Università degli Studi di Napoli Federico II

Dipartimento di Fisica "E. Pancini"

Dottorato di Ricerca in Fisica

CICLO XXXVIII (2022–2025)

SETTORE SCIENTIFICO DISCIPLINARE FIS/08

Tesi di Dottorato

A Deep Learning Approach to the Analysis of Complex Datacubes

Tutore:
Prof. Giuseppe Longo

Candidato: Andrea Dosi

Acknowledgments

Thanks to my research group for the support and encouragement. It was not easy to believe me despite my full-time job.
I am particularly grateful to Prof. Giuseppe Longo for his guidance throughout my Ph.D. Professor Longo has been a *master* in the ancient Greek meaning: he has provided scientific education and also human development. I am also grateful to Prof. Longo who has become my friend and whose generosity and humanity have enriched this journey.

Abstract

Modern scientific imaging increasingly relies on the analysis of high-dimensional data acquired through complex and often indirect measurement processes. Hyperspectral Earth observation and radio astronomy are two representative examples of this trend: both generate structured datacubes whose interpretation is far from trivial and requires models able to combine local correlations with global context, often under significant uncertainty. Against this background, this thesis focuses on the design of efficient deep learning architectures for hyperspectral image analysis and frames these developments within the broader perspective of inverse problems in scientific imaging.

The first part of the thesis introduces **AMBER**, an advanced SegFormer-based architecture [1] designed for semantic segmentation of hyperspectral datacubes. This chapter reports on the corresponding paper already published for *Neural Computing and Applications, Springer Nature*[2]. One of AMBER's main features is its ability to operate directly on hyperspectral datacubes, without preprocessing with dimensionality reduction methods (e.g., SVD). The architecture (the transformer) was designed to jointly encode spectral and spatial information: a 3D convolution combined with multi-head self-attention.

The thesis then extends the preprint "**Less is More: AMBER-AFNO – a New Benchmark for Lightweight 3D Medical Image Segmentation**[3]". In this lightweight AMBER variant, Multi-Head Self-Attention is replaced by Adaptive Fourier Neural Operators[4], and the decoder is suited for 3D segmentation. Despite significantly reducing model complexity -far fewer parameters compared with the SOTA models in this field- this design retains the ability to capture global dependencies, making it particularly attractive for high-dimensional inputs.

The work subsequently moves toward the **foundation model paradigm** by pretraining the AMBER-AFNO encoder in a self-supervised manner (using DINO [5]) on the large-scale *SpectralEarth* hyperspectral dataset, and then evaluates the semantic segmentation performance on the DESIS-CDL dataset [6]. In this setting, AMBER-AFNO emerges as a hyperspectral foundation model, capable of learning representations that transfer across sensors, geographic regions, and downstream semantic segmentation tasks.

Finally, these outcomes are put into perspective with respect to the inverse problems in astrophysics [7]. Although the physics of the two problems is obviously distinct, hyperspectral imaging and astrophysical observation have in common that they are measurement processes governed by an operator which gives rise to high-dimensional data cubes. From this perspective, this thesis highlights some computational patterns and suggests that there is value in architectural solutions that aim at capturing global structure with a reasonable computational cost.

Contents

Chapter 1

Introduction

Scientific discovery is increasingly shaped by the availability of high-dimensional data acquired through complex measurement systems. Across domains such as astronomy, Earth observation, and biomedical imaging, advances in instrumentation have enabled the routine collection of massive datasets whose interpretation requires sophisticated computational models. In many modern pipelines, the central challenge is no longer the acquisition of data, but the extraction of reliable and generalizable knowledge from observations that are indirect, noisy, and often incomplete.

A useful and unifying perspective for understanding this challenge is the language of *inverse problems*. In a broad class of scientific settings, the measured data are related to an underlying physical quantity of interest through a forward operator encoding the acquisition process. Recovering the latent quantity from the observation corresponds to an inverse problem, which is frequently ill-posed in the sense of Hadamard: solutions may fail to exist, may not be unique, and can be unstable under perturbations of the data. This thesis adopts this viewpoint as a conceptual thread—not because the work is primarily about solving a single inverse problem, but because operator-induced ill-posedness is a common structure underpinning many imaging tasks, and it motivates architectural choices that recur throughout the thesis.

In radio astronomy, inverse problems constitute the core of synthesis imaging: the measured visibilities correspond to an incomplete sampling of the Fourier transform of the sky brightness distribution, and image reconstruction is intrinsically ill-posed due to missing spatial frequencies, noise, and instrumental effects. The forward model can be expressed as a linear operator and, under shift-invariance assumptions, as a convolution with the instrument point spread function. This operator-theoretic formulation, together with its Fourier representation, provides a principled foundation for regularization and uncertainty-aware inference, and will be discussed in depth in Chapter 5. Such ideas are not restricted to astronomy: they appear whenever measurements are mediated by physics, sensors, or acquisition protocols that filter or compress information.

Another example from the remote sensing field is the hyperspectral imaging (HSI). HSI collects hundreds of adjacent bands of spectral data per pixel, making a 3-D volume data cube. HSI has been widely used to identify land cover types and mineralogical signatures, detecting weathering and mineral alterations, oil detection, and characterizing vegetation. However, HSI data has its own set of

challenges, including: (1) the high dimensionality of data; (2) redundancy between the features; (3) noises from atmosphere and sensor; (4) the difficulty and expensiveness of acquiring labeled samples; and (5) the robustness of the trained model to different geographical areas, acquisition times and sensors.

Historically, hyperspectral analysis relied on physics-inspired models and statistical techniques, including dimensionality reduction, spectral unmixing, and handcrafted features. The rise of deep learning introduced end-to-end approaches capable of learning spectral–spatial features directly from data. Convolutional networks, in particular, improved classification and segmentation performance by exploiting local spatial context and hierarchical feature extraction. Nevertheless, convolutional inductive biases remain fundamentally local, and purely CNN-based architectures can struggle to capture long-range dependencies and global context that are important for structured scenes and large-scale segmentation.

The emergence of transformer architectures has reshaped representation learning by enabling explicit global interaction modeling. Vision Transformers (ViTs), inspired by sequence models in natural language processing, demonstrated that attention-based token mixing can compete with or surpass convolutional designs when trained at scale [8, 9]. These encouraging results on 3-channel images motivated a growing literature on transformer-based designs for hyperspectral images[10, 11, 12]. But applying attention-based architectures directly to hyperspectral data is non-trivial. Self-attention scales quadratically with the number of tokens, and hyperspectral imagery naturally increases token dimensionality if spectral structure is preserved. Consequently, many approaches resort to spectral dimensionality reduction, band grouping, or hybrid CNN–transformer designs. While effective in specific settings, these strategies may compromise spectral fidelity or limit scalability.

In recent years, machine learning has undergone a conceptual shift toward *foundation models*. Foundation models are pretrained on large, diverse datasets using self-supervised objectives, producing general-purpose representations that can be adapted to downstream tasks with minimal supervision. This paradigm has transformed natural language processing and has rapidly gained traction in computer vision and remote sensing. In geospatial imagery, the availability of massive open multispectral and RGB archives has enabled training at scale, and pretrained backbones have shown strong generalization across tasks and regions. Hyperspectral imaging has historically lagged behind due to the lack of large, globally representative pretraining corpora.

Recent spaceborne missions have changed the landscape. EnMAP, DESIS, and PRISMA have increased the availability of hyperspectral observations at a global scale [13, 14, 15], with CHIME expected to further expand temporal coverage [16]. Currently, large datasets such as SpectralEarth[6] facilitate self-supervised pretraining, thereby enabling the development of hyperspectral foundation models. This thesis is motivated by the hypothesis that hyperspectral foundation models require architectures explicitly designed to: (i) preserve spectral–spatial structure without aggressive preprocessing; (ii) capture global context efficiently at scale; and (iii) generalize across sensors and acquisition conditions. To this end, the thesis introduces the *AMBER* family of models.

None of the current state-of-the-art foundation models in hyperspectral analy-

sis use a transformer that jointly processes both spectral and spatial information; instead, they handle these types of information separately. This separation is primarily motivated by the aim to prevent a substantial increase in model parameters[6]. In this context, our approach differs by using AMBER-AFNO, which replaces multi-head self-attention with an Adaptive Fourier Neural Operator (AFNO) token mixer, as described in previous work[4, 3]. This token mixer can process spatial and spectral information together, addressing the parameter increase issue observed in other models. This modification is more than an engineering optimization. It highlights the link between operator structure and representation learning. Fourier-domain mixing provides global receptive fields and scales computations efficiently. This matches the operator-theoretic view common in inverse problems, where Fourier representations and convolution operators play a central role. While AMBER-AFNO is mainly for efficient hyperspectral segmentation, its architecture also learns global transformations that mirror operator mappings.

The thesis shifts from task-specific modeling to the foundation model paradigm. The AMBER-AFNO encoder is pretrained using the DINO self-supervised framework [5], yielding a hyperspectral backbone adaptable to downstream tasks. This step addresses a long-standing limitation in hyperspectral learning: the scarcity of large-scale, globally diverse data suitable for self-supervised pretraining. By combining a hyperspectral-native architecture with a large, multi-temporal dataset, the thesis aims to demonstrate that strong, transferable hyperspectral representations can be learned at scale.

Last but not least, it is worth mentioning that many of the current problems of astrophysics can be posed as inverse problems. An inverse problem is a procedure to retrieve a latent physical entity given a finite set of noisy indirect measurements. The most common examples are deconvolution and radio-interferometric imaging. In these two cases the measurements are some samples of the Fourier transform of the sky brightness distribution, that makes the reconstruction problem ill-posed. I do not consider these techniques here as a different topic, but more like a common scientific background. In the astrophysical context we learn that scientific data are often the result of an instrumental acquisition procedure, and that we need powerful models of the global structure that underlie these data in order to perform robust inference in presence of errors and limited data.

Methodologically, the relation with hyperspectral imaging is not in the similarity of the physics involved but in the similarity of the computational traits of the *data modality*. The data of hyperspectral imaging form an intrinsically structured *datacube*, each point of which is a data point with a high-dimensional spectrum. The semantics of the data is in the union of the spatial and spectral domains. In this sense, measurement products of astrophysical observations (e.g., multi-frequency images and spectral cubes) are high-dimensional datacubes like datacubes of hyperspectral Earth observation, and both are suited to the usage of deep learning methods capable of modeling local associations and non-local correlations. This work utilizes this angle by exploring and investigating the deep learning models capable of efficient and accurate processing of such high-dimensional, multimodal datacubes, while the primary interest of this work remains in hyperspectral image segmentation and representation learning.

Organization of the thesis. Chapter 2, titled **AMBER: advanced SegFormer for multi-band image segmentation**, is devoted to the particular design decisions made for hyperspectral images, along with the associated experimental analysis. We propose using **AMBER-AFNO** as a baseline model and assess its efficiency for segmenting 3D medical data in Chapter 3, using the Adaptive Fourier Neural Operator (AFNO) instead of multi-head self-attention to reduce the number of parameters without sacrificing accuracy. We discuss the use of **AMBER-AFNO as a hyperspectral backbone** in Chapter 4, as well as large-scale DINO pretraining on the *SpectralEarth* dataset and initial downstream evaluations. Chapter 5, we look at the big picture, and show how to **solve astrophysical inverse problems using deep learning**, highlighting how current astrophysical and remote-sensing data-reduction pipelines are progressively based on learning-based approaches to invert high-dimensional datacubes.

Chapter 2

AMBER: advanced SegFormer for multi-band image segmentation—an application to hyperspectral[2]

Hyperspectral imaging (HSI) records, for each spatial location, a dense spectrum over dozens to hundreds of contiguous bands. This representation enables fine-grained material discrimination and supports a wide range of remote-sensing applications, from environmental monitoring to agriculture and mineral exploration. At the same time, the high dimensionality of hyperspectral data introduces substantial challenges for representation learning: the spectral axis is informative but redundant, and the joint modeling of spatial and spectral structure is computationally demanding.

Convolutional Neural Networks (CNNs) have been widely employed in the classification and segmentation of hyperspectral images (HSI) in recent years, owing to their superior capability of extracting meaningful spectral-spatial information from the data [10, 11]. Nevertheless, CNNs are inherently local: they are based on small size filters which can only take into account local spatial context and generally need stacking up many layers to model the contextual information to cover a larger region in the field of view. In many hyperspectral real-world images, however, classes of interest might depend on spatial relationships that are defined at a larger scale or long-range dependencies which can not be captured by a simple stack of convolutional layers.

This issue is alleviated in transformer-based models that implicitly learn global contextual information by self-attention mechanisms. Vision Transformers (ViTs) [46] and its variants that model long-range dependencies can be important for robust scene analysis. However, it is not necessarily the best practice to directly apply models originally designed for RGB images to HSI data. Most of the existing approaches perform dimensionality reduction or ad hoc preprocessing to transform the input into a three-channel image that may lose valuable spectral information for segmentation.

In this chapter, we introduce **AMBER**, a SegFormer-based [1] architecture designed for *multi-band* segmentation, with hyperspectral imaging as the primary target domain. AMBER integrates three-dimensional convolutions, custom ker-

nel configurations, and a dedicated *Funnelizer* layer to explicitly model joint spectral–spatial structure while producing standard two-dimensional segmentation masks [2]. A central design choice is that AMBER processes hyperspectral cubes in their native form, avoiding spectral dimensionality reduction as a mandatory preprocessing step (e.g., SVD or Gabor filtering [17, 18]). The chapter develops the architectural design, describes the experimental protocol, and reports results on widely used airborne benchmarks—Salinas, Indian Pines, and Pavia University—as well as on spaceborne PRISMA data. Across these settings, AMBER demonstrates strong accuracy and robustness, supporting its use as a general foundation for multi-band segmentation.

This chapter is based on the paper **AMBER: Advanced SegFormer for Multi-Band Image Segmentation—An Application to Hyperspectral Imaging**[2], published in *Neural Computing and Applications* (Springer Nature). The candidate is the first author of this publication. Parts of the text and figures are reproduced with permission from Springer Nature.

2.1 Introduction

Hyperspectral image classification and segmentation aim to assign a semantic label to each pixel based on its spectral signature and spatial context. The task is challenging for at least three reasons: (i) the spectral dimension is high and often redundant; (ii) class boundaries may be spatially complex and depend on context beyond local neighborhoods; and (iii) practical datasets frequently exhibit strong class imbalance and limited labeled samples.

CNN-based pipelines have achieved strong results by learning hierarchical spectral–spatial representations [10, 11]. Yet, their difficulty in capturing long-range dependencies has motivated a growing literature on transformer-based solutions for hyperspectral data [19]. ViTs [9] provide a principled mechanism to integrate global context via self-attention, and recent hyperspectral transformer variants have demonstrated competitive results (e.g., [12, 20, 21, 22, 23, 24, 25]). These models confirm that global context is beneficial, but also highlight the need for architectures that are explicitly adapted to multi-band data, rather than merely repurposed from RGB settings.

AMBER is introduced to address this gap. Starting from SegFormer [1], AMBER introduces three key modifications: (1) *3D convolutions* to jointly process spectral and spatial information; (2) *custom kernel sizes and strides* to better preserve spatial resolution; and (3) a *Funnelizer* layer that reduces spectral dimensionality *within the network* to produce a 2D segmentation mask. This design streamlines the processing chain and preserves spectral information end-to-end, delegating feature extraction to the model rather than to preprocessing heuristics.

The evaluation in this chapter considers four datasets: Salinas, Indian Pines, Pavia University, and PRISMA. The first three are standard airborne benchmarks; PRISMA represents a realistic spaceborne scenario with different noise characteristics and acquisition conditions. This combination allows assessing not only peak performance on benchmarks, but also robustness across sensors and platforms.

2.2 Related Works

Deep learning methods have substantially advanced hyperspectral pixel classification and segmentation by learning discriminative features hierarchically. CNN-based architectures remain a cornerstone because they can extract local spectral–spatial cues efficiently and scale to large images ([26]). A number of CNN variants have been proposed specifically for hyperspectral imagery. Zhong et al. (2017) [27] introduced a spectral–spatial residual network based on residual blocks. Paoletti et al. (2018) [28] proposed a pyramidal residual network to diversify higher-level features across layers. Zhu et al. (2020) [29] introduced a residual spectral–spatial attention network (RSSAN) that cascades spectral and spatial attention modules. Chhapariya et al. (2022) [30] designed a multi-level 3D CNN with varied kernels to extract multi-scale features. Despite their effectiveness, these CNN approaches can struggle to model global relationships and long-range dependencies due to their inherently local operations.

Transformers, originally developed for sequence modeling in NLP, address long-range dependency modeling via self-attention [8]. Vision Transformers [9] extend this paradigm to images by tokenizing an image into patches. In semantic segmentation, SegFormer [1] combines a hierarchical transformer encoder with an efficient decoder, enabling strong segmentation performance while remaining computationally tractable. Multiple works have adapted transformer ideas to hyperspectral analysis, including SpectralFormer [12] and subsequent approaches emphasizing joint global–local modeling [29, 31, 32, 33]. These studies motivate architectures that explicitly treat hyperspectral data as volumetric (spectral–spatial) objects, rather than as RGB-like inputs.

AMBER follows this motivation: it retains the strengths of SegFormer while introducing volumetric processing and an explicit spectral reduction mechanism within the network, targeting practical multi-band segmentation.

2.3 Methodology

This section presents AMBER, an adaptation of SegFormer [1] for multi-band semantic segmentation. AMBER retains the overall structure of SegFormer: (i) a hierarchical transformer encoder that extracts multi-resolution features, and (ii) a lightweight All-MLP decoder that fuses those features into a segmentation prediction. The key difference is that AMBER is designed to process three-dimensional data cubes ($D \times H \times W$) with explicit spectral–spatial coupling, while ultimately outputting a standard two-dimensional mask.

A schematic overview of AMBER is shown in Fig. 2.1.

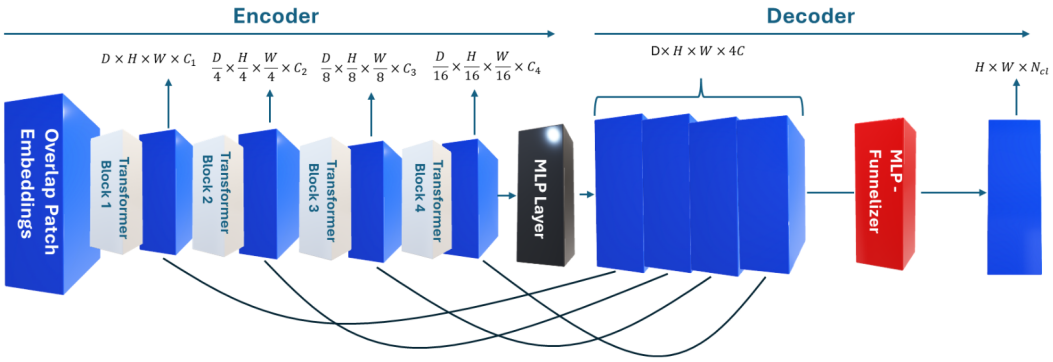

Figure 2.1: Overview of AMBER. A hierarchical Transformer encoder extracts multi-scale spectral–spatial representations from a $D \times H \times W$ input. A lightweight All-MLP decoder fuses multi-level features. The Funnelizer reduces the spectral dimension to enable the prediction of a $H \times W \times N_{\text{cls}}$ segmentation mask.

Given an input cube of size $D \times H \times W$, AMBER first partitions the data into overlapping volumetric patches using a 3D convolution. The resulting token-like representations are processed by the hierarchical transformer encoder, producing multi-level features at resolutions $\{1/4, 1/8, 1/16\}$ of the input. These features are fused by the decoder and mapped to a dense prediction. In contrast to pipelines that reduce D prior to learning, AMBER preserves the full spectral axis at input and performs spectral compression internally via the Funnelizer. To provide a detailed explanation of our model architecture, we will follow almost the same steps as in the original paper[1].

2.3.1 Hierarchical Transformer Encoder

AMBER constructs a series of Mix Transformer encoders (MiT) adapted to volumetric (spectral–spatial) data.

Hierarchical Feature Representation.

Given a $D \times H \times W$ input, patch merging produces hierarchical feature maps F_i at resolution $\frac{D}{2^{i+1}} \times \frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}} \times C_i$, where $i \in \{1, 2, 3, 4\}$ and channel width increases with stage.

Overlapped Patch Merging.

To avoid explicit positional encodings, AMBER employs overlapping patch merging. Let K denote the 3D kernel size (patch size), S the stride, and P the padding. In our experiments, we set $K = 3$ and consider $S \in \{1, 2\}$ with appropriate padding, using small kernels to preserve detail without excessive parameter growth. Notably, $S = 1$ preserves the spatial dimensions H and W, avoiding the early $1/4$ reduction typical of many segmentation backbones.

Efficient Self-Attention.

For multi-head self-attention, each head uses queries, keys, and values with size $N \times C$, where $N = DHW$ is the sequence length. Self-attention is computed as

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_{head}}}\right)V. \tag{2.1}$$

Following [1], the complexity is reduced from $O(N^2)$ to $O(N^2/R)$ by spatial-reduction attention. We set R to $[64, 16, 4, 1]$ from stage-1 to stage-4.

Mix-FFN.

AMBER employs Mix-FFN [1], which incorporates a $3 \times 3 \times 3$ convolution inside the feed-forward block:

$$x_{out} = \text{MLP}(\text{GELU}(\text{Conv}_{3\times3\times3}(\text{MLP}(x_{in})))) + x_{in}, \tag{2.2}$$

where x_{in} denotes the output of the attention block.

2.3.2 Lightweight All-MLP Decoder

The All-MLP decoder follows the SegFormer design: multi-level encoder features are projected via MLP layers to a common channel dimension, upsampled to a common resolution, concatenated, and fused by an additional MLP.

A key AMBER-specific component is the *Funnelizer*: a convolution with kernel size $(D \times 1 \times 1)$ that collapses the spectral dimension, transforming the representation into a 2D feature map while retaining spatial resolution. A final 1×1 2D convolution produces the segmentation mask with size $H \times W \times N_{cls}$.

2.4 Experiments and Analysis

2.4.1 Data and experimental environment description

To evaluate AMBER, experiments are conducted on three widely used public airborne hyperspectral benchmarks—Salinas, Indian Pines, and Pavia University [34]—as well as on a PRISMA spaceborne scene. Together, these datasets allow assessing both controlled benchmark performance and real satellite robustness.

Salinas, Indian Pines and Pavia University datasets

The Salinas, Indian Pines, and Pavia University datasets consist of hyperspectral images with water absorption bands removed.

The Salinas dataset was acquired using AVIRIS over the Salinas Valley (California) and has high spatial resolution (3.7 m). After removing absorption bands, it contains 204 spectral bands and size 512×217.

The Indian Pines dataset (AVIRIS) covers a mixed agricultural/forested region in Northwestern Indiana (USA). It has size 145×145, spectral range 0.40–2.50 μm, 200 bands, and 20 m spatial resolution.

The Pavia University dataset (ROSIS) was acquired over Pavia (Italy), with size 610×340, spectral range 0.43–0.86 μm, 103 bands, and 1.3 m resolution.

Salinas and Indian Pines include 16 classes; Pavia University includes 9 classes. False-color images and ground-truth maps are shown in Figs. 2.2, 2.3, and 2.4. Class statistics are reported in Tables 2.1, 2.2, and 2.3.

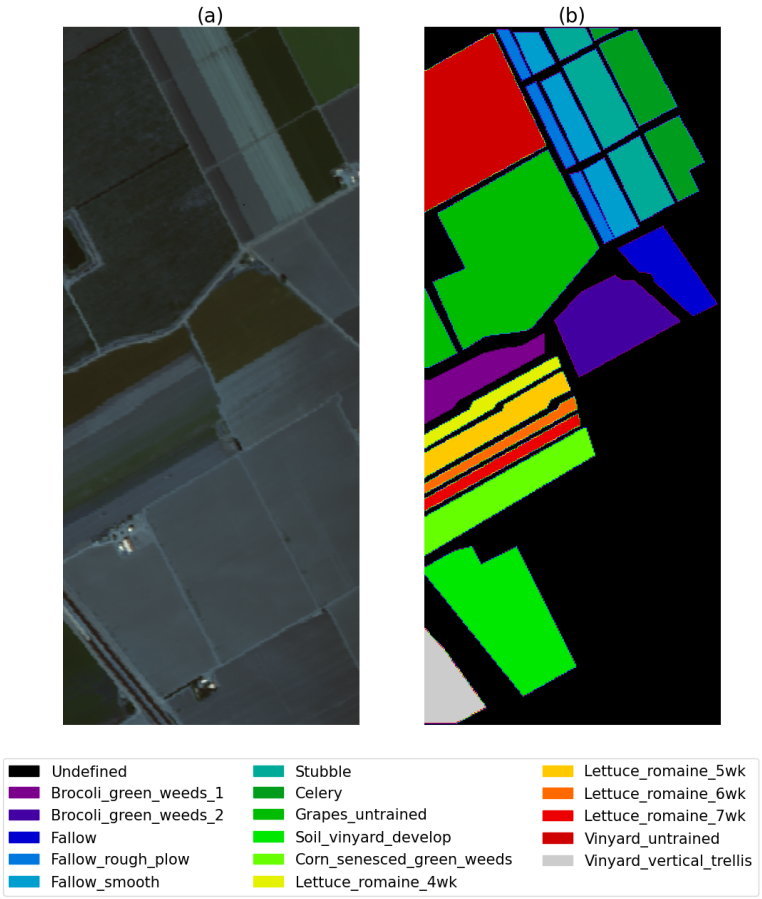

Figure 2.2: Salinas dataset. (a) False-color map; (b) Ground-truth map.

Table 2.1: Groundtruth classes for the Salinas scene and their respective samples number.

#	Class	Samples
1	Brocoli green weeds 1	2009
2	Brocoli green weeds 2	3726
3	Fallow	1976
4	Fallow rough plow	1394
5	Fallow smooth	2678
6	Stubble	3959
7	Celery	3579
8	Grapes untrained	11271
9	Soil vinyard develop	6203
10	Corn senesced green weeds	3278
11	Lettuce romaine 4wk	1068
12	Lettuce romaine 5wk	1927
13	Lettuce romaine 6wk	916
14	Lettuce romaine 7wk	1070
15	Vinyard untrained	7268
16	Vinyard vertical trellis	1807

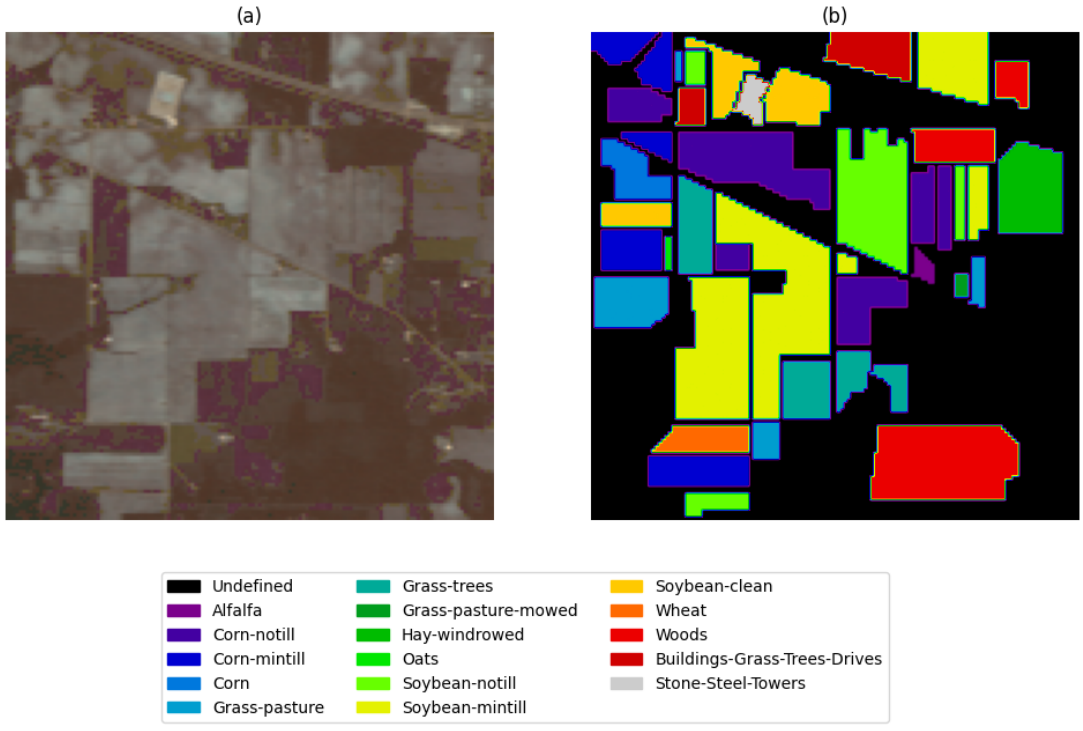

Figure 2.3: Indian Pines dataset. (a) False-color map; (b) Ground-truth map.

Table 2.2: Groundtruth classes for the Indian Pines scene and their respective samples number.

#	**Class**	**Samples**
1	Alfalfa	46
2	Corn-notill	1428
3	Corn-mintill	830
4	Corn	237
5	Grass-pasture	483
6	Grass-trees	730
7	Grass-pasture-mowed	28
8	Hay-windrowed	478
9	Oats	20
10	Soybean-notill	972
11	Soybean-mintill	2455
12	Soybean-clean	593
13	Wheat	205
14	Woods	1265
15	Buildings-Grass-Trees-Drives	386
16	Stone-Steel-Towers	93

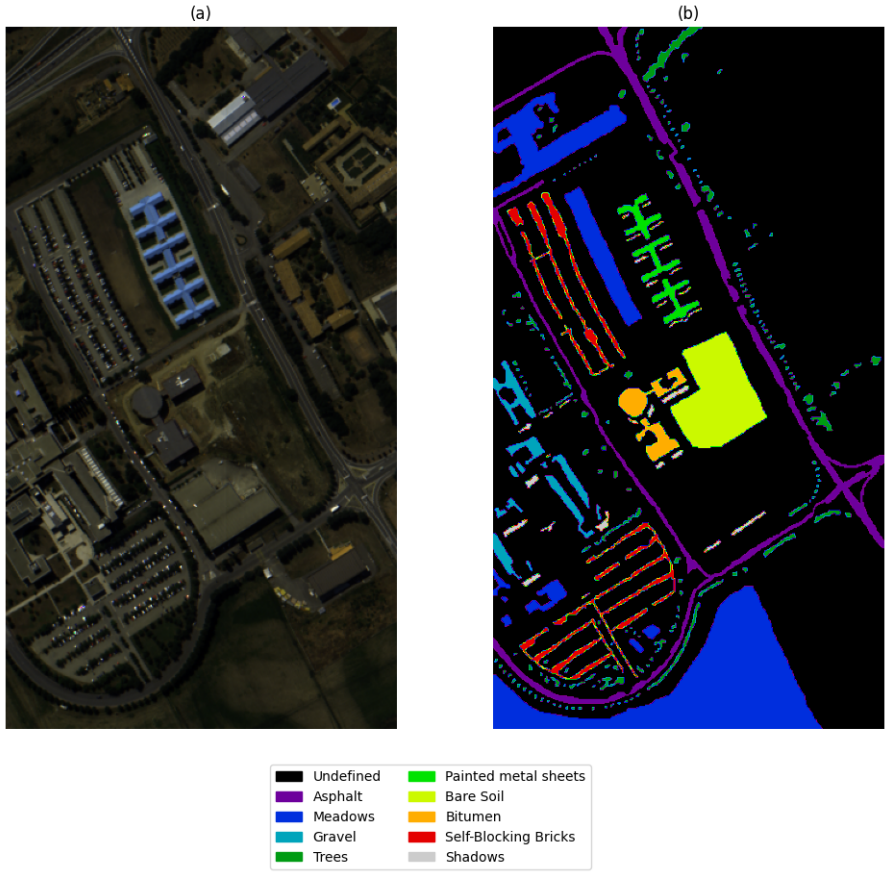

Figure 2.4: Pavia University dataset. (a) False-color map; (b) Ground-truth map.

Table 2.3: Groundtruth classes for the Pavia University scene and their respective samples number.

#	Class	Samples
1	Asphalt	6631
2	Meadows	18649
3	Gravel	2099
4	Trees	3064
5	Painted metal sheets	1345
6	Bare Soil	5029
7	Bitumen	1330
8	Self-Blocking Bricks	3682
9	Shadows	947

PRISMA dataset

PRISMA is a spaceborne hyperspectral sensor developed by ASI, acquiring 231 spectral bands between 400 and 2500 nm at 30 m spatial resolution. In this chapter, AMBER is evaluated on a PRISMA Level 2C Bottom-Of-Atmosphere (BOA) reflectance scene over the Quadrilátero Ferrífero region (Minas Gerais, Brazil), acquired on May 21, 2022 at 13:05:52 UTC.

Level 2C data were accessed from http://prisma.asi.it/. VNIR and SWIR cubes were preprocessed with ENVI 5.6.1, including correction of geolocation errors via the Refine RPCs (HDF-EOS5) task within the PRISMA Toolkit. Due to atmospheric absorption, only 193 out of 231 bands were retained.

Ground truth was derived by combining the ESA WorldCover-Map (https://esa-worldcover.org/en) with Global-scale mining polygons (Version 1) [35]. To reduce redundancy, the following classes were considered: vegetation, water, mining areas, and built-up areas. Since vegetation is strongly overrepresented (Table 2.4), a random subset of 700,000 vegetation pixels was set to the undefined class to balance training and evaluation. The resulting false-color image and ground truth are shown in Fig. 2.5.

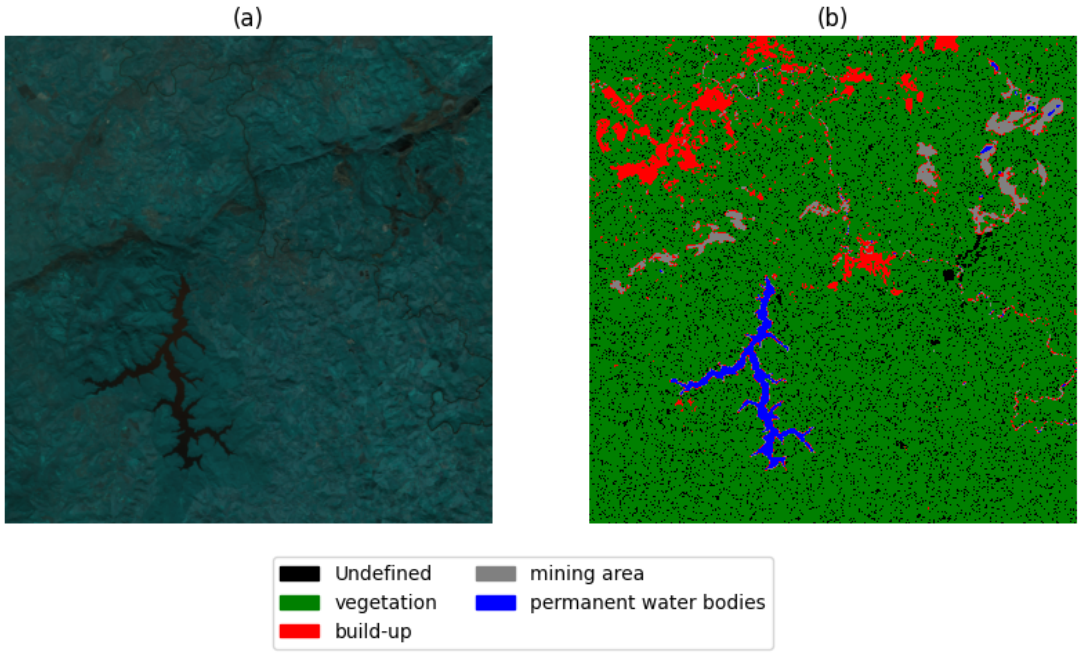

Figure 2.5: PRISMA dataset. (a) False-color map; (b) Ground-truth map.

Table 2.4: Groundtruth classes for the PRISMA scene and their respective samples number.

#	Class	Samples
1	Vegetation	793572
2	Mining Area	36227
3	Permanent Water bodies	19638
4	Build-up	15463

2.4.2 Implementation details

Training and inference were executed on the IBiSCo Data Center (https://ibiscohpc-wiki.scope.unina.it/). AMBER was trained using four Tesla V100 GPUs (32 GB each). Training time depended on dataset size and complexity: ∼6 hours for Salinas, 5 hours for Indian Pines, and ∼7 hours for Pavia University and PRISMA. Unlike the original SegFormer [1], AMBER was trained from scratch, without encoder pretraining and with random decoder initialization.

During training, random flipping was applied to patches. The undefined class was excluded from loss computation. The crop size was set to $D \times 32 \times 32$ for all datasets after empirical evaluation: larger patches reduced fine-detail accuracy, while smaller patches reduced contextual consistency.

An SGD optimizer was used in all experiments. Batch size was set to 6 (Salinas), 3 (Indian Pines), 10 (Pavia University), and 4 (PRISMA), reflecting GPU memory constraints and dataset characteristics. The learning rate was fixed at 0.01. AMBER was trained for 30 epochs on Salinas and Pavia University, and 50 epochs on Indian Pines and PRISMA, as these values achieved stable

convergence while limiting overfitting. Categorical Cross Entropy was used as the loss function.

2.4.3 Evaluation metrics

Segmentation performance is evaluated using Overall Accuracy (OA), Kappa Coefficient (Kappa), and Average Accuracy (AA). OA is defined as

$$\mathrm{OA} = \frac{\sum_{i=1}^{n} h_{ii}}{N} \times 100/\% \tag{2.3}$$

where h_{ii} are the diagonal entries of the confusion matrix, N is the total number of samples, and n is the number of classes.

The Kappa coefficient is computed as

$$\mathrm{Kappa} = \frac{N \sum_k h_{kk} - \sum_k \left(\sum_j h_{kj} \sum_i h_{ik} \right)}{N^2 - \sum_k \left(\sum_j h_{kj} \sum_i h_{ik} \right)}. \tag{2.4}$$

Average Accuracy is defined as

$$\mathrm{AA} = \frac{\sum_{i=1}^{C} \left(M_{ii} / \sum_{j=1}^{C} M_{ij} \right)}{C} \times 100/\%, \tag{2.5}$$

where C is the number of classes and M_{ij} counts samples of true class i predicted as class j.

2.4.4 Data Preprocessing

Datasets were constructed using random patches, rather than using explicit balancing strategies (e.g., [36]). This choice assesses AMBER's ability to learn under realistic class imbalance and standardizes the number of patches.

To mitigate information leakage due to overlap between training and test pixels [37, 38], only a small fraction of patches was used for training, following common practice (e.g., [12, 39, 40]). For Salinas, Indian Pines, and Pavia University, 20% of patches were used for training and 80% for testing. For PRISMA, 10% of patches were used for training and 90% for testing.

Finally, while dimensionality reduction (e.g., PCA, SVD, Gabor filters [39, 17, 18]) is frequently applied to reduce redundancy along the spectral axis, AMBER is trained on full-spectrum patches. This design explicitly tests the network's ability to learn an effective spectral representation internally.

2.5 Results

2.5.1 Classification Results

AMBER is compared against four representative baselines: SVD/Unet [18], HSI-CNN [41], 3D-CNN [42], and contextual deep CNN [43]. Among these, SVD/Unet, contextual deep CNN, and AMBER perform pixel-wise prediction. HSI-CNN and 3D-CNN assign labels based on the central pixel of each patch,

which can inflate reported metrics relative to full pixel-wise segmentation. The comparisons below should be interpreted with this distinction in mind.

Each baseline was trained using the configurations described in its corresponding reference [18, 41, 42, 43]. Patch sizes differ across methods: AMBER uses $D \times 32 \times 32$; SVD/Unet uses $3 \times 32 \times 32$; HSI-CNN uses $D \times 3 \times 3$; 3D-CNN uses $D \times 7 \times 7$; contextual deep CNN uses $D \times 5 \times 5$. All methods were evaluated with five independent runs using five-fold Monte Carlo cross-validation [44]. Tables 2.5–2.8 report mean per-class accuracy and aggregate metrics (OA, Kappa, AA), including standard deviations ($\pm$).

Table 2.5: Classification accuracy of Salinas Dataset.

Class	SVD/Unet [18]	HSI-CNN [41]	3D-CNN [42]	Cont Deep CNN [43]	**AMBER**
1	0.00	99.10	99.49	97.14	**99.95**
2	99.99	97.29	99.99	99.82	**100**
3	34.24	84.52	**99.97**	99.31	99.65
4	91.47	96.91	98.60	99.52	**99.68**
5	99.55	98.49	98.09	94.17	**99.65**
6	99.96	99.98	**100**	99.99	99.99
7	65.26	99.69	99.96	**99.99**	99.98
8	97.96	86.49	96.09	90.95	**99.46**
9	84.33	99.12	99.85	**99.95**	96.99
10	79.15	91.84	96.76	95.63	**99.74**
11	31.48	96.92	98.91	95.98	**99.63**
12	32.15	**100**	99.37	99.92	99.93
13	99.57	98.90	95.29	**99.67**	99.57
14	99.05	95.56	**99.83**	95.52	99.16
15	0.013	56.94	77.89	78.32	**97.60**
16	64.54	96.34	99.11	99.16	**99.97**
OA/%	72.28 ± 2.94	90.20 ± 0.65	96.08 ± 0.36	84.42 ± 2.02	$\mathbf{99.73 \pm 0.053}$
Kappa $\times$ 100	67.87 ± 3.42	88.99 ± 0.029	95.59 ± 0.42	81.05 ± 2.56	$\mathbf{99.69 \pm 0.061}$
AA/%	67.42 ± 6.80	93.63 ± 0.24	97.46 ± 0.19	71.40 ± 9.02	$\mathbf{99.75 \pm 0.021}$

Table 2.6: Classification accuracy of Indian Pines Dataset.

Class	SVD/Unet [18]	HSI-CNN [41]	3D-CNN [42]	Cont Deep CNN [43]	**AMBER**
1	24.56	18.28	94.01	83.79	**99.27**
2	**99.70**	30.78	93.58	77.10	99.52
3	99.36	5.29	83.44	22.10	**99.42**
4	57.24	7.66	77.85	48.69	**99.49**
5	96.27	51.96	94.94	71.89	**99.16**
6	99.96	96.09	99.69	99.25	**99.98**
7	0.00	0.00	94.25	86.79	**96.02**
8	91.98	95.30	99.63	93.49	**99.97**
9	0.0029	1.56	93.65	34.85	**96.95**
10	75.17	52.75	88.94	87.74	**99.35**
11	99.93	89.80	95.83	89.77	**99.86**
12	94.022	21.35	88.38	49.17	**99.16**
13	73.19	97.36	**100**	98.46	99.95
14	99.97	95.49	99.83	99.19	**99.97**
15	72.0055	32.03	96.22	60.46	**97.42**
16	73.58	79.07	98.94	99.90	**99.54**
OA/%	94.18 ± 4.22	68.15 ± 7.36	95.01 ± 0.79	84.42 ± 2.02	$\mathbf{99.74 \pm 0.097}$
Kappa × 100	92.77 ± 5.40	60.55 ± 10.029	94.03 ± 0.94	81.05 ± 2.56	$\mathbf{99.68 \pm 0.12}$
AA/%	70.99 ± 4.22	47.059 ± 9.28	93.25 ± 1.10	71.40 ± 9.02	$\mathbf{99.18 \pm 0.34}$

Table 2.7: Classification accuracy of Pavia University Dataset.

Class	SVD/Unet [18]	HSI-CNN [41]	3D-CNN [42]	Cont Deep CNN [43]	**AMBER**
1	89.19	96.79	98.94	97.48	**99.79**
2	96.47	99.09	99.37	99.51	**100**
3	0.17	74.44	90.70	82.06	**99.27**
4	95.18	94.58	98.87	97.21	**99.60**
5	94.15	99.89	**100**	99.98	99.99
6	91.50	98.14	97.81	95.05	**99.99**
7	44.09	96.90	97.06	97.12	**99.88**
8	**99.91**	96.34	98.29	96.05	99.85
9	99.75	99.70	99.86	99.96	**99.98**
OA/%	91.12 ± 8.2	96.96 ± 0.77	98.51 ± 0.31	97.40 ± 0.84	$\mathbf{99.94 \pm 0.0098}$
Kappa × 100	86.72 ± 11.82	96.15 ± 0.98	98.12 ± 0.39	96.49 ± 1.14	$\mathbf{99.91 \pm 0.015}$
AA/%	70.99 ± 4.22	95.11 ± 1.66	97.87 ± 0.40	96.04 ± 0.64	$\mathbf{99.82 \pm 0.022}$

Table 2.8: Classification accuracy of PRISMA Dataset.

Class	SVD/Unet [18]	HSI-CNN [41]	3D-CNN [42]	Cont Deep CNN [43]	**AMBER**
1	90.57	**95.34**	93.59	77.97	81.70
2	94.99	68.97	81.75	89.52	**96.60**
3	**92.65**	74.32	78.69	87.80	92.21
4	**96.71**	80.35	83.71	90.54	93.18
OA/%	$\mathbf{93.52 \pm 0.090}$	89.88 ± 3.57	88.11 ± 0.21	88.52 ± 0.26	90.90 ± 0.84
Kappa × 100	$\mathbf{91.06 \pm 0.12}$	74.62 ± 0.99	80.46 ± 0.51	84.06 ± 0.34	87.41 ± 1.48
AA/%	$\mathbf{93.73 \pm 0.10}$	79.74 ± 0.69	84.44 ± 0.68	87.66 ± 0.54	90.92 ± 1.32

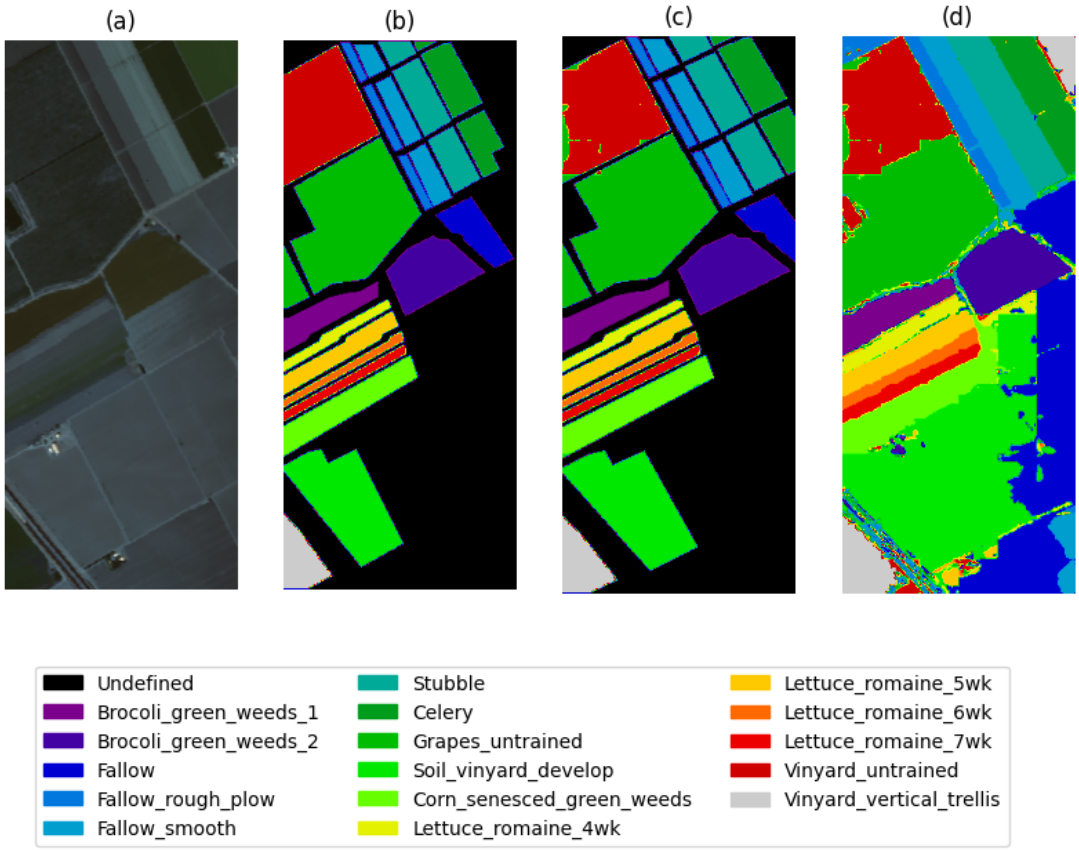

Figure 2.6: Salinas prediction. (a) False-color map. (b) Ground-truth map. (c) AMBER prediction with the undefined mask. (d) AMBER prediction without the undefined mask. Reproduced by the author based on results published in [2].

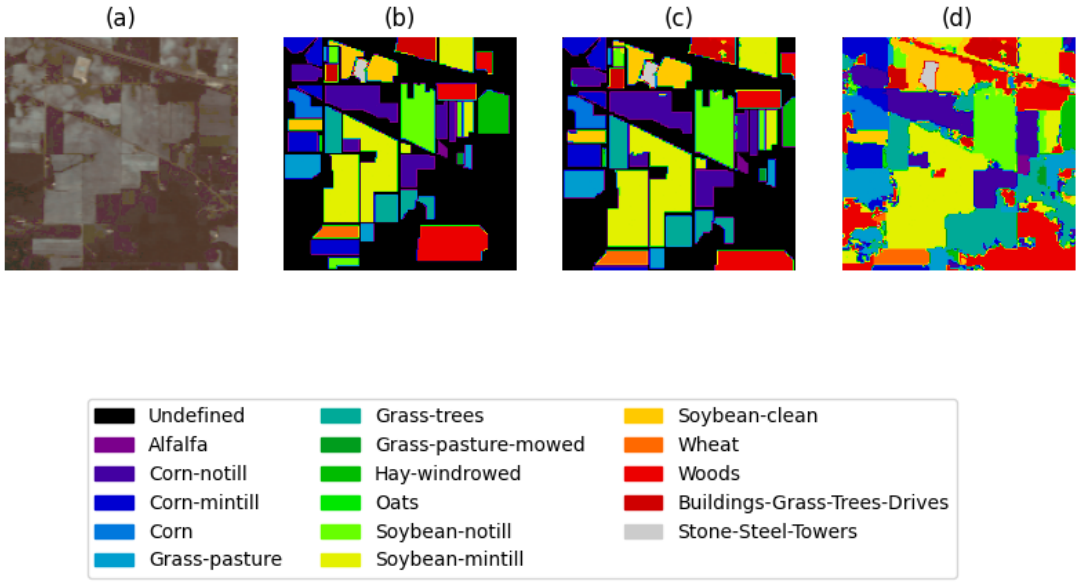

Figure 2.7: Indian Pines prediction. (a) False-color map. (b) Ground-truth map. (c) AMBER prediction with the undefined mask. (d) AMBER prediction without the undefined mask. Reproduced by the author based on results published in [2].

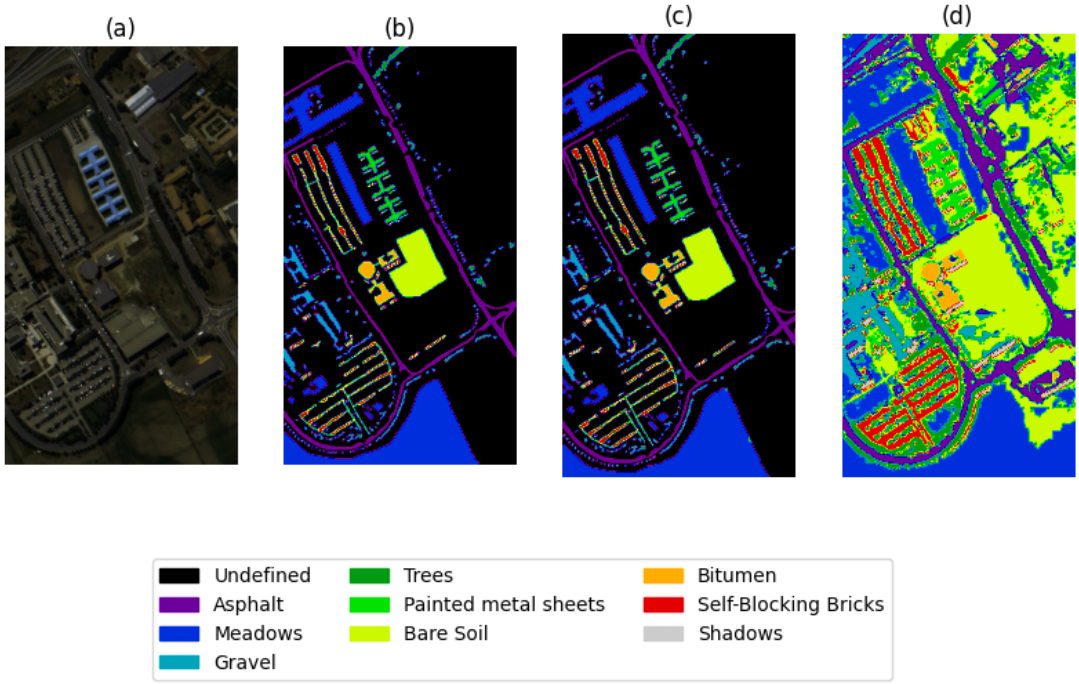

Figure 2.8: Pavia University prediction. (a) False-color map. (b) Ground-truth map. (c) AMBER prediction with the undefined mask. (d) AMBER prediction without the undefined mask. Reproduced by the author based on results published in [2].

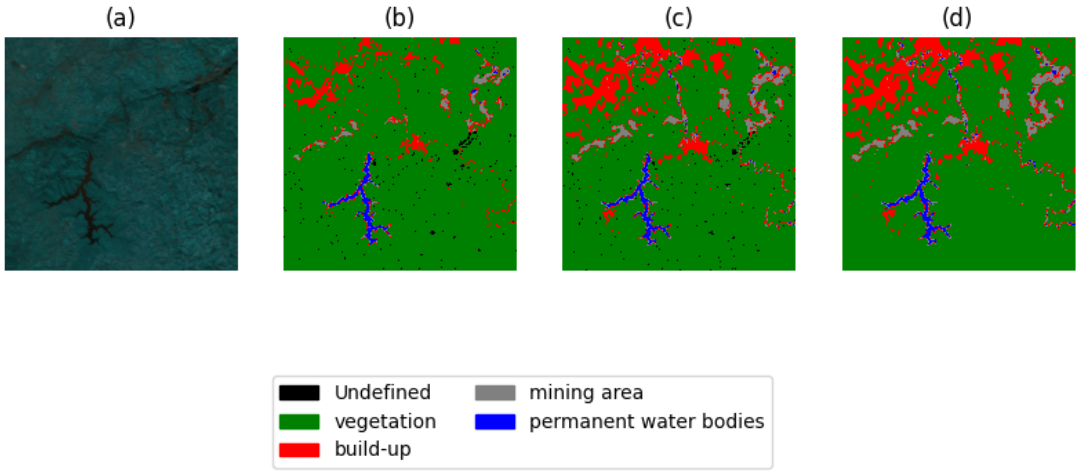

Figure 2.9: PRISMA prediction. (a) False-color map. (b) Ground-truth map. (c) AMBER prediction with the undefined mask. (d) AMBER prediction without the undefined mask. Reproduced by the author based on results published in [2].

The results in Tables 2.5–2.7 indicate that AMBER consistently achieves the highest OA, Kappa, and AA on the Salinas, Indian Pines, and Pavia University benchmarks, with very small variability across Monte Carlo folds. In particular, AMBER performs strongly on minority classes and on classes with subtle spectral signatures, suggesting that the joint spectral–spatial modeling and the preservation of full-spectrum input are beneficial in practice.
On PRISMA (Table 2.8), AMBER achieves competitive performance and improves upon several baselines, while remaining slightly below SVD/Unet in OA and Kappa. This gap is informative for future work: PRISMA data introduce additional sources of variability (spaceborne acquisition, different noise statistics, and heterogeneous ground truth derived from external products), and further improvements may require targeted regularization or domain adaptation strategies.
Qualitative predictions for each dataset are shown in Figs. 2.6, 2.7, 2.8, and 2.9. For Salinas, Indian Pines, and Pavia University, AMBER's predictions closely match the ground truth visually. The paired panels with and without the undefined mask illustrate how undefined pixels (excluded from loss computation) influence visual assessment without affecting the quantitative metrics.
Figure 2.10 illustrates a representative example of AMBER's segmentation behavior on a complex asphalt structure in the Pavia University dataset, including regions labeled as undefined in the ground truth.

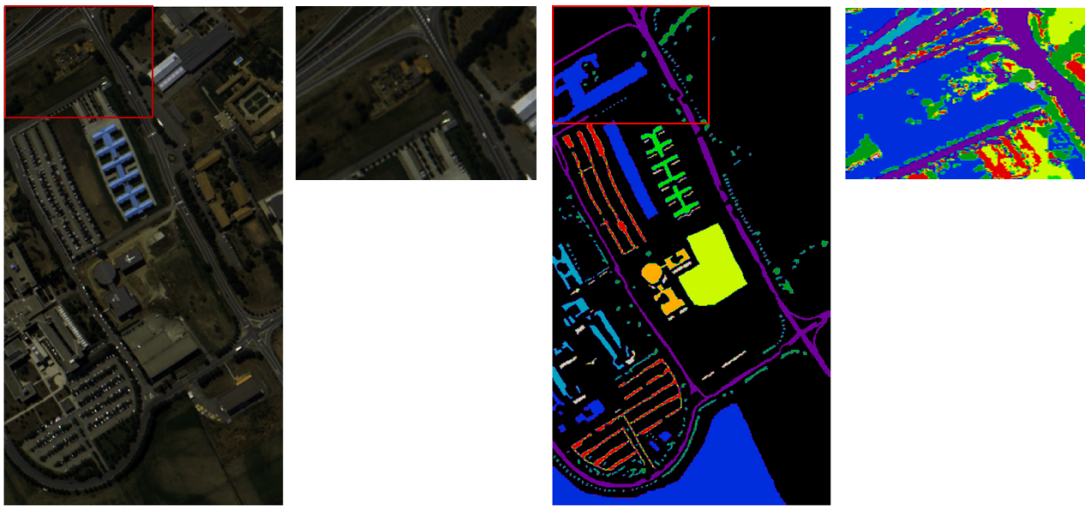

Figure 2.10: AMBER prediction for undefined pixels: example of complex asphalt-structure segmentation in the Pavia University dataset. Reproduced by the author based on results published in [2].

2.6 Conclusion

This chapter introduced AMBER, a transformer-based architecture for multi-band semantic segmentation derived from SegFormer [1]. AMBER is designed to operate directly on hyperspectral cubes through three-dimensional convolutions, while preserving spatial resolution and performing spectral dimensionality reduction internally via the Funnelizer. The experimental evaluation on Salinas, Indian Pines, Pavia University, and PRISMA demonstrates that AMBER provides strong and consistent performance on standard benchmarks and competitive results on real satellite data.

From a methodological standpoint, the results support two conclusions. First, explicitly modeling spectral–spatial structure in a volumetric representation can improve segmentation robustness, especially under class imbalance. Second, avoiding mandatory spectral dimensionality reduction can preserve discriminative information and simplify preprocessing pipelines, shifting feature selection to the learning model.

At the same time, the PRISMA results indicate that spaceborne hyperspectral segmentation remains challenging and that performance may depend on domain-specific factors such as noise, calibration, and ground-truth uncertainty. Future work will therefore focus on improving robustness under strong spectral variability through regularization, adaptive attention mechanisms, and strategies that better account for label noise.

Finally, although AMBER is developed here for hyperspectral remote sensing, the architecture targets a broader class of multi-dimensional data cubes. The same design principles can be applied to other domains where volumetric data and multi-band observations are central, including astronomy (e.g., multi-frequency radio imaging products) and medical imaging (e.g., MRI/CT volumes).

Chapter 3

Less is More: AMBER-AFNO - a New Benchmark for Lightweight 3D Medical Image[3]

This chapter extends the methodological developments introduced in Chapter 2 to the domain of three-dimensional medical image segmentation. In particular, it investigates the transfer of the *AMBER* architecture [2], originally conceived for multiband remote sensing imagery, to volumetric medical data such as MRI and CT scans. The central idea explored in this chapter is that global contextual modeling in 3D medical images can be achieved without relying on computationally expensive self-attention mechanisms.

To this end, we introduce *AMBER-AFNO*, a variant of AMBER in which the multi-head self-attention blocks are replaced by Adaptive Fourier Neural Operators (AFNO). Rather than computing explicit token-to-token interactions, AFNO performs global feature mixing in the frequency domain, offering a substantially more parameter-efficient alternative [4, 3]. This architectural choice results in a reduction of more than 80% in the number of trainable parameters compared to transformer-heavy baselines such as UNETR++, while preserving comparable computational complexity in terms of FLOPs.

The proposed architecture is evaluated on three widely used 3D medical imaging benchmarks—ACDC, Synapse, and BraTS—using standard segmentation metrics including the Dice Similarity Coefficient (DSC) and the 95th percentile Hausdorff Distance (HD95). Beyond raw accuracy, this chapter places particular emphasis on efficiency, robustness, and deployability, which are critical considerations in real-world clinical environments. The results demonstrate that AMBER-AFNO achieves competitive or superior segmentation performance while significantly reducing memory usage, inference latency, and architectural complexity.

This chapter is based on the preprint **Less is More: AMBER-AFNO – A New Benchmark for Lightweight 3D Medical Image Segmentation**[3], currently available on arXiv and under review for publication in *Expert Systems with Applications* (Elsevier). The candidate is the first author of this work. The material has been expanded and further developed for inclusion in this thesis.

3.1 Introduction

According to the World Health Statistics 2024 report [45], heart disease was the leading cause of death globally in 2021, accounting for approximately 9.1 million deaths, with kidney disease, lung cancer, lower respiratory infections, and stroke also ranking among the top ten global causes of mortality.
In all cases, an early diagnosis is crucial to reduce mortality rates as well as to optimize the choice among different treatment possibilities such as different therapies or even surgery. This early diagnosis heavily depends upon on the proper exploitation of advanced technologies such as, for instance, endoscopy, MRI, and CT Scans which allow the detection of tissue abnormalities [46]. MRI and CT Scans generate reports capturing slices at various depths, resulting in three-dimensional images (hereafter data cubes) that capture more information than the 2D images (b/W or RGB) produced by traditional X-ray machines [47]. As we shall describe in more detail in the next section, most healthcare datacubes represent a 3D volume of the anatomy. Since a single 2D slice alone often fails to capture the full anatomical context, relying solely on such data can compromise model performance [48]. It is therefore no surprise that, in recent years, considerable effort has been devoted to developing advanced methods, often based on artificial intelligence, that can fully exploit the richer information embedded in these volumetric datasets.
In medical image segmentation, the U-Net [49] has been widely used as a popular deep network architecture [50, 51] thanks to its encoder-decoder structure, which learns both local and global contextual features. Furthermore, the encoder-decoder structure with skip connections enables U-Net to obtain the lost spatial information during the down-sampling operation in the encoder. Therefore, the U-Net architecture has been commonly adopted for pixel-wise segmentation tasks. Based on U-Net, various deep models [52, 53, 54, 55, 56, 57, 58] have been proposed to overcome some intrinsic weaknesses of U-Net, such as fixed-size receptive fields and the inability to learn multi-scale features. All these variants are still built upon the encoder-decoder structure but introduce different modules to improve the capability of the U-Net architecture, such as adaptive fusion modules, dual-branch architecture, attention modules, and multi-resolution learning pathways.
Even though, CNN based architectures, including U-Net and its variants, have made tremendous progress in modeling spatial information, they are still bounded by the size of the kernel and stride to learn long-range dependencies. Several architectural improvements have been introduced to better model the global context by learning with the gradient information (e.g., LMISA [59]). However, they are still not fully able to circumvent these architectural constraints, especially when dealing with 3D data such as CT and MRI, where inter-slice spatial dependencies are substantial.
To address these limitations, the transformer architectures [8] are designed for the vision tasks using Vision Transformers (ViT) [9]. Different from CNNs, transformers are based on self-attention instead of convolution, which is able to model the long-range dependencies over the whole images. Based on the vision transformer, a series of transformer-based models for the medical image segmentation have been proposed, such as RT-UNet [60], SWTRU [61], LMIS [62] and SDV-

TUNet [63]. These models often adopt the transformer blocks in the U-Net like architectures to simultaneously exploit the spatial resolution and global context modeling. However, they generally involve deep stacks of encoders (usually 8 to 12 transformer layers) and are dominated by the computationally intensive attention module, which makes them less memory and inference efficient than those lightweight CNN based models.
SegFormer [19] is proposed to solve two main problems in the semantic segmentation. Firstly, the SegFormer has a better model structure, including a lightweight mix transformer block and a simple MLP-based decoder. This replaces the original position encoding with a feed-forward network (FFN), and makes the attention mechanism simplified, which significantly reduces the computation cost. Secondly, the encoder of the SegFormer consists of only 4 hierarchical mix transformer blocks, making the encoder structure far smaller than the traditional architectures.
On top of that, AMBER [2] has adapted SegFormer to the case of multi-band image segmentation, including also 3D convolutional layers, variable kernel size, and a Funnelizer to tackle the complex spatial-spectral information.

In this chapter, we propose AMBER-AFNO (AFNO: Adaptive Fourier Neural Operators, see later for details), a version of AMBER specifically adapted for 3D medical datacube segmentation. This methodological transfer from remote sensing to medical imaging is based on an adaptation of AMBER, which was originally developed for multiband images (e.g., hyperspectral images), to 3D medical images.
In AMBER-AFNO, we replace the conventional multi-head self-attention mechanism with Adaptive Fourier Neural Operators [4], enabling global context modeling through frequency-domain mixing rather than attention. This results in a substantial reduction in architectural complexity, cutting the number of trainable parameters by over 80% compared to UNETR++, while preserving a number of FLOPs comparable to other state-of-the-art models.
We trained and validated our AMBER-AFNO approach by conducting comprehensive experiments on three benchmarks: ACDC [64], Synapse [65] and BraTS [66] datasets. Both the evaluation metrics, the Dice score (DSC) and Hausdorff distance (HD) score, along with the model architecture in terms of total number of trainable parameters, show the efficiency of AMBER-AFNO compared to the existing methods in the literature.

3.2 Related Work

Three-dimensional medical-image segmentation has evolved rapidly since the success of the encoder–decoder paradigm initiated by **U–Net**[2]. Current methods can be grouped into three non-exclusive research lines that progressively trade architectural simplicity for richer multi-scale context awareness:

1. **Single-branch encoder–decoder networks**
2. **Multi-branch encoder networks**
3. **Efficient Attention Methods**

Single-branch models remain attractive for their efficiency in both training and inference. *V-Net* introduced residual 3-D convolutions and Dice loss to tackle volumetric data directly [67]. The *DeepLab* family incorporates Atrous convolutions and the Atrous Spatial Pyramid Pooling (ASPP) module to enlarge the receptive field without increasing the number of parameters [68]. *nnU-Net* argues against over-engineered designs, proposing a self-configuring pipeline that automatically adapts its depth, patch size and learning schedule to each dataset and has become the de-facto baseline in many challenges [69].

Multi-branch encoders explicitly separate semantic scales to improve both global context perception and boundary precision. *DS-TransUNet* employs dual-scale ViT encoders combined with a Token-Interaction Fusion module, achieving superior Dice on abdominal CT [70]. *DHR-Net* couples a multi-scale branch with a detail-enhancement branch selected by reinforcement learning to boost small-structure recall [71]. Further examples include *MILU-Net*, which mitigates information loss with dual up-sampling paths [72], and *BSC-Net*, a dual-resolution lightweight design that matches U-Net++ accuracy with roughly half the parameters [73]. Despite their merits, many two-branch schemes are hard to migrate across datasets and remain computationally heavy.

Efficient Attention Methods aim to capture long-range dependencies that CNNs struggle with in high-resolution volumes. *UNETR* re-implements U-Net with a ViT encoder feeding skip tokens to a CNN decoder [74]. *Swin-UNETR* inherits the shifted-window self-attention of Swin-Transformer and extracts features at five resolutions, outperforming UNETR on the MSD organ sets [75]. *nnFormer* interleaves local convolution and global self-attention while introducing skip-attention to replace classical long skip connections, reducing parameters and improving BRATS Dice [76]. Recognizing that spatial and channel signals should not be conflated, *UNETR++* proposes an Efficient Paired Attention (EPA) block to disentangle them [77]. Building on this, *DS-UNETR++* introduces a *dual-branch* feature encoder together with a Gated Shared-Weight Pairwise Attention (G-SWPA) module and a Gated Dual-Scale Cross-Attention Module (G-DSCAM), yielding consistent 2–4, % Dice improvements across diverse MSD organs while keeping the overall network lightweight [78].

Vision Transformers have achieved remarkable success in representation learning due to their ability to perform effective token mixing through self-attention. However, the quadratic complexity of self-attention with respect to input resolution becomes prohibitive for high-resolution or volumetric data [79]. To address this limitation, Adaptive Fourier Neural Operators (AFNO) [4] have emerged as an efficient alternative tokenizer that performs global mixing directly in the Fourier domain.While conventional attention requires computing token-token similarity, AFNO models token mixing as a continuous global convolution that can be efficiently approximated in the frequency domain, thus capturing long-range dependencies with quasi-linear complexity and linear memory complexity. This idea has its root in operator learning and the Fourier Neural Operator (FNO) [80] which was originally proposed to tackle a variety of hard PDE problems, fluid dynamics, turbulent flow, and climate modeling. AFNO generalizes the FNO to visual representation learning with architectural adaptations, including, but not limited to, block-diagonal channel-mixing weights, adaptive weight-sharing across tokens, frequency mode sparsification with soft-thresholding and shrink-

age to remove modes with small magnitude, etc., to deal with discontinuities, high-resolution inputs and structured multi-dimensional data.
Unlike other fast attention mechanisms, such as Linformer [81] and Performer [82], which leverage low-rank factorizations or kernel-based approximations, AFNO removes the attention matrix altogether, and instead projects features into the Fourier basis, applies learnable, adaptive filtering to a reduced set of frequency modes, and reconstructs the output through an inverse transform. While the latter acts as an implicit global pooling mechanism, AFNO achieves a similar receptive field as the full self-attention at the cost of far fewer learnable parameters, reduced memory footprint, and better numerical conditioning on high-dimensional 3D grids. Existing frequency token mixers, like FNet [83], rely on a similar set of Fourier transforms, but apply fixed, non-learnable frequency filters, and lack the capacity to model the structured nature of 3D signals.
Building on these strengths, this work introduces AFNO as a fourth research direction within transformer-based 3D medical image segmentation. To the best of our knowledge, this is the first application of AFNO to volumetric medical imaging, demonstrating its potential to establish a new, computationally efficient paradigm for transformer-based 3D segmentation.
In summary, the evolution from streamlined single-branch CNNs to increasingly elaborate multi-branch and attention-centric transformers has steadily enhanced volumetric segmentation performance, albeit at the cost of growing model bulk and deployment friction. Our *AFNO-based* variant, *AMBER-AFNO*, recovers the global-context benefits of self-attention while replacing token-token interactions with light-weight spectral mixing, reducing the number of parameters without reducing performance.

3.3 Methodology

In this section, we introduce AMBER-AFNO, an extension of the AMBER model [2], which replaces the traditional self-attention mechanism with Adaptive Fourier Neural Operators (AFNO) to reduce model complexity and improve computational efficiency. While staying close to the original AMBER design, AMBER-AFNO adapts the architecture to address the 3D volumetric data. As illustrated in Figure 3.1, the architecture consists of two main modules: a hierarchical Transformer encoder with 3D patch embedding and AFNO-based feature mixing; and a lightweight MLP decoder that fuses multi-scale features and predicts the final 3D segmentation mask. Unlike AMBER, which uses a dimensionality reduction layer ("*funnalizer*") to collapse 3D features into 2D outputs, AMBER-AFNO operates entirely in 3D and directly outputs a volumetric segmentation mask of size $D \times H \times W \times N_{cls}$, where N_{cls} is the number of classes. Furthermore, the decoder integrates a deconvolutional layer to up-sample feature maps and recover the original spatial resolution.

3.3.1 Hierarchical Transformer Encoder

We designed a series of Mix Transformer encoders (MiT) for semantic segmentation of 3D images, replacing standard self-attention layers with Adaptive Fourier

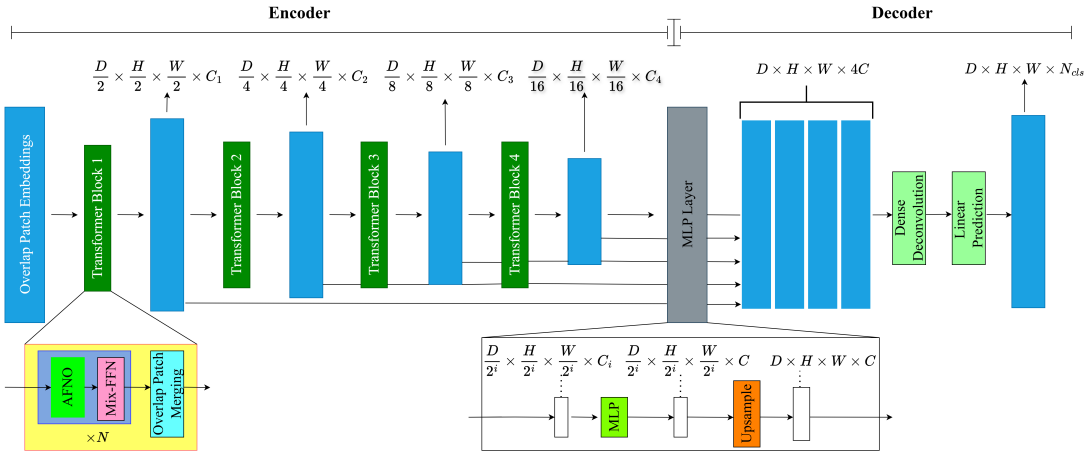

Figure 3.1: **The Proposed AMBER-AFNO framework** consists of two main modules: A hierarchical Transformer encoder to extract coarse and fine features; and a lightweight MLP decoder to directly fuse these multi-level features and predict the semantic segmentation mask. **FFN** indicates a feed-forward network.

Neural Operators (AFNO) to reduce the number of parameters without compromising accuracy. Unlike conventional attention mechanisms—which are memory-intensive and scale quadratically with input size—AFNO performs global feature mixing in the frequency domain with quasi-linear complexity. This enables the encoder to capture both local and global context while significantly lowering computational and memory costs.

Hierarchical Feature Representation. The goal of this module is, given an input image, to generate CNN-like multi-level features. These features provide high-resolution coarse features and low-resolution fine-grained features that usually boost the performance of semantic segmentation [1]. More precisely, given an input 3D image with $D \times H \times W$, we perform patch merging to obtain a hierarchical feature F_1 with a resolution of $\frac{D_{qw}}{2^i} \times \frac{H}{2^i} \times \frac{W}{2^i} \times C_1$, where $i \in \{0, 1, 2, 3\}$ and C_{i+1} is larger than C_i [1].

Overlapped Patch Merging. We utilize merging overlapping patches to avoid the need for positional encoding. To this end, we define K, S, and P, where K is the three-dimensional kernel size (or patch size), S is the stride between two adjacent patches, and P is the padding size. Unlike the original SegFormer, in our experiments, we set $K = 3$, $S = 1$, $P = 1$, and $K = 3$, $S = 2$, $P = 2$ to perform overlapping patch merging. The patch size is intentionally kept small to preserve image details and avoid parameter explosion. $S = 1$ preserves the original image spatial dimensions H and W, avoiding the reduction of spatial dimensions by 1/4 [1].

Adaptive Fourier Neural Operators (AFNO). The self-attention mechanism, while effective in capturing global dependencies, is widely recognized as the primary computational bottleneck in transformer-based encoder architectures due to its quadratic complexity with respect to the input sequence length [1].

In AMBER [2], a multi-head self-attention scheme is used, where each attention layer incorporates a reduction factor R to mitigate this cost. This reduces the computational complexity from the standard $O(N^2)$ to $O\left(\frac{N^2}{R}\right)$, offering a more tractable solution for high-resolution inputs [1].

In the proposed AMBER-AFNO architecture, the entire attention mechanism is replaced with the AFNO block. AFNO uses token mixing in the frequency domain, leveraging the Fast Fourier Transform (FFT) to achieve global interactions with quasi-linear complexity. Here, the input tokens are transformed into the frequency domain using FFT, capturing global contextual information efficiently. In [4] the authors introduced AFNO for the 2D image segmentation task. Using similar methodologies, we have extended the approach for the 3D image segmentation task.

The input to the AFNO block is a 5-D tensor which can be represented as $x \in \mathbb{R}^{B\times D\times H\times W\times C}$ where B is the batch size, D, H, and W are the spatial dimensions (depth, height, and width), and C is the embedding dimension.

We then apply a real-valued 3D Fast Fourier Transform (RFFT) over the spatial dimensions:

$$\hat{x} = \text{RFFT}_3(x) \in \mathbb{C}^{B\times D\times H\times (W/2+1)\times C} \tag{3.1}$$

The channel dimension C is partitioned into K frequency blocks:

$$\hat{x} \to \hat{x}_{\text{blk}} \in \mathbb{C}^{B\times D\times H\times (W/2+1)\times K\times \frac{C}{K}} \tag{3.2}$$

Each frequency block undergoes a learnable complex-valued two-layer MLP:

$$\hat{x}^{(i)}_{\text{blk}} \leftarrow W_2^{(i)} \cdot \phi\left(W_1^{(i)} \cdot \hat{x}^{(i)}_{\text{blk}} + b_1^{(i)}\right) + b_2^{(i)}, \quad \forall i \in \{1, \ldots, K\} \tag{3.3}$$

After the operation of two layer MLP, the MLP output is reshaped back to $\hat{x} \in \mathbb{C}^{B\times D\times H\times (W/2+1)\times C}$ and then soft shrinkage is applied to attenuate small-magnitude frequency responses:

$$\hat{x}' = \text{SoftShrink}(\hat{x}) \tag{3.4}$$

Finally, the inverse 3D FFT (IRFFT) is applied, and the result is combined with the original input via residual addition:

$$\tilde{x} = \text{IRFFT}_3(\hat{x}') + x \tag{3.5}$$

Thus, the final output of the AFNO-3D block is as follows:

$$\boxed{\tilde{x} = \text{IRFFT}_3\left(\text{SoftShrink}\left(\text{MLP}_{\mathbb{C}}\left(\text{RFFT}_3(x)\right)\right)\right) + x} \tag{3.6}$$

Algorithm 1 shows the pseudo-implementation of the AFNO-3D block in the AMBER-AFNO architecture.

Mix-FFN. Likewise, in the AMBER [2] and SegFormer [1], we also used the Mix-FFN, which considers the effect of zero padding using a $3\times3\times3$ Conv in the feed-forward network (FFN). We used Mix-FFN instead of positional encoding because Mix-FFN combines both Depthwise Convolution and MLP layers to capture both local and global context in the seen [1].

$$x_{\text{out}} = \text{MLP}\left(\text{GELU}\left(\text{Conv}_{3\times3\times3}\left(\text{MLP}(x_{\text{in}})\right)\right)\right) + x_{\text{in}} \tag{3.7}$$

where x_{in} is the feature from the self-attention module. Mix-FFN mixes a $3\times3\times3$ convolution and an MLP into each FFN.

Algorithm 1: AFNO-3D with Adaptive Weight Sharing and Adaptive Masking

Input: Tensor $x \in \mathbb{R}^{b \times d \times h \times w \times c}$
Output: Transformed tensor y

$bias \leftarrow x$
$x \leftarrow \text{RFFT3}(x)$
$x \leftarrow \text{reshape}(x, [b, d, h, \frac{w}{2} + 1, k, \frac{c}{k}])$
$x \leftarrow \text{BlockMLP}(x)$
$x \leftarrow \text{reshape}(x, [b, d, h, \frac{w}{2} + 1, c])$
$x \leftarrow \text{SoftShrink}(x)$
$x \leftarrow \text{IRFFT3}(x)$
return $x + bias$

Function `BlockMLP(`x`):`
    $x \leftarrow xW_1 + b_1$
    $x \leftarrow \text{ReLU}(x)$
    return $xW_2 + b_2$

3.3.2 Lightweight All-MLP Decoder

The four feature maps produced by the MiT encoder are first channel-projected to a common embedding size d by position-agnostic MLP layers. Then every projected tensor is trilinearly upsampled to the finest encoder resolution and concatenated along the channel axis. A $1 \times 1 \times 1$ convolution with ReLU activation and 3-D batch normalization fuses this aggregate into a compact representation. A single transposed 3-D convolution subsequently doubles the spatial dimensions, bringing the volume back to the native voxel grid while reducing the channels to N_{cls}. A final $1 \times 1 \times 1$ convolution sharpens the logits, yielding the segmentation mask $M \in \mathbb{R}^{N_{\text{cls}} \times D \times H \times W}$. Compared with the original five-stage AMBER decoder, this variant omits the dedicated spectral-reduction block and integrates the final MLP into the transposed convolution, lowering memory cost without sacrificing accuracy.
The figure 3.2 shows a more simplified and straightforward workflow diagram of the AMBER-AFNO Architecture.

3.4 Experiments

To assess the effectiveness of the proposed AMBER-AFNO architecture, we performed a comprehensive benchmark against a selection of representative convolutional and transformer - based segmentation models. These include U-Net [49], TransUNet [84], Swin-UNet [75], UNETR [74], MISSFormer [85], Swin-UNETR [86], nnFormer [76], UNETR++ [77], LeVit-UNet [87], and PCCTrans [88]. While this list is not exhaustive, it represents a diverse and competitive set of baselines from both CNN and transformer families.
All models are evaluated on three heterogeneous public benchmarks: the ACDC cardiac MR challenge, the Synapse multiorgan abdominal CT segmentation task and the Brain Tumor dataset (BraTS). The following subsections detail the

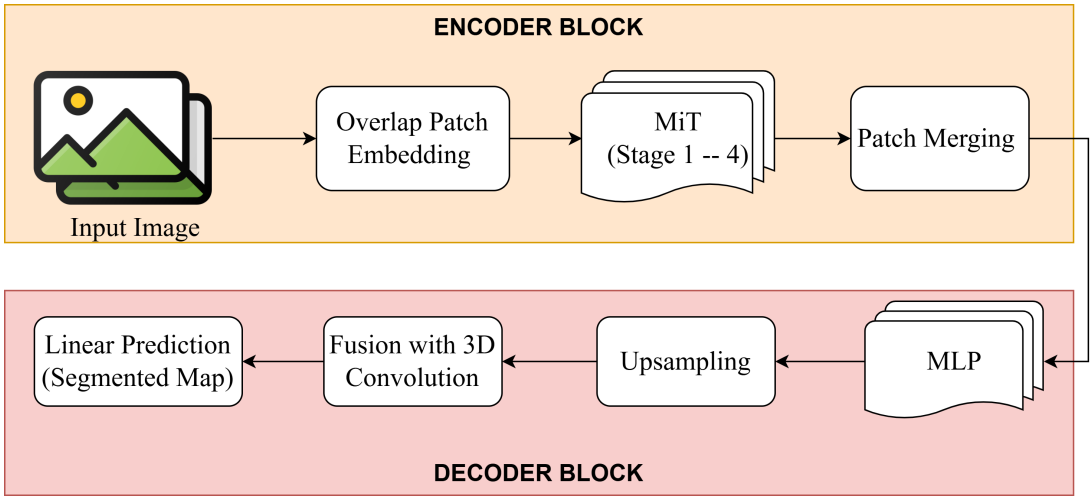

Figure 3.2: Simplified Work Flow Diagram of AMBER-AFNO Architecture

datasets and the pre-processing protocols used, the training and inference procedures adopted throughout this study, and the quantitative criteria — Hausdorff distance (HD95) and Dice similarity coefficient (DSC) — used for performance comparison.

3.4.1 Dataset Overview

Automated cardiac diagnostic segmentation dataset (ACDC)

The ACDC dataset [64] consists of 3D cardiac MRI images with multi-class annotations. A total of 200 labeled samples were used, split into 160 for training and 40 for testing. The annotated classes include the right ventricle (RV), myocardium (MYO), and left ventricle (LV). In this study, the Dice Similarity Coefficient (DSC) was used as the evaluation metric for model comparison.

Multi-organ CT Segmentation Dataset (Synapse)

The Synapse dataset [65] contains 30 abdominal CT scans and is commonly used for benchmarking multi-organ segmentation methods. Following the UNETR++ preprocessing protocol [77], the dataset is split into 18 training volumes and 12 validation volumes. It provides annotations for eight abdominal organs: spleen, right kidney, left kidney, gallbladder, liver, stomach, aorta, and pancreas. For quantitative evaluation, we adopt two standard metrics in medical image segmentation: the Dice Similarity Coefficient (DSC) to measure volumetric overlap, and the 95th percentile Hausdorff Distance (HD95) to assess boundary accuracy.

Brain Tumor Segmentation (BraTS)

The BraTS dataset [66] contains multimodal 3D brain MRI scans of patients with gliomas, which is a type of brain tumor. It contains 387 MRI scan images for training and 73 scans for testing. Here, each image consists of four channels, such as FLAIR, T1w, T1gd, and T2w. The images have four classes, including whole tumor (WT), enhancing tumor (ET), tumor core (TC) and background. In this

Table 3.1: High-level description of the datasets.

Dataset	Spatial Dimension	Depth Dimension	Modality	Class	Training Samples	Test Samples
ACDC	320 × 320 (resampled)	90–130 slices	1 (Cine MRI)	4	160 cases	40 cases
Synapse	512 × 512	75–250 slices	1 (CT)	9	18 scans	12 scans
Brain Tumor (BraTS)	240 × 240	155 slices	4 (MRI)	4	387 scans	73 scans

experiment, the Dice Similarity Coefficient (DSC) is used to measure volumetric overlap, and the 95th percentile Hausdorff Distance (HD95) to assess boundary accuracy.

3.5 Experimental Settings

We have used the same preprocessing steps which are used such as Resampling, Intensity Normalization, Z- Score Normalisation, Cropping and Padding, in the UNETR++ [77], UNETR [74], DS-UNETR++ [78], SAM [89] Models.
The program is executed on a single NVIDIA Tesla V100-SXM2 32G GPU with Thermal Design Power (TDP) of 300W. The CUDA version used is cu12.2. For all data sets, the learning rate was set to 0.01 and weight decay to 3e-5. The Deep Supervision [90] technique was used to formulate the hybrid loss function combining the Dice and Cross Entropy Loss.

3.6 Loss Function

For the semantic segmentation task, one possible problem that arises is class imbalance. We have considered three datasets with multilabel segmentation problems. All the datasets have class imbalance issues. The background class dominates over other classes, as the ROI for medical images is very low compared to the original shape of the image. One loss function alone, such as focal loss, dice loss or cross entropy loss can not handle this situation. To address this issue we have used a custom weighted loss function using the concept of Deep Supervision, which is a combination of cross-entropy loss and dice loss. Instead of relying on the final output of the model, the intermediate feature maps or predictions at different resolutions of the decoder block are also considered while calculating the loss and backpropagation.

$$L(G,P) = 1 - \frac{2}{J}\sum_{j=1}^{J}\frac{\sum_{i=1}^{I} G_{i,j}P_{i,j}}{\sum_{i=1}^{I} G_{i,j}^2 + \sum_{i=1}^{I} P_{i,j}^2} - \frac{1}{I}\sum_{i=1}^{I}\sum_{j=1}^{J} G_{i,j}\log P_{i,j} \tag{3.8}$$

Where **G** refers to the set of ground truth labels, and **P** refers to the set of predicted probabilities. $P_{i,j}$ and $G_{i,j}$ represent the predicted probability and the one-hot encoded true value of class j at voxel i, respectively. I denotes the total number of voxels, and J denotes the number of classes [78].

3.7 Evaluation Metrics

We have adopted two primary evaluation metrics to assess segmentation performance: the Dice Similarity Coefficient (DSC), which quantifies the overlap between predicted and ground truth regions, and the 95th percentile Hausdorff Distance (HD95), which measures the spatial distance between boundary surfaces while mitigating the influence of outliers. Detailed definitions and computation procedures for these metrics are provided in the following subsection.

Hausdorff Distance (HD95)

The HD 95 is a boundary-based metric that evaluates segmentation quality by computing the 95th-percentile distance between the predicted volume's boundary voxels and those of the ground-truth segmentation.

$$HD_{95}(Y, P) \;=\; \max\Big(d_{95}(Y, P),\, d_{95}(P, Y)\Big) \tag{3.9}$$

Here, $d_{95}(Y, P)$ is the maximum 95th percentile distance between the ground truth and predicted voxels, and $d_{95}(P, Y)$ is the maximum 95th percentile distance between the predicted and ground truth voxels.

Dice Similarity Coefficient (DSC).

The Dice Similarity Coefficient (DSC) measures the similarity between two sets, returning values from 0 to 1, with 1 representing perfect similarity. It is computed using the following formula:

$$DSC(G, P) = \frac{2\,|G \cap P|}{|G| + |P|} = \frac{2\sum_{i=1}^{I} G_i P_i}{\sum_{i=1}^{I} G_i + \sum_{i=1}^{I} P_i} \tag{3.10}$$

Where G is the set of real results, P refers to the set of predicted results, G_i and P_i represent the true and predicted values of the voxel i, respectively, and I is the number of voxels.

3.8 Results

In this section, we summarise the results obtained on three different datasets. As previously said, we have compared the performance of our model with other convolution-based image segmentation models, including Swin-UNet [75], UNETR [74], UNETR++ [77], TransUNet [84], MISSFormer [85], nnFormer [76], LeVit-UNet [87], PCCTrans [88] and others. A numerical and visual comparison among the results of the AMBER-AFNO model with these other modelsis performed.

3.8.1 ACDC Dataset

Tab. 3.2 and 3.3 show that AMBER–AFNO achieves the highest overall Dice score in the ACDC validation set (92. 85%), outperforming UNETR ++ (92. 83%) while using fewer than half of its parameters. The proposed model delivers the best myocardium segmentation (90.74%) and ranks second on both the right-ventricle (91.60%) and left-ventricle (96.21%) classes, demonstrating balanced accuracy across all cardiac structures. These results confirm that the AFNO encoder, combined with a lightweight SegFormer-style decoder, can match, or even surpass, state-of-the-art baselines in ACDC with substantially reduced computational demands.

Table 3.2: Dice Similarity Coefficient (%) on the ACDC validation set for the right ventricle (RV), myocardium (Myo) and left ventricle (LV); the column **DSC** reports the mean of the three structures. **Bold** values are the best in each column, while underlined values are the second best.

Methods	**RV**	**Myo**	**LV**	**DSC**
TransUNet[84]	88.86	84.54	95.73	89.71
Swin-UNet[75]	88.55	85.62	95.83	90.00
UNETR[74]	85.29	86.52	94.02	86.61
MISSFormer[85]	86.36	85.75	91.59	87.90
nnFormer[76]	90.94	89.58	95.65	92.06
UNETR++[77]	**91.89**	90.61	96.00	92.83
LeVit-UNet[87]	89.55	87.64	93.76	90.32
PCCTrans[88]	90.55	90.57	**96.22**	92.45
AMBER–AFNO (ours)	91.60	**90.74**	96.21	**92.85**

Table 3.3: Comparison on ACDC. AMBER-AFNO achieves best segmentation results(DSC), while being efficient (Params in millions)

Methods	**Params**	**FLOPs**	**DSC**
UNETR++[77]	81.55	52.14	92.83
AMBER–AFNO (ours)	**14.77**	163.27	**92.85**

3.8.2 Synapse Dataset

As summarized in Tab. 3.4 and 3.5, AMBER–AFNO attains a mean Dice score of 83.76%, outperforming every competing baseline except the much larger nnFormer (86.57%) and UNETR++ (87.22%). Crucially, this performance is reached with only 14.77 M parameters and *without* any task-specific tuning of network depth, patch size, or input resolution.

3.8.3 BraTS Dataset

Our quantitative comparison across the BraTS benchmark (Tab. 3.6) shows that AMBER–AFNO consistently delivers state-of-the-art performance across all tumor subregions. The model achieves the best average DSC (82.82%), outperforming UNETR++ (82.49%) while maintaining a substantially lighter architecture.

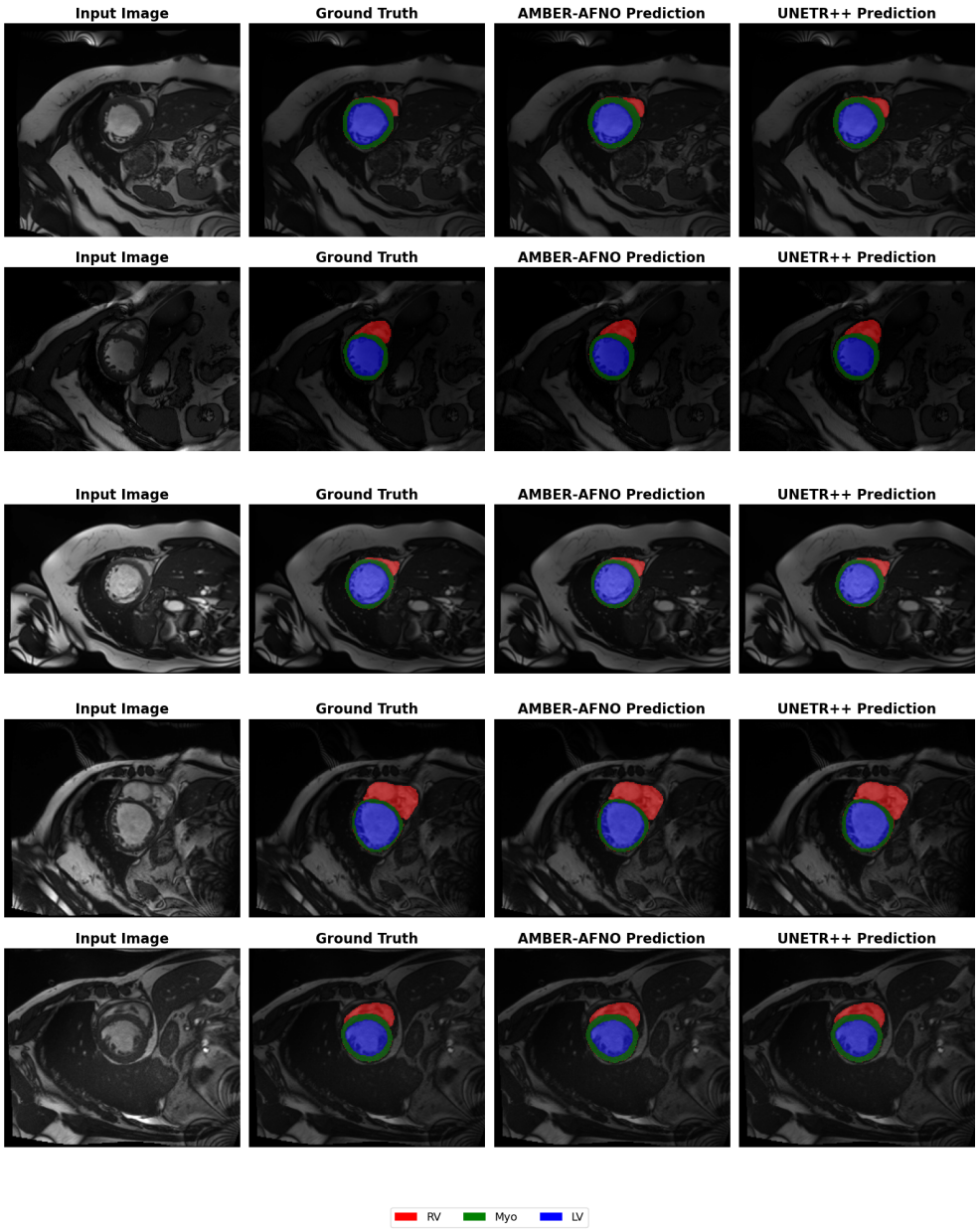

Figure 3.3: Visual Comparison of AMBER-AFNO and UNETR++ Model's Prediction on ACDC Dataset

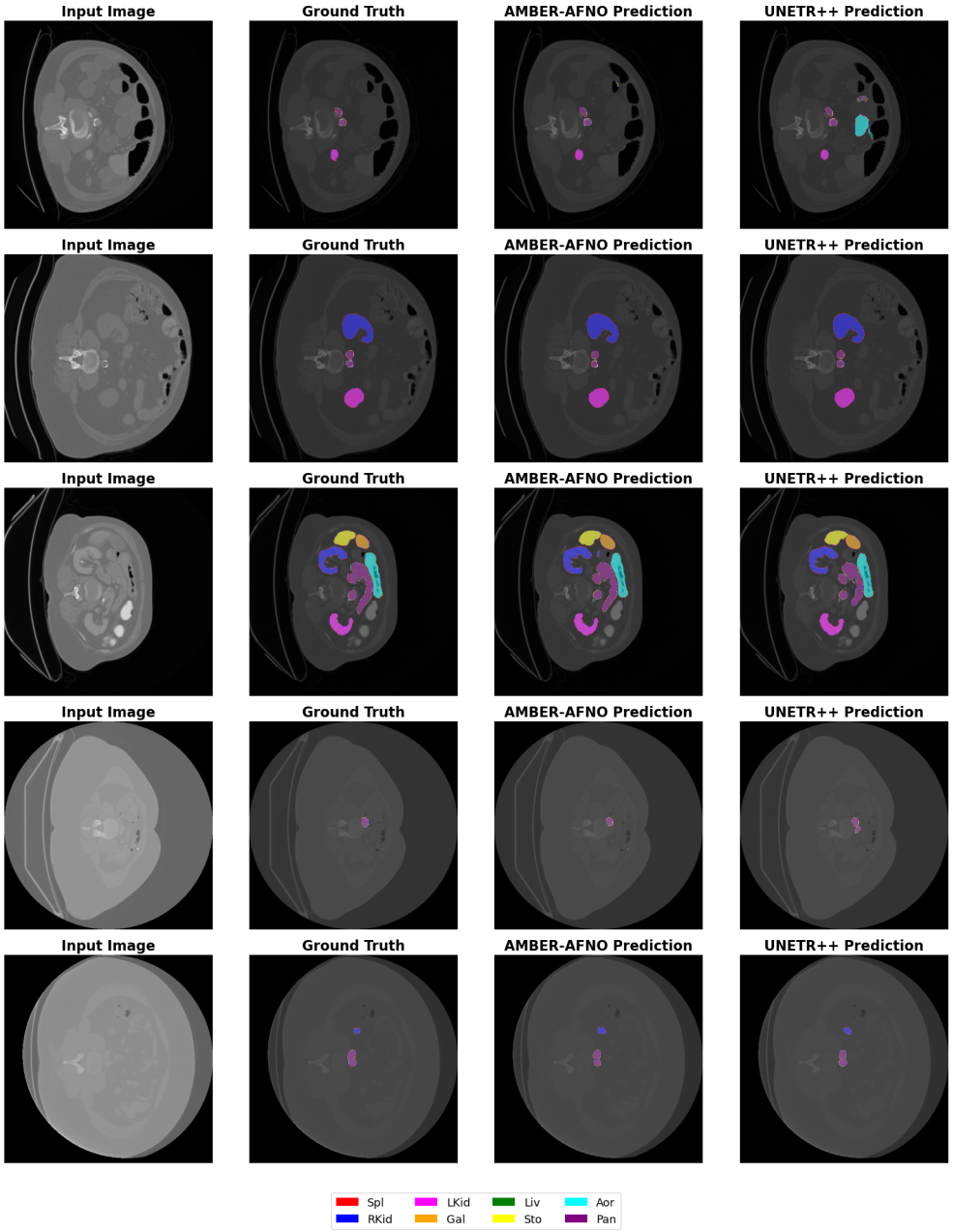

Figure 3.4: Visual Comparison of AMBER-AFNO and UNETR++ Model's Prediction on Synapse Dataset

Table 3.4: Dice scores for eight abdominal organs and HD95 on the Synapse validation set. **Bold** values denote the best result in each column, while underlined values denote the second best.

Methods	Spl	RKid	LKid	Gal	Liv	Sto	Aor	Pan	HD95	DSC
U-Net[49]	86.67	68.60	77.77	69.72	93.43	75.58	89.07	53.98	–	76.85
TransUNet[84]	85.08	77.02	81.87	63.16	94.08	75.62	87.23	55.86	31.69	77.49
Swin-UNet[75]	90.66	79.61	83.28	66.53	94.29	76.60	85.47	56.58	21.55	79.13
UNETR[74]	85.00	84.52	85.60	56.30	94.57	70.46	89.80	60.47	18.59	78.35
MISSFormer[85]	91.92	82.00	85.21	68.65	94.41	80.81	86.99	65.67	18.20	81.96
nnFormer[76]	90.51	86.25	86.57	70.17	**96.84**	**86.83**	92.04	**83.35**	10.63	86.57
Swin-UNETR[86]	95.37	86.26	86.99	66.54	95.72	77.01	91.12	68.80	10.55	83.48
UNETR++[77]	**95.77**	**87.18**	**87.54**	**71.25**	96.42	86.01	**92.52**	81.10	**7.53**	**87.22**
LeVit-UNet[87]	88.86	80.25	84.61	62.23	93.11	72.76	87.33	59.07	16.84	78.53
PCCTrans[88]	88.84	82.64	85.49	68.79	93.45	71.88	86.59	66.31	17.10	80.50
AMBER–AFNO (ours)	87.82	86.26	87.36	61.33	96.02	80.50	91.42	79.36	16.96	83.76

Table 3.5: Comparison on Synapse. AMBER-AFNO achieves the third best segmentation results (DSC), while being more efficient (Params in millions)

Methods	Params	FLOPs	DSC
TransUNet[84]	96.07	88.91	77.49
UNETR[74]	92.49	75.76	78.35
Swin-UNet[75]	62.83	384.2	83.48
nnFormer[76]	150.5	213.4	86.57
UNETR++[77]	42.96	47.98	**87.22**
AMBER–AFNO (ours)	**14.86**	161.24	83.76

AMBER–AFNO also provides the best Dice scores on the Enhancing Tumor (ET) and second best on the Tumor Core (TC) regions, reaching 87.01% and 77.26%, respectively. Figure 3.5 shows a visual comparison of prediction of the AMBER-ANFO model with the UNETR++ model.
In terms of boundary accuracy, AMBER–AFNO attains the second-best average HD95 for all three tumor regions. These results demonstrate that the proposed AFNO-based encoder, paired with a compact decoder, not only improves segmentation fidelity but also achieves balanced and robust performance across all tumor components while remaining computationally efficient.

3.8.4 Discussion: Performance vs Efficiency Trade-off

The experimental results reported in Tables 3.2, 3.3, 3.4, 3.5 and 3.6 highlight the strong performance of the proposed AMBER-AFNO model when compared with both convolutional and transformer-based segmentation architectures. On the ACDC dataset, AMBER-AFNO achieves the highest overall Dice Similarity Coefficient (92.85%), slightly surpassing even the more complex UNETR++ model, which requires more than five times the number of parameters. This result is particularly significant as it confirms that frequency-domain mixing via AFNO can capture global contextual information without the computational burden of traditional self-attention.
On the Synapse dataset, which includes a greater anatomical variety and more heterogeneous imaging conditions, AMBER-AFNO remains highly competitive.

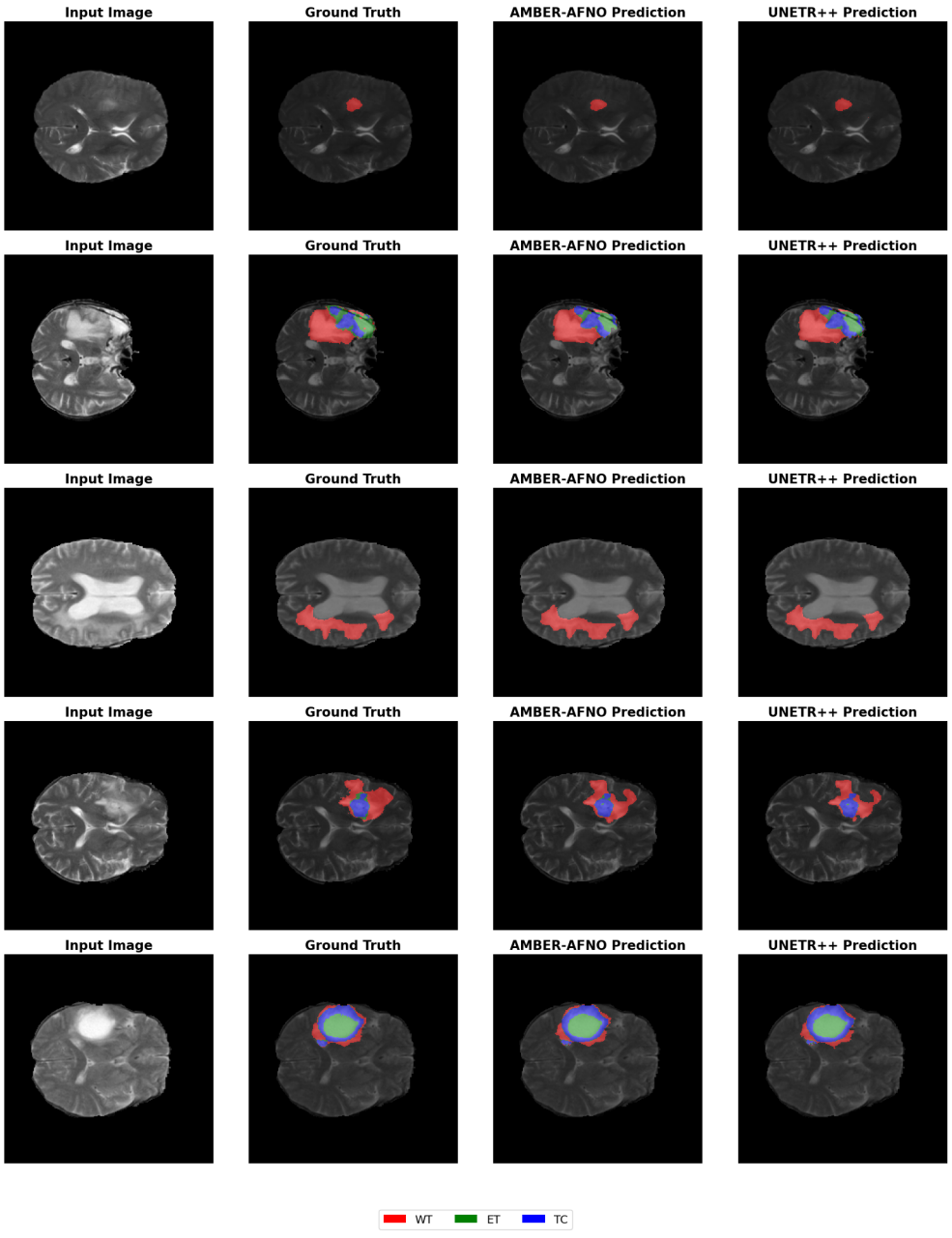

Figure 3.5: Visual Comparison of AMBER-AFNO and UNETR++ Model's Prediction on BraTS Dataset

Table 3.6: HD95 (mm) and Dice Similarity Coefficient (%) for Whole Tumor (WT), Enhancing Tumor (ET), and Tumor Core (TC). The last two columns report the mean HD95 and DSC across the three structures. **Bold** indicates the best performance, while underlined indicates the second-best.

Methods	**WT**		**ET**		**TC**		**Average**	
	HD95	**DSC**	**HD95**	**DSC**	**HD95**	**DSC**	**HD95**	**DSC**
SETR NUP [91]	14.419	69.70	11.72	54.40	15.19	66.90	13.78	63.70
SETR PUP [91]	15.245	69.60	11.76	54.90	15.023	67.00	14.01	63.80
SETR MLA [91]	15.503	69.80	10.24	55.40	14.72	66.50	13.49	63.90
TransUNet [84]	14.03	70.60	10.42	54.20	14.50	68.40	12.98	64.40
TransBTS [92]	10.03	77.90	9.97	57.40	8.95	73.50	9.65	69.60
CoTr w/o CNN encoder [93]	11.49	71.20	9.59	52.30	12.58	69.80	11.22	64.40
CoTr [93]	9.20	74.60	9.45	55.70	10.45	74.80	9.70	68.30
UNETR [74]	8.27	78.90	9.35	58.50	8.85	76.10	8.82	71.10
UNETR++ [77]	**4.76**	**91.07**	**4.86**	78.31	8.19	**78.08**	**5.94**	82.49
IMS2Trans [94]	–	86.57	–	58.28	–	75.67	–	73.51
Hsienchih Ting's Model [95]	–	87.24	–	62.47	–	77.90	–	75.87
AMBER–AFNO (ours)	5.76	90.86	5.54	**80.33**	**7.98**	77.26	6.42	**82.82**

While it is marginally outperformed by the heaviest models in terms of overall DSC, it still achieves a very solid performance (83.76%), only a few points behind state-of-the-art architectures like UNETR++ and nnFormer, but with a significantly lower parameter count and similar inference-time complexity. This reinforces the model's capacity to generalize well across different domains and target structures without the need for extensive task-specific tuning.

On the BraTS dataset, which poses a substantially more challenging scenario due to the heterogeneous appearance of brain tumors and the requirement to segment multiple subregions, AMBER-AFNO continues to demonstrate strong generalization capability. The model achieves an impressive overall Dice score of 82.82%, surpassing UNETR++ despite having a significantly lighter architecture. Notably, AMBER-AFNO attains the best performance on the Enhancing Tumor (ET) region (80.33%) and achieves one of the second-best scores on Whole Tumor (WT) segmentation (90.86%), confirming its ability to capture both fine-grained boundaries and large-scale tumor extents. These results highlight that the AFNO-based encoder effectively handles the complex spatial variability inherent to brain tumor anatomy, enabling high segmentation fidelity across all subregions while maintaining favorable computational efficiency.

The observed trade-off between performance and efficiency strongly supports the core motivation of this study. AMBER-AFNO demonstrates that it is possible to design architectures for 3D medical image segmentation that retain high representational power while dramatically reducing the number of trainable parameters. This is achieved through the use of Adaptive Fourier Neural Operators, which perform learnable filtering in the spectral domain and eliminate the quadratic scaling associated with attention mechanisms. Such a design makes AMBER-AFNO particularly suitable for deployment in resource-constrained environments, including clinical systems with limited memory and compute budgets.

Furthermore, the model achieves these results without relying on complex decoder branches, multi-scale fusion strategies, or deep attention hierarchies. The simplicity and modularity of the architecture make it easy to adapt to different

datasets and segmentation tasks with minimal overhead. Taken together, these aspects make AMBER-AFNO not only a promising benchmark for lightweight segmentation, but also a practical choice for real-world applications where efficiency and robustness are equally important as accuracy.
In particular, its compact design and low memory footprint enable deployment in time- or resource-constrained clinical settings—such as real-time organ delineation during image-guided interventions, integration into portable imaging systems, or use in hospitals with limited computational infrastructure. These scenarios highlight the clinical relevance of AMBER-AFNO's lightweight design, which provides accurate segmentation without the need for high-end GPU resources typical of large transformer architectures.

The proposed model secures second-best results for both kidney classes (**R. kidney:** 86.26%, **L. kidney:** 87.36%) and remains within 1.5 percentage points of the best method on aorta (91.42%) and liver (96.02%), while posting a solid Dice of 79.36% on the notoriously challenging pancreas in the Synapse dataset.
Tab. 3.7 reports the memory profile and the latency of AMBER-AFNO across four hardware configurations combining two modern GPUs (NVIDIA H100 and NVIDIA L40) and two different CPUs (AMD EPYC 7413-24c and AMD EPYC 9534-64c). AMBER-AFNO exhibits an extremely light memory footprint, requiring only 2.96 GB of GPU memory for full-resolution 3D inference, making the model deployable even on mid-range or shared accelerator environments. In terms of speed, the model achieves sub-100 ms latency on the NVIDIA L40 and remains below 160 ms on the H100, demonstrating high throughput for the BraTS 128^3 volumetric inputs. CPU-only execution is naturally slower, but still processes each volume within 3–3.5 s, which remains practical for non-real-time clinical workflows. Overall, the measurements highlight the efficiency of Amber-AFNO, combining low memory demand with fast inference suitable for scalable deployment.

Table 3.7: Computational performance of Amber-AFNO on the BraTS tumor input ($1 \times 128 \times 128 \times 128$). GPU T. and CPU T. denote the average forward-pass latency over 20 runs. Mem indicates the peak GPU memory usage during inference.

GPU	CPU	Params (M)	FLOPs (G)	Mem (GB)	GPU T. (ms)	CPU T. (ms)
NVIDIA H100 PCIe	AMD EPYC 7413 (24c)	9.84	122.9	2.96	159.6	3305
NVIDIA L40	AMD EPYC 9534 (64c)	9.84	122.9	2.96	84.2	3453

3.9 Ablation Study

To isolate the performance attributable to *Adaptive Fourier Neural Operators* (AFNO), we train two encoder variants on the ACDC dataset using **identical** hyper-parameters—network depth, embedding width, number of heads, MLP expansion ratio, and deep-supervision strategy (cf. §3.3.1). The only difference lies in the attention mechanism.

- **AMBER (MHSA).** A light configuration whose AFNO blocks are replaced by standard multi-head self-attention (MHSA); the embedded dimensions are:
  [32, 64, 128, 256].
- **AMBER-AFNO (light).** Uses the AFNO block described in 3.3.1 with the *same* light configuration above.

Although *AMBER-AFNO (light)* uses only **8.7 M** parameters and **58.29 GFLOPs**—less than half the parameters and compute of the MHSA variant (**19.01 M**, **132.07 GFLOPs**)—it achieves a mean Dice score of **92.83 %**, matching the performance of the much larger UNETR++, while requiring only 10% of its parameters and a comparable number of FLOPs. In contrast, the MHSA-based model achieves a lower mean Dice score of **92.03 %**, reflecting a performance drop of approximately 0.8 despite its higher complexity. Table 3.8 presents a detailed comparison of all models, reporting parameter count, FLOPs, and Dice Similarity Coefficient (DSC). The results highlight the superior efficiency–accuracy trade-off enabled by AFNO for 3-D medical image segmentation.
To further assess the stability and reproducibility of the proposed AFNO-based model, we performed a sensitivity analysis by varying key hyperparameters, including learning rate, batch size, embedding dimension, dropout ratio, number of blocks per stage, and use of deep supervision. Table 3.9 summarizes the performance under different configurations. The results indicate that the model maintains consistent Dice scores (0.91–0.93) across a broad hyperparameter range, demonstrating robustness to moderate changes in training settings. While deeper architectures and slightly higher embedding dimensions offer marginal improvements, these gains come at increased computational cost. Similarly, deep supervision and small learning rate adjustments contribute to faster convergence without significantly affecting the final segmentation accuracy. Overall, these observations confirm that the proposed architecture is stable and easily transferable to new datasets with minimal hyperparameter tuning.

Notes on parameters: **Deep Supervision** indicates whether additional prediction heads are used, enabling loss computation at multiple resolutions. **Batch Size** denotes the number of 3D volumes processed per iteration; **Epochs** represents total training cycles; **LR** is the initial learning rate; **Embed Dim.** lists the embedding dimensions across encoder stages; **Dropout** specifies the dropout probability applied in the MLP layers; and **Blocks/Stage** refers to the number of AFNO blocks used in each hierarchical encoder stage.

Table 3.8: Efficiency–accuracy trade-off on the ACDC validation set. AFNO delivers state-of-the-art Dice scores with markedly fewer parameters. Best results are highlighted in **bold**, and second-best results are underlined.

Method	Params	FLOPs	DSC (%)
UNETR++[77]	81.55	**52.14**	92.83
AMBER-AFNO	14.77	163.27	**92.85**
AMBER-AFNO (light)	**8.70**	58.29	92.83
AMBER (MHSA)	19.01	132.07	92.03

Table 3.9: Sensitivity analysis of key hyperparameters on the ACDC validation set. ED = embedding dimensions per stage; B/S = blocks per stage. The model shows stable Dice scores across configurations, confirming robustness.

Exp.	DS	BS	Ep.	LR	ED	DO	B/S	Dice
1	T	4	1000	1e2	(32,64,128,256)	0.1	(2,2,4,2)	**0.93**
2	F	1	1500	3.7e3	(32,64,128,256)	0.1	(3,4,6,3)	0.92
3	T	4	1000	1e2	(64,128,320,512)	0.0	(3,4,6,3)	**0.93**
4	F	4	1000	1e2	(64,128,320,512)	0.0	(3,4,6,3)	0.91
5	F	4	1000	1e2	(64,128,320,512)	0.0	(3,4,6,3)	**0.93**
6	F	4	1000	5.3e3	(64,128,320,512)	0.0	(3,6,36,3)	0.92

3.10 Conclusions

In this study, we introduced AMBER-AFNO, a lightweight yet high-performing architecture for 3D medical image segmentation that replaces conventional attention mechanisms with Adaptive Fourier Neural Operators (AFNO). Originally developed for multiband remote sensing tasks, the AMBER framework was adapted to volumetric medical imaging, resulting in a model that achieves state-of-the-art performance with a fraction of the parameters required by leading transformer-based methods.

We have conducted comprehensive experiments on 3 publicly available datasets (ACDC, Synapse, and BraTS), and show that AMBER-AFNO offers competitive or better performance compared to much larger models (UNETR++ and nnFormer), with 80% fewer parameters. This leads to faster training, lower inference latency, and less memory usage, desirable properties for clinical and edge use cases.

Although AMBER-AFNO yields state-of-the-art results on ACDC and BraTS datasets, we are just a notch below the best models on the more diverse Synapse benchmark. We are currently looking into the matter to push the capacity of our model to generalize even better on a larger set of anatomies and contrasts.

Despite its first place ranking in ACDC and BraTS, the performance of AMBER-AFNO on the more anatomically and radiologically heterogeneous dataset Synapse, is marginally inferior to the topmost performing architectures. This could be attributed to a larger anatomical and intensity variation in abdominal organs in Synapse where the inter organ (say pancreas and stomach) boundary is ambiguous and quite irregular in the frequency domain. The global frequency representation of AFNO might be not as useful for such fine-grained local organ specific information where organs are small and their shapes are irregular.

To overcome these issues, future efforts will be made towards learning a hybrid representation that incorporates both the global mixing power of AFNO and local spatial attention mechanisms, better multi-scale feature integration, and contrastive or domain-adaptive learning to improve dataset generalizability. These efforts will help make the AMBER-AFNO framework even more robust in various clinical imaging environments.

In conclusion, this study demonstrates the potential of frequency-domain mixing in 3D medical imaging and presents AMBER-AFNO as an attractive option for efficient and accurate segmentation. The small footprint of AMBER-AFNO, however, makes it an interesting candidate for real-world clinical integration —

due to its fast inference time, low power consumption, and hardware agnosticism. Future applications may involve on-device segmentation for hand-held scanners, surgical navigation systems, or environments with restricted computational capabilities. As such, we believe that AMBER-AFNO is not only a research-efficient architecture but also a clinically compatible model for real-world deployment.

Chapter 4

AMBER-AFNO as a Hyperspectral Foundation Model: Pretraining on the SpectralEarth Dataset

The emergence of foundation models has profoundly reshaped representation learning in computer vision, natural language processing, and, more recently, Earth observation. In hyperspectral imaging (HSI), this paradigm shift has been delayed not by architectural limitations alone, but primarily by the lack of large-scale, globally representative datasets suitable for self-supervised pretraining.
This gap has now been decisively addressed by *SpectralEarth*[6], a large-scale, multi-temporal hyperspectral dataset derived from the Environmental Mapping and Analysis Program (EnMAP). SpectralEarth comprises 538,974 hyperspectral image patches covering 415,153 unique locations extracted from 11,636 globally distributed EnMAP scenes over a two-year archive. Approximately 17.5% of these locations include multiple temporal observations, enabling multi-temporal hyperspectral analysis [6]. Leveraging this large-scale dataset, we pretrain AMBER-AFNO as a hyperspectral foundation model using the DINO [5] self-supervised learning framework, which is particularly well suited for Vision Transformer–based architectures.
Instead of integrating spectral adapters into classical vision backbones, as proposed in the *SpectralEarth* paper, we adopt the AMBER-AFNO architecture described in the previous chapter, which replaces multi-head self-attention with an Adaptive Fourier Neural Operator (AFNO) to reduce model complexity while preserving global contextual modeling, jointly capturing spatial and spectral information [3].
This chapter builds upon the framework introduced in **SpectralEarth: Training Hyperspectral Foundation Models at Scale** [6], adapting it to evaluate AMBER-AFNO [3] as a hyperspectral foundation model.

4.1 Introduction

In recent years, the volume and accessibility of hyperspectral data have increased substantially with the deployment of new spaceborne missions. Key ex-

amples include Germany's Environmental Mapping and Analysis Program (EnMAP) [13, 6], its precursor mission DESIS mounted on the International Space Station (DESIS) [14, 6], Italy's Hyperspectral Precursor and Application Mission (PRISMA) [15, 6], and the upcoming Copernicus Hyperspectral Imaging Mission for the Environment (CHIME) [16, 6].

The growing availability of large-scale, spaceborne hyperspectral observations creates new opportunities for applying deep learning—and in particular self-supervised learning—to hyperspectral image analysis at unprecedented scale.

Foundation models pretrained through self-supervised learning have shown strong generalization capabilities with minimal task-specific fine-tuning in computer vision [96]. This success has motivated the rapid development of geospatial foundation models, particularly for multispectral and high-resolution RGB imagery [6], where the availability of large, openly accessible satellite archives—such as those provided by the Copernicus Sentinel-2 mission—has enabled training at scale.

In contrast, progress in hyperspectral foundation models has been substantially slower, largely due to the limited availability of large-scale datasets suitable for pretraining. Hyperspectral imagery has traditionally been expensive to acquire, and most benchmark datasets have been restricted to small, localized airborne scenes [11]. Although recent initiatives have begun to increase the scale of unlabeled hyperspectral data [97] [98], they still lack the volume and global diversity required to support the training of truly general-purpose hyperspectral foundation models.

SpectralEarth [6], a large-scale hyperspectral dataset derived from the EnMAP satellite mission, addresses this limitation by providing 538,974 hyperspectral image patches extracted from 415,153 unique locations, with global spatial coverage and cloud contamination below 10%. Compiled from 11,636 EnMAP scenes, the dataset exceeds 3 TB of hyperspectral imagery and represents a substantial leap in scale compared to existing HSI datasets. *SpectralEarth* spans a wide diversity of landscapes, capturing the full variability of spectral signatures observed on Earth. Notably, 17.5% of the locations include multi-temporal acquisitions, enabling the learning of temporally consistent hyperspectral representations.

The dataset includes time series observations that enable multi-temporal hyperspectral analysis. To leverage this rich spectral–spatial information, we pretrain the AMBER-AFNO backbone, described in the previous chapter, using the DINO [5] self-supervised learning on *SpectralEarth*. Unlike approaches that adapt classical vision backbones with spectral adapters, AMBER-AFNO natively processes hyperspectral data by jointly modeling spatial and spectral dimensions while reducing model complexity through the Adaptive Fourier Neural Operator.

In this preliminary work, we focus exclusively on semantic segmentation tasks. To evaluate the effectiveness of the learned representations, we conduct experiments on downstream hyperspectral segmentation benchmarks constructed from EnMAP, DESIS, and Hyperion EO-1 imagery, paired with land-cover, crop-type, and tree-species annotations. The results demonstrate that large-scale self-supervised pretraining of AMBER-AFNO enables strong generalization across diverse hyperspectral datasets and sensors, including those with differing spectral characteristics.

4.2 SpectralEarth Dataset

SpectralEarth is a large-scale hyperspectral dataset derived from imagery acquired by the EnMAP satellite mission. The data were sourced from the DLR GeoPortal and comprise 11,636 EnMAP tiles spanning the full archive between April 27, 2022 and April 24, 2024. All scenes were manually selected through visual inspection to ensure cloud coverage below approximately 10%. Radiometric, geometric, and atmospheric corrections were applied using the official EnMAP L2A processing chain.

Each EnMAP tile was subdivided into non-overlapping geospatial patches of size 128×128 pixels. Each pixel initially contains 224 spectral bands. Bands strongly affected by water absorption—specifically indices [127–141] and [161–167]—were removed due to frequent missing values, following standard practice in hyperspectral dataset construction. After band removal, each patch contains 202 spectral bands.

To exploit repeated acquisitions over the same geographic areas, *SpectralEarth* incorporates multi-temporal hyperspectral views. Overlapping EnMAP acquisitions were identified and used to extract time series of spatially aligned patches, while ensuring that patches corresponding to fixed geo-locations never overlap spatially. This procedure required identifying spatial intersections between tiles and extracting patches from their common regions. Patch locations were tracked using an R-tree structure to prevent spatial overlap.

Due to the combinatorial complexity arising from areas with many overlapping acquisitions, several optimization strategies were employed, including breadth-first search–based subset selection, elimination of redundant computations, and parallel processing across connected components of the overlap graph.This led to 73,307 unique locations with multiple timestamps. For locations with a single acquisition, we extracted non-overlapping patches. In total, SpectralEarth contains 415,153 unique locations, corresponding to 538,974 hyperspectral patches.

The collected data set spans the globe and includes a variety of land cover types and scenes, see Fig. 4.1. The temporal distribution is skewed to the needs of the EnMAP mission: around 17.5% of sites contain data from more than one time stamp. The majority of these multi-temporal sites consist of two data takes, the most common time difference between two consecutive acquisitions is 4 days (EnMAP's target revisit frequency), and the median time difference is 27 days.

Looking at the months of the year, all of them are covered by SpectralEarth albeit with some moderate seasonal bias. Most of the winter months, especially January, are relatively less frequently observed, due to generally cloudier weather conditions but also to short temporary outages of the satellite in December and January 2023 and 2024, while most of the summer months, July and August, are relatively more densely sampled. These properties emerge naturally from the tasking-based acquisition planning of EnMAP, and resemble the situation for real hyperspectral satellite missions. Future missions like ESA's CHIME [16] constellation are likely to offer even higher temporal density and coverage.

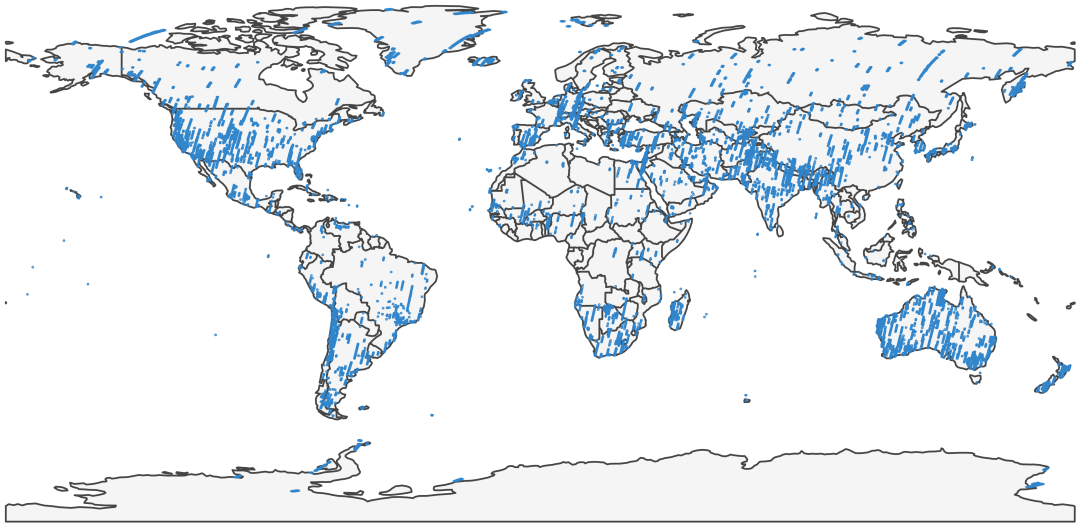

Figure 4.1: Geographical Distribution of SpectralEarth. The map depicts the global coverage of hyperspectral images within our dataset, demonstrating its extensive geographical scope. Reproduced from [6] under the CC BY 4.0 license.

4.3 Experimental Setup

This section describes the experimental protocols adopted for self-supervised pretraining and downstream evaluation. A schematic overview of the overall pipeline is provided in Fig. 4.2. While multiple downstream tasks are commonly employed in the literature to benchmark hyperspectral foundation models, the present work focuses exclusively on *semantic segmentation* as a preliminary evaluation of the proposed architectures.

4.3.1 Self-Supervised Pretraining

Self-supervised pretraining is performed on the *SpectralEarth* dataset using transformer-based spectral encoders. In related studies, a variety of architectures and learning paradigms have been explored, including spectral ResNet and Vision Transformer backbones trained with MoCo-v2, DINO, and masked autoencoding (MAE). In this thesis, we restrict our attention to transformer-based encoders compatible with the AMBER-AFNO design, and adopt the DINO framework for self-supervised learning.

To preserve the physical integrity of hyperspectral signatures, data augmentations are carefully selected. We employ spatial augmentations inspired by SimCLR, including random resized cropping and horizontal and vertical flipping, while explicitly excluding color jittering and grayscale transformations, which would distort spectral information. When available, multi-temporal acquisitions of the same geographic location are exploited as positive pairs during pretraining, providing a natural form of temporal augmentation.

Pretraining is conducted for a fixed number of epochs using large batch sizes and the AdamW optimizer, with a cosine annealing learning rate schedule. These settings are chosen to ensure stable convergence and to facilitate the learning of transferable spectral–spatial representations at scale.

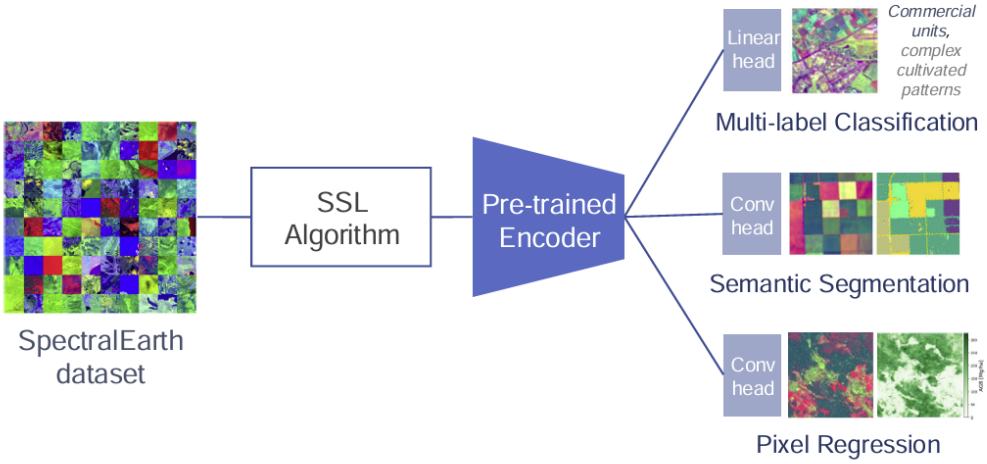

Figure 4.2: Schematic overview of the pre-training/fine-tuning pipeline. An encoder is pre-trained on SpectralEarth, and adapted to the several downstream tasks with task-specific heads. Reproduced from [6] under the CC BY 4.0 license.

4.3.2 Downstream Task: Semantic Segmentation

As a first step toward assessing the quality of the learned representations, we evaluate the pretrained models on semantic segmentation tasks derived from spaceborne hyperspectral imagery. Semantic segmentation provides a particularly stringent test of representation quality, as it requires both accurate local classification and coherent global spatial structure.
Following a lightweight transfer protocol commonly adopted in recent literature, the pretrained encoders are converted into segmentation models by attaching a shallow decoder head. For transformer-based backbones, the token embeddings from the final encoder layer are reshaped into a two-dimensional feature map and passed through a small number of convolutional layers with upsampling operations to recover the original spatial resolution.
All segmentation models are trained using a pixel-wise cross-entropy loss and optimized with the AdamW optimizer. Training is performed for a fixed number of epochs, with batch sizes adapted to the spatial resolution and memory requirements of each dataset. Data augmentation is limited to spatial transformations, namely random resized cropping and horizontal and vertical flipping, in order to avoid altering spectral content.

4.3.3 Pre-training Algorithms

Hyperspectral data exhibit characteristics that differ significantly from RGB images, and as a result, the relative performance of self-supervised learning methods established in conventional computer vision benchmarks does not necessarily transfer to the hyperspectral domain. While the literature commonly explores a broad range of SSL approaches to establish comparative baselines (as MoCo-V2 [99] or MAE [100]), the present work focuses exclusively on the DINO self-

supervised learning framework [5].
DINO is particularly well suited to Vision Transformer–based architectures, providing strong representation quality in frozen-encoder settings and enabling effective learning of global contextual features without reliance on negative samples. These properties make DINO a natural choice for pretraining the AMBER-AFNO architecture on large-scale hyperspectral data.

4.3.4 Evaluation Protocols

To assess the quality and transferability of the representations learned during self-supervised pretraining, we consider multiple evaluation protocols that differ in the degree of adaptation allowed during downstream training. These protocols are designed to probe complementary aspects of representation quality, ranging from pure feature reuse to task-specific adaptation.

Frozen Encoder. In the frozen-encoder setting, the parameters of the pretrained backbone are kept fixed and only the task-specific head is trained on the downstream dataset. This protocol evaluates the intrinsic quality of the learned representations without any adaptation of the backbone and is widely adopted in the self-supervised learning literature as a standardized measure of representation usefulness [96, 101]. Strong performance in this regime indicates that the pretrained features are linearly separable and transferable across datasets.

Full Fine-Tuning. In the full fine-tuning protocol, all model parameters are updated during downstream training. This setting assesses the maximum achievable performance when the pretrained weights are used as initialization and adapted to the target dataset. While this protocol often yields the best absolute performance, it is more sensitive to overfitting, particularly when the downstream dataset is small or exhibits a distribution shift relative to the pretraining data.

Adapter Fine-Tuning. While frozen-encoder evaluation provides a conservative estimate of representation quality, it does not always yield optimal results in remote sensing applications. Hyperspectral data are strongly influenced by factors such as geographic location, atmospheric conditions, seasonal variability, and sensor-specific spectral responses. Moreover, the relevance of individual spectral bands may vary significantly across downstream tasks [6].
To account for these factors, we consider an intermediate adaptation strategy in which only the initial spectral adapter block is fine-tuned, while the remainder of the backbone is kept frozen. This *adapter fine-tuning* protocol enables limited task-specific adaptation while preserving the majority of the pretrained representations, offering a favorable trade-off between flexibility and robustness. In practice, this approach often achieves performance comparable to full fine-tuning, while reducing the risk of overfitting and the computational cost of training.

4.3.5 Scope of the Preliminary Evaluation

Although hyperspectral foundation models can be evaluated across a wide range of downstream tasks—including multi-label classification, parameter re-

gression, and pixel-wise regression—such experiments are beyond the scope of the present preliminary study. The focus on semantic segmentation is motivated by its relevance to land-cover mapping and by its sensitivity to both spectral discrimination and spatial consistency. More extensive downstream evaluations, including regression and cross-task transfer, are left to future work.

4.4 AMBER-AFNO Model

This section motivates and discusses the architectural and methodological design choices underlying the proposed approach, with a particular focus on the rationale for adopting the AMBER-AFNO architecture.

4.4.1 Network Architecture

To support large-scale self-supervised pretraining on spaceborne hyperspectral imagery, the adopted architecture must operate on large spatial patches while jointly modeling spectral and spatial information. Traditional hyperspectral approaches, which are often designed at the pixel level or rely on small spatial contexts, do not scale effectively to this setting. Moreover, classical vision architectures based on convolutional networks or Vision Transformers typically treat spectral bands independently [6] or rely on multi-head self-attention, which becomes computationally expensive when applied to high-dimensional hyperspectral data.

The *AMBER-AFNO*[3] architecture addresses these limitations by combining three-dimensional convolutions with an Adaptive Fourier Neural Operator (AFNO) [4]. Three-dimensional convolutions enable explicit spectral–spatial tokenization, allowing local spectral correlations and spatial structures to be captured jointly at early stages of the network. This design eliminates the need for explicit spectral dimensionality reduction and preserves fine-grained spatial detail throughout the encoding[3].

AMBER-AFNO also implements an efficient representation of the global context. Specifically, it replaces the multi-head self-attention with frequency-domain mixing, as proposed in the AFNO [4] framework. While the complexity of the standard self-attention is quadratic in the number of tokens, the complexity of the frequency-domain mixing in the Fourier space is linear w.r.t. the token dimension. This allows for global receptive fields without introducing a proportional number of model parameters. Consequently, this greatly reduces the size of the overall model, while still being able to learn long-range dependencies over the hyperspectral cube [3].

Hence, the 3D convolution for local spectral-spatial encoding and the AFNO for global spectral-spatial encoding together form an efficient and parameter-effective way for hyperspectral representation learning, and the module is very suitable for large-scale self-supervised pre-training, which makes the *AMBER-AFNO* have a high efficiency in processing the high dimensional hyperspectral data and have a good representation capability.

A critical module specific to AMBER [2] is the *Funnelizer* which connects the 3D spectral and spatial feature representations to 2D semantic maps. It is realized by a 3D convolution with a kernel of size $(D \times 1 \times 1)$ where D is the spectral

depth of the feature volume. This reduces the feature volume to 2D encoding while retaining the spatial resolution.This operation combines the learned spectral signatures to create a 2D image. A subsequent 2D convolution with a 1×1 kernel produces the final segmentation image of size $H \times W \times N_{\text{cls}}$ where N_{cls} is the number of semantic classes.

4.5 Downstream Evaluation: DESIS-CDL (Preliminary Results)

To assess the effectiveness and generalizability of the representations learned during self-supervised pretraining on *SpectralEarth*, a number of downstream benchmarks have recently been proposed for hyperspectral foundation models. These benchmarks typically combine hyperspectral imagery acquired by different sensors—such as EnMAP, DESIS, EO-1 Hyperion, or airborne platforms—with geospatial products for land-cover mapping, agricultural monitoring, and forest analysis. Although the associated labels may contain a degree of uncertainty due to spatial resolution mismatches or annotation noise, they are widely regarded as sufficiently reliable for comparative evaluation and have become standard in the remote-sensing foundation-model literature.

A key question for hyperspectral foundation models is *cross-sensor transferability*: the ability of representations learned from one sensor to generalize to data acquired by different instruments with distinct spectral response functions, noise characteristics, and acquisition geometries. Recent studies have investigated this aspect by evaluating models pretrained on SpectralEarth across multiple unseen sensors, including EO-1 Hyperion (EO1-CDL), DESIS (DESIS-CDL), and airborne datasets emulating the spectral characteristics of upcoming missions such as Intuition-1 (Hyperview). These datasets differ substantially from EnMAP in terms of spectral sampling, band placement, and signal-to-noise ratio, and therefore constitute a stringent test of representation robustness.

In this thesis, we restrict our analysis to a *preliminary cross-sensor evaluation* based exclusively on the **DESIS-CDL** dataset. This choice is motivated by two considerations. First, DESIS provides spaceborne hyperspectral imagery with spectral characteristics that differ from EnMAP, while remaining sufficiently close in spatial resolution and acquisition modality to allow a controlled assessment of transferability. Second, limiting the scope to a single downstream dataset allows us to isolate architectural and representation-level effects specific to *AMBER-AFNO*, without conflating them with extensive dataset-specific tuning or large-scale benchmarking across multiple sensors.

The DESIS-CDL dataset contains images captured for the summers of 2018–2023 from the DLR Earth Sensing Imaging Spectrometer Mission (DESIS) instrument, matched with the corresponding CDL mask for crop segmentation. DESIS images are generated at 30 m spatial resolution with 235 bands covering wavelengths from 400 nanometers through a micron. The resulting dataset comprises 1000 images with 14 classes [6](see Fig. 4.3).

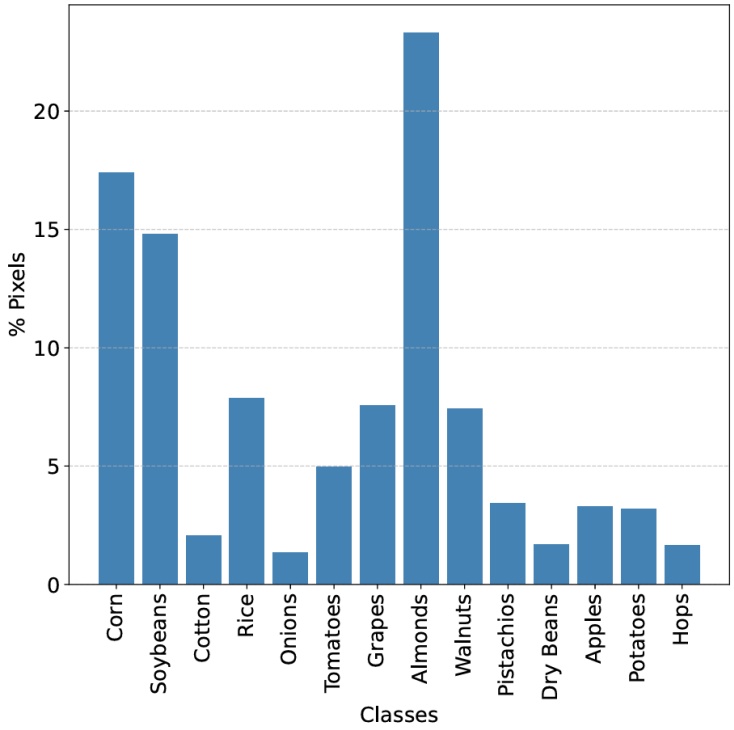

Figure 4.3: Class distributions for the downstream dataset DESIS-CDL. Reproduced from [6] under the CC BY 4.0 license.

4.5.1 Self-Supervised Pretraining

In this preliminary study, AMBER-AFNO pretrained on SpectralEarth is evaluated on DESIS-CDL using standard fine-tuning protocols. The results indicate that the pretrained model consistently outperforms training from scratch, confirming that the representations learned from EnMAP-based SpectralEarth data transfer effectively to DESIS imagery. Notably, this transferability is achieved despite the *substantially lower parameter count* of AMBER-AFNO relative to transformer-based hyperspectral foundation models relying on multi-head self-attention.

Specifically, AMBER-AFNO contains approximately **8.1 million trainable parameters**, whereas commonly used spectral Vision Transformers pretrained with MAE comprise approximately **86 million parameters** for Spec. ViT-B and **304 million parameters** for Spec. ViT-L. Despite this order-of-magnitude reduction in model size, the performance of AMBER-AFNO on DESIS-CDL remains *comparable to significantly larger architectures*. This result highlights the effectiveness of the Adaptive Fourier Neural Operator in capturing global spectral–spatial dependencies without the quadratic scaling and parameter overhead associated with multi-head self-attention.

At this stage, we do not perform a full comparative evaluation against alternative hyperspectral foundation models such as DOFA or HyperSigma across multiple sensors. Instead, the DESIS-CDL experiment is intended to establish a proof of concept: namely, that *AMBER-AFNO learns transferable spectral–spatial representations that remain effective under sensor shift, even in a highly parameter-efficient regime*. A more exhaustive cross-sensor benchmarking—including EO-1

Hyperion and Hyperview—is left to future work and represents a natural extension of the framework developed in this thesis.
It is worth noting that, unlike several transformer-based baselines, the current implementation of the AMBER-AFNO backbone does not explicitly incorporate a dedicated spectral adapter module. Consequently, the *adapter fine-tuning* protocol is not applicable in the present experiments, and the evaluation is restricted to frozen-encoder and full fine-tuning settings. The introduction of adapter-based adaptation mechanisms within the AMBER architecture constitutes an interesting direction for future development.
In summary, this early analysis supports the understanding of AMBER-AFNO as a competent and scalable hyperspectral foundation model. It shows that large-scale self-supervised pretraining on SpectralEarth gives rise to representations that are not sensor-specific, but reveal sensor-invariant patterns within hyperspectral datacubes. This is done with a lightweight architecture, which confirms the core hypothesis of this paper: efficient operator-inspired architectures can achieve the representational capacity of significantly larger transformer models while providing better scalability and practical applicability in the context of diverse Earth observation missions.
The most relevant training hyperparameters used for this baseline assessment are reported below. The hyperspectral image blocks input to AMBER-AFNO are obtained via a **spatial patch size of** 4×4, which represents a good compromise between image spatial resolution and computational complexity, while keeping most of the local spatial-spectral information. In the course of the training stage for the DESIS-CDL dataset, the models are fitted via the `AdamW` optimizer with a learning rate of 1×10^{-3} and a weight decay of 1×10^{-4}. Such a setting turned out to be consistent in all the experimental protocols, and is consistent with the one adopted in previous works on fine-tuning lightweight transformers in remote sensing applications. This setup, even in the presence of the small number of parameters of AMBER-AFNO, guarantees convergence of the model and competitive results under the cross-sensor transfer protocol, which once again proves the efficiency of the proposed architecture.

4.6 Conclusions

This chapter investigated *AMBER-AFNO* as a hyperspectral foundation model, focusing on large-scale self-supervised pretraining on the *SpectralEarth* dataset and on a preliminary evaluation of cross-sensor transferability. The central objective was to assess whether a hyperspectral-native, parameter-efficient architecture can learn representations that generalize beyond the sensor used during pretraining, while avoiding the computational overhead typical of large attention-based transformers.
Leveraging the scale and diversity of *SpectralEarth*, we pretrained AMBER-AFNO using the DINO self-supervised framework, demonstrating that meaningful spectral–spatial representations can be learned without explicit supervision or aggressive spectral preprocessing. Unlike approaches that retrofit classical vision backbones with spectral adapters, AMBER-AFNO natively processes hyperspectral datacubes, jointly modeling spatial structure and spectral continuity through three-dimensional convolutions and global Fourier-domain mixing. This design

Backbone	Protocol	DESIS (mIoU)
AMBER-AFNO	From Scratch	64.12 ± 0.32
	Frozen	63.63 ± 0.58
	Full FT	67.13 ± 0.43
SpecViT-B	From Scratch	66.18 ± 0.50
	Frozen	62.22 ± 0.61
	Full FT	68.50 ± 0.78
	Adapter	67.74 ± 0.29
SpecViT-L	From Scratch	66.06 ± 0.57
	Frozen	62.78 ± 0.31
	Full FT	**69.05** ± 0.50
	Adapter	<u>68.45</u> ± 0.16
DOFA-B	Frozen	57.92 ± 0.37
	Full FT	63.43 ± 0.72
DOFA-L	Frozen	58.26 ± 0.36
	Full FT	62.88 ± 0.34
SpatSigma-B	Frozen	59.15 ± 0.76
	Full FT	62.48 ± 0.29
SpatSigma-L	Frozen	56.72 ± 0.35
	Full FT	62.41 ± 0.35

Table 4.1: Cross-sensor transferability on CDL-based benchmarks. Models are pretrained on SpectralEarth and evaluated on DESIS CDL. Best results are shown in **bold**, second-best results are <u>underlined</u>.

choice enables the preservation of fine-grained spectral information while maintaining favorable computational scaling.

The downstream evaluation on the **DESIS-CDL** benchmark provides an initial but informative assessment of cross-sensor transferability. Despite the spectral and radiometric differences between EnMAP and DESIS, AMBER-AFNO pretrained on SpectralEarth consistently outperforms training from scratch and achieves performance comparable to substantially larger hyperspectral Vision Transformer models. Interestingly, we achieve this with about **8.1 million trainable parameters**, whereas attention-based models have tens or hundreds of millions of parameters. This demonstrates that global spectral and spatial dependencies can indeed be modeled using Adaptive Fourier Neural Operators, without using quadratic-complexity self-attention.

Apart from the numerical results, a more practical conclusion can be drawn from the two protocols. That is, the adapter-tuning strategy can be considered as a good trade-off between robustness and generalizability. It allows for task- and sensor-specific fine-tuning of small parts of the network while keeping most of the knowledge learned from the pretraining source. This feature is of great value in real-world hyperspectral problems where the amount of labeled samples is often limited and data collection settings may differ significantly in terms of spatial location, time and sensors.

Collectively, the results in this chapter provide evidence for the plausibility of the notion of AMBER-AFNO as a *feasible and scalable hyperspectral foundation model*. They show that one can learn representations by self-supervised pretrain-

ing on a large dataset from a single spaceborne instrument that generalize to new, sensor-specific settings and that are sensor-invariant in the sense that they reveal structure of the hyperspectral datacube that is shared across different sensors. This ability to generalize is achieved by a computationally light model, which confirms the claim that in the field of operator-inspired design of efficient architectures one can find models with a high representational power that is equivalent to the one of much larger transformer models.

Meanwhile, this work is just a starting point. To truly characterize hyperspectral foundation models, one needs more cross-sensor comparison, more downstream tasks (e.g. regression, unmixing), and a comparison against other self-supervised learning goals and architectures. All of these are directions for the future. Nevertheless, the results shown here, however, demonstrate that it is reasonable to consider that AMBER-AFNO is a simple and versatile architecture which can serve as a useful starting point for large-scale hyperspectral representation learning, connecting the foundation-model universe to the Earth-observation deep-learning universe.

Chapter 5

Resolution of Astrophysical Inverse Problems through Deep Learning

This chapter introduces astrophysical inverse problems and motivates their resolution with modern deep learning methods. To set the stage, we first frame imaging as the inversion of a forward operator—often linear, convolutional, and ill-posed. Building on this foundation, we then focus on radio-interferometric image reconstruction as a central case study. Finally, bringing these ideas together, we discuss how astrostatistics and astroinformatics, exemplified by the ESO BRAIN study (RESOLVE and DeepFocus), provide scalable pathways toward next-generation ALMA imaging. This chapter includes material previously published in [7], an open access article distributed under the Creative Commons Attribution (CC BY 4.0) license. The candidate is a co-author of this publication.

5.1 Introduction

Many problems in science and engineering concern the inference of the causal mechanisms that give rise to a set of observations. More formally, the goal is to determine the relationship linking the physical parameters of a model to a set of measured data, under the assumption that a causal mapping connects the two. Let m denote the mathematical model describing the physical system of interest, $b \in \mathbb{R}^n$ the observable quantities (or functions thereof), and G the operator mapping the model to the observations. This relationship can be written as

$$G(m) = b. \tag{5.1}$$

In this framework, the *forward problem* consists of computing the observations b given a model m, that is, predicting the data from the underlying physical description. Numerical simulations are typical examples of forward problems. Conversely, *inverse problems* aim at recovering the model m from the observations b.

The operator G, commonly referred to as the *forward operator*, encodes the physics of the measurement process. Depending on the application, it may take the form of an ordinary differential equation (ODE), a partial differential equation (PDE), or a linear or nonlinear system of equations. In many imaging systems,

and in particular in astronomical imaging, the noiseless observation can be well approximated as a linear function of the underlying model. Under this assumption, the operator G is linear.
Moreover, most astronomical imaging instruments are approximately space invariant. As will be discussed later in this chapter, this property implies that the forward operator can be modeled as a convolution with the point spread function (PSF) of the instrument.
The formulation in Eq. (5.1) does not explicitly account for the fact that real measurements are inevitably affected by noise and that models are only approximations of the true physical processes. A more realistic formulation of inverse problems is therefore

$$b = G(m) + \varepsilon, \tag{5.2}$$

where ε represents uncertainties arising from measurement noise, modeling errors, and unmodeled physical effects. This expression does not necessarily imply strictly additive noise, but rather accounts for the discrepancy between the observed data and the ideal noiseless prediction.

5.1.1 Ill-posedness of Inverse Problems

The defining characteristic that makes inverse problems challenging is that they are often *ill-posed*, in the sense that they do not satisfy the Hadamard conditions:

1. **Existence**: a solution may not exist;
2. **Uniqueness**: if a solution exists, it may not be unique;
3. **Stability**: small perturbations in the data may lead to arbitrarily large perturbations in the solution[102].

The lack of existence may arise when the model is insufficiently expressive or when the data are contaminated by noise. Non-uniqueness occurs when multiple models explain the observations equally well. Instability implies that even when a solution exists and is unique, it may be extremely sensitive to noise, rendering the inversion numerically unstable.

5.1.2 Operator-Theoretic Formulation

Let X and Y be functional spaces associated with the model and the observations, respectively. The forward operator can be written as

$$G : X \to Y, \qquad b = G(m). \tag{5.3}$$

If an inverse operator G^{-1} exists,

$$G^{-1} : G(X) \to X, \qquad m = G^{-1}(b), \tag{5.4}$$

then the problem admits a unique solution. In this case, G is bijective and the inverse problem is well-posed with respect to existence and uniqueness. The set $G(X) \subset Y$ is referred to as the range of the operator.

In this thesis, we restrict the discussion to linear operators, which satisfy

$$G(\alpha m_1 + m_2) = \alpha G(m_1) + G(m_2), \tag{5.5}$$

for all $m_1, m_2 \in X$ and $\alpha \in \mathbb{R}$.

A linear operator G is said to be bounded if there exists a constant $C > 0$ such that

$$\|G(m)\| \leq C\|m\| \quad \forall m \in X. \tag{5.6}$$

If a linear operator is both bounded and closed, it is compact. A compact linear operator is invertible only if its range has finite dimension. Furthermore, if the inverse operator is also bounded, the solution depends continuously on the data, and the problem is said to be well-conditioned.

5.1.3 Integral Formulation and Fredholm Equations

Given the continuous nature of the problems considered in this thesis, we further restrict our attention to linear integral operators. Equation (5.1) can then be written as

$$\int_a^b g(x, \xi)\, m(\xi)\, d\xi = b(x), \tag{5.7}$$

where $g(x, \xi)$ is the kernel of the operator, corresponding to the PSF of the imaging system. Equations of this form are known as *Fredholm integral equations of the first kind*, with $m(x)$ as the unknown function.

Such equations are generally ill-conditioned unless additional constraints are imposed on the functional spaces of m and b. In many practical applications, these constraints are introduced through regularization or prior information.

When the kernel depends only on the difference $x - \xi$, the integral equation reduces to a convolution:

$$\int_{-\infty}^{+\infty} g(x - \xi)\, m(\xi)\, d\xi = b(x). \tag{5.8}$$

5.1.4 Linear Time-Invariant Systems and Convolution

A fundamental property of linear time-invariant (or space-invariant) systems is that their forward operator can always be expressed as a convolution:

$$b(t) = \int_{-\infty}^{+\infty} g(t - \tau)\, m(\tau)\, d\tau. \tag{5.9}$$

Using the sifting property of the Dirac delta function,

$$m(t) = \int_{-\infty}^{+\infty} m(\tau)\, \delta(t - \tau)\, d\tau, \tag{5.10}$$

and the linearity of G, any linear system can be rewritten in the convolutional form of Eq. (5.9).

5.1.5 Fourier Representation and the Convolution Theorem

The Fourier transform of a function $g(t)$ is defined as

$$G(f) = \int_{-\infty}^{+\infty} g(t)\, e^{-i2\pi ft}\, dt, \tag{5.11}$$

with inverse transform

$$g(t) = \int_{-\infty}^{+\infty} G(f)\, e^{i2\pi ft}\, df. \tag{5.12}$$

The Convolution Theorem states that convolution in the time (or image) domain corresponds to multiplication in the Fourier (frequency) domain:

$$\mathcal{F}[m * g](f) = M(f)\, G(f). \tag{5.13}$$

Applying this result to Eq. (5.1) yields

$$B(f) = G(f)\, M(f), \tag{5.14}$$

where $G(f)$ is the *transfer function* of the system. Although a band-limited transfer function would, in principle, lead to a well-posed problem, instrumental limitations and noise — which is generally not band-limited — render most practical inverse problems ill-posed.

5.2 Classes of Inverse Problems in Astrophysics

As discussed in the previous section, the data acquisition process in most astronomical instruments can be accurately modeled as a linear forward process. As a consequence, a large fraction of data processing tasks in astrophysics can be reformulated as linear inverse problems of the form

$$Y = HX \bullet N, \tag{5.15}$$

where Y denotes the observed data, X the unknown signal or image to be recovered, H the linear forward operator, and N an unknown noise or perturbation term. The operator $\bullet$ represents the mechanism by which noise contaminates the observations. Depending on the context, this contamination may be additive, multiplicative, or more generally non-linear.

The noise term N may arise from stochastic processes intrinsic to the detector (e.g. thermal or photon noise) or from deterministic effects such as calibration errors, instrumental imperfections, or inaccuracies in the forward model. The inverse problem consists in estimating X from the measurements Y given partial or imperfect knowledge of both H and N.

Several classes of inverse problems commonly arise in astrophysics. In the following, we briefly describe the most relevant ones for the purposes of this thesis.

5.2.1 Deconvolution

When the observing system is linear and shift invariant, the relationship between the measured data Y and the unknown signal X reduces to a convolution:

$$Y = HX = X * h, \tag{5.16}$$

where h is the point spread function (PSF) of the instrument. The inverse problem, known as *deconvolution*, consists in recovering X from the blurred observations Y.

Deconvolution problems are intrinsically ill-posed, as the PSF typically acts as a low-pass filter, suppressing high-frequency information. In the presence of noise, direct inversion leads to severe amplification of measurement errors, making regularization indispensable.

5.2.2 Radio-Interferometric Image Reconstruction

Radio-interferometric imaging constitutes a particular and especially challenging instance of a deconvolution problem. In this case, the observed data correspond to a sparse and incomplete sampling of the Fourier transform of the sky brightness distribution X. The forward operator H maps the image domain to the visibility domain through a Fourier transform followed by a sampling operation.

The corresponding PSF, commonly referred to as the *dirty beam*, is non-compact in the image domain and exhibits strong sidelobes. Moreover, its Fourier transform contains large regions of zero response, implying that entire spatial frequency bands are unobserved. These properties make radio-interferometric image reconstruction a severely ill-posed inverse problem, requiring sophisticated regularization strategies.

This problem constitutes the primary focus of the present thesis and will be discussed in detail in the following sections.

5.2.3 Blind Source Separation and Detection

Another important class of inverse problems in astrophysics is *blind source separation*, also referred to as signal detection, source detection, or image segmentation, depending on the application. In this framework, each observation is modeled as a linear mixture of multiple latent source processes:

$$Y = AS + N, \tag{5.17}$$

where S denotes the set of unknown source signals and A the mixing matrix. Different measurements correspond to different mixtures of the same underlying sources.

The goal of blind source separation techniques is to recover the original source components using minimal prior information about A and S. This class of problems is inherently ill-posed and requires the introduction of strong assumptions, such as statistical independence, sparsity, or non-negativity of the sources.

5.2.4 Regularization Principles

As discussed previously, inverse problems in astrophysics are generally ill-posed and do not admit stable solutions without additional constraints. Regularization provides a systematic framework to overcome these difficulties by incorporating prior information about both the observation process and the expected properties of the solution.

Typical forms of a priori knowledge include non-negativity of intensities, spatial smoothness, sparsity in a suitable transform domain, or physical constraints derived from the underlying astrophysical model. Although numerous regularization techniques have been developed, they are all rooted in the same fundamental principle: among all models consistent with the data, the preferred solution is the simplest one capable of explaining the observations.

In the remainder of this chapter, we focus on radio-interferometric image reconstruction. We first provide a detailed formulation of the problem and subsequently review the traditional and state-of-the-art approaches developed for its solution.

5.3 Introduction to Radio Astronomy and the Radio-Interferometric Image Reconstruction Problem

Radio astronomy is the study of astronomical sources through their emission in the radio portion of the electromagnetic spectrum. Although no unique and universally accepted definition of the radio regime exists, radio astronomical observations typically span frequencies from approximately 10 MHz to 1 THz. This range is not arbitrary but is instead determined by the interaction of electromagnetic radiation with the Earth's atmosphere.

Radio waves with frequencies between roughly 30 MHz and 50 GHz can propagate through the atmosphere with relatively minor attenuation. Below the characteristic plasma frequency of the ionosphere, approximately 10 MHz, radio waves are reflected back into space, while at frequencies above $\sim$ 50 GHz strong absorption features due to water vapor in the troposphere become dominant. Low-frequency observations can therefore only be carried out from space, whereas high-frequency radio and submillimeter observations are preferentially performed from high-altitude, arid sites on Earth. A prominent example is the Atacama Large Millimeter/submillimeter Array (ALMA), located in the Atacama Desert in Chile [103].

The wide frequency range accessible to radio astronomy enables the study of a broad variety of physical emission mechanisms, including synchrotron radiation, atomic and molecular line transitions, thermal blackbody emission, free–free radiation, and inverse Compton scattering. As a result, radio observations probe an extensive range of astrophysical objects and environments, such as radio galaxies, active galactic nuclei, quasars, the interstellar and intergalactic medium, the Galactic center, supernova remnants, compact objects, stars, planets, and even potential signatures of extraterrestrial intelligence. This diversity highlights the importance of developing robust and efficient data reduction and image recon-

struction techniques in the radio domain.

5.3.1 Single-Dish Radio Telescopes and Angular Resolution

Radio waves are detected by antennas, which can be implemented either as dipoles or as parabolic reflectors (dishes). Most modern radio telescopes employ parabolic dishes, which focus incoming electromagnetic radiation onto a receiver and provide high efficiency over broad frequency ranges. A radio dish acts as an electromagnetic transducer, converting incident electromagnetic waves into measurable electrical signals.

The region of the sky to which an antenna is sensitive is known as the field of view (FoV). The directional sensitivity of the antenna across the FoV is described by its gain pattern, or beam. For a parabolic dish, the beam consists of a dominant main lobe, which determines the angular resolution of the instrument, and a series of sidelobes arising from diffraction and interference effects.

At a given observing wavelength λ, the angular resolution of a circular aperture with diameter D is approximately given by the Rayleigh criterion:

$$\theta = 1.22\,\frac{\lambda}{D}, \tag{5.18}$$

where θ is the smallest angular separation (in radians) at which two point-like sources can be distinguished. This relation illustrates that achieving high angular resolution at long wavelengths would require impractically large dishes. For instance, resolving a source at $\lambda = 1$ m with an angular resolution of 1 arcsec would require a dish diameter of approximately 250 km.

At the time of writing, the largest single-dish radio telescope is the Five-hundred-meter Aperture Spherical Telescope (FAST) in China, with an effective diameter of 500 m. While such instruments provide excellent sensitivity, their angular resolution remains fundamentally limited.

5.3.2 Radio Interferometry and Aperture Synthesis

To overcome the resolution limitations of single-dish telescopes, radio astronomers have, since the 1960s, developed interferometric techniques that combine signals from multiple antennas. A radio interferometer consists of an array of antennas observing the same source simultaneously at the same frequency ν. The imaging technique enabled by such arrays is known as aperture synthesis.

Due to the finite separation between antennas, the incoming wavefront reaches each antenna at slightly different times. This geometric delay introduces a phase difference between the received signals, which depends on the baseline vector $\mathbf{B}$ connecting the antennas. To first order, the angular resolution of an interferometer is obtained by replacing the dish diameter D in Eq. (5.18) with the maximum baseline length B.

After compensating for geometric delays, the signals from each antenna pair are correlated, averaged in time and frequency, and stored. Each antenna pair produces an interference fringe pattern modulated by the individual antenna beams,

and the collection of all baselines defines the interferometer point spread function (PSF).

Interferometers offer considerable flexibility compared to single-dish telescopes. By changing antenna configurations or exploiting Earth rotation synthesis, different baseline lengths can be sampled, allowing sensitivity to both compact and extended structures.

5.3.3 The Radio Interferometric Measurement Equation

Consider a pair of antennas observing a distant astronomical source, such that the incoming wavefronts can be approximated as planar. The electric fields measured at the two antennas at frequency ν can be written as

$$E_1(t) = E_0(t, \hat{\mathbf{x}}) \, \exp[2\pi i \, (\nu t + \mathbf{x} \cdot \hat{\mathbf{x}})] \, , \tag{5.19}$$

$$E_2(t) = E_0(t, \hat{\mathbf{x}}) \, \exp[2\pi i \, (\nu t + (\mathbf{x} - \mathbf{b}) \cdot \hat{\mathbf{x}})] \, , \tag{5.20}$$

where $\hat{\mathbf{x}}$ denotes the direction of arrival, $\mathbf{b}$ is the baseline vector between the antennas, and $\mathbf{x}$ is a reference position.

Since radiation is received from the entire FoV, the measured signals are obtained by integrating over all sky directions:

$$E_1(t) = \int E_0(t, \hat{\mathbf{x}}) \, \exp[2\pi i(\nu t + \mathbf{x} \cdot \hat{\mathbf{x}})] \, d^2\hat{\mathbf{x}}, \tag{5.21}$$

$$E_2(t) = \int E_0(t, \hat{\mathbf{x}}) \, \exp[2\pi i(\nu t + (\mathbf{x} - \mathbf{b}) \cdot \hat{\mathbf{x}})] \, d^2\hat{\mathbf{x}}. \tag{5.22}$$

The interferometric observable is the time-averaged cross-correlation of the two signals,

$$W_{12} = \frac{1}{T} \int \overline{E_1(t)} \, E_2(t) \, dt, \tag{5.23}$$

where the overline denotes complex conjugation. Assuming the incoming radiation is spatially incoherent, as is the case for astronomical sources, the visibility can be written as

$$W(\mathbf{u}) = \int I(\hat{\mathbf{x}}) \, \exp[2\pi i \, \mathbf{u} \cdot \hat{\mathbf{x}}] \, d^2\hat{\mathbf{x}}, \tag{5.24}$$

where $\mathbf{u} = \mathbf{b}/\lambda$ is the baseline vector expressed in units of wavelength, and $I(\hat{\mathbf{x}}) = |E_0(\hat{\mathbf{x}})|^2$ is the sky brightness distribution.

Equation (5.24) is known as the *radio interferometric measurement equation*. It formally resembles a Fourier transform and defines the visibility space, or (u, v)-plane. Inverting this equation corresponds to solving a Fredholm integral equation of the first kind and therefore constitutes an ill-posed inverse problem.

5.3.4 The Dirty Image and the Deconvolution Problem

In practice, visibilities are only measured at discrete points in the (u, v)-plane corresponding to the available baselines[104]. Due to the finite number of antennas and their geometric constraints, the visibility space is only partially sampled. The unmeasured regions introduce a sampling function that renders the inverse problem ill-posed.

If the Fourier transform of the sampled visibilities is computed directly, the resulting image, known as the *dirty image*, is related to the true sky brightness distribution $T(l, m)$ through

$$I_D(l, m) = T(l, m) * G(l, m), \tag{5.25}$$

where $G(l, m)$ is the synthesized beam, or dirty beam, defined as the Fourier transform of the sampling and weighting function applied in visibility space.

The dirty beam is generally non-compact and exhibits strong sidelobes, which introduce oscillatory artifacts around bright sources and obscure faint structures. Consequently, the dirty image is not suitable for scientific analysis.

5.3.5 Interferometric Image Reconstruction

Recovering the true sky brightness distribution from interferometric data requires solving a severely ill-posed deconvolution problem. Simple linear filtering techniques, such as smoothing the dirty image, suppress high-frequency noise but also erase fine-scale astrophysical information and cannot recover unmeasured spatial frequencies.

As a result, radio interferometric imaging relies on nonlinear, iterative reconstruction algorithms that incorporate prior information about the sky. The problem can be formulated as the solution of a linear system

$$\mathbf{V} = \mathbf{F}\,\mathbf{S}_{dd}\,\mathbf{T}, \tag{5.26}$$

where $\mathbf{V}$ denotes the calibrated visibilities, $\mathbf{F}$ the Fourier transform operator, $\mathbf{S}_{dd}$ a sampling operator possibly including direction-dependent effects, and $\mathbf{T}$ a parameterization of the sky brightness.

Given the ill-posed nature of the problem, reconstruction algorithms proceed iteratively, alternating between the image and Fourier domains. At each iteration, a sky model is used to predict visibilities, residuals are computed, and a residual image is formed. The agreement with the data is typically quantified through a statistical metric such as a χ^2.

The reconstruction process is commonly divided into major cycles, which update the model visibilities, and minor cycles, which perform deconvolution in the image domain. Different algorithms correspond to different choices of sky parameterization and priors, enabling flexibility in how astrophysical structures are modeled.
The development and analysis of such reconstruction techniques form the central focus of this thesis.

5.4 Astrostatistics and Astroinformatics for Next-Generation ALMA Imaging

The ill-posedness of the radio-interferometric imaging problem is probably the most important motivation to come back to synthesis imaging. As already explained in earlier sections, classical imaging techniques will not be able to provide trustworthy results in the case of incomplete (u, v)-coverage, noise and instrument

effects. Since the ALMA telescope will have increasingly better and better sensitivity with the foreseen upgrades in the near future, it is necessary to investigate data analysis techniques that are inherently statistical (and data-driven when possible) and not relying only on deconvolution algorithms.
To address this, an internal ESO ALMA development study, *BRAIN*, was started, taking on synthesis imaging through astrostatistics and astroinformatics. The latter are disciplines that combine observational astronomy with statistical inference, algorithmic development, and data science [7], providing formal concepts and tools for tackling ill-posed inverse problems in the face of large dimensionality, sparse sampling, and noise.
In the context of the BRAIN project, we used ALMA-calibrated science data sets observed under different array configurations and observational settings, with the objective of increasing the efficiency of operations and the quality of the science data. We considered two distinct, yet synergic, perspectives: on the one hand, the probabilistic modeling approach of *RESOLVE*; on the other hand, the learning-based reconstruction technique of *DeepFocus* [105, 106]. We showed how these novel image reconstruction schemes can be used to generate better quality and more informative data products for the ALMA Science Archive, as well as to speed-up image processing in the operational pipelines.
These methods have the potential to affect current and future imaging pipelines, while also offering a roadmap to handling the expected order of magnitude increases in both data volume and rate that are expected with forthcoming electronic upgrades. In addition, the BRAIN study has developed and released a new version of an enhanced ALMA simulator, dubbed *ALMASim*, for use in algorithm development, testing, and training in the synthesis imaging community.

5.4.1 ALMA2030 and the Data Deluge Challenge

ALMA has been one of the leading facilities in millimeter and submillimeter astronomy since its start in 2011, delivering a wealth of breakthrough results in all fields of astrophysics. Building on this success, To further increase ALMA's scientific capabilities and competitiveness, an ambitious plan was published for future developments and improvements of ALMA up to 2030, known as the ALMA Development Roadmap (hereafter referred to as ALMA2030)[107, 108].
It highlights several main directions for improving the receiver bandwidth, sensitivity, baseline length, and the ALMA Science Archive. Scientifically, the scientific drivers for these developments are the observational studies necessary for addressing some of the big questions: when and how did galaxies form, how did the complexity of chemistry develop, and how did planetary systems form. In parallel, ALMA is expanding its frequency scope from 35–950 GHz with new receiver bands to cover a wider variety of astrophysical tracers.
WSU is one of the key components of the ALMA2030 roadmap, which proposes to double or quadruple the present IF bandwidth (8 GHz) of ALMA. Along with the receivers, electronics, and 2nd generation correlator, this will potentially realize a factor of 10 improvement in speed of observing. Considering that 1.7 PB of visibility data have been generated with ALMA by Cycle 9, it is foreseen that this data volume may be raised by a factor of 100.
All of these enhancements will have significant implications for both continuum

and spectral-line science. While the increased bandwidth and sensitivity will allow for the observation of intrinsically weaker sources and lines, it will also make continuum subtraction, line identification, and calibration issues more critical. In the transition periods between upgrades, observations might need to be performed with a smaller number of antennas, which will result in poorer (u, v)-coverage and make imaging even more challenging.

5.4.2 Implications for Inverse Problems and Imaging Algorithms

The expected jump in data rate, volume, and complexity will require commensurate advances in imaging and data processing. Current deconvolution algorithms are based on ad hoc assumptions and insufficiently informed prior knowledge. They are ill-equipped to cope with these challenging conditions, notably in terms of making the most of the available data information and delivering sufficient uncertainty quantification.
The aforementioned difficulties can be overcome by using astrostatistical tools, in particular those related to Bayesian inference and probabilistic modeling, for inference that accounts for noise, instrumentation, and prior information. Astroinformatics tools such as machine learning and deep learning can be used to build probabilistic models to extract patterns and correlations from the data without physical understanding, taking advantage of the computational capabilities of recent computer hardware such as graphics processing units (GPUs).
The data-science inspired image reconstruction methods, that are being explored in the BRAIN, incorporate these two elements. By exploiting a learned data base or library, obtained from ALMA observations or simulations, these methods are able to (a) adjust to changing observing conditions, (b) handle missing information in the visibility space, and (c) incorporate a data-driven regularization of the ill-posed inverse problem. Last but not least, they also have the potential to achieve, at least, real or near-real time processing, that will be required to handle the data-rates that will be introduced by ALMA2030.
In this subsection, we describe these two methods in the following order: casting radio-interferometric imaging into a statistical inference problem and, secondly, the practical details and empirical success of these learned reconstructions. Both offer a complementary, algorithmic perspective on a coherent road map towards the future of synthesis imaging for next-generation radio astronomy.

5.5 Methods and Results: Artificial Intelligence for Synthesis Imaging with BRAIN

The ESO internal ALMA development study *BRAIN* (Bayesian Reconstruction through Adaptive Image Notion)[7] started in late 2019 and concluded at the end of 2023. Its goal is to find and study alternatives to the classical synthesis imaging and to evaluate their suitability within the framework of the ALMA2030 vision. Specifically, the study considers methods and tools from the astrostatistics and astroinformatics domains, which nowadays are already an unavoidable part of the solution of the modern astronomical inverse problems, due to their

dimensionality and ill-posedness [105].
While astrostatistics is the science of quantitative inference, probabilistic modeling and uncertainty quantification of astro/nuclear/particle physics and cosmology from incomplete and noisy data, astroinformatics deals with massive, complex and distributed data sets, often requiring high performance computing and networking resources. Both of these discipline areas are essential for scientific discovery, but also play a key role in the operations of large observatories, such as real time data quality evaluation and pipeline imaging.
As part of the BRAIN project, two frameworks were implemented and tested on ALMA calibrated science data: the IFT framework *RESOLVE* [109, 110, 111] and the deep learning framework *DeepFocus* [106]. In addition, a companion simulation framework, *ALMASim*, has been created to support the training, testing, and benchmarking of machine learning algorithms for radio-interferometric imaging.

5.5.1 RESOLVE: Bayesian Imaging via Information Field Theory

RESOLVE operates directly on calibrated ALMA measurement sets in the visibility domain. The observed data vector $\mathbf{d}$ is modeled as

$$\mathbf{d} = \mathbf{R}\, e^{\mathbf{s}} + \mathbf{n}, \tag{5.27}$$

where $\mathbf{s}$ represents the logarithmic sky brightness field, $\mathbf{R}$ denotes the instrument response operator (including Fourier sampling and the dirty beam), and $\mathbf{n}$ accounts for both stochastic and systematic noise contributions.
Image reconstruction is formulated as a Bayesian inference problem, where the posterior probability density function of the sky signal is given by

$$P(\mathbf{s}|\mathbf{d}) = \frac{\exp[-H(\mathbf{s},\mathbf{d})]}{Z(\mathbf{d})}, \tag{5.28}$$

with $H(\mathbf{s},\mathbf{d})$ being the information Hamiltonian and $Z(\mathbf{d})$ the partition function. Variational inference techniques are employed to approximate this posterior efficiently in high-dimensional signal spaces.
The output of RESOLVE consists of posterior samples of both the reconstructed sky image and its power spectrum. From these samples, summary statistics such as the posterior mean image and pixel-wise uncertainty maps can be derived, providing quantitative uncertainty estimates that are absent in traditional deconvolution approaches.
An illustrative example is shown in Fig. 5.1 for the protoplanetary disk Elias 27 from the ALMA Disk Substructures at High Angular Resolution Project (DSHARP) [112, 113].
Compared to the fiducial CASA-based imaging products, RESOLVE delivers enhanced structural fidelity and explicit uncertainty quantification. Current developments extend RESOLVE to additional science cases, including self-calibration, full-Stokes polarization imaging, single-dish reconstruction, and very long baseline interferometry (VLBI) applications such as time-resolved imaging of black hole environments [114, 115]. These capabilities are expected to be further adapted to ALMA datasets.

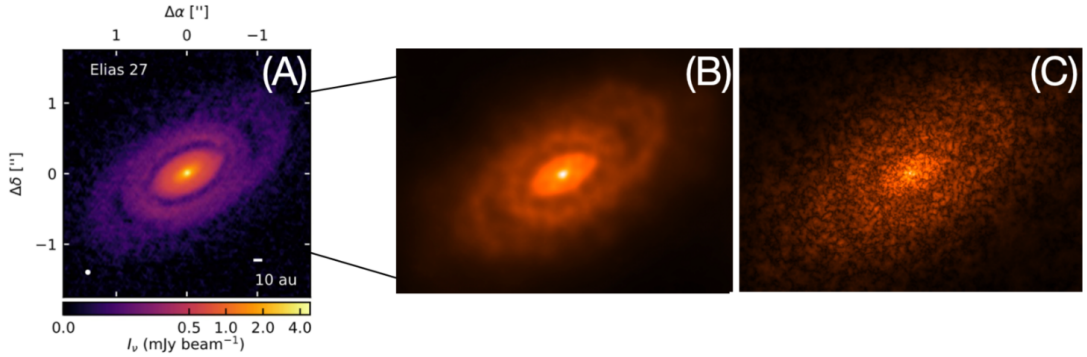

Figure 5.1: Application of the RESOLVE algorithm to the protoplanetary disk Elias 27 from the DSHARP ALMA survey at 240 GHz (1.25 mm) continuum. **(A)** Fiducial self-calibrated image released by the DSHARP collaboration [113]. **(B)** RESOLVE posterior mean sky map of Elias 27. **(C)** RESOLVE posterior uncertainty map. Reproduced from [7] under the Creative Commons Attribution (CC BY 4.0) license.

5.5.2 DeepFocus: Learning-Based Synthesis Imaging

DeepFocus is a deep learning pipeline designed to perform deconvolution, denoising, source detection, and characterization in radio-interferometric images [105, 106]. Initially developed for the detection of faint compact sources, the framework has been extended to handle extended emission through the introduction of a meta-learning strategy.

The meta-learner explores multiple neural network architectures, including convolutional autoencoders (CAE), variational autoencoders (VAE), U-Net, and ResNet models. Architecture selection and hyperparameter tuning are performed using Bayesian optimization, enabling automated adaptation to different imaging tasks and data characteristics.

The optimization procedure relies on surrogate modeling via Gaussian processes,

$$y \sim \mathcal{GP}(\mu(x), \Sigma(x, x')), \tag{5.29}$$

where μ and Σ denote the mean and covariance functions, respectively. A Matérn kernel is adopted to model smoothness and correlations in the parameter space:

$$\Sigma(x, x') = \sigma^2 \left(1 + \frac{(x - x')^2}{2\alpha l}\right)^{-\alpha}, \tag{5.30}$$

with α controlling smoothness and l the characteristic length scale.

An expected improvement (EI) acquisition function guides the exploration–exploitation trade-off,

$$EI(x) = \mathbb{E}\left[\max\left(f(x) - f(x^\star), 0\right)\right], \tag{5.31}$$

where $f(x^\star)$ is the best observed performance so far. This strategy minimizes the number of expensive deep learning model evaluations while converging toward optimal configurations.

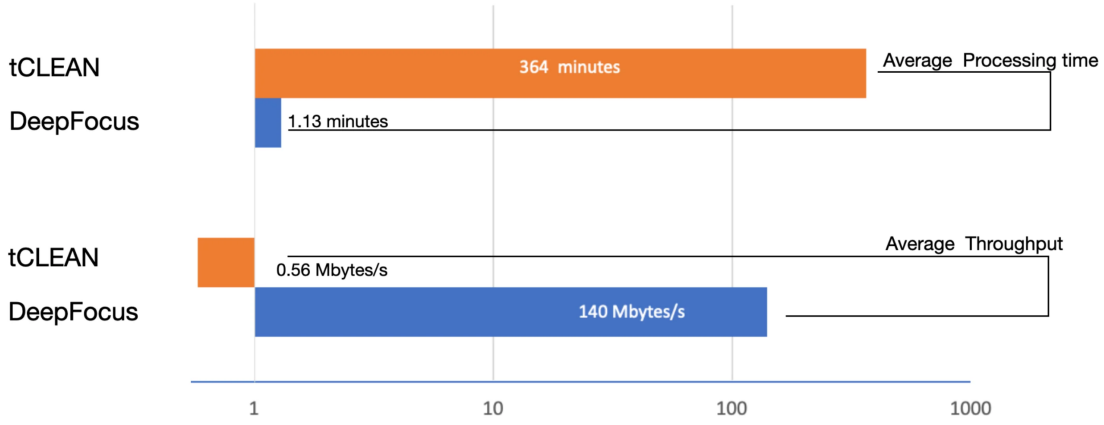

Figure 5.2: Comparison of processing time and computing throughput with tCLEAN and DeepFocus on 29×103 archived cube data from cycles 7, 8, and 9. This represents a rough estimate because at this stage of development, it is challenging to make a robust comparison between the two techniques. Reproduced from [7] under the Creative Commons Attribution (CC BY 4.0) license.

5.5.3 Benchmarking on Archived ALMA Data

DeepFocus was benchmarked on 2.9×10^4 archived ALMA data cubes from Cycles 7–9 using the European ALMA Regional Centre (EU ARC) computing cluster. The average processing time was 1.13 minutes per cube, corresponding to a throughput of approximately 140 MB s^{-1}.
To put these numbers in context, we also ran CASA's `tCLEAN` on the same data with `niter`=1000. The average throughput was 0.56 MB s^{-1}, about 100 times slower than DeepFocus. We find that DeepFocus can be anywhere from 280 to more than 5000 times faster than classical deconvolution (see Fig. 5.2). This makes it possible to use GPU-accelerated machine learning for large-scale synthesis imaging.

5.5.4 ALMASim: A Refined ALMA Simulation Framework

To support supervised learning workflows and algorithm validation, the ALMASim package was developed as an extension of the CASA simulator [116]. ALMASim integrates additional software components, including MARTINI [117], Illustris Python tools [118], and the Numerical Information Field Theory (NIFTy) framework [119].
ALMASim enables the generation of realistic ALMA observations across the full range of array configurations, frequencies, and source morphologies. Supported sky models include point-like, Gaussian, diffuse, and extended emission, with diffuse structures generated using information field theory techniques (see Fig. 5.3).
The output products include noise-free sky models, dirty image cubes, dirty beams, measurement sets, and diagnostic plots, all provided in standard FITS formats.
The package is designed for high-performance computing environments and sup-

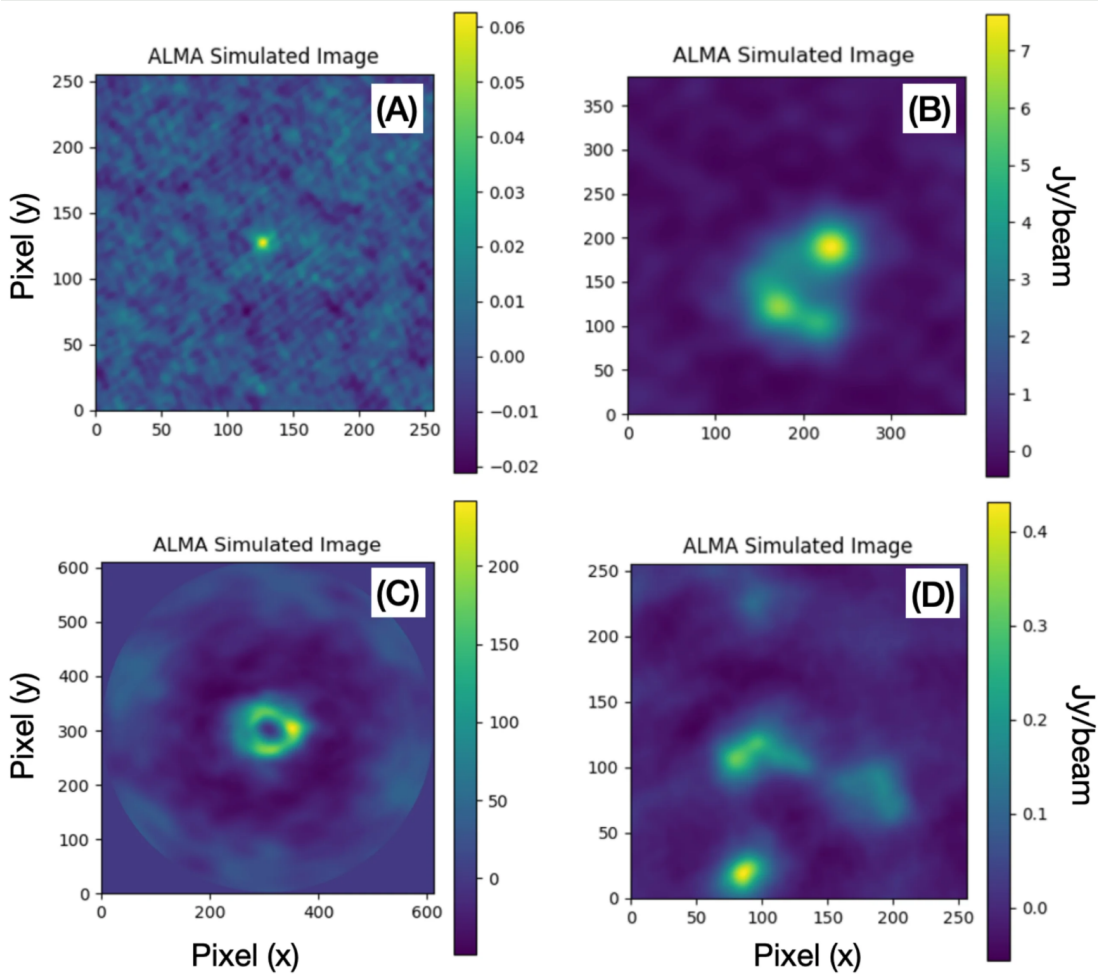

Figure 5.3: Examples of ALMA-simulated dirty images generated with the ALMASim package. **(A)** Point-like source. **(B)** Gaussian-shaped emission. **(C)** Extended emission. **(D)** Diffuse emission. Reproduced from [7] under the Creative Commons Attribution (CC BY 4.0) license.

ports MPI-based parallelization, allowing the efficient generation of thousands of simulated datasets. Both deterministic and fully randomized observation configurations are supported.

5.5.5 Empirical Noise Modeling

To enhance realism, ALMASim incorporates both synthetic and empirical noise modeling. Synthetic noise can be injected at the visibility level using the CASA simulator, including thermal noise, atmospheric attenuation, gain fluctuations, and polarization leakage. However, such models may not fully capture correlated or structured noise patterns.
An empirical approach was therefore developed to extract noise characteristics directly from real ALMA images. High-frequency spatial noise components are isolated from observed data and statistically characterized, preserving pixel-wise flux distributions while excluding astrophysical signal regions. These empirically derived noise patterns are then injected into simulated sky models, yielding synthetic images that closely resemble real ALMA observations. In Fig. 5.4, a simulated ALMA image characterized by noise is shown and extracted from a real ALMA image.
This hybrid noise modeling strategy significantly improves the fidelity of simulated datasets, making ALMASim a powerful tool for training, validating, and stress-testing advanced synthesis imaging algorithms.

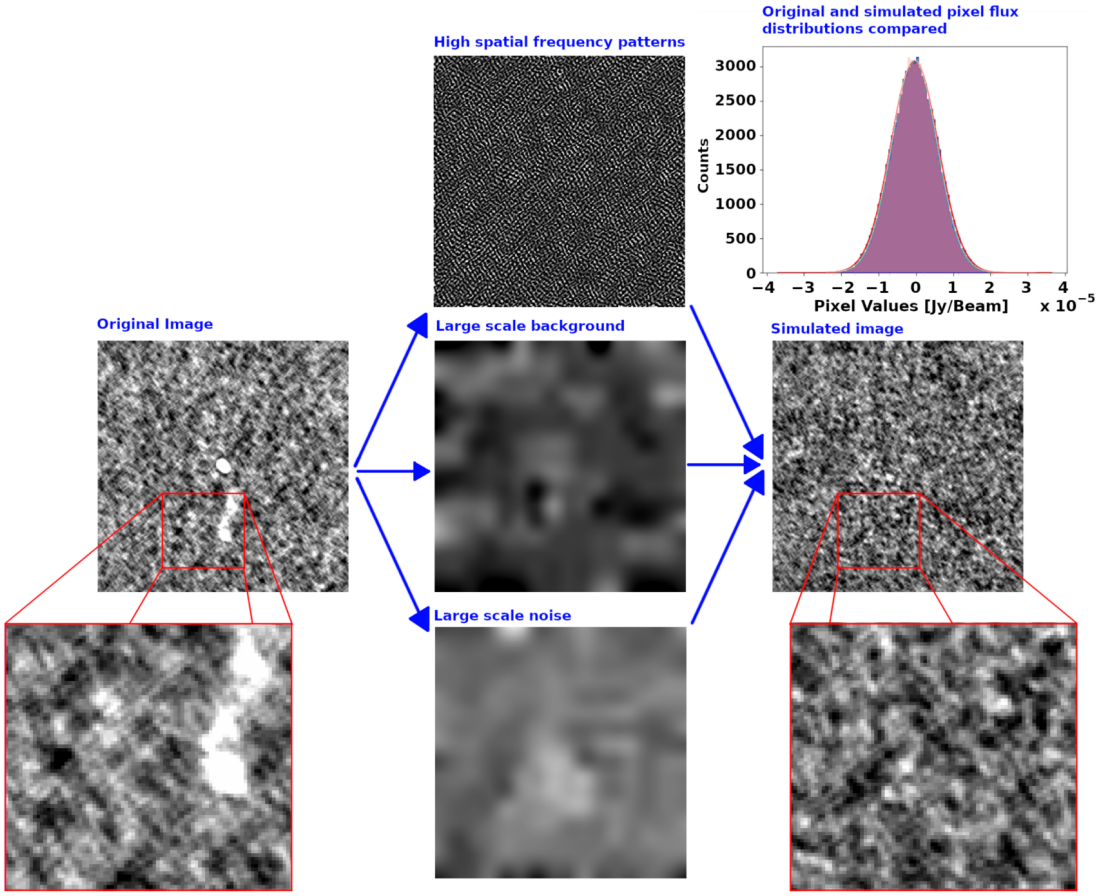

Figure 5.4: Simplified visual explanation of the empirical approach to noise modeling. Different background and noise components measured at scales larger and shorter than the typical beam scale (central panels) are isolated from a real ALMA image (e.g., an ALMA calibrator (left panel)) and then added to a simulated image (right panel). In this example, we considered local fluctuations (center, bottom panel), a large-scale background (central panel), and high spatial frequency patterns (center, top panel). Instead of using a theoretical model to simulate noise and instrumental response, the same effects were directly measured from real observations obtained in comparable situations (telescope configuration and atmospheric conditions). Reproduced from [7] under the Creative Commons Attribution (CC BY 4.0) license.

5.6 Outlook and Conclusions

The ALMA2030 roadmap will introduce revolutionary changes to the observatory, most of which are due to the large data rates and data volumes. In this framework, the ESO internal ALMA development study *BRAIN* has carried out a comprehensive investigation of advanced imaging techniques for tackling the next decade's challenges in synthesis imaging. Two techniques tested within this study, namely *RESOLVE* and *DeepFocus*, show encouraging performances in terms of scalability to large data and reduced human intervention.
RESOLVE is characterized by its ability to perform statistically robust diffuse source separation and long-tailed emission deconvolution. Its Bayesian frame-

work inherently allows for the simultaneous analysis of several measurement sets, which makes it a key tool for *group-level imaging* of data from different ALMA instruments (12 m, 7 m and total power), which is necessary to be sensitive to a wide range of angular resolutions in a single reconstruction. These features make RESOLVE the go-to method in the field. Current efforts are towards improving the computational time required for large ALMA data cubes.
In comparison, DeepFocus is better tailored to the operational needs. It focuses on speed, real-time usage, and processing of multi-spectral data cubes. The pipeline is evolving to identify and characterise multiple source morphologies and it is trained on newly available ALMA data. The accompanying tool, *ALMASim*, provides the simulated data to be used for the development of science-driven imaging algorithms.
While the exact design of the ALMA2030 is yet to be fully determined, it is already clear that learning-based techniques like DeepFocus could potentially be integrated into real-time processing pipelines soon after the calibration process. In this context, data arriving in the ALMA Science Archive could potentially be imaged in a matter of minutes, and provide virtually immediate feedback to operators and users. The delivery of a real-time imaging service as part of an archive-based science model is therefore already plausible.
Third, if automated imaging products are available, an alert system could in principle be employed to detect anomalies or artifacts, for example due to instrumentation problems, in an image. This alert could be sent to the system astronomers, for investigation and remediation in following observations, or to the next-generation CASA developers, for improvement of the pipeline. If the archive is the focal point for importing calibrated data, multiple imaging algorithms can be employed in a systematic way, even group imaging algorithms, if other relevant data are available.
Finally, increased efficiency of the observatory and improved quality and completeness of data products in the archive will greatly facilitate large-scale mining of data, and will be a key part of maximizing the scientific return of the ALMA observations, and enabling the discoveries in the ALMA2030 timeframe. Finally, increased efficiency of the observatory and improved quality and completeness of data products in the archive will greatly facilitate large-scale mining of data, and will be a key part of maximizing the scientific return of the ALMA observations, and enabling the discoveries in the ALMA2030 timeframe.

Chapter 6

Conclusions

We have presented a collection of concepts, techniques, and models for the design, extension, and optimization of deep learning models specifically tailored for the processing of high-dimensional data cubes, using hyperspectral imaging (HSI) as a use case. One of the main goals of all the chapters in this dissertation is to achieve deep architectures that are able to (i) retain and simultaneously take advantage of the spatial and spectral information present in the input data, (ii) be computationally efficient and scalable, and (iii) be transferable to other data, sensors, or even disciplines.
Overall, this paper addresses a key problem in contemporary scientific imaging: designing representation-learning frameworks that can be applied to high-dimensional, structured, partially redundant, and possibly indirectly measured data. In this respect, hyperspectral imaging and astrophysical synthesis imaging should not be seen as mere applications, but rather as illustrative use cases of a general computational framework that is characterised by structured datacubes and operator-induced ill-posedness.

From RGB Transformers to Hyperspectral-Native Architectures

The first contribution is to present **AMBER** [2], a specially designed SegFormer version for multi-band semantic segmentation. Unlike traditional computer vision models originally developed for 3-band RGB data, the proposed AMBER model operates on hyperspectral data cubes, eliminating the need to reduce the spectral dimensions prior to segmentation.
In contrast to many standard HSI pipelines, which reduce the dimensionality of the spectral axis using techniques such as PCA or band clustering prior to model training, the proposed 3D convolutional layers in the input end of AMBER, combined with the volumetric spectral-spatial coupling in the encoding path, enable the model to jointly learn to compress and represent the spectral axis.
This design leads to two consequences: 1) the fine-level spectral information is not lost too early, and 2) the feature learning is fully learnable, which meets the spirit of the contemporary deep learning techniques. By demonstrating experimental results on benchmark datasets, the proposed method achieves competitive, consistent segmentation performance, especially under class imbalance and

low-discriminative spectral characteristics. The promising results on the spaceborne dataset also demonstrate the suitability of the proposed architecture for this dataset, which is less ideal than the airborne benchmark datasets.

Efficiency Through Operator-Theoretic Design: AMBER-AFNO

The second contribution builds on the architecture mentioned above. The multi-head self-attention modules are replaced with Adaptive Fourier Neural Operators (AFNO)[4], yielding the model **AMBER-AFNO**[3]. Note that this replacement is not intended to reduce the number of parameters. Rather, the authors believe that representation learning can be more intrinsically linked with operator learning. The self-attention module explicitly models interactions among all pairs of input tokens; it has quadratic memory complexity with respect to the input length. For high-dimensional data cubes, this becomes prohibitive. On the other hand, AFNO enables global mixing of tokens in the frequency domain, posing itself as a spectral approximation of a global convolution operator. This leads to a linear memory complexity and a substantial reduction in the number of learnable parameters. Importantly, it retains the network's global receptive field. In experiments on hyperspectral segmentation and volumetric medical image analysis, the authors show that frequency-domain token mixing achieves a favorable trade-off between accuracy and complexity. More significantly, they point out that explicit attention is not a necessary module to model global context. Instead, frequency-domain operators provide a theoretically grounded and computationally efficient framework.

Toward a Hyperspectral Foundation Model

The third contribution is to push AMBER-AFNO into a hyperspectral foundation model, which, through massive self-supervised pretraining of globally distributed, multi-temporal hyperspectral datasets, as well as DINO-based objectives [5], shows the potential of learning transferable spectral–spatial representations.
Contrary to RGB data, hyperspectral image learning was previously constrained by data quantity, sensor variability, and computational resources. Large-scale self-supervised pre-training solves all of these problems by allowing the network to learn transferable features on a large dataset before fine-tuning for a specific task. Although this task is not yet being evaluated properly, the results show the potential of the transfer across space, instruments, and the need for fewer labels. This work is a move towards general-purpose, off-the-shelf hyperspectral backbones, much like foundation models in natural language processing and RGB computer vision. Crucially, the findings show that such hyperspectral foundation models require a particular structural model that directly accounts for the volumetric spectral-spatial dependency.

Datacubes as a Unifying Scientific Abstraction

Although the main focus of this thesis is its application to hyperspectral image segmentation, it is also useful to place the thesis's contributions in the context of other scientific imaging applications. Indeed, many similarities exist between the problem of segmenting a hyperspectral image and some astrophysical inverse problems. Radio-interferometric imaging in astronomy and spectral cube analysis are examples of inverse problems similar to the hyperspectral imaging problem in Earth observation. All these problems involve high-dimensional data, indirect measurements, undersampling, and noise.
Reconstruction in astrophysics is often performed using linear operators and a Fourier-space formulation. In addition, Bayesian inference is used for uncertainty quantification. The frequency-domain token mixing in AMBER-AFNO is, in a sense, a manifestation of this operator-theoretic perspective. Frequency-domain representations are common in inverse problems and in neural network architectures.
Hence, astrophysical use cases are not introduced as a research direction of its own, but rather as a tool to demonstrate that fast, global representation learning is a structural requirement in the imaging sciences of the 21st century. This unifying datacube is indeed a valid abstraction; it might very well be that the architectural paradigms established here will find applications in other scientific fields dominated by structured transformations.

Scientific Contributions and Future Perspectives

In conclusion, this thesis offers the following:

- A hyperspectral-native transformer architecture (AMBER[2]) preserving full spectral–spatial structure.
- An operator-inspired, memory-efficient alternative to self-attention (AMBER-AFNO[3]).
- A showcase for large-scale self-supervised pre-training of hyperspectral foundation models[6].
- A datacube unifying view of remote sensing and astrophysical inverse problems[106].

Still, some avenues remain. We may increase the number of pretraining sensors and their time spans to get even more robust representations. An interesting follow-up research is to design hybrid models that involve frequency operators for global modeling and spatial attention for local modeling. The transferability of the proposed models to other hyperspectral, medical, and astrophysical images can be systematically tested. It is also important to estimate the uncertainty of the operator-based networks.

In conclusion,

In summary, this thesis makes the main statement that for effective and generalizable representation learning from scientific data cubes, a model should be

spatial–spectral aware, operator-inspired, computationally efficient, and scalable for large-scale self-supervised learning.

This is the motivation for AMBER and AMBER-AFNO . The direct application lies in hyperspectral segmentation, but the algorithmic message is more general: there are benefits to scientific imaging with techniques that model a global representation while preserving the computational complexity. In this light, AMBER-AFNO is not just a segmentation architecture, but a template for a fast scientific datacube processor, unifying hyperspectral remote sensing, volumetric medical imaging, and operator-based astrophysical reconstruction.

The road ahead for scientific imaging is to further bridge the gap between deep learning and operator theory, and to move closer to foundation models, providing a common model for all kinds of structured, high-dimensional data.

Bibliography

[1] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers, 2021.

[2] Andrea Dosi, Massimo Brescia, Stefano Cavuoti, Mariarca D'Aniello, Michele Delli Veneri, Carlo Donadio, Adriano Ettari, Giuseppe Longo, Alvi Rownok, Luca Sannino, and Maria Zampella. Amber: advanced segformer for multi-band image segmentation—an application to hyperspectral imaging. *Neural Computing and Applications*, 37(22):17273–17291, 2025.

[3] Andrea Dosi, Semanto Mondal, Rajib Chandra Ghosh, Massimo Brescia, and Giuseppe Longo. Less is more: Amber-afno – a new benchmark for lightweight 3d medical image segmentation, 2025.

[4] John Guibas, Morteza Mardani, Zongyi Li, Andrew Tao, Anima Anandkumar, and Bryan Catanzaro. Adaptive fourier neural operators: Efficient token mixers for transformers. *ArxiV*, 11 2021.

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *CoRR*, abs/2104.14294, 2021.

[6] Nassim Ait Ali Braham, Conrad M Albrecht, Julien Mairal, Jocelyn Chanussot, Yi Wang, and Xiao Xiang Zhu. Spectralearth: Training hyperspectral foundation models at scale. *arXiv preprint arXiv:2408.08447*, 2024.

[7] Fabrizia Guglielmetti, Michele Delli Veneri, Ivano Baronchelli, Carmen Blanco, Andrea Dosi, Torsten Enßlin, Vishal Johnson, Giuseppe Longo, Jakob Roth, Felix Stoehr, Łukasz Tychoniec, and Eric Villard. A brain study to tackle image analysis with artificial intelligence in the alma 2030 era. *Physical Sciences Forum*, 9(1), 2023.

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image

is worth 16x16 words: transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2021.

[10] Muhammad Ahmad, Sidrah Shabbir, Swalpa Kumar Roy, Danfeng Hong, Xin Wu, Jing Yao, Adil Mehmood Khan, Manuel Mazzara, Salvatore Distefano, and Jocelyn Chanussot. Hyperspectral image classification—traditional to deep models: A survey for future prospects. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:968–999, 2022.

[11] Nicolas Audebert, Bertrand Le Saux, and Sebastien Lefevre. Deep learning for classification of hyperspectral data: A comparative review. *IEEE Geoscience and Remote Sensing Magazine*, 7(2):159–173, 2019.

[12] Danfeng Hong, Zhu Han, Jing Yao, Lianru Gao, Bing Zhang, Antonio Plaza, and Jocelyn Chanussot. Spectralformer: Rethinking hyperspectral image classification with transformers. *IEEE Trans. Geosci. Remote Sens.*, 60:1–15, 2022. DOI: 10.1109/TGRS.2021.3130716.

[13] H. Kaufmann, S. Förster, H. Wulf, K. Segl, L. Guanter, M. Bochow, U. Heiden, A. Müller, W. Heldens, T. Schneiderhan, et al. Science plan of the environmental mapping and analysis program (enmap). Technical report, German Aerospace Center (DLR), 2012.

[14] G. Kerr, J. Avbelj, E. Carmona, A. Eckardt, B. Gerasch, L. Graham, B. Günther, U. Heiden, D. Krutz, H. Krawczyk, et al. The hyperspectral sensor desis on muses: Processing and applications. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 268–271. IEEE, 2016.

[15] S. Pignatti, A. Palombo, S. Pascucci, F. Romano, F. Santini, T. Simoniello, A. Umberto, C. Vincenzo, N. Acito, M. Diani, et al. The PRISMA hyperspectral mission: Science activities and opportunities for agriculture and land monitoring. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 4558–4561. IEEE, 2013.

[16] M. Rast, J. Nieke, J. Adams, C. Isola, and F. Gascon. Copernicus hyperspectral imaging mission for the environment (CHIME). In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 108–111. IEEE, 2021.

[17] Qu Shenming, Li Xiang, and Gan Zhihua. A new hyperspectral image classification method based on spatial-spectral features. *Scientific Reports*, 12, 01 2022.

[18] Andrea Dosi, Michele Pesce, Anna Di Nardo, Vincenzo Pafundi, Michele Delli Veneri, Rita Chirico, Lorenzo Ammirati, Nicola Mondillo, and Giuseppe Longo. Prisma hyperspectral image segmentation with u-net convolutional neural network using singular value decomposition for mapping mining areas: Preliminary results. In Cosimo Ieracitano, Nadia Mammone, Marco Di Clemente, Mufti Mahmud, Roberto Furfaro, and Francesco Carlo

Morabito, editors, *The Use of Artificial Intelligence for Space Applications*, pages 327–340, Cham, 2023. Springer Nature Switzerland.

[19] Jian Fang, Jingxiang Yang, Abdolraheem Khader, and Liang Xiao. Mimosst: Multi-input multi-output spatial-spectral transformer for hyperspectral and multispectral image fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–20, 2024.

[20] Xuming Zhang, Yuanchao Su, Lianru Gao, Lorenzo Bruzzone, Xingfa Gu, and Qingjiu Tian. A lightweight transformer network for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–17, 2023.

[21] Muhammad Ahmad, Muhammad Hassaan Farooq Butt, Manuel Mazzara, Salvatore Distefano, Adil Mehmood Khan, and Hamad Ahmed Altuwaijri. Pyramid hierarchical spatial-spectral transformer for hyperspectral image classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17:17681–17689, 2024.

[22] Muhammad Hassaan Farooq Butt, Jian Ping Li, Muhammad Ahmad, and Muhammad Adnan Farooq Butt. Graph-infused hybrid vision transformer: Advancing geoai for enhanced land cover classification. *International Journal of Applied Earth Observation and Geoinformation*, 129:103773, 2024.

[23] Preetam Ghosh, Swalpa Kumar Roy, Bikram Koirala, Behnood Rasti, and Paul Scheunders. Hyperspectral unmixing using transformer network. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2022.

[24] Chenyu Li, Bing Zhang, Danfeng Hong, Jun Zhou, Gemine Vivone, Shutao Li, and Jocelyn Chanussot. Casformer: Cascaded transformers for fusion-aware computational hyperspectral imaging. *Information Fusion*, 108:102408, 2024.

[25] Haoyang Yu, Zhen Xu, Ke Zheng, Danfeng Hong, Hao Yang, and Meiping Song. Mstnet: A multilevel spectral–spatial transformer network for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2022.

[26] Bing Liu, Xuchu Yu, Pengqiang Zhang, Xiong Tan, Anzhu Yu, and Zhixiang Xue. A semi-supervised convolutional neural network for hyperspectral image classification. *Remote Sensing Letters*, 8(9):839–848, 2017.

[27] Zilong Zhong, Jonathan Li, Zhiming Luo, and Michael Chapman. Spectral–spatial residual network for hyperspectral image classification: A 3-d deep learning framework. *IEEE Transactions on Geoscience and Remote Sensing*, 56(2):847–858, 2018.

[28] Mercedes E. Paoletti, Juan Mario Haut, Ruben Fernandez-Beltran, Javier Plaza, Antonio J. Plaza, and Filiberto Pla. Deep pyramidal residual networks for spectral–spatial hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 57(2):740–754, 2019.

[29] Feng Zhao, Junjie Zhang, Zhe Meng, Hanqiang Liu, Zhenhui Chang, and Jiulun Fan. Multiple vision architectures-based hybrid network for hyperspectral image classification. *Expert Systems with Applications*, 234:121032, 2023.

[30] Koushikey Chhapariya, Krishna Mohan Buddhiraju, and Anil Kumar. Spectral-spatial classification of hyperspectral images with multi-level cnn. In *2022 12th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, pages 1–5, 2022.

[31] Le Sun, Guangrui Zhao, Yuhui Zheng, and Zebin Wu. Spectral–spatial feature tokenization transformer for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2022.

[32] Shaohui Mei, Chao Song, Mingyang Ma, and Fulin Xu. Hyperspectral image classification using group-aware hierarchical transformer. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2022.

[33] Xin He, Yushi Chen, and Zhouhan Lin. Spatial-spectral transformer for hyperspectral image classification. *Remote Sensing*, 13(3), 2021.

[34] Ehu.eus. Hyperspectral remote sensing scenes (2017).[online].

[35] Victor Maus, Stefan Giljum, Jakob Gutschlhofer, Dieison M da Silva, Michael Probst, Sidnei L B Gass, Sebastian Luckeneder, Mirko Lieber, and Ian McCallum. Global-scale mining polygons (Version 1), 2020.

[36] M.E. Paoletti, J.M. Haut, J. Plaza, and A. Plaza. A new deep convolutional neural network for fast hyperspectral image classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 145:120–147, 2018. Deep Learning RS Data.

[37] Hao Feng, Yongcheng Wang, Zheng Li, Ning Zhang, Yuxi Zhang, and Yunxiao Gao. Information leakage in deep learning-based hyperspectral image classification: A survey. *Remote Sensing*, 15(15), 2023.

[38] Reaya Grewal, Singara Singh Kasana, and Geeta Kasana. Hyperspectral image segmentation: a comprehensive survey. *Multimedia Tools and Applications*, 82:20819 – 20872, 2022.

[39] Ying Li, Haokui Zhang, and Qiang Shen. Spectral–spatial classification of hyperspectral imagery with 3d convolutional neural network. *Remote Sensing*, 9(1), 2017.

[40] Heming Liang and Qi Li. Hyperspectral imagery classification using sparse representations of convolutional neural network features. *Remote Sensing*, 8(2), 2016.

[41] Yanan Luo, Jie Zou, Chengfei Yao, Tao Li, and Gang Bai. Hsi-cnn: A novel convolution neural network for hyperspectral image, 2018.

[42] Mingyi He, Bo Li, and Huahui Chen. Multi-scale 3d deep convolutional neural network for hyperspectral image classification. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3904–3908, 2017.

[43] Hyungtae Lee and Heesung Kwon. Contextual deep cnn based hyperspectral classification. In *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 3322–3325, 2016.

[44] Jakub Nalepa, Michal Myller, and Michal Kawulok. Validating hyperspectral image segmentation. *IEEE Geoscience and Remote Sensing Letters*, 16(8):1264–1268, August 2019.

[45] World Health Organization. World health statistics 2024: Monitoring health for the sdgs, sustainable development goals, 2024. Accessed: 2025-08-03.

[46] Foong Way David Tai, Nicholas Wray, Reena Sidhu, Andrew Hopper, and Mark McAlindon. Factors associated with oesophagogastric cancers missed by gastroscopy: A case-control study. *Frontline Gastroenterol.*, 11, 2019.

[47] M. C. Florkow, K. Willemsen, V. V. Mascarenhas, E. H. G. Oei, M. van Stralen, and P. R. Seevinck. Magnetic resonance imaging versus computed tomography for three-dimensional bone imaging of musculoskeletal pathologies: A review. *Journal of Magnetic Resonance Imaging*, 56(1):11–34, July 2022. Epub 2022-01-19.

[48] Zongwei Zhou, Vatsal Sodha, Md Mahfuzur Rahman Siddiquee, Ruibin Feng, Nima Tajbakhsh, Michael B. Gotway, and Jianming Liang. Models genesis: Generic autodidactic models for 3d medical image analysis, 2019.

[49] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, volume 9351, pages 234–241, 10 2015.

[50] Niranjan Sathianathen, Nicholas Heller, Tejpaul, et al. Automatic segmentation of kidneys and kidney tumors: The kits19 international challenge. *Frontiers in Digital Health*, 3, 01 2022.

[51] Spyridon Bakas, Mauricio Reyes, Andras Jakab, Bauer, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *arXiv e-prints*, page 38, 03 2019.

[52] Zhenzhou Zheng, Ming Yang, Chao Wang, Wei Li, Lixin Zhou, Yun Zhang, and Jianping Li. Affu-net: Attention feature fusion u-net with hybrid loss for winter jujube crack detection. *Computers and Electronics in Agriculture*, 202:107049, 2022.

[53] H. J. Song, Y. F. Wang, S. J. Zeng, X. Y. Guo, and Z. H. Li. Oau-net: Outlined attention u-net for biomedical image segmentation. *Biomedical Signal Processing and Control*, 80:104038, 2023.

[54] Q. Y. Hou, Y. J. Zhao, H. Zhou, Y. Li, and L. Yu. Istdu-net: Infrared small-target detection u-net. *IEEE Geoscience and Remote Sensing Letters*, 2022.

[55] C. F. Lan, L. Zhang, Y. Q. Wang, and C. D. Liu. Research on improved dnn and multiresu_net network speech enhancement effect. *Multimedia Tools and Applications*, 81:26163–26184, 2022.

[56] M. Chen, Y. Zhang, L. Wang, J. Liu, and Q. Li. Sau-net: A novel network for building extraction from high-resolution remote sensing images by reconstructing fine-grained semantic features. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17:6747–6761, 2024.

[57] J. M. J. Valanarasu, Vishwanath A. Sindagi, Ilker Hacihaliloglu, and Vishal M. Patel. Kiu-net: Overcomplete convolutional architectures for biomedical image and volumetric segmentation. *IEEE Transactions on Medical Imaging*, 41:965–976, 2022.

[58] Q. Sun, Y. Zhang, H. Liu, X. Wang, and J. Li. Ucr-net: U-shaped context residual network for medical image. *Computers in Biology and Medicine*, 149:106203, 2022.

[59] M. Jafari, S. Francis, J. M. Garibaldi, and X. Chen. Lmisa: a lightweight multi-modality image segmentation network via domain adaptation using gradient magnitude and shape constraint. *Medical Image Analysis*, 80:102536, 2022.

[60] B. Li, Y. Wang, X. Zhang, J. Chen, and L. Liu. Rt-unet: an advanced network based on residual network and transformer for medical image segmentation. *International Journal of Intelligent Systems*, 37:8565–8582, 2022.

[61] Jiayu Zhang, Yihong Liu, Qiang Wu, Yaxin Wang, Yi Liu, Xiaoxu Xu, and Bo Song. Star-shaped window transformer reinforced u-net for medical image segmentation. *Computers in Biology and Medicine*, 150:105954, 2022.

[62] Zhiqin Zhu, Kun Yu, Guanqiu Qi, Baisen Cong, Yuanyuan Li, Zexin Li, and Xinbo Gao. Lightweight medical image segmentation network with multi-scale feature-guided fusion. *Computers in Biology and Medicine*, 182:109204, 2024.

[63] Zhiqin Zhu, Mengwei Sun, Guanqiu Qi, Yuanyuan Li, Xinbo Gao, and Yu Liu. Sparse dynamic volume transunet with multi-level edge fusion for brain tumor segmentation. *Computers in Biology and Medicine*, 172:108284, 2024.

[64] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, Gerard Sanroma, Sandy Napel, Steffen Petersen, Georgios Tziritas, Elias Grinias, Mahendra Khened, Varghese Alex Kollerathu, Ganapathy Krishnamurthi, Marc-Michel Rohé, Xavier

Pennec, Maxime Sermesant, Fabian Isensee, Paul Jäger, Klaus H. Maier-Hein, Peter M. Full, Ivo Wolf, Sandy Engelhardt, Christian F. Baumgartner, Lisa M. Koch, Jelmer M. Wolterink, Ivana Išgum, Yeonggul Jang, Yoonmi Hong, Jay Patravali, Shubham Jain, Olivier Humbert, and Pierre-Marc Jodoin. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE Transactions on Medical Imaging*, 37(11):2514–2525, 2018.

[65] Bennett Landman, Zhoubing Xu, J Igelsias, Martin Styner, T Langerak, and Arno Klein. Multi-atlas labeling beyond the cranial vault. *arXiv*, 2015.

[66] Bjoern H. Menze, Andras Jakab, Stefan Bauer, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, 2015.

[67] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 565–571, 2016.

[68] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2018.

[69] Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnu-net: A self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods*, 18:1–9, 2021.

[70] Ailiang Lin, Bingzhi Chen, Jiayu Xu, Zheng Zhang, Guangming Lu, and David Zhang. Ds-transunet: Dual swin transformer u-net for medical image segmentation. *IEEE Transactions on Instrumentation and Measurement*, 71:1–15, 2022.

[71] Qinyan Bai, Xiaobo Luo, Yaxu Wang, and Tengfei Wei. Dhrnet: A dual-branch hybrid reinforcement network for semantic segmentation of remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17:4176–4193, 2024.

[72] Zhenxiang He, Xiaoxia Li, Nianzu Lv, Yuling Chen, and Yong Cai. Retinal vascular segmentation network based on multi-scale adaptive feature fusion and dual-path upsampling. *IEEE Access*, 12:48057–48067, 2024.

[73] Quan Zhou, Linjie Wang, Guangwei Gao, Bin Kang, Weihua Ou, and Huimin Lu. Boundary-guided lightweight semantic segmentation with multi-scale semantic context. *IEEE Transactions on Multimedia*, 26:7887–7900, 2024.

[74] Ali Hatamizadeh, Dong Yang, Holger Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. *ArXiv*, 03 2021.

[75] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In Leonid Karlinsky, Tomer Michaeli, and Ko Nishino, editors, *Computer Vision – ECCV 2022 Workshops*, pages 205–218, Cham, 2023. Springer Nature Switzerland.

[76] Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Xiaoguang Han, Lequan Yu, Liansheng Wang, and Yizhou Yu. nnformer: Volumetric medical image segmentation via a 3d transformer. *IEEE Transactions on Image Processing*, 32:4036–4045, 2023.

[77] Abdelrahman Shaker, Muhammad Maaz, Hanoona Abdul Rasheed, Salman Khan, Ming-Hsuan Yang, and Fahad Khan. Unetr++: Delving into efficient and accurate 3d medical image segmentation. *IEEE Transactions on Medical Imaging*, PP:1–1, 05 2024.

[78] Chunhui Jiang, Yi Wang, Qingni Yuan, Pengju Qu, and Heng Li. A 3d medical image segmentation network based on gated attention blocks and dual-scale cross-attention mechanism. *Scientific Reports*, 15, 02 2025.

[79] Fahad Shamshad, Salman Khan, Syed Waqas Zamir, Muhammad Haris Khan, Munawar Hayat, Fahad Shahbaz Khan, and Huazhu Fu. Transformers in medical imaging: A survey, 2022.

[80] Kamyar Azizzadenesheli, Nikola Kovachki, Zongyi Li, Miguel Liu-Schiaffini, Jean Kossaifi, and Anima Anandkumar. Neural operators for accelerating scientific simulations and design. *Nature Reviews Physics*, 6(5):320–328, 2024.

[81] Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity, 2020.

[82] Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, David Belanger, Lucy Colwell, and Adrian Weller. Rethinking attention with performers, 2022.

[83] James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontanon. Fnet: Mixing tokens with fourier transforms, 2022.

[84] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation, 2021.

[85] Xiaohong Huang, Zhifang Deng, Dandan Li, and Xueguang Yuan. Missformer: An effective transformer for 2d medical image segmentation. *IEEE Transactions on Medical Imaging*, 42(5):1484–1494, 2023.

[86] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R. Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In Alessandro Crimi and Spyridon

Bakas, editors, *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 272–284, Cham, 2022. Springer International Publishing.

[87] Yuncong Feng, Jianyu Su, Jian Zheng, Yupeng Zheng, and Xiaoli Zhang. A parallelly contextual convolutional transformer for medical image segmentation. *Biomedical Signal Processing and Control*, 98:106674, 2024.

[88] Guoping Xu, Xuan Zhang, Xinwei He, and Xinglong Wu. Levit-unet: Make faster encoders with transformer for medical image segmentation. In *Pattern Recognition and Computer Vision*, pages 42–53, Singapore, 2024. Springer Nature Singapore.

[89] Jiazhong Cen, Jiemin Fang, Zanwei Zhou, Chen Yang, Lingxi Xie, Xiaopeng Zhang, Wei Shen, and Qi Tian. Segment anything in 3d with radiance fields, 2024.

[90] Renjie Li, Xinyi Wang, Guan Huang, Wenli Yang, Kaining Zhang, Xiaotong Gu, Son N. Tran, Saurabh Garg, Jane Alty, and Quan Bai. A comprehensive review on deep supervision: Theories and applications, 2022.

[91] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H.S. Torr, and Li Zhang. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6877–6886, 2021.

[92] Wenxuan Wang, Chen Chen, Meng Ding, Hong Yu, Sen Zha, and Jiangyun Li. Transbts: Multimodal brain tumor segmentation using transformer. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 109–119, Cham, 2021. Springer International Publishing.

[93] Yutong Xie, Jianpeng Zhang, Chunhua Shen, and Yong Xia. Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 171–180, Cham, 2021. Springer International Publishing.

[94] Dongsong Zhang, Changjian Wang, Tianhua Chen, Weidao Chen, and Yiqing Shen. Scalable swin transformer network for brain tumor segmentation from incomplete mri modalities. *Artificial Intelligence in Medicine*, 149:102788, 2024.

[95] Hsienchih Ting and Manhua Liu. Multimodal transformer of incomplete mri data for brain tumor segmentation. *IEEE Journal of Biomedical and Health Informatics*, 28(1):89–99, 2024.

[96] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024.

[97] Medical Segmentation Decathlon. *Medical Segmentation Decathlon (MSD) Challenge Rank*, 2018.

[98] D. Wang, M. Hu, Y. Jin, Y. Miao, J. Yang, Y. Xu, X. Qin, J. Ma, L. Sun, C. Li, C. Fu, H. Chen, C. Han, N. Yokoya, J. Zhang, M. Xu, L. Liu, L. Zhang, C. Wu, B. Du, D. Tao, and L. Zhang. Hypersigma: Hyperspectral intelligence comprehension foundation model. *arXiv preprint arXiv:2406.11519*, 2024.

[99] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. *CoRR*, abs/1911.05722, 2019.

[100] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. *CoRR*, abs/2111.06377, 2021.

[101] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicregl: Self-supervised learning of local visual features. In *Advances in Neural Information Processing Systems*, volume 35, pages 8799–8810, 2022.

[102] Heinz W. Engl, Otmar Scherzer, and Masahiro Yamamoto. Uniqueness and stable determination of forcing terms in linear partial differential equations with overspecified boundary data. *Inverse Problems*, 10(6):1253–1270, December 1994.

[103] Jonathan M. Marr, Ronald L. Snell, and Stanley E. Kurtz. *Fundamentals of Radio Astronomy: Observational Methods*. CRC Press, 1 edition, 2015.

[104] A. Richard Thompson, James M. Moran, and George W. Jr. Swenson. *Interferometry and Synthesis in Radio Astronomy*. Astronomy and Astrophysics Library. Springer, Cham, 3 edition, 2017.

[105] Francesco Guglielmetti, Philipp Arras, Marco Delli Veneri, Torsten Enßlin, Giuseppe Longo, Łukasz Tychoniec, and Eric Villard. Bayesian and Machine Learning Methods in the Big Data Era for Astronomical Imaging. *Physical Sciences Forum*, 5:50, 2022.

[106] Marco Delli Veneri, Łukasz Tychoniec, Francesco Guglielmetti, Giuseppe Longo, and Eric Villard. 3D detection and characterization of ALMA sources through deep learning. *Monthly Notices of the Royal Astronomical Society*, 518(3):4131–4147, 2023.

[107] John Carpenter, Dario Iono, Leonardo Testi, Nick Whyborn, and Al Wootten. The ALMA Development Roadmap. https://www.eso.org/sci/facilities/alma/developmentstudies/ALMA_Development_Roadmap_public.pdf, 2018. ESO technical document, accessed 23 November 2023.

[108] John Carpenter, Dario Iono, Fred Kemper, and Al Wootten. The ALMA Development Program: Roadmap to 2030. *arXiv e-prints*, 2020.

[109] Henrik Junklewitz, Matthew R. Bell, Martin Selig, and Torsten A. Enßlin. RESOLVE: A new algorithm for aperture synthesis imaging of extended emission in radio astronomy. *Astronomy & Astrophysics*, 586:A76, 2016.

[110] Łukasz Tychoniec, Francesco Guglielmetti, Philipp Arras, Torsten Enßlin, and Eric Villard. Bayesian Statistics Approach to Imaging of Aperture Synthesis Data: RESOLVE Meets ALMA. *Physical Sciences Forum*, 5:52, 2022.

[111] Jakob Roth, Philipp Arras, Martin Reinecke, Richard A. Perley, Rüdiger Westermann, and Torsten A. Enßlin. Bayesian radio interferometric imaging with direction-dependent calibration. *Astronomy & Astrophysics*, 678:A177, 2023.

[112] Sean M. Andrews, Jane Huang, Laura M. Pérez, Andrea Isella, Cornelis P. Dullemond, Nicolas T. Kurtovic, Viviana V. Guzmán, John M. Carpenter, David J. Wilner, Shangjia Zhang, et al. The Disk Substructures at High Angular Resolution Project (DSHARP). I. Motivation, Sample, Calibration, and Overview. *The Astrophysical Journal Letters*, 869:L41, 2018.

[113] Jane Huang, Sean M. Andrews, Laura M. Pérez, Zhaohuan Zhu, Cornelis P. Dullemond, Andrea Isella, Myriam Benisty, Xue-Ning Bai, Tilman Birnstiel, John M. Carpenter, et al. The Disk Substructures at High Angular Resolution Project (DSHARP). III. Spiral Structures in the Millimeter Continuum of the Elias 27, IM Lup, and WaOph 6 Disks. *The Astrophysical Journal Letters*, 869:L43, 2018.

[114] Philipp Arras, Philipp Frank, Philipp Haim, Jakob Knollmüller, Reimar Leike, Martin Reinecke, and Torsten Enßlin. Variable structures in M87* from space-, time-, and frequency-resolved interferometry. *Nature Astronomy*, 6:259–269, 2022.

[115] Jakob Knollmüller, Philipp Arras, and Torsten Enßlin. Resolving Horizon-Scale Dynamics of Sagittarius A*. *arXiv e-prints*, 2023.

[116] The CASA Team. CASA: The Common Astronomy Software Applications for Radio Astronomy. *Publications of the Astronomical Society of the Pacific*, 134(1041), 2022.

[117] Kyle A. Oman. MARTINI: Mock Spatially Resolved Spectral Line Observations of Simulated Galaxies. Astrophysics Source Code Library, 2019. ascl:1911.005.

[118] The Illustris Collaboration. illustris_python: Python tools for accessing and analyzing Illustris and IllustrisTNG simulations. https://github.com/illustristng/illustris_python, 2018. Software repository, accessed 23 November 2023.

[119] The NIFTy Team. NIFTy: Numerical Information Field Theory. https://gitlab.mpcdf.mpg.de/ift/nifty, 2021. Software repository, accessed 23 November 2023.

List of Figures

List of Tables