Technische Universität Berlin

Faculty of Electrical Engineering and Computer Science
Dept. of Computer Engineering and Microelectronics
Remote Sensing Image Analysis Group

Multi-Modal Vision Transformers for Multi-Label Remote Sensing Image Classification

Master of Science in Computer Science

September, 2022

Hoffmann, David Sebastian

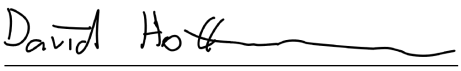
Matriculation Number: 380850

First Supervisor:	Prof. Dr. Begüm Demir
Second Supervisor:	Prof. Dr. Volker Markl
Advisor:	Kai Norman Clasen

Declaration

I hereby declare that the thesis submitted is my own, unaided work, completed without any unpermitted external help. Only the sources and resources listed were used.

Berlin, September 16, 2022



David Sebastian Hoffmann

Abstract

The fusion of images captured by different sensor modalities is a highly researched subject in the field of remote sensing. Deriving information about a scene from multiple modalities holds the potential to significantly improve the performance of deep learning models to accomplish various tasks in the remote sensing domain. In this context, the thesis investigates the capabilities of the novel Vision Transformer (ViT) architecture to perform multi-modal fusion of remote sensing images. To this end, several multi-modal fusion methods based on the ViT architecture are proposed and compared on a multi-label remote sensing dataset. The experiments are conducted on the BigEarthNet-MM dataset with multi-spectral optical and Synthetic Aperture Radar images serving as the modalities to be fused. The proposed fusion methods adapt different architectural components of the standard ViT architecture to improve their respective potential to facilitate multi-modal fusion. The results show that, while all investigated fusion methods can improve upon training on single modality data, substantial performance differences among the considered fusion methods could be observed. A comprehensive analysis is conducted to evaluate these differences and identify their underlying causes. The investigation includes the comparison of the overall as well as the class-wise performance of these methods on a scene classification task. Additionally, detailed ablation studies are performed to assess the impact of architectural hyper-parameters on the fusion performance.

Zusammenfassung

Ein wichtiges Forschungsgebiet innerhalb der Fernerkundungsforschung beschäftigt sich mit der effektiven Fusion von Bilddaten, welche von unterschiedlichen Sensorsystemen aufgenommen wurden. Solche visuellen Sensormodalitäten effizient zu kombinieren, bietet großes Potenzial die Fähigkeiten von tiefen Neuronalen Netzen zu verbessern, welche zur Analyse solcher Daten eingesetzt werden. In diesem Zusammenhang erforscht diese Arbeit die Eignung der neuartigen Vision Transformer (ViT) Architektur zur Fusion von multimodalen Satellitenbilddaten. Dazu werden mehrere multimodale Fusionsmethoden basierend auf der ViT Architektur entwickelt und auf einem Satellitenbilddatensatz mit multipler Klassenzuordung verglichen. Zu diesem Zwecke wird der BigEarthNet-MM Datensatz verwendet, welcher multispektrale und Synthetic Aperture Radar Bilddaten zur Analyse bereitstellt. Die untersuchten Fusionsmethoden adaptieren verschiedene Komponenten der ViT Architektur mit dem Ziel eine bessere und einfachere Fusion multimodaler Sensordaten zu ermöglichen. Die Ergebnisse zeigen, dass die unterschiedlichen Methoden eine bessere Leistung erzielen als Modelle, welche nur auf einer Modalität trainiert wurden. Gleichzeitig konnten jedoch starke Unterschiede in der Klassifizierungsleistung der verschiedenen Methoden festgestellt werden. Diese Unterschiede werden in einer detaillierten Analyse sowohl der insgesamten als auch der klassenspezifischen Klassifizierungsleistung ausgewertet. Zusätzlich werden ausführliche Ablationstudien durchgeführt, um den Einfluss bestimmter Hyperparameter auf die Klassifizierungsleistung zu ermitteln.

Contents

List of Figures

List of Tables

List of Acronyms

AP	Average Precision
CBIR	Content-based Image Retrieval
CLC	CORINE Land Cover
CNN	Convolutional Neural Network
CV	Computer Vision
GELU	Gaussian Error Linear Unit
GPU	Graphics Processing Unit
GRD	Ground Range Detected
HL	Hamming Loss
HSI	Hyper-Spectral Image
LIDAR	Light Detection and Ranging
LN	Layer Normalisation
LULC	Land Use Land Cover
MLP	Multilayer Perceptron
MSA	Multi-Head Self-Attention
MSI	Multi-Spectral Image
NIR	Near Infrared
NLP	Natural Language Processing
ReLU	Rectified Linear Unit
RGB	Red-Green-Blue
RS	Remote Sensing
S1	Sentinel-1
S2	Sentinel-2
SAR	Synthetic Aperture Radar
SCT	Synchronised Class Token
SWIR	Short-wave Infrared

ViT	Vision Transformer
VRAM	Video Random Access Memory

1 Introduction

In recent decades the field of Remote Sensing (RS) has seen a rapid increase in the amount of satellite data available for earth analysis tasks. This trend has been powered by continuously decreasing costs to launch satellite systems to low earth orbit. Subsequently, the number of deployments of satellites equipped with sensors for visual data acquisition has consistently risen in recent years [1]. Concurrently, improvements in the technological capabilities of imaging sensors have enabled the acquisition of data at increasingly higher resolutions [1]. These developments have considerably increased the amount and complexity of available image data captured by satellites necessitating the development of dedicated methods to analyse and process such large quantities of data.

Many applications from the research, commercial and military sector rely on the imaging capabilities of sensors deployed on satellite systems to obtain detailed information about the earth's surface. To suit the objectives of different stakeholders, a great variety of specialised satellite missions utilising different technologies to acquire image data have been developed. The sensor systems of these satellite missions can differ in many aspects ranging from the spectral or spatial resolutions to the fundamental physical principles on which they operate. The data captured by a particular type of sensor system is referred to as a *sensor modality*, often abbreviated as modality in the context of this thesis. The most prominent types of sensor modalities for satellite imaging include panchromatic, multi-spectral and hyper-spectral optical sensors as well as Synthetic Aperture Radar (SAR) systems [2].

Historically, the data obtained from different systems has often been analysed separately partly due to the restrictions in available data for a specific location. However, modern satellite systems can provide wide-ranging coverage of most geographic locations at high revisiting times. This could enable the utilisation of multiple modalities acquired over the same location for various earth analysis tasks. Nonetheless, combining different sensor modalities can introduce significant challenges to any analysis procedure necessitating the development of dedicated methods to effectively derive information from multiple modalities [3].

Concurrently to the progress in RS technology, considerable leaps in the field of machine learning and especially Computer Vision (CV) have resulted in various new models for analysing visual data with significantly increased performance compared to classical methods. Especially the field of computer vision has largely been dominated by Convolutional Neural Network (CNN) architectures [4] in the last decade since the inception of the AlexNet architecture [5]. Many more advances in the field have steadily led to better and more sophisticated CNN models achiev-

ing outstanding performances on many image analysis tasks [6–8]. Due to these developments, deep learning methods have also become invaluable in satellite image analysis. Deep artificial neural networks allow the extraction of semantic information from an RS image by mapping it to a highly abstract but semantically meaningful feature representation. This enables tasks such as scene classification, Land Use Land Cover (LULC) classification or Content-based Image Retrieval (CBIR) to be performed with very high accuracy by automated procedures [9].

Therefore, a significant amount of research has focused on the multi-modal fusion of satellite data by integrating the fusion process into CNN architectures [2]. While these approaches are often able to achieve high performances the rapid advancements in machine learning constantly require further exploratory studies to identify potential new methods to perform multi-modal fusion.

Recent advances have shown that convolution-free architectures can match and even outperform their convolutional counterparts. This was prominently shown with the proposal of the Vision Transformer (ViT) architecture by Dosovitskiy et al. [10] by achieving new state-of-the-art classification results on the ImageNet dataset [11]. The ViT architecture is based on the Transformer architecture, which was proposed by Vaswani et al. [12] for tasks in the field of natural language processing. The ViT architecture introduced a radically different approach to image analysis tasks by not integrating a local inductive bias which is an intrinsic component of CNN models. Instead, the ViT architecture employs an *attention mechanism* to compute a set of attention scores denoting the relevance of each part of an input sequence to the current classification task.

Such an approach could be beneficial for combining separate modalities by determining which features from each modality have the highest relevance for a specific task. Therefore, Transformer-based multi-modal fusion methods have been widely researched in the broader deep learning field [13].

In the previously described context, this thesis aims to investigate the capabilities of the ViT architecture for the fusion of multi-spectral and SAR satellite images. To this end, multiple fusion methods based upon the ViT architecture are proposed in this thesis and evaluated on a scene classification task. The proposed fusion methods introduce various modifications to the ViT architecture with the aim to identify a modification strategy best suited for the fusion of multi-spectral and SAR satellite images. The training and experiments are conducted on the BigEarthNet-MM dataset proposed by Sümbül et al. [14]. BigEarthNet-MM presents one of the most extensive datasets available for deep learning RS research. It provides sufficient quantities of high-resolution multi-spectral and SAR images to train and evaluate the investigated multi-modal fusion methods effectively. The fusion performance of the proposed and investigated methods is analysed in detail by comparing their respective performance on the same multi-label classification task.

Outline

The thesis is structured as follows: Firstly, Chapter 2 introduces the types of multi-modal data utilised for the multi-modal fusion task. Next, Chapter 3 provides a detailed introduction of the ViT architecture on which all investigated multi-modal fusion methods are based. Chapter 4 then gives an outline of the state-of-the-art in multi-modal fusion of RS image data with deep learning techniques. Additionally, this chapter contains an overview of Transformer-based multi-modal fusion methods from the general deep learning domain. Afterwards, Chapter 5 provides a detailed introduction of the multi-modal fusion methods investigated and proposed in this thesis. For each method, the specific modifications performed on the ViT architecture are defined with an analysis of their intended effect on the fusion process. Chapter 6 then details relevant properties of the BigEarthNet-MM dataset and describes the experimental setup used to conduct and evaluate the experiments to determine the performance of the various fusion methods. An analysis of the results obtained by the presented fusion methods is provided in Chapter 7. This includes an analysis of the performance on fusing the multi-spectral and SAR information present in the BigEarthNet-MM dataset and the performance on a reduced fusion task only utilising Red-Green-Blue (RGB) and SAR data. The chapter further includes detailed ablation studies on the impact of various hyper-parameter settings on the fusion performance. Finally, Chapter 8 concludes with a discussion of relevant insights obtained from the results as well as a small outlook on potential future research.

2 Fundamentals of Remote Sensing Data

This thesis's primary focus is to analyse and improve the capabilities of the ViT architecture in fusing two multi-modal data representations from the field of remote sensing for a classification task. The following sections briefly introduce the type of satellite data utilised in this thesis. The main focus is on their inherent properties arising due to different acquisition processes.

2.1 Multi-Spectral Satellite Images

Multi-spectral imaging is the process of capturing an image over multiple spectral bands simultaneously, usually with an array of dedicated sensors. These sensors are typically specialised for a specific range of wavelengths from the electromagnetic spectrum. Similar to classical RGB cameras, the individual channels contain information on the spectral intensity received in the spectrum corresponding to that particular sensor. Multiple multi-spectral imaging satellite missions are currently in operation. Some of the most well-known include the Sentinel-2 mission [15, pp. 9-12], whose images are utilised in this thesis, as well as the Landsat and MODIS missions [16]. The overall amount of spectral bands captured typically ranges from 3 to up to a few dozen. Commonly, most missions operate in the electromagnetic spectrum of visible light, Near Infrared (NIR) and Short-wave Infrared (SWIR) [16].

A Multi-Spectral Image (MSI) allows the distinguishment of different surface properties based on specific spectral responses. This is possible because the chemical compositions of various materials exhibit distinct spectral signatures. Previously, the varying spectral responses relating to differing materials have often been exploited to perform a basic form of classification by computing a so-called *index* from specific channels. For such an index, the value at a pixel location corresponds with the presence of particular materials such as water or vegetation [16]. In recent years, however, deep learning has largely replaced such approaches due to its ability to model more complex relationships and achieve better results in a variety of tasks.

The spatial and spectral resolution are critical properties of any multi-spectral system employed for remote sensing. The spatial resolution refers to the physical size of the area on the ground captured by one pixel of the sensor. It can range from sub-meter resolution for very advanced systems up to a few dozen meters. Conversely, the spectral resolution denotes the detail at which a system captures spectral bands. More specifically, it refers to the difference between a sensor's maximum and minimum wavelength to which it is sensitive. Fig. 2.1 shows the

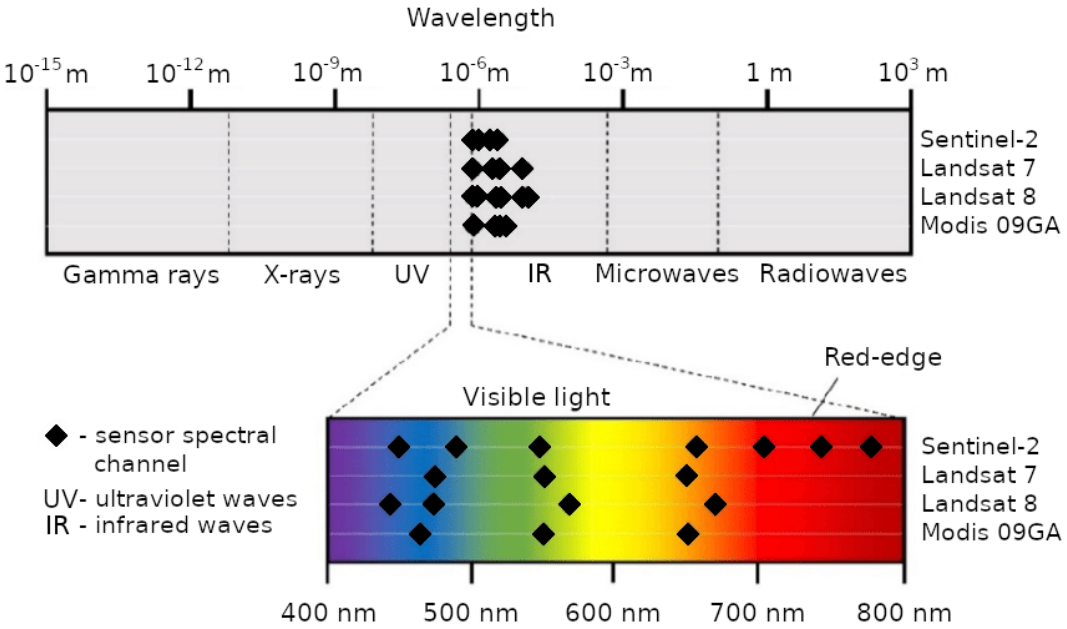


Figure 2.1: Overview on the spectral bands captured by a selection of relevant multi-spectral satellite missions and their position on the electromagnetic spectrum [16].

spectral bands and their position on the electromagnetic spectrum for a selection of multi-spectral missions.

Continuous advancements in spatial and spectral resolution allow modern multi-spectral satellite systems to provide a high amount of detail in their captured images. This information enables a plethora of different research and commercial applications to conduct detailed analyses of the Earth's surface. Operating very similarly to classical camera systems with more channels also simplifies the interpretation of the provided data.

However, as multi-spectral systems rely on optical observation, they are strongly impacted by changes to ground illumination and visibility. Therefore, such systems can neither operate at night nor when cloud cover prevents the direct observation of the surface. The problem of relying on daylight illumination can be alleviated by placing the satellite in a sun-synchronous orbit. However, the captured data will still experience a significant variance in the intensity and the angle of illumination. In addition, the presence of clouds at the time of capture usually prevents the corresponding images from being considered for further analysis.

Overall, multi-spectral satellite images provide highly detailed and valuable information about the Earth's surface. However, the aforementioned limitations can be inhibiting for some applications. Therefore, to reduce the effects of clouds and illumination on the resulting image data, other imaging technologies have been developed, such as the SAR technology introduced in the next section.

2.2 Synthetic Aperture Radar Data

Synthetic Aperture Radar is a widely utilised technology in RS to acquire images of the Earth's surface. While it also relies on electromagnetic radiation in the imaging process, its underlying principle and operation differ substantially from spectral satellite imaging.

SAR systems mainly utilise microwave radiation for scanning the surface. Such microwave radiation has significantly larger wavelengths than the spectral bands used in optical imaging technologies. Additionally, SAR systems do not measure the spectral response for a pixel location but instead rely on the radar principle to obtain measurements of distances to objects.

A radar system sends out bursts of radiation and measures the backscattered signal caused by objects reflecting the bursts of radiation back. The distance to a target can be determined by evaluating the time difference between responses. Additionally, the strength of the backscattered signal directly corresponds to the surface properties of a target due to its interaction with the radiation. Fig. 2.2 visualises the principle of SAR data acquisition.

At this point, it is important to introduce a specific terminology for certain spatial directions relative to the satellite's movement. The direction along the satellite's flight path is called the *azimuth* direction and the direction orthogonal to the satellite's flight path is called the *range* direction. The differentiation is relevant because different mechanisms are utilised to determine the source location of a measured backscatter signal in azimuth and range direction from the satellite's position.

The range distance can be computed from the time difference between sending and receiving a signal, as described previously. However, it requires the imaging to be performed at an angle to the surface from the satellite's point of view. Otherwise, the time difference would not precisely correspond to a specific distance in range direction as multiple objects might be at the same distance but on different sides of the satellite's flight path. Therefore, a SAR imaging satellite can never obtain an image directly in the nadir direction. Fig. 2.2 also visualises the angled scanning of the surface from a SAR satellite.

However, to precisely discern different objects in the azimuth direction, a radar imaging system would need to send out very narrow beams to determine where precisely a signal was backscattered. Because a narrower beamwidth would require a significantly larger aperture, radar imaging systems relying on a real aperture are highly impractical to build and deploy with currently available technologies. Therefore, a SAR imaging system constructs a synthetic aperture, hence the name of the imaging technique, by sending and receiving a multitude of wider bursts over the satellite's flight path. By taking such measurements at different points in time, an object on the surface produces multiple backscatter responses that differ in range depending on the satellite's position. Considering the properties of the satellite's orbit and the burst frequency, the various measurements can

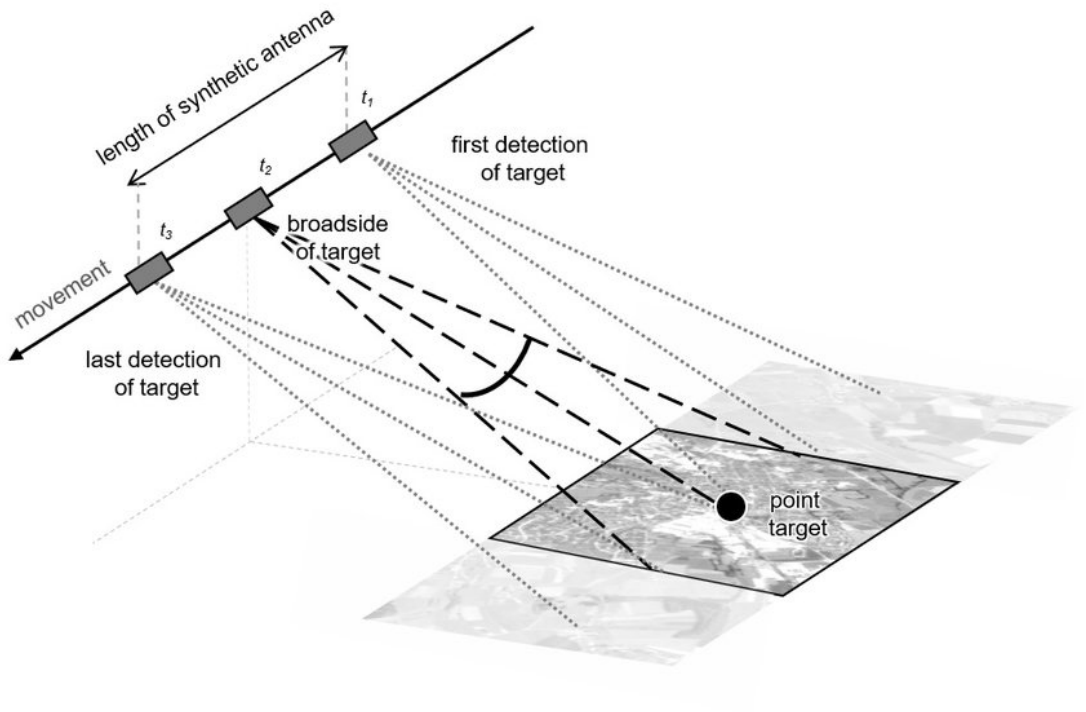


Figure 2.2: Visualisation of the flight path of a SAR imaging satellite showing the angled acquisition of the imaged swath. The satellite measures backscatter responses at time points t1, t2 and t3. The measured signals are used to compute the precise location of the point target in both the range and azimuth direction relative to the satellite [17, p. 21].

be combined to infer an object's position relative to the satellite accurately. The acquisition of multiple signals for a single surface location is also visualised in Fig. 2.2.

The resulting single radar image has the same spatial resolution as would be achieved with a theoretically larger aperture. Such images are often acquired in so-called *swaths*, which denotes a strip on the surface repeatedly illuminated by the beam bursts and scanned as the satellite travels along its orbit. The synthetic aperture corresponds to the distance travelled by the satellite while a given point on the surface is in view of the beam bursts.

Interestingly, the beam width of a burst increases with range distance, and the amount of time a point location stays within the burst's view increases simultaneously. Therefore, the synthetic aperture for a point location further distanced in range direction from the satellite is larger than for a point location closer to the satellite. This variation in the aperture's size has the effect that the resulting image has a uniform pixel resolution in both azimuth and range direction independent of the distance of the corresponding location relative to the satellite.

SAR systems often rely on the radar signal's polarisation to obtain more detailed information about the surface. Different ground geometries produce various backscatter signals depending on the polarisation type of the original signal, enabling a greater variety of detail to be captured simultaneously.

The fundamental information of SAR images relies on the fact that varying properties of the Earth's surface lead to different backscatter behaviour. Rough surfaces, such as a forest cover, create a lot of signal scattering when hit by a radar signal with a large amount of radiation sent back to the sensor. Surfaces with flat textures, however, mainly reflect the signal away from the receiver. Therefore, rough surfaces usually appear bright in the resulting images, while flatter ones appear darker. However, variations in the signal strength are common because changes in the angle between the surface and the signal direction can considerably affect the backscattered signal strength. In rare cases, such as the perfect alignment of a mountain slope and the burst beam angle, a flat surface might reflect most of the burst's radiation back to the receiver resulting in high signal strength for a smooth surface.

Therefore, SAR measurements are not necessarily similar for the same ground geometry when imaged from different angles. To provide another example, when a satellite observes a mountain from one side, it can eclipse a large area behind it for which no measurements would be available. Consequently, the same terrain can lead to very different SAR images for varying viewing angles, which is a crucial property to consider when interpreting SAR data. The whole procedure is also inherently very susceptible to noise from multiple sources. Such noise can, for example, come from backscatter interference or complex scattering behaviour leading to a delayed response compared to other time steps.

The various properties of SAR imaging introduce much complexity to interpreting the resulting image data. This is where deep learning methods can aid the process of analysing such data due to their inherent ability to detect and handle complex dependencies in input data. Fig. 2.3 shows some example SAR images from the BigEarthNet [14] dataset.

SAR imaging systems possess a multitude of advantages over classical optical camera systems. Firstly, due to their active component, they can operate entirely independent of illumination by the sun. Additionally, SAR systems are not disrupted by cloud cover as microwave radiation can pass through cloud layers undisturbed. Therefore, such systems are suited for applications where reliability is a crucial requirement. For specific wavelengths, the radiation can even partly penetrate the upper layers of soil and reveal submerged artificial and natural structures. However, the amount of information a SAR system can obtain about a certain location on the ground is limited. While optical systems measure the reflected radiation at multiple wavelengths, SAR systems are mostly restricted to a specific wavelength chosen when building the system. Additionally, the resulting images suffer from a higher amount of noise than optical data, which can also be seen in Fig. 2.3. The increased complexity in processing and interpreting the data compared

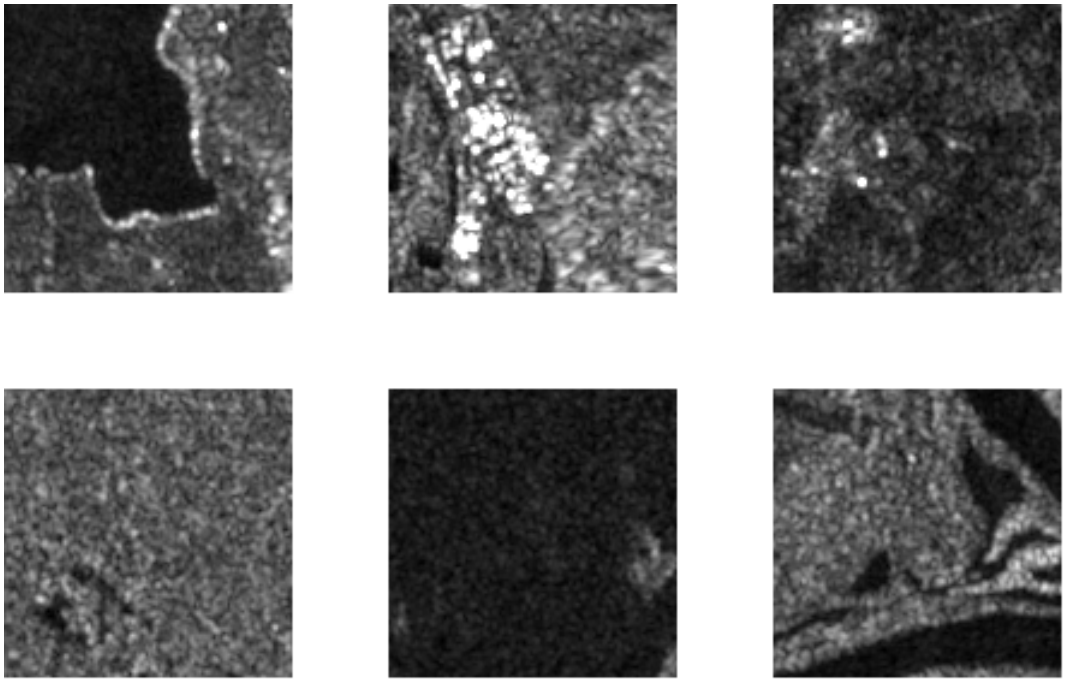


Figure 2.3: Visualisation of some example SAR images from the BigEarthNet dataset. Only images received at vertical polarisation are shown.

to standard spectral imaging can also present a significant challenge.

However, their reliability and ability to obtain structural information can be crucial for many applications [18]. SAR imaging systems are, therefore, an invaluable source of data widely utilised in many modern research fields.

3 Vision Transformer Architecture

The ViT architecture [10] has garnered significant interest from the research community since its original proposal. It achieved an outstanding classification performance on the ImageNet dataset [11] at its release.

This chapter provides a detailed description of the various components of the ViT architecture, as it forms the primary underlying model of the multi-modal fusion methods analysed in this thesis. First, the Transformer Encoder as its main building block is introduced, accompanied by a mathematical definition of the attention mechanism. After that, the ViT architecture is explained in detail. Finally, the chapter concludes with a discussion of the general advantages of the ViT architecture.

3.1 Transformer Encoder

The Transformer Encoder constitutes one of the central processing blocks of the original Transformer architecture. It was proposed by Vaswani et al. [12] and is originally used to compute attention-based feature representations for sequential input data from the Natural Language Processing (NLP) domain. However, the Transformer Encoder can be extended to process arbitrary sequential data representations, which gave rise to multiple Transformer-based architectures in other fields. Notable examples include the ViT architecture utilised in this thesis, as well as the Audio Spectrogram Transformer [19] for audio data or the Point Cloud Transformer [20].

The sequential inputs to the Transformer Encoder are formed by so-called *tokens*. A token is a vector embedding generated from the input data to represent an abstract subset of the input's feature information. For NLP tasks, each token usually represents a singular word from a sentence. However, various input data types, such as images, can be represented as sequences of token embeddings [10, 19, 20].

The attention mechanism utilised in the Transformer Encoder functions by computing attention scores reciprocal between all the tokens in the input sequence. The resulting scores represent the relevance of the relationship between any two tokens for the task the model is trained to solve. The Transformer Encoder specifically utilises the so-called *Scaled Dot-Product Attention* function proposed by Vaswani et al. [12]. Throughout this thesis, the term Scaled Dot-Product Attention will be abbreviated to *Attention* for the sake of simplicity.

To formally define the Attention function, let $l \in \mathbb{N}$ be the number of token embeddings in a sequence and let $d_e \in \mathbb{N}$ be the size of each token embedding.

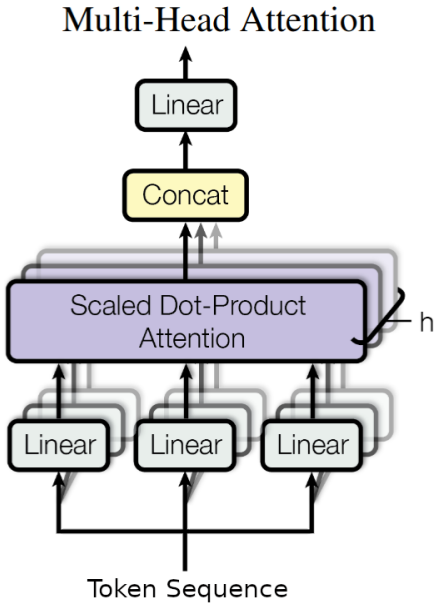


Figure 3.1: Visualisation of the Multi-Head Self-Attention procedure. The Scaled Dot-Product Attention block refers to the computation of Attention as defined in Eq. 3.1. h denotes the number of attention heads [12].

With this, let $Z \in \mathbb{R}^{l \times d_e}$ be a matrix representation of an input sequence with each row corresponding to a token embedding in the input token sequence. Let then $Q, K, V \in \mathbb{R}^{l \times d_e}$ be matrices obtained by applying learned linear transformations to Z with weight matrices $W^Q, W^K, W^V \in \mathbb{R}^{d_e \times d_e}$. Q, K and V are respectively referred to as *query*, *key* and *value*. With this, the Attention function can be defined as seen in Eq. 3.1.

$$\text{Attention}\,(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_e}}\right)V \tag{3.1}$$

As seen in the definition of Attention, Q and K^T are multiplied by matrix multiplication, resulting in an intermediary result matrix with dimension $l \times l$. It can be observed that this intermediary result directly assigns a score to each token to token relationship. By scaling with $\sqrt{d_e}$ and applying the Softmax function, these scores are mapped to a range between $0-1$. The resulting matrix is then multiplied with V, which combines the computed scores with features directly derived from the input tokens. This procedure strengthens features with higher relevance scores while less relevant features are diminished. After multiple layers, the repeated selection of the most relevant features should influence the model to only attend to the most pertinent tokens for the task the model is trained to solve.

The entire procedure of computing the attention scores can be scaled to perform Multi-Head Self-Attention (MSA) [12] which consists of multiple Attention calcu-

lations in parallel. MSA maintains multiple sets of linear transformation layers to map the same token sequence to different query, key and value matrices and independently compute attention scores for them. Each layer of distinct learned linear transformations and corresponding Attention computation is referred to as an *attention head*.

Let $h \in \mathbb{N}$ be a model's number of attention heads. With this for each attention head $i \in [1, \cdots, h]$ a set of linear transformations can be defined with independently learned weight matrices W_i^Q,W_i^K, $W_i^V \in \mathbb{R}^{d_e \times d_e}$. By applying each of these linear transformations to Z distinct $Q_i, K_i, V_i \in \mathbb{R}^{l \times d_e}$ can be obtained for each attention head i. The attention scores are then generated separately for each attention head i by computing the Attention function with the corresponding Q_i, K_i and V_i matrices. This allows each attention head to learn filters to attend to different properties of the input, which can boost the overall performance of the model. Fig. 3.1 visualises the computation of multiple Attention calculations for multiple heads.

In practice, the weight matrices W_i^Q,W_i^K and W_i^V of each attention head are combined to form unified weight matrices $W_c^Q, W_c^K, W_c^V \in \mathbb{R}^{d_e \times (d_e * h)}$ with h referring to the number of attention heads. Applying the linear transformations corresponding to these weight matrices to Z generates matrices $Q_c, K_c, V_c \in \mathbb{R}^{l \times (d_e * h)}$. Q_c, K_c and V_c can be split along their second dimension into h matrices to obtain Q_i, K_i and V_i for each attention head. Combining the linear transformations into one matrix multiplication in the previously described manner allows the MSA layer to compute the attention scores for all attention heads in parallel.

The tokens generated by the MSA layer are then passed through a Multilayer Perceptron (MLP) which consists of two linear layers with one hidden dimension. Each feature token in the output sequence from the Attention function is passed through the MLP layer individually, with weights shared for all tokens generated by a specific attention head.

Additionally, Layer Normalisation (LN) [21] is applied in each Transformer Encoder layer before passing the inputs to the MSA and the MLP layer. LN computes normalisation statistics for the individual feature tokens in a sequence and maps the feature values in the tokens to follow a normal distribution. LN improves the performance of a model due to advantageously influencing gradient computation during gradient descent and has been shown to improve training stability and generalisation capabilities [21]. To further improve performance, residual connections [6] are employed around the MSA and the MLP layer. Residual connections can simplify the optimisation during gradient descent and aid in preventing the problem of vanishing gradients which could hinder overall performance when scaling the number of layers in the Transformer Encoder. A detailed overview of the aforementioned components of the Transformer Encoder is shown in Fig. 3.2. The Transformer Encoder forms the primary building block of the ViT architecture introduced in the next section.

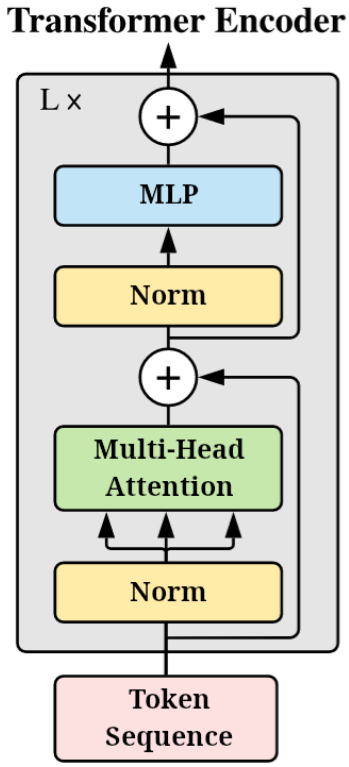


Figure 3.2: The Transformer Encoder constitutes one of the main building blocks of the Transformer architecture. It consists of a sequence of layers, each containing an MSA layer followed by an MLP layer. Before each of these layers, a normalisation operation is applied, and residual connections are introduced. The image shows one layer in the Transformer Encoder, which is stacked multiple times to form the full architecture [12].

3.2 Vision Transformer

The Vision Transformer architecture [10] is a neural network architecture specifically designed for the task of image classification. Fig. 3.3 gives an overview of the different components of the architecture. It is an adaptation of the original Transformer architecture proposed by Vaswani et al. [12] for the NLP domain.

The ViT architecture operates on a sequence of feature tokens derived from images by a specialised embedding layer. Unlike words or other data representations, images consist of structured information in the form of pixels, which do not directly translate to a form of sequential input as is required by the Transformer architecture.

Therefore, in the ViT model, an embedding layer processes an input image by

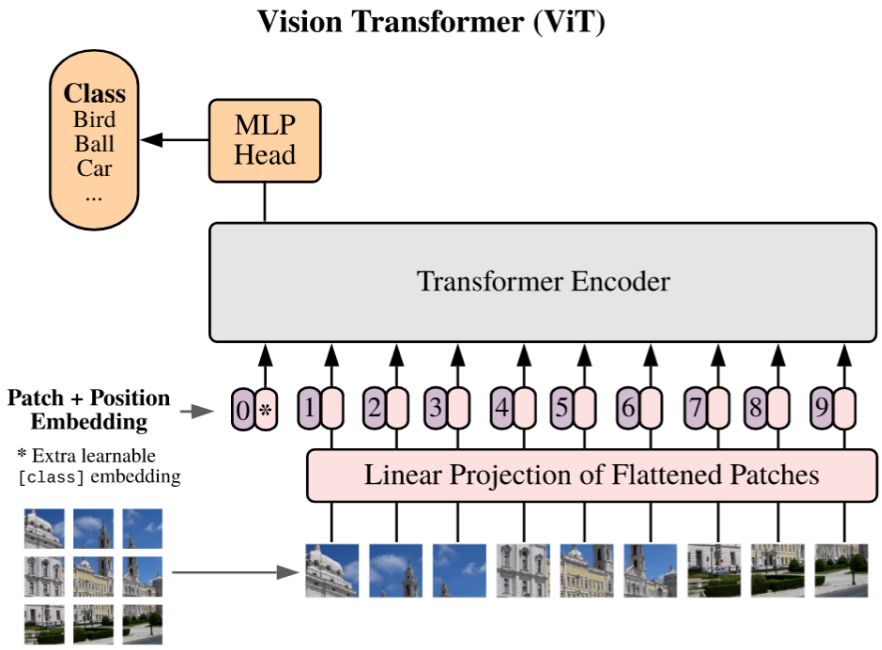


Figure 3.3: A general overview of the ViT architecture. The embedded patch tokens and a class token are passed through a Transformer Encoder. The final class token is utilised for classification by an additional MLP [10].

dividing it along its spatial dimensions into a predefined number of patches. For a given image $x \in \mathbb{R}^{h \times w \times c}$ with $h, w \in \mathbb{N}$ being the with and height of the image and $c \in \mathbb{N}$ denoting the number of channels in the image, the resulting patches $x_p \in \mathbb{R}^{p \times p \times c}$ will have a spatial dimension of $p \in \mathbb{N}$ so that $\frac{h}{p} \in \mathbb{N}$ and $\frac{w}{p} \in \mathbb{N}$.

Each of these patches is then processed with a linear transformation to derive a patch embedding token $z \in \mathbb{R}^{d_e}$ with $d_e \in \mathbb{N}$ denoting the embedding dimension. The embedding token condenses the information present at the patch's respective image location into a lower dimensional feature space. In addition to the patch tokens, a special learnable class token $z_{cls} \in \mathbb{R}^{d_e}$ is defined and appended to the token sequence. After processing the patch tokens and the class token in consecutive attention layers only the class token is extracted and employed for the final classification step. This forces the network to map the most relevant information from the patch tokens to the class token. Due to the limited size of the class token, mapping all information to it is intended to have a condensing effect and lead to more accurate classification results. The entire input sequence to the Transformer Encoder can hence be defined as $Z = [z_1, z_2, \cdots, z_l, z_{cls}]$ with $l \in \mathbb{N}$ referring to the number of patches generated from an input image.

To assist the model in encoding the positional information of the patch tokens in the input sequence, a positional encoding is added to all individual tokens.

The positional encoding ensures that the model can easily discern the location of a specific patch and aids in deriving spatially related information from the whole sequence. It is implemented by a learnable vector with the same size as the embedding dimension d_e, which is added to all patch tokens and allows the architecture to learn an encoding for the position of a token in the sequence.

The resulting sequence of tokens directly derived from the input data is then passed through a Transformer Encoder, which forms one of the main components of the original Transformer model and has been introduced in Section 3.1. The Transformer Encoder performs a computation of the attention mechanism in each of its layers. It outputs a sequence with the same length as the input sequence, with the features now corresponding to higher-level information about the semantic content of the input image. The class token is then extracted from the final output sequence and individually utilised to perform the classification, while the rest of the tokens are discarded. Next, the class token is processed by a dedicated classification MLP called the *classification head*. The classification head maps directly from the embedding dimension d_e of the class token to the class dimension $d_{class} \in \mathbb{N}$ which is equal to the number of classes present in the current classification task. Finally, a Softmax function is applied to these class logits to obtain a confidence score for the presence of a specific class in the input image. For multi-label classification tasks, as is performed in this thesis, this Softmax function is replaced with a Sigmoid function.

A significant modification is introduced in the Transformer Encoder used by the ViT model. While the standard Transformer relied on the more simplistic Rectified Linear Unit (ReLU) function [22] in its encoder, in the ViT architecture, this is replaced with the Gaussian Error Linear Unit (GELU) activation function. The GELU function smooths the transition activation outputs around zero and exhibited the potential to increase the performance of models on various deep learning tasks [23].

When the ViT model was initially proposed, it achieved a performance improvement on the ImageNet dataset [11] compared to rivalling methods. Remarkably, the ViT model was the first convolution-free architecture to outperform the more established CNN architectures dominant in the field of computer vision in the last decade. However, it required extensive pretraining of the model on a very large-scale image dataset before fitting the model on the ImageNet dataset. The additional pretraining demanded a vast amount of additional processing resources. The significant computational requirements are caused in particular by the complexity of the Attention function. With l, d_e denoting the length of the processed sequence and the embedding dimension, the complexity of the Attention function is $O(l^2 * d_e)$. Therefore, the computation scales quadratically with the length of the input token sequence.

When CNN models were first proposed, their high performance on visual data was partially explained by their strong local inductive bias imposed on the model due to the limited number of pixels observed by any convolutional filter. This was deemed

beneficial because it allows deeper layers to learn more abstract features than shallow layers. In contrast, the ViT architecture exhibits a very weak local bias and attends to all input tokens equally from the first layer onward. The computation of attention scores between all patch tokens allows the model to immediately correlate features from the entire image without first having to abstract local features from a specific image location. Such a property might be beneficial when analysing multi-modal data in the field of RS because images captured by different sensors can exhibit significant levels of correlation at the feature level. Additionally, because the Attention function inherently assigns relevancy scores to components of an input sequence, it could potentially be extended to filter the most relevant information provided by multi-modal data. Therefore, the ViT model serves as the main building block on which all multi-modal fusion methods investigated in this thesis are based.

4 Related Work

The following sections give an overview of various approaches to performing multi-modal fusion within deep learning architectures. The first section provides an overview of the state-of-the-art in deep-learning-based multi-modal fusion of remote sensing data. Specifically, the description focuses on approaches dedicated to scene classification or LULC classification tasks, as these relate most closely to the scene classification task of interest in this thesis. Afterwards, an outline of various state-of-the-art Transformer-based multi-modal fusion methods from the general deep learning field is provided to embed the work of this thesis in the broader research landscape.

4.1 Multi-Modal Fusion in Remote Sensing Image Classification

In the field of RS, data captured by different sensor systems are abundantly available. Such sensor modalities are commonly used in various applications due to the distinct types of information they can provide for a geographical location. However, combining multiple modalities effectively still presents a challenging subject in modern remote sensing research. Therefore, multiple different methods have been developed to effectively achieve the fusion of separate RS modalities [2].

Some early works in the field have often relied on statistical methods [24, 25] to fuse features from separate modalities, while others employed shallow neural networks for this purpose [26, 27]. Other approaches that have received significant attention are direct subspace learning from multi-modal data representations [28, 29] and manifold alignment methods, which seek to map different modality representations of the same underlying physical source to a common representation on the same manifold [30, 31].

However, with the recent success of deep-learning-based methods in the field of RS, most research has focused on fusion techniques fully integrated with deep learning models. Many of these fusion methods rely on CNN architectures due to their outstanding performance in the CV field for many years.

Hang et al. [32] propose a relatively shallow CNN model with two coupled modality encoders for the fusion of Hyper-Spectral Image (HSI) and Light Detection and Ranging (LIDAR) data. In both encoders, the last layers share weights to produce more similar feature encodings between the modalities and reduce the model size.

Other publications have used an encoder-decoder design to map multi-modal

inputs to a shared latent space representation of lower dimensionality than the input features [33, 34]. These methods often perform an additional step during the training phase, where the original inputs are reconstructed from the fused feature embedding. The reconstruction forces the model to learn a mapping that preserves as much information as possible in a condensed latent space.

A very different approach is explored by Li et al. [35], who conceived a fusion model where feature embeddings are extracted from HSI and SAR data by the same encoder. During training and inference and depending on the relevance of a channel feature to the current classification operation, a corresponding feature from the other modality is injected to substitute that feature. This reduces redundancy and has subsequently been shown to improve classification performance compared to other models.

Another method is investigated by Audebert et al. [36], who use a fully convolutional architecture to achieve the fusion of multi-spectral and LIDAR data for LULC classification. The respective modality encoders are modified to include a reoccurring fusion layer between convolutional layers in the encoder to exchange information from both modalities perpetually. Therefore, the output from one encoder is assumed to contain a fused representation of both modalities and is passed to a decoder to construct an LULC map. Comparisons to a late fusion model indicate that integrating such sophisticated fusion steps into a model architecture can significantly improve the performance capacity of a deep learning model.

Hong et al. [37] conduct a comprehensive comparison between multiple CNN-based fusion methods for the fusion of multi-spectral and SAR data. These fusion methods include concatenation-based early, middle and late fusion but also a more sophisticated encoder-decoder fusion model and a *cross fusion* method which functions by sharing weights between the CNN encoders for the different modalities. The conducted experiments indicate that models with more sophisticated fusion approaches directly integrated into the model architecture can outperform methods relying on more straightforward fusion approaches.

Due to the beneficial properties of the attention mechanism for the fusion of modalities, many works have sought to incorporate it into the fusion process. Furthermore, because of the general prevalence of CNN architectures in RS, various models have been proposed which integrate attention-based fusion modules with CNN models in a limited capacity [38, 39].

Other works have focused on integrating Transformer models to specifically perform the fusion of modalities, while dedicated CNN encoders are still used to derive feature embeddings from input data. Such a model is proposed by Fan et al. [40], which employs separate encoder modules to extract features from high-resolution RS images and population movement data. A standard Transformer Encoder then directly performs the fusion of the extracted features on a combined input sequence derived from both modalities. In a more sophisticated manner, Ma et al. [41] conceive a specialised Transformer-based model with adapted attention layers to better facilitate the fusion of multi-modal and multi-scale features. These

models receive features derived from CNN encoders as inputs.

However, due to the recent success the ViT architecture achieved in the classical CV domain, much research has been focused on the multi-modal fusion of RS data solely relying on Transformer models in all steps of the processing pipeline. Wang et al. [42] tested the capacity of the standard ViT architecture to perform self-supervised learning on the BigEarthNet-MM dataset [14] for a scene classification task. The fusion is performed by concatenating the channel information from Sentinel-1 (S1) and Sentinel-2 (S2) patches. The embedding generation procedure from the standard ViT architecture is then used to extract patch embeddings from the combined image tensors. The results indicate that Transformer-based self-supervised learning could potentially match or outperform supervised training on RS datasets.

Xue et al. [43] propose another approach to self-supervised learning with Transformers on multi-modal RS data. Here the authors train a Transformer-based encoder-decoder architecture to map HSI and high-resolution RGB data to feature embeddings in a constrained latent space. The model is pre-trained in a self-supervised manner by masking parts of the input images and reconstructing the original images from the feature embeddings. After pre-training, a classifier is trained on these feature embeddings to generate accurate classification maps of the respective area.

A supervised approach to the fusion of LIDAR and HSI data has been conceived by Roy et al. [44]. While the token embeddings from the HSI data form the main input sequence to a Transformer model, the information from the LIDAR data is injected through a modified class token. This class token is generated from the LIDAR data instead of being initialised randomly, as would be the case in the standard Transformer model. The modalities are then fused in the attention layers of a slightly modified Transformer Encoder which incorporates an adapted attention computation between the class token and the feature tokens.

Another more classical fusion model is conceived by Xue et al. [45]. Their architecture contains multiple modified Transformer encoders dedicated to specific modalities which map the multi-modal inputs to an encoded embedding token sequence. The resulting sequences are concatenated and fused by a specialised attention operation in the last layer before employing a linear classifier to obtain LULC maps.

It can be observed that a considerable amount of fusion methods have been conceived specifically for RS applications. However, many relevant approaches to Transformer-based multi-modal fusion are designed for applications from other deep learning fields. Such approaches could provide valuable insights on how to best analyse multiple modalities with Transformer-based models. Therefore, the following section provides an overview of relevant works on multi-modal fusion based on the Transformer architecture dedicated to tasks outside the field of RS.

4.2 Transformer-based Multi-Modal Fusion

This section introduces a multitude of Transformer-based methods to perform multi-modal fusion in the general deep learning domain to obtain insights on the applicability of Transformers in multi-modal contexts outside the RS domain. In the general deep learning field, the term modality usually refers to different types of data representations corresponding to similar underlying information such as text, audio or image data. Data from many such modalities are available in large quantities and utilised for various types of tasks [46].

However, the definition above of the term modality differs slightly from the definition used throughout the rest of this thesis. Therefore, in this section, the term modality refers to the definition used in the general deep learning domain instead of the stricter definition as a sensor modality.

Fusing different modalities is a highly researched subject within the general deep learning community [42, 46–48]. However, to effectively fuse modalities such as images or text, sophisticated embedding procedures are required to map their inherent information to feature representations with similar dimensionality.

Interestingly, Transformer-based architectures could demonstrate impressive results on various tasks while processing different types of modalities [10, 19, 20]. This could imply that the design of the Transformer architecture is inherently modality-agnostic, and its internal layers can adapt to various representation types for different input sequences, as has been hypothesised by Xu et al. [13].

Due to the abundance of modality-specific Transformer models, many procedures exist to map different modalities to standard Transformer input embedding tokens, simplifying the integration of different modalities for multi-modal learning. Therefore, some fusion approaches map two modalities to sequences of feature tokens that can be directly concatenated to serve as an input to the standard Transformer model. Shvetsova et al. [49] employ such an approach to fuse video, audio and text data. The authors generated multiple latent space representations from the input modalities and utilised a contrastive loss to maximise the similarity for the same sample. Relying on a similar principle, Gabeur et al. [50] conceive a multi-modal Transformer for video retrieval tasks. A different approach is explored by Yao et al. [51]. The authors propose a model which performs the fusion of text and image modalities during the self-attention computation. While the query contains information from both modalities, the key and value vectors are only derived from the text modality.

Numerous other works substitute the self-attention computation step in some or all of the Transformers layers with a Cross-Attention step [47, 52–55]. Cross-Attention is a type of attention computation where the query vector is derived from one modality while the key and value vectors are derived from another modality. For a detailed description the reader is referred to Section 5.5. It represents one of the most prevalent methods for Transformer-based multi-modal fusion due to the simplicity of incorporating it into the standard Transformer architecture.

Additionally, Cross-Attention presents an intuitive approach to fusing separate modalities because it directly computes relevancy scores between distinct token sequences corresponding to different respective modalities.

A notable architecture mainly relying on Cross-Attention is the Perceiver IO architecture proposed by Jaegle et al. [47]. Perceiver IO is designed to perform the fusion of arbitrary types of modalities by mapping all modalities to a joint latent space embedding of similar dimensionality. Most computations are then performed directly in the latent space to reduce computational demand. A standardised latent size simplifies the integration of multiple modalities and allows the architecture to process both uni-modal and multi-modal inputs. A multitude of experiments proves the capacity of the architecture to adapt to different modalities while retaining the ability to achieve competitive results on various tasks.

A radically different multi-modal fusion method is proposed by Nagrani et al. [48], which adds specialised bottleneck tokens to facilitate the fusion between RGB and spectrogram images. Both modalities are mapped to sequences, with each token in these sequences having the same embedding dimension. Additionally, a class token is appended to each modality sequence. The arrangements are then concatenated, and bottleneck tokens are added to the combined sequence input. While the architecture is mainly based on the default Transformer Encoder, it is modified to only allow information flow between both modalities through the bottleneck tokens. Therefore, the bottleneck tokens have the effect of condensing and exchanging the most relevant information from both input modalities. Finally, the class tokens corresponding to each modality are combined by averaging and then utilised by a linear classifier to perform the final classification. The results show that the addition of bottleneck tokens can significantly increase the performance and decrease processing requirements on multi-modal tasks compared to a more simplistic early fusion of modality-specific tokens.

As can be seen, many adaptations of Transformer-based fusion methods have been conceived for various deep learning tasks. The multi-modal fusion methods introduced in this thesis will incorporate and advance upon some of the procedures above to facilitate the fusion of multi-spectral and SAR images.

5 Multi-Modal Fusion Methods

Some modifications to the standard information flow in the model are required to apply the ViT architecture to multi-modal input data. As multi-modal information is present in different data representations, it is not trivial to combine these into a joint representation. This necessitates introducing additional processing steps to fuse modalities and obtain such a joint representation for further processing by a deep learning model. The fusion can be performed at either a low feature level or with abstract features derived from multiple preceding deep learning layers.

Multiple multi-modal fusion methods based on the ViT architecture are conceived and investigated in this thesis to assess their capabilities in fusing multi-spectral and SAR image data for a classification task. These methods can be separated into two categories:

1. Fusion methods, which modify the generation of token embeddings but rely on the standard ViT architecture to perform the classification task.

2. Methods which improve the ViT architecture design by employing separate modality encoders and introducing various techniques to influence and facilitate the inter-modal information exchange between them.

The investigated methods have different benefits and can vary considerably in terms of complexity, performance capabilities and computation requirements. The following sections give a detailed introduction of each method and provide an analysis of its intended effects.

5.1 Early Fusion by Modality Channel Concatenation

Early fusion approaches represent probably the most simplistic and widely adapted fusion method in many neural-network-based fusion approaches for multi-modal data sets. It usually refers to the concatenation of unprocessed modality inputs or low-level features derived from these before analysing them with a model architecture. Consequently, an Early Fusion approach is also investigated in this thesis for the fusion of the multi-spectral and SAR images. The same method has previously been employed by Wang et al. [42] for self-supervised multi-modal classification on the BigEarthNet-MM dataset.

The procedure is especially viable when both modalities can be easily projected into a feature space with the same dimensionality. In the case of the multi-spectral

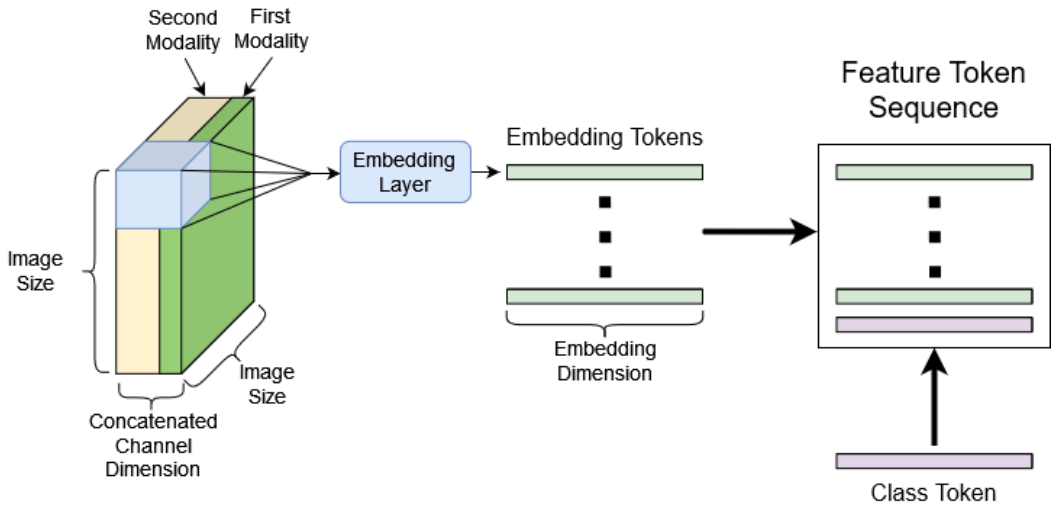


Figure 5.1: The token generation process for the Early Fusion approach. A patch from the concatenated modalities tensor is mapped to an embedding token by a linear embedding layer. The resulting tokens are assembled into one sequence and a class token is appended. The feature token sequence forms the input to the Transformer Encoder.

and SAR data utilised in this thesis, they are both available at a similar spatial dimension. While some bands of the multi-spectral data require an up-sampling and interpolation step to match the spatial dimension of all other bands, no information loss should occur during this step. Therefore, a direct concatenation can combine both modalities along their channel dimension. The concatenated input tensor can then be processed by the default patch embedding layer employed in the ViT architecture. The resulting feature tokens represent the combined information from both modalities at a specific patch location.

Because the mapping from inputs to feature embeddings is performed by a standard linear layer, there are two possible ways for data fusion to occur. To some degree, the modality-specific information can already be merged by the linear filter weights to produce combined features in the token embeddings. Conversely, the linear filters might learn to map information from both modalities to different feature positions in the token embedding vector. This would then leave the fusion process of the modality information to the attention mechanism in the encoder layers. Most likely, the fusion of modalities is performed by both of these processes simultaneously, as it is difficult to determine how exactly information is processed in a trained architecture. Fig. 5.1 visualises the working principle of the Early Fusion method.

To define the process formally, let $h, w \in \mathbb{N}$ be the height and width of the input images and let $c_{mod1}, c_{mod2} \in \mathbb{N}$ be the channel dimension of the respective modalities to be fused. Let further $x_{mod1} \in \mathbb{R}^{h \times w \times c_{mod1}}$ and $x_{mod2} \in \mathbb{R}^{h \times w \times c_{mod2}}$ be

the images from the two separate modalities. The combined input to a model can then be defined as $x \in \mathbb{R}^{h \times w \times (c_{mod1}+c_{mod2})}$ whereas the channel information of x consists of the channel information of x_{mod1} and x_{mod2} concatenated along the channel dimension. Each patch embedding token is then derived from an input patch $x_p \in \mathbb{R}^{p \times p \times (c_{mod1}+c_{mod2})}$ with $p \in \mathbb{N}$ again denoting the spatial dimension of a patch.

The patches are processed by a set of linear filters similarly to the process for a single modality as described in Section 3.2. It should be noted that the number of learned filter weights increases to account for the additional channels compared to single modality processing. The information of the combined modalities can either be mapped to a feature embedding with the same dimension used when training with single modality data or the feature dimension can be increased to account for the additional information. However, increasing the embedding dimension also influences other layers in the architecture, hampering the comparability between models.

To conclude, the Early Fusion method presents a straightforward and efficient approach to fusing MSI and SAR data. However, the combined generation of embeddings from both modalities might limit the beneficial properties of the attention layers in the fusion process. Therefore, the next sections present further embedding-based fusion methods which fully incorporate the attention mechanism into the fusion process.

5.2 Proposed Modality Token Fusion

By separating the generation of embedding tokens between the two modalities an additional simplistic fusion method can be designed. Such a Modality Token Fusion method is proposed in this thesis for the fusion of multi-spectral and SAR images. It operates by generating the token embeddings for each modality separately with the same process as is described in Section 3.2.

The generated tokens are then concatenated to form the input sequence to a Transformer Encoder. This avoids that features from both modalities are fused at the patch embedding step as can occur in the aforementioned Early Fusion approach. Therefore, all fusion has to take place within the attention layers of the Transformer Encoder. Because the Attention function measures the relative importance of all token embeddings relative to one another, separating the token generation step for the input modalities could potentially improve the selection of the most relevant features from each modality. While the direct fusion by concatenation of tokens generated from modalities has been explored for other types of modalities [56, 57], to the best of my knowledge this work is the first to apply such an approach to the fusion of multi-spectral and SAR images.

Additionally, the patches from each modality do not have to share the same embedding token due to separating the token embedding step. This allows more

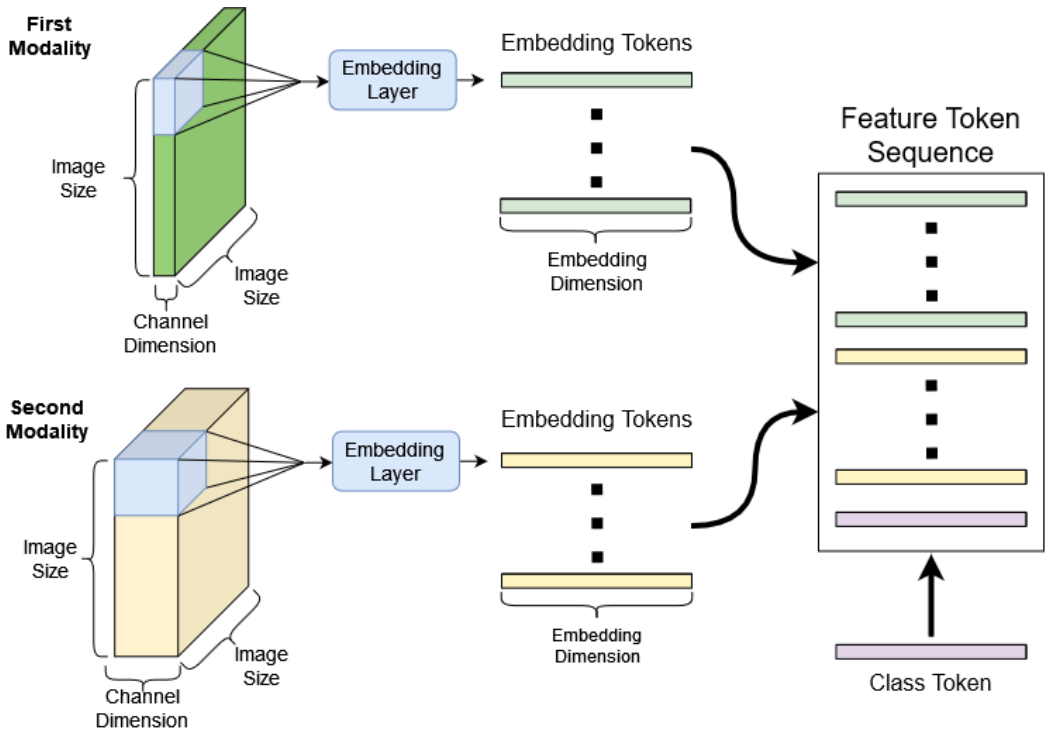


Figure 5.2: The token generation process for the Modality Token Fusion approach. Separate patches from each modality are individually mapped by a modality-specific embedding layer to embedding tokens. The resulting token sequences corresponding to each modality are combined and a class token is appended to form the feature token sequence. The resulting sequence serves as the input to the Transformer Encoder.

features to be passed on to the attention layers. However, processing more tokens in the attention layers comes at the cost of increasing the computational requirements. Since Modality Token Fusion generates twice the number of tokens as Early Fusion, its processing requirements should significantly increase when a similar embedding dimension and patch size are used.

To formally define Modality Token Fusion let again $x_{mod1} \in \mathbb{R}^{h \times w \times c_{mod1}}$ and $x_{mod2} \in \mathbb{R}^{h \times w \times c_{mod2}}$ be the two modalities with $h, w \in \mathbb{N}$ being the height and width and $c_{mod1}, c_{mod2} \in \mathbb{N}$ denoting the channel dimension of each modality.

Each patch embedding token is then derived from an input patch either defined as $x_p^{mod1} \in \mathbb{R}^{p \times p \times c_{mod1}}$ or as $x_p^{mod2} \in \mathbb{R}^{p \times p \times c_{mod2}}$ with $p \in \mathbb{N}$ again denoting the spatial dimension of a patch. The token sequence, which serves as an input to the ViT model, is then formed by concatenating all x_p^{mod1} and x_p^{mod2} and appending a class token at the end. Fig. 5.2 also visualises the generation of the token embeddings from the input tensors.

In summary, Modality Token Fusion is built on the simple concatenation of tokens

generated from separate MSI and SAR modalities. This fully allows the attention mechanism to select the most relevant features from both modalities which could prove advantageous, especially when compared to Early Fusion. The principle can be further expanded to directly let the attention mechanism perform the fusion of the individual channel information as is introduced in the following section.

5.3 Proposed Channel Token Fusion

An additional embedding-based fusion method is proposed with the Channel Token Fusion method. It functions by dividing patches utilised for the generation of the feature tokens not just along the spatial but also along the channel dimension. A separate filter is applied to each channel of a patch from the multi-channel image of one modality to derive a token embedding independent from all other channels. For multi-spectral data, this corresponds to the individual spectral bands, while for the SAR data, this corresponds to the different polarisation channels. To the best of my knowledge, utilising tokens derived from the individual channels of multiple modalities has not been employed previously for the fusion of multi-spectral and SAR images. While a similar concept is explored by Hong et al. [58], their approach is dedicated specifically to HSI analysis and does not extend to multi-modal fusion.

Again let $x_{mod1} \in \mathbb{R}^{h \times w \times c_{mod1}}$ and $x_{mod2} \in \mathbb{R}^{h \times w \times c_{mod2}}$ be the two modalities with height and width $h, w \in \mathbb{N}$ and channel dimensions $c_{mod1}, c_{mod2} \in \mathbb{N}$. Then each patch can be defined as $x_p \in \mathbb{R}^{p \times p \times 1}$ with p denoting the spatial dimension of the patch. The embedding layer effectively iterates over all patches in the spatial dimensions of both modalities and employs a separate filter for each channel to map channel patches to separate token embeddings.

Interestingly, the overall number of parameters in all channel-wise filters is the same as with the classical patch embeddings employed by the standard ViT model. At the same time, a significantly higher number of tokens is generated from the same input. The resulting tokens, therefore, correspond to a smaller number of input features respectively. Hence, the condensing of information at the embedding generation step is not as severe, which means that more information should be available in the Transformer Encoder and its attention layers. Finally, the input to the Transformer Encoder is formed by concatenating all generated channel feature embeddings and appending a class token similar to the aforementioned methods. A visualisation of the procedure is provided in Fig. 5.3.

Notably, there are some relevant downsides to the fusion of multi-modal images with channel tokens in such a manner. One of the main problems is the overly large amount of generated input tokens. The number of tokens scales directly with the number of channels present in both modalities. This induces a very high processing workload to compute the attention scores for all these input tokens because the computation of attention scores scales quadratically with the length of the input sequence. With multi-spectral data as one modality, such an approach can become

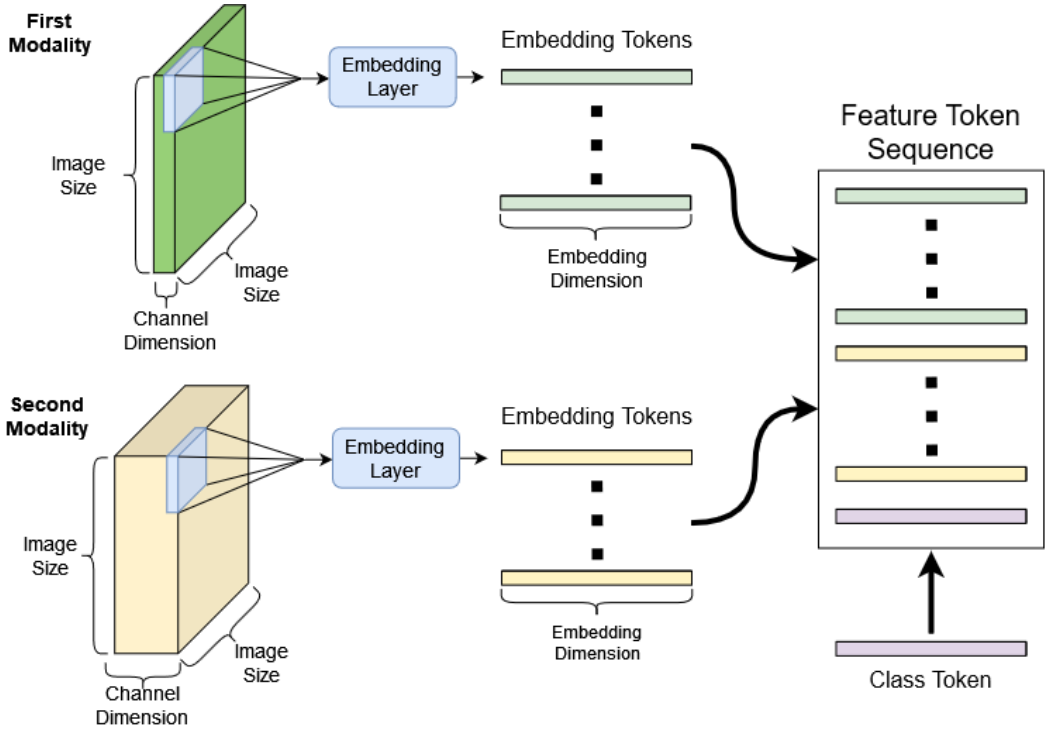


Figure 5.3: The generation of embedding tokens from patches in Channel Token Fusion. A separate token is generated for each patch and each channel of the patch separately. The final token sequence consists of tokens corresponding to each channel of patches in both modalities, as well as a class token. The resulting sequence is then processed by a Transformer Encoder.

infeasible as it leads to an explosion of memory and processing requirements.

However, Channel Token Fusion does not result in a larger model size of the utilised ViT model compared to the other aforementioned fusion methods. While the sequence length increases, the size of the individual tokens does not change, which is one of the main factors contributing to the model size. In fact, the embedding size can potentially even be reduced as a considerably smaller amount of input features has to be mapped to each token embedding. Additionally, such an amount of token embeddings could also have a regularising effect on the model's training because the same number of internal model weights are required to process a higher amount of feature inputs with a likely overall higher variance.

To summarise, in Channel Token Fusion the channel information from MSI and SAR data is fused directly in the attention layers of a Transformer Encoder. The advantageous properties of the attention mechanism might, therefore, enable a more effective fusion of the channel information compared to the other aforementioned fusion methods.

All previously introduced fusion methods specifically modify the process of generating the token input sequence but rely on a standard Transformer Encoder for the feature extraction and classification. Conversely, the following sections introduce fusion methods which employ modality-specific encoders to extract features from both modalities and incorporate various procedures to fuse multi-modal information between these encoders.

5.4 Proposed Middle Fusion with Separate Modality Encoders

Based on the idea that features of deeper layers of a neural network contain semantically more abstract and relevant information for a given input a Middle Fusion method derived from the ViT architecture is proposed for the fusion of multi-spectral and SAR images. It functions by merging the abstract feature representations generated by separate encoders dedicated to different modalities. Similar methods have previously been investigated for CNN architectures, such as the one by Hong et al. [37]. Consequently, as both CNN and ViT architectures derive more abstract features at deeper layers, the principle can also be extended to Transformer-based multi-modal fusion. While similar approaches have been conceived previously, as described by Xu et al. [13], to the best of my knowledge this is the first time such a fusion method is specifically designed for the fusion of multi-spectral and SAR image data.

The concrete architecture proposed in this thesis for multi-modal fusion of multi-spectral and SAR images is composed of three separate Transformer Encoders. The first two encoders, called *modality encoders*, are dedicated to each of the modalities and extract modality-specific features. A dedicated fusion layer performs the fusion of the feature token sequences generated by each modality encoder. This layer separately fuses the standard feature tokens and the class tokens from both encoders through a set of fully connected layers. The third encoder, called the *fusion encoder*, receives the fused feature token sequences from both modality encoders and passes these through additional attention layers. Processing the fused feature tokens with further attention layers serves to extract fused abstract features relevant to the classification task.

Fusing the class tokens separately ensures that all class-related information contained in the class tokens is preserved and does not get mixed with the other feature token information during the fusion step. Afterwards, the fused token sequence is of equal length as both input sequences received from the modality encoders. Therefore, the fusion step essentially halves the total capacity of the whole sequence, which reduces processing requirements for the fusion encoder. Simultaneously, the reduction of feature information at the fusion step might result in a selection of abstract features of higher relevance for the classification task.

Afterwards, the fusion encoder processes the fused sequence of feature tokens

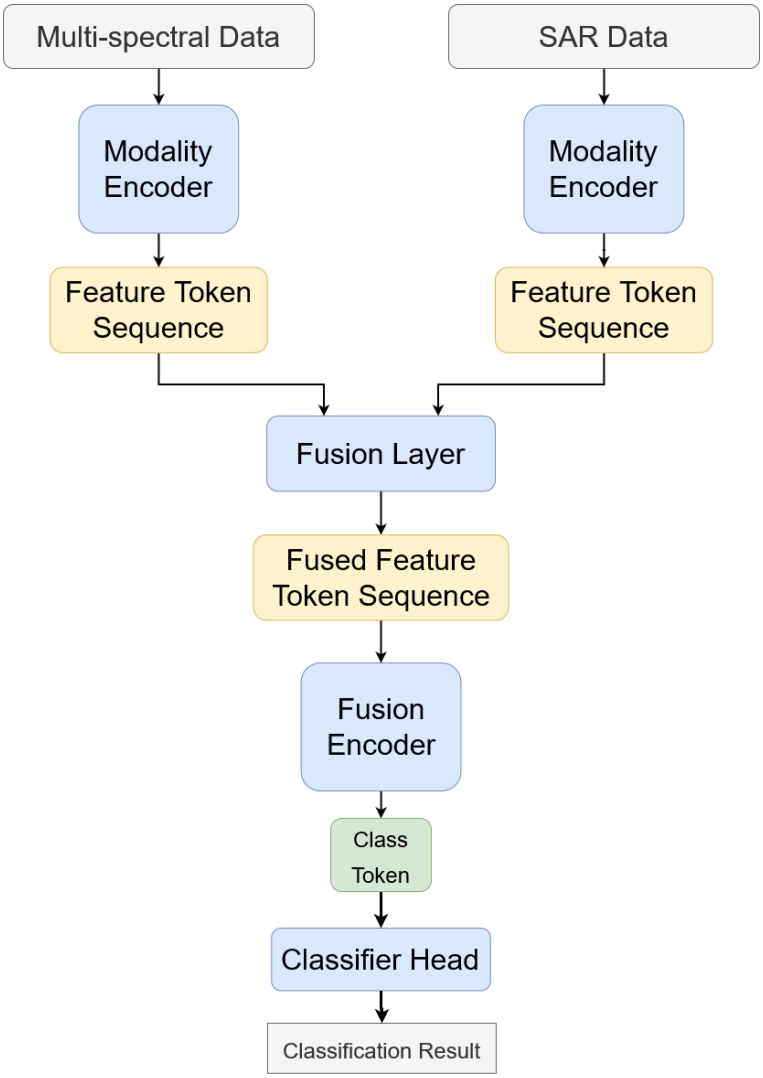


Figure 5.4: Visualisation of the architecture design for the Middle Fusion approach. The modalities are processed by the modality encoders to extract abstract feature representations. The resulting feature tokens are fused and processed by a further fusion encoder. Finally, the last class token is extracted to perform the classification.

through further attention layers. Finally, as in the standard ViT model, the last class token is extracted to perform the classification. The full architecture design is visualised in Fig. 5.4.

The number of Transformer Encoder layers can be set independently from one another in the modality encoders and the fusion encoder. The architecture can, therefore, be configured to fuse the feature tokens at different layer depths. The

fusion could, for example, be performed at an early step using small modality encoders or at a later step by employing very deep modality encoders. Setting different depths for the encoder types additionally allows for an analysis of how many Transformer Encoder layers might be required to extract sufficiently abstract features from both modalities. This might reveal relevant insights on the multi-modal fusion of MSI and SAR data in general.

In summary, the Middle Fusion method fuses information derived from two modalities by dedicated encoders in a specialised fusion encoder. Therefore, abstract features from both modalities are combined in dedicated attention layers, which is expected to positively impact the classification performance. However, further advanced architectures can be conceived that fuse modalities in between the layers of two Transformer Encoders as is shown in the following sections.

5.5 Proposed Cross-Attention Fusion

Another advanced architecture design is proposed in this thesis with a Cross-Attention Fusion method to fuse multi-spectral and SAR image data. The fusion method relies on Cross-Attention which is an established method to combine multiple modalities in deep neural networks, as has been described previously in Chapter 4. The proposed Cross-Attention Fusion approach is based in particular on the method proposed by Lu et al. [59]. To the best of my knowledge, this thesis is the first work to extend this specific implementation of Cross-Attention to the fusion of multi-spectral and SAR image data.

The Cross-Attention Fusion architecture features two separate encoders for extracting features from both modalities, with each encoder dedicated to a specific modality. However, in contrast to the standard ViT architecture, the attention layers of these encoders compute Cross-Attention instead of self-attention. Cross-Attention derives the query input to the Attention function from another input sequence than the key and value inputs.

Let z_{mod1} and z_{mod2} be sequence inputs of two separate modalities derived from two input images x_{mod1} and x_{mod2}. The generation of z_{mod1} and z_{mod2} follows the same procedure as has been introduced in Section 3.2 for each modality separately. Let then $Q_{mod1}, K_{mod1}, V_{mod1} \in \mathbb{R}^{l \times d_e}$ and $Q_{mod2}, K_{mod2}, V_{mod2} \in \mathbb{R}^{l \times d_e}$ be the respective query, key and value matrices derived from each modality. As in Section 3.1, $l \in \mathbb{N}$ denotes the number of feature embedding tokens and $d_e \in \mathbb{N}$ denotes the embedding dimension. Using Eq. 3.1, Cross-Attention can then be defined as shown in Eq. 5.1 and Eq. 5.2.

$$\text{Cross-Attention}_{mod1} = \text{Attention}\left(Q_{mod2}, K_{mod1}, V_{mod1}\right) \tag{5.1}$$

$$\text{Cross-Attention}_{mod2} = \text{Attention}\left(Q_{mod1}, K_{mod2}, V_{mod2}\right) \tag{5.2}$$

It can be seen in Eq. 5.1 and Eq. 5.2 that the information from the opposite modality encoder only influences the calculation of the attention scores through

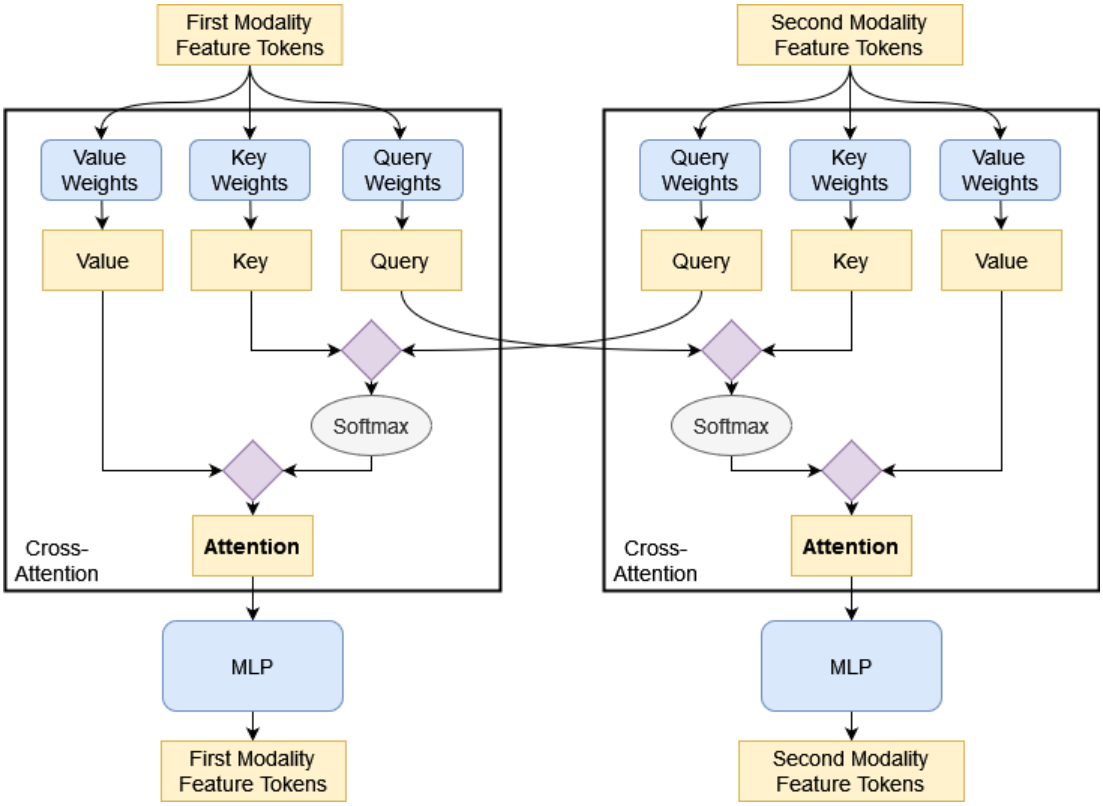


Figure 5.5: Visualisation of the information flow in the Cross-Attention layers of the two encoders employed in Cross-Attention Fusion. Note the crossing information flow from the query matrices to the opposite encoder.

the query matrices Q_{mod1} and Q_{mod2}. Therefore, through the computation of the Cross-Attention scores, both matrices only influence which features of the opposite modality encoder are strengthened or diminished in relevance. Consequently, the output feature sequence of each encoder layer is still directly derived from the respective modality of that encoder.

The aforementioned information flow is intended to facilitate a strong inter-modal information exchange to select the features from both modalities with the highest relevance for classification. However, both encoders still maintain a separate class token and extract features for classification separately. Fig. 5.5 visualises the information exchange between the Cross-Attention layers of both encoders.

While the two encoders are largely separated, it is essential that the embedding dimension of the processed tokens and the length of the token sequence remain the same for both encoders. Otherwise, the computation of the Cross-Attention scores would not be possible as the resulting matrices would not have the required shapes.

After passing the two sequences through all encoder layers, their respective class tokens are extracted. These contain highly condensed features, which should correlate with the classes present in the input patch. To combine their individual information, the class tokens are passed through a dedicated fusion layer. The fusion layer consists of a set of learned linear filters and maps both class tokens to a combined class token with the same dimension as the input tokens individually. This additional fusion step combines the highly abstract features from both modalities directly instead of just influencing the evaluation of their respective relevance, as was the case in previous layers. The final class token is then utilised by a classification head to perform the classification in the same manner as in the base ViT model by mapping the class token features to class prediction scores.

To conclude, the Cross-Attention Fusion method incorporates Cross-Attention into the fusion process which is a well-established procedure to perform multi-modal fusion as shown in Chapter 4. It utilises two encoders and employs Cross-Attention between them to exchange information about the two modalities to improve the selection of features for the classification tasks.

5.6 Proposed Synchronised Class Token Fusion

Focusing on the classification task of interest in this thesis, a Synchronised Class Token (SCT) Fusion method is proposed to fuse multi-spectral and SAR images. It relies on the dedicated fusion of class tokens attending to different modalities, by exploiting the class tokens to facilitate an information exchange between two modality-specific encoders. While both encoders maintain separate class tokens, their information is continuously synchronised between each layer of the encoders to condense and exchange the most relevant features from both modalities. To the best of my knowledge, a Transformer-based multi-modal fusion approach relying on a repeated synchronisation of class tokens has not been investigated previously in the RS domain.

In the standard ViT architecture, as described in Section 3.2, a special class token is concatenated to the feature embedding sequence generated from the input data. This class token is passed through all attention layers and correlated with all other tokens during the attention computation. However, unlike the standard feature tokens, at the final classification step, only the class token is forwarded to the classifier head. The classifier head is an MLP layer and directly maps the class token features to classification scores for each class. The idea behind this design is to condense features of high relevance for classification into a relatively restricted feature space. The same principle is exploited by SCT Fusion to facilitate multi-modal fusion for a classification task.

Employing different encoders for separate modalities, as the two aforementioned architectures did, can be very beneficial for modelling the varying features present in each modality. In such an approach, both encoders maintain separate class

tokens for both modality sequences. Therefore, the features most relevant for classification remain separated until the last layer. However, the information contained in these class tokens could already be relevant in the computation of the attention scores at earlier layers. Therefore, with SCT Fusion, a fusion method is proposed to fuse class tokens at earlier layers.

The fusion is achieved by an additional class token synchronisation step between each Transformer Encoder layer. Two separate Transformer Encoders with distinct input token sequences compute independent attention scores for the features from each modality. Each encoder receives a randomly initialised class token, which first separately passes through an attention layer of each modality-specific encoder. The attention computation maps features of relevance for the later classification from the feature tokens to each class token. Afterwards, the class tokens are extracted from the feature token sequences generated by each encoder. The extracted class tokens are then concatenated along the feature dimension and passed through a specialised linear fusion layer which maps to a token with the same embedding dimension as all other tokens in both sequences. This reduces the available space for the combined class token information and forces the model to keep only the most relevant features from both modalities.

To formally define the fusion procedure, let $z_{cls}^{mod1}, z_{cls}^{mod2} \in \mathbb{R}^{d_e}$ be the two class tokens from each modality with $d_e \in \mathbb{N}$ denoting the embedding dimension. The concatenated combined class token then becomes $z_{cls}^{concat} \in \mathbb{R}^{2d_e}$. The fusion step performed in each class token synchronisation layer can then be defined as shown in Eq. 5.3.

$$z_{cls} = W^{cls} z_{cls}^{concat} + b^{cls} \tag{5.3}$$

Here $W^{cls} \in \mathbb{R}^{d_e \times (2d_e)}$ and $b^{cls} \in \mathbb{R}^{d_e}$ refer to the learned weights and bias parameters of the linear fusion layers between each Transformer Encoder layer. The fused class token is then passed back to both encoders and concatenated to the current sequence of feature tokens at the class token position. It replaces the original class tokens from each sequence and introduces information from the opposite modality to the other modality-specific encoder. During the attention computation of each layer, the class token contains the same information for both encoders but influences the attention scores calculation separately. Therefore, different features from both modalities can be selected in each encoder, influenced by the information the class token contributes from the other modality. The fusion step for the class token is repeated with each layer of the Transformer Encoders. Finally, in the last layer, the fused class token is passed to a classifier head to compute the class prediction scores. An overview of the entire architecture is visualised in Fig. 5.6.

Due to the fact that trained linear layers perform the fusion step, their weights can adapt to the type of features present at a specific encoder layer. Additionally, a learned fusion mapping should encourage the selection of features with higher importance for the classification task.

The class token only has the same dimensionality as all other feature tokens. Therefore, the model is encouraged to extract information from both modalities

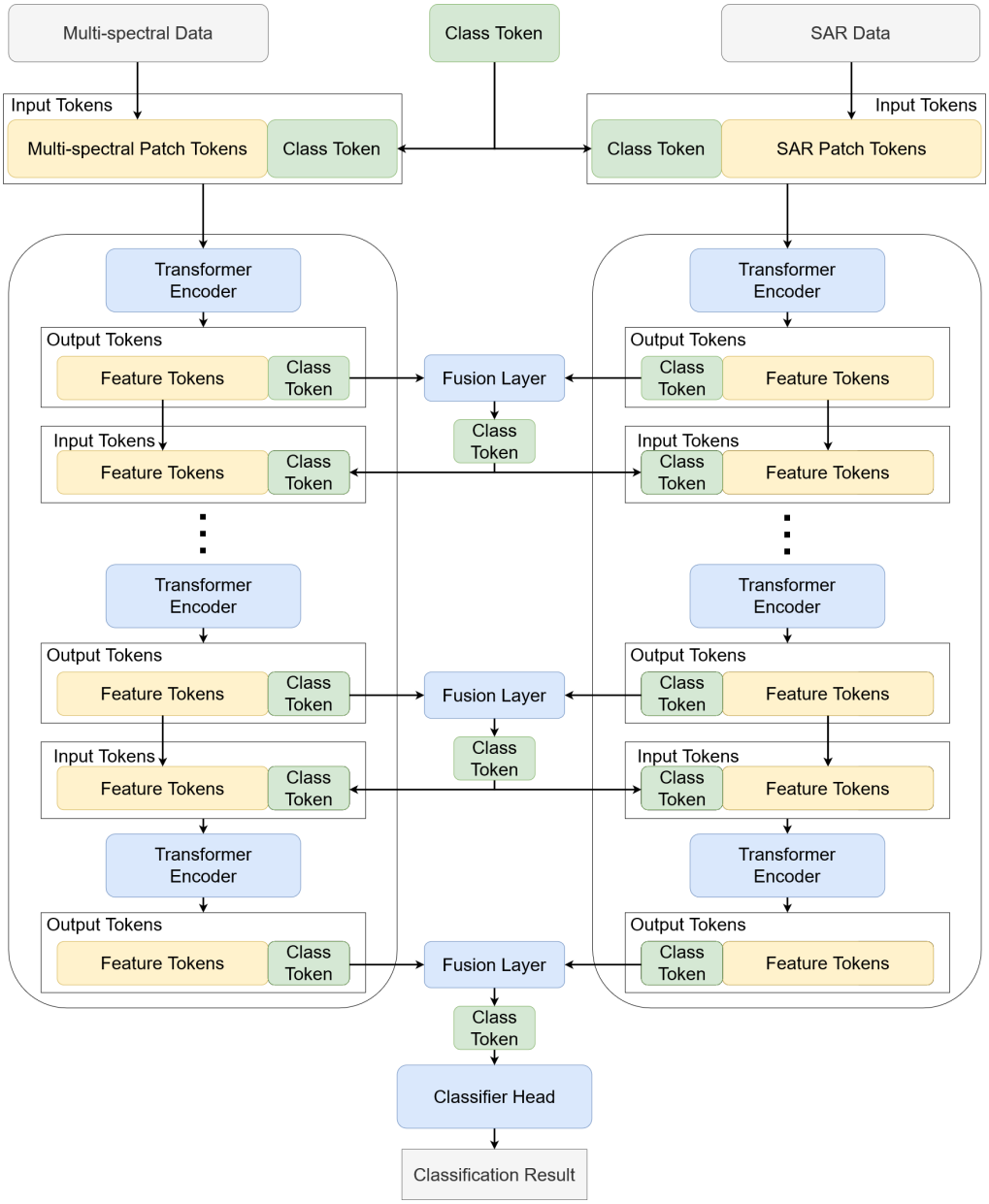


Figure 5.6: Visualisation of the full architecture design for the Synchronised Class Token Fusion model. The class token is consecutively extracted and fused by dedicated fusion layers. The fused class token is then passed back to each encoder and used in the next layer to compute attention scores. Finally, the last class tokens are extracted, fused and employed for classification.

more efficiently because both encoders effectively share the same feature space to store information. Additionally, this strongly limits the inter-modal information exchange to the most important features, which should have a further condensing effect on the feature selection. In that sense, the class token functions similarly to the bottleneck tokens proposed by Nagrani et al. [48]. However, in contrast to these bottleneck tokens, a more efficient fusion of the respective information is expected by SCT Fusion due to using the same token for the fusion and the classification task. Additionally, the utilisation of fully separate encoders should create a stronger requirement for inter-modal information flow. Similarly, Chen et al. [60] also discussed an information exchange through the class token between two modalities. However, their concept only combines the class tokens in a simplistic addition operation, which does not allow extraction and filtering of the most relevant features from both class tokens.

In conclusion, the repeated fusion of class tokens in SCT Fusion is intended to condense the most important information from the two modalities for later classification. Therefore, SCT Fusion might have an advantage for the classification task of interest in this thesis over the aforementioned fusion methods. Whether such an advantage exists is analysed in detail in the comparison between the investigated fusion methods conducted in the result's analysis.

5.7 Summary

The chapter introduced six different Transformer-based multi-modal fusion methods for the fusion of multi-spectral and SAR images which are classified into two principal categories. The three models in the first category are Early Fusion, Modality Token Fusion and Channel Token Fusion which have the advantage of only requiring minor modifications to the underlying Transformer Encoder. Only the embedding generation procedure is modified. Early Fusion, as the most simplistic approach, has the lowest demand for processing resources. Conversely, Modality Token Fusion and Channel Token Fusion have a significantly higher demand for computational resources due to the longer feature token sequences derived from the input modalities. Especially Channel Token Fusion generates feature token sequences of such length that training a model with it can become infeasible. However, it boasts the advantage of separating the fusion of channel information up to the attention layers in the Transformer Encoder, which could positively impact the feature extraction from these channels.

The second category of approaches includes Middle Fusion, Cross-Attention Fusion and SCT Fusion. These models rely on modality-specific encoders to extract abstract features from each modality separately. The added encoder approximately doubles the number of learnable parameters of these models, with the exact increase dependent on various hyper-parameter settings. Therefore, the processing requirements double because the whole computation is performed

once for each modality encoder. Consequently, the processing requirements for all three approaches with modality-specific encoders are lower than for Modality Token Fusion and Channel Token Fusion, as the impact of an enlarged feature token sequence is more significant than the added encoder.

Middle Fusion derives abstract features from each modality with distinct Transformer Encoders and employs an additional encoder to combine these features. It works on the idea that abstract feature representations from each modality provide more relevant information for the fusion process than low-level features.

Cross-Attention Fusion interconnects the two encoders with a Cross-Attention computation for the feature sequences derived from each modality. Because Cross-Attention directly correlates the token sequences from two modalities, it is expected to consider both modalities equally during the selection of features most relevant for the classification task.

SCT-Fusion extends the functionality of the class token in the architecture by introducing a repeated synchronisation step between the layers of each modality-specific encoder. This modifies the class token to perform a limited information exchange between both modality-specific encoders. Additionally, the repeated exchange steps should allow an SCT Fusion model to effectively condense the most relevant information from both modalities for the later classification task.

Overall, all introduced fusion methods modify a specific principle of the information processing of the two modalities with a Transformer-based model. Therefore, a comparison between all methods should reveal valuable insights on the validity of Transformer-based multi-modal fusion for multi-spectral and SAR images.

6 Dataset and Experimental Setup

This chapter first provides a detailed description of the dataset utilised for training the multi-modal fusion methods. Afterwards, the experimental setup for training the models is described by listing the utilised hardware and software components as well as the settings for important training options and hyper-parameters. To conclude, the relevant metrics used for the later evaluation and data augmentation strategies are introduced in detail.

6.1 BigEarthNet-MM Dataset

The BigEarthNet-MM dataset, released by Sümbül et al. [14], is a remote sensing dataset with one of the largest collections of labelled satellite images available. It was built with the intent to be used for training deep neural network architectures and evaluating their performance in the field of remote sensing. The dataset consists of a collection of labelled ground patches sampled from different regions within Europe. For each patch two types of sensor modalities captured by different satellite missions are available. In total, it is made up of 590 326 patches with each patch corresponding to a geographic region of about 1200m $\times$ 1200m respectively. The two modalities provided for each patch divide the dataset into two separate subsets. The first subset, BigEarthNet-S2, solely contains multi-spectral satellite images while the second subset, BigEarthNet-S1, consists of SAR images. The following sections give a detailed description of the data provided by the two subsets as well as a description of the classes assigned to each patch.

BigEarthNet-S2

The BigEarthNet-S2 is a subset of the full BigEarthNet-MM dataset but was released before the full multi-modal version as a standalone dataset by Sümbül et al. [61]. The individual patches were compiled from 125 Sentinel-2 tiles which were captured within a time frame from June 2017 to August 2018. The Sentinel-2 mission consists of a constellation of two satellites in a sun-synchronous polar orbit capturing images from the vast majority of geographical regions on the planet [15, p. 10]. The mission satellites were launched with the aim to provide high-resolution multi-spectral imagery at a relatively high revisiting time for research and civil applications.

The satellites capture 13 spectral bands from the electromagnetic spectrum of which 12 are utilised to form the BigEarthNet-S2 dataset. The band B10 has been

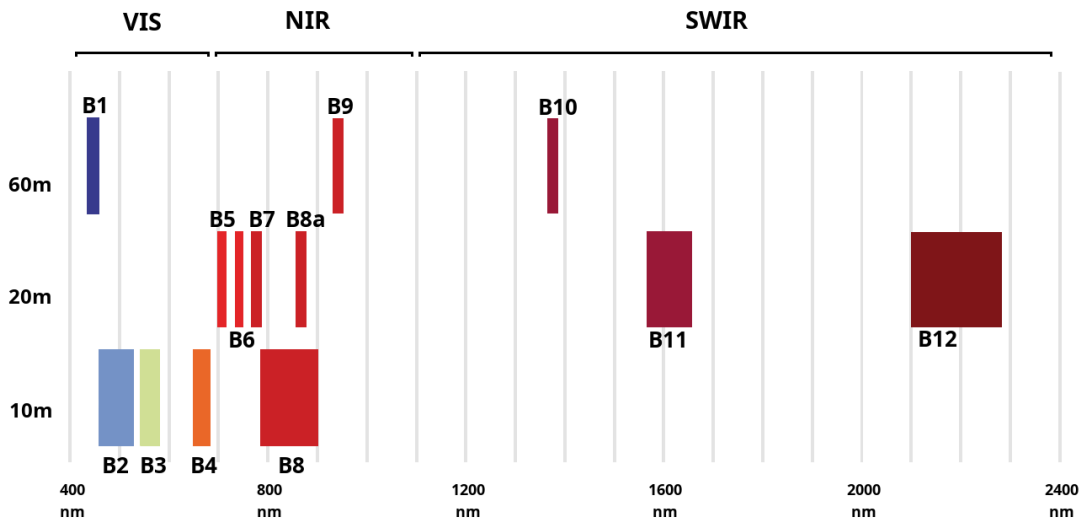


Figure 6.1: All spectral bands captured by the Sentinel-2 mission and their position and extent on the electromagnetic spectrum [15, p. 13].

discarded as it only provides atmospheric measurements which have no relevance for land cover classification tasks. An overview of the spectra captured by each sensor can be seen in Fig. 6.1.

The bands differ in their spatial resolution with the ground level resolution of a single pixel being either 10m, 20m or 60m respectively. The 10m bands consist of the bands B2, B3, B4 and B8. The bands B2, B3 and B4 capture at spectral ranges corresponding to the classical RGB colours while band B8 is sensitive to a relatively broad range of NIR wavelengths. The 20m bands consist of the bands B5, B6, B7, B8a, B11 and B12 which are all sensitive to various spectra in the NIR and SWIR. They offer a better spectral resolution for the NIR spectrum than band B8 which can be beneficial to differentiate vegetation types [62].

The 60m bands B1 and B9 are not considered for the training of models in this thesis, because their spatial resolution is often too low to discern many spatial features on the surface.

To serve as input to deep learning models the 10m and 20m bands must be transformed to have the same spatial dimensions. This necessitates the up-scaling of the 20m bands and the corresponding interpolation of their pixel values. Up-scaling could theoretically impact the capabilities of a model to extract information from the 20m bands. However, it is expected that this impact would be marginal as effectively no significant information loss should occur. An alternative option to align the spatial dimensions of all spectral bands would be to down-sample the 10m bands to the same resolution as the 20m bands. However, such an operation would incur a significant loss of information in the 10m bands and is therefore

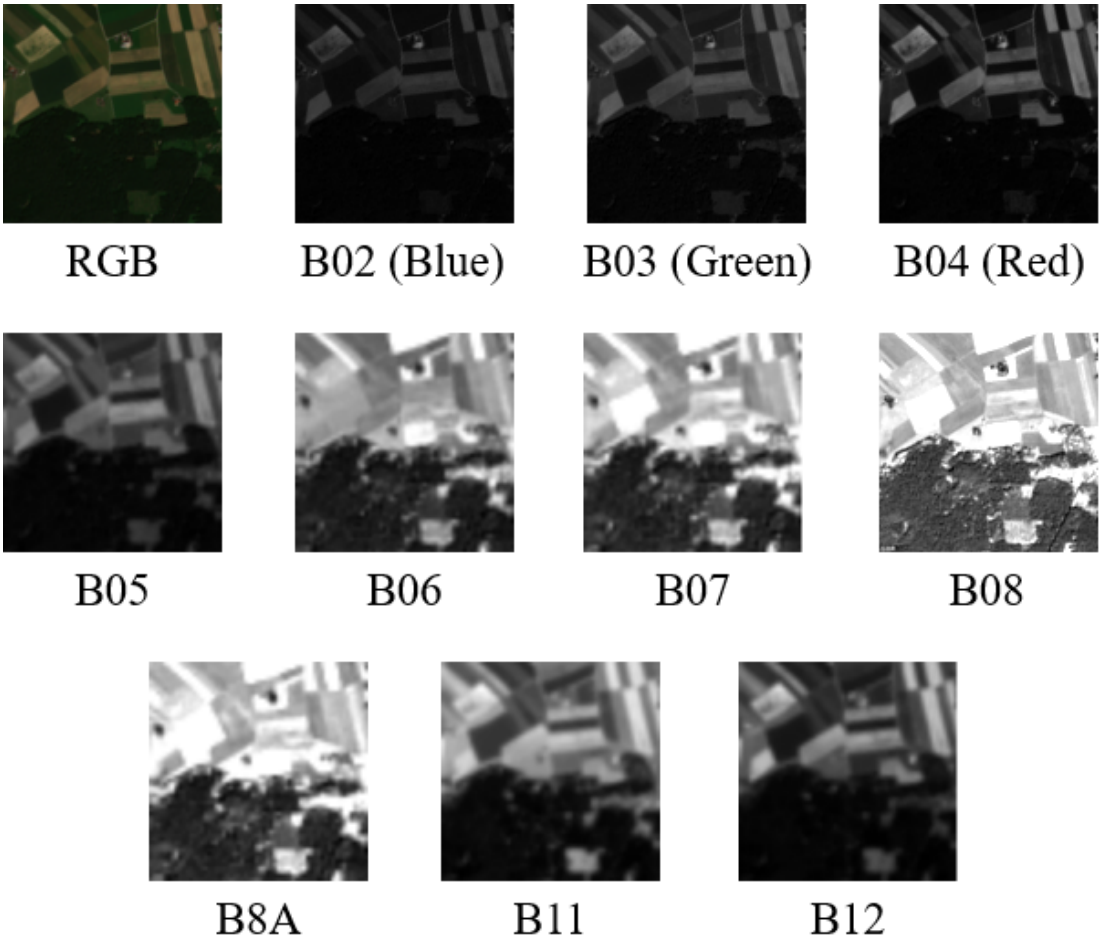


Figure 6.2: Visualisation of the 10m and 20m bands for an example S2 patch. Note the varying contrast of some visible features between the bands relating to different spectral reflectance behaviour of materials.

undesirable. Fig. 6.2 shows all 10 bands which are utilised in this thesis for an example patch.

Due to the time span over which the patches were collected, they can contain significant variations in their depicted surface conditions depending on the acquisition time. Many patches in the BigEarthNet-S2 dataset depict a significant amount of snow cover which strongly influences the spectral response compared to other variations in surface conditions. Additionally, despite having filtered all patches to have a small cloud cover percentage, some patches still contain clouds and cloud shadows which can also strongly influence spectral signatures. In total the two aforementioned conditions affect 70 987 patches in the whole dataset. Due to significantly altering the spectral response of such patches, Sümbül et al. [61] recommend removing them when training deep neural network architectures.

Overall, the S2 patches in the BigEarthNet-S2 provide relevant information on spectral surface properties to differentiate between various land cover types, making it an ideal candidate for deep learning research in the RS domain.

BigEarthNet-S1

The other subset of the BigEarthNet-MM dataset is the BigEarthNet-S1, which is composed of patches captured by the Sentinel-1 mission satellites. The Sentinel-1 mission is an imaging radar mission with the aim to provide high-resolution SAR data over a vast geographic area covering almost all major landmasses [63, pp. 9-10]. As raw SAR data is usually not directly usable, the mission's data is available at multiple processing levels. Each processing level pertains to different amounts and types of processing steps performed on the raw SAR measurements. The BigEarthNet-S1 dataset was entirely constructed from patches processed to the Ground Range Detected (GRD) product level. For all GRD products, multi-looking was performed and all signals were mapped to ground range by utilising an ellipsoid approximation model of the Earth's surface [63, p. 78]. It is noteworthy that in the GRD format all phase information is lost entirely and only the amplitude information is available for each pixel.

All data was originally collected using the Interferometric Wide Swath mode by the satellites. This mode scans the ground in a series of strips by repeatedly shifting the beam direction in both azimuth and range directions. It generates three strips which together cover the entire swath. The Interferometric Wide Swath mode presents the main acquisition mode of the Sentinel-1 mission satellites above land.

The Sentinel-1 satellites can send and capture signals in two polarisation channels. These are either polarised horizontally or vertically and are denoted as H and V respectively. The images used for the BigEarthNet-S1 were all captured with dual polarisation. They are referred to as the VV and VH channels correspondingly. The V at the beginning indicates that for both channels the radiation was sent with vertical polarisation. The V and H at the end denote that the response was separately received at vertical and horizontal polarisation.

The archive was constructed from 321 full GRD patches which overlapped with the acquisition areas of all patches in BigEarthNet-S2. These were then separated into smaller images with spatial dimensions of 120 by 120 pixels with a spatial resolution of 10m per pixel. Each of these patches precisely matches the geographic location of a corresponding patch in BigEarthNet-S2. Overall, patches from both Sentinel-1 satellites were utilised. All patches in the dataset were acquired between June 2017 and May 2018 which largely overlaps with the acquisition time frame of the corresponding Sentinel-2 patches. However, it should be noted, that the Sentinel-2 and Sentinel-1 patches showing the same geographic location in the BigEarthNet-MM dataset, were not acquired on the exact same data.

All patches were additionally processed by Sümbül et al. [14]. These processing procedures included the application of orbit files and geometric correction steps to

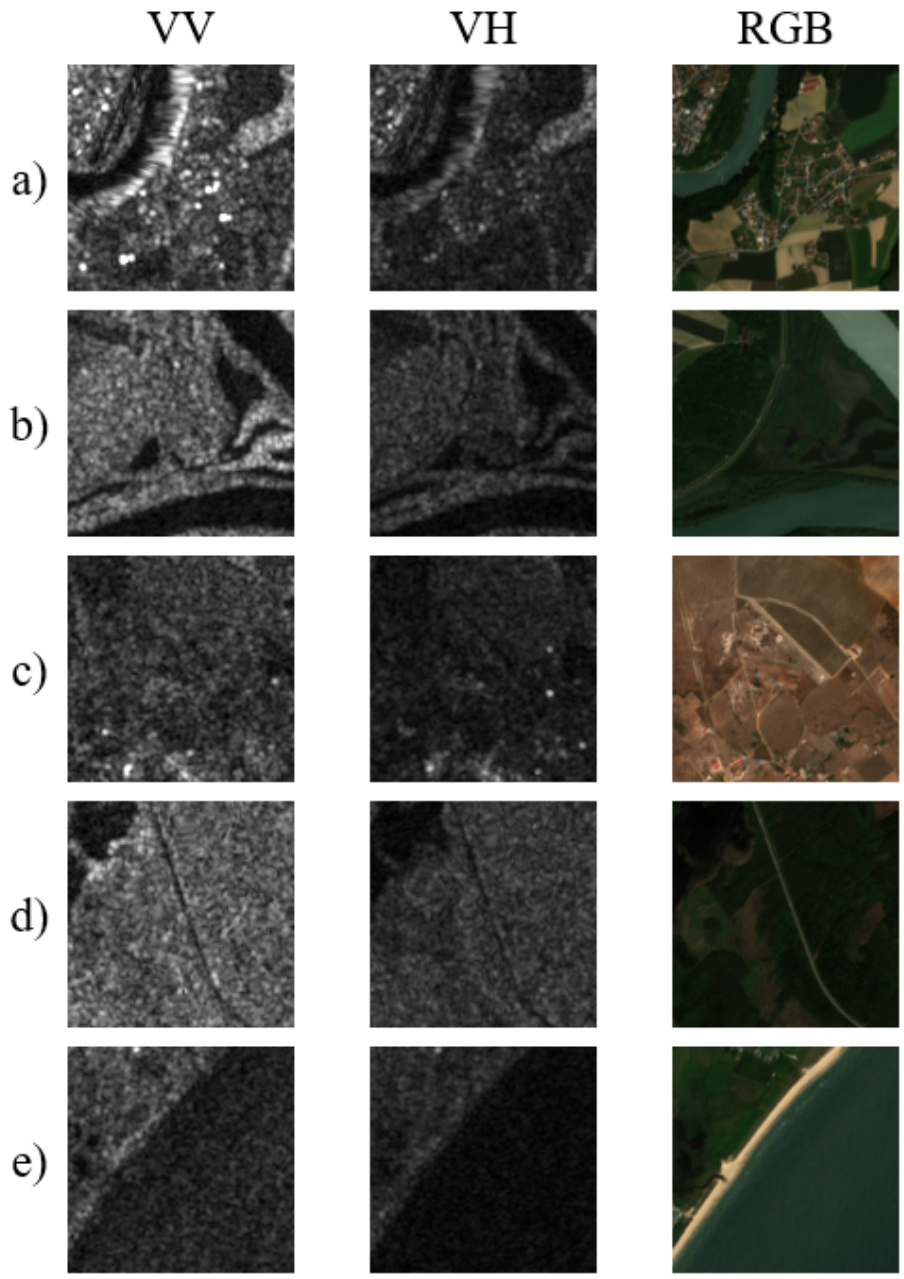


Figure 6.3: Visualisation of the VV and VH channels for a selection of S1 patches from the BigEarthNet-S1 and the corresponding RGB images from the BigEarthNet-S2. Note how some features are visible in the RGB images but do not appear in the SAR images.

map from range images to ground ranges by utilising a digital elevation model. Addi-

tionally, the images were radio-metrically calibrated and border and thermal noise were removed. The final values were converted from linear backscatter responses to decibel values. A speckle filter was not applied to the patches composing the dataset as the choice of a specific speckle filter is application dependent. In Fig. 6.3 a few SAR patches from the dataset can be seen together with the corresponding RGB images from the BigEarthNet-S2 dataset.

In conclusion, the S1 images present a valuable addition to the BigEarthNet dataset, as they substantially extend the information provided by the S2 patches. The resulting BigEarthNet-MM dataset is, therefore, well suited for research on multi-modal fusion of RS image data.

Classes and Class Distribution in the BigEarthNet dataset

In the BigEarthNet-MM dataset, every patch is assigned to a set of classes depending on the type of land cover present at that respective geographical location. Two types of labelling nomenclatures are provided for the dataset but in this work, only the newer and recommended 19 class nomenclature is used. The original label information was derived from the CORINE Land Cover (CLC) database from the label collection for the year 2018 to be aligned with the acquisition period for the original Sentinel-2 patches in the BigEarthNEt-S2 dataset. The original CLC labelling is available at three levels of detail with the third level of labels serving as the basis for the first set of labels proposed for the BigEarthNet dataset. In a later revision with the proposal of the BigEarthNet-MM[14] the more broadly defined 19 class labelling nomenclature was then created by collapsing many classes with similar spectral signatures from the previous CLC-3 classes into one class. Additionally, some classes were also removed entirely. The number of classes was therefore reduced from 43 original classes to the 19 classes used in this thesis. Table A.1 shows all 19 classes and the number of images belonging to each class.

Most classes correspond to a relatively dominant type of land cover present at a geographic location. For example, all types of human buildings and settlements belong to the “Urban fabric” and “Industrial or commercial units” classes. Additionally, only three types of forest classes are differentiated depending on their respective leaf type, regardless of the type of tree species present at a location. Additionally, there are multiple classes for different types of agricultural uses and different types of water bodies and wetlands.

As can be seen from Table A.1 the whole dataset exhibits a strong class imbalance which necessitates the use of adequate evaluation metrics to measure the performance of models. It can also be seen that some classes should have a relatively distinguishable spectral signature. Such a property might incline a model to rely more on Sentinel-2 data for classification in multi-modal scenarios. However, the addition of the Sentinel-1 data could possibly contribute to better identifying classes such as the various types of forests or the different artificial building types. These classes should generate relatively distinguishable backscatter signatures

corresponding to structural differences present on the surface.

Note that the classes are not evenly distributed over all geographic regions present in the dataset. Patches from the class “Agro-forestry areas” for example only occur in Portugal while patches belonging to the class “Urban fabric” are relatively evenly distributed over the entire dataset but centred around urban centres [64, pp. 35-36].

To summarise, the BigEarthNet-MM dataset provides a vast collection of labelled multi-spectral and SAR satellite images at a high resolution. Therefore, it is highly suitable for training and evaluating the multi-modal fusion methods investigated in this thesis.

6.2 Experimental Setup

The implemented multi-modal fusion methods as well as the classical ViT model utilised in this thesis are all implemented in PyTorch [65]. PyTorch is an advanced tensor processing library specifically designed for implementing and training neural networks. More specifically, all methods investigated in this thesis are derived from an implementation of the ViT architecture in the PyTorch Image Models library [66]. This library provides a multitude of efficient and tested implementations for various state-of-the-art deep learning architectures for CV tasks. The implementation of the ViT model provided by the library served as the basis for the implementation of the encoders utilised in the advanced multi-modal fusion methods.

All training runs for all models were performed on a Nvidia® V100 Graphics Processing Unit (GPU) with 32 Gigabytes of Video Random Access Memory (VRAM). The V100 GPU can perform a high number of floating point tensor operations and is optimised for deep learning applications.

Due to training models for multi-label predictions, the last layer activation function is a logistic sigmoid function. The binary cross entropy function is used as the loss function for training all models. The multi-label classification problem is thereby interpreted as a problem with multiple binary classifiers for each label in the dataset.

For training the models the well established Adam optimiser [67] is utilised with the default parameters of $\beta_1 = 0.9$ and $\beta_2 = 0.999$. All runs were initialised with a learning rate of 0.001 and a cosine annealing learning rate scheduler [68] is employed to optimise the learning rate during the training process. The number of epochs for all training runs is set to 60 epochs as a significant degree of convergence to an optimum could normally be registered at that point. The batch size is set to 1024 input images per batch if possible but could be set to lower values for models with an extensive demand for GPU VRAM. The input image dimension of 120 is fixed for all runs with the 20m bands of the BigEarthNet-S2 being up-scaled to fit this size.

The original train, validation and test split proposed for the BigEarthNet-MM dataset [14] are used throughout all training runs. This ensures that the results are directly comparable to previous research conducted on the classification of BigEarthNet patches.

The following sections introduce the metrics utilised to evaluate all models in this thesis. Additionally, a set of data augmentations for both single-modality and multi-modal training scenarios is defined.

Utilised Classification Metrics

To accurately evaluate the performance of the multi-modal fusion methods for the classification task a set of established classification metrics is selected to serve as the standard comparison method for further results analysis.

The fact that the BigEarthNet-MM dataset is a multi-label dataset introduces numerous challenges for the selection of evaluation metrics. In multi-label classification, a given prediction cannot be easily determined to be accurate. Some classes might be identified correctly while others are either missed or wrongly assigned to an image. Multiple strategies exist to handle multi-label predictions simultaneously. Some metrics are inherently capable to evaluate the quality of one-hot-encoded vectors representing multiple class predictions. An example of such a metric would be the Hamming Loss which is also employed in this thesis. Other metrics measure the performance of a model by evaluating the ratio between correct and wrong predictions for different classes.

Additionally, due to the class imbalance present in the BigEarthNet-MM dataset, the performance of a model must be evaluated while taking the performance on the individual classes into account. This is required, as a model might for example perform exceptionally well on the larger classes but more poorly on smaller ones, which is usually not desirable.

Therefore, different averaging procedures are employed for some metrics to evaluate the impact of class imbalance on the performance of the multi-modal fusion methods. While multiple averaging strategies exist, for evaluations in this thesis, only micro and macro averaging are employed.

When micro averaging a metric, the amounts of correct and wrong predictions are computed for all class labels for all images in an evaluation set. All these results are then directly averaged to compute a metric over the whole evaluation set. Due to giving equal weight to all predictions for all images, a model performing well in a micro averaged metric generally performs better for a larger number of images from the whole dataset.

In contrast, when macro averaging, the amounts of correct and wrong predictions are computed for each class label individually and aggregated to compute a mean metric score for a specific class label. The resulting scores for each class label are then averaged while giving equal weight to each class label. Therefore, a high macro averaged metric indicates that the model performs better over all types

of classes, regardless of the number of images belonging to each class. Macro averaging can be quite relevant for evaluations on the BigEarthNet-MM as it gives more weight to the smaller classes which would otherwise only weakly impact the overall metric. Noteworthy, when no imbalance is present in a dataset both the micro and macro averaged scores for a metric would be equal.

Another problem in multi-label classification pertains to the confidence of a model in whether a specific class is present in an input. Models for multi-label classification output a vector where each vector position corresponds to a score denoting how strongly the classifier detected the presence of a specific class in the given image. The confidence score is the output of a Sigmoid function, which outputs values between $0 - 1$, but does not restrict the size of the individual scores relative to each other. This necessitates the selection of an arbitrary threshold to determine if a score corresponds to a detection of a class. All scores above the threshold are then interpreted as a positive class prediction. However, the threshold is not present during the training process since the model would not be differentiable anymore, which is a necessary condition to perform gradient descent. Therefore, the threshold has to be selected as an additional parameter for the inference step after training. For the metrics utilised in this thesis, which require definitive predictions, a threshold of 0.5 is used. While the threshold parameter could be selected in an additional optimisation step, it would introduce further complexity. Instead, the problem of optimal threshold selection will be addressed by the introduction of a specialised metric, which evaluates a model over multiple thresholds simultaneously.

To provide a generalised definition of each metric, the term *sample* is used throughout this chapter to denote an arbitrary individual input to a model. A sample can, therefore, either consist of the data from only one or both modalities, depending on the corresponding use case. Overall, the metrics utilised in this thesis are the F2 score, the Hamming Loss and the Average Precision. The following sections introduce each metric and provide details on its advantages and downsides.

Precision, Recall and F2 Score

Precision and recall are well-established standard metrics in the field of classification. Given a set of samples and corresponding class predictions the precision measures the fraction of correct predictions for a class out of all samples that actually belong to that class. This directly relates to how many samples belonging to a class the classifier missed in its predictions. Conversely, recall measures the fraction of correct predictions out of all predicted samples for a specific class. In other words, it corresponds to how many samples have wrongly been assigned a certain class label. The definitions for both metrics are given in Eq. 6.1 and Eq. 6.2 respectively.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{6.1}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{6.2}$$

Here TP or true positives refers to the number of all samples that belong to a class and were predicted as such by the classifier. FP or false positives denotes the number of samples that were assigned to a class by the classifier but are not actually labelled as such. Lastly, FN or false negatives is the number of samples belonging to a class but which were not classified as such by the classifier.

Individually, precision and recall measure important qualities of the prediction capabilities of a classifier. However, optimising for only one of these metrics would usually result in a model performing poorly in the other metric. Therefore, to measure the performance of a classifier in both metrics they are combined in the F_β score to balance their respective influences. The definition of the F_β score is given in Eq. 6.3

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{Precision} \cdot \text{Recall}}{(\beta^2 \cdot \text{Precision}) + \text{Recall}} \tag{6.3}$$

Here, β denotes a weight parameter which can be used to influence the impact of the recall on the output score. Higher values for β increase the impact of recall while lower values decrease it. Throughout the evaluation of all models in this thesis, a value of $\beta = 2$ will be utilised. The recall metric is considered slightly more important in evaluating the classification performance on the BigEarthNet-MM dataset due to the significant level of imbalance in the number of samples belonging to each class. Therefore, good precision scores for the large classes might overestimate the overall performance of a model, necessitating a higher emphasis on the recall in the computation of the F_β score.

In conclusion, due to combining the advantages of precision and recall effectively, the F_β score is employed as a standard metric to evaluate the multi-modal fusion methods explored in this thesis.

Hamming loss

The Hamming Loss (HL) is often utilised to evaluate multi-label classification tasks due to it inherently being able to evaluate the quality of one-hot encoded class predictions without requiring the computation of ratios.

It is derived from the Hamming distance which, given two input sequences of equal length, is defined as the number of non-matching positions in both sequences. To define it, let $X, Y \in \{0, 1\}^{n_{smp} \times n_{class}}$ be two matrices with n_{smp} denoting the amount of samples to evaluate and n_{class} denoting the number of classes in the dataset. The rows of these matrices then represent the individual samples and the columns represent the respective class assignments. Let X contain the predictions from a model for all samples while Y contains all true label assignments for the same set

of samples. The definition of the HL can then be given as shown in Eq. 6.4.

$$\text{Hamming Loss} = \frac{1}{n_{smp} \cdot n_{class}} \cdot \sum_{i=1}^{n_{smp}} \sum_{j=1}^{n_{class}} x_{i,j} \oplus y_{i,j} \tag{6.4}$$

Here $x_{i,j}$ and $y_{i,j}$ denote the individual entries of X and Y respectively while $\oplus$ is the standard XOR operation. The XOR operation can be applied directly as all $x_{i,j}$ and $y_{i,j}$ are either 0 or 1. It is important to note that the HL is a loss function and therefore lower scores denote a better classification performance, which is different to all other metrics employed in this thesis.

A significant limitation of the HL is that the metric relies on true or false predictions and is unable to determine the performance of the model solely based on logit outputs. It therefore also requires the selection of a threshold like the aforementioned metrics and is susceptible to evaluating a model's predictions at a sub-optimal threshold value. To remedy this effect the following section introduces a metric capable of evaluating the performance of a model over multiple threshold values.

Average Precision

The Average Precision (AP) differs from the aforementioned metrics significantly as it does not require definitive class predictions for each sample to be computed. As has been described previously, one big problem with precision and recall is that they can only be determined for a specific threshold.

AP instead is computed over a multitude of thresholds. Its value corresponds to the area under the precision-recall curve. The precision-recall curve is defined by a function $p(r)$ which assigns a precision value to each recall value computed at the same threshold. This function expresses the relationship between precision and recall values for multiple thresholds. However, because comparing plotted function curves directly introduces further complexity, the performance achieved over all thresholds is summarised by the AP in a single metric score. Eq. 6.5 gives the formal definition for the AP.

$$\text{AP} = \int_0^1 p(r)dr \tag{6.5}$$

In practice, the integral in Eq. 6.5 is only approximated by using a specific set of threshold values between 0 and 1. Therefore, let p_i and r_i be the precision and recall computed for the same fixed threshold i and let n_t be the number of overall threshold scores. Then the approximate AP can be defined as shown in Eq. 6.6.

$$\text{AP} = \sum_{i=1}^{n_t} (r_i - r_{i-1}) \cdot p_i \tag{6.6}$$

As can be seen, the computed precision score for a threshold is weighted by the achieved improvement in the recall score compared to the previous threshold. This connects both metrics so that an improvement in one of them can not come at the cost of the other without decreasing the resulting AP score.

The main advantage of the AP is of course the evaluation of a model over multiple thresholds. Considering multiple thresholds in the evaluation is advantageous because the optimal threshold for a specific model is not determined by the loss function and can therefore differ for different models. Additionally, the metric combines both precision and recall into one metric. This allows the resulting AP score to represent both metrics equally without a manually selected weighting as is defined for the F_2 score. However, AP still requires an averaging procedure, because its computation is dependent on precision and recall which are based on ratios between the number of true positives, false positives and false negatives. Therefore, throughout the later evaluations both micro and macro averaged AP scores are provided to identify performance differences among the different classes present in the BigEarthNet-MM dataset.

To conclude, due to the aforementioned properties, the AP is utilised as the primary metric for evaluating the classification performance of the various multi-modal fusion methods investigated in this thesis.

Data Augmentations

Data Augmentations are artificial augmentations applied to the input data of a deep learning model to create a more difficult learning scenario. Due to their success in improving the generalisation capabilities of many models and their effectiveness in introducing regularisation to a training process, data augmentations have become a standard method in many deep learning applications. Unlike other regularisation techniques, data augmentations do not directly impact the information flow internal to the model and only influence it indirectly through introducing more variation in the training inputs.

The data augmentations considered for model training in this thesis, are restricted to only modifying the arrangement of values in an image without modifying the actual values themselves except for eliminating them entirely. Many common data augmentations in other CV tasks modify the individual channel values. Such augmentations potentially encourage a model to mainly rely on the spatial information in an image to derive a prediction. For many CV tasks, such an augmentation is reasonable because many classes in standard CV datasets such as ImageNet [11] can exhibit large variability in their internal colour distribution and are mostly identifiable by the spatial features present in an image.

However, when classifying satellite images, most land cover classes of interest often exhibit clearly identifiable spectral signatures while spatial feature composition takes a secondary role. Therefore, not considering such augmentations is intended to preserve the spectral information contained in the multi-spectral in-

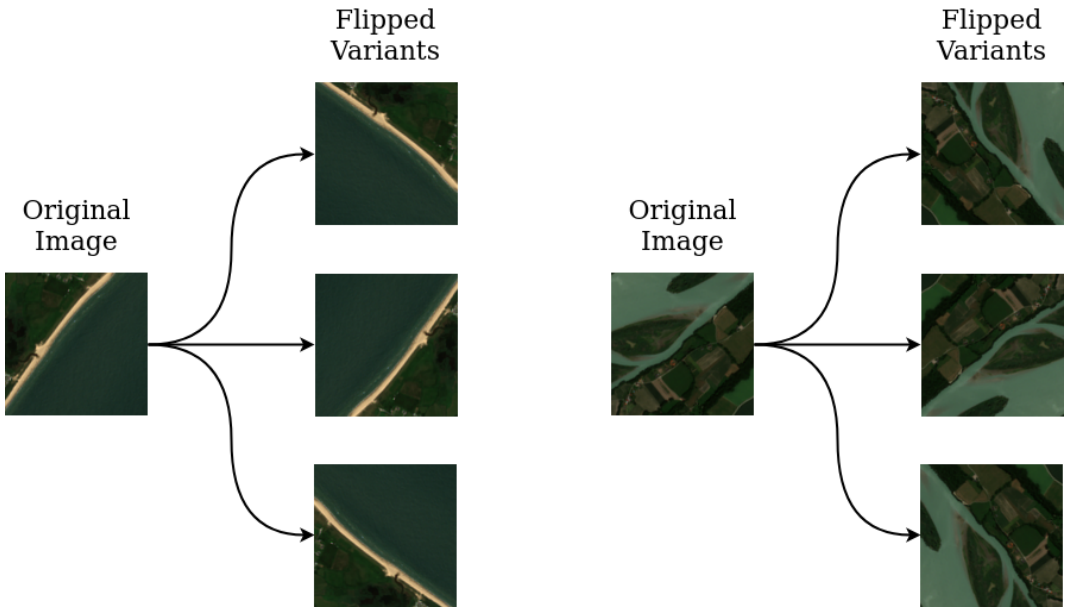


Figure 6.4: Example visualisations for the flipping data augmentation. Two patches from the BigEarthNet dataset are shown and their possible corresponding variants which could be produced by the flipping augmentation.

puts. The following sections introduce various types of data augmentations applied to the multi-spectral and SAR images for training the multi-modal fusion methods investigated in this thesis.

Random Flipping and Cropping

Flipping and cropping augmentations are two standard data augmentations utilised in this thesis. Both are randomly applied to the samples in all training scenarios throughout the conducted experiments.

The flipping augmentation flips the image and all its channel values on either the horizontal or vertical axis. Such an operation is equivalent to rotating the image by a random amount of 90° degree steps with an equal likelihood for all positions. The flipping augmentation is intended to encourage the model to learn features, which are independent of rotations of the input data. It can have a significant impact on the generalisation capabilities of a deep neural network by preventing it from overfitting on the spatial features present in an image. However, as satellite images are inherently not oriented, the effect is likely diminished compared to traditional CV applications. On the other hand, rotating satellite images does effectively increase the variability in images presented to a model, which should have a positive effect on the generalisation capability of a model. More complex

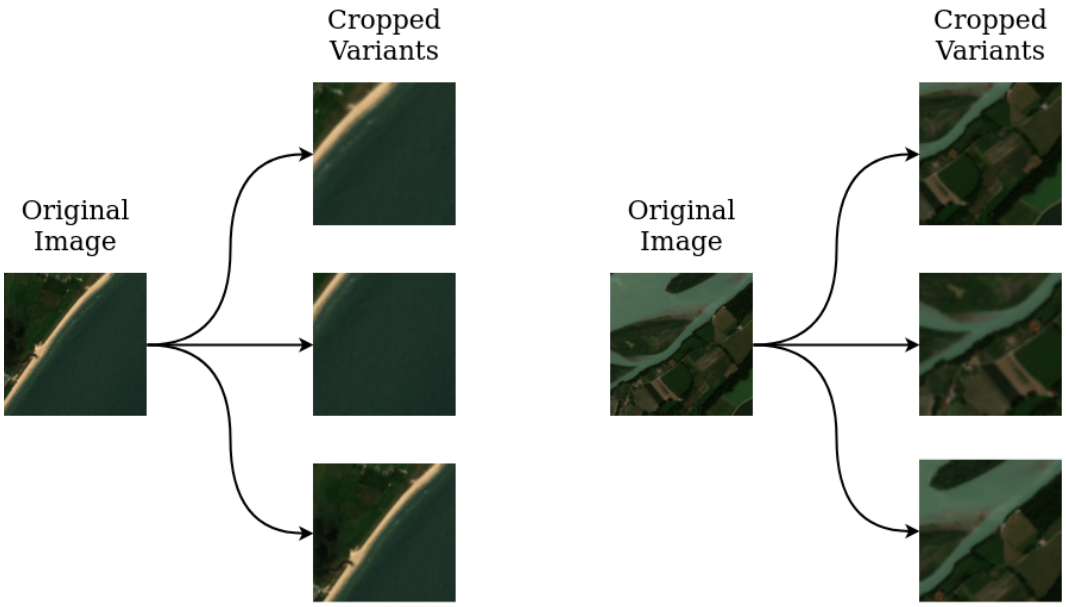


Figure 6.5: Example visualisations for the cropping data augmentation. Two example patches from the BigEarthNet dataset are shown and some possible resulting variants which could be produced by the cropping augmentation.

rotations with arbitrary degrees of rotation could also be considered but were not investigated further in this thesis as such rotations would require the interpolation of spectral values. Fig. 6.4 visualises the possible flipping results for some example patches.

Another standard augmentation applied in all training runs in this thesis is the cropping augmentation. Here a random sub-region within the image is selected and then resized to the spatial resolution of the original image. The size of the sub-region is determined by a scaling parameter, which is randomly selected from a preset range of possible values. The augmentation, therefore, scales the features present at a specific location in the image and cuts off all other features. Some regions containing features corresponding to a specific class present in the image might potentially be cut off during the operation. However, this should not occur so frequently to negatively impact the capabilities of a model as most land cover classes occupy a relatively large portion of an image. Occasionally dropping classes from images, might also have a regularising effect on the training as it prevents overfitting on specific class features. Overall, the augmentation is intended to encourage the model to learn features which are invariant to the scale of spatial features present in the input samples. It should also have a strong regularising effect as it significantly alters the composition of features in these inputs. Fig. 6.5 visualises some possible cropping results on satellite image data as an example.

For multi-modal training, both aforementioned augmentations can either be applied combined to both modalities or individually in a desynchronised manner to each modality. Desynchronising the augmentation could be advantageous as it decouples the spatial features in both modality inputs. Especially for the cropping augmentation, the desynchronisation should encourage a model to exchange information about missing regions in each modality.

Therefore, the desynchronised flipping and cropping of the multi-spectral and SAR images could have a positive effect on the performance of the multi-modal fusion methods investigated in this thesis. Additionally, another augmentation which affects the two modalities unequally is introduced in the next section to further strengthen the inter-modal information exchange within a model.

Random Modality Dropping

The Random Modality Drop augmentation forces a model to rely equally on both modalities during the training process. When training deep learning models on multi-modal data, it may occur that one modality dominates the training process and a model will learn to only rely on one modality to derive predictions. The issue can especially arise with datasets, where one modality provides significantly more features than the other modality, as is the case in the BigEarthNet-MM dataset. The Random Modality Drop prevents such an overreliance on one modality by setting all channels of a randomly selected modality to zero while leaving the other modality unchanged. It was originally proposed for multi-modal training on the BigEarthNet dataset by Wang et al. [42].

While the augmentation is intended to eliminate the reliance on one specific modality in single encoder models, it should have two additional effects on models with dedicated modality encoders. Firstly, dropping one modality forces one of the encoders to only rely on the information provided by the constraint information exchange, necessitating the adaptation of internal weights to facilitate such information flow adequately. Secondly, the final classification step cannot always rely on the output of one encoder for classification as it might have insufficient information to provide an accurate prediction. Therefore, the layers facilitating both of these steps are forced to learn weights that equally consider both modalities for predictions.

Overall, introducing augmentations which modify both modalities independently is expected to positively impact the performance of the multi-modal fusion methods. To ascertain such an effect, experiments with different augmentation settings are compared in Section 7.4.

Speckle Filtering

Synthetic Aperture Radar images are inherently noisy due to the mechanism by which SAR data is acquired. The beam bursts send out by a SAR imaging satellite

are initially in-phase but can become out-of-phase due to complex backscattering behaviour or time delays in the receiving of the response signal. As a result, at the time the radar sensor receives the backscattered signals, the out-of-phase radiation can constructively and destructively interfere with one another. This interference is noticeable in the generated SAR images as a salt-and-pepper noise, called *speckle*, uniformly distributed over the image. It affects all areas of the resulting image because the interference occurs over the entire acquisition area. To mitigate the influence of the aforementioned noise on the performance of classification models and other analysis tasks, various speckle filtering techniques have been proposed [69–71]. Most speckle filters function by identifying pixels with a significant difference from surrounding pixels and either removing them or smoothing their effects over a larger area.

To investigate the potential of speckle filtering for the multi-modal fusion of MSI and SAR data, a speckle filtering step is incorporated into the preprocessing pipeline of images from BigEarthNet-S1. Specifically, a median filter was selected due to its simplicity. A median filter functions by applying a median filter kernel to the entire image in which the centre pixel is replaced with the median of surrounding pixel values. This has the effect of removing outlier values which strongly differ from the surrounding region. However, legitimate spikes in backscattering strength and general sharp edges are also removed or blurred by the median kernel. The strength of the blurring can be controlled by setting the size of the kernel to relatively low values. Fig. 6.6 shows some example patches from BigEarthNet-S1 after processing with a median filter with varying kernel sizes.

Nevertheless, the blurring can incur an information loss on the SAR image which might eliminate features of relevance to distinguish between classes. Therefore, most applications which require speckle filtering of SAR data rely on more sophisticated filtering procedures such as the one proposed by Lee [69]. While other types of speckle filters might not suffer from such a strong blurring effect, all methods still inherently remove information from the SAR image. Additionally, more advanced filters are relatively complex in nature and a positive effect on the training of deep neural networks is not certain because such models could inherently learn better filters to adapt to the noise.

Therefore, a simplistic method was chosen with the median filter for initial tests to ascertain the general influence of speckle filtering on the overall performance. Experiments were conducted by training models on the S1 data with varying kernel sizes to determine the effect of speckle filtering on the performance of a model. The corresponding results are discussed in the following chapter in Section 7.4.

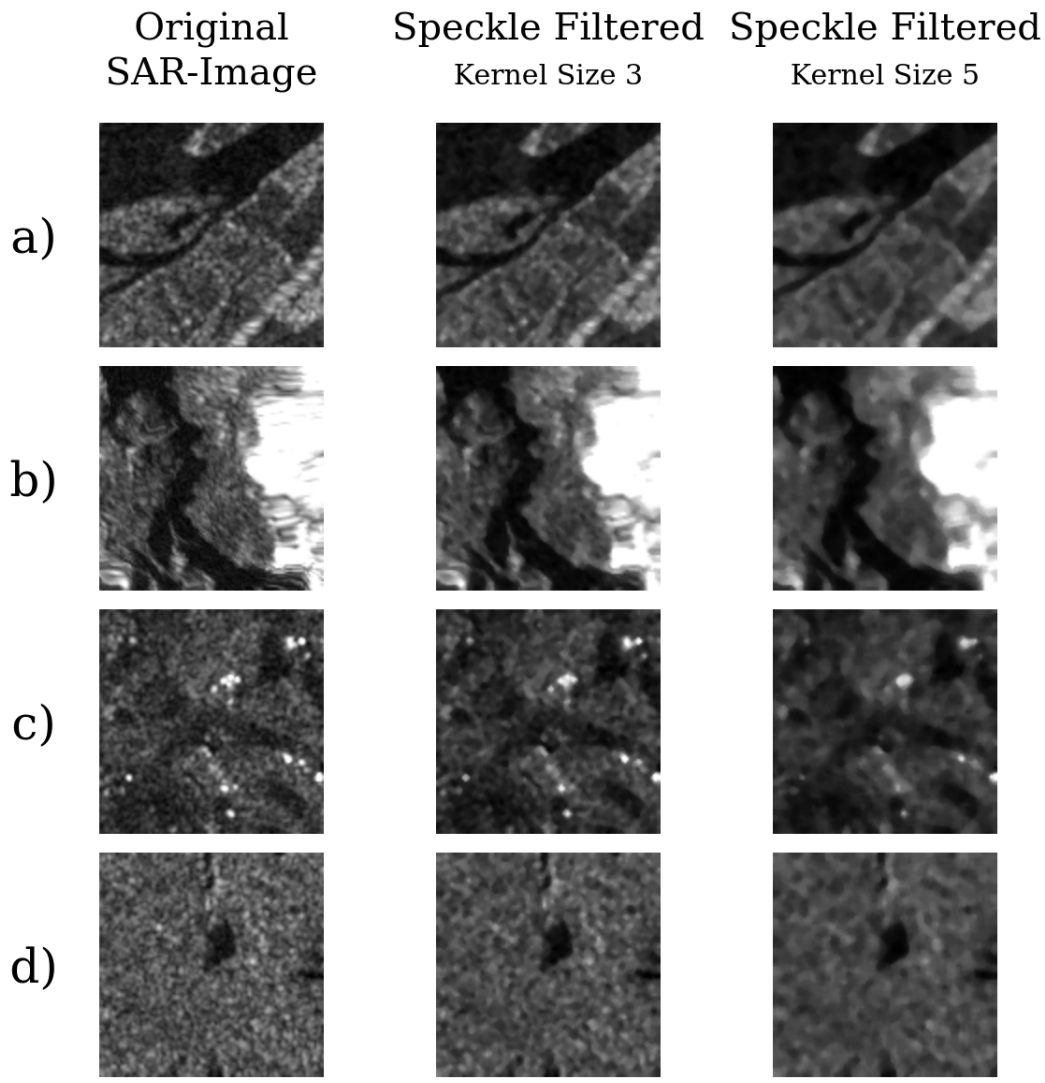


Figure 6.6: Visualisation of S1 patches and their corresponding filtered variants which were generated by a median filter with a kernel size of 3 and 5 respectively. Only images received at vertical polarisation are shown. Note the significant blurring and reduced intensity which can be observed in the images processed with a kernel size of 5.

7 Results

The chapter provides a detailed analysis of the multi-modal fusion methods described in Chapter 5. The internal layer composition and information flow within the ViT architecture and its derived multi-modal fusion models are governed by a set of hyper-parameters. These hyper-parameters can have a significant effect on the overall performance of a model. Therefore, the hyper-parameter settings used to train a specific model are provided throughout the whole chapter. One of the most important hyper-parameters is the *patch size* of the patches extracted from the input image tensor as has been defined in Section 3.2. Another one is the *depth* of the encoders utilised within a model, which defines the number of layers within a Transformer Encoder. Additionally, the *embedding dimension* and the *number of heads* used in the computation of MSA are also highly relevant to the performance of a model. Table 7.1 shows the standard settings of hyper-parameters as they are used for the following comparison experiments. The choice of these parameters was experimentally determined with the corresponding results provided in Section 7.4.

Another set of important parameters is the collection of different regularisation methods used during the training process. Two standard regularisation methods were utilised over all experiments. The first one is the *dropout rate* [72], and the second one is the *stochastic depth* [73]. However, following the results obtained by Steiner et al., [74] such regularisation methods were only used conservatively, and a greater emphasis was put on data augmentations. Therefore, by default, no dropout was employed, and *stochastic depth* was set to 0.25 for the training of all models. The positive influence of these regularisation methods and data augmentations on the multi-modal fusion performance is analysed in a dedicated section in the ablation studies.

Throughout the following evaluations, both the micro and macro averaged AP scores are provided. Therefore, it can be directly compared how well a model performs on either the different classes or over all samples in the dataset. At the same time, all provided F_2 score results are always micro averaged to reduce the visual overload.

First, the performance of the standard ViT architecture on singular modality inputs is analysed. Afterwards, a comprehensive analysis of all multi-modal fusion methods investigated in this thesis is conducted with different combinations of multi-modal input data. Lastly, the chapter concludes with detailed ablation studies to evaluate the impact of hyper-parameter settings as well as data augmentations and regularisation methods on the fusion performance of the investigated fusion methods.

Table 7.1: The hyper-parameter settings used throughout the majority of experiments.

Hyper-parameter	Value
Patch Size	20
Depth	8
Embedding Dimension	256
Number of Attention Heads	8

7.1 Analysis of Modality-Specific Performance

In the following, the performance of the standard ViT architecture on inputs from only one modality is analysed to establish a baseline performance to which to compare the fusion methods. The modalities analysed here naturally consist of the multi-spectral Sentinel-2 image data and the SAR Sentinel-1 data described in Chapter 6 but also includes the RGB subset of channels from the Sentinel-2 data. These images with reduced channel dimensions are included to evaluate the effect of the amount of information in the multitude of channels present in Sentinel-2 data on the classification performance. This is especially important for the later analysis of the performance of multi-modal fusion methods as the amount of information present in Sentinel-2 bands could already generate good results, seemingly diminishing the effect of the second modality on performance. For all modalities, a standard ViT model was employed. The architecture-dependent hyper-parameters were set to the standard configuration as introduced previously. Additionally, standard flipping and cropping augmentations were applied. Table 7.2 presents the overall results obtained on the three types of modalities analysed.

The results show that the best overall performance was achieved by a model relying on the full multi-spectral information provided by the S2 data. This could be explained by the breadth of information provided by the different spectral bands, which should allow for a fine distinguishment of the various classes in the BigEarthNet dataset. The effect is further emphasised when directly comparing the results obtained on the S2 and RGB bands. While the AP under micro averaging does not fall too significantly, a more severe drop can be observed in the AP score under macro averaging, indicating that the removal of specific bands disproportionately affects the capability of the model to identify specific classes. A similar effect can be observed when analysing the results obtained on the S1 modality. Here again, the micro averaged metrics show a less significant difference in performance than the macro averaged metrics. Here the drop in performance can likely be explained by the fact that some classes in the BigEarthNet dataset might exhibit similar features in the SAR data. As the underlying information SAR data represents strongly relates to the geometric shape of objects on the surface or the surface itself, classes with no significant surface features should generate similar backscatter responses. Such

Table 7.2: The performance achieved by a ViT model trained on either full Sentinel-2, Sentinel-1 or RGB channels data.

Modality	AP (micro)	AP (macro)	F_2	HL
S2	**0.8924**	**0.8164**	**0.7819**	**0.0602**
RGB	0.8710	0.7739	0.7519	0.0666
S1	0.8106	0.6754	0.6705	0.0822

an effect could, for example, occur if the general surface geometry is mainly flat, as would be the case for different types of fields.

The aforementioned results already indicate that separate modalities provide different features which can aid the distinguishment of specific classes. To further investigate these differences, Table 7.3 shows the results obtained by the previously introduced models on the individual classes present in the BigEarthNet dataset.

Here it can be seen that similar to the overall metrics, for nearly all classes, training on S2 data still outperforms training on RGB data while both outperform training on S1 data. However, the differences in class-wise performance are substantial.

When comparing the performance on S2 and RGB data, it becomes apparent that some of the classes most notably affected are the classes “Coastal wetlands”, “Inland wetlands”, “Moors, heathland and sclerophyllous vegetation” and “Permanent crops”. This could be explained by the fact that the missing bands constitute all of the NIR bands, which provide vital information for the differentiation of vegetation types in a specific location [62].

When comparing the performances to the model trained on S1 data, one can again identify significant differences among classes. Here, the strongest divergence occurs for the classes “Beaches, dunes, sands”, “Coastal wetlands”, “Inland wetlands” and “Moors, heathland and sclerophyllous vegetation”, “Natural grassland and sparsely vegetated areas” and “Permanent crops”. As observed, the same classes as previously discussed for the S2 and RGB data are again included. However, this time other classes are also affected. Most likely, all of these classes generate a relatively similar backscatter response to at least one other class present in the dataset. It would also seem that these classes all correspond to some type of flatland, but it could also be that the wetland classes again appear similar in the data due to the water content. Interestingly, it seems that the model trained on S1 data performs well on classes with distinct objects on the surface such as the ones corresponding to forests or human buildings.

Overall, most differences in performance can likely be attributed to the inherent properties of the respective modalities, and no unexpected variations could be identified. The performance on the individual modalities defines a lower bound, which has to be exceeded by the multi-modal fusion methods to fuse the discussed modalities successfully.

Table 7.3: The class-wise AP scores generated by a ViT model trained on either full Sentinel-2, Sentinel-1 or RGB channels data.

Class name	S2	RGB	S1
Agro-forestry areas	**0.8744**	0.8346	0.7248
Arable land	**0.9490**	0.9367	0.9005
Beaches, dunes, sands	**0.6617**	0.6331	0.4873
Broad-leaved forest	**0.8917**	0.8638	0.7703
Coastal wetlands	**0.7177**	0.5393	0.4112
Complex cultivation patterns	**0.8014**	0.7762	0.6931
Coniferous forest	**0.9546**	0.9423	0.8931
Industrial or commercial units	**0.5774**	0.5222	0.4850
Inland waters	**0.9287**	0.8964	0.9041
Inland wetlands	**0.7457**	0.6969	0.5337
Land principally occupied by agriculture, with significant areas of natural vegetation	**0.7647**	0.7492	0.6804
Marine waters	**0.9991**	0.9979	0.9930
Mixed forest	**0.9211**	0.9017	0.8361
Moors, heathland and sclerophyllous vegetation	**0.7778**	0.6753	0.4746
Natural grassland and sparsely vegetated areas	**0.6412**	0.5970	0.3633
Pastures	**0.8830**	0.8609	0.7817
Permanent crops	**0.7708**	0.7015	0.4537
Transitional woodland, shrub	**0.7817**	0.7532	0.6796
Urban fabric	**0.8697**	0.8256	0.7662

7.2 Analysis of Multi-Spectral and SAR Data Fusion

The performance of the investigated fusion methods is compared on the same scene classification task as the aforementioned single modality experiments. Again for the training of all methods, the standard hyper-parameter configuration and a *stochastic depth* of 0.25 was used. Unlike all other models, Early Fusion achieved the best performance with a *dropout* setting of 0.1 during training. The same effect could not be observed for the other methods, which therefore did not utilise any dropout.

The modality-specific data augmentations introduced in Section 6.2 were applied in all experiments, which includes the desynchronised flipping and cropping of the modality input patches as well as randomly dropping one of the modalities during training. The positive effect of these augmentations on the training performance was experimentally verified and is discussed further in Section 7.4. The overall performance achieved by each multi-modal fusion method is given in Table 7.4.

When comparing the performance of the various methods, it is immediately apparent that all methods achieve scores within close proximity to each other. Therefore, additional experiments were conducted to show that random fluctuations in performance between different experiments are not responsible for the observed differences. The corresponding results can be found in Table A.3 and Table A.4 and support the findings reported in Table 7.4.

Channel Token Fusion and SCT Fusion achieve the best classification performance for all metrics. The difference between all fusion methods is marginal when only

Table 7.4: The results achieved by each of the investigated multi-modal fusion methods on fusing MSI and SAR data to classify images in the BigEarthNet dataset.

Model name	AP (micro)	AP (macro)	F_2	HL
Early Fusion	0.8963	0.8253	0.7768	0.0591
Modality Token Fusion	0.8937	0.8202	0.7779	0.0598
Channel Token Fusion	**0.9004**	**0.8295**	**0.7870**	**0.0578**
Middle Fusion	0.8955	0.8244	0.7815	0.0595
Cross-Attention Fusion	0.8910	0.8193	0.7765	0.0606
SCT Fusion	0.8990	0.8292	0.7849	0.0582

considering the micro averaged AP or the F_2 score. However, when comparing the macro averaged AP score and the HL, Channel Token Fusion and SCT Fusion outperform the competition. Therefore, the performance of the fusion methods must differ considerably in their individual performance for each of the classes present in the dataset.

Overall, all investigated multi-modal fusion methods show an improvement over the single modality training cases discussed previously. However, the observed improvement is usually within a range of about one per cent, with the maximum improvement achieved sitting at 1.6% for the macro averaged AP. This can likely be attributed to the class imbalance present within the BigEarthNet dataset. Some classes in the dataset are assigned to a large number of patches. Therefore, if these classes are easily identifiable, it should increase the aggregated metrics to such a point where improvements for the smaller classes are only marginally recognisable.

Due to these shortcomings of the aggregated metrics and to better analyse the performance of the fusion methods on individual classes Table 7.5 presents the class-wise performance of all fusion methods for each of the classes in the BigEarthNet dataset.

In Table 7.5, it can be noticed that for most classes, Channel Token Fusion achieves the best scores. At the same time, SCT Fusion generates the best results for the second-highest number of classes, which aligns with the aggregated results presented previously. However, when analysing the absolute difference between all methods, the improvements among different methods are relatively small for most classes. The most significant absolute difference can be observed for the class “Coastal wetlands” where SCT Fusion generates the best score. However, the class “Coastal wetlands” contains only a very small amount of samples, as is shown in Table A.1. Therefore, random fluctuations might have a more noticeable influence on the class than would be the case for other classes. The same is true for the class “Beaches, dunes, sands” where, interestingly, Early Fusion, Modality Token Fusion and Cross-Attention Fusion outperform the other three methods.

Table 7.5: The class-wise performance of the different multi-modal fusion methods on fusing MSI and SAR images. All scores given are the AP.

Class name	Early Fusion	Modality Token Fusion	Channel Token Fusion	Middle Fusion	Cross-Attention Fusion	SCT Fusion
Agro-forestry areas	0.8781	0.8780	0.8798	0.8791	0.8744	**0.8888**
Arable land	0.9508	0.9497	**0.9540**	0.9497	0.9480	0.9522
Beaches, dunes, sands	0.7026	**0.7086**	0.6989	0.6997	0.7052	0.6892
Broad-leaved forest	0.8945	0.8915	**0.8989**	0.8936	0.8903	0.8972
Coastal wetlands	0.7378	0.7068	0.7340	0.7409	0.7481	**0.7565**
Complex cultivation patterns	0.8063	0.8042	**0.8114**	0.8055	0.8012	0.8089
Coniferous forest	0.9568	0.9546	**0.9575**	0.9560	0.9539	0.9571
Industrial or commercial units	0.5806	0.5738	**0.5959**	0.5883	0.5651	0.5915
Inland waters	0.9348	0.9327	**0.9428**	0.9383	0.9304	0.9384
Inland wetlands	0.7603	0.7594	**0.7719**	0.7605	0.7469	0.7698
Land principally occupied by agriculture, with significant areas of natural vegetation	0.7738	0.7676	**0.7771**	0.7694	0.7628	0.7750
Marine waters	0.9993	0.9993	0.9995	0.9993	0.9989	**0.9995**
Mixed forest	0.9221	0.9209	0.9251	0.9218	0.9197	**0.9258**
Moors, heathland and sclerophyllous vegetation	0.7850	0.7781	0.7895	0.7888	0.7768	**0.7962**
Natural grassland and sparsely vegetated areas	0.6659	0.6493	**0.6709**	0.6521	0.6482	0.6653
Pastures	0.8857	0.8831	**0.8886**	0.8846	0.8819	0.8864
Permanent crops	0.7867	0.7751	0.7859	0.7786	0.7728	**0.7879**
Transitional woodland, shrub	0.7869	0.7831	**0.7911**	0.7820	0.7790	0.7903
Urban fabric	0.8734	0.8685	**0.8871**	0.8746	0.8633	0.8794

Overall, the class-wise results achieved by almost all fusion methods exceed the performance of the ViT model trained only on S2 data. This affirms that the fusion process contributes positively to the capabilities of a model to discern most classes in the dataset while not diminishing it for others.

When examining the performance of the fusion methods, specifically on the classes that showed the most significant divergence in the single modal training case, some interesting things can be noticed. For these classes, adding the SAR modality likely provides the least amount of additional discriminative features. For three of these classes, SCT Fusion generated the best overall performance. These are namely the classes ”Coastal wetlands”, ”Moors, heathland and sclerophyllous vegetation” and ”Permanent crops”. Such results might hint at an advantage of SCT Fusion over the other methods in extracting discriminative features from both modalities.

While it seems that Channel Token Fusion has a slight edge on SCT Fusion, it should be noted that Channel Token Fusion requires considerably more computational resources than all other multi-modal fusion methods analysed in this thesis. The increased processing requirements are caused by the excessive amount of feature tokens generated by Channel Token Fusion, for which attention scores need to be computed. The marginal gains in classification performance do not justify such an increase in computational demand, significantly impeding the usefulness

of Channel Token Fusion.

Consequently, SCT Fusion is likely the best out of the investigated fusion methods suitable for the particular task of fusing MSI and SAR data. It combines a high performance on the BigEarthNet-MM dataset at a reasonable processing demand.

Early Fusion and Middle Fusion seem to perform similarly well, with both methods achieving good overall scores. Notably, for Middle Fusion, the depth of the modality encoders and the feature encoder can be scaled separately, resulting in an additional hyper-parameter which can significantly impact the overall performance. The impact of varying the depth settings for both encoder types is analysed in the ablation studies.

Notably, Cross-Attention Fusion and Modality Token Fusion perform considerably worse than the other fusion methods when considering the aggregated metrics. Analysing their class-wise performance reveals that this drop is related to drops in the performance on specific classes such as, for example, “Moors, heathland and sclerophyllous vegetation” or “Urban fabric”. These results indicate that the fusion of modality features does not work as effectively as possible in both of these fusion strategies.

As discussed previously, the multi-spectral images contain considerably more features than the SAR images. Therefore, to analyse the impact of this difference in features on the fusion performance, the following section conducts a comparison of all investigated fusion methods on fusing RGB and SAR images for the same classification task.

7.3 Analysis of RGB and SAR Data Fusion

The S1 and S2 modalities exhibit significant variation in the quality and quantity of information they can provide to differentiate the classes present in the BigEarthNet dataset. As shown previously, such a difference leads to considerable variability in the achieved performance when training models on a single modality.

As a result, the substantial contrast in the complexity of both modalities could lead to S2 data largely dominating the fusion process. Therefore, the fusion performance of the fusion methods was evaluated on an additional fusion task where the S2 modality is replaced with only the RGB channels extracted from the S2 data. An RGB image should provide much less information for the classification than a multi-spectral image, especially for specific classes as was analysed previously.

For these experiments, the same hyper-parameter settings used for S2 and S1 fusion were utilised to ensure the comparability of results. Table 7.6 shows the aggregated results obtained for the fusion of RGB and SAR remote sensing images.

Again it can be observed that compared to the training scenario employing only RGB data, all fusion methods improve the overall performance. Therefore, the methods seem to again successfully derive relevant features from both modalities.

Apparently, for the reduced fusion task, the Channel Token Fusion again dom-

Table 7.6: The results achieved by each of the investigated multi-modal fusion methods on fusing RGB and SAR data to classify images in the BigEarthNet dataset.

Model name	AP (micro)	AP (macro)	F_2	HL
Early Fusion	0.8792	0.7921	0.7570	0.0643
Modality Token Fusion	0.8818	0.7977	0.7607	0.0635
Channel Token Fusion	**0.8893**	**0.8119**	**0.7715**	**0.0614**
Middle Fusion	0.8810	0.7938	0.7601	0.0638
Cross-Attention Fusion	0.8806	0.7951	0.7611	0.0638
SCT Fusion	0.8804	0.7942	0.7572	0.0639

inates the overall performance, clearly outperforming all other methods. The difference in performance between Channel Token Fusion and the other methods is considerably larger than in the previously examined S1 and S2 fusion. Additionally, the other fusion methods show markedly different results compared to the previous fusion task. Both Early Fusion and SCT Fusion have lost their performance advantage and now generate results very similar to the remaining three methods. When not considering Channel Token Fusion, the variance in the metric scores between the other five methods is significantly lower than for the S1 and S2 fusion task. Especially the macro averaged AP scores exhibit a low variance which implies that only slight fluctuations in the class-wise performance occurred for these methods. As a result, the impact of random variations could be stronger, which would further support the observation that all methods except Channel Token Fusion perform similarly well. The class-wise results show a similar overall effect as the aggregated metrics. For most classes, Channel Token Fusion achieves the best results, while all other methods perform similarly well on most classes. Therefore, the class-wise performance is not discussed in detail as no significant further details of interest could be identified. For the sake of completeness the class-wise results are provided in Table A.2.

Overall, Channel Token Fusion seems to be more viable for such a reduced fusion task due to its outstanding performance. However, it still requires more computational resources than the other methods, even though the total number of channels and therefore feature tokens that need to be processed is considerably lower. For the other methods, the results show that Early Fusion and SCT Fusion are unable to reproduce their performance advantage, which they achieved on S2 and S1 fusion. Compared to the remaining fusion methods, their performance results are relatively similar over all metrics. However, the similarity in results might be explained by the fact that SCT Fusion and Early Fusion could be better suited to fusing modalities with a considerable difference in the number of features in each modality provides. Therefore, the results would indicate that Early Fusion and SCT Fusion are not better fusion methods in general but are instead specifically

well suited for the fusion of multi-spectral and SAR images.

The results for Modality Token Fusion and Cross-Attention Fusion show that both methodologies present viable fusion methods for general multi-modal fusion tasks. However, they seem to suffer from an inability to effectively fuse multi-modal data with a significant differential in the number of discriminative features provided by each modality which could explain their weak performance on fusing multi-spectral and SAR images.

To summarise, while Channel Token Fusion clearly outperforms the competition on the reduced fusion task no second-best method could be identified. Instead, all other five methods produce very similar results, which stands in contrast to the MSI and SAR fusion task analysed previously. It is hypothesised that the divergence is caused by the stark difference in the number of features multi-spectral and RGB images can provide but further research is required to ascertain the underlying causes.

7.4 Ablation Studies

The following ablation studies are conducted to determine the impact of hyper-parameter settings on the performance capabilities of the investigated multi-modal fusion methods. This includes models utilising the standard ViT architecture as an encoder as well as the more advanced fusion architectures, which significantly alter the information flow within their internal encoders. All tests were conducted on the full multi-modal BigEarthNet data set with all channels from the S2 and S1 data.

Additionally, these ablation studies enable the comparison of the different fusion methods under various hyper-parameter settings. Because these hyper-parameters directly govern the size of internal layers within the architecture and the amount of information processing performed by a model, varying them changes the overall computational requirements of a model. This might reveal specific advantages or disadvantages of the analysed fusion approaches for operations on systems with reduced or improved processing capabilities.

To reduce the amount of information discussed here, only the results for Early Fusion and SCT Fusion are presented for all hyper-parameters. These were selected according to their best overall performance on the full S1 and S2 fusion task when not considering Channel Token Fusion due to its excessive computation requirements. However, for specific hyper-parameters, the results for other fusion methods are presented as well, if they are deemed to provide valuable insights. In general, during the training of the models, the full range of data augmentations was utilised. The *stochastic depth* was set to 0.25 and no dropout was employed to ensure comparability to the results used for the previous comparisons. As a result, the performance of Early Fusion is slightly lower in most cases due to it profiting from using dropout during training.

Table 7.7: Classification performance of Early Fusion and SCT Fusion over varying *patch sizes*.

Patch size	Early Fusion		SCT Fusion	
	AP (micro)	AP (macro)	AP (micro)	AP (macro)
10	**0.8987**	**0.8269**	0.8995	0.8204
15	0.8963	0.8227	**0.9018**	**0.8345**
20	0.8949	0.8197	0.8990	0.8286

First, the influence of varying the patch size is analysed. The results obtained by training the two aforementioned models under different patch size configurations are presented in Table 7.7. It should be noted that a patch size of 10 presents the lowest bound at which the training of most of these architectures remains feasible with the available hardware. By lowering the patch size with the standard patch embedding procedure, the amount of generated feature tokens effectively quadruples, which increases the processing requirements. With other patch embedding methods, such as modality-specific patches or channel patches, the effect is even more drastic.

Apparently, lowering patch sizes directly corresponds to an improved classification performance, which aligns with initial expectations because the original ViT model trained on single modal inputs exhibits the same behaviour [10]. Interestingly, SCT Fusion did not generate the best results at a patch size of 10. However, by increasing the dropout rate for the S2 encoder to 0.1 for SCT Fusion at a patch size of 10 leads to an increase in performance. Similarly, increasing the dropout rate for Early Fusion also improves its performance, hinting at a general trend. The corresponding results can be found in Table A.5. However, the improvement for SCT Fusion with dropout still does not significantly improve upon the macro averaged AP score, indicating that a patch size of 15 might be an optimal setting for SCT Fusion on the MSI and SAR fusion task. Overall, the presented results show that the improved fusion performance of the respective models is mostly maintained over different patch sizes.

Next, the impact of changing the *depth* hyper-parameter, which controls the number of layers in the encoders, is investigated. Table 7.8 shows the results over a range of *depth* values from 2 to 12 for Early Fusion, Cross-Attention Fusion and SCT Fusion.

The results show that scaling the depth of the Transformer Encoder for each of the multi-modal fusion methods directly impacts the overall performance. For all shown models, higher depth values do not seem to induce a higher performance of the respective fusion method. Such behaviour was unexpected as higher *depth* values result in deeper encoders with more parameters which have usually been shown to achieve better performances on other classification tasks [10]. Potentially, deeper models might require more regularisation. However, further research is

Table 7.8: Classification performance of Early Fusion, Cross-Attention Fusion and SCT Fusion over varying *depth* settings.

Depth	Early Fusion		Cross-Attention Fusion		SCT Fusion	
	AP (micro)	AP (macro)	AP (micro)	AP (macro)	AP (micro)	AP (macro)
2	0.8871	0.8081	0.8892	0.8087	0.8887	0.8080
4	0.8929	0.8170	**0.8976**	0.8277	0.8947	0.8200
8	**0.8949**	0.8197	**0.8976**	**0.8281**	**0.8990**	**0.8286**
12	0.8935	**0.8229**	0.8974	0.8280	0.8985	**0.8286**

required to analyse the benefit of deeper models on the fusion performance.

Interestingly, when analysing the effect of lowering the *depth* value, it can be observed that the performance is only degrading slowly. Relatively shallow models with only 4 encoder layers are still able to produce strong results. A possible explanation could be the fact that some classes present in the BigEarthNet dataset are naturally identifiable by distinct spectral responses. Further support for this hypothesis is provided by the stronger degradation on the macro averaged than on the micro averaged AP scores. All presented results reinforce the selection of a *depth* value of 8 for the multi-modal fusion comparison.

The performance of Cross-Attention Fusion is shown to highlight an interesting observation. Cross-Attention Fusion seems to achieve a better performance than the other approaches at low depth values. Therefore, Cross-Attention might have an advantage in fusing cross-modal features at earlier layers than the other fusion approaches. Such a property might make the Cross-Attention Fusion method more applicable for an environment with low processing resources, but further research would be required to ascertain the effect.

Another series of experiments was conducted to analyse the impact of different *depth* settings for the two encoder types employed in the Middle Fusion architecture. The results of these experiments are provided in Table A.7. Based on these results, the *depth* of the modality encoders was set to 6 and the *depth* of the fusion encoder was set to 2 for all other experiments. Interestingly, the Middle Fusion method seems to produce better results with deeper modality encoders which could imply that multiple self-attention computations on modality-specific features before the fusion step is beneficial for multi-modal fusion on the BigEarthNet-MM dataset.

Additionally, the impact of varying the *embedding dimension* was also assessed. The corresponding results can be found in Table A.6. However, no major implications could be observed for the tested fusion methods. Notably, SCT Fusion's performance dropped significantly at higher *embedding dimension* values. Such a performance drop could be mitigated by imposing stronger regularisation during the training phase but was not investigated further.

Table 7.9: Experimental results with Early Fusion and SCT Fusion to select the optimal value for the stochastic depth regularisation method.

Stochastic depth	Early Fusion		SCT Fusion	
	AP (micro)	AP (macro)	AP (micro)	AP (macro)
0	0.8874	0.8084	0.8922	0.8169
0.25	0.8949	0.8197	**0.8990**	**0.8286**
0.5	**0.8971**	**0.8236**	0.8987	0.8262

Impact of Regularisation Methods and Data Augmentations

Two main groups of methods to introduce regularisation during training of the various models were utilised in this thesis. The first group consists of the data augmentation techniques introduced in Section 6.2. The second group constitutes the regularisation methods stochastic depth [73] and dropout [72].

Dropout refers to the standard dropout applied between layers in the model [72] as well as the attention dropout, which drops outputs from the attention computation randomly [75]. The second type of regularisation method is called *stochastic depth*, and it was proposed by Huang et al. [73]. The value for *stochastic depth* defines the probability of omitting entire layers during training and replacing them with the identity operation.

Dropout was only used conservatively during the training of most models because a generally positive impact on all fusion methods could not be identified. Especially when employing data augmentations, dropout seemed to inhibit the performance potential of the models. Such an effect could also be identified by Steiner et al. [74] who showed that data augmentations should be preferred when training ViT models. In accordance with these results, a larger emphasise was put on data augmentations, and other regularisation methods were only employed conservatively. The value of 0.25 used for *stochastic depth* throughout all experiments was selected due to results showing an overall performance improvement even under the utilisation of data augmentations. The corresponding results are provided in Table 7.9.

All multi-modal fusion methods compared previously were trained with the same data augmentation settings. These data augmentations are the desynchronised flipping and cropping augmentation as well as the random dropping of whole modality inputs. Experiments were conducted to evaluate the impact of these augmentations on the fusion performance with the corresponding results provided in Table 7.10. Overall, the results show a generally positive impact of employing the aforementioned data augmentations, validating their use during training the multi-modal fusion methods for the previous comparisons.

Table 7.10: The classification performance of Early Fusion and SCT Fusion under varying data augmentation settings. The standard augmentation type refers to using combined flipping and cropping augmentations.

Augmentation Type	Early Fusion		SCT Fusion	
	AP (micro)	AP (macro)	AP (micro)	AP (macro)
Standard	0.8916	0.8171	0.8980	0.8253
Desynchronised Flipping & Cropping	0.8919	0.8180	0.8980	0.8256
Modality Dropping	0.8938	**0.8205**	0.8988	**0.8294**
All Augmentations	**0.8949**	0.8197	**0.8990**	0.8286

Impact of speckle filtering

Speckle filtering of the S1 data was also investigated as a potential method to improve the fusion performance with SAR data. Speckle filtering represents a highly utilised tool to improve the interpretability of SAR data for various applications [70]. To this end, a median speckle filter was integrated into the preprocessing pipeline for S1 data as has been described in Section 6.2. Table 7.11 shows the results obtained by training a ViT model on S1 data without filtering and with a median speckle filter with a kernel size of 3 and 5 respectively.

Evidently, the classification performance of the trained models suffers in all experiments where the speckle filter was applied. Potentially, the models are able to learn an internal filter to differentiate between noise and relevant information that performs better than the median filter employed here. A possible explanation would be that a classical speckle filter inherently removes information present in input images by blurring or removing pixel values even if they are not caused by noise. Deep learning models, on the other hand, might be able to discern between valuable information and noise at such a level by learning filters that can dynamically extract relevant features and drop irrelevant ones.

Additionally, applying a speckle filter significantly increases the computational demand for preprocessing during training compared to the standard preprocessing pipeline. While the preprocessing could likely be optimised or even performed once for all samples before starting the training, it still introduces more complexity into the training procedure which is undesirable.

Potentially, a more sophisticated speckle filter such as a Lee filter [69] could result in a better performance. However, all speckle filters effectively reduce the available information from an input image. More sophisticated speckle filters require significantly more processing resources, further increasing the complexity of the training procedure. Therefore, a positive effect on the performance was not deemed likely, and this direction was not further pursued.

Table 7.11: Performance comparison of ViT models trained on S1 data with and without speckle filtering. The models all used the same standard hyper-parameter settings without any regularisation settings.

Speckle Filtering Setting	AP (micro)	AP (macro)	F_2	HL
Without Speckle Filtering	**0.8068**	**0.6676**	**0.6684**	**0.0833**
Speckle Filtering Kernel Size 3	0.8044	0.6605	0.6682	0.0839
Speckle Filtering Kernel Size 5	0.7962	0.6536	0.6576	0.0857

Impact of S1 data normalisation

SAR and optical image data inherently represent different types of information. As a result, the range of values present in both data types used in this thesis varies considerably. While the multi-spectral data consists of linearly scaling sensor readings, the SAR data is provided in decibels which are logarithmic measurements.

Therefore, utilising a different normalisation procedure might be beneficial for SAR data. When standardising an image, the pixel values are mapped to a range of values so that all pixels follow a standard normal distribution with a mean of zero and unit variance. These properties are beneficial for training neural networks because they result in better gradient computation during gradient descent.

However, other types of normalisation procedures are also utilised for various machine learning tasks depending on the type of data. One of the most prominent ones is the min-max-scaling operation, where all values are scaled to a range between $0 - 1$. This is achieved by computing the minimum and maximum values for all samples in the training set and setting these values as the minimum and maximum bounds for the mapping operation. All pixel values are then mapped to a range between $0 - 1$ with their ratio to the minimum and maximum value staying equal before and after the mapping operation.

To test which type of normalisation procedure is most beneficial for the BigEarthNet's SAR data, experiments were conducted to compare the classification performance under different normalisation schemes. Table 7.12 presents the results obtained from these experiments. The three configurations tested are no normalisation, a min-max scaling normalisation as well as a standardisation normalisation as would typically be performed in the CV domain.

Apparently, standardisation outperforms both other methods to such a degree that no further testing with multi-modal fusion methods is performed. As a result, all other experiments conducted for this thesis employed a standardisation normalisation. Other normalisation techniques would likely negatively impact the performance of models.

Table 7.12: Comparison of the results obtained by ViT models trained on S1 data under different normalisation strategies. The models used the standard hyper-parameter settings without any regularisation settings.

Normalisation Technique	AP (micro)	AP (macro)	F_2	HL
Standardisation	**0.8068**	**0.6670**	**0.6680**	**0.0833**
Min-Max Scaling	0.6861	0.4760	0.4710	0.1101
Without Normalisation	0.7916	0.6430	0.6440	0.0867

Summary

Overall, the ablation experiments validate the hyper-parameter selection used for the previously conducted comparison between the different fusion methods. While a potential for improvements in the overall performance could be identified by using a reduced *patch size*, such a modification would incur a significant increase in computational demand. Interestingly, reducing the *depth* parameter does not significantly reduce the performance of the tested fusion models until the number of layers is below a threshold of four.

The positive impact of data augmentations and stochastic depth could also be ascertained. Additionally, some experiments showed that modifications to the hyper-parameter settings can require a reevaluation of the employed regularisation strategies, especially the dropout rate. However, further research would be required to accurately evaluate the extent of this effect.

The results obtained with a simple median speckle filter showed that such filtering of the SAR data negatively impacts the classification performance of a ViT model trained on S1 images. Similarly, training a ViT model on standardised S1 images, as is common practice in the CV domain, outperforms both training on unprocessed images as well as on min-max scaled images.

In general, it could be shown that the multi-modal fusion methods investigated in this thesis achieve good results over varying hyper-parameter settings. Furthermore, the obtained results support the selection of hyper-parameters for the previous comparisons of the multi-modal fusion methods.

8 Conclusion and Discussion

This thesis investigates the capabilities of utilising the ViT architecture to fuse multi-spectral and SAR remote sensing images. To this end, multiple multi-modal fusion methods based on the standard ViT architecture are proposed and evaluated by comparing their performance on a scene classification task. All proposed approaches adapt specific properties of the ViT architecture to identify a modification strategy that yields the best overall performance on the classification task.

The investigated fusion methods can be divided into two principal categories. The first category of methods functions by modifying the procedure by which feature embeddings are generated from both modalities but rely on the standard ViT model for processing the generated input sequences. Early Fusion derives input embeddings directly from a tensor of both modalities concatenated in the channel dimension. Conversely, Modality Token Fusion generates separate patch embeddings from each input modality and combines all derived patches into one input sequence. Channel Token Fusion further extends this separation by deriving patch embeddings from individual channel patches for all channels in both modalities and combining them in one input sequence.

The second type of fusion methods encompasses three approaches which perform multi-modal fusion by exchanging information between separated modality-specific encoders based on the ViT model. Middle Fusion relies on separate Transformer Encoders to derive abstract features from each modality and employs a third encoder to combine the abstract feature representations. Cross-Attention Fusion replaces the self-attention in two Transformer Encoders with a Cross-Attention computation between the feature sequences derived from each modality. The newly proposed SCT Fusion exchanges information between two modality-specific encoders by continuously fusing the class tokens between the attention layers to extract and condense the most relevant information for the classification task.

The performance of these fusion methods is evaluated in detail on the large-scale remote sensing archive BigEarthNet-MM, which provides considerable quantities of labelled multi-spectral and SAR images. Because all fusion methods receive the same data as inputs but process it differently, a fusion model's overall performance on the classification task provides insights into its capabilities in effectively deriving information from multiple modalities.

Results showed that Channel Token Fusion and SCT Fusion achieve the best overall performance for fusing MSI and SAR data. Early Fusion and Middle Fusion follow closely by generating good but subpar scores on all metrics. However, the performance of Modality Token Fusion and Cross-Attention Fusion could only marginally improve upon results from a model trained on single modality input

data.

From all investigated fusion methods, Channel Token Fusion is considerably less viable for most applications due to its excessive computational requirements and only marginal improvements it could provide over SCT Fusion. Therefore, SCT Fusion presents the best-performing candidate for the fusion of multi-spectral and SAR images with practical applicability among the investigated methods. Early Fusion and Middle Fusion also provide decent overall results, with Early Fusion having the additional advantage of requiring even less computational resources than all other models.

Additional comparisons on a reduced fusion task consisting of RGB and SAR images were conducted to assess the impact of the difference in features provided by each modality on the fusion performance. While Channel Token Fusion achieved the best overall results on the reduced fusion task, all other methods generated results in a very similar range to each other. This implies that the performance advantages achieved by SCT Fusion on the MSI and SAR fusion task potentially stem from a higher capability of SCT Fusion to discern important features from modalities with highly disparate amounts of features. On the contrary, Modality Token Fusion and Cross-Attention Fusion might be more susceptible to such differences, which would explain their poor performance on the MSI and SAR fusion task.

Detailed ablation studies were conducted to assess the impact of specific hyperparameters on the fusion performance of the different fusion methods. The results show that reducing the *patch size* of a fusion method holds the highest potential for further performance improvements. Interestingly, experiments on reducing the number of layers in the Transformer Encoders indicate that the investigated fusion methods can still reliably generate competitive results even with relatively shallow models.

It could be shown that the application of data augmentations and other regularisation strategies has a positive impact on the overall performance achieved by the fusion methods. However, regularisation was used conservatively during the training, and future exploratory studies on the selection of optimal dropout rate and data augmentations might reveal further potential for improvements. Experiments on speckle filtering or applying different normalisation strategies to the SAR images were also performed, but no significant potential for improvement could be identified here.

A possible direction for future improvements could be to combine the fusion methodologies from the best-performing methods, namely Channel Token Fusion and SCT Fusion. Here it would be of significant interest to investigate how the processing requirements of Channel Token Fusion could be reduced while still retaining its performance advantage. Furthermore, even better performance improvements might be enabled by incorporating Channel Token Fusion with the condensing of class-related information from both modalities performed in the SCT Fusion approach.

Additionally, the experiments revealed that data augmentations that modify both modalities differently could significantly impact the generalisation capabilities of the various fusion methods. However, the data augmentation strategies employed in this thesis are relatively simplistic. Future research could focus on the integration of more sophisticated data augmentation techniques. Potential ideas for data augmentations of interest could be the partial or complete masking of channels or the distortion of pixel information at a specific location in an image from one modality to force the model to derive inputs from the other modality.

Similarly, other research in the multi-modal fusion domain for remote sensing images has incorporated a reconstruction step of the original modalities from the encoded features into the training phase. Including such a step into the training of the investigated fusion methods could also improve the overall performance.

Furthermore, due to the inherent variability of SAR images depicting the same location, it could be beneficial to add additional SAR samples to the dataset depicting each location captured from different directions. Such additions might better convey the relationship between specific backscatter responses and their underlying ground geometry. The generation of artificial adversarial SAR samples could also be considered to aid in learning the connection between backscatter responses and their underlying causes. Moreover, testing the multi-modal fusion methods on other relevant tasks from the remote sensing domain, such as LULC classification or CBIR, could provide more insight into the generalisability of the results obtained in this thesis.

Bibliography

[1] Alan S. Belward and Jon O. Skøien. “Who launched what, when and why; trends in global land-cover observation capacity from civilian earth observation satellites”. In: *ISPRS Journal of Photogrammetry and Remote Sensing* 103 (2015). Global Land Cover Mapping and Monitoring, pp. 115–128.

[2] Jiaxin Li et al. “Deep learning in multimodal remote sensing data fusion: A comprehensive review”. In: *International Journal of Applied Earth Observation and Geoinformation* 112 (2022), p. 102926.

[3] Michael Schmitt and Xiao Xiang Zhu. “Data Fusion and Remote Sensing: An ever-growing relationship”. In: *IEEE Geoscience and Remote Sensing Magazine* 4.4 (2016), pp. 6–23.

[4] Y. LeCun et al. “Backpropagation Applied to Handwritten Zip Code Recognition”. In: *Neural Comput.* 1.4 (Dec. 1989), pp. 541–551.

[5] Alex Krizhevsky et al. “ImageNet Classification with Deep Convolutional Neural Networks”. In: *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*. 2012.

[6] Kaiming He et al. “Deep Residual Learning for Image Recognition”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 770–778.

[7] Sergey Ioffe and Christian Szegedy. “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift”. In: *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*. ICML’15. Lille, France: JMLR.org, 2015, pp. 448–456.

[8] Karen Simonyan and Andrew Zisserman. “Very Deep Convolutional Networks for Large-Scale Image Recognition”. In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. 2015.

[9] John E. Ball et al. “Comprehensive survey of deep learning in remote sensing: theories, tools, and challenges for the community”. In: *Journal of Applied Remote Sensing* 11.4 (Sept. 23, 2017), p. 042609.

[10] Alexey Dosovitskiy et al. “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”. In: *International Conference on Learning Representations*. OpenReview.net, 2021.

[11] Olga Russakovsky et al. “ImageNet Large Scale Visual Recognition Challenge”. In: *Int. J. Comput. Vis.* 115.3 (2015), pp. 211–252.

[12] Ashish Vaswani et al. “Attention is All you Need”. In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017.

[13] Peng Xu et al. “Multimodal Learning with Transformers: A Survey”. In: *CoRR* abs/2206.06488 (2022). arXiv: 2206.06488 [cs.CV].

[14] Gencer Sümbül et al. “BigEarthNet-MM: A Large-Scale, Multimodal, Multilabel Benchmark Archive for Remote Sensing Image Classification and Retrieval [Software and Data Sets]”. In: *IEEE Geoscience and Remote Sensing Magazine* 9.3 (Sept. 2021), pp. 174–180.

[15] European Space Agency. *Sentinel-2: ESA’s Optical High-Resolution Mission for GMES Operational Services*. Ed. by K. Fletcher. ESTEC, PO Box 299, 2200 AG Noordwijk, The Netherland: ESA Communications, 2012.

[16] Dorijan Radočaj et al. “Global Open Data Remote Sensing Satellite Missions for Land Monitoring and Conservation: A Review”. In: *Land* 9.11 (2020), p. 402.

[17] Andreas Braun. “Radar satellite imagery for humanitarian response. Bridging the gap between technology and application”. PhD thesis. Eberhard Karls Universität Tübingen, Aug. 2019.

[18] C. Palmann et al. “Earth observation using radar data: an overview of applications and challenges”. In: *International Journal of Digital Earth* 1.2 (2008), pp. 171–195.

[19] Yuan Gong et al. “AST: Audio Spectrogram Transformer”. In: *Proc. Interspeech 2021*. 2021, pp. 571–575.

[20] Meng-Hao Guo et al. “PCT: Point cloud transformer”. In: *Computational Visual Media* 7.2 (Apr. 2021), pp. 187–199.

[21] Lei Jimmy Ba et al. “Layer Normalization”. In: *CoRR* abs/1607.06450 (2016). arXiv: 1607.06450 [stat.ML].

[22] Abien Fred Agarap. “Deep Learning using Rectified Linear Units (ReLU)”. In: *CoRR* abs/1803.08375 (2018). arXiv: 1803.08375 [cs.NE].

[23] Dan Hendrycks and Kevin Gimpel. “Bridging Nonlinearities and Stochastic Regularizers with Gaussian Error Linear Units”. In: *CoRR* abs/1606.08415 (2016). arXiv: 1606.08415 [cs.LG].

[24] J.A. Benediktsson and P.H. Swain. “A Method of Statistical Multisource Classification with a Mechanism to Weight the Influence of the Data Sources”. In: *12th Canadian Symposium on Remote Sensing Geoscience and Remote Sensing Symposium,* vol. 2. 1989, pp. 517–520.

[25] A.H.S. Solberg et al. “Multisource classification of remotely sensed data: fusion of Landsat TM and SAR images”. In: *IEEE Transactions on Geoscience and Remote Sensing* 32.4 (1994), pp. 768–778.

[26] L. Bruzzone et al. “A neural-statistical approach to multitemporal and multisource remote-sensing image classification”. In: *IEEE Transactions on Geoscience and Remote Sensing* 37.3 (1999), pp. 1350–1359.

[27] J.A. Benediktsson et al. “Neural Network Approaches Versus Statistical Methods In Classification Of Multisource Remote Sensing Data”. In: *IEEE Transactions on Geoscience and Remote Sensing* 28.4 (1990), pp. 540–552.

[28] Behnood Rasti and Pedram Ghamisi. “Remote sensing image classification using subspace sensor fusion”. In: *Information Fusion* 64 (2020), pp. 121–130.

[29] Danfeng Hong et al. “CoSpace: Common Subspace Learning From Hyperspectral-Multispectral Correspondences”. In: *IEEE Transactions on Geoscience and Remote Sensing* 57.7 (2019), pp. 4349–4359.

[30] Jingliang Hu et al. “A Comparative Review of Manifold Learning Techniques for Hyperspectral and Polarimetric SAR Image Fusion”. In: *Remote Sensing* 11.6 (2019), p. 681.

[31] Danfeng Hong et al. “Learnable manifold alignment (LeMA): A semi-supervised cross-modality learning framework for land cover and land use classification”. In: *ISPRS Journal of Photogrammetry and Remote Sensing* 147 (2019), pp. 193–205.

[32] Renlong Hang et al. “Classification of Hyperspectral and LiDAR Data Using Coupled CNNs”. In: *IEEE Transactions on Geoscience and Remote Sensing* 58.7 (2020), pp. 4939–4950.

[33] Danfeng Hong et al. “Multimodal Convolutional Neural Networks with Cross-Channel Reconstruction”. In: *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*. 2021, pp. 282–285.

[34] Danfeng Hong et al. “X-ModalNet: A semi-supervised deep cross-modal network for classification of remote sensing data”. In: *ISPRS Journal of Photogrammetry and Remote Sensing* 167 (2020), pp. 12–23.

[35] Wei Li et al. “Asymmetric Feature Fusion Network for Hyperspectral and SAR Image Classification”. In: *IEEE Transactions on Neural Networks and Learning Systems* (2022), pp. 1–14.

[36] Nicolas Audebert et al. “Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks”. In: *ISPRS Journal of Photogrammetry and Remote Sensing* 140 (2018). Geospatial Computer Vision, pp. 20–32.

[37] Danfeng Hong et al. “More Diverse Means Better: Multimodal Deep Learning Meets Remote-Sensing Imagery Classification”. In: *IEEE Transactions on Geoscience and Remote Sensing* 59.5 (2021), pp. 4340–4354.

[38] Qinghui Liu et al. “Multi-modal land cover mapping of remote sensing images using pyramid attention and gated fusion networks”. In: *International Journal of Remote Sensing* 43.9 (2022), pp. 3509–3535.

[39] Tao Lei et al. “Multi-Modality and Multi-Scale Attention Fusion Network for Land Cover Classification from VHR Remote Sensing Images”. In: *Remote Sensing* 13.18 (2021), p. 3771.

[40] Runyu Fan et al. “Urban informal settlements classification via a transformer-based spatial-temporal fusion network using multimodal remote sensing and time-series human activity data”. In: *International Journal of Applied Earth Observation and Geoinformation* 111 (2022), p. 102831.

[41] Xianping Ma et al. “A Crossmodal Multiscale Fusion Network for Semantic Segmentation of Remote Sensing Data”. In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 15 (2022), pp. 3463–3474.

[42] Yi Wang et al. “Self-supervised Vision Transformers for joint SAR-optical representation learning”. In: *IGARSS 2022*. 2022, pp. 1–4.

[43] Zhixiang Xue et al. “Self-Supervised Feature Learning for Multimodal Remote Sensing Image Land Cover Classification”. In: *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), pp. 1–15.

[44] Swalpa Kumar Roy et al. “Multimodal Fusion Transformer for Remote Sensing Image Classification”. In: *CoRR* abs/2203.16952 (2022). arXiv: 2203.16952 [cs.CV].

[45] Zhixiang Xue et al. “Deep Hierarchical Vision Transformer for Hyperspectral and LiDAR Data Classification”. In: *IEEE Transactions on Image Processing* 31 (2022), pp. 3095–3110.

[46] Dana Lahat et al. “Multimodal Data Fusion: An Overview of Methods, Challenges, and Prospects”. In: *Proceedings of the IEEE* 103.9 (2015), pp. 1449–1477.

[47] Andrew Jaegle et al. “Perceiver IO: A General Architecture for Structured Inputs & Outputs”. In: *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

[48] Arsha Nagrani et al. “Attention Bottlenecks for Multimodal Fusion”. In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., 2021, pp. 14200–14213.

[49] Nina Shvetsova et al. “Everything at Once - Multi-modal Fusion Transformer for Video Retrieval”. In: *CoRR* abs/2112.04446 (2021). arXiv: 2112.04446 [cs.CV].

[50] Valentin Gabeur et al. “Multi-modal Transformer for Video Retrieval”. In: *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part IV*. Ed. by Andrea Vedaldi et al. Vol. 12349. Lecture Notes in Computer Science. Springer, 2020, pp. 214–229.

[51] Shaowei Yao and Xiaojun Wan. “Multimodal Transformer for Multimodal Machine Translation”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 4346–4350.

[52] Peizhao Li et al. “SelfDoc: Self-Supervised Document Representation Learning”. In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 5648–5656.

[53] Chun-Mei Feng et al. “Multi-Modal Transformer for Accelerated MR Imaging”. In: *IEEE Transactions on Medical Imaging* (2022), pp. 1–1.

[54] Yanan Zhang et al. “CAT-Det: Contrastively Augmented Transformer for Multi-Modal 3D Object Detection”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2022, pp. 908–917.

[55] Wei Zhang et al. “Transformer-Based Multimodal Information Fusion for Facial Expression Analysis”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. June 2022, pp. 2428–2437.

[56] Chen Sun et al. “VideoBERT: A Joint Model for Video and Language Representation Learning”. In: *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 2019, pp. 7463–7472.

[57] Renjie Zheng et al. “Fused Acoustic and Text Encoding for Multimodal Bilingual Pretraining and Speech Translation”. In: *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, 2021, pp. 12736–12746.

[58] Danfeng Hong et al. “SpectralFormer: Rethinking Hyperspectral Image Classification With Transformers”. In: *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), pp. 1–15.

[59] Jiasen Lu et al. “ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks”. In: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*. Ed. by Hanna M. Wallach et al. 2019, pp. 13–23.

[60] Chun-Fu (Richard) Chen et al. “CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification”. In: *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, 2021, pp. 347–356.

[61] Gencer Sumbul et al. “Bigearthnet: A Large-Scale Benchmark Archive for Remote Sensing Image Understanding”. In: *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*. 2019, pp. 5901–5904.

[62] Jesús Delegido et al. “Evaluation of Sentinel-2 Red-Edge Bands for Empirical Estimation of Green LAI and Chlorophyll Content”. In: *Sensors* 11.7 (2011), pp. 7063–7081.

[63] European Space Agency. *Sentinel-1: ESA’s Radar Observatory Mission for GMES Operational Services*. Ed. by K. Fletcher. ESTEC, PO Box 299, 2200 AG Noordwijk, The Netherland: ESA Communications, 2012.

[64] Kai Norman Clasen. “Large-scale contrastive self-supervised representation learning for content-based remote sensing image retrieval”. MA thesis. Berlin: Technische Universität Berlin, 2021.

[65] Adam Paszke et al. “PyTorch: An Imperative Style, High-Performance Deep Learning Library”. In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach et al. Curran Associates, Inc., 2019, pp. 8024–8035.

[66] Ross Wightman. *PyTorch Image Models*. `https://github.com/rwightman/pytorch-image-models` (accessed Aug. 20, 2022).

[67] Diederik P. Kingma and Jimmy Ba. “Adam: A Method for Stochastic Optimization”. In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. 2015.

[68] Ilya Loshchilov and Frank Hutter. “SGDR: Stochastic Gradient Descent with Warm Restarts”. In: *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[69] Jong-Sen Lee. “Speckle analysis and smoothing of synthetic aperture radar images”. In: *Computer Graphics and Image Processing* 17.1 (1981), pp. 24–32.

[70] Fang Qiu et al. “Speckle Noise Reduction in SAR Imagery Using a Local Adaptive Median Filter”. In: *Giscience & Remote Sensing - GISCI REMOTE SENS* 41 (Sept. 2004), pp. 244–266.

[71] Zhenghao Shi and K.B. Fung. “A comparison of digital speckle filters”. In: *Proceedings of IGARSS ’94 - 1994 IEEE International Geoscience and Remote Sensing Symposium*. Vol. 4. 1994, 2129–2133 vol.4.

[72] Nitish Srivastava et al. “Dropout: A Simple Way to Prevent Neural Networks from Overfitting”. In: *Journal of Machine Learning Research* 15.56 (2014), pp. 1929–1958.

[73] Gao Huang et al. “Deep Networks with Stochastic Depth”. In: *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV*. Ed. by Bastian Leibe et al. Vol. 9908. Lecture Notes in Computer Science. Springer, 2016, pp. 646–661.

[74] Andreas Steiner et al. “How to train your ViT? Data, Augmentation, and Regularization in Vision Transformers”. In: *CoRR* abs/2106.10270 (2021). arXiv: `2106.10270 [cs.CV]`.

[75] Zehui Lin et al. “DropAttention: A Regularization Method for Fully-Connected Self-Attention Networks”. In: *CoRR* abs/1907.11065 (2019). arXiv: 1907.11065 [cs.CL].

A Appendix

Table A.1: The different classes present in the BigEarthNet-MM dataset as well as the total number of samples belonging to each class. Note that patches often belong to multiple classes and therefore count towards each of these separately.

Class name	Number of patches
Agro-forestry areas	30 649
Arable land	194 148
Beaches, dunes, sands	1536
Broad-leaved forest	141 300
Coastal wetlands	1566
Complex cultivation patterns	104 203
Coniferous forest	164 775
Industrial or commercial units	11 865
Inland waters	67 277
Inland wetlands	22 100
Land principally occupied by agriculture, with significant areas of natural vegetation	130 637
Marine waters	74 877
Mixed forest	176 567
Moors, heathland and sclerophyllous vegetation	16 267
Natural grassland and sparsely vegetated areas	12 022
Pastures	98 997
Permanent crops	29 350
Transitional woodland, shrub	148 950
Urban fabric	74 891

Table A.2: The class-wise performance of the different multi-modal fusion methods on fusing RGB and SAR images. All scores given are the AP.

Class name	Early Fusion	Modality Token Fusion	Channel Token Fusion	Middle Fusion	Cross-Attention Fusion	SCT Fusion
Agro-forestry areas	0.8501	0.8530	**0.8685**	0.8545	0.8528	0.8511
Arable land	0.9416	0.9429	**0.9478**	0.9428	0.9428	0.9422
Beaches, dunes, sands	0.6619	0.6594	**0.7091**	0.6408	0.6680	0.6728
Broad-leaved forest	0.8732	0.8759	**0.8835**	0.8747	0.8737	0.8724
Coastal wetlands	0.6174	0.6477	**0.6700**	0.6216	0.6283	0.6206
Complex cultivation patterns	0.7835	0.7884	**0.7994**	0.7854	0.7869	0.7836
Coniferous forest	0.9453	0.9469	**0.9499**	0.9459	0.9461	0.9462
Industrial or commercial units	0.5531	0.5649	**0.5806**	0.5610	0.5569	0.5609
Inland waters	0.9189	0.9225	**0.9337**	0.9261	0.9218	0.9252
Inland wetlands	0.7166	0.7230	**0.7444**	0.7204	0.7222	0.7196
Land principally occupied by agriculture, with significant areas of natural vegetation	0.7561	0.7603	**0.7697**	0.7588	0.7590	0.7576
Marine waters	0.9987	0.9989	**0.9991**	0.9990	0.9989	0.9989
Mixed forest	0.9064	0.9084	**0.9152**	0.9071	0.9075	0.9069
Moors, heathland and sclerophyllous vegetation	0.7079	0.7176	**0.7404**	0.7090	0.7115	0.7136
Natural grassland and sparsely vegetated areas	0.6222	0.6322	**0.6498**	0.6238	0.6261	0.6159
Pastures	0.8692	0.8708	**0.8778**	0.8683	0.8705	0.8690
Permanent crops	0.7226	0.7290	**0.7517**	0.7289	0.7264	0.7207
Transitional woodland, shrub	0.7625	0.7669	**0.7735**	0.7648	0.7644	0.7662
Urban fabric	0.8433	0.8474	**0.8623**	0.8496	0.8432	0.8466

Table A.3: The results of additional training runs performed for Early Fusion to verify the stability of training the multi-modal fusion methods. The Early Fusion results are inferior to the ones reported in the comparison because no dropout was employed.

Experiment	AP (micro)	AP (macro)	F_2	HL
Experiment 1	0.8949	0.8197	0.7802	0.0594
Experiment 2	0.8942	0.8212	0.7814	0.0597
Experiment 3	0.8938	0.8203	0.7823	0.0599

Table A.4: The results of additional training runs performed for SCT Fusion to verify the stability of training the multi-modal fusion methods.

Experiment	AP (micro)	AP (macro)	F_2	HL
Experiment 1	0.8993	0.8295	0.7848	0.0582
Experiment 2	0.8990	0.8286	0.7848	0.0582
Experiment 3	0.8990	0.8292	0.7849	0.0582

Table A.5: The performance of Early Fusion and SCT Fusion with a *patch size* of 10 and a dropout rate of 0.1.

Fusion Method	AP (micro)	AP (macro)
Early Fusion	0.9016	0.8318
SCT Fusion	**0.9031**	**0.8343**

Table A.6: Classification performance of Early Fusion and SCT Fusion over varying *embedding dimensions*.

Embedding Dimension	Early Fusion		SCT Fusion	
	AP (micro)	AP (macro)	AP (micro)	AP (macro)
128	0.8948	**0.8208**	0.8978	0.8263
256	**0.8949**	0.8197	**0.8990**	**0.8286**
384	0.8922	0.8182	0.8305	0.6772

Table A.7: Classification performance of Middle Fusion under varying *depth* settings for the modality encoders and the fusion encoder.

Modality Encoder Depth	Fusion Encoder Depth	AP (micro)	AP (macro)
1	7	0.8929	0.8194
2	6	0.8926	0.8199
4	4	0.8953	0.8222
6	2	**0.8955**	**0.8244**
7	1	0.8952	0.8207