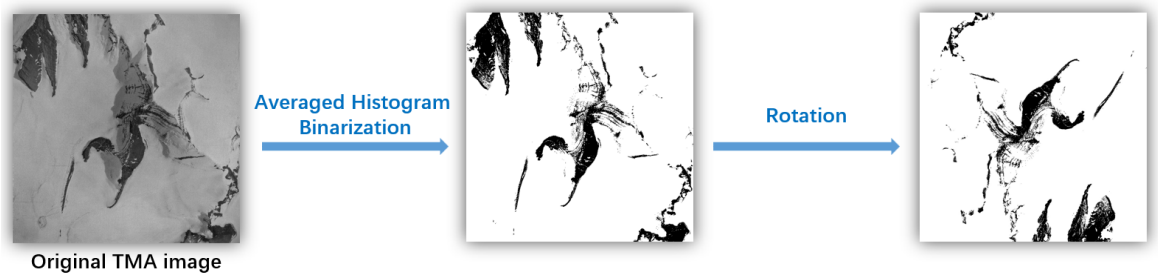
the subsequent fully connected layer consists of 128 neurons, signifying that the highest learned feature vector from each side of SigNet is a 128-dimensional vector.

Layer	Size	Parameters
Convolution	96×11×11	stride = 1
Batch Normalization	–	$\epsilon = 10^{-6}$, momentum = 0.9
Pooling	96×3×3	stride = 2
Convolution	256×5×5	stride = 1, pad = 2
Batch Normalization	–	$\epsilon = 10^{-6}$, momentum = 0.9
Pooling + Dropout	256×3×3	stride = 2, p = 0.3
Convolution	384×3×3	stride = 1, pad = 1
Convolution	256×3×3	stride = 1, pad = 1
Pooling + Dropout	256×3×3	stride = 2, p = 0.3
Fully Connected + Dropout	1024	p = 0.5
Fully Connected	128	–

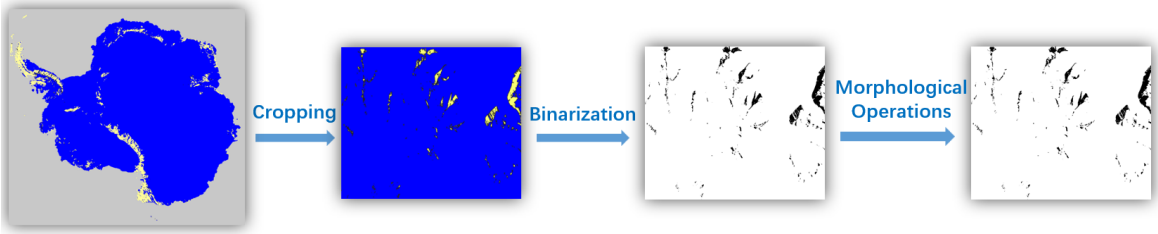
Table 4.1: Overview of the constituting CNNs in SigNet

4.2.2. Data Preparation

The image processing procedures are consistent between creating the training dataset and processing test images intended for input into the trained model for predictions. The image pairs are sourced from two distinct data repositories: the TMA dataset and the Quantarctica Rock Outcrop dataset. Processing the TMA images involves several sequential steps. Initially, the process involves cropping out the informational frame within the image. Subsequently, binarization occurs based on the threshold obtained from the averaged histogram thresholding method detailed in chapter 4.1.1. Following this, the TMA image is rotated using the azimuth angle derived from the additional documentation in the TMA dataset. Utilizing the coordinates of the Areas of Interest (AoIs), the next step involves extracting the AoI from the rock mask. This extraction results in a shapefile containing rocks represented as polygonal geometry objects. Subsequently, the shapefile is converted into a PNG image, with rocks depicted as black and other areas as white. Finally, various morphological operations are applied to the rock mask. The process of image processing for SigNet is visually depicted in Fig. 4.11.



(a) Processing procedure of TMA images.



(b) Processing procedure of AoI images.

Figure 4.11: Data preparation for SigNet.

Creation of training dataset

The dataset consists of a total of 188 image triplets, where half (94) stem from the augmentation of the other half. Through data augmentation, the 94 image triplets are each rotated by 90 degrees, effectively doubling the dataset size from 94 to 188 image triplets. The generation of the image triplet dataset involves several steps. Initially, each TMA image undergoes processing as illustrated in Fig. 4.11a. Subsequently, the AoI is computed, and the rock mask is processed as outlined in Fig. 4.11b. The subsequent step involves manual inspection to identify and extract the scene from the TMA image corresponding to the AoI. This cropped scene from the AoI and the processed TMA image constitute an image pair. The images in are sized at 820×820 pixels, applicable to both the TMA images and the rock mask. The dataset originally consisted only of image pairs with the anchor and the positive image. However, since the model also requires negative pairs, the dataset was expanded by adding a negative image to each image pair. This was achieved by randomly selecting another image, different from the positive image, from the dataset and forming a triplet. Several examples of such image triplets are shown in Fig. 4.12.

(a) Image triplet 1 (Left: Anchor | Middle: Positive | Right: Negative)

(b) Image triplet 2 (Left: Anchor | Middle: Positive | Right: Negative)

(c) Image triplet 3 (Left: Anchor | Middle: Positive | Right: Negative)

(d) Image triplet 4 (Left: Anchor | Middle: Positive | Right: Negative)

Figure 4.12: Examples of image triplets in the TMA-Rock Triplets dataset for SigNet.

The dataset is further partitioned into training, validation, and testing sets, with an 8:1:1 proportion. Given the dataset's size, the emphasis is placed on allocating a significant portion to the training set to ensure the network is sufficiently trained.

4.2.3. Training Process

The entire framework was built using the Keras library with TensorFlow as the backend. As previously outlined, SigNet went through two training phases in our study. The initial training occurred on the Hindi Signature dataset, a sizable collection comprising over 60,000 signature triplets. In this phase, the SigNet model was trained using RMSprop for 100 epochs, employing a momentum rate of 0.9 and a mini-batch size of 128. The initial learning rate (LR) was set at 1e-4, with hyperparameters $\rho = 0.9$ and $\epsilon = 1e - 8$. These values are summarized in Table 4.2. This training process took around 1 hour on Python 3 Google

Compute Engine backend (TPU) provided by Google Colab.

Parameter	Value
Initial Learning Rate (LR)	1e-4
Learning Rate Schedule	LR ←LR×0.1
Weight Decay	0.0005
Momentum (ρ)	0.9
Fuzz factor (ϵ)	1e-8
Batch Size	128

Table 4.2: Training Hyper-parameters for SigNet

After completing the initial training phase, the model underwent a second training phase known as adaptive training. In this phase, the model's weights obtained from the first training were used as a starting point, and all layers were retrained using the TMA-Rock Mask Triplet dataset. The training parameters remained consistent with the initial training settings (refer to Table 4.2). However, due to the limited size of the TMA-Rock Mask Triplet dataset, the batch size was reduced to 1 to ensure efficient learning.

The adaptive training process was completed remarkably quickly, taking only 303 seconds to train on the TMA-Rock Mask Triplet dataset, and reaching an early stop at 18 epochs. The training loss vs. validation loss figure presented below illustrates that both the training loss and validation loss decreased steadily throughout the adaptive training process. Additionally, the small generalization gap, which is the difference between the training loss and the validation loss, indicates that the model's performance is not significantly affected by overfitting. This observation suggests that the model is effectively learning patterns from the new dataset while maintaining its generalization ability.

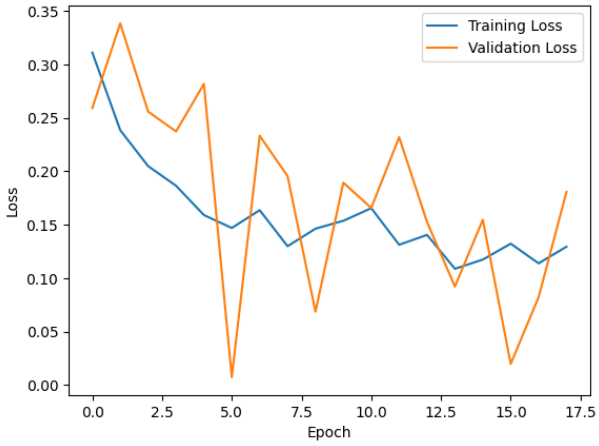


Figure 4.13: Training loss VS Validation loss for SigNet

Both the pre-training on the Hindi Signature dataset and the adaptive training phase leveraged the Python 3 Google Compute Engine backend (TPU) provided by Google Colab. The pre-training on the Hindi Signature dataset took approximately 1 hour on the TPU, while the adaptive training process took only 303 seconds to complete training on the dataset consisting of 188 image triplets generated specifically for this study.

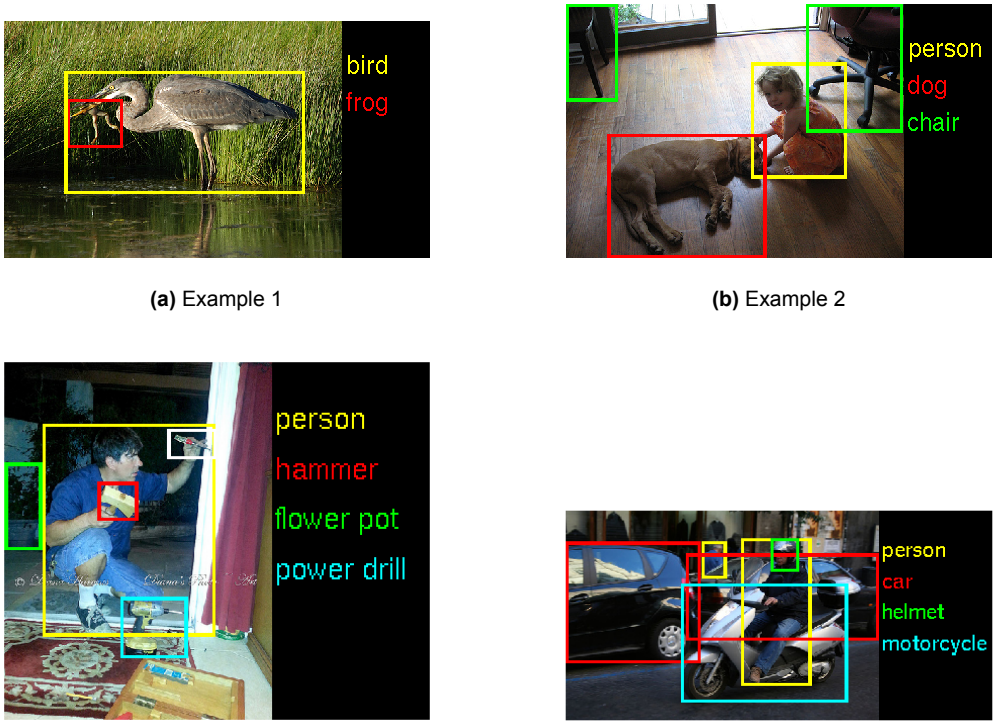
4.3. ResNet50 based Siamese Network

Siamese networks might benefit from deeper architectures to learn complex representations from input pairs, therefore using a relatively deep and sophisticated network as the backbone is a popular choice in many studies. Networks like AlexNet [55], ResNet [62], VGG [63], Inception [64], EfficientNet [65],

MobileNets [66], are some common examples to be used as base network or backbone of a Siamese network. Among which, ResNet's design enables training of very deep networks and it's state of art performance makes it a popular choice in many studies[67], [68]. Also, the unique residual connections (Skip Connections) introduced by ResNet facilitate the flow of gradients during training, allowing information to propagate more easily through the network, this is also a big advantage of ResNet.

ResNet models pre-trained on various datasets are readily available. In this study, we utilize ResNet-50, pre-trained on the ImageNet dataset, as the backbone of the Siamese network. ImageNet comprises millions of labeled everyday images across thousands of object categories, as illustrated in Fig. 4.14. Pre-training on ImageNet has become a common practice for transfer learning in computer vision. Models pre-trained on ImageNet have demonstrated strong generalization capabilities across various downstream tasks. This allows researchers to fine-tune these models for specific applications using smaller datasets.

ImageNet consists of everyday images. Through transfer learning, specifically adaptive training, knowledge acquired from the source domain (ImageNet) can be effectively transferred to the target domain. For instance, in this study, knowledge was adapted from ImageNet to a remote sensing image dataset, specifically the TMA-Sentinel triplet dataset created herein. This adaptation is possible due to the shared underlying features between the domains.



(a) Example 1

(b) Example 2

(c) Example 3

(d) Example 4

Figure 4.14: Examples of images and labels in ImageNet[1].

4.3.1. Network Architechture

ResNet-50, a variant of the Residual Network (ResNet) architecture introduced by He et al. in [62], is characterized by its 50-layer depth, organized into a sequence of four stages, each comprised of varying numbers of 'bottleneck' blocks. Within ResNet-50, these 'bottleneck' blocks consist of three key convolutional layers: 1×1, 3×3, and 1×1 convolutions. Notably, the initial convolutional layer, responsible

[1]Image source: https://www.image-net.org/challenges/LSVRC/2014/

for processing input images, lies outside the count of residual blocks. Following these blocks, the network concludes with additional layers, including global average pooling and a fully connected layer tailored for classification tasks.

In the architectural design of the Siamese network, a ResNet-50 model pre-trained on ImageNet interfaces with two fully connected layers, illustrated in Fig. 4.15. The primary function of the ResNet-50 model is to generate embeddings for the input image triplets. Subsequently, two fully connected layers aim to discern and differentiate these embeddings. During the adaptive training phase, the upper layers, which include some fully connected layers from the original ResNet-50 model, are discarded. Furthermore, the weights of the ResNet-50 layers up to 'conv5_block1_out' are frozen, leaving only the two lower convolutional blocks ('conv5_block2' and 'conv5_block3') trainable. These blocks, each comprising three convolutional layers (1×1, 3×3, and 1×1), collectively account for the six trainable convolutional layers. After the convolutional layers, the output, shaped $7\times7\times2048$, is flattened into one-dimensional features with a shape of 1×100352. Subsequently, two fully connected layers, housing 512 and 256 neurons, are applied to the flattened layer. These layers provide 256 feature representations crucial for computing the distance within the loss function. Each of these fully connected layers is followed by a ReLU activation function and a batch normalization step.

During adaptive training, freezing the weights before the 'conv5_block1_out' layer is a strategic choice. These early layers typically capture fundamental features such as edges, textures, and basic patterns, which are transferable across diverse datasets. By preserving these layers during adaptive training, the model focuses on acquiring task-specific features in the later layers. This strategy of freezing the layers also helps mitigate overfitting by constraining the number of adaptable parameters, a critical consideration because of the constrained size of our training data.

The selection of 512 and 256 neurons in the fully connected layers determines the dimensionality and complexity of the network's learned representations during the transition from convolutional features to higher-level abstractions. Given the relatively straightforward nature of our dataset, devoid of intricate attributes necessitating higher-level representations such as semantic or temporal complexities, our focus remains on spatial aspects like structure, texture, and arrangement. Hence, opting for 512 and 256 neurons in the fully connected layers is expected to provide sufficient representation for our dataset.

Triplet loss [61] is used in the ResNet-50 based network. It's particularly popular in scenarios where you have sets of triplets: an anchor, a positive example, and a negative example. The objective of triplet loss is to make sure the distance between the anchor and the positive example in the embedding space is smaller than the distance between the anchor and the negative example, by at least a certain margin. The loss function encourages the model to minimize the distance between the anchor and the positive example while simultaneously maximizing the distance between the anchor and the negative example. This creates a 'margin' between the positive and negative examples, allowing the model to learn better embeddings that capture the desired similarity relationships. The formula for triplet loss is often expressed as:

$$L(s_0 s_1, s_2) = max(d(s_0, s_1) - d(s_1, s_2) + margin, 0) \tag{4.2}$$

Here, $d(s_0, s_1)$ is the distance between the anchor and the positive example in the embedding space. $d(s_0, s_2)$ is the distance between the anchor and the negative example in the embedding space. Margin is a hyper-parameter that specifies the minimum difference between the distances. The loss is computed for each triplet in the training set, and the overall objective is to minimize this loss function across all triplets. By using triplet loss, the Siamese network learns to produce embeddings where similar examples are closer together, and dissimilar examples are farther apart in the embedding space, improving its ability to perform similarity matching. Similar to SigNet Siamese network, the network branches are interconnected by a layer that computes the Euclidean distance between two points in the embedded space. To obtain the similarity score between image pairs, the distance is normalized between 0 and 1.

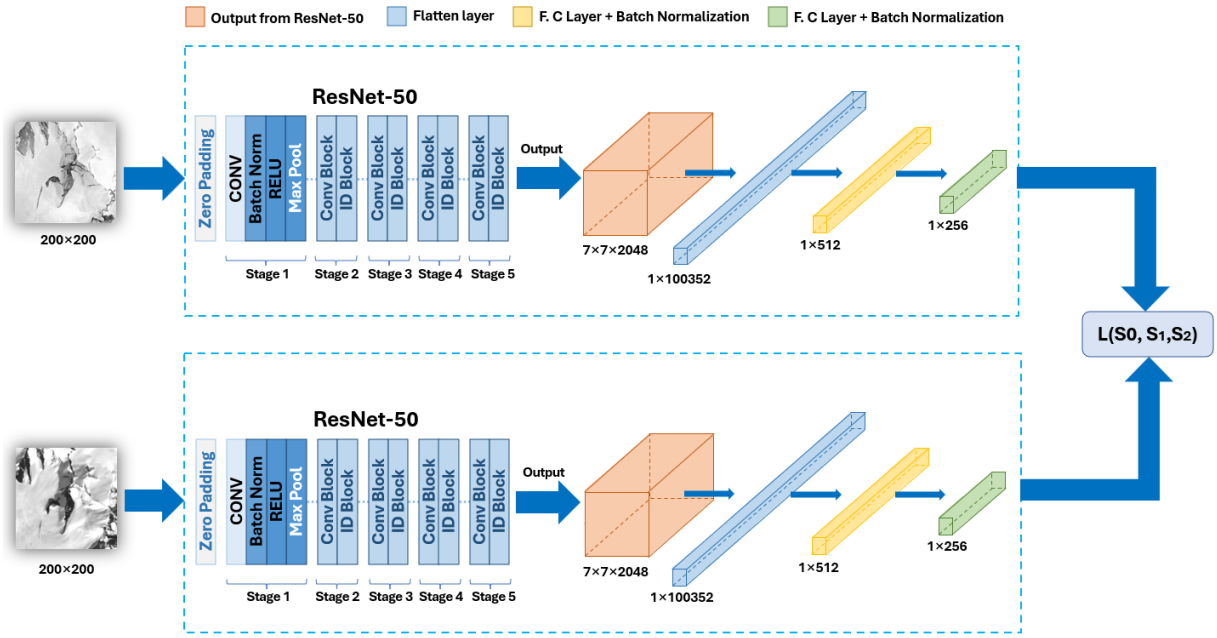


Figure 4.15: ResNet-50 based Siamese network architecture.

4.3.2. Data Preparation

In the context of the ResNet-50 based Siamese network, the image triplets are sourced from two distinct data repositories: the TMA dataset and the Sentinel-2 dataset. The pre-processing techniques outlined in chapter 4.1.2 are applied uniformly to both the training and additional testing image pairs. It is crucial to note that the processing of TMA and Sentinel-2 images is interdependent and should not be viewed as separate processes.

The processing sequence for both image types comprises several sequential steps. Initially, the information frame is removed from the TMA images. Simultaneously, using the Area of Interest (AoI) coordinates computed during the preliminary localization step, the corresponding AoI within the Sentinel-2 images is cropped from the larger dataset. Subsequently, the AoI Sentinel-2 image undergoes grayscale conversion, accompanied by enhancement techniques to highlight rocks while mitigating shadows, as elucidated in chapter 4.1.2. Further in the process, the AoI Sentinel-2 image is rotated to align with the azimuth angle of the TMA images. Additionally, histogram matching is employed on the grayscale TMA image to harmonize it with the AoI Sentinel-2 grayscale images, enhancing their resemblance. This workflow is visualized in Fig. 4.16.

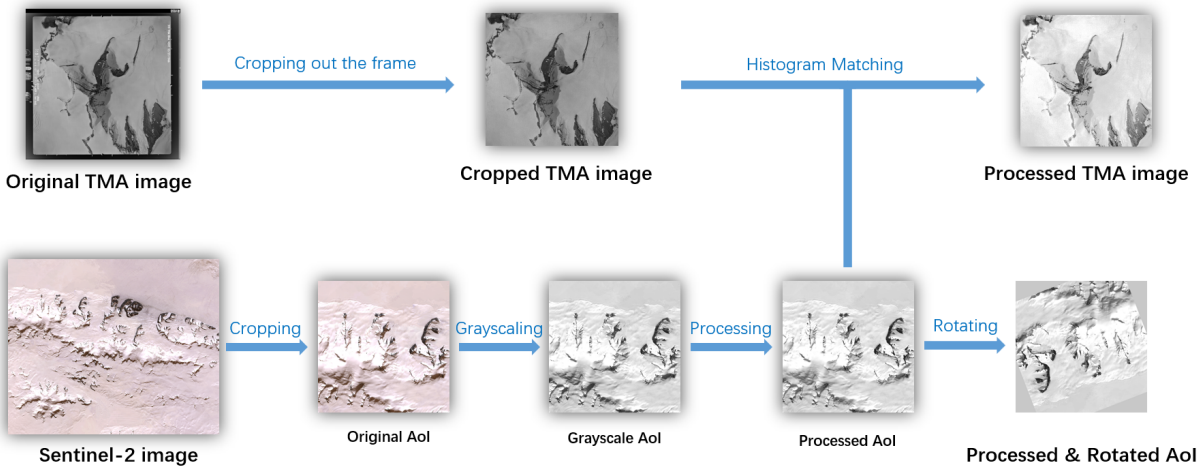
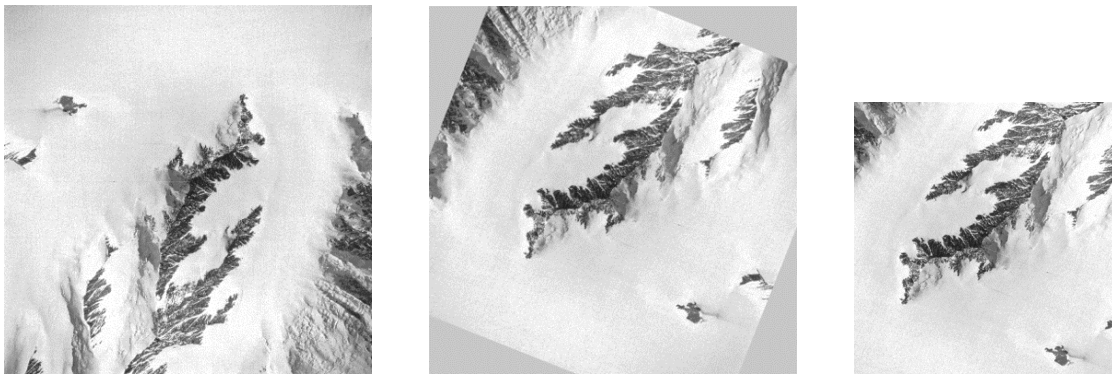


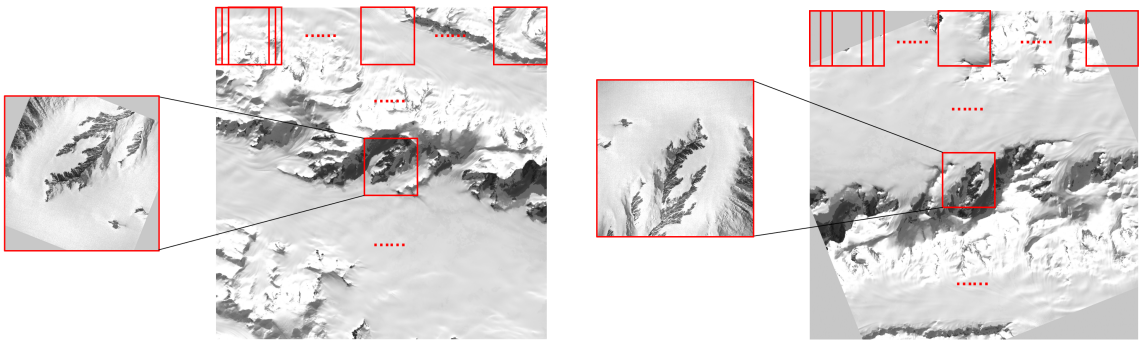
Figure 4.16: Data preparation for ResNet-50 based network.

The rationale behind rotating the AoI images instead of the TMA images, in contrast to the image processing methodology for SigNet, is driven by the grayscale nature of the images. Rotating the TMA images presents a challenge in preserving edge data without introducing false information by uniformly filling empty areas, which can significantly impact the accuracy of similarity estimation. As depicted in Fig. 4.17, a dilemma emerges between preserving information and introducing false data in these two rotation strategies. This dilemma becomes particularly critical when considering rotating the TMA images. Since each TMA image is paired and compared with every sub-image extracted from the AoI Sentinel-2 images, any false or lost information in the TMA image significantly affects the similarity estimation for all image pairs, as illustrated in Fig. 4.18a. Conversely, rotating the AoI Sentinel-2 images confines false information only to the sub-images on the corners of the AoI image, ensuring that similarity scores for most pairs predominantly reflect genuine information, as demonstrated in Fig. 4.18b. However, rotating the AoI image necessitates an additional step of computing the location pre-rotation in the final stages.



(a) original TMA image **(b)** Rotation method 1 **(c)** Rotation method 2

Figure 4.17: Two rotation methods. Method 1 crops the parts of the rotated image that extend beyond the original boundaries, while filling the empty areas with a specific value (e.g., 128). While preserving much of the original information, this method introduces false information by adding uniformly colored padded areas. Method 2 maintains the original image ratio and contains only the areas covered by the original image without introducing any additional padded areas. However, it may result in significant information loss as it strictly limits the rotated image to the covered regions of the original image.



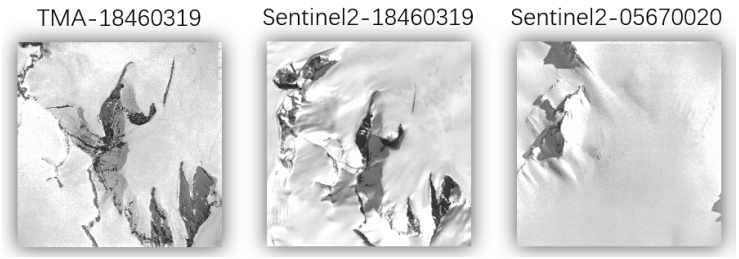
(a) Rotating the TMA image. False information in the TMA image is introduced into all similarity estimation process between each sub-images on the AoI.

(b) Rotating the AoI image. The TMA image has no false information, therefore false information is only introduced in the similarity estimation process between the TMA and the sub-images on the corners of the AoI image.

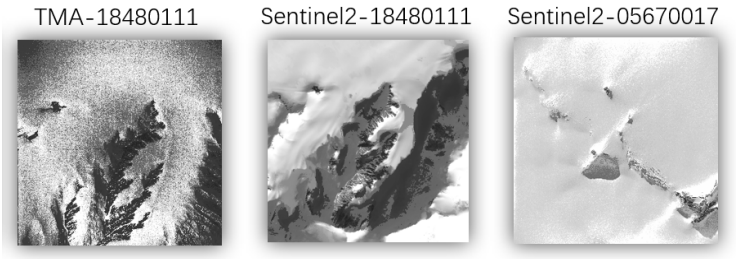
Figure 4.18: Comparison of similarity estimation for rotating the TMA image and the AoI image.

Creation of training dataset

Similar to training dataset creation for SigNet, the training dataset employed for the ResNet-50 based network comprises the same 188 image triplets as for SigNet, wherein half (94) are augmented versions of the other half. Employing data augmentation techniques, the original 94 image triplets are individually rotated by 90 degrees, effectively augmenting the dataset size from 94 to 188 image pairs. The model necessitates image triplets as its input data structure. Each image triplet consists of an anchor, a positive, and a negative image. Anchor images serve as the initial reference TMA images, while positive images are sub-images extracted from AoI Sentinel-2 images, presenting a similar scene to the corresponding anchor. Conversely, negative images represent sub-images portraying dissimilar scenes from the anchor. In this study, negative images are generated by randomly selecting an image from the positive set, and ensuring it does not correspond to the true match of the reference TMA anchor image. Examples of image triplets are depicted in Fig. 4.19.

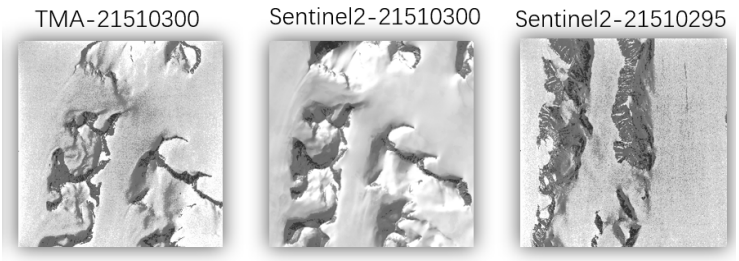


(a) Image triplet 1 (Left: Anchor | Middle: Positive | Right: Negative)



(b) Image triplet 2 (Left: Anchor | Middle: Positive | Right: Negative)

(c) Image triplet 3 (Left: Anchor | Middle: Positive | Right: Negative)



(d) Image triplet 4 (Left: Anchor | Middle: Positive | Right: Negative)

Figure 4.19: Examples of image triplets in the TMA-Sentinel Triplets dataset for ResNet-50 based network.

4.3.3. Training Process

Similar to SigNet, the framework of the ResNet-50 based network was also constructed using the Keras library with TensorFlow as the backend. The pre-trained ResNet-50 model is readily available in the Keras library, pre-trained on the ImageNet dataset. Hence, the model is directly applicable for adaptive training. During adaptive training, the model was optimized using the Adam optimizer with a learning rate of 10^{-4} for 500 epochs. The adaptive training process concluded in approximately 1 hour and 45 minutes. The training versus validation loss plot is illustrated in Fig. 4.20. Despite the apparent fluctuations in the validation loss, all fluctuations remained below 0.4, indicating their acceptability. Furthermore, both training and validation losses notably stabilized and converged to small values after 400 epochs, indicating the efficacy of the model.

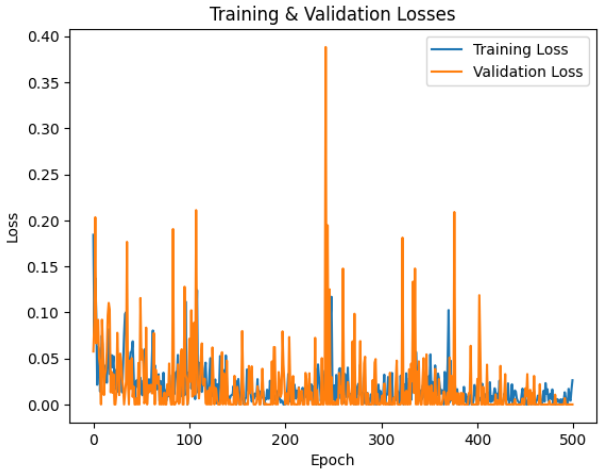


Figure 4.20: Training loss VS Validation loss for ResNet-50 based network

4.4. Heatmap Prediction Confidence Assessment

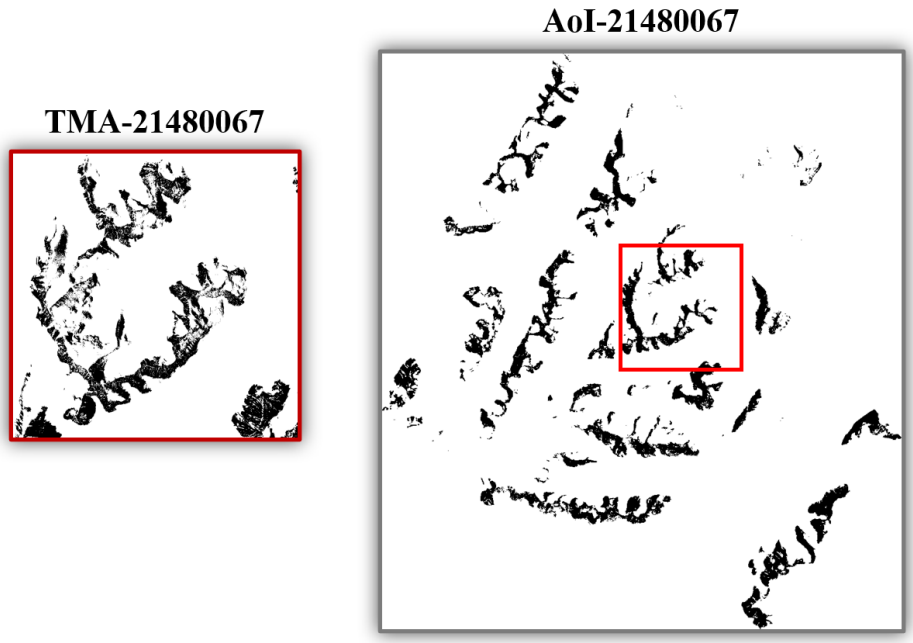
A heatmap is a graphical representation of data where the values of a matrix are represented as colors. Typically, each cell in the matrix corresponds to a specific data point, and the color intensity of the cell reflects the magnitude of the data value it represents. Heatmaps are commonly used to visualize the distribution, density, or intensity of data across a two-dimensional space.

In this study, the pixel values on the heatmaps depict similarity scores between the reference TMA image and the sub-images located at specific positions. Consequently, the hotspots in these heatmaps denote regions bearing the closest resemblance to the reference image, indicating the approximate locations of the reference TMA image. In essence, pixels with the highest similarity scores on a heatmap correspond to matched sub-images of the referenced TMA image. Accordingly, to pinpoint the location on the heatmap, the search initiates from the pixels exhibiting the highest similarity score, followed by a downward search. It is plausible to have one or multiple matched sub-images, considering that the same scene might appear across several neighboring sub-images. However, the accuracy of these hotspots in pinpointing the real location of the reference TMA image relies heavily on the performance of the similarity estimation models.

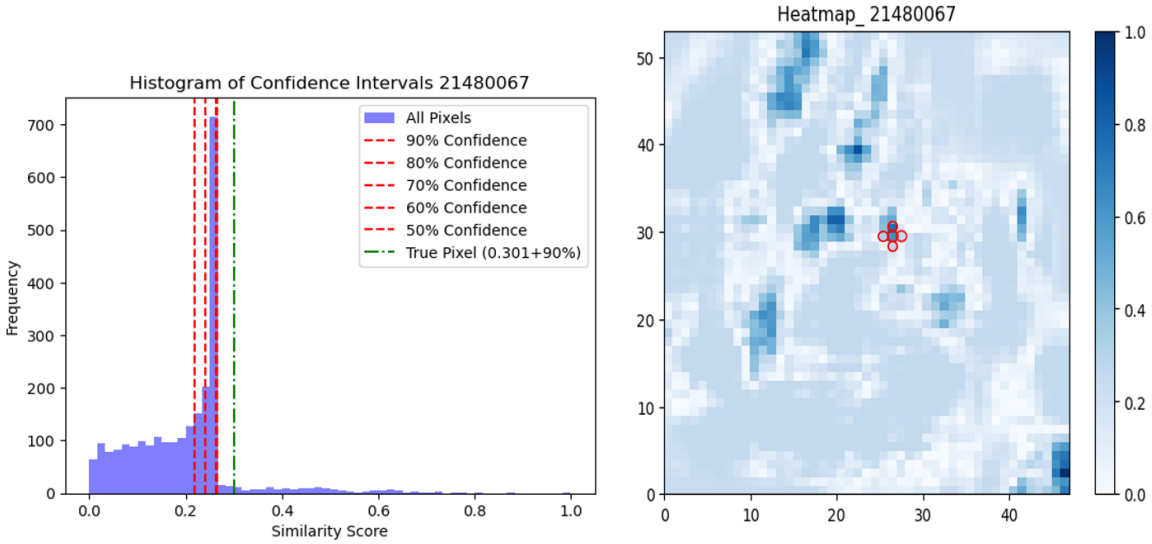
The precision of geo-localization of the reference TMA image is closely tied to the effectiveness of the models. Typically, evaluation metrics are employed to gauge model accuracy. However, due to the small dataset size, consisting of approximately 15 image triplets in the test set, the accuracy metric may not be reliable. An alternative approach to assess the performance of this geo-locating approach involves testing on the heatmaps. This evaluation method, referred to as heatmap prediction confidence assessment in this study, offers a more robust measure of performance.

The process involved in heatmap prediction confidence assessment begins with the selection of 51 TMA images as testing samples. For each image, a heatmap is generated within the workflow. An additional step is incorporated where, for every TMA image, a manual selection on the Area of Interest (AoI) is made, marking the true location of the TMA image. This serves as the ground truth for this geo-localization method. Then the pixels on the heatmap that correspond to the accurately matched sub-images of the TMA image is identified. Notably, the surrounding four pixels are also considered as ground truth sub-images because there is only a small shift between adjacent sub-images. The average similarity score of these four pixels are considered to be the similarity score of the TMA and ground truth sub-image pairs.

Comparing the average similarity scores of these true matched pixels with the others allows us to compute the percentage they represent. This percentage signifies the confidence level of the heatmap in accurately predicting the genuine location of the TMA image. By computing confidence levels for all testing TMA images and determining their average, the measure of confidence in the effectiveness of this geo-localization method is established. Below in Fig. 4.21 is an example illustrating the computation of one such confidence level.



(a) For TMA image 21480067 (left), we manually delineate an area within the corresponding AoI (right), marked by a red bounding box in the AoI illustration. This selected area might encompass one or several sub-images obtained using the sliding window method from the AoI.



(b) In the heatmap (right), the red circles indicate two pixels representing the sub-images within the selected red bounding box in the AoI. The histogram (left) shows an average similarity score of 0.301 for these two pixels, placing it within the top 90% of the highest similarity scores. This suggests a 90% confidence level in locating the correct position of the TMA image using this heatmap.

Figure 4.21: Example for confidence level computation.

However, it is important to provide specific clarification regarding the interpretation of a model's average confidence level: suppose a model's average confidence level is computed as 90% by the heatmap prediction confidence assessment. This means that, for any input TMA image, the sub-image with the true match will typically have a similarity score higher than that of approximately 90% of other sub-images in the heatmap generated by the model. However, this does not imply that the probability of the highest-scored hotspot in the heatmap being the true match is 90%. As mentioned earlier, after this study, it is necessary to combine image matching to systematically search through sub-images for a more precise match. Interpreted in this context, the 90% indicates that, on average, the true match will be found after

searching through approximately 10% of the sub-images, assuming the search starts from the sub-image with the highest similarity score. However, this is just an average value; for some TMA images, the true match may be found in the first sub-image searched, while for others, it may only be found in the last sub-image. Therefore, the 90% should be interpreted as follows: this model can reduce the search time for subsequent image matching by an average of 90%.

5

Results and Discussions

In this chapter, evaluation of the testing results of both the SigNet and ResNet-50 based networks are presented, supplemented with examples. The evaluation includes highlighting three favorable outcomes and three unfavorable outcomes for each model. However, only two unfavorable outcomes are illustrated for the ResNet-50 based model, as there were only two instances of unfavorable outcomes observed out of a total of 51 testing samples. Subsequently, a comparative analysis will be conducted to address various aspects of the models.

5.1. SigNet Binary Image Geo-localization

The SigNet model is tested on 51 TMA images. The testing images are selected from four different flight lines and are selected in the way that it is made sure that each image covers some rock structures. Specific image numbers can be found in table Table A.2. The heatmaps are evaluated by manually picking the ground truth from the Area of Interest (AoI) sub-images as described in chapter 4.4.

In this chapter, three examples demonstrating high confidence levels and three examples showcasing low confidence levels are presented. Systematic analysis is conducted to examine the reasons behind the high or low confidence levels observed in each example. This analysis aims to evaluate both the effectiveness and limitations of SigNet.

5.1.1. Overview of Results

The average confidence level of SigNet predictions on the 51 test samples is 70.7%, with a variance of 0.0439 across the confidence levels. The variance indicates the degree of dispersion or spread of the confidence levels around the average value. In other words, it quantifies the extent to which the confidence levels deviate from the average. Detailed results are provided in Table A.2 in the appendix. The average confidence level suggests that SigNet is delivering a moderate performance. While this variance alone cannot provide much information, it will be later used in comparison to the ResNet-50 based network. Subsequent discussions will delve into examining different cases to explore the factors influencing SigNet's overall performance.

As explained in Chapter 4.4, the 70.7% average confidence level does not imply that the pixel or sub-image with the highest similarity score on the heatmap has a 70.7% probability of being the true match. Instead, it indicates that, on average, the similarity score of the true match pixel or sub-image is higher than that of 70.7% of the pixels or sub-images. For any given TMA image case, the true match pixel or sub-image may have either a high or low similarity score.

5.1.2. High Confidence Level Cases

In this chapter, three cases exhibiting high confidence levels are presented and assessed to comprehend the strengths and weaknesses of SigNet.

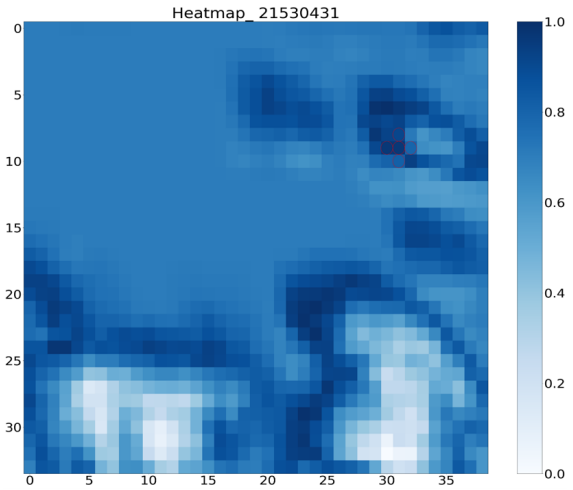
High confidence level case 1

Just as in this case, for each instance, the processed binary TMA reference image is presented alongside the processed binary Area of Interest (AoI), the corresponding heatmap, and a histogram illustrating the distribution of similarity scores. Subsequently, the heatmap is superimposed onto the AoI with a

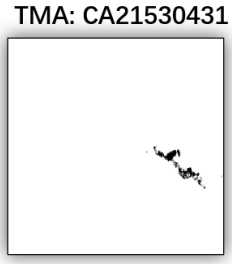
transparency setting of 50%. This enables a comprehensive visualization of the model's performance, revealing the magnitude of similarity scores at various positions within the AoI. This approach provides a nuanced understanding of the model's effectiveness at each location on the AoI, enhancing the overall assessment.

In the first case, the results are presented below in Fig. 5.1, where Fig. 5.1c is the pre-processed black and white TMA reference image (CA21530431), Fig. 5.1a, Fig. 5.1b, and Fig. 5.1d are the AoI (Area of Interest), heatmap, histogram of the heatmap, respectively, generated based on the TMA reference image. The confidence level of this case is 96%, meaning that the similarity score between the ground truth sub-image and the reference TMA image is higher than 96% of all the similarity scores.

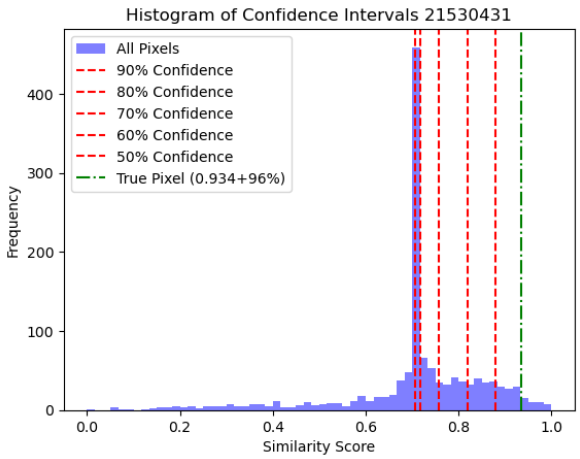
(a) Binary Area of interest (AoI) computed and cropped from the Quantarctica ADD Rock Mask dataset based on the reference TMA image.



(b) Heatmap indicating the similarity scores between the reference TMA image and sub-image on each location on the AoI image.



(c) TMA reference image which has been pre-processed and converted to binary format.



(d) Histogram of the heatmap, indicating the confidence level of the heatmap. The green line indicates where the similarity score of the true pixel stands among all similarit scores.

Figure 5.1: Result example: CA21530431

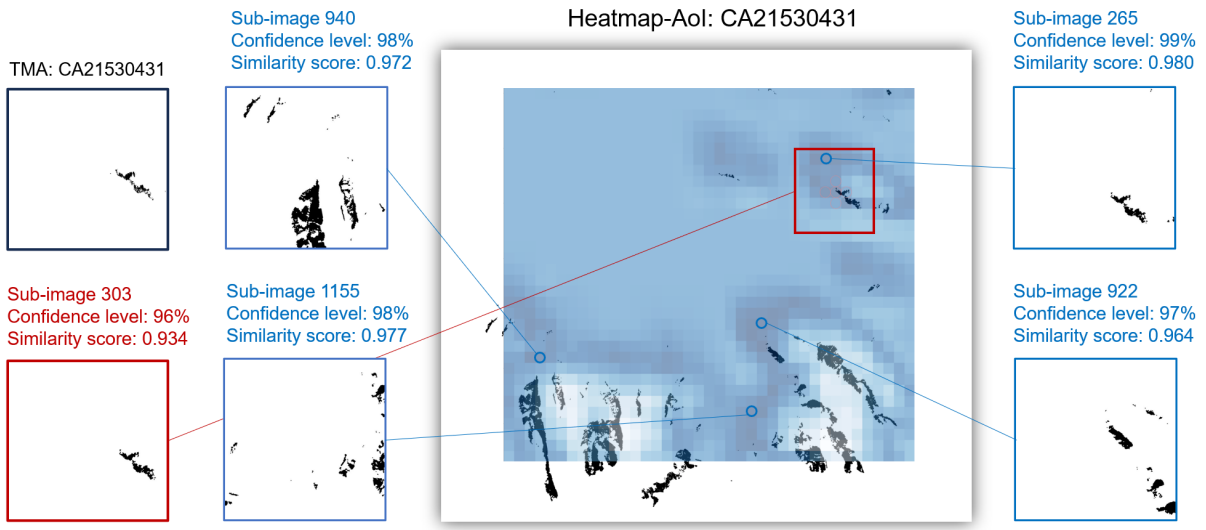


Figure 5.2: Demonstration of several false hotspots on the heatmap. The heatmap is overlaid on the AoI image with a transparency set to 50% to reveal both the heatmap and the underlying AoI. The red box delineates the coverage area of the manually selected ground truth sub-image. The blue circles represent sub-images (pixels) that, while not part of the ground truth, exhibit high similarity scores.

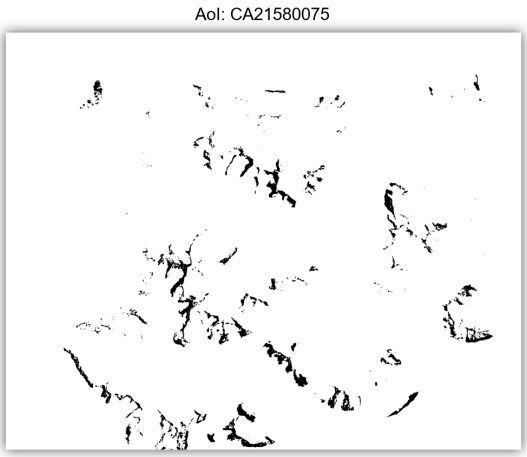
To comprehensively assess SigNet's performance in this case, four sub-images that exhibited higher similarity scores with the reference TMA than the ground truth sub-image are selected. The blue text above each sub-image indicates the similarity scores and the corresponding confidence levels. The ground truth sub-image, denoted as sub-image 303, is outlined by a red boundary line and located at the upper right corner on the Heatmap-AoI image. The designation '303' signifies that this is the 303rd sub-image generated by the sliding window method on the AoI. The same naming scheme applies to all other sub-images presented in this study. The similarity score between this ground truth sub-image 303 and the reference TMA image directly above it is 0.934, resulting in a confidence level of 96%. Upon visual inspection, a notable similarity between these two images is evident, especially in the characteristic black rock formations. Despite this similarity, approximately 4% of sub-images (53 in total) exhibit higher similarity scores than the ground truth sub-image 303. Four of these instances are selected for detailed discussion on why their similarity scores with the reference TMA image surpass that of the ground truth sub-image.

Firstly, sub-image 265 in the right top corner, prominently features the same rock structure as the reference TMA and the ground truth sub-image303. The only discernible difference lies in the rock's slightly lower position and more centralized placement within the image, justifying the model's attribution of a higher similarity score. Moving to sub-image 922 in the right bottom corner, the rock structures concentrate predominantly in the lower-right corner, extending diagonally to the upper left—resembling the pattern observed in the ground truth sub-image 303. Both sub-image 265 and sub-image 922 exhibit clear resemblances of their rock structures to the reference TMA image.

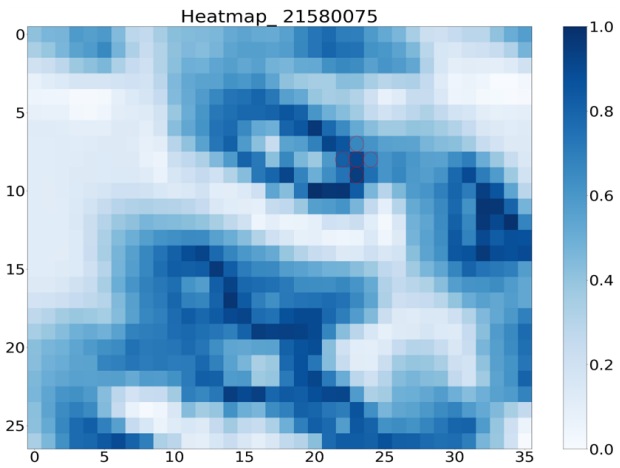
However, for sub-image 940 and sub-image 1155 situated to the left of the heatmap image, there is no apparent similarity in the rock structures to the reference TMA image. The high similarity scores for these sub-images may be attributed to the concentration of rock structures mostly on their lower right corners. Upon closer examination of the heatmap and the AoI images, it becomes evident that high similarity scores are predominantly observed along the left side of the rock structures, causing the rocks in sub-images to appear concentrated on the right side of the images.

High confidence level case 2

In Case 2, as depicted in Fig. 5.3, the confidence level reaches 90%. This indicates that the similarity score between the ground truth sub-image and the reference TMA image (measuring 0.807) surpasses that of most similarity scores. Theoretically, this implies a substantial reduction in search time, emphasizing the efficiency gained when relying on such a high confidence level.

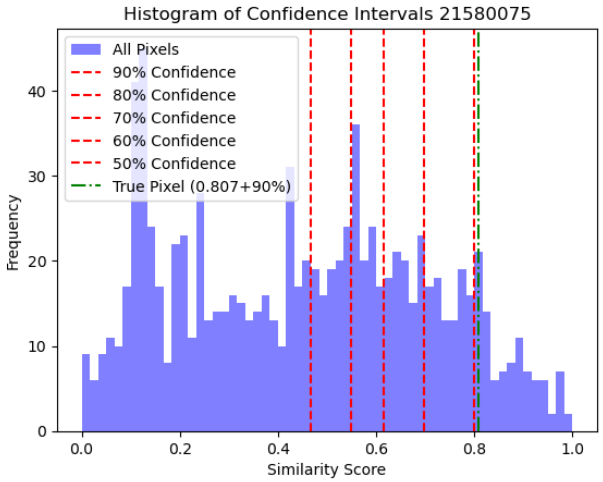


(a) Area of interest (AoI) of the TMA reference image.



(b) Heatmap indicating the similarity scores of each pixel.

(c) TMA reference image.



(d) Histogram of the heatmap, indicating the confidence level of the heatmap.

Figure 5.3: Result example: CA21580075

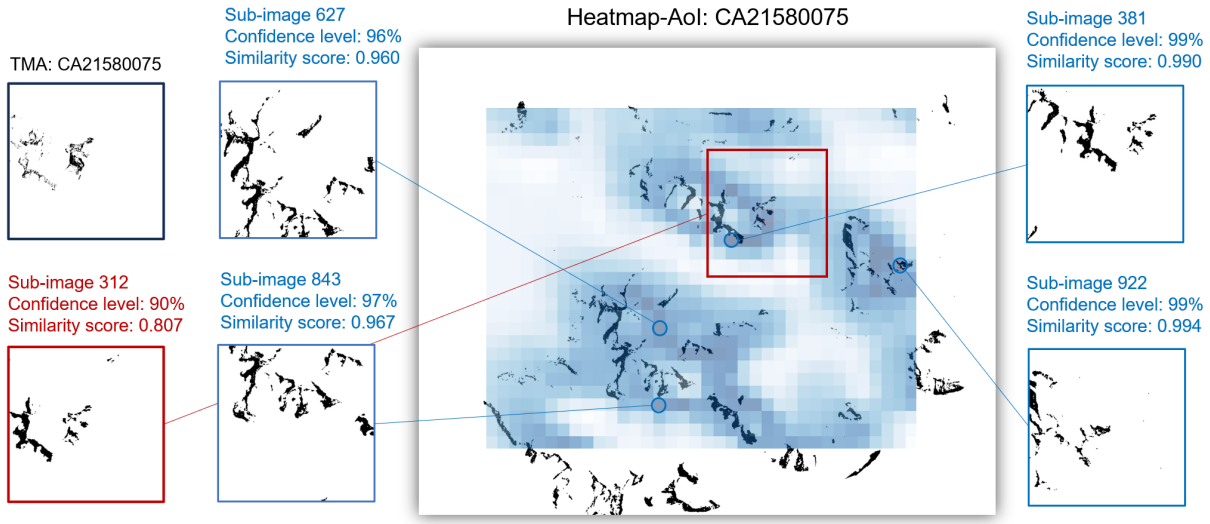


Figure 5.4: Demonstration of several false hotspots on the heatmap. The heatmap is overlaid on the AoI image with a transparency set to 50% to reveal both the heatmap and the underlying AoI. The red box delineates the coverage area of the manually selected ground truth sub-image. The blue circles represent sub-images (pixels) that, while not part of the ground truth, exhibit high similarity scores.

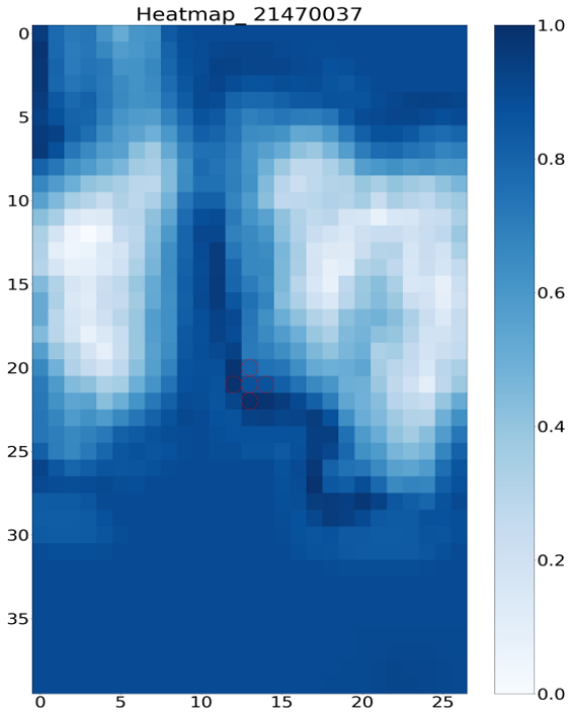
In this case, ground truth sub-image 312 achieves a significant similarity score of 0.807, with a confidence level of 90%. The rock structure within sub-image 312 closely resembles the reference TMA, displaying a semicircular shape with an upward opening and branches extending from it. This distinct rock formation recurs in all four falsely matched sub-images, with sub-image 381 exhibiting the same rock structure as seen in ground truth sub-image 312 and the reference TMA image.

Upon examining the Heatmap-AoI image at the center, it becomes apparent that the darker blue pixels predominantly cluster to the right side of the primary rock structures on the AoI. This observation aligns with the explanation provided for Case 1: given the concentration of rock structures on the left side of the reference TMA image, sub-images positioned to the right of the main rock structures result in the appearance of the rock structure on the left side in those sub-images. When the rocks align in comparable positions across images, there is a tendency for higher similarity scores.

High confidence level case 3

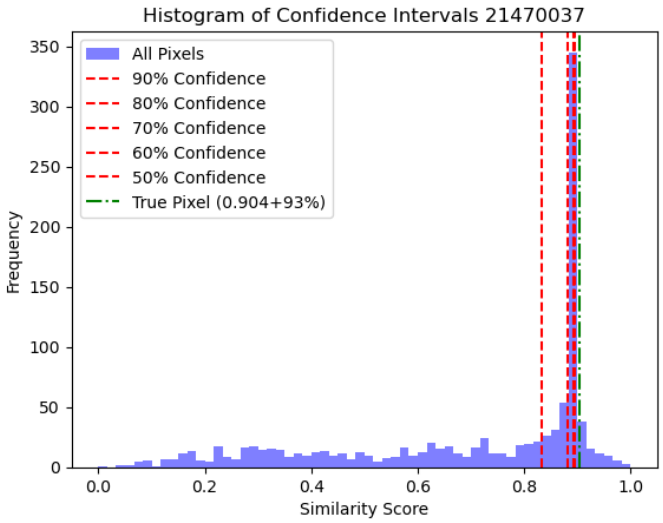
In Case 3, as illustrated in Fig. 5.5, the confidence level reaches 93%. This signifies that the similarity score between the ground truth sub-image and the reference TMA image, amounting to 0.904, exceeds that of 93% of all similarity scores. In essence, this suggests a prospective 93% reduction in search time, emphasizing the enhanced efficiency associated with a confidence level of 93%.

(a) Area of interest (AoI) of the TMA reference image.



(b) Heatmap indicating the similarity scores of each pixel.

(c) TMA reference image.



(d) Histogram of the heatmap, indicating the confidence level of the heatmap.

Figure 5.5: Result example: CA21470037

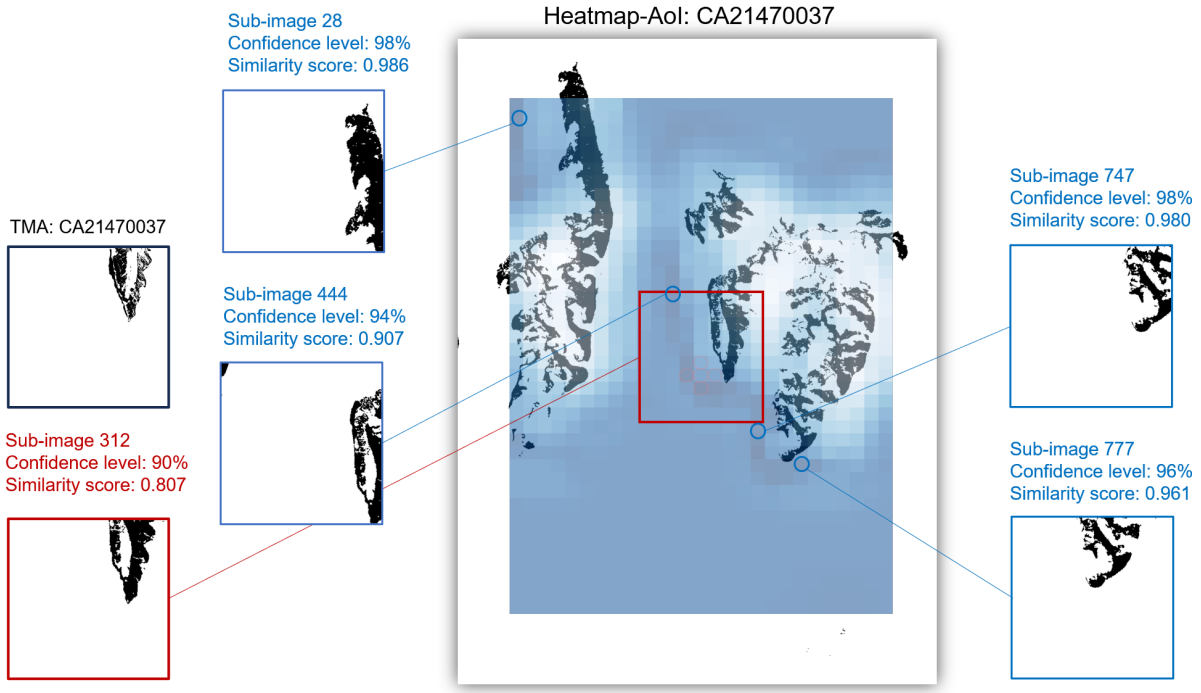


Figure 5.6: Demonstration of several false hotspots on the heatmap. The heatmap is overlaid on the AoI image with a transparency set to 50% to reveal both the heatmap and the underlying AoI. The red box delineates the coverage area of the manually selected ground truth sub-image. The blue circles represent sub-images (pixels) that, while not part of the ground truth, exhibit high similarity scores.

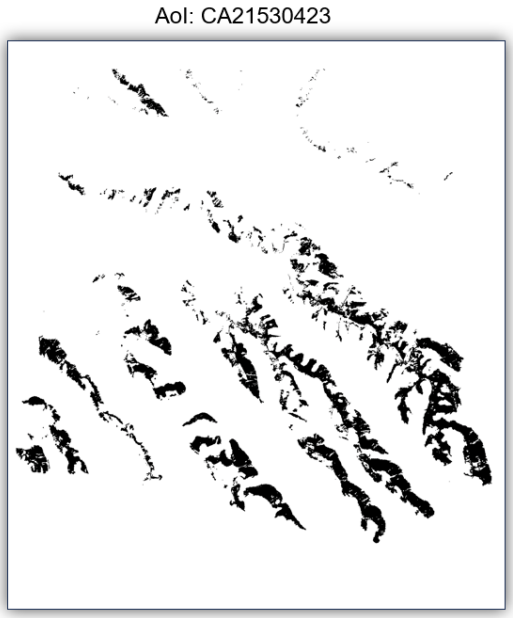
In Fig. 5.6, four non-ground truth sub-images with high similarity scores are presented. Firstly, it is observed in the reference TMA image that a rock with a downward-pointing tip and a central blank area is situated in the upper right portion of the image. In sub-image 312, the same rock appears in a visually similar position. The distinct edges and relatively simple shape of this rock suggest that it is easily recognizable by the model. However, approximately 10% of other sub-images exhibit similarity scores higher than that of the ground truth sub-image 312 and the reference TMA image. Four such sub-images are highlighted within blue rectangles. Among them, sub-image 747 and sub-image 777 contain the same rock structure, exhibiting some resemblance to the target rock: both have a downward-pointing tip and a central hollow area, situated in the upper right portion of the sub-images, mirroring the configuration in the reference TMA image. On the other hand, for sub-image 28 and sub-image 444, the rocks do not exhibit an obvious similarity in shape to the reference TMA. One conceivable explanation for the high similarity scores in these two sub-images could be the rightward positioning of rocks in these images. From the Heatmap-AoI image (or Fig. 5.5b), it can also be observed that pixels with high similarity scores are predominantly located in the lower-left area of the rock region, once again indicating the model's sensitivity to the spatial positioning of rocks in images.

5.1.3. Low Confidence Level Cases

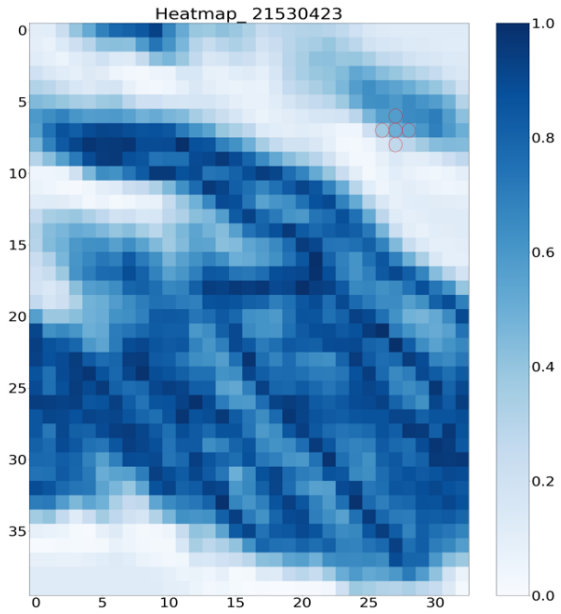
In this chapter, three cases that yielded relatively low confidence levels are introduced. Similar to the approach taken for cases with high confidence levels, the examination focuses on sub-images with the highest similarity scores to understand the factors behind these low confidence levels.

Low confidence level case 1

The first case with a low confidence level is depicted in Fig. 5.7. The similarity score between the ground truth sub-image and the reference TMA image is 0.444, accompanied by a notably low confidence level of 35%. Subsequently, the factors contributing to this low confidence level are investigated.

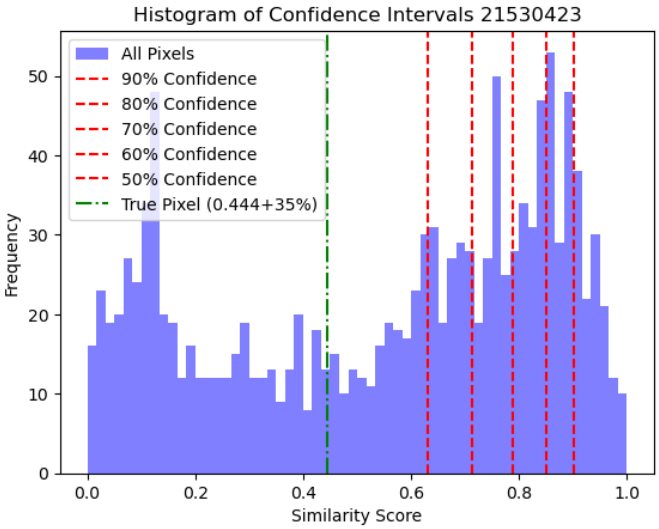


(a) Area of interest (AoI) of the TMA reference image.



(b) Heatmap indicating the similarity scores of each pixel.

(c) TMA reference image.



(d) Histogram of the heatmap, indicating the confidence level of the heatmap.

Figure 5.7: Result example: CA21530423

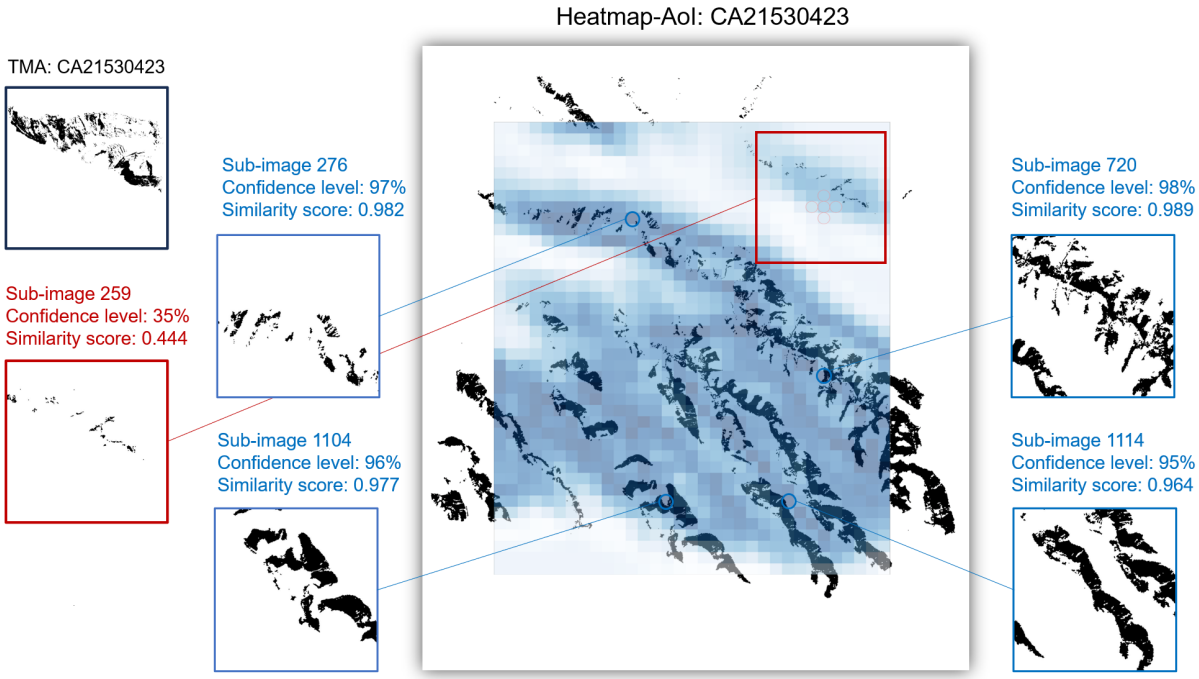


Figure 5.8: Demonstration of several false hotspots on the heatmap. The heatmap is overlaid on the AoI image with a transparency set to 50% to reveal both the heatmap and the underlying AoI. The red box delineates the coverage area of the manually selected ground truth sub-image. The blue circles represent sub-images (pixels) that, while not part of the ground truth, exhibit high similarity scores.

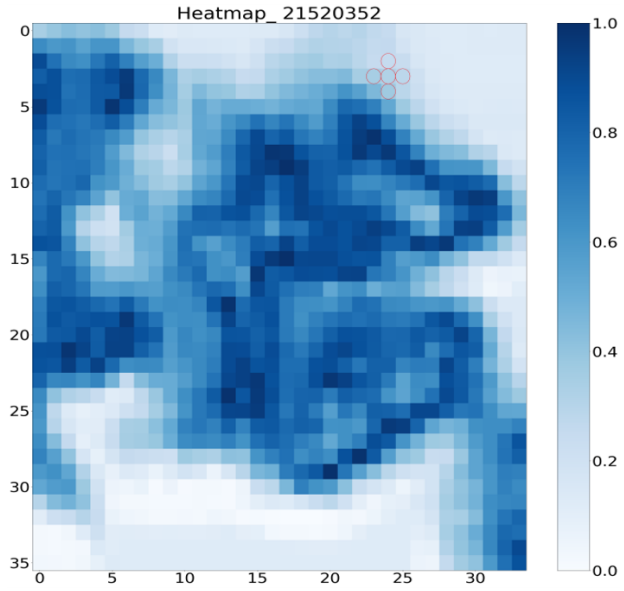
In Fig. 5.8, the shape of the rock in the reference TMA gradually converges from the lower right to the upper left, forming an approximate angle of 30 degrees, and culminating in a pointed end at the left edge of the image. However, in the ground truth sub-image 259 within the Area of Interest (AoI), there is no distinct representation of this rock shape; instead, there are scattered instances of rocks. From a visual perspective, sub-image 259 appears dissimilar to the reference TMA, even though it indeed occupies the same location in the AoI. This dissimilarity can account for the lower similarity score (0.444) and the associated low confidence level.

Upon examining the four sub-images with high similarity scores and confidence levels, a commonality emerges: the orientation of rocks in these images is akin to that in the reference TMA. Specifically, rocks in sub-image 720, sub-image 1104, and sub-image 1114 are predominantly located in the upper right part of the images. Although the rock in sub-image 276 is not positioned in the upper right, its shape closely resembles that of the reference TMA's rock. This observation suggests that the SigNet model is not only highly sensitive to the spatial positioning of rocks in images but also exhibits discernment regarding the shape and orientation of rocks.

Low confidence level case 2

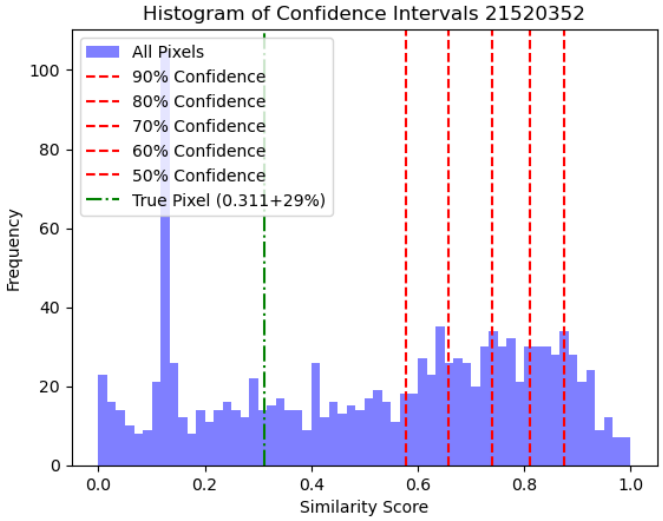
In this instance, depicted in Fig. 5.9, the similarity score between the ground truth sub-image and the reference TMA image is notably low at 0.311, with a confidence level of 29%. The reasons for this low similarity score and confidence level will be explored later in the discussion.

(a) Area of interest (AoI) of the TMA reference image.



(b) Heatmap indicating the similarity scores of each pixel.

(c) TMA reference image.



(d) Histogram of the heatmap, indicating the confidence level of the heatmap.

Figure 5.9: Result example: CA21520352

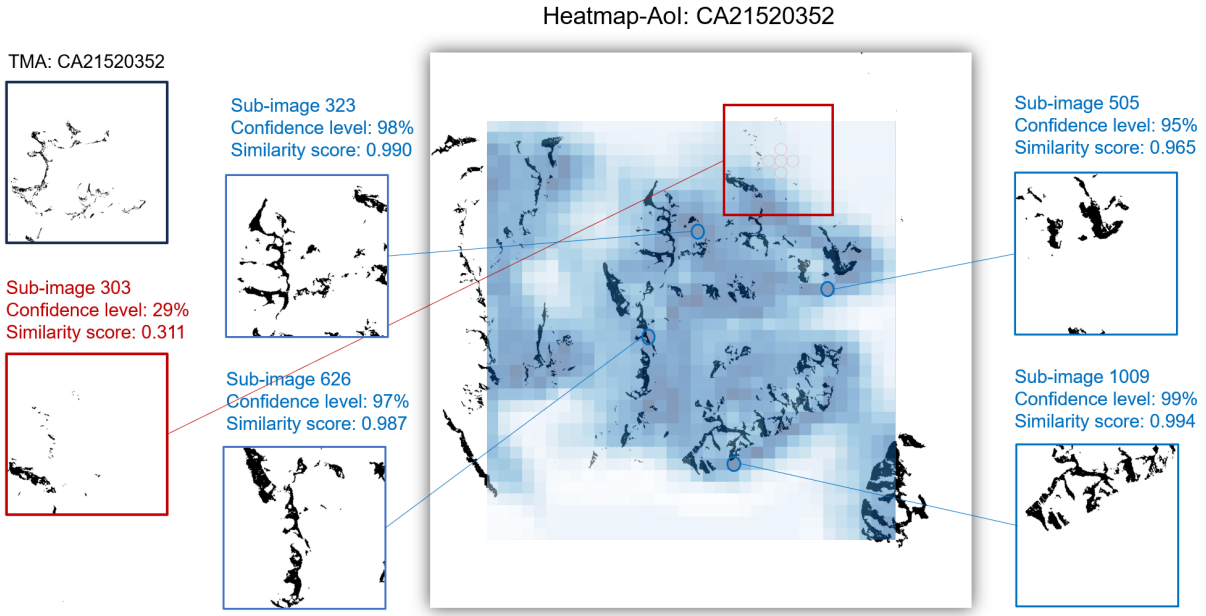


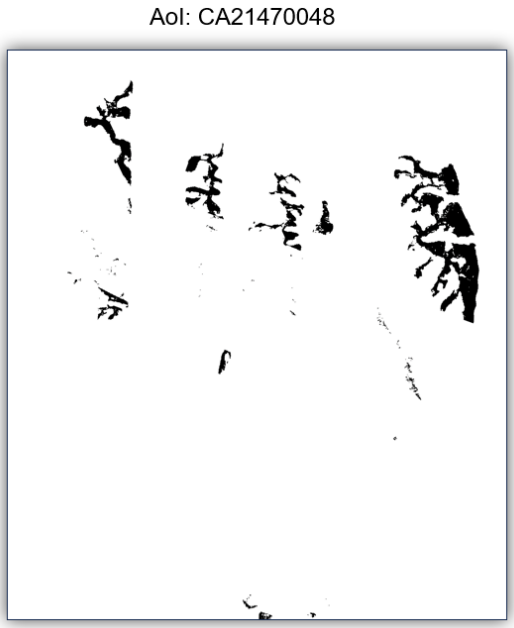
Figure 5.10: Demonstration of several false hotspots on the heatmap. The heatmap is overlaid on the AoI image with a transparency set to 50% to reveal both the heatmap and the underlying AoI. The red box delineates the coverage area of the manually selected ground truth sub-image. The blue circles represent sub-images (pixels) that, while not part of the ground truth, exhibit high similarity scores.

In this illustration, the reference TMA image showcases a rock structure extending from the lower left corner. The main body of this rock ascends vertically, with two branches extending to the right from the main trunk. The overall delineation of these rock structures lacks sharp clarity, and the representation of the rock in black pixels is somewhat subtle, potentially introducing challenges for the model. Upon comparison with the ground-truth sub-image 303, it becomes evident that the rock in this section is minimally identified in the rock mask. Only a few ambiguous rock formations are visible, with a horizontally extending rock structure in the lower left being relatively conspicuous. However, overall, discerning significant similarities between sub-image 303 and the rock structure in the reference TMA image proves challenging for the naked eye.

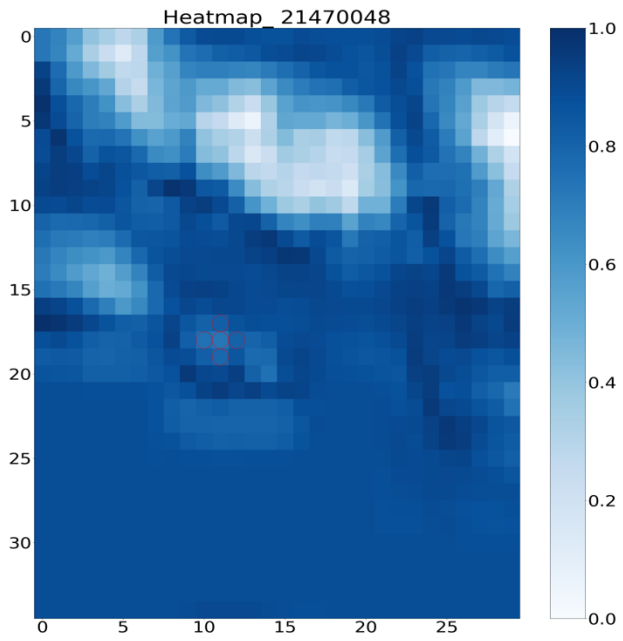
Regarding the four false match sub-images, notably in sub-image 323, a rock structure similar to that in the reference TMA image is observed. Both exhibit a vertically extending core from the lower left, accompanied by some horizontally extending branches. This resemblance elucidates the relatively high similarity score for sub-image 323. The other three sub-images do not show clear similarities. However, a plausible explanation could be considered: in these three sub-images, most parts of the rock structures are predominantly on the left side of the images, resembling the configuration in the reference TMA image. This sensitivity of the SigNet model to the spatial positioning of rocks in images may account for these observations.

Low confidence level case 3

In this example, illustrated in Fig. 5.11, the similarity score between the ground truth sub-image and the reference TMA image is 0.776, while the confidence level is recorded at 24%. Detailed discussions explaining the reasons behind this low confidence level are presented subsequently, accompanied by Fig. 5.12.

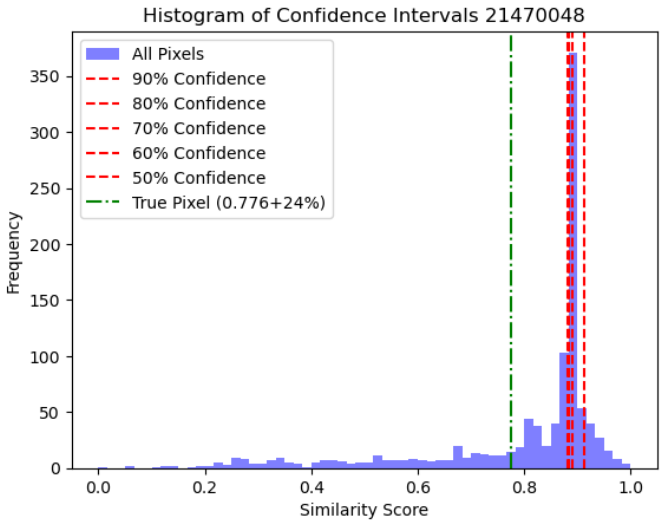


(a) Area of interest (AoI) of the TMA reference image.



(b) Heatmap indicating the similarity scores of each pixel.

(c) TMA reference image.



(d) Histogram of the heatmap, indicating the confidence level of the heatmap.

Figure 5.11: Result example: CA21470048

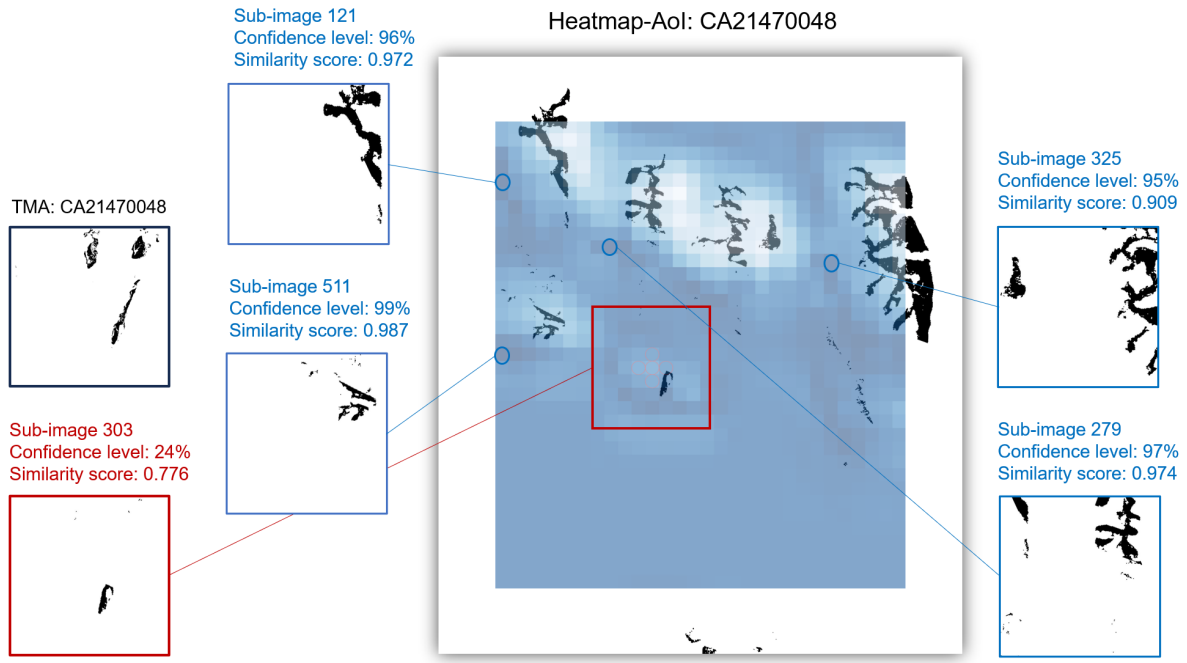


Figure 5.12: Demonstration of several false hotspots on the heatmap. The heatmap is overlaid on the AoI image with a transparency set to 50% to reveal both the heatmap and the underlying AoI. The red box delineates the coverage area of the manually selected ground truth sub-image. The blue circles represent sub-images (pixels) that, while not part of the ground truth, exhibit high similarity scores.

In Fig. 5.12, it is observed that in the reference TMA image, the rock structures are concentrated at the upper-right corner, featuring an elongated rock structure diagonally extending upwards. However, in the ground truth sub-image 303, most of the rock structures present in the reference TMA image are not annotated in the Quantarctica Rock Outcrop Mask dataset. Consequently, the similarity score between sub-image 303 and the reference TMA image is only 0.776, falling below 76% of the sub-images.

The four sub-images outlined with blue borders exhibit relatively high similarity scores primarily because the main rock structures in these sub-images are concentrated in the upper-right portion, similar to the configuration in the reference TMA image.

5.1.4. Evaluation and Adaptability Overview

In summary, across all six examples, it is evident that the SigNet model is sensitive to the general location of rocks in the sub-images, particularly highlighted in high confidence level case 3 (Fig. 5.6). However, the model does not exhibit significant sensitivity to the specific shapes of rocks. This tendency results in the model assigning high similarity scores to many sub-images where rocks are in the 'correct' positions, leading to challenges in achieving a consistently high average confidence level.

Another crucial factor contributing to the decline in model performance stems from the significant time gap between the capture of images in the Quantarctica Rock Outcrop dataset and historical TMA images. This temporal difference causes variations in snow cover, rendering some rocks unannotated in the rock mask and thereby preventing meaningful comparisons. This situation is particularly evident in the three cases with low confidence levels. Generally, the snow cover on the rock mask tends to be more extensive than that on the TMA images, presenting a challenge that SigNet cannot effectively address.

Furthermore, another phenomenon is observed across these cases. Sub-images containing entirely blank regions with no rock structure were assigned non-zero, and sometimes even relatively high, similarity scores. This resulted in large uniform blue regions in the heatmap, corresponding to the blank areas in the Area of Interest (AoI) images. From a human logic perspective, this seems illogical. However, for this model, during the initial training with the Hindi signature dataset and the subsequent adaptive training with the binary TMA-Rock Mask dataset, image triplets did not include areas completely devoid of rock. As a

result, the model lacks the capability to handle entirely blank images.

In conclusion, the predictive ability of the SigNet model for similarity in this study, adapted through training with binary TMA and rock masks, is not entirely reliable. Although it demonstrates an average reduction in search time by 70.7% in subsequent image matching processes, the notable variance in confidence levels suggests a substantial range of differences in reduced search times for individual cases. This variability introduces uncertainty, making it challenging to consistently gauge the model's effectiveness. Furthermore, given the computational time needed for heatmap generation, there are instances where the search time might even increase, emphasizing the need for careful consideration when implementing this approach.

5.1.5. Computational Efficiency

The computational time for pre-training and adaptive training for SigNet is detailed in chapter 4.2.3. Here, the computational time required for predicting similarity scores between the TMA image and each sub-image cropped from the Area of Interest (AoI) image is presented under two different configurations.

The first configuration involves utilizing only the CPU with a system RAM of 12.7GB. Under this configuration, it takes approximately 0.4 seconds to predict the similarity score for one sub-image. Given that the number of generated sub-images for different AoI images may vary from around 600 to 2000, the total prediction time ranges from four to fourteen minutes to predict for all the sub-images on one AoI image and generate heatmaps.

The second configuration employs the T4 GPU provided by Google Colab, which boasts a higher RAM of 51 GB. Under this configuration, it takes around 0.07 seconds to predict the similarity score for one sub-image. Consequently, the total prediction time required for creating a heatmap for an AoI image may vary from 42 seconds to 1 minute and 20 seconds. This approach significantly outperforms the CPU-only configuration.

Generally, considering the widespread availability of GPUs in academic environments, the prediction time for SigNet is deemed acceptable.

5.2. ResNet-50 based Grayscale Image Geo-localization

The ResNet-50 based model underwent evaluation using the identical test set, consisting of 51 TMA images as used for SigNet. The evaluation methodology outlined in chapter 4.4 was also applied to analyze the test results.

The assessment of the results follows the same procedures as those applied to the SigNet results. Three cases with high confidence levels and two cases with low confidence levels are presented to discuss potential reasons for the effective or suboptimal performance of the ResNet-50 based model.

5.2.1. Overview of Results

Across the 51 test samples, the model exhibited an impressively high average confidence level of 95.5%, accompanied by a low variance of 0.00393 (for SigNet the variance is 0.0439). Notably, the confidence level achieved by the ResNet-50 based model surpasses that of the SigNet model, which recorded a confidence level of 70.7%. This elevated confidence level indicates that the ResNet-50 based model can effectively contribute to reducing search time for image matching in the geo-referencing of TMA images.

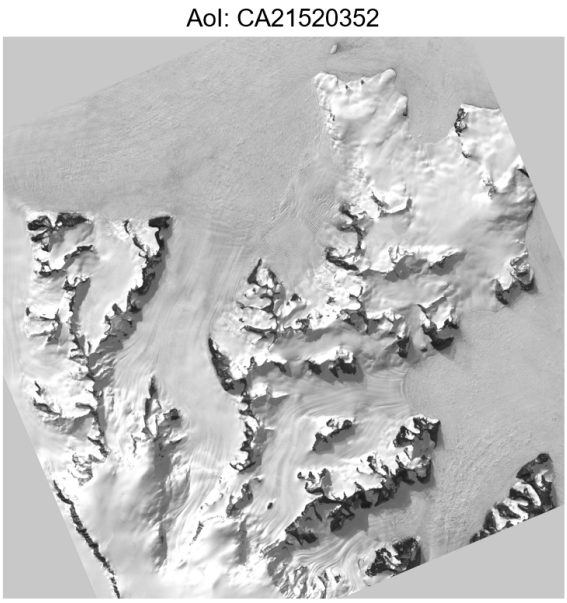
However, despite the high average confidence level, there are instances where the confidence level is relatively low. In the subsequent analysis, both successful and less optimal results will be assessed. The evaluations will be presented with a similar arrangement as in the SigNet analysis: firstly, the figures of the Area of Interest (AoI), heatmap, reference TMA, and histogram will be presented together to provide an overview of the model's performance in each specific case. Subsequently, there will be a demonstration of several falsely matched hotspots, including the overlay of the heatmap on the AoI image.

5.2.2. High Confidence Level Cases

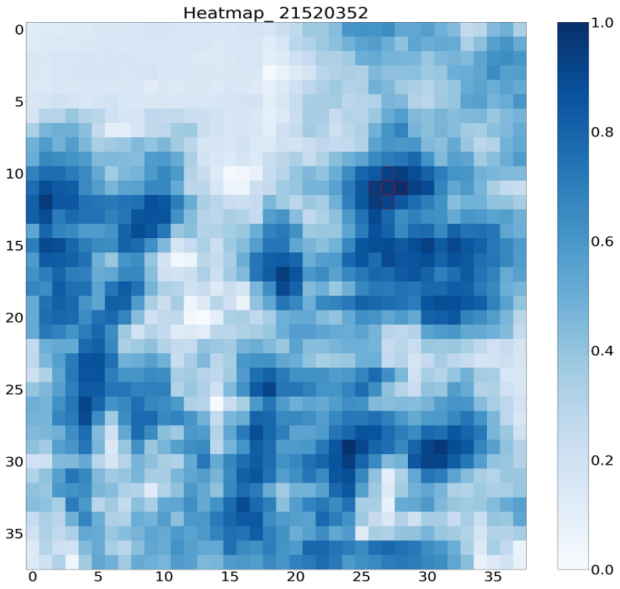
In this section, three cases with high confidence levels are showcased to highlight the robustness of the model. The first two cases correspond to the same reference TMA images as those in the low confidence level case for SigNet, facilitating direct comparisons. This comparative analysis reveals that instances which yielded low confidence levels with the SigNet model exhibit notably positive outcomes when evaluated by the ResNet-50 based model.

High confidence level case 1

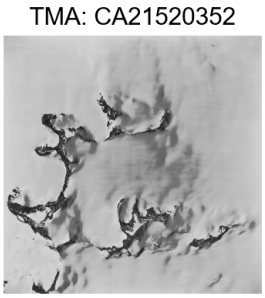
In this scenario, as depicted in Fig. 5.13, the similarity score between the reference TMA image and the ground truth sub-image stands at 0.955, accompanied by a confidence level of 99%, the highest possible confidence level. Nevertheless, there exist some sub-images that yield high similarity scores, some even higher than this ground truth sub-image. Four such sub-images are presented below in Fig. 5.14, and an assessment will be conducted to understand the reasons behind this phenomenon.



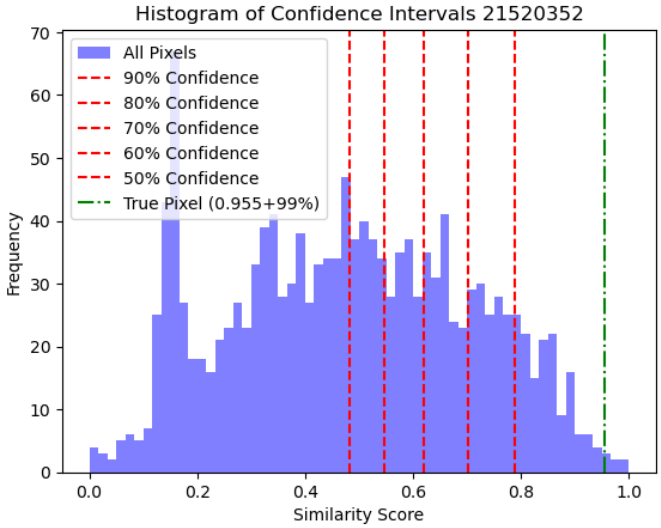
(a) Area of interest (AoI) of the TMA reference image.



(b) Heatmap indicating the similarity scores of each pixel.



(c) TMA reference image.



(d) Histogram of the heatmap, indicating the confidence level of the heatmap.

Figure 5.13: Result example: CA21520352

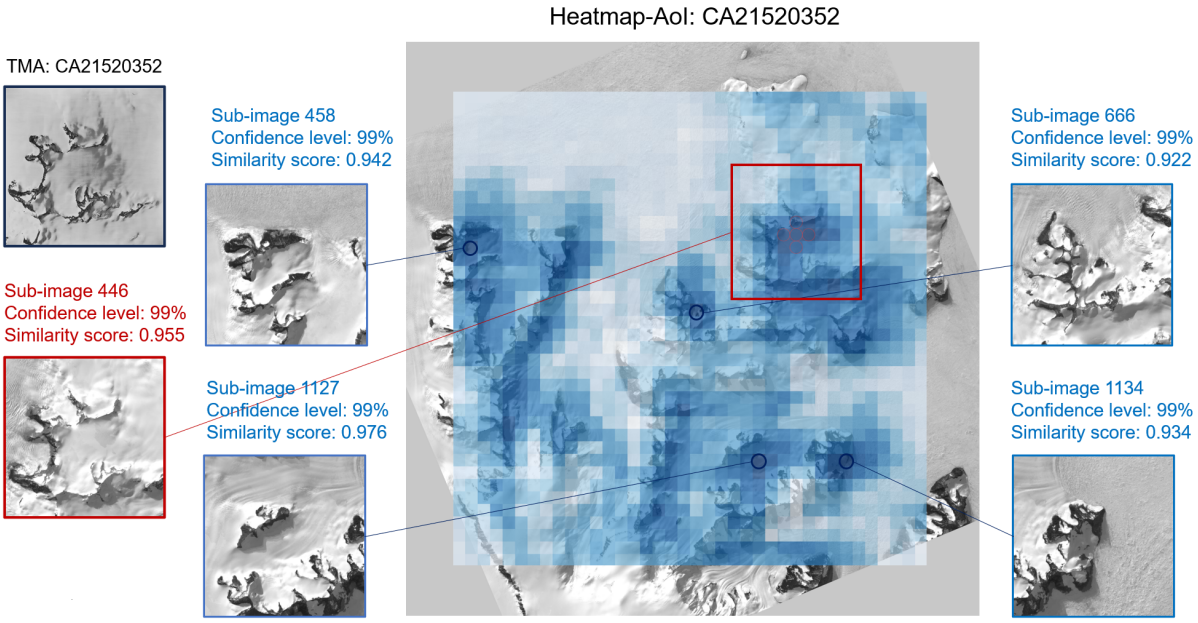


Figure 5.14: Demonstration of several false hotspots on the heatmap. The heatmap is overlaid on the AoI image with a transparency set to 50% to reveal both the heatmap and the underlying AoI. The red box delineates the coverage area of the manually selected ground truth sub-image. The blue circles represent sub-images (pixels) that, while not part of the ground truth, exhibit high similarity scores.

As illustrated in Fig. 5.14, the reference TMA image in the top-left corner exhibits a rock structure with distinctive features. The characteristic shape of this rock has been previously discussed in the description of the rock structure in the reference TMA image shown in Fig. 5.10. However, for the ResNet-50 based model, both the Area of Interest (AoI) and the TMA image are processed grayscale images, unlike the binary nature of the SigNet model's images.

Upon comparing these two reference TMA images, derived from the same original TMA image but subject to different processing procedures, it becomes evident that the grayscale TMA image produced by the ResNet-50 model displays a more complete representation of the rock structure. The edges are also clearer, and notable features such as snow textures and shadows are discernible, characteristics not observable in the binary images generated by the SigNet model.

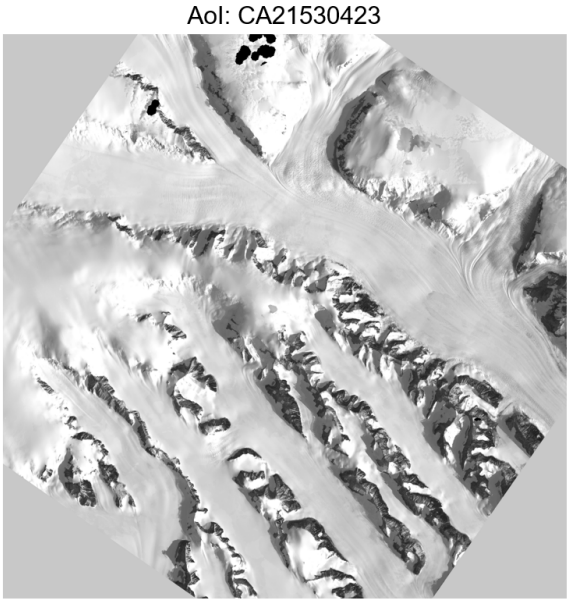
The ground truth sub-image 446 is presented beneath the TMA image. Through visual inspection, it is evident that the rock structures depicted in both images are identical. Although, in sub-image 446, the overall color tone may not closely resemble that of the TMA image, and it appears that, possibly due to the angle of the photograph, the rock in sub-image 446 is slightly horizontally elongated compared to the rock in the TMA image. Nevertheless, even with these differences, the ResNet-50 based model successfully identifies a high degree of similarity between sub-image 446 and the TMA image.

In each of the four falsely matched sub-images, a discernible resemblance to the rock structure in the reference TMA image is observed. Firstly, in sub-image 1127 with the highest similarity score, both upper and lower sections exhibit horizontally elongated rock structures. The smaller rock above combines with the larger one below to form a shape highly reminiscent of the two rock branches in the reference TMA image. The elevated similarity score, compared to the ground truth sub-image, may be attributed to the relatively brighter region in the ground truth sub-image 446, caused by reflections, resulting in an overall brighter tone. Consequently, sub-image 1127 may appear more similar in terms of color tone.

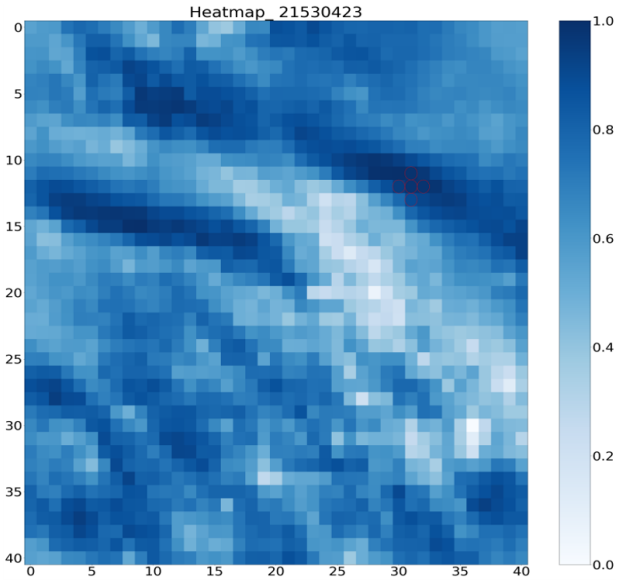
Furthermore, in sub-images 458 and 666, rock structures closely resemble those in the reference TMA, featuring two horizontally extending rock branches. In contrast, sub-image 1134 may seem less similar to the reference TMA image upon visual inspection. However, considering its overall color tone, which closely aligns with the reference TMA image, one can understand why the model assigned a relatively

high similarity score, even though the perceived visual similarity may be limited compared to the other three sub-images.

High confidence level case 2
In this instance, as depicted in Fig. 5.15, the similarity score between the reference TMA image and the ground truth sub-image is 0.958, corresponding to a confidence level of 98%.

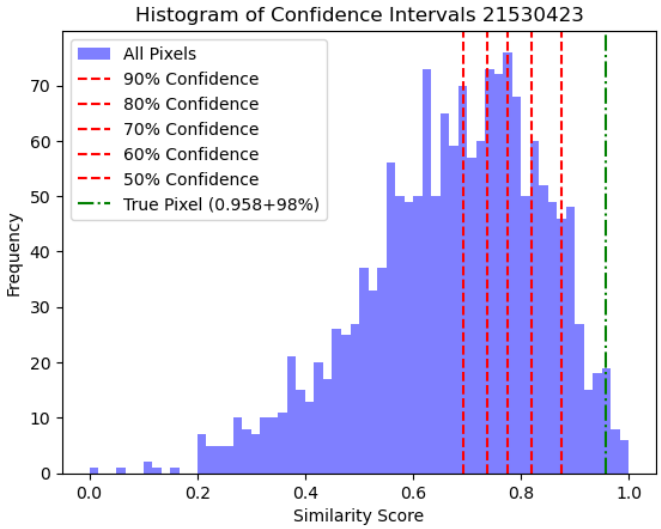


(a) Area of interest (AoI) of the TMA reference image.



(b) Heatmap indicating the similarity scores of each pixel.

(c) TMA reference image.



(d) Histogram of the heatmap, indicating the confidence level of the heatmap.

Figure 5.15: Result example: CA21530423

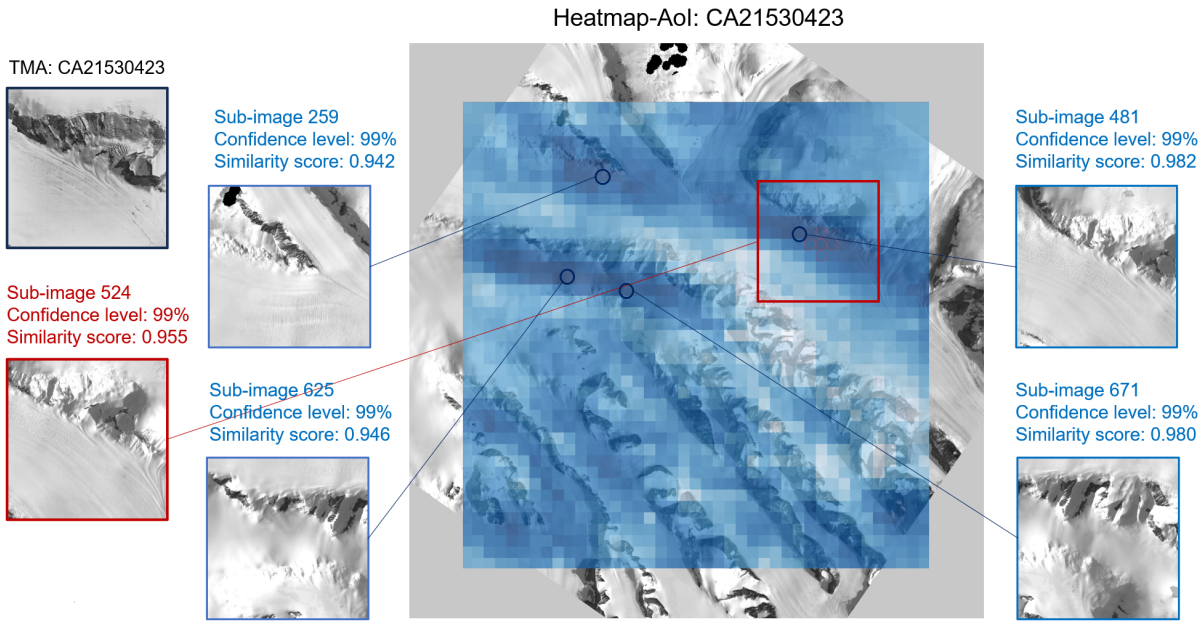


Figure 5.16: Demonstration of several false hotspots on the heatmap. The heatmap is overlaid on the AoI image with a transparency set to 50% to reveal both the heatmap and the underlying AoI. The red box delineates the coverage area of the manually selected ground truth sub-image. The blue circles represent sub-images (pixels) that, while not part of the ground truth, exhibit high similarity scores.

As illustrated in Fig. 5.16, the reference TMA image in the top-left corner exhibits a rock structure concentrated in the upper half of the image. The specific shape appears to narrow gradually from the bottom right to the top left. However, upon closer examination, it becomes apparent that a portion of the dark area in the upper-left section of the rock structure is, in fact, a shadow rather than rock material. This issue has the potential to impact the predictive performance of the model. In the corresponding black and white TMA image generated by SigNet (refer to Fig. 5.7c), some of these shadowed areas are incorrectly identified as rock, resulting in the display of black pixels. This misclassification contributes to the suboptimal performance of SigNet in this particular example.

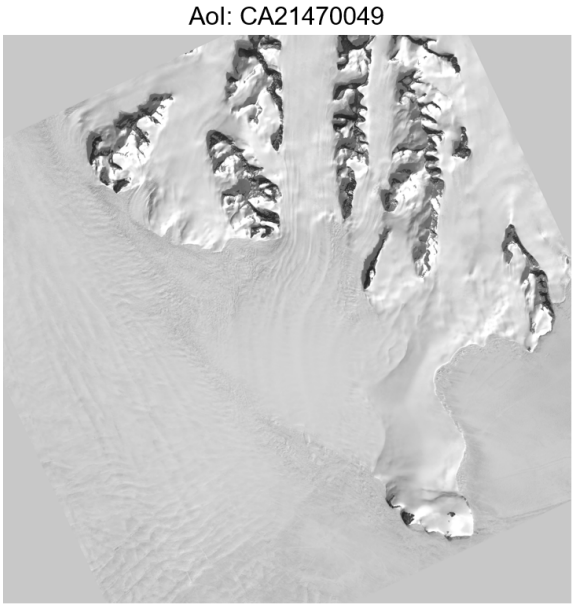
Below the reference TMA image is the manually selected ground truth sub-image 524. In this image, certain rocks that seem visible in the reference TMA image appear to be covered by snow here. However, due to the presence of shadows, a terrain similar to that in the reference TMA image is still discernible, and the texture of the snow in the lower-left part of the image also exhibits similarity.

Given the substantial time gap between the capture of historical TMA images and Sentinel-2 images, which can span several decades, significant changes in snow cover are inevitable, as evident in this example. Despite this variation, the ResNet-50 based model accurately identifies a high level of similarity between the two images. In contrast, examining sub-image 259 in Fig. 5.8, where almost no rocks are annotated, it is plausible that during the creation of the Quantarctica Rock Outcrop dataset, the rocks in this location were already covered by snow and hence not annotated. Consequently, the SigNet model struggles to make accurate predictions.

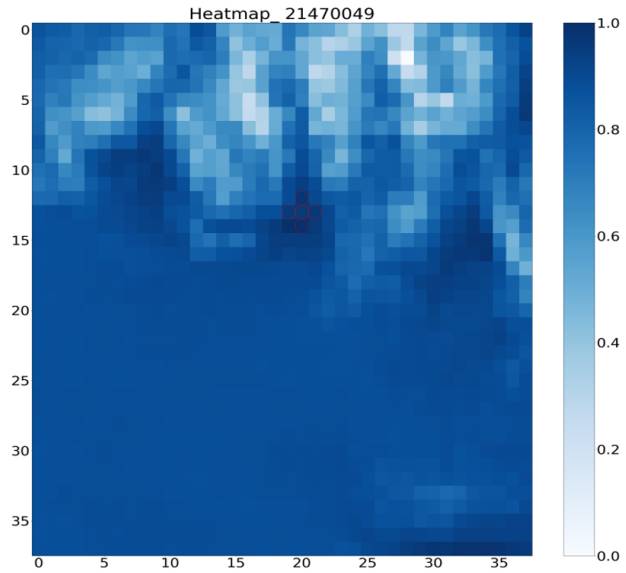
In almost all four falsely matched sub-images, rock structures that bear a striking resemblance to the reference TMA image can be observed, except for sub-image 259. In sub-image 259, one can understand why the model perceives it as similar to the reference TMA image. This is because the rocks in sub-image 259 also extend in a similar direction and are positioned above in the image. Therefore, overall, in this case, even though the rocks that should be visible in the ground truth sub-image 524 are covered by snow, the model demonstrates the ability to make reliable predictions by considering shadows, snow textures, and the limited exposed rock structures.

This indicates that, in this particular case, the ResNet-50 based model is capable of learning and utilizing information beyond bare rock, showcasing its ability to incorporate features such as shadows and snow textures into its predictions.

High confidence level case 3
In this case, as depicted in Fig. 5.17, the similarity score between the reference TMA image and the ground truth sub-image is 0.970, and the confidence level is 99%.

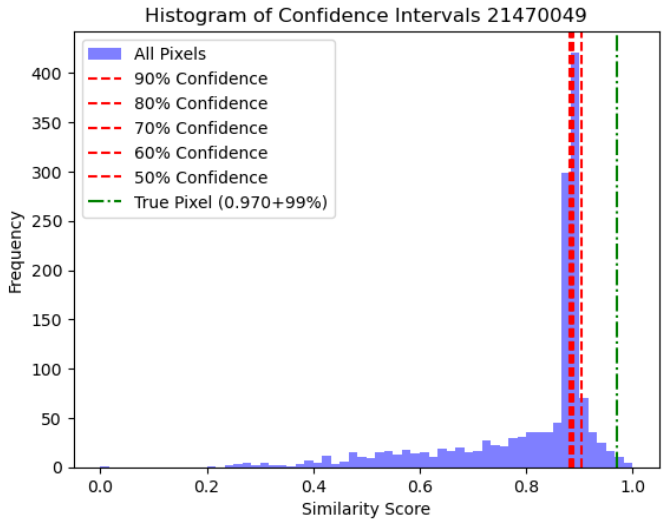


(a) Area of interest (AoI) of the TMA reference image.



(b) Heatmap indicating the similarity scores of each pixel.

(c) TMA reference image.



(d) Histogram of the heatmap, indicating the confidence level of the heatmap.

Figure 5.17: Result example: CA21470049

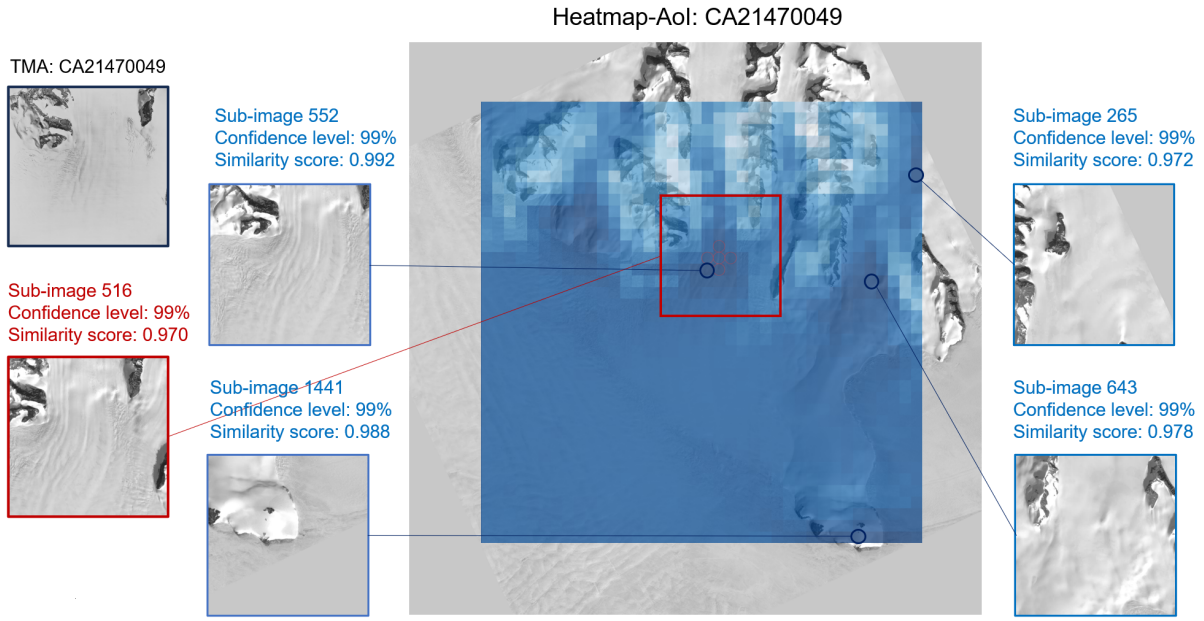


Figure 5.18: Demonstration of several false hotspots on the heatmap. The heatmap is overlaid on the AoI image with a transparency set to 50% to reveal both the heatmap and the underlying AoI. The red box delineates the coverage area of the manually selected ground truth sub-image. The blue circles represent sub-images (pixels) that, while not part of the ground truth, exhibit high similarity scores.

In Fig. 5.18, let us first examine the reference TMA image: it features two distinct rock structures. The leftmost rock is situated at the top-left corner of the image, while a smaller vertical rock is positioned at the top-right corner. In ground truth sub-image 516, slight positional shifts can be observed of these two rocks, resulting in the emergence of a new rock in the lower right area of the image, absent in the TMA image. Despite these positional deviations and the appearance of an additional rock in sub-image 516, the model correctly identifies the image as highly similar to the reference TMA.

Reflecting on the analysis of the testing results cases for the SigNet model, it is conceivable that the SigNet model may face difficulties in accurately predicting this example owing to notable positional disparities of the rocks. From this standpoint, although the ResNet-50 based model also takes into account rock positions as a feature, it does not excessively depend on this aspect. Instead, it simultaneously incorporates other features such as texture and shape.

Turning to the four falsely matched sub-images, firstly, for sub-image 552, it is positioned diagonally below ground truth sub-image 516. Consequently, sub-image 552 actually contains the same rocks as sub-image 516, although in different positions. Visually, the positions of the rocks in sub-image 552 seem more aligned with those in the TMA image, suggesting that sub-image 552 could have been a more suitable choice as the ground truth sub-image.

For sub-image 1441, a rock area in the top-left corner of the image is mostly covered by snow. Despite the snow cover, it can be discerned that the shape of this rock is similar to those in the TMA image, indicating that the model might have considered them to be the same rock, albeit covered in sub-image 1441.

In sub-image 643, the rocks exhibit noticeable similarities to those in the TMA image. The rock in the top-left corner features a complex and convoluted shape, while the one in the top-right corner has a simpler shape and clear outline.

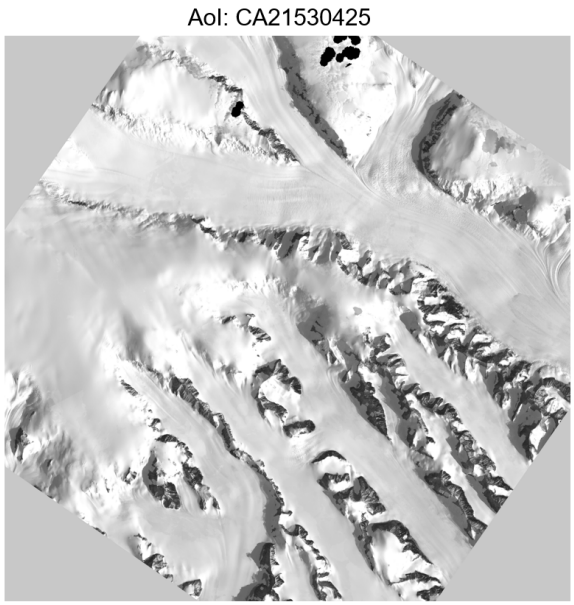
Lastly, for sub-image 265, the rocks are concentrated on the left side of the image. The leftmost rock bears a resemblance in shape to the rock in the TMA image, both featuring horizontal branches. Additionally, the vertical rock closer to the center in sub-image 265 might have been considered similar to the rightmost rock in the TMA image. This observation suggests that the ResNet-50 based model assigns less importance to the feature of rock position in the image compared to the SigNet model.

5.2.3. Low Confidence Level Cases

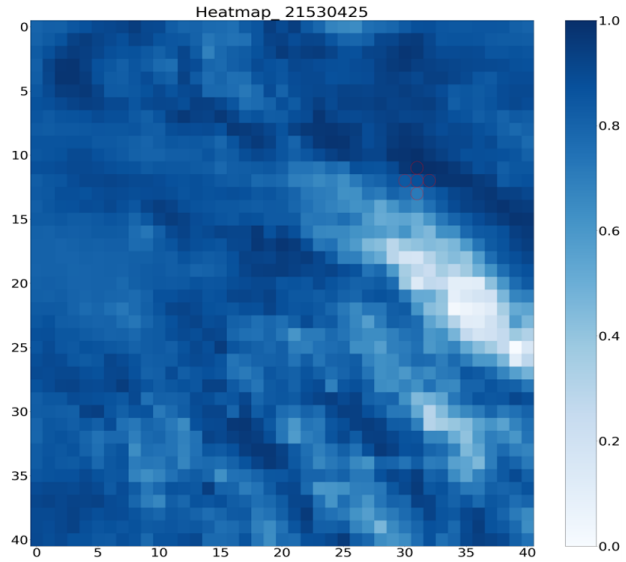
As shown in Table A.4 in the appendix, 46 out of the 51 test samples exhibited a confidence level exceeding 90%. Even for the five samples where the confidence level did not surpass 90%, their confidence levels were not notably low. This indicates the robustness of the ResNet-50 based model. In this section, the discussion and analysis focus on three cases with low confidence levels. The aim is to discern the reasons behind these occurrences and gain a deeper understanding of the limitations of the model.

Low confidence level case 1

In this instance, the similarity score between the reference TMA image and the ground truth sub-image stands at 0.875, with a confidence level of 67%. Notably, this case represents the lowest confidence level among all 51 testing results obtained with the ResNet-50 based model, although for SigNet, this result may be considered as moderate.

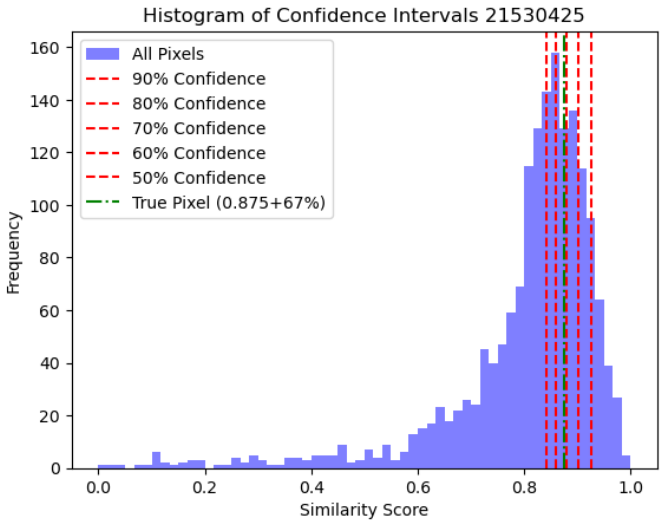


(a) Area of interest (AoI) of the TMA reference image.



(b) Heatmap indicating the similarity scores of each pixel.

(c) TMA reference image.



(d) Histogram of the heatmap, indicating the confidence level of the heatmap.

Figure 5.19: Result example: CA21530425

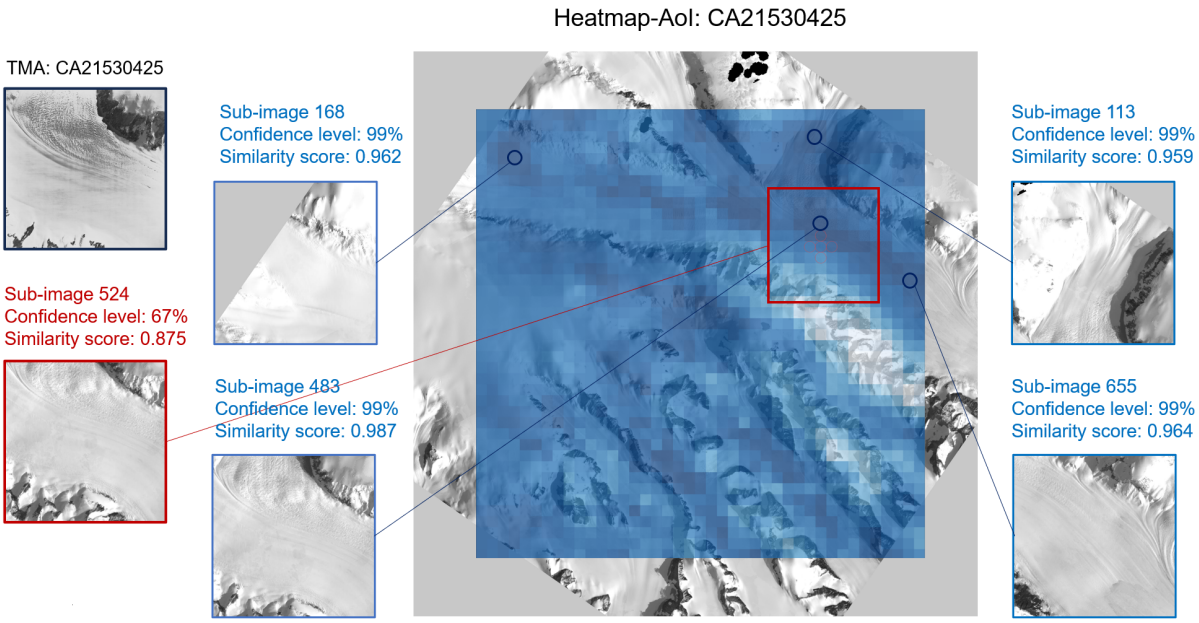


Figure 5.20: Demonstration of several false hotspots on the heatmap. The heatmap is overlaid on the AoI image with a transparency set to 50% to reveal both the heatmap and the underlying AoI. The red box delineates the coverage area of the manually selected ground truth sub-image. The blue circles represent sub-images (pixels) that, while not part of the ground truth, exhibit high similarity scores.

In Fig. 5.20, the examination begins with the reference TMA image positioned in the top-left corner. Notably, an arc-shaped rock occupies the top-right corner of the frame, while scattered rocks are dispersed in the bottom-left area. Additionally, pronounced shadows in the upper-right rock area contribute to a darker appearance. In contrast, in the ground truth sub-image 524, due to error in selecting the ground truth sub-image, the position of rocks in the image is overall shifted towards the upper-right, resulting in a reduction in the arc-shaped rock area in the top-right corner, while the distribution of rocks in the bottom-left corner increases. Consequently, rocks in the bottom-left corner appear more prominent in sub-image 524 than the rocks in the top-right corner. This alteration in the focal point of the image leads to a lower similarity between sub-image 524 and the reference TMA image, hence resulting in lower similarity scores and confidence levels. The less-than-ideal outcome in this example can primarily be attributed to errors made during the manual selection of the ground truth.

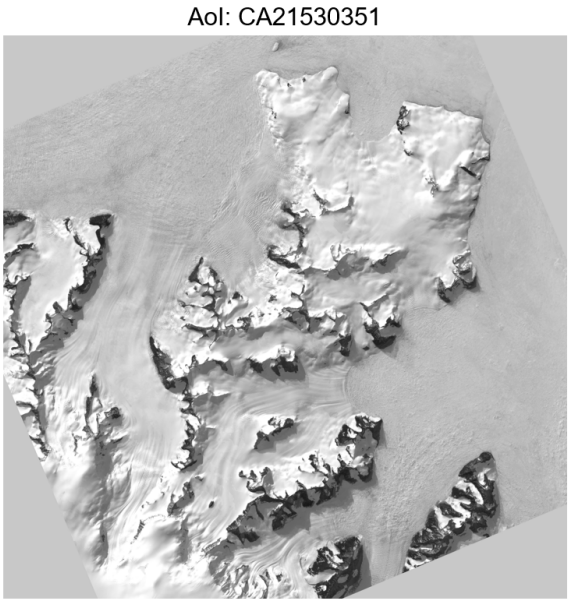
Moving on to sub-image 483, a similar pattern of rock formations can be observed, closely resembling those depicted in the TMA image. Although the rock in the top-right corner of sub-image 483 appears less dark due to the absence of shadow coverage, the model still discerns a strong similarity between the rock formations in sub-image 483 and those in the TMA image, mainly based on their shapes. While previous discussions acknowledged that the ResNet-50 based model may not exhibit strong sensitivity to the precise positioning of rocks in the image, the significant disparity in similarity scores between sub-image 483 and sub-image 524, relative to the TMA image, can be mainly attributed to several factors. These include the insufficient representation of the distinctive arc-shaped feature at the upper-right corner of the sub-image 524, while in sub-image 483 the feature is well extracted. Similarly, since more rocks are visible in the bottom-left corner of sub-image 524, the morphological patterns are clearer, resulting in different feature representations from those in the TMA images. Conversely, in sub-image 483, the number of visible rocks in the bottom-left corner is generally consistent with that in the TMA image, implying that the extracted features may not be as sophisticated or information-rich as those in sub-image 524.

In sub-image 655, rock shapes and arrangements closely resemble those depicted in the TMA image. Moving to sub-image 168, despite a missing portion in the top-left corner resulting from a rotation operation, discernible rock structures are faintly perceivable within the remaining expanse, with some rocks seemingly covered by snow. Nevertheless, the model still attributes higher similarity scores to this image. This suggests that the model possesses a certain level of capability to identify changes in snow coverage.

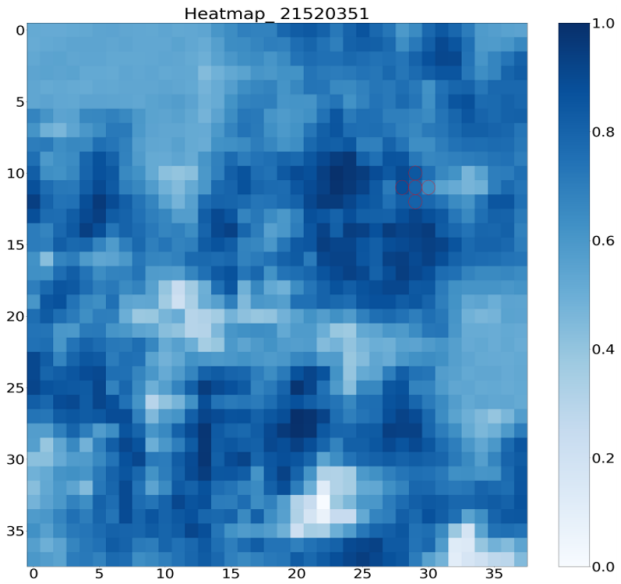
Sub-image 113 presents an intriguing phenomenon where an arc-shaped rock now resides in the bottom-right corner of the frame. However, visualizing a counterclockwise rotation of 90 degrees would reveal rock structures within the image, exhibiting shapes and positions more closely resembling those in the TMA image. Hence, it can be infered that the ResNet-50 based model is capable of accommodating rotated rocks within the image. This adaptability may have been acquired during the initial training utilizing ImageNet, as rotational transformations were systematically applied to all image triplets in the TMA-Sentinel 2 dataset during adaptive training, thereby ensuring consistent orientations of rock structures.

Low confidence level case 2

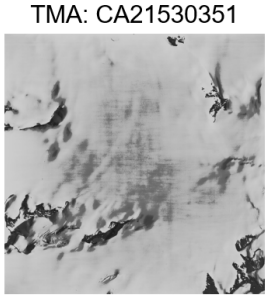
As depicted in Fig. 5.21, the similarity score between the reference TMA image and the ground truth sub-image in this instance is 0.796, accompanied by a confidence level of 71%. Although the confidence level is not excessively low, for the ResNet-50 based model, this performance is considered relatively poor.



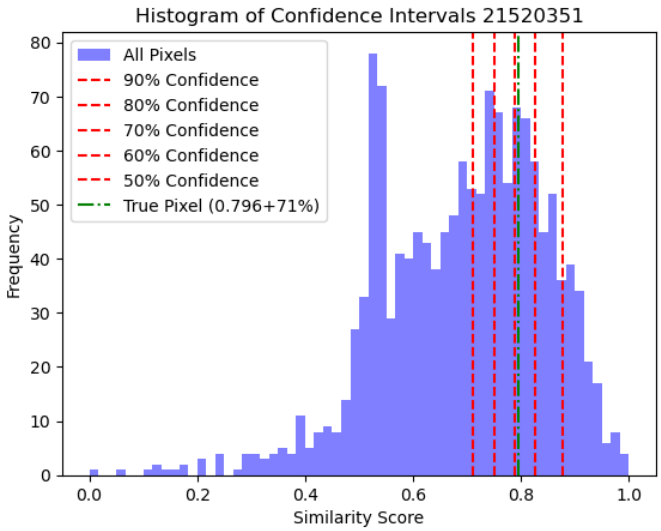
(a) Area of interest (AoI) of the TMA reference image.



(b) Heatmap indicating the similarity scores of each pixel.



(c) TMA reference image.



(d) Histogram of the heatmap, indicating the confidence level of the heatmap.

Figure 5.21: Result example: CA21530351

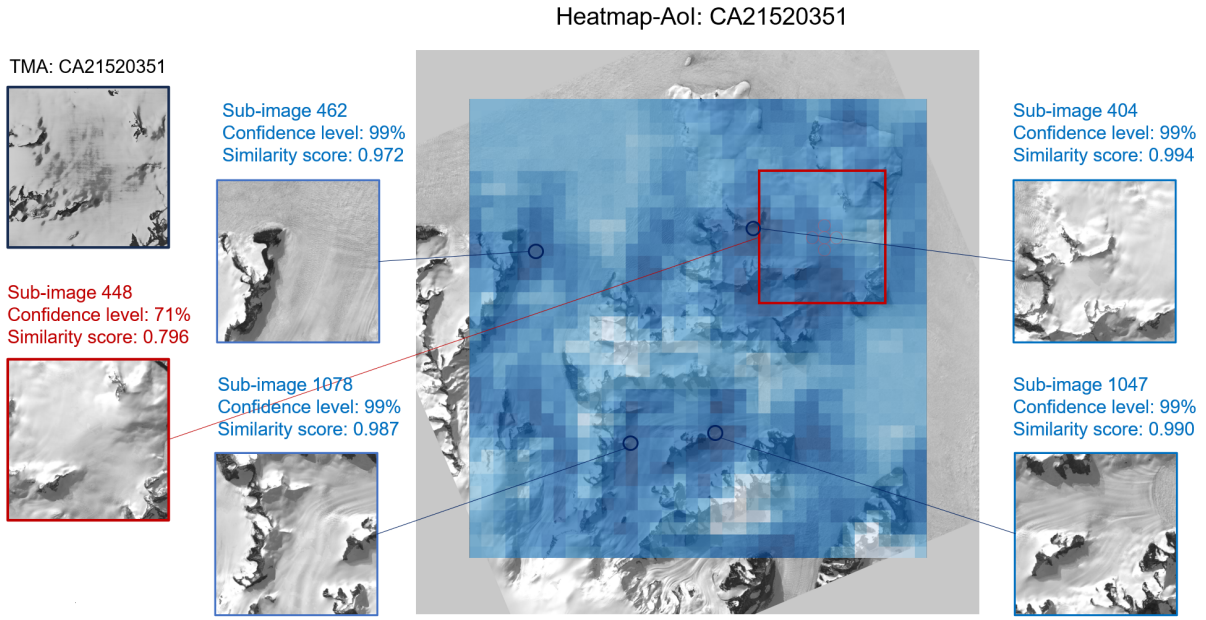


Figure 5.22: Demonstration of several false hotspots on the heatmap. The heatmap is overlaid on the AoI image with a transparency set to 50% to reveal both the heatmap and the underlying AoI. The red box delineates the coverage area of the manually selected ground truth sub-image. The blue circles represent sub-images (pixels) that, while not part of the ground truth, exhibit high similarity scores.

Below, Fig. 5.22 is utilized to analyze the results of this case. Firstly, in the top-left corner is the reference TMA image, where the bare rocks appear scattered without distinct shapes, distributed throughout the image. Additionally, there are some unusual, relatively dark and mottled areas in the central part of the image. These areas do not seem to be shadows but are more likely the result of histogram matching operations during image processing. In ground truth sub-image 448, the rocks revealed in the TMA image are not very prominent. However, visual inspection reveals terrain variations and allows us to roughly infer the presence of rocks beneath the snow, based on the interplay of light and shadow. It appears that the ResNet-50 based model also possesses a certain ability to infer the positions of rocks through the interplay of light and shadow. Therefore, the confidence level of sub-image 448 is at a relatively moderate level of 71%.

In the four falsely matched sub-images, apart from sub-image 462, rock structures similar to those in the TMA image can be observed. The main characteristic of these rock structures is that they are scattered and lack large continuous rock formations, presenting a mottled appearance. However, in sub-image 462, the rocks do not conform to the aforementioned characteristics. Regarding why it has such a high similarity level, one speculation is that the bifurcated shape of the rock in sub-image 462 resembles two rock formations in the TMA image: firstly, there is a rock in the upper-left corner of the TMA image that forks towards the upper-left direction, and secondly, there is a rock in the upper-right corner of the TMA image that forks upwards. The model may have assigned a high similarity score to sub-image 462 based on the shape similarity of the rock formations.

5.2.4. Evaluation and Adaptability Overview

In summary, the ResNet-50 based model attains a notable 95.5% average confidence level, signifying satisfactory performance. By analyzing five examples, one can discern the strengths, weaknesses, and overall applicability of the ResNet-50 based network.

Primarily, owing to its deep architecture, the ResNet-50 based model exhibits robustness in feature extraction. It effectively captures a wide array of information from the grayscale image triplets, encompassing not only low-level features such as the shapes and positions of bare rocks in darker pixel areas but also higher-level features like strong reflections, shadows, snow texture, relative rock positions, and image tones. Surprisingly, the model demonstrates adaptability to image rotation and variations in snow coverage.

This adaptability enables the model to assign relatively high similarity scores even when slight rotations or differences in snow coverage occur between highly similar image pairs.

However, the model shares a limitation with the SigNet model regarding feature extraction for blank sub-images devoid of rocks. This deficiency arises from the absence of blank images in the adaptive training samples, resulting in suboptimal feature extraction for areas without rocks. Consequently, pixels corresponding to blank areas in the Area of Interest (AoI) images display uniformly non-zero values in the heatmap.

Despite initial concerns that shadows and snow texture might impede model performance, the examples indicate that the model can effectively differentiate between shadows and rocks. This capability may be partially attributed to preprocessing steps that attenuate shadows and enhance rock features. Further investigation into whether these enhancements contribute to the model's overall performance will be undertaken in subsequent chapters.

5.2.5. Computational Efficiency

The training time for adaptive training of the ResNet-50 based network is detailed in chapter 4.3.3. Here, the prediction time for the ResNet-50 based network to determine the similarity score between a TMA image and a sub-image cropped from the Area of Interest (AoI) is presented. This prediction was conducted under two different configurations.

In the first configuration, utilizing only the CPU with a system RAM of 12.7GB, it takes approximately 1 second to predict the similarity score for one sub-image. Given the varying number of generated sub-images for different AoI images, ranging from around 1000 to 1500, the total prediction time ranges from seventeen to 25 minutes to predict for all the sub-images on one AoI image and generate heatmaps.

For the second configuration, the T4 GPU provided by Google Colab was employed, boasting a higher RAM of 51 GB. Here, it takes around 0.1 seconds to predict the similarity score for one sub-image. Consequently, the total prediction time required for creating a heatmap for an AoI image may vary from 1 minute 40 seconds to 2 minutes and 30 seconds.

Despite the ResNet-50 based network being deeper and more complex than SigNet, resulting in slightly longer training and prediction times, its significant performance improvement is noteworthy. With GPUs, the time difference is minimal, yet the ResNet-50 based network demonstrates much better performance than SigNet. Therefore, considering all factors, the ResNet-50 based network is deemed more practical and effective.

5.3. Comparative Analysis

In this chapter, a comparative analysis is conducted on three aspects. Firstly, the performance of SigNet and the ResNet-50 based model is compared, and possible reasons for any differences are analyzed from multiple perspectives. Following this, a comparison is made to demonstrate the effectiveness of adaptive training on both models, as well as the efficacy of the data preprocessing procedure for the ResNet-50 based network.

5.3.1. SigNet VS ResNet-50 based networks

In terms of general performance, the ResNet-50 based model achieves an average confidence level of 95.5% with a variance of 0.00393, while SigNet achieves an average confidence level of 70.7% with a variance of 0.0432. Undoubtedly, the ResNet-50 based model outperforms the SigNet model in all aspects, including robustness and accuracy. Although the ResNet-50 based network may have a slower prediction speed compared to the SigNet model, this speed difference can be largely overlooked when using GPU acceleration. Therefore, for the tasks conducted in this study, the ResNet-50 based network is the optimal choice. Next, the discussion will focus on why the ResNet-50 based network excels over the SigNet network from various perspectives such as network architecture, receptive field, and training dataset.

Firstly, concerning the training datasets utilized by both networks, the SigNet model undergoes pre-training with a binary Hindi signature dataset, succeeded by adaptive training using binary TMA-Rock Mask image triplets tailored specifically for this model. Images in both datasets are sized at 155×200 pixels. In contrast, the ResNet-50 based network is initially trained on the RGB ImageNet dataset and subsequently adapted with grayscale TMA-Sentinel 2 image triplets customized for this model, with images in both datasets

sized at 200×200 pixels. It is evident that RGB and grayscale images inherently encompass significantly more information than black and white images. Additionally, SigNet, by nature, constitutes a relatively simplistic network unable to handle complex images. Initially, the choice to try SigNet was made under the presumption that, typically, simpler networks tend to exhibit greater effectiveness in scenarios characterized by lower problem complexity. However, upon comparative analysis of the results yielded by both networks, it became apparent that the task complexity in this study is relatively high. Consequently, the informational content inherent in solely black and white training images proves insufficient to fulfill the task requirements. Conversely, grayscale training images offer supplementary information, such as texture and tone, which are advantageous for addressing the task objectives.

To analyze the extent to which the increased information in the grayscale training dataset compared to the binary training dataset contributes to the effectiveness of the ResNet-50 based model, a set of control experiments were conducted. The binary TMA-Rock Mask image triplet training dataset was used for adaptive training on the ResNet-50 based model and then tested it with the same binary testing samples used for the SigNet model testing. In essence, the SigNet adaptation dataset was applied to the ResNet-50 based model. The testing results under this scenario can be observed in Appendix Table A.6. The average confidence level is 88.7%, which is significantly lower than when using the grayscale TMA-Sentinel 2 dataset (95.5%). This indicates that the additional information contained in grayscale images compared to binary images plays a crucial role in enhancing the performance of the ResNet-50 based model.

Considering the network architecture perspective, the SigNet network comprises only four convolutional layers, whereas the ResNet-50 based network encompasses 49 convolutional layers, indicating a substantial difference in network depth. This significant contrast in network depth leads to considerable variations in receptive field sizes. Receptive field refers to the specific spatial area of the input data that influences the activation of a particular neuron in the hidden layers of the network, essentially representing the region from which the neuron gathers information. The receptive field size at each layer i can be computed using the formula:

$$RF_i = (RF_{i+1} - 1) \times stride_i + K_{size_i} - dilation_i \tag{5.1}$$

Here, RF_i denotes the size of the receptive field at the i-th layer, stride represents the stride (step size) of the convolution operation at the i-th layer, K_{size_i} is the size of the convolution kernel at the i-th layer, $dilation_i$ signifies the dilation factor, which is typically set to 1 for standard convolutions but can be adjusted for dilated convolutions.

Using the formula, the size of the receptive field for the SigNet network can be computed as 44×44 and for the ResNet-50 based network to be around 483×483, indicating a significant difference. The training datasets for the SigNet network consist of 155×220 images, while the ResNet-50 based network uses 200×200 images. Consequently, the receptive field for SigNet does not cover the entire input image, whereas the receptive field for the ResNet-50 based network is much larger than the input image. Generally, larger receptive fields are associated with better network performance, especially for classification tasks, where it is typically necessary to ensure that the receptive field size is greater than or equal to the input image size. For our task, the ideal scenario would be to maintain a receptive field size approximately equal to the input image size. However, currently, the receptive field of SigNet is too small, while that of the ResNet-50 based network is unnecessarily large. An excessively large receptive field may lead to the wastage of computational resources, unnecessary computational complexity, and an overemphasis on local details at the expense of global context information, thereby affecting the network's generalization ability and understanding of the overall structure. This could result in overfitting or instability, particularly when the training dataset is small or the input images exhibit significant variations. However, despite the large receptive field of the ResNet-50 based model, it has not exhibited overfitting or instability issues, and its performance has been satisfactory. Nevertheless, it is advisable to consider reducing the receptive field appropriately to reduce the number of parameters used during training and increase training efficiency while keeping the input image size constant. Conversely, for the SigNet model, a smaller receptive field means it cannot capture the global information of the input data, thereby affecting its overall performance. This could lead to difficulties in handling complex tasks as it fails to grasp the overall structure and context information of the data, which is why SigNet's performance is suboptimal.

5.3.2. Effectiveness of adaptive training

For both networks, since the models are pre-trained on datasets not specific to our task, adaptive training is adopted to realize domain adaptation. The two models used the same set of 188 adaptation training image triplets, with the only difference being the pre-processing methods applied to the images in the adaptation training datasets.

To assess the effectiveness of adaptive training, two sets of controlled experiments were conducted. Using SigNet and ResNet-50 based models before domain adaptation, they were tested on 51 testing samples, which had also undergone the two different types of image pre-processing procedures applied to the images in the adaptive training dataset. Detailed results for SigNet before domain adaptation can be found in Table A.1, and for ResNet-50 based model in Table A.3. The average confidence level for SigNet before domain adaptation was 54.3%, which increased to 70.7% after domain adaptation. For the ResNet-50 based model, the average confidence level increased from 84.9% before domain adaptation to 95.5% after domain adaptation. Domain adaptation resulted in an approximate 15% increase in average confidence level for SigNet and a 10% increase for the ResNet-50 based model. This comparison highlights the significant improvement in model performance achieved through domain adaptation in this study.

5.3.3. Effectiveness of data pre-processing

To validate that the data pre-processing procedure described in chapter 4.1.2 indeed has a positive impact on the predictive performance of the ResNet-50 based network, adaptive training and testing are conducted using RGB image triplets that had not undergone pre-processing. The testing results are presented in A.5. The average confidence level obtained was 86.6%. In contrast, when using pre-processed image triplets, the confidence level increased to 95.5%. This indicates that the image pre-processing step boosted the model's average confidence level by approximately 9%, which represents a significant improvement. Hence, it can be concluded that the data pre-processing steps designed in this study are quite effective.

6

Conclusions and Recommendations

In this chapter, the responses to each research question are provided and elucidated first, followed by the presentation of recommendations for further studies.

6.1. Conclusions

In this chapter, the primary research question is addressed initially, followed by the answers to the six sub-questions that collectively contribute to answering the main research query.

Main research question: How can deep learning-based image similarity estimation algorithms be used to locate historical aerial imagery with modern remote sensing datasets?

In prior research, we developed a methodology to approximate footprints using historical aerial imagery and additional files from the Antarctic Single Frames dataset. Leveraging these footprints, we can delineate an Area of Interest (AoI) for each TMA image. By employing deep learning-based image similarity algorithms, we can identify ground structures within the AoI that closely resemble features in the TMA images. This involves initially utilizing a sliding window approach to segment the AoI into sub-images, followed by comparison with the TMA image using the deep learning-based similarity algorithm. The resultant similarity scores between the TMA image and all sub-images generate a heatmap, with each pixel representing the similarity score for the corresponding sub-image position. The location of the TMA image is inferred by identifying the position on the heatmap with the highest similarity score, indicating the presence of the corresponding ground structure. Consequently, the hot spots on the heatmap serve as potential locations for the TMA image. However, this study is limited to providing potential locations, and further verification of the true location is necessary, which will be addressed using image matching algorithms.

1. How to establish a structured workflow for the project.

The project workflow can be outlined as follows: Initially, for each individual TMA image, we examine whether there exist other TMA images along the same flight line within a proximity of no more than 5 adjacent camera shooting points. These images must have undergone the geo-location and subsequent image matching procedures outlined in this study, yielding a determined location. If such images exist, subsequent steps can be bypassed, and the location of the current TMA image can be directly inferred based on the positional relationships provided in the additional files within the Antarctic Single Frames dataset. However, if no such images are found, the following steps are undertaken: First, the approximate coordinates of the TMA image footprint are computed, generating an Area of Interest (AoI) image. Subsequently, a sliding window approach is employed to segment the AoI image, generating sub-images. These sub-images are then paired with the TMA image and inputted into the image similarity estimation network to compute image similarity scores. The resulting similarity scores are represented as a heatmap, with the hotspots indicating potential final locations. Finally, an image matching algorithm is utilized to examine the sub-images from the highest to lowest similarity scores, aiming to ascertain the actual location. It's noteworthy that the final step is not encompassed within this study's scope.

2. How to create a suitable training dataset? In this study, the investigated deep learning-based networks, including SigNet and ResNet-50, are all Siamese networks. Typically, Siamese networks are trained on datasets containing image pairs that are either similar or triplets consisting of an anchor image, a positive image similar to the anchor, and a negative image dissimilar to the anchor. Accordingly, for

this study, we constructed a training dataset comprising image triplets to facilitate adaptive training of the model. Each image triplet includes a TMA image, a Sentinel-2 image representing the same ground area, and another Sentinel-2 image representing a different ground area. These images underwent specific pre-processing steps outlined in chapters 4.2.2 and 4.3.2 to meet the requirements of the different networks. The adaptive training datasets for both networks consisted of 188 image triplets, with 94 triplets generated through image augmentation (rotation of the image clockwise by 90 degrees). Although the size of the adaptive datasets is relatively small, it is sufficient for our purposes due to the time-consuming and labor-intensive nature of dataset creation, particularly considering the pre-trained status of the model.

3. How to assess image similarity using deep neural networks?

In this study, Siamese networks are utilized for addressing image similarity estimation problems. The principles of Siamese networks are elaborated in detail in Chapter 2.3.1. Broadly, the process of assessing image similarity using Siamese deep neural networks involves several key steps: Firstly, feature extraction, where a pre-trained deep neural network (such as SigNet or ResNet-50 in this study) is employed to extract features from the images. These features encapsulate the visual characteristics of the images within a high-dimensional space. Subsequently, a distance metric (specifically Euclidean distance in this study for both models) is applied to calculate the similarity between images based on the extracted features. Following this, the feature vectors are normalized to ensure that each dimension contributes equally to the similarity calculation, and the similarity score is derived from the distance. Additionally, in this study, the pre-trained networks undergo fine-tuning or adaptive training on task-specific datasets, namely the dataset created within this study.

4. Based on the model predictions, how to locate the rough footprint?

The model generates similarity scores for each of the sub-images derived from the Area of Interest (AoI) of a TMA image, producing a heatmap based on these predicted scores. The regions with higher scores on the heatmap represent potential locations of the footprint. However, in this study, we are unable to definitively pinpoint the location; instead, we provide potential locations. This is because the method developed in this study lacks the precision to directly ascertain a location. Further steps involving image matching algorithms are necessary to identify, among the potential locations, the actual footprint location.

5. How can the relative positions between historical images on the same flight lines be utilized in the algorithm?

As outlined in the workflow, for each TMA image, we initially verify if there exists another TMA image on the same flight line that is within a distance of five or fewer shooting points. If such an image is identified, we utilize the positional relationships between the TMA images on the same flight line to deduce the location of the current TMA image. These relationships are detailed in the Antarctic Single Frames dataset, which provides information such as the interval in miles between each picture and the direction in which the airplane flew for each flight line.

6. How to evaluate the performance of the algorithm?

The evaluation method, detailed in chapter 4.4, is primarily based on confidence levels. We assess the performance of the models (SigNet and ResNet-50 based) using a testing dataset consisting of 51 TMA images spanning four different flight lines. For each TMA image, we manually designate the sub-image that contains the scene depicted in the TMA image. We then compare the predicted similarity score of this ground truth sub-image with the similarity scores of the other sub-images relative to the TMA image. The resulting percentage, indicating by how much the predicted similarity score exceeds the others, represents the confidence level for localizing the TMA image. Subsequently, an average confidence level is computed across all 51 TMA test images. This percentage serves as an indicator of the algorithm's or the model's confidence in locating a TMA image, thereby reflecting the method's performance.

6.2. Recommendations

Below are several aspects that can be further investigated in this study.

Train with larger images and try different base networks

In this study, both networks (SigNet and ResNet-50 based networks) are limited to processing and predicting images significantly smaller than the original size. The original TMA images typically measure around 8000

x 8000 pixels with a Ground Sampling Distance (GSD) below 1 meter. Sentinel-2 images, with a GSD of 10 meters, cover approximately 800 x 800 pixels, matching the area of TMA images. Rock mask images share the same GSD as Sentinel-2 images, resulting in identical image sizes due to parameter settings.

However, images in the training datasets are resized to 155 x 224 for SigNet and 200 x 200 for the ResNet-50 based network. This resizing is necessary due to limitations in the networks' capabilities to effectively handle larger images. SigNet's shallow architecture struggles to capture features adequately on larger images due to the overwhelming amount of information. ResNet-50, designed for images around 224 x 224 pixels, faces challenges when scaling to handle larger images, such as 800 x 800, which significantly increases computational load and training time. Additionally, this results in excessively large weight files due to the increased number of parameters, as evidenced by initial attempts during this project.

However, smaller GSD implies more information that a network should be capable of extracting. Therefore, to fully utilize all available information in the images, employing or designing another network, such as U-Net, would be necessary to effectively train with larger images.

Create larger adaptive training datasets

The current adaptive training datasets each contain 188 image triplets, which is a relatively small number. Generally, larger adaptive training datasets tend to improve the predictive capability of models, particularly when the model architecture is sufficiently complex. For the ResNet-50 based network, the complexity of the model architecture is adequate. Under this premise, increasing the size of the adaptive training dataset can enhance the model's predictive performance. Expanding the dataset introduces greater variability, which can address issues such as the one previously discussed: the absence of blank image samples in the training dataset leads to non-zero, sometimes even quite large similarity scores between blank sub-images and TMA images, which is deemed unreasonable. Enlarging the size and variability of the dataset has the potential to mitigate such issues.

Investigate the method's generalizability

The ResNet-50 based model demonstrates strong predictive capabilities for similarity scores on our existing dataset. However, it is crucial to investigate whether this method possesses generalizability. This is essential for assessing the reliability and applicability of the method in real-world scenarios. Further studies can explore the method's performance on other datasets, such as historical aerial images of the Alps combined with Sentinel-2 images of the Alps region.

Enhance understanding of the networks principles to better explain its prediction behavior

Currently, regarding these two networks (SigNet and the ResNet-50 based network), in cases of both high and low confidence levels, there have been observations where certain sub-images, despite not appearing highly similar to the TMA reference image to the naked eye, are assigned relatively high similarity scores. While efforts have been made to explain some of these cases, there are still instances where a clear reason is hard to find. This is because there is a lack of thorough understanding regarding the criteria or underlying logic that the model utilizes to predict similarity scores. Although one characteristic of Convolutional Neural Networks (CNNs) is their ability to automatically generate features without the need for manual feature generating, there may still be value in future research efforts to investigate certain principles of the network, such as the distance matrix and loss function used. By doing so, insights into the basis upon which the model calculates similarity scores for these image pairs may be obtained.

References

[1] *Antarctic Seismic Data Library System (SDLS)*. https://sdls.ogs.trieste.it/cache/index.jsp. Accessed: 2024-01-08.

[2] *NASA Operation IceBridge*. https://icebridge.gsfc.nasa.gov/. Accessed: 2024-01-08.

[3] Anders A Bjørk et al. "An aerial view of 80 years of climate-related glacier fluctuations in southeast Greenland". In: *Nature Geoscience* 5.6 (2012), pp. 427–432.

[4] Yushan Liu. "Information extraction and geolocalization of historical aerial imagery". Additional MSc thesis, Delft University of Technology. 2023.

[5] Paul-Edouard Sarlin et al. "Superglue: Learning feature matching with graph neural networks". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 4938–4947.

[6] Raul Mur-Artal et al. "ORB-SLAM: a versatile and accurate monocular SLAM system". In: *IEEE transactions on robotics* 31.5 (2015), pp. 1147–1163.

[7] Ebrahim Karami et al. "Image matching using SIFT, SURF, BRIEF and ORB: performance comparison for distorted images". In: *arXiv preprint arXiv:1710.02726* (2017).

[8] Tasha Wade et al. *A to Z GIS, An illustrated dictionary of geographic information systems*. 2006.

[9] François Kayitakire et al. "Retrieving forest structure variables based on image texture analysis and IKONOS-2 imagery". In: *Remote sensing of environment* 102.3-4 (2006), pp. 390–401.

[10] Christopher SR Neigh et al. "Deciphering the precision of stereo IKONOS canopy height models for US forests with G-LiHT airborne LiDAR". In: *Remote Sensing* 6.3 (2014), pp. 1762–1782.

[11] Jesús Cascón Katchadourian et al. "The georeferencing of old cartography in geographic information systems (GIS): review, analysis and comparative study of georeferencing software". In: (2021).

[12] Lucas Santos Santana et al. "Influence of flight altitude and control points in the georeferencing of images obtained by unmanned aerial vehicle". In: *European Journal of Remote Sensing* 54.1 (2021), pp. 59–71.

[13] Julián Tomaštík et al. "UAV RTK/PPK method—an optimal solution for mapping inaccessible forested areas?" In: *Remote sensing* 11.6 (2019), p. 721.

[14] Canh Le Van et al. "Experimental investigation on the performance of DJI phantom 4 RTK in the PPK mode for 3D mapping open-pit mines". In: *Inżynieria Mineralna* 1.2 (2020), pp. 65–74.

[15] Martin Štroner et al. "Evaluation of the georeferencing accuracy of a photogrammetric model using a quadrocopter with onboard GNSS RTK". In: *Sensors* 20.8 (2020), p. 2318.

[16] Paul Ryan Nesbit et al. "Enhancing UAV–SFM 3D model accuracy in high-relief landscapes by incorporating oblique images". In: *Remote Sensing* 11.3 (2019), p. 239.

[17] Yuri Taddia et al. "Coastal mapping using DJI Phantom 4 RTK in post-processing kinematic mode". In: *Drones* 4.2 (2020), p. 9.

[18] Lorenzo Teppati Losè et al. "Boosting the timeliness of UAV large scale mapping. Direct georeferencing approaches: Operational strategies and best practices". In: *ISPRS International Journal of Geo-Information* 9.10 (2020), p. 578.

[19] Abdoulaye A Diakite et al. "Automatic geo-referencing of BIM in GIS environments using building footprints". In: *Computers, Environment and Urban Systems* 80 (2020), p. 101453.

[20] Patria Rachman Hakim et al. “Autonomous Image Georeferencing Based on Database Image Matching”. In: *2018 IEEE International Conference on Aerospace Electronics and Remote Sensing Technology (ICARES)*. IEEE. 2018, pp. 1–6.

[21] Francisco Eugenio et al. “Automatic satellite image georeferencing using a contour-matching approach”. In: *IEEE transactions on geoscience and remote sensing* 41.12 (2003), pp. 2869–2880.

[22] Gianpaolo Conte et al. “Vision-based unmanned aerial vehicle navigation using geo-referenced information”. In: *EURASIP Journal on Advances in Signal Processing* 2009 (2009), pp. 1–18.

[23] Aurelien Yol et al. “Vision-based absolute localization for unmanned aerial vehicles”. In: *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE. 2014, pp. 3429–3434.

[24] Lu Chen et al. “Extracting and matching lines of low-textured region in close-range navigation for tethered space robot”. In: *IEEE Transactions on Industrial Electronics* 66.9 (2018), pp. 7131–7140.

[25] Yicheng Li et al. “Image sequence matching using both holistic and local features for loop closure detection”. In: *IEEE Access* 5 (2017), pp. 13835–13846.

[26] Yuanxin Ye et al. “A local descriptor based registration method for multispectral remote sensing images with non-linear intensity differences”. In: *ISPRS Journal of Photogrammetry and Remote Sensing* 90 (2014), pp. 83–95.

[27] Jianwei Fan et al. “SAR and optical image registration using nonlinear diffusion and phase congruency structural descriptor”. In: *IEEE Transactions on Geoscience and Remote Sensing* 56.9 (2018), pp. 5368–5379.

[28] Qiuze Yu et al. “High-performance SAR image matching using improved SIFT framework based on rolling guidance filter and ROEWA-powered feature”. In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 12.3 (2019), pp. 920–933.

[29] Tsung-Yi Lin et al. “Learning deep representations for ground-to-aerial geolocalization”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 5007–5015.

[30] Sebastien Lefevre et al. “Toward seamless multiview scene analysis from satellite to street level”. In: *Proceedings of the IEEE* 105.10 (2017), pp. 1884–1899.

[31] Tobias Weyand et al. “Planet-photo geolocation with convolutional neural networks”. In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*. Springer. 2016, pp. 37–55.

[32] Gabriele Berton et al. “Deep visual geo-localization benchmark”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 5396–5407.

[33] Sixing Hu et al. “Image-based geo-localization using satellite imagery”. In: *International Journal of Computer Vision* 128.5 (2020), pp. 1205–1219.

[34] Yicong Tian et al. “Cross-view image matching for geo-localization in urban environments”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 3608–3616.

[35] Sudong Cai et al. “Ground-to-aerial image geo-localization with a hard exemplar reweighting triplet loss”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 8391–8400.

[36] Aysim Toker et al. “Coming down to earth: Satellite-to-street view synthesis for geo-localization”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 6488–6497.

[37] Shiv Ram Dubey. “A decade survey of content based image retrieval using deep learning”. In: *IEEE Transactions on Circuits and Systems for Video Technology* 32.5 (2021), pp. 2687–2704.

[38] Qian Zhao et al. “AdaSAN: Adaptive cosine similarity self-attention network for gastrointestinal endoscopy image classification”. In: *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2021, pp. 1855–1859.

[39] Ganesh Gopalrao Patil et al. “A hybrid SUGWO optimization for partial face recognition with new similarity index”. In: *Multimedia Tools and Applications* 82.12 (2023), pp. 18097–18116.

[40] Srikar Appalaraju et al. “Image similarity using deep CNN and curriculum learning”. In: *arXiv preprint arXiv:1709.08761* (2017).

[41] Navneet Dalal et al. “Histograms of oriented gradients for human detection”. In: *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR’05)*. Vol. 1. Ieee. 2005, pp. 886–893.

[42] David G Lowe. “Distinctive image features from scale-invariant keypoints”. In: *International journal of computer vision* 60 (2004), pp. 91–110.

[43] Yun Liu et al. “Del: Deep embedding learning for efficient image segmentation.” In: *IJCAI*. Vol. 864. 2018, p. 870.

[44] Lawrence Hubert et al. “Comparing partitions”. In: *Journal of classification* 2 (1985), pp. 193–218.

[45] Scott Deerwester et al. “Indexing by latent semantic analysis”. In: *Journal of the American society for information science* 41.6 (1990), pp. 391–407.

[46] Raymond T. Ng et al. “CLARANS: A method for clustering objects for spatial data mining”. In: *IEEE transactions on knowledge and data engineering* 14.5 (2002), pp. 1003–1016.

[47] Richard W Hamming. “Error detecting and error correcting codes”. In: *The Bell system technical journal* 29.2 (1950), pp. 147–160.

[48] Chen Shen et al. “Deep siamese network with multi-level similarity perception for person re-identification”. In: *Proceedings of the 25th ACM international conference on Multimedia*. 2017, pp. 1942–1950.

[49] Elad Hoffer et al. “Deep metric learning using triplet network”. In: *Similarity-Based Pattern Recognition: Third International Workshop, SIMBAD 2015, Copenhagen, Denmark, October 12-14, 2015. Proceedings 3*. Springer. 2015, pp. 84–92.

[50] Seyed Vahid Moravvej et al. “A method based on an attention mechanism to measure the similarity of two sentences”. In: *2021 7th International Conference on Web Research (ICWR)*. IEEE. 2021, pp. 238–242.

[51] Norwegian Polar Institute. *Quantarctica*. https://www.npolar.no/quantarctica/. Accessed: December 10, 2023. 2023.

[52] A. Burton-Johnson et al. “An automated methodology for differentiating rock from snow, clouds and sea in Antarctica from Landsat 8 imagery: a new rock outcrop map and area estimation for the entire Antarctic continent”. In: *The Cryosphere* 10.4 (2016), pp. 1665–1677. DOI: 10.5194/tc-10-1665-2016. URL: https://tc.copernicus.org/articles/10/1665/2016/.

[53] Sounak Dey et al. “Signet: Convolutional siamese network for writer independent offline signature verification”. In: *arXiv preprint arXiv:1707.02131* (2017).

[54] Yann LeCun et al. “Gradient-based learning applied to document recognition”. In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.

[55] Alex Krizhevsky et al. “Imagenet classification with deep convolutional neural networks”. In: *Advances in neural information processing systems* 25 (2012).

[56] Md. Rakibul Islam et al. “Detection of COVID 19 from CT Image by The Novel LeNet-5 CNN Architecture”. In: *2020 23rd International Conference on Computer and Information Technology (ICCIT)*. 2020, pp. 1–5. DOI: 10.1109/ICCIT51783.2020.9392723.

[57] Ali Abd Almisreb et al. “Utilizing AlexNet Deep Transfer Learning for Ear Recognition”. In: *2018 Fourth International Conference on Information Retrieval and Knowledge Management (CAMP)*. 2018, pp. 1–5. DOI: 10.1109/INFRKM.2018.8464769.

[58] Xiaobing Han et al. “Pre-trained alexnet architecture with pyramid pooling and supervision for high spatial resolution remote sensing image scene classification”. In: *Remote Sensing* 9.8 (2017), p. 848.

[59] Ishani Kathuria. *Handwritten Signature Datasets*. 2024/01/01. 2021. URL: https://www.kaggle.com/datasets/ishanikathuria/handwritten-signature-datasets/data.

[60] Sumit Chopra et al. “Learning a similarity metric discriminatively, with application to face verification”. In: *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR’05)*. Vol. 1. IEEE. 2005, pp. 539–546.

[61] Florian Schroff et al. “Facenet: A unified embedding for face recognition and clustering”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 815–823.

[62] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[63] Karen Simonyan et al. “Very deep convolutional networks for large-scale image recognition”. In: *arXiv preprint arXiv:1409.1556* (2014).

[64] Christian Szegedy et al. “Going deeper with convolutions”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1–9.

[65] Mingxing Tan et al. “Efficientnet: Rethinking model scaling for convolutional neural networks”. In: *International conference on machine learning*. PMLR. 2019, pp. 6105–6114.

[66] Andrew G Howard et al. “Mobilenets: Efficient convolutional neural networks for mobile vision applications”. In: *arXiv preprint arXiv:1704.04861* (2017).

[67] Yishu Liu et al. “Similarity-based unsupervised deep transfer learning for remote sensing image retrieval”. In: *IEEE Transactions on Geoscience and Remote Sensing* 58.11 (2020), pp. 7872–7889.

[68] Kai Qiu et al. “Siamese-ResNet: Implementing loop closure detection based on Siamese network”. In: *2018 IEEE Intelligent Vehicles Symposium (IV)*. IEEE. 2018, pp. 716–721.

A

Appendix

A.1. 51 Test Results on SigNet (before and after fine-tuning)

Table A.1: 51 test results on SigNet Before Fine-tuning (average confidence level: 54.3%)

TMA_num	photo_id	Similarity Score	Confidence Level
2147	24	0.865	0.406
2147	25	0.866	0.436
2147	27	0.977	0.735
2147	28	0.990	0.876
2147	29	0.956	0.500
2147	30	0.689	0.118
2147	31	0.831	0.152
2147	32	0.935	0.654
2147	33	0.875	0.217
2147	34	0.352	0.138
2147	35	0.856	0.248
2147	36	0.962	0.866
2147	37	0.990	0.957
2147	38	0.560	0.091
2147	39	0.802	0.738
2147	40	0.982	0.934
2147	41	0.839	0.251
2147	42	0.891	0.239
2147	43	0.547	0.076
2147	44	0.989	0.941
2147	45	0.884	0.212
2147	46	0.987	0.905
2147	47	0.881	0.908
2147	48	0.973	0.951
2147	49	0.667	0.189
2147	50	0.924	0.674
2147	51	0.809	0.263
2152	351	0.918	0.411
2152	352	0.873	0.866

Table A.1: 51 test results on SigNet Before Fine-tuning (average confidence level: 54.3%)

TMA_num	photo_id	Similarity Score	Confidence Level
2152	353	0.949	0.540
2153	400	0.864	0.259
2153	401	0.816	0.123
2153	420	0.929	0.807
2153	421	0.966	0.789
2153	422	0.937	0.785
2153	423	0.943	0.620
2153	424	0.921	0.599
2153	425	0.902	0.310
2153	426	0.944	0.623
2153	427	0.963	0.825
2153	428	0.987	0.948
2153	429	0.980	0.870
2153	430	0.978	0.879
2153	431	0.972	0.842
2158	72	0.938	0.798
2158	73	0.556	0.058
2158	74	0.770	0.538
2158	75	0.500	0.039
2158	77	0.915	0.605
2158	78	0.845	0.534
2158	84	0.788	0.369

Table A.2: 51 test results on SigNet After Fine-tuning (average confidence level: 70.7%).

TMA_num	photo_id	Similarity Score	Confidence Level
2147	24	0.598	0.425
2147	25	0.773	0.655
2147	27	0.790	0.730
2147	28	0.838	0.821
2147	29	0.488	0.362
2147	30	0.848	0.842
2147	31	0.889	0.866
2147	32	0.881	0.880
2147	33	0.855	0.933
2147	34	0.776	0.888
2147	35	0.924	0.983
2147	36	0.893	0.913
2147	37	0.904	0.934
2147	38	0.807	0.883
2147	39	0.730	0.803
2147	40	0.895	0.921
2147	41	0.738	0.418

Table A.2: 51 test results on SigNet After Fine-tuning (average confidence level: 70.7%).

TMA_num	photo_id	Similarity Score	Confidence Level
2147	42	0.949	0.901
2147	43	0.775	0.796
2147	44	0.735	0.797
2147	45	0.679	0.746
2147	46	0.824	0.742
2147	47	0.852	0.806
2147	48	0.776	0.250
2147	49	0.914	0.420
2147	50	0.871	0.408
2147	51	0.934	0.464
2152	351	0.775	0.591
2152	352	0.311	0.297
2152	353	0.627	0.407
2153	400	0.812	0.657
2153	401	0.744	0.556
2153	420	0.352	0.436
2153	421	0.545	0.559
2153	422	0.640	0.655
2153	423	0.444	0.359
2153	424	0.947	0.907
2153	425	0.943	0.959
2153	426	0.916	0.719
2153	427	0.875	0.865
2153	428	0.762	0.772
2153	429	0.797	0.863
2153	430	0.884	0.955
2153	431	0.934	0.968
2158	72	0.861	0.729
2158	73	0.826	0.845
2158	74	0.856	0.813
2158	75	0.807	0.909
2158	77	0.731	0.404
2158	78	0.837	0.602
2158	84	0.632	0.667

A.2. 51 test results on ResNet-50 based network (before and after fine-tuning)

Table A.3: 51 test results on ResNet50-based Network Before Fine-tuning (average confidence level: 84.9%).

TMA_num	photo_id	Similarity Score	Confidence Level
2147	24	0.721	0.305
2147	25	0.601	0.158
2147	27	0.947	0.997
2147	28	0.911	0.965
2147	29	0.898	0.994
2147	30	0.872	0.922
2147	31	0.784	0.540
2147	32	0.898	0.903
2147	33	0.819	0.584
2147	34	0.732	0.341
2147	35	0.536	0.181
2147	36	0.893	0.962
2147	37	0.951	0.979
2147	38	0.988	0.995
2147	39	0.907	0.912
2147	40	0.858	0.945
2147	41	0.917	0.995
2147	42	0.955	0.991
2147	43	0.950	0.994
2147	44	0.905	0.961
2147	45	0.900	0.863
2147	46	0.956	0.984
2147	47	0.942	0.980
2147	48	0.745	0.898
2147	49	0.961	0.991
2147	50	0.926	0.392
2147	51	0.882	0.979
2152	351	0.800	0.805
2152	352	0.894	0.926
2152	353	0.863	0.690
2153	400	0.952	0.933
2153	401	0.946	0.992
2153	420	0.951	0.946
2153	421	0.920	0.936
2153	422	0.988	0.998
2153	423	0.938	0.811
2153	424	0.961	0.899
2153	425	0.891	0.745
2153	426	0.936	0.988
2153	427	0.904	0.993
2153	428	0.873	0.987
2153	429	0.909	0.993

Table A.3: 51 test results on ResNet50-based Network Before Fine-tuning (average confidence level: 84.9%).

TMA_num	photo_id	Similarity Score	Confidence Level
2153	430	0.934	0.825
2153	431	0.935	0.862
2158	72	0.862	0.972
2158	73	0.881	0.671
2158	74	0.842	0.671
2158	75	0.970	0.996
2158	77	0.947	0.996
2158	78	0.866	0.996
2158	84	0.883	0.942

Table A.4: 51 test results on ResNet50-based Network After Fine-tuning (average confidence level: 95.5%).

TMA_num	photo_id	Similarity Score	Confidence Level
2147	24	0.948	0.927
2147	25	0.834	0.945
2147	27	0.909	0.981
2147	28	0.944	0.979
2147	29	0.957	0.984
2147	30	0.966	0.991
2147	31	0.859	0.927
2147	32	0.965	0.994
2147	33	0.945	0.983
2147	34	0.853	0.925
2147	35	0.848	0.902
2147	36	0.964	0.995
2147	37	0.955	0.975
2147	38	0.986	0.996
2147	39	0.908	0.974
2147	40	0.812	0.886
2147	41	0.985	0.988
2147	42	0.942	0.99
2147	43	0.94	0.93
2147	44	0.979	0.997
2147	45	0.931	0.963
2147	46	0.885	0.955
2147	47	0.963	0.992
2147	48	0.971	0.967
2147	49	0.97	0.991
2147	50	0.919	0.957
2147	51	0.979	0.982
2152	351	0.796	0.717

Table A.4: 51 test results on ResNet50-based Network After Fine-tuning (average confidence level: 95.5%).

TMA_num	photo_id	Similarity Score	Confidence Level
2152	352	0.955	0.996
2152	353	0.957	0.994
2153	400	0.989	0.999
2153	401	0.945	0.977
2153	420	0.923	0.957
2153	421	0.924	0.971
2153	422	0.966	0.994
2153	423	0.958	0.985
2153	424	0.934	0.889
2153	425	0.875	0.674
2153	426	0.975	0.982
2153	427	0.909	0.936
2153	428	0.86	0.859
2153	429	0.912	0.952
2153	430	0.91	0.941
2153	431	0.87	0.901
2158	72	0.981	0.992
2158	73	0.945	0.963
2158	74	0.966	0.994
2158	75	0.932	0.987
2158	77	0.963	0.993
2158	78	0.918	0.981
2158	84	0.933	0.974

A.3. Fine-tuning and testing with unprocessed image pairs on ResNet-50 based network.

Table A.5: Results of testing and fine-tuning with unprocessed image pairs on ResNet50-based Network (average confidence level: 86.6%).

TMA_num	photo_id	Similarity Score	Confidence Level
2147	24	0.933	0.918
2147	25	0.866	0.734
2147	27	0.904	0.935
2147	28	0.943	0.983
2147	29	0.814	0.924
2147	30	0.838	0.904
2147	31	0.721	0.814
2147	32	0.842	0.918
2147	33	0.940	0.976
2147	34	0.828	0.945
2147	35	0.707	0.708

Table A.5: Results of testing and fine-tuning with unprocessed image pairs on ResNet50-based Network (average confidence level: 86.6%).

TMA_num	photo_id	Similarity Score	Confidence Level
2147	36	0.950	0.997
2147	37	0.959	0.970
2147	38	0.987	0.995
2147	39	0.815	0.888
2147	40	0.838	0.909
2147	41	0.964	0.994
2147	42	0.954	0.965
2147	43	0.928	0.991
2147	44	0.917	0.981
2147	45	0.917	0.937
2147	46	0.878	0.504
2147	47	0.828	0.378
2147	48	0.876	0.406
2147	49	0.932	0.975
2147	50	0.815	0.849
2147	51	0.892	0.288
2152	351	0.910	0.715
2152	352	0.934	0.982
2152	353	0.931	0.985
2153	400	0.960	0.996
2153	401	0.954	0.980
2153	420	0.890	0.844
2153	421	0.911	0.908
2153	422	0.933	0.932
2153	423	0.920	0.930
2153	424	0.924	0.911
2153	425	0.866	0.753
2153	426	0.944	0.924
2153	427	0.925	0.928
2153	428	0.879	0.819
2153	429	0.916	0.921
2153	430	0.893	0.685
2153	431	0.826	0.583
2158	72	0.977	0.973
2158	73	0.954	0.947
2158	74	0.943	0.864
2158	75	0.925	0.920
2158	77	0.954	0.985
2158	78	0.953	0.961
2158	84	0.953	0.954

A.4. Testing the test dataset made for SigNet on ResNet-50 based network.

Table A.6: Results of fine-tuning and testing the black and white image pairs made for SigNet on ResNet50-based Network (average confidence level: 88.7%).

TMA_num	photo_id	Similarity Score	Confidence Level
2147	24	0.656	0.190
2147	25	0.780	0.490
2147	27	0.908	0.817
2147	28	0.950	0.881
2147	29	0.898	0.922
2147	30	0.907	0.990
2147	31	0.930	0.990
2147	32	0.924	0.986
2147	33	0.942	0.980
2147	34	0.945	0.993
2147	35	0.939	0.993
2147	36	0.970	0.996
2147	37	0.953	0.993
2147	38	0.931	0.615
2147	39	0.884	0.953
2147	40	0.940	0.978
2147	41	0.958	0.979
2147	42	0.957	0.918
2147	43	0.942	0.981
2147	44	0.971	0.997
2147	45	0.938	0.980
2147	46	0.994	0.988
2147	47	0.905	0.577
2147	48	0.889	0.825
2147	49	0.935	0.873
2147	50	0.978	0.983
2147	51	0.917	0.881
2152	351	0.933	0.889
2152	352	0.920	0.976
2152	353	0.873	0.942
2153	400	0.887	0.701
2153	401	0.806	0.340
2153	420	0.986	0.995
2153	421	0.973	0.996
2153	422	0.966	0.978
2153	423	0.966	0.976
2153	424	0.899	0.896
2153	425	0.918	0.940
2153	426	0.942	0.967

Table A.6: Results of fine-tuning and testing the black and white image pairs made for SigNet on ResNet50-based Network (average confidence level: 88.7%).

TMA_num	photo_id	Similarity Score	Confidence Level
2153	427	0.946	0.992
2153	428	0.955	0.997
2153	429	0.962	0.998
2153	430	0.975	0.993
2153	431	0.992	0.990
2158	72	0.856	0.887
2158	73	0.961	0.994
2158	74	0.946	0.906
2158	75	0.911	0.825
2158	77	0.914	0.845
2158	78	0.925	0.871
2158	84	0.777	0.616