University of California
Santa Barbara

Building Efficient Vision Models for Ecological and Earth Observation Studies

A dissertation submitted in partial satisfaction
of the requirements for the degree

Doctor of Philosophy
in
Electrical and Computer Engineering

by

Satish Kumar

Committee in charge:

Professor B.S. Manjunath, Chair
Professor Kenneth Rose
Professor ShivKumar Chandrasekaran
Professor Ramtin Pedarsani

September 2024

The Dissertation of Satish Kumar is approved.

Professor Kenneth Rose

Professor ShivKumar Chandrasekaran

Professor Ramtin Pedarsani

Professor B.S. Manjunath, Committee Chair

July, 2024

Building Efficient Vision Models for Ecological and Earth Observation Studies

by

Satish Kumar

Dedicated to my family and friends.

Acknowledgements

I would like to express my deepest and most heartfelt gratitude to all those who have supported me throughout this remarkable journey of my dissertation.

First and foremost, my sincerest appreciation goes to Professor B.S. Manjunath. His invaluable guidance, patience, and constant encouragement have been the cornerstone of my research. Professor Manjunath believed in me even when I struggled or made mistakes. His unwavering support, even in difficult situations, has been a source of strength. There were times when I messed up, and though he was understandably upset, he always stood by me and supported me through it all. His mentorship has not only shaped my academic pursuits in the fields of computer vision and machine learning but has also instilled in me the rigor and passion for conducting impactful research. I am truly grateful for the wisdom, insight, and steadfast belief he has imparted to me.

I am also immensely grateful to my dissertation committee members, Professor Kenneth Rose, Professor ShivKumar Chandrasekaran, and Professor Ramtin Pedarsani. Their thoughtful feedback and insightful suggestions have been instrumental in refining my research and broadening my perspectives. I am also thankful for the financial support provided by NSF SI2-SSI award number 1664172 and NSF award number 1934641, which were critical in enabling my Ph.D. studies. I extend special thanks to Professor Dar Roberts and Dr. Alana Ayasse for their invaluable assistance in conducting interdisciplinary research.

Collaborating with Dr. Jared Stabach and Dr. Lacey Hughey on interdisciplinary projects has been one of the most rewarding experiences of my academic career. I am deeply appreciative of Dr. Tom Bullock's contributions to our collaborative work on Brain Sciences. My heartfelt thanks also go to the members of the Vision Research Lab (VRL) – Michael Goebel, Amil Khan, Joaquin Giorgi, Po-Yu Kao, Devendra Ku-

mar, Raphael Santos, Connor Levenson, Chandrakanth Gudavalli, Shailja, Abu Saleh Mohammed Iftekhar Niloy, Bowen Zhang, Ivan Arevalo, Oytun Ulutan, Austin Mcever, Angela Zhang, Ekta Prashnani, Rahul Vishwakarma – for their unwavering support, camaraderie, and the stimulating discussions that have made this journey memorable.

I also want to extend my heartfelt thanks to Abu Saleh Mohammed Iftekhar Niloy. Throughout my PhD journey, he has been more than just a colleague; he has been my steadfast companion in every project. Together, we navigated countless challenges, fought through each submission, and even got mad at each other along the way. Yet, through all the ups and downs, we remained a strong team, pushing each other to achieve more. We've come a long way together, and I am deeply grateful for finding a brother like friend in my journey.

I owe a special debt of gratitude to Jaicy Vallapurackal and Anagha Uppal, whose meticulous reviews and corrections of every single line of my thesis ensured its clarity and precision. Their dedication and keen eye for detail have been invaluable, and I am incredibly fortunate to have had their support.

Lastly, but most importantly, I want to thank my family, especially my parents, whose love, sacrifices, and unwavering support have been my greatest source of strength and motivation. I am profoundly grateful for their belief in me. My Mom and Dad, has been my guiding star. This journey would not have been possible without them. They sacrificed everything to ensure my success, working day and night to provide me with the opportunities that have allowed me to stand where I am today. Their dedication and hard work have been the foundation upon which I have built my academic career. I would also like to express my deep gratitude to my younger brother, Bhuvnesh Yadav, who has been one of the strongest pillars in my life. Because of him, I was able to focus on my PhD work without the stress of worrying about our parents' well-being at home. Knowing that my brother was always there gave me the peace of mind to pursue my

research with full dedication. To all my friends who have stood by me throughout this journey, your support has been a beacon of light, and I cherish the moments we've shared.

Thank you all for being a part of this incredible journey. Your support has meant more to me than words can express.

Curriculum Vitæ

Satish Kumar

Education

July 2024	**Doctor of Philosophy** Electrical and Computer Engineering University of California, Santa Barbara, USA.
June 2018	**Master of Science** Electrical and Computer Engineering University of California, Santa Barbara, USA.
March 2013	**Bachelor of Technology** Electrical Engineering National Institute of Technology, Kurukshetra.

Honors & Awards

2024	Schmidt Science Fellow 2024
2024	First Place in New Venture Competition at UCSB
2024	ECE Department Dissertation Fellowship, UCSB
2024	Climate Innovation Grant Proof of Concept, UCSB
2024	Invited Subject matter expert on Climate Policy Board at Amsterdam, Netherlands
2024	Invited keynote talk at ACS Spring 2024 Division of Energy & Fuels (ENFL) on Methane detection
2023	The Ohrstrom Family Foundation, and the Smithsonian National Zoo and Conservation Biology Institute (SNZCBI) fellowship
2023	Highlight (top 2.5%) Paper at Computer Vision and Pattern Recognition (CVPR 23)
2023	UCSB Doctoral Travel Grant for CVPR 2023
2023	Conference Travel Grant, UCSB Graduate Student Association
2023-2022	Interviewed by several newspapers, magazines and articles for Methane Detection work
2019	Outstanding Teaching Assistant Award from ECE department of UC, Santa Barbara
2009-2011	Merit Scholarship for four years in B.Tech from NIT Kurukshetra

Publications and Patents

S Kumar, B Zhang, ..., B S Manjunath *WildlifeMapper: Aerial Image Analysis for Multi-Species Detection and Identification*, Computer Vision and Pattern Recognition (CVPR 2024)

S Kumar, I Arevalo, A S M Iftekhar, B S Manjunath *MethaneMapper: Spectral Absorption aware Hyperspectral Transformer for Methane Detection*, Computer Vision and Pattern Recognition acepted to (CVPR 2023) selected for **Hightlights**.

S Kumar, W Kingwill, R Mouton, W Adamczyk, R Huppertz, E Sherwin *Guided Transformer Network for Detecting Methane Emissions in Sentinel-2 Satellite Imagery*, (NeurIPS 2022) Tackling Climate Change with ML, New Orleans, December 2022.

S Kumar, C Torres, O Ulutan, A Ayasse, D Robert, B Manjunath, *Deep Remote Sensing Methods for Methane Detection in Overhead Hyperspectral Imagery*, The IEEE Winter Conference on Applications of Computer Vision (WACV 20), Aspen, Colorado, March 2020.

S Kumar, B Zhang, ASM Iftekhar *EyeMethane: Unveiling Methane Emissions with spectral Transformer-Enabled Detection*, Fall Meeting (AGU 2023) Meeting.

S Kumar, A S M Iftekhar, E Prashnani, B Manjunath *LOCL: Learning Object-Attribute Compositionality using Localization*, British Machine Vision Conference (BMVC 22) 2022.

S Kumar, R Kou, J Lempges, H Hill, T Qian, V Jayaram, *In-situ Water quality monitoring in Oil and Gas operations*, Algorithms, Technologies, and Applications for Multispectral and Hyperspectral Imaging XXIX at SPIE Defense + Commercial Sensing (2023)

R Kou, **S Kumar**, S Zhang, T Qian, V Jayaram, *Application of High-Performance Paralleled Numerical Simulation in Uncertainty Quantification of Methane Emission from Oil and Gas Operation Facilities*, Interpretation, Advancing the world of Petroleum Geoscience (AAPG)

S Kumar, S Shailja, A Caetano, A Pandey, *Scaffolding AI research projects increases self-efficacy of high school students in learning neural networks (Fundamental)*, American Society for Engineering Education (ASEE 2024), Accepted

A S M Iftekhar, Raphael Ruschel, **S Kumar**, Suya You, B S Manjunath *DDS: Decoupled Dynamic Scene-Graph Generation Network*, IEEE Transactions on Image Processing. under review

S Kumar, A Iftekhar, M Goebel, B Manjunath, *StressNet : Detecting stress from multispectral images*, The IEEE Winter Conference on Applications of Computer Vision (WACV 21) 2021.

A S M Iftekhar, **S Kumar**, R. McEver, Suya You, B Manjunath, *GTNet: Guided Transformer Network for Detecting Human-Object Interactions*, Pattern Recognition and Tracking XXXIV at SPIE commerce+ defence Program 2023.

Calvin Xia, Vikram Bhagavatula, .., Tom Bullock, **S Kumar** *StressVision : Non-Invasive Stress Detection from Thermal Videos*, International Telemetring Conference 2023.

Geffen Cooper, Vincent Benenati, .., **S Kumar** *Autonomous System for Sorting Objects at the Edge*, International Telemetring Conference 2022.

Dr. Keerti Kaushik, **S Kumar**, *COMPLIANCE MONITORING DEVICE" for dental application*, Patent no: 202011017313. Submitted to Intellectual Property India (IPI).

Experience

01/2019 - 07/2024	Graduate Student Researcher, Vision Research Lab, UCSB
09/2018 - 12/2019	Teaching Assistant, Electrical and Computer Engineering, UCSB
07/2013 - 09/2017	Lead Research Engineer, Samsung Research & Development, India
06/2020 - 09/2020	Research Intern, Pioneer Natural Resources, Texas
06/2018 - 09/2018	Research Intern, Mayachitra Inc.

Technical Skills

Python, Matlab, C++, OOP, C, Bash, PyTorch, TensorFlow, Numpy, Data Structure and Algorithms, OpenCV, Git, Linux, Docker Containers, Kubernetes, Rancher, Openstack, Django

Service

Reviewer for CVPR (*2024, 2023, 2022, 2021*), NeurIPS (*2022, 2023*), ICCV (*2023, 2021*), ECCV (*2022, 2024*), WACV (*2020, 2021, 2022, 2023, 2024*), BMVC (*2022, 2023*), TIP (*2023, 2024*), Interpretation AAPG conference (*2022, 2023*).

Abstract

Building Efficient Vision Models for Ecological and Earth Observation Studies

by

Satish Kumar

Numerous large vision models for natural images, such as SAM, Florence-2, and GPT-4, have achieved state-of-the-art (SOTA) performance, largely due to vast amounts of image and text data available online. Smaller models like EfficientSAM and CLIP have also shown the potential of achieving significant results with comparatively less data. However, real-world scientific problems, particularly in remote sensing, present unique challenges due to the complexity of the data and scarcity of annotations. These problems often require data from multiple sources, such as hyperspectral sensors on airplanes and multispectral sensors on satellites, which are expensive and time-consuming to acquire.

This dissertation addresses the key question: how can large vision models be built and trained effectively under data constraints? The proposed solution involves integrating domain-specific knowledge into large vision models, specifically vision transformers, to optimize their performance and training efficiency. By incorporating core signal processing techniques, domain-specific knowledge is encoded as prior information, guiding the feature extraction process and refining randomly initialized queries via a query refiner module. This approach accelerates convergence with limited training data.

Three key applications are explored: (1) methane detection in remote sensing from aerial imagery, (2) animal detection and classification in large grasslands for ecological studies, and (3) estimation of physiological signals such as ECG and ISTI for stress assessment in biomedical contexts.

This research establishes an optimal methodology for embedding domain-specific

knowledge into deep learning models, thereby enhancing performance in data-limited environments. It provides valuable insights for improving the applicability of vision transformer-based models across various domains, contributing to advancements in computer vision research and its practical real-world applications.

Contents

Curriculum Vitae **viii**

Abstract **xi**

List of Figures **xv**

List of Tables **xx**

1 Introduction **1**
- 1.1 Motivation . . . 2
- 1.2 Challenges . . . 3
- 1.3 Summary of Contributions . . . 5
- 1.4 Organization of Thesis . . . 6

2 MethaneMapper **11**
- 2.1 Introduction . . . 12
- 2.2 Related Works . . . 14
- 2.3 Hyperspectral Mask-RCNN (H-MRCNN) . . . 17
- 2.4 MethaneMapper . . . 19
- 2.5 MethaneMapper (MM) Architecture . . . 21
- 2.6 Methane Hot Spots (MHS) dataset . . . 27
- 2.7 Experimental settings . . . 31
- 2.8 Results . . . 33
- 2.9 Conclusion . . . 38

3 Methane SatelliteMapper **39**
- 3.1 Introduction . . . 40
- 3.2 Related Work . . . 42
- 3.3 Approach . . . 44
- 3.4 Training and Inference . . . 48
- 3.5 Results . . . 50
- 3.6 Conclusion . . . 52

4 WildlifeMapper **53**
4.1 Introduction . . . 54
4.2 Related Works . . . 57
4.3 Mara-Wildlife Dataset . . . 59
4.4 WildlifeMapper Architecture . . . 62
4.5 Experiments . . . 68
4.6 Results . . . 69
4.7 Conclusion . . . 73

5 StressNet **77**
5.1 Introduction . . . 78
5.2 Related Works . . . 81
5.3 Method . . . 83
5.4 Experiments . . . 90
5.5 Results . . . 93
5.6 Conclusion . . . 98

6 Discussions and Future Work **100**
6.1 Conclusions . . . 100
6.2 Future Work . . . 101
6.3 Long-term Impact and Vision . . . 105

A MethaneMapper Appendix **107**
A.1 Introduction . . . 107

B WildlifeMapper Appendix **116**
B.1 Introduction . . . 116

Bibliography **122**

List of Figures

1.1 *Data collection processing using aerial imagery and detection of methane plumes* 7

1.2 *Data collection processing using aerial imagery and detection of wild animals. The bottom image is a zoomed in version of the aerial imagery. The Green boxes represents cattle and the Red boxes represents shoats (sheep and goats).* 9

1.3 *Representation of the Cold Pressor Test (CPT) for StressNet, Left image shows video recording of the face of person, Right image shows physiological signal (ECG) collection* 10

2.1 *Representation of information in visible and beyond visible(multispectral/hyperspectral domain) domain* 12

2.2 *Biogenic and Anthroproenic sources of methane emissions* 13

2.3 *Relation between dataset A (χ_A) and dataset B (χ_B). The 432-bands data from dataset B are processed through a matched filter to yield dataset A. Detecting plumes using this information poor dataset (Dataset A) is challenging. H-mrcnn addresses this challenge by modeling terrain absorption using ensemble and decision fusion methods.* 18

2.4 *Depiction of data collection process. Each flightline is* $\sim$ 300 *km long. An array of 598 sensors records data at 1.5m/pixel spatial resolution. All flightlines are ortho-corrected. Each data-cube is of dimension* $\sim 25000\times \sim 1500 \times 432$. 21

2.5 *Overview of MethaneMapper (MM) architecture. Given a hyperspectral image, our RGB ($400nm - 700nm$) and SWIR ($2000nm - 2500nm$) band-pass filters passes a subset of channels in desired wavelength range and feed them to CNN backbones (ResNet) to extract features. These features are concatenated and fed to Transformer Encoder. Parallelly, our Spectral Feature Generator (SFG) modules takes in all channels of input image and generate methane candidates features. Next these candidates are sent to Query Refiner (QR) to refine queries. Then these queries decoded using encoded feature from Transformer Encoder. Finally each decoded query is used to predict a plume mask via Mask Prediction and, bounding box and class via FFNs (Feed Forward Network).* 22

2.6 *Sample images from MHS dataset. The colormap in black circle shows concentration maps corresponding to the plume mask shown in red. We are showing different types of leakage sources and land cover types. For better visualization, we plotted the binary mask on color image created using visible bands of hyperspectral image.* 28

2.7 *Comparison of SLF with traditional filter in SFG module. White pixels represent methane and black no-methane. Red boundary represents ground-truth plume mask. SLF module generates better CH_4 candidates* . 36

2.8 *Sample ground truths and predictions on MHS dataset. We show robustness of MethaneMapper predictions on different kind of ground terrain, rows 1 and 3 shows leakage at a refinery, row 2 shows leakage from pipeline in agricultural land, row 4 shows leakage from storage tank with concrete background.* 37

3.1 *Qualitative visualization of SFG intermediate steps and final estimation.* 41

3.2 *Overview of Methane SatelliteMapper (MSM) architecture. Given a multispectral image, our RGB (B1, B2, B3) is passed to Feature Extractor which is a CNN backbone (ResNet) to extract features. Parallely, all channels (B1-B12) of the multispectral input as passed to Spectral Feature Generator (SFG). The SFG module generates methane candidates features. Next these candidates are sent to Query Refiner (QR) to refine queries. Then these queries decoded using encoded feature from Transformer Encoder. Finally each decoded query is used to predict a plume mask via Mask Prediction and, bounding box and class via FFNs (Feed Forward Network* . . 45

3.3 *Sites of Oil and Gas operator in Los Angeles where pilot was conducted. The images (10km × 10km in size) represent the RGB channels from Sentinel-2 satellite. The red pin represents the location of interest* 50

3.4 *Detection of methane emissions at the sites of Oil and Gas operator where pilot was conducted. The images shows an overlay of concentration mask on the RGB imagery. The red pin represents the location of interest. Total instances of detections were 11, we are only showing two here for site 1 and one for site 2. Site 2 shows an unknown detection in a neighboring area of almost 1200kg/hr concentration* 51

4.1 *Summary of Mara-Wildlife dataset. (a) Satellite view indicating the four flight trajectories, each represented in a different color. (b, c, d, e, f, g) Annotations of (b) zebra, (c) hartebeest, (d) cattle, (e) shoats (sheep and goats), and (f) zebra. These are the zoomed in verion of aerial images. The pixel footprint of object of interest is* $\leq 0.001\%$ *of the image. Best viewed in color.* 55

4.2 *Distribution of* ($\geq 15kg$) *mammals identified in digital imagery collected across the Masai Mara Ecosystem, Kenya.* 57

4.3 *Overview of WildlifeMapper (WM) architecture. Given an input image of size* $1024 \times 1024 \times 3$*, the High-Frequency Feature Generator (HFG) module generates information about potential location of object of interest. The Feature Refiner (FR) takes these potential location along with contextual features from Patch Embed layer and sent output to Image Encoder. In parallel, the Query Refiner (QR) incorporates the output of FR to refine learnable queries. Finally these queries are decoded using encoded features from Image Encoder and predict bounding box and class.* 60

4.4 *The sample output visualization from the High-Frequency Feature Generator (**HFG**) module. The illustration shows the effectiveness of the module in suppressing the homogeneous and dominant background, while highlighting objects of interest (i.e., animals). The top image shows bomas, natural structures constructed to contain livestock, and paths that have been suppressed. Animals, however, are clearly identified, especially inside the boma. The bottom image shows a water body (a dam created for livestock) that has been suppressed by the module. Animals can again be highlighted throughout the image.* 66

4.5 *Qualitative results. The top row highlights examples of crowded and partially occluded scenes. Row-1, Column-1 & Row-3, Column-3 shows examples where animals are obstructed by shadows. The zoomed-in box in Row-3, Column-3 shows a zebra partially occluded. The bounding box color is coded according to class names: shoats-"hot pink", cattle-"deep sky blue", zebra-"light yellow".* 75

4.6 *Failure cases. Left shows an example where animals are occluded by shadows and are difficult to detect. Right shows an example of rock detected as an impala, emphasizing the difficulty in differentiating objects in the image from animals of interest.* 76

5.1 Example of ECG and $\partial Z/\partial t$ waveforms computed from the present data. $\partial Z/\partial t$ represents the change in impedance recorded by ICG (Z) signal with time. After each ECG peak value there exists an $\partial Z/\partial t$ peak value. The time difference between these two values is known as the initial systolic time interval (ISTI). 79
5.2 Model Architecture. Green boxes are the different modules of the model. Yellow boxes are the variables throughout the model. The Emission model processes the raw input data which is then fed into spatial and temporal modules. The Detector network predicts ISTI value for each of the frames from the output of these modules. This ISTI signal is used as input in our stress detection network. 80
5.3 Discrete ISTI values are plotted against the peak positions of the ECG signal for a single participant. The "Base", "Prep", "Immersion" and "Recovery" labels refer to different phases of our stress induction protocol, whereby participants immerse their feet in either ice-water ("stress" condition) or lukewarm water ("no-stress" condition). The data shown were randomly selected from the "no-stress" condition. See section 5.4.1 for a detailed description of the protocol. 83
5.4 Stress detection network. Estimated ISTI signal is directly fed into the classifier network to predict the probability that the subject is under stress. 87
5.5 CPT/WPT Setup and Protocol. An example of a fully instrumented participant is shown. Participants followed instructions for the protocol presented on a computer monitor. After the baseline period the participant is instructed to position both feet on the edge of the bucket and prepare for immersion (prep). They then immerse the feet for 90s, then withdraw the feet and rest them on a towel for a 40 s recovery period. 89
5.6 Quality of our predicted ISTI signal in stress and no-stress conditions. Data shown are examples from a single participant's data (selected at random). The "Base", "Prep", "Immersion" and "Recovery" labels refer to the different phases of the CPT/WPT procedure. 94
5.7 Importance of ISTI signal in detecting stress. Ground truth ISTI data from a single participant (randomly selected) are shown. Clearly, ISTI signal in the stress condition is different from the ISTI signal in no-stress condition. The "Base", "Prep", "Immersion" and "Recovery" labels refer to the different phases of the CPT/WPT procedure. 95
5.8 Example StressNet failure cases. Network performance is impaired when the face is outside the video frame or obscured. 98

6.1 *Depiction of visible morphological and pigment change in plants due to pollution. It can be seen that the green color of the leaves turning white and brown shows pigment change in the plant due to presence of certain pollutants around, in this specific case, this is due to presence of SO_2 in the atmosphere as shown in Red boxes.* 103

A.1 *Depiction of data collection process. Each flightline is* $\sim$ 300 *kms long. An array of 598 sensors records data at 1.5m/pixel spatial resolution. All flightlines are ortho-corrected. Each data-cube is of dimension* $\sim 23k \times \sim 1.5k \times 432$. 108

A.2 *Spectral absorption pattern of CH_4 gas. The x-axis show the channel number ranging from 0-400 corresponding to wavelength range (*$400nm - 2500nm$*). It is obtained from the public repository HITRAN [1].* 109

A.3 *Sample ground truths and predictions on MHS dataset. We are showing different type of terrains and CH_4 predictions on them. The type of emission source in all samples varies too.* 114

A.4 *Samples where MM fails to detect the CH_4 plume. We observed that these samples were recorded during the evening time and hence reflectance from the ground terrain is very weak. Therefore the absorption of reflected solar radiations by CH_4 is very low and hence the emissions goes undetected.* . 115

B.1 *Good cases. Each column shows different category of detection. Column-1 shows large animals: cattle, buffalo; Column-2 shows detection of small animals (warthog, topi), Column-3 shows detection of animals hidden or occluded.* 120

B.2 *Failure cases. The animals hiding in the shade are difficult to detect. Additional examples of misclassification also provided.* 121

List of Tables

2.1 *Statistics shows MHS dataset comparison with JPL-CH4-detection-V1.0 [2] dataset. Each flightline have multiple large and small plume sites. Each flightline have atleast 4 plume sites. The Point Source represents high concentration (*$300kg/hr$*) to leakage from sources like pipeline leak, storage tanks, oil and gas refineries. Diffused Source represent low concentration leakages from sources like biomass degradation in landfills. Our dataset is covers more diverse type of terrain over 6 states.* 29

2.2 *Comparison with baselines. "-" represent Not Applicable and "*" represent no* ***SFG*** *module and a random query used for transformer decoder. The top section shows performance on JPL-*CH_4 *dataset [2]. MethaneMapper achieves better results than heavily tuned H-mrcnn with* $\sim 5\times$ *fewer parameters. The overall detection accuracy is higher on this dataset because the type of ground terrain is uniform across all flightlines. In MHS dataset, MM outperforms multiple baselines as shown in rows 4-12. MM accuracy is lower in MHS than JPL-*CH_4 *dataset because MHS dataset has more variety of ground terrain spreading over 6 states* 33

2.3 *Comparison with classical machine learning methods. "-" represent Not Available. The classical ML methods are not suited for the* CH_4 *detection task. MethaneMapper outperforms all methods on JPL dataset [2]* 34

4.1 *Comparison of Mara-Wildlife dataset with other publicly available dataset. Mara-Wildlife dataset has* $\times 3$ *more unique species than the total of all other datasets. Each image is significantly larger and higher ground resolution making* $77k$ *unique images of size* 1024×1024 *with* 21 *different animal species. GSD: ground sampling distance; DRC: Democratic Republic of Congo.* 57

4.2 *Comparison with baseline models. The top section shows performance on species detection on Mara-wildlife dataset and low section shows performance on the mixed dataset from Virunga-Garamba-AED dataset. The overall detection accuracy is generally higher in Virunga-Garamba-AED dataset because there are only 6 species and the terrain is quite similar in all images.* 70
4.3 *HFG module effectiveness in refining the image features and queries. "✗" represents not used, "✓" represents used and "*" represents that random queries are used but there was not refining with HFG features.* 71

5.1 StressNet's performance in predicting ISTI signal. The performance is measured on Pearson-Correlation Coefficient(PC Coefficient) and mean square error. Our model clearly outperforms the existing methods by a good margin. 93
5.2 StressNet can classify stress state with greater AP using contact-less ISTI estimates when compared to other commonly used contact-less signal estimates (HR and HRV). 96
5.3 Comparison of different backbones' performance. In the task of estimating ISTI signal resnet50 is better than all other backbones. 97

B.1 *Comparison with baseline models. The top section shows performance on species detection on Mara-wildlife dataset and low section shows performance on the mixed dataset from Virunga-Garamba-AED dataset. The overall detection accuracy is generally higher in Virunga-Garamba-AED dataset because there are only 6 species and the terrain is quite similar in all images.* 116

Chapter 1

Introduction

The field of computer vision has witnessed unprecedented growth over the past decade, driven by the advent of deep learning and the availability of large-scale datasets. Large vision models, such as the Segment Anything Model (SAM) [3], Florence-2 [4], and GPT-4 [5], have set new benchmarks in various computer vision tasks, including object detection, image segmentation, and image generation. These models have been trained on vast amounts of data, with SAM leveraging 11 million images and 1 billion annotations, Florence-2 using 126 million images and 5.4 billion annotations, and GPT-4 utilizing 13 trillion tokens. The sheer volume of data has enabled these models to learn robust features and achieve high performance across diverse applications.

The primary challenge in applying large vision models to scientific problems lies in the scarcity of annotated data and the inherent complexity of the data itself. For example in the remote sensing domain, the problem of methane detection have very limited annotated dataset available, acquiring such dataset and annotation is requires expensive instruments and we need a subject matter experts to interpret and annotate such high-dimensional data. State-of-the-art large vision models in natural image domain that rely on massive amounts of labeled data are impractical in scientific domains, ne-

cessitating alternative strategies to build and train models effectively. The limitations in data availability are further exacerbated by the specialized nature of the data for example hyperspectral data requires an domain-expert knowledge to interpret and utilize effectively.

Addressing these challenges requires a paradigm shift in how large vision models are designed and trained for scientific applications. Instead of relying solely on large volumes of data, there is a need to integrate domain-specific knowledge (e.g. need to know the chemical reactive properties of methane gas with solar radiations) into the models to enhance their performance with limited data. This involves developing techniques to extract and encode prior information about the domain into the model, enabling it to focus on the most relevant features and achieve faster convergence during training.

1.1 Motivation

Motivated by the above mentioned problems, we set out to develop models dealing with complex data in different problem contexts and modalities.

Remote Sensing Applications : In the context of remote sensing, for example, the data obtained from hyperspectral and multispectral sensors is vastly different from the natural images commonly used to train large vision models. These sensors capture detailed spectral information that can provide insights into material composition, vegetation health, and other environmental factors. However, this richness in data comes at the cost of increased complexity, making it difficult to apply conventional computer vision techniques directly.

Biomedical Applications : Similar challenges are faced in the field of biology, particularly in the detection and analysis of physiological signals. For instance, multispectral

biomedical imaging techniques captures complex data that require expert interpretation. The detection of physiological signals, such as heart rate variability from electrocardiograms (ECGs) or brain activity patterns from electroencephalograms (EEGs), involves intricate data that is difficult to annotate accurately and abundantly. The complexity of physiological data, combined with the scarcity of annotated datasets, makes it challenging to develop robust models for tasks such as disease diagnosis, monitoring of physiological conditions, and personalized medicine.

Ecology Applications : Another challenge lies in the field of ecology, specifically in the detection and monitoring of animals in the wild, such as in large open grasslands. Remote sensing techniques, including satellite imagery and drone footage, are increasingly used to monitor wildlife populations and track animal movements. However, these images often have varying resolutions and conditions, such as different times of day and weather, making it difficult to identify and count animals accurately. The annotated datasets required to train models for these tasks are limited, as capturing and annotating data in the wild is resource-intensive and logistically challenging. The sparse and noisy nature of ecological data further complicates the task, necessitating sophisticated models that can effectively discern animals from their natural habitats and deal with occlusions and varying backgrounds.

1.2 Challenges

One of the most pressing issues is the limited availability of annotated data. In many scientific domains, the process of annotating datasets is both expensive and time-consuming. This is particularly true in specialized areas such as methane emission detection, where obtaining accurate annotations requires sophisticated hyperspectral sensors

and expert knowledge.

Furthermore, scientific computer vision tasks often necessitate the integration of multiple sources of information. For instance, in the context of methane emissions, hyperspectral imagery from airplane based sensors, multispectral imagery from satellite based sensors, ground release information, and wind speed data. Each of these data sources provides unique insights, but combining them effectively poses a significant challenge. Integrating these diverse datasets requires advanced data fusion techniques and robust algorithms capable of handling the inherent variability and noise present in each modality.

Another critical challenge is the acquisition of annotated datasets. Given the specialized nature of scientific computer vision problems, creating comprehensive and high-quality datasets involves significant logistical and financial hurdles. Additionally, the expertise required to annotate such datasets accurately is scarce, adding another layer of complexity to the process. Combining data from multiple modalities also presents technical difficulties. Different sensors and measurement instruments have varying spatial and temporal resolutions, as well as differing levels of sensitivity and accuracy. Aligning and integrating these disparate data sources into a cohesive framework requires sophisticated preprocessing and calibration methods. Moreover, the computational demands of processing and analyzing multimodal data can be substantial, necessitating the use of high-performance computing resources.

In summary, the challenges in scientific computer vision are multifaceted and stem from the inherent complexity of the problems, the limited availability of annotated data, and the difficulties associated with integrating multiple data modalities.

1.3 Summary of Contributions

This thesis research focuses on making large vision models by incorporating domain-specific knowledge by using advanced signal processing techniques. By leveraging the unique characteristics and requirements of specific scientific domains, We aim to enhance the performance and applicability of these models. This approach not only improves the accuracy of detection, segmentation, quantification, and localization tasks but also reduces the computational cost required. Additionally, this work extends to the application of these models across various modalities of data such as hyperspectral, multispectral and thermal imagery along with other modalities. The contributions are as follows:

1. In the context of methane emission detection, we developed a series of novel works to detect and segment methane gas plumes from satellites as-well-as drones, covering multiple spatial scales. We can handle and synthesize data from different sources. Using such imagery we can segment methane gas plume mask, estimate concentration and potential emission source. This is discussed further in detail in **chapter 2** and **3** in the thesis

2. In the context of ecological applications, we collaborated with the Smithsonian Institution, Kenya Wildlife Trust, and Wildlife Research and Training Institute in Kenya to develop a novel method for detecting and classifying animal species from aerial imagery in the vast open grasslands of the Masai Mara Conservancy in Kenya. As discussed in **chapter 4**, this innovative work significantly enhances wildlife accounting, enabling faster and more accurate monitoring of animal populations. By providing detailed and timely data on species distribution and abundance, our approach plays a crucial role in accelerating conservation efforts and informing effective wildlife management strategies.

3. In the context of biomedical applications, we worked on developing a series of innovative methods for detecting stress experienced by individuals using thermal imagery as discussed in **chapter 5**. By analyzing thermal videos, we were able to reconstruct various physiological signals such as Impedance Cardiography (ICG), Electrocardiography (ECG), and Initial Systolic Time Interval (ISTI) through our models. These reconstructed signals allowed us to estimate the amount and type of stress experienced by individuals with high accuracy. This non-invasive approach to stress detection provides valuable insights into the physiological responses to stress, offering a promising tool for both research and practical applications in health monitoring and stress management.

This versatility ensures that the proposed can be applied to a wide range of scientific problems, from environmental monitoring to biomedical imaging to ecological monitoring, thereby demonstrating the broad impact and potential of my research.

1.4 Organization of Thesis

The organization of the thesis is as follows:

Chapter 2, **MethaneMapper** discusses the development of large vision models for remote sensing applications, specifically targeting methane detection from hyperspectral data. Our research explores methane detection using hyperspectral imagery collected from airplanes, addressing this challenge through the lenses of signal processing, computer vision, and machine learning. We developed the spectral linear filter as a part of initial exploration of methane detection using deep learning. This filter is an integral to MethaneMapper, our large vision model designed for methane detection. MethaneMapper leverages the vision transformer architecture to approach the problem as a segmentation task. By using the spectral linear filter in the feature extraction pro-

cess, MethaneMapper achieves faster convergence, improving the efficiency and accuracy of methane detection from hyperspectral imagery.

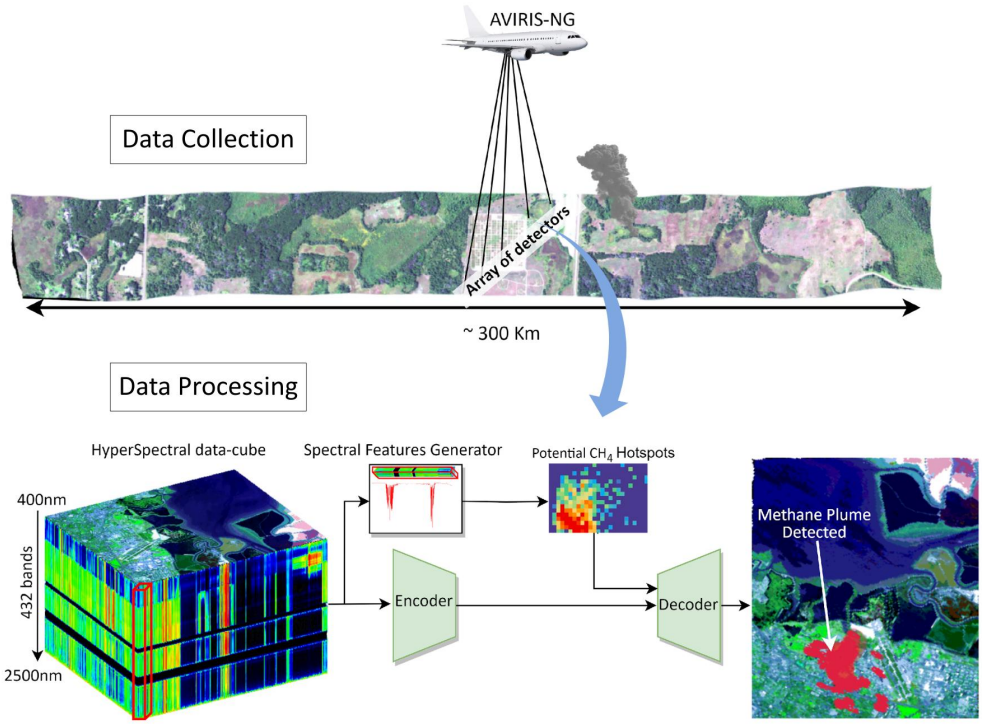

Figure 1.1: *Data collection processing using aerial imagery and detection of methane plumes*

Chapter 3: In this chapter, we present the development of **MethaneMapper** for processing multispectral data from **satellites**. Our primary focus is on multispectral data from Sentinel-2 and LandSat-8 satellites. MethaneMapper incorporates a pretrained vision transformer model into its architecture, with a redesigned spectral feature generator based on the Beer-Lambert law. This generator identifies potential methane hotspots, which are then refined by the Query Refiner module to guide the feature extraction process from the encoded features provided by the transformer encoder. We demonstrate the effectiveness of MethaneMapper through test results on data from the Sentinel-2 satellite over Los Angeles County and LandSat-8 imagery from Algeria, showcasing its capability

to accurately detect methane hotspots from multispectral satellite data.

Chapter 4: **WildlifeMapper**, this chapter discusses the application of our large vision models in ecology, specifically for detecting and identifying animals in aerial imagery. We introduce WildlifeMapper, a model designed to identify animals in large open grasslands, where the animals occupy less than 0.01% of the image pixels. We explore methods to guide the feature extraction process to capture relevant features from the imagery. Our approach includes developing a high-frequency feature generator and utilizing patch-level filtering to accurately locate animals and capture their context within the image. We also cover the real-world application of WildlifeMapper. Park rangers will use our online platform, BisQue [6], to detect and identify animals, thereby supporting conservation efforts.

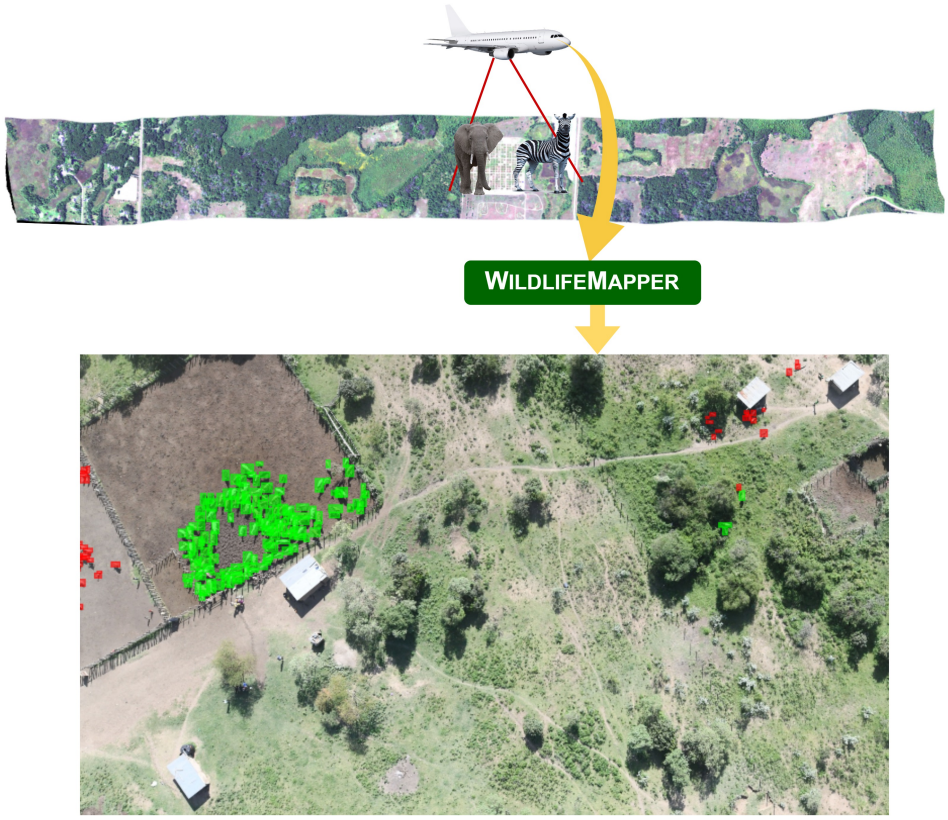

Figure 1.2: *Data collection processing using aerial imagery and detection of wild animals. The bottom image is a zoomed in version of the aerial imagery. The Green boxes represents cattle and the Red boxes represents shoats (sheep and goats).*

Chapter 5: **StressNet**, This chapter explores the biomedical applications of our large vision models, focusing on our work with StressNet. StressNet processes multispectral data (RGB + thermal) of the human face to estimate physiological signals such as Electro Cardio Graph (ECG), Impedance Cardio Graph (ICG), and Initial Systolic Time Interval (ISTI). These signals are used to determine whether a subject is experiencing physical stress. To build StressNet, we developed an emission representation model to simulate the reflection, refraction, and absorption of light by different sources. We integrated this emission representation model to our large vision model to guide the feature extraction process within the vision transformer model. We validated StressNet through

experiments involving 60 subjects from UCSB, using the Cold Pressor Test (CPT), where individuals experience physical stress by immersing their hands in cold water.

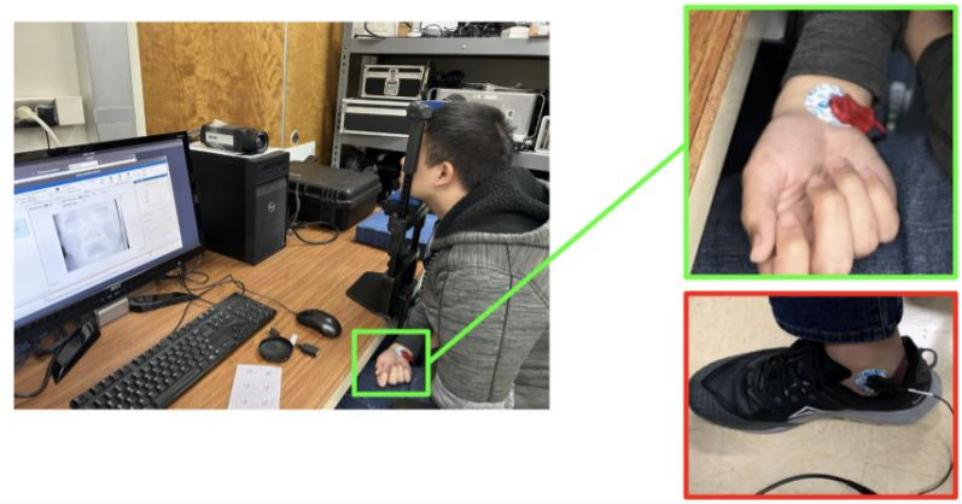

Figure 1.3: *Representation of the Cold Pressor Test (CPT) for StressNet, Left image shows video recording of the face of person, Right image shows physiological signal (ECG) collection .*

Chapter 6: We discuss some future directions and conclude this dissertation in this chapter.

In summary, this dissertation introduces novel methodologies for building large vision models tailored to diverse scientific imaging problem contexts and applications. By leveraging the unique characteristics of specific scientific domains, these models improve accuracy in detection, segmentation, quantification, and localization tasks, while reducing computational costs. This research extends across various data modalities, including hyperspectral, multispectral, and thermal imagery, demonstrating the versatility and impact of large vision models in environmental monitoring, wildlife conservation, and healthcare.

Chapter 2

MethaneMapper

We consider the problem of detecting and localizing methane (CH_4) plumes from multi-spectral/hyperspectral imaging data. Fig. 2.1 represents an example source of methane emission and its presence in visible and beyond visible domain. Detecting and localizing potential CH_4 hot spots is a necessary first step in combating global warming due to greenhouse gas emissions. Methane gas is estimated to contribute 20% of global warming induced by greenhouse gasses with a Global Warming Potential (GWP) 86 times higher than carbon dioxide (CO_2) in a 20 year period. To put into perspective, the amount of environmental damage that CO_2 can do in 100 years, CH_4 can do in 1.2 years. Hence it is critical to monitor and curb the CH_4 emissions. The longstanding CH_4 has a mean atmospheric residence of 7.9 years [7] and its presence in the atmosphere has been increasing since the industrial revolution [8]. In this chapter, we explain the methane detection using hyperspectral imagery from aerial survey. We first introduce the spectral linear filter as part of Hyperspectral Mask-RCNN (H-MRCNN) [9] work, and later discuss about how spectral linear filter is used in MethaneMapper [10] to accurately delineate methane plumes. The results of this chapter are published in Computer Vision and Pattern Recognition (CVPR) 2023 conference and Winter Conference on Applications of

Computer Vision (WACV) 2020.

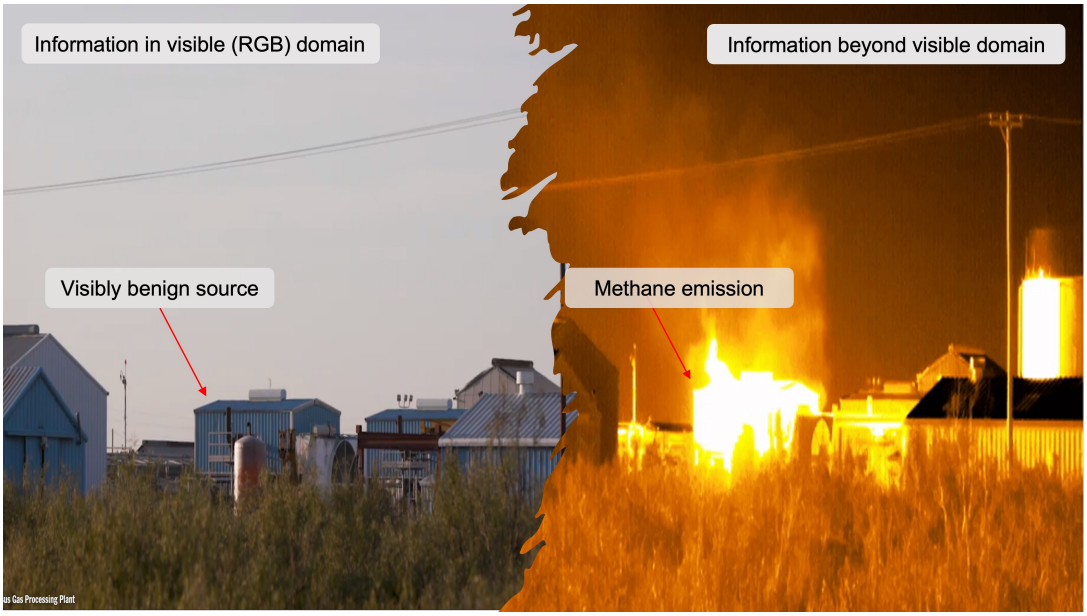

Figure 2.1: *Representation of information in visible and beyond visible(multispectral/hyperspectral domain) domain*

2.1 Introduction

CH_4 emission has many sources as shown in Fig. 2.2, the ones of particular interest are those from oil and natural gas industries. According to the United States Environmental Protection Agency report, CH_4 emissions from these industries accounts to 84 million tons per year [11,12]. These CH_4 emissions emanate from specific locations, mainly from pipeline leakages, storage tank leak or leakage from oil extraction point.

Current efforts to detect these sources mostly depend on aerial imagery. The Jet Propulsion Laboratory (JPL) has conducted thousands of aerial surveys in the last decade to collect data using an airborne sensor AVIRIS-NG [13]. Several methods have been proposed to detect potential emission sites from such imagery, for example, see [14–19].

However, these methods are in general very sensitive to background context and land-cover types, resulting in a large number of false positives that often require significant domain expert time to correct the detections. The primary reason is that these pixel-based methods are solely dependent on spectral correlations for detection. Spatial information can be very effective in reducing these false positives as CH_4 plumes exhibit a plume-like structure morphology. There has been recent efforts in utilizing spatial correlation using deep learning methods [9, 20], however, these works don't leverage spectral properties to filter out confusers. For example, methane has similar spectral properties as white-painted commercial roofs or paved surfaces such as airport asphalts [21]. This chapter presents a novel deep-network based solution to minimize the effects of such confusers in accurately localizing methane plumes.

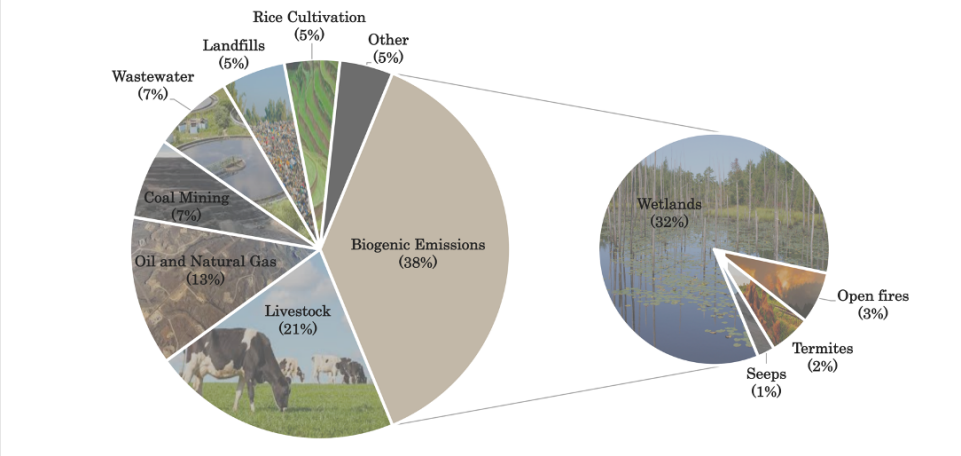

Figure 2.2: *Biogenic and Anthroproenic sources of methane emissions*

In this chapter we explore the development of large vision models for methane detection from hyperspectral data. We use hyperspectral imagery collected from airplanes, addressing this challenge through the lenses of signal processing, computer vision, and machine learning. We will discuss two works named: Hyperspectral Mask-RCNN (H-MRCNN) [9] and MethaneMapper(MM) [10]. H-MRCNN is the basis of MM. During

the development of H-MRCNN, we developed a matched filtering approach and used a Mask-RCNN based deep convolutional neural network to solve the problem as a segmentation mask task. In MM, we improved on the limitations of H-MRCNN, and curated a comparatively larger and diverse dataset and efficient transformer based detection algorithm. We will discuss H-MRCNN first in brief and have a detailed discussion about MM in the following chapter.

2.2 Related Works

Existing machine learning-based hyperspectral image analysis methods primarily focus on classification, with a smaller subset dedicated to target detection, as discussed in [22]. One commonly used method in this domain is logistic regression, particularly for land cover classification in remote sensing applications, where it performs pixel-wise classification [23]. Despite its widespread use, this method often suffers from high false-positive rates due to its sensitivity to noise and variations in the data. To address this, more advanced methods such as multinomial logistic regression (MLR) [24] have been developed. MLR, a discriminative approach, directly models the posterior class distributions and is particularly effective in applications involving the linear spectral unmixing process, where it can provide more accurate classifications by focusing on the most relevant spectral features.

Support vector machines (SVMs) are another popular choice in hyperspectral data analysis, widely recognized for their ability to generate decision boundaries with the maximum margin of separation between data samples from different classes [25]. However, when it comes to target detection, SVM's related algorithm, support vector data description (SVDD) [26, 27], is often employed. SVDD generates a minimum enclosing hypersphere containing the targets, which can be particularly useful in identifying anoma-

lies or specific targets within a hyperspectral image. Nonetheless, a significant limitation of SVDD is its failure to account for the underlying distribution of the scene data, which can result in an inability to distinguish targets from the background distribution, leading to false positives.

Gaussian mixture models (GMMs) have also been applied to hyperspectral data analysis, where they represent the probability density of the data as a weighted summation of a finite number of Gaussian densities, each with distinct means and standard deviations. This approach allows for the clustering of hyperspectral data and the segmentation of images into homogeneous areas [28], providing a more structured analysis of the data. Latent linear models, such as principal component analysis (PCA), are also commonly used in hyperspectral imagery. PCA performs a linear transformation to find a latent representation of the data, projecting it onto an orthogonal set of axes to reduce dimensionality while preserving the most significant variance in the data [29–31]. This technique is particularly valuable as a preprocessing tool, helping to simplify the data before applying more complex analysis methods.

Ensemble learning, another powerful approach, involves combining the predictions of multiple base models to produce a more accurate overall result. This technique has been successfully applied to hyperspectral classification tasks, where it leverages the strengths of various models to improve prediction accuracy [32]. In addition, kernelized PCA, followed by deep learning methods, offers a promising solution for target detection in hyperspectral imagery [33–35]. For example, the work by Chen et al. [36] introduces a three-dimensional convolutional neural network (CNN) that outperforms two-dimensional CNNs by directly learning spatial-spectral features. This three-dimensional CNN approach, which spans both spatial and spectral axes, significantly improves the model's ability to capture complex patterns within the data, though it requires large training datasets to perform effectively.

In the specific context of methane plume detection using airborne imaging spectrometer data like AVIRIS-NG [13], traditional methods have been widely employed. These include the Iterative Maximum a Posterior Differential Optical Absorption Spectroscopy algorithm (IMAP-DOAS) [19,37] and matched filters [14–18]. IMAP-DOAS, while effective, requires data from both airborne and ground-based hyperspectral sensors, making it impractical for many real-world applications due to the logistical challenges and resource requirements. On the other hand, matched-filter methods, which normalize spectral signals using background statistics and match them with the CH_4 spectral signature at each spatial location, are more commonly used. However, these methods are highly sensitive to surface albedo and land cover, often resulting in spurious detections that resemble methane plumes. Consequently, domain experts must manually inspect each flight line to distinguish between real CH_4 plumes and false positives [2].

To mitigate the high rate of false positives, clustering techniques have been introduced, such as the cluster-tuned matched filter [16, 38], which involves clustering pixels with similar spectral properties using methods like k-means clustering. While these techniques improve detection accuracy, both IMAP-DOAS and matched filters remain prone to false positives due to their reliance on pixel-wise processing, which does not adequately capture the complex interactions between different elements in the scene.

Machine learning approaches have also been explored for methane detection in hyperspectral imagery, leveraging algorithms like SVMs, GMMs, and deep learning models. For instance, Methanet [20], a more recent model, focuses on estimating methane concentrations from matched-filter data, yet it still faces limitations in effectively addressing confusers in CH_4 spectral signatures.

To address the limitations of existing methods, our proposed works: H-MRCNN 2.3 and MethaneMapper 2.4 advances methane plume detection by integrating both spectral and spatial correlations, thereby providing a more robust and accurate delineation of CH_4

plumes. These approaches effectively mitigates the issues of false positives and enhances the reliability of methane detection in complex environments.

Datasets: The only dataset publicly available with annotation for CH_4 plume detection is JPL-CH_4-detection2017-V1.0 dataset [2]. It contains only 46 AVIRIS-NG [13] flight lines in the US Four-Corners region. Deep learning architectures require a large number of annotated samples, and for this reason we introduce the new MHS dataset with over 1200 annotated flightlines and $\sim$ 4000 plume sites discussed in section 2.5.1.

2.3 Hyperspectral Mask-RCNN (H-MRCNN)

Effective analysis of hyperspectral imagery is essential for gathering fast and actionable information of large areas affected by atmospheric and green house gases. Existing methods, which process hyperspectral data to detect amorphous gases such as CH_4 require manual inspection from domain experts and annotation of massive datasets. These methods do not scale well and are prone to human errors due to the plumes' small pixel-footprint signature. The first proposed Hyperspectral Mask-RCNN (H-mrcnn) uses principled statistics, signal processing, and deep neural networks to address these limitations. H-mrcnn introduces fast algorithms to analyze large-area hyper-spectral information and methods to autonomously represent and detect CH_4 plumes. H-mrcnn processes information by match-filtering sliding windows of hyperspectral data across the spectral bands. This process produces information-rich features that are both effective plume representations and gas concentration analogs.

2.3.1 Proposed H-mrcnn solution

The proposed approach tackles two datasets derived from AVIRIS-NG instrument by Jet Propulsion Laboratory (JPL) [39]; **Dataset A** is a rectified 4-band dataset defined

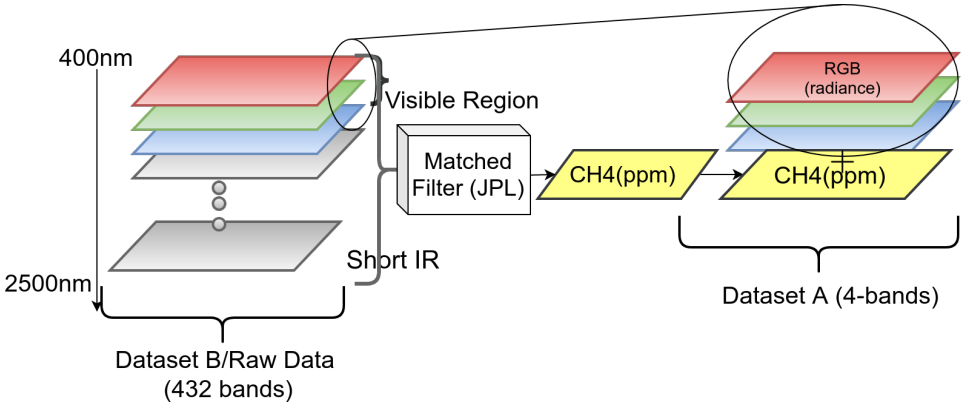

Figure 2.3: *Relation between dataset A (χ_A) and dataset B (χ_B). The 432-bands data from dataset B are processed through a matched filter to yield dataset A. Detecting plumes using this information poor dataset (Dataset A) is challenging. H-mrcnn addresses this challenge by modeling terrain absorption using ensemble and decision fusion methods.*

in [39]. This data contains 4-band datum with three bands comprising red, green, and blue reflectance intensities and a fourth band comprising CH_4 relative concentration in ppm per meter (parts per million per meter). The fourth channel is generated from 432-bands. The 432-band measurements are processed into one single-channel array using conventional match-filtering techniques with the CH_4 signature as the target. The conventional match-filtering technique takes 180 minutes per datapoint to process 432bands into 1 single channel output. The optimized implementation has reduced this processing time to 15 minutes per datapoint. The single channel array is stacked with three other bands, each selected from the visual red, blue, and green wavelengths. The proposed naive single-band solution uses dataset A to evaluate and validate the initial findings and tune a binary plume detector.

Dataset B is an unrectified, 432-band (i.e., raw data) dataset. It is acquired in VSWIR(Visible Shortwave Infrared) range, measuring over 432 spectra of color channels ranging from ultraviolet ($380nm$) to shortwave infrared ($2510nm$). The images are taken over large areas, creating a three-dimensional data cube of radiance, where two dimensions are the spatial domain (i.e., 2D-image) and the third one is in the spectral domain

(i.e., wavelength) as shown in Figure 2.3, which visualizes the relationship between the two datasets. This data is collected in "Four Corner Area" (FCA), the geographical US-Mexico border. This dataset is used to design, develop, and evaluate the proposed H-mrcnn solution, which is the formalized naive single-band detector. H-mrcnn is a combination of an optimized matched filter and Mask-RCNN that identifies the correlation both in spectral and spatial domains respectively and detects the presence and shape CH_4 plume.

2.4 MethaneMapper

The H-MRCNN approach has limitations in scalability, geographic generalizability, and efficiency. To address these challenges, we propose a novel end-to-end spectral absorption wavelength aware transformer network, MethaneMapper, to detect and quantify the emissions. MethaneMapper introduces two novel modules that help to locate the most relevant methane plume regions in the spectral domain and uses them to localize these accurately. Thorough evaluation shows that MethaneMapper achieves 0.63 mAP in detection and reduces the model size (by 5×) compared to the current state of the art. In addition, we also introduce a large-scale dataset of a methane plume segmentation mask for over 1200 AVIRIS-NG flight lines from 2015-2022. It contains over 4000 methane plume sites. Our dataset will provide researchers the opportunity to develop and advance new methods for tackling this challenging green-house gas detection problem with significant broader social impact.

Our proposed approach, referred to as the MethaneMapper (MM), adapts the DETR [40], a transformer model that combines the spectral and spatial correlations in the imaging data to generate a map of potential methane (CH_4) plume candidates. These candidates reduce the search space for a hyperspectral decoder to detect CH_4 plumes and

remove potential confusers. MM is a light-weight end-to-end single-stage CH_4 detector and introduces two novel modules: a *Spectral Feature Generator* and a *Query Refiner*. The former generates spectral features from a linear filter that maximizes the CH_4-to-noise ratio in the presence of additive background noise, while the latter integrates these features for decoding.

A major bottle neck for development of CH_4 detection methods is the limited availability of public training data. To address this, another significant contribution of this research is the introduction of a new Methane Hot Spots (MHS) dataset, largest of its kind available for computer vision researchers. MHS is curated by systematically collecting information from different publicly available datasets (airborne sensor [41], Non-profits [42, 43] and satellites [44]) and generating the annotations as described in Section 2.6.2. This curated dataset contains methane segmentation masks for over 1200 AVIRIS-NG flight lines from years 2015 to 2022. Each flight line contains anywhere from 3-4 CH_4 plume sites for a total of 4000 in the MHS dataset.

Our contributions can be summarized as follows:

1. We introduce a novel single-stage end-to-end approach for methane plume detection using a hyperspectral transformer. The two modules, *Spectral Feature Generator* and *Query Refiner*, work together to improve upon the traditional transformer design and enable localization of potential methane hot spots in the hyperspectral images using a Spectral-Aware Linear Filter and refine the query representation for better decoding.

2. A new *Spectral Linear Filter (SLF)* improves upon traditional linear filters by strategically picking correlated pixels in spectal domain to better whiten background distribution and amplify methane signal.

3. A new benchmark dataset, MHS, provides the largest ($\sim 35\times$) publicly available

dataset of annotated AVIRIS-NG flight lines from years 2015-2022.

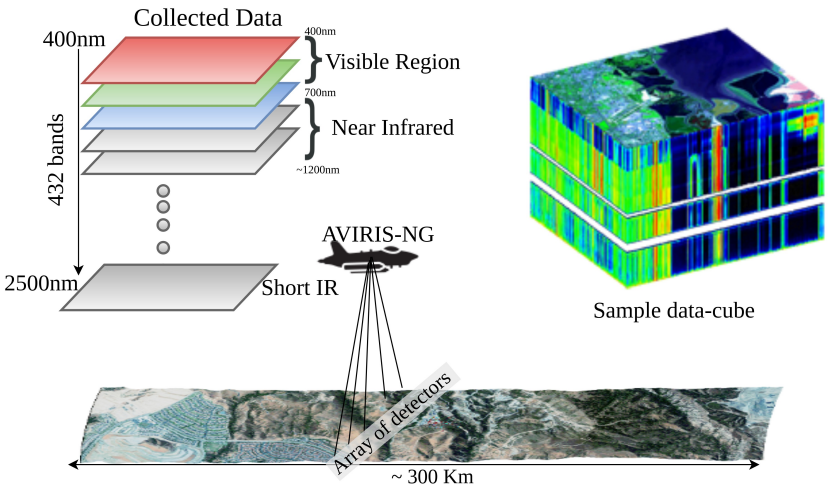

Figure 2.4: *Depiction of data collection process. Each flightline is* ~ 300 *km long. An array of 598 sensors records data at 1.5m/pixel spatial resolution. All flightlines are ortho-corrected. Each data-cube is of dimension* $\sim 25000\times \sim 1500 \times 432$.

Our work is at the intersection of hyperspectral data for CH_4 detection, deterministic linear filtering methods for spectral features and encoder-decoder based transformer. MethaneMapper uses both spectral and spatial correlation to accurately delineates CH_4 plumes.

2.5 MethaneMapper (MM) Architecture

2.5.1 Data Overview

AVIRIS-NG hyperspectral imaging sensors capture spectral radiance values from N_0 ($N_0 = 432$) channels corresponding to wavelengths ranging from $400nm - 2500nm$ as shown in Fig. 2.4. The complete hyperspectral image is represented as $\mathbf{x} \in \mathbb{R}^{H_0 \times W_0 \times N_0}$ where H_0, W_0 are the height & width, respectively, and $N_0 = 432$ is number of channels. This hyperspectral data includes a very weak signature of CH_4 around 2100-2400nm,

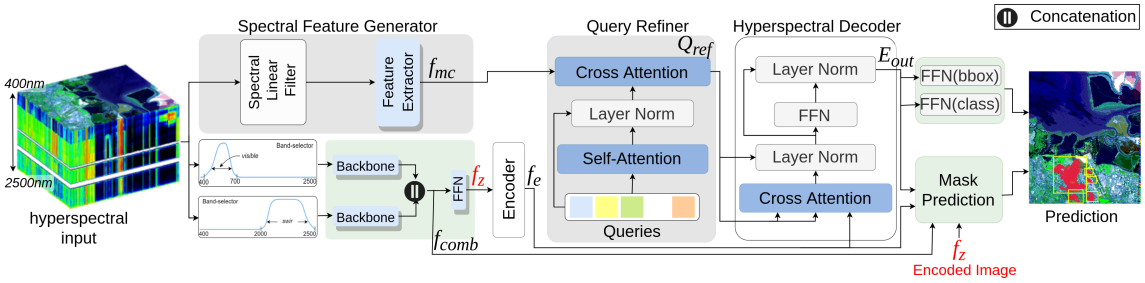

Figure 2.5: *Overview of MethaneMapper (MM) architecture. Given a hyperspectral image, our RGB* $(400nm - 700nm)$ *and SWIR* $(2000nm - 2500nm)$ *band-pass filters passes a subset of channels in desired wavelength range and feed them to CNN backbones (ResNet) to extract features. These features are concatenated and fed to Transformer Encoder. Parallelly, our Spectral Feature Generator (SFG) modules takes in all channels of input image and generate methane candidates features. Next these candidates are sent to Query Refiner (QR) to refine queries. Then these queries decoded using encoded feature from Transformer Encoder. Finally each decoded query is used to predict a plume mask via Mask Prediction and, bounding box and class via FFNs (Feed Forward Network).*

conflated with radiations from the surrounding land cover and background clutter. A single flight-line could be over a couple miles long (about 25K pixels in one of the dimensions), with an array of sensors recording the data at 1.5m/pixel resolution. The images are orthorectified before processing.

2.5.2 Technical Overview

Referring to Fig.4.3, MM contains the following main components: (i) 2 CNN backbones to extract a compact feature representation of the spectral regions of interest from the hyperspectral image, (ii) a *Spectral Feature Generator* (**SFG**), and (iii) a *Query Refiner* (**QR**) in between an encoder-decoder pair (inspired by GTNet [45], SSRT [46]). The hyperspectral image is first processed through two separate band-pass filters to select the channels in visible $(400-700nm)$ and short-wave infrared (SWIR)$(2000-2500nm)$ wavelength regions, and are then passed through CNN backbones. Output of these backbones are concatenated together and then encoded using a transformer encoder.

The **SFG** (Sec. 2.5.4) takes in all channels of the hyperspectral image and process them through a spectral linear filter. The **SFG** exploits the spectral correlation to generate methane candidates feature maps and passes them to **QR**. The **QR** (Sec. 2.5.5) uses these methane candidates to refine the learnable queries. Our hyperspectral decoder takes the encoded features from the encoder and refined queries from **QR** to generate the embeddings. The mask-prediction layer processes these embeddings along with the feature pyramid from the backbone layers to generate the final methane-plume segmentation prediction.

These individual blocks are discussed in more detail below.

2.5.3 Bandpass filtering for the Encoder

The HSI is processed by two parallel band-pass filters; a visible wavelength ($400 - 700nm$) (RGB) and a short-wave infrared wavelength ($2000 - 2500nm$) (SWIR) band-pass filter. The RGB filter results in a 3 channel output corresponding to the normal red, green, and blue wavelengths. The SWIR generates channels, approximately 5nm apart. The filtered outputs are $\mathbf{x}_{rgb} \in \mathbb{R}^{H_0 \times W_0 \times 3}$ and $\mathbf{x}_{swir} \in \mathbb{R}^{H_0 \times W_0 \times 100}$. Using $\mathbf{x}_{rgb}$ and $\mathbf{x}_{swir}$, two conventional CNN backbones (e.g. ResNet-50 [47, 48]) generate two feature maps respectively of size $\in \mathbb{R}^{H \times W \times N}$. Here $H = \frac{H_0}{32}$, $W = \frac{W_0}{32}$ and $N = 2048$ typically. We concatenate these feature maps along channel dimension and project through a 1×1 convolution layer to retain channel dimension of N. The resulting output is $f_{comb} \in \mathbb{R}^{H \times W \times N}$.

Following the standard architecture of transformer encoder from previous works [40, 45, 46, 49, 50], we reduce the channel dimension of f_{comb} using 1×1 convolution to $f_z \in \mathbb{R}^{H \times W \times d}$ and supplement position information by adding a fixed positional embedding $p \in \mathbb{R}^{H \times W \times d}$. The encoder consists of a stack of multi-head self-attention modules and

feed-forward networks (FFN). The encoded feature map is $f_e \in \mathbb{R}^{H \times W \times d}$:

$$f_e = \text{Encoder}(f_z, p) \tag{2.1}$$

2.5.4 Spectral Feature Generator (SFG)

In parallel, the input hyperspectral image is processed by the **SFG** module to generate methane candidates feature map f_{mc}, providing the **QR** module with spatial information to help the network delineate the methane plumes.

The **SFG** consist of a spectral linear filter (SLF) and a Feature Extractor (e.g.ResNet-50 [47]). The most common linear filtering approach for detecting CH_4 is to take each pixel from the input hyperspectral image $\{\mathbf{x}_{ij} \mid \mathbf{x}_{ij} \in \mathbb{R}^{1\times1\times N_0}\}_{i,j=1}^{H_0,W_0}$ and project it onto a CH_4 spectral absorption signature vector of same size [1]. This is to reduce the interference from ground terrain and amplify the CH_4 visibility in that pixel. Accurately modeling SLF is critical given that it is designed to reduce ground terrain interference.To model **SLF** we use the most common approach to matched filtering from information theory [51].

Spectral Linear Filter (SLF): The design of SLF is dependent on the spectral absorption pattern of CH_4 gas [1] and distribution of ground terrain. Since our signal of interest, CH_4, is very weak, traditional methods of linear filtering [16,17] are not effective. The conventional methods to whiten the ground terrain noise includes calculating the covariance ($\mathbf{Cov} \in \mathbb{R}^{N_0 \times N_0}$) of background by selecting a set of 10-15 adjacent columns $\{\mathbf{x}_i \mid \mathbf{x}_i \in \mathbb{R}^{1\times H_0\times N_0}\}_{i=1}^{W_0}$). However, in a given flight-line, the terrain changes frequently, from water bodies to bare soil, vegetation, buildings and other urban structures. Therefore single approximation of the covariance can not provide correct estimate of CH_4 and a localized context-based whitening will be more effective. To address this problem, we

took a very simple and effective approach of doing land cover classification and segmentation [52–54], and then compute covariance per class from the land cover. More details in supplementary materials. This improves the quality of methane candidates in presence of confusers (materials with similar spectral absorption patterns as CH_4) and also in cases where CH_4 concentration is low. The final **SLF** design with per class covariance is:

$$\mathbf{SLF}(\mathbf{x}_{ij}) = \frac{(\mathbf{x}_{ij} - \mu_k)^T \mathbf{Cov}_k^{-1} t}{\sqrt{t^T \mathbf{Cov}_k^{-1} t}} \; \forall \; (i,j) \in \text{class } k \tag{2.2}$$

where t represents the spectral absorption pattern [1] of CH_4 gas, and $\mathbf{Cov}_k$, μ_k are the covariance and mean of k^{th} class respectively. $\mathbf{x}_{ij}$ represents the pixel in input hyperspectral image at (i,j) index in k^{th} class. This operation generates a 2-D spatial CH_4 candidates map of size $\mathbb{R}^{H_0 \times W_0}$. Next this CH_4 candidates map is fed to a Feature Extractor to generate CH_4 candidates feature map f_{mc}. Details of the land cover segmentation/classification and complete SLF derivation are in the Supplementary materials.

$$f_{mc} = \text{FeatureExtractor}(\; \mathbf{SLF}(\mathbf{x}_{ij}) \; \forall \; i,j) \tag{2.3}$$

2.5.5 Query Refiner (QR)

Next the methane candidate feature map $f_{mc} \in \mathbb{R}^{H \times W \times d}$ is fed to the **QR** module along with a set of 100 learnable queries $Q \in \mathbb{R}^{100 \times d}$. The f_{mc} refines the learnable queries via cross-attention mechanism. This operation provides a narrow search space for the queries. The **QR** module follows a transformer decoder-like architecture inspired from [45, 46]. The randomly initialized queries $Q \in \mathbb{R}^{100 \times d}$ are first passed through a self-attention layer to attend to themselves. Next, these queries attend to our methane candidates feature map f_{mc} from **SFG** module through a cross-attention layer. The

methane candidates feature map serves as key-values pairs in our attention architecture. The output of **QR** is Q_{ref}.

$$Q_{ref} = \mathbf{QR}(f_{mc}, Q) \tag{2.4}$$

2.5.6 Hyperspectal Decoder

The Q_{ref} is fed to the decoder module along with encoder output f_e to generate output embeddings. Our hyperspectral decoder follows the standard architecture with a minor difference. There are no self-attention layers, just stack of multi-headed cross attention layers. The refined queries are transformed into output embeddings $E_{out} \in \mathbb{R}^{100 \times d}$.

$$E_{out} = \text{Decoder}(f_e, p, Q_{ref}) \tag{2.5}$$

2.5.7 Box and Mask Prediction

The decoder output embeddings (E_{out}) are fed to two Feed Forward Network (FFNs) and a Mask prediction layer. The outputs of the FFNs are the bounding boxes covering each CH_4 plume and a confidence score corresponding to each box. The mask-prediction module follows the standard segmentation head of DETR [40]. It computes multi-head attention scores of each embedding over the f_e (Eq. 2.1), generating a low-resolution heatmap for each embedding. To make the final prediction a Feature Pyramid Network [55] like structure is used. Each heatmap is designed to capture one methane plume. A simple thresholding is used to merge the heatmaps as final segmentation mask.

$$mask = \text{Mask\_pred}(E_{out}, f_e, f_{comb}) \tag{2.6}$$

2.5.8 Training and Inference

We train MethaneMapper in two stages; first we train bounding box detection corresponding to each CH_4 plume, and second by freezing the box detection network and training only the mask prediction module. We also trained both box and mask prediction modules end-to-end and achieved similar performance. We use a similar two-stage loss strategy for training MethaneMapper as that used in DETR [40]: first stage is the bipartite matching between the predictions and the ground truths both in bounding box and mask prediction, and then second stage is loss calculation for the matched pairs. The bipartite matching employs the Hungarian algorithm [40] to find the optimal matching between the predictions and the ground truths. After this matching, every prediction is associated with a ground truth. Next, we calculate the l_1 and $GIoU$ loss on both box and mask predictions and cross entropy loss for class prediction [40].

Inference: The inference pipeline is similar to training pipeline and can be implemented using approximately 50 lines of code. During inference, we first filter the detections with confidences below 50% and a per-pixel max to determine which pixels are predicted to belong to a CH4 plume.

2.6 Methane Hot Spots (MHS) dataset

Another significant contribution of this work is a large scale curated MHS dataset. It contains the AVIRIS-NG spectral data with wavelength ranging from $380nm$ to $2510nm$, a $5nm$ sampling [13], and capturing 432 channels per pixel. The images from the flightline are orthorectified and of size $\sim 23K \times \sim 1.5K \times 432$. The only currently publicly-available dataset with methane plume segmentation masks is the JPL-CH4-detection-V1.0 [2] dataset released by JPL-NASA in 2017.

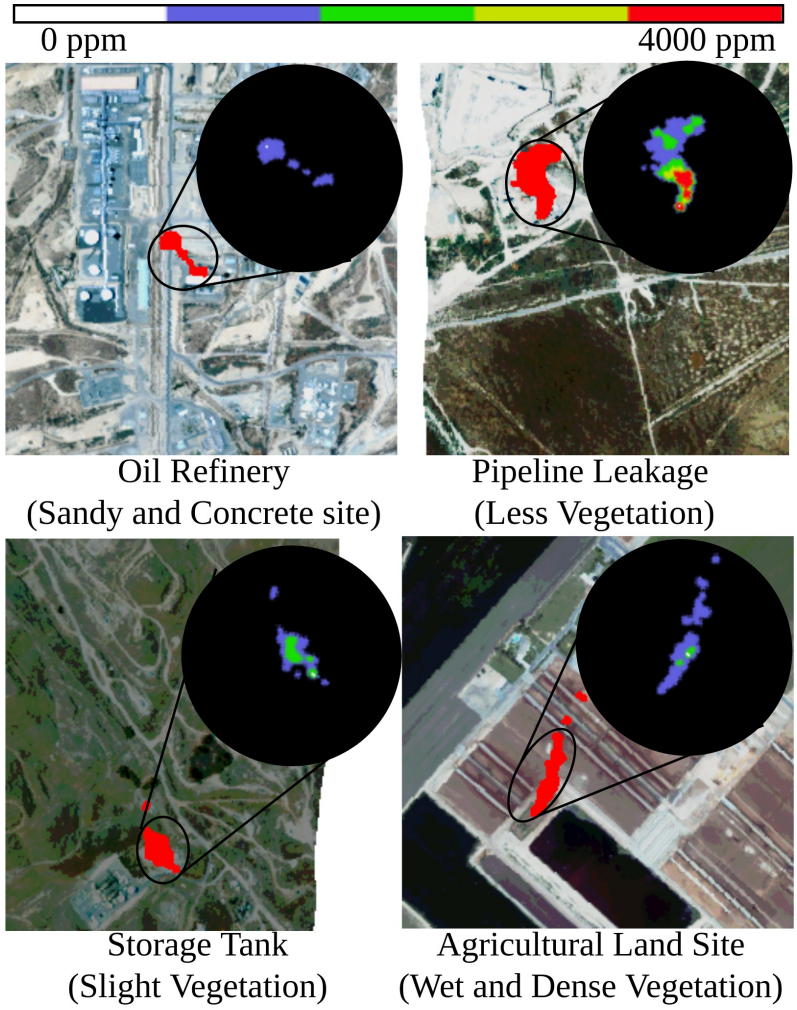

Figure 2.6: *Sample images from MHS dataset. The colormap in black circle shows concentration maps corresponding to the plume mask shown in red. We are showing different types of leakage sources and land cover types. For better visualization, we plotted the binary mask on color image created using visible bands of hyperspectral image.*

The MHS dataset has approximately 4000 plume sites corresponding to approximately 1200 AVIRIS-NG flightlines as shown in Table 2.1. MHS also has higher diversity data with flight lines spanning from 2015-2022 and covering terrain from 6 states– California, Nevada, New Mexico, Colorado, Midland Texas, and Virginia.

Dataset	MHS (Ours) Dataset	JPL-CH4 detection-V1.0 [2]
# plume sites	**3961**	161
# flightlines	**1185**	46
# point source	**3675**	114
# diffused source	**286**	57
Time period	**2015 - 2022 (8 years)**	2015 (1 year)
Segmentation Mask	**Yes**	**Yes**
Bonding box	**Yes**	No
Concentration map	**Yes**	No
Number of Regions	**6**	1

Table 2.1: *Statistics shows MHS dataset comparison with JPL-CH4-detection-V1.0 [2] dataset. Each flightline have multiple large and small plume sites. Each flightline have atleast 4 plume sites. The Point Source represents high concentration (*$300kg/hr$*) to leakage from sources like pipeline leak, storage tanks, oil and gas refineries. Diffused Source represent low concentration leakages from sources like biomass degradation in landfills. Our dataset is covers more diverse type of terrain over 6 states.*

Data Pruning: We selected AVIRIS-NG flight lines over varying regions as it covers a wide variety of CH_4 plume sources, such as leaks in oil and gas refineries, oil and gas extraction points, natural seeps, leaking underground storage tank, coal mines, dairy farms, landfill sites, and pipeline leaks. Along with varying emission sources, we selected regions with different types of ground terrains like, bare soil, rocks, mountains, light vegetation, water bodies and dense vegetation as shown with few samples in Fig. 2.6. Different types of ground terrain exhibit widely varying albedo and thus have a major impact on the quality of CH_4 detections as shown in Fig. 2.7. Given this, training models with diverse ground terrain data leads to a more robust model.

2.6.1 Concentration map and Segmentation mask

Concentration map is provided in the form of a matrix of spatial dimensions same as the flightline ($\sim 23k \times \sim 1.5k \times 1$). There is one concentration map per flight-line (orthorectified). It shows methane concentration in parts-per-million (ppm) per-pixel on

the ground. Pixel-regions with no methane presence are set to zero.

Segmentation mask provided in the format of a *"png"* image file with three channels and of the same spatial dimension as the corresponding flight line ($\sim 23k\times \sim 1.5k \times 3$). The segmentation mask is obtained from the concentration mask file by setting all pixel values above zero to represent methane plumes. We manually annotated *Point Source* and *Diffused Source* based on the type of ground terrain and concentration of methane gas. Following the benchmark dataset [2], three channels are used to color code *Point Source* (Red) and *Diffused Source* (Green). The distinction of *Point Source* and *Diffused Source* is derived from the JPL-CH4-detection-V1.0 benchmark dataset [2]. Our annotation style is also consistent with the JPL-CH4-detection-V1.0 benchmark dataset [2], so that both datasets can be merged seamlessly.

2.6.2 Constructing Concentration map

Concentration maps are generated by mapping expert-annotated methane-plume concentration maps to the ortho-corrected AVIRIS-NG flightlines. These methane plume annotations are systematically collected from a non-profit [42] entity. They provide concentration masks of methane emissions in 150×150 size patches along with location information from different sources (airborne sensors [41], satellites [44]). In order to map these patches from different sources to the AVIRIS-NG flight-lines, we use the pixel coordinate locations provided for both the annotations and flight-lines. We use this information to create a homography transformation to map each pixel to its corresponding location in the flight-line. Fig. 2.6 shows a sample of varying types of terrains with CH_4 segmentation mask in red and concentration mask in black circle. Details about matching the resolution, ortho-correction, and transformation are discussed in supplementary materials. The patch annotations are verified by experts visiting the physical location

of emission the same day [56]. Most of the regions in California are verified by physical visits by California Air Resource Board [42, 56].

2.6.3 MHS Statistics

MHS statistics and properties are summarized in Table 2.1.

Annotations: MHS provides both segmentation masks and concentration maps which enable development of deep learning algorithms than can produce both CH_4 plume location and concentration predictions.

Diversity: MHS dataset includes AVARIS-NG flightlines spanning 8 years (2015 - 2022) from six states in the U.S.: California, Nevada, New Mexico, Colorado, Texas, and Virginia.

Data Split: We divide MHS dataset into train/test splits of 80-20% with overlapping time periods and locations. Our dataset covers 6 states. Each state has subregions/locations (e.g. Permian basin) that are covered by multiple non-overlapping flightlines ($25k \times 1.5k \times 432$ pixels). These flightlines are split into train and test sets. In each set, we create patches ($256 \times 256 \times 432$ pixels) from the corresponding flightlines. From the patches/tiles, we take all positives patches (methane (CH_4)) and randomly sample equal number of negative (no-CH_4) patches. This is done for both train and test sets separately to balance the data and we refer to Section 2.8.2 for detailed ablation studies.

2.7 Experimental settings

Evaluation Metrics: Following the evaluation protocol of H-mrcnn [9] we report our performance in mean intersection-over-union (mIOU). Here, mIOU indicates the overlap between the predicted and the ground truth CH_4 plume masks. ED represents the

accuracy in plume core prediction. Additionally, as first stage of our two stage training procedure contains bounding box prediction, we also report our performance in predicting plume bounding boxes in terms of mean Average Precision (mAP) which tells us the effectiveness of MethaneMapper in eliminating the false positives in plume prediction.

Data Pre-Processing: Each input hyperspectral image is approximately of size $25000 \times 1500 \times 432$ taking up memory space of $55-60$ GB. We create tiles of each image in spatial domain, each tile is of size $256 \times 256 \times 432$ [9] with an overlap of 128. The CH_4 plume is available in very few pixels in the whole image, 90% of the tiles are negative samples (no methane, just ground terrain). We can not use the whole hyperspectral image because of GPU memory limitations

Implementation Details: The band-selectors module takes 432-channels hyperspectral image as input, the RGB band-selector picks 60 channel from $400nm - 700nm$ wavelength range and creates a 3-channel RGB image, the SWIR band-selector picks 100 channel from wavelength range $2000nm - 2500nm$. These input images are passed to two ResNet-50 [47] feature extractor backbones. The backbone networks are initialized with DETR [40] trained on COCO dataset [57] and input layer initialized randomly [58]. The transformer encoder-decoder and our query refiner have 6 layers and 8 heads. We initialized the transformer encoder-decoder with weights extracted and stripped from DETR [40] model. The dimension of transformer architecture is 256 and number of queries is 100. The **SFG** module takes in all 432-channels hyperspectral image and generates 1-channel output map of same spatial dimension as input. The feature extractor in **SFG** is ResNet-50 [47] initialized with DETR [40] trained on COCO dataset [57]. The decoder output embeddings are of size 512. The feature pyramid network in mask prediction module has 3 layers. More details are mentioned in supplementary materials.

	Methods	Back bone	SFG F.Ext.	#params	mAP	mIOU
JPL-CH4-detection-v1.0 Dataset						
1	Hu et. al	R-50	-	75M	0.26	0.48
2	H-mrcnn	R-50	-	353M	0.53	0.86
3	MM	R-50	R-50	**80M**	**0.63**	**0.91**
MHS (Ours) Dataset						
4	SpectralFormer	R-50	-	84M	0.33	0.41
5	UPSNet (stuff)	R-50	-	69M	0.32	0.38
6	UPSNet (stuff + things)	R-50	-	69M	0.29	0.35
7	DETR	R-18	*	33M	0.37	0.56
8	DETR	R-50	*	59M	0.44	0.59
10	MM	R-18	Linear Layer	39M	0.45	0.60
11		R-18	R-18	44M	0.52	0.63
12		R-50	R-50	**80M**	**0.59**	**0.68**

Table 2.2: *Comparison with baselines. "-" represent Not Applicable and "*" represent no **SFG** module and a random query used for transformer decoder. The top section shows performance on JPL-CH_4 dataset [2]. MethaneMapper achieves better results than heavily tuned H-mrcnn with $\sim 5\times$ fewer parameters. The overall detection accuracy is higher on this dataset because the type of ground terrain is uniform across all flightlines. In MHS dataset, MM outperforms multiple baselines as shown in rows 4-12. MM accuracy is lower in MHS than JPL-CH_4 dataset because MHS dataset has more variety of ground terrain spreading over 6 states*

2.8 Results

In this section we will discuss and validate all the design choices for MethaneMapper (MM) with ablations. We show that MM achieves state-of-the-art results in overall performance compared all other methods shown in Tables 4.2 & 2.3.

2.8.1 Performance comparison

Deep Learning methods: We trained MM with ResNet-50 [47] backbone on the same dataset that H-mrcnn [9] (JPL-CH4-detection-V1.0 [2]) was trained on for fair comparison. To align with H-mrcnn we used the same split and input image size. The MM model with 80M parameters trained for 250 epochs outperforms by significant margin

Methods	mAP	mIOU
LogReg [23]	-	0.05
SVM [28]	-	0.29
PCA + LogReg	-	0.06
PCA + SVM	-	0.31
MM (R-50)	**0.63**	**0.91**

Table 2.3: *Comparison with classical machine learning methods. "-" represent Not Available. The classical ML methods are not suited for the CH_4 detection task. MethaneMapper outperforms all methods on JPL dataset [2]*

the H-mrcnn model with 352M parameters. Results are summarized in Table 4.2 that includes the performance of MM on the new larger MHS dataset. We note that though the code for H-mrcnn is available, many of the modules are deprecated and can not be reproduced. The 'Backbone' column represents backbones used for feature extraction from input image,'SFG F.Ext.' represents the feature extractor in **SFG** module in MethaneMapper. We observed (qualitatively) that H-mrcnn fails to detect small CH_4 plumes with concentration lower than 100kg/hr while MM detects those.

We did evaluation by implementing 3 baseline models [40, 59, 60] shown rows 4-8 of Table 4.2. These methods were not designed for CH_4 detection task, therefore we needed to modify their input channel size. The poor performance of these methods may be attributed to the weak signal of interest in a high dimensional data, high number of confusers, and limited annotated data. Additionally, the only hyperspectral baseline method SpectralFormer [59] has low efficiency due its pixel-wise training scheme.

Classical ML methods: We trained and tested multiple existing machine learning based approaches that are used for methane detection, performance shown in Table 2.3. Logistic regression (LogReg) [23] and multinomial logistic regression (MLR) [24] failed to produce any meaningful detection with 90% false positive detections. We also trained a Support Vector Machine (SVM) [26,28] based classifier, it performed slightly better than

LR and MLR methods with an IOU of 21%. SVMs are prone to false positives detections same as Gaussian Mixture Models [28]. We observed that all traditional methods are not suited for the task of CH_4 detection. We also tested reducing the dimension using principal component analysis (PCA) or just taking bands which shows maximum CH_4 absorption. In the later case, the traditional methods performed better than using all 432 bands, this backs our idea of just using bands from SWIR region.

Qualitative results. Fig. 4.5 shows comparison of MM's mask and bounding box prediction with ground truth mask on different ground terrains. The Leakages are from different type of sources such as, oil refinery, pipeline and storage tank. MM makes correct predictions in varying scenarios. A detailed discussion on good and bad cases of detection is shown in Appendix MethaneMapper, chapter 7.

2.8.2 Ablation Studies

We did the experiments for ablation on MHS dataset with ResNet-50 as backbone and validate the design choices. One parameter is changed for each ablation and others kept at best settings. More ablations in Supplementary.

Spectral Feature Generator Module: In Table 4.2 lower section, we show the effectiveness of our **SFG** module for the query refiner block. Our baseline is standard implementation of DETR [40] for segmentation task represented by row-1 and row-2 of Tab. 4.2 lower section. Using CH_4 candidates feature from **SFG** improves the bounding box detection performance by 0.14 mAP and mask prediction by 0.09 mIOU. This demonstrate that guiding queries with CH_4 candidates feature generated by **SFG** produces better embeddings as compared to random queries.

Along with this, we explored the provision of CH_4 candidates feature at 2 places, (i) at input level concatenating it with f_{comb}; and (ii) as input to query refiner. We see

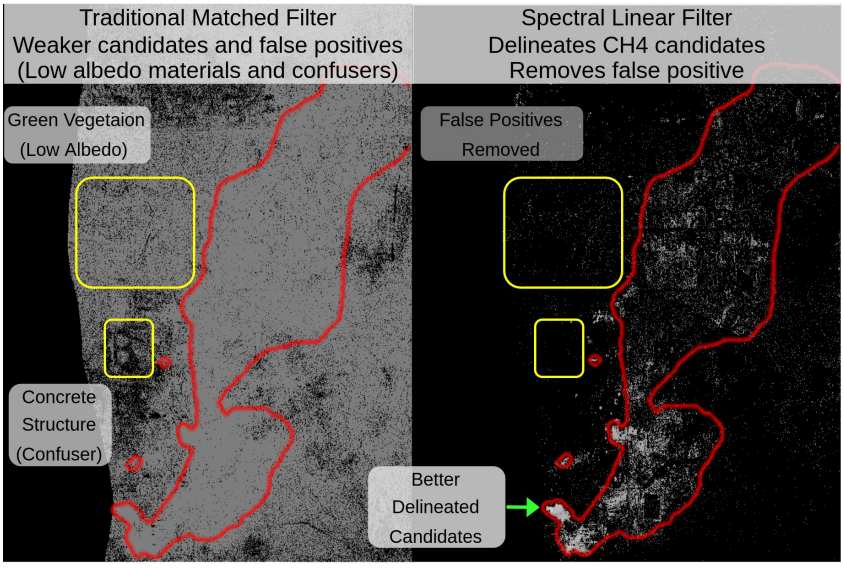

Figure 2.7: *Comparison of SLF with traditional filter in SFG module. White pixels represent methane and black no-methane. Red boundary represents ground-truth plume mask. SLF module generates better CH_4 candidates*

an improvement of 0.09 mAP and 0.08 mIOU when **SFG** module output is passed to query refiner. We hypothesize that this is because on concatenating with input features, the CH_4 candidates feature information gets lost, while as cross-attention with queries reduces the search space for decoder and generate better embeddings.

We also experimented with different types of feature extractors for **SFG** module, and observed that a Resnet18 or Resnet50 [47] is more effective than a 2 linear layer feature extractor as shown in Table 4.2.

Spectral Linear Filter: We experimented with SLF for computing covariance (*Cov*) using different subset of columns in the input hyperspectral image. We observed that the SLF is most effective when covariance is computed class-wise based on land cover. Class-wise *Cov* ensures that the radiance absorption by ground terrain is same for all the pixels while computing CH_4 enhancement. As can be seen in Fig. 2.7, SLF amplifies CH_4 candidate detection and reduces false positives. SLF leads to a 0.03 mAP improved

in detection compared to traditional filters. The prediction from MM is shown row-1 of Fig. 4.5.

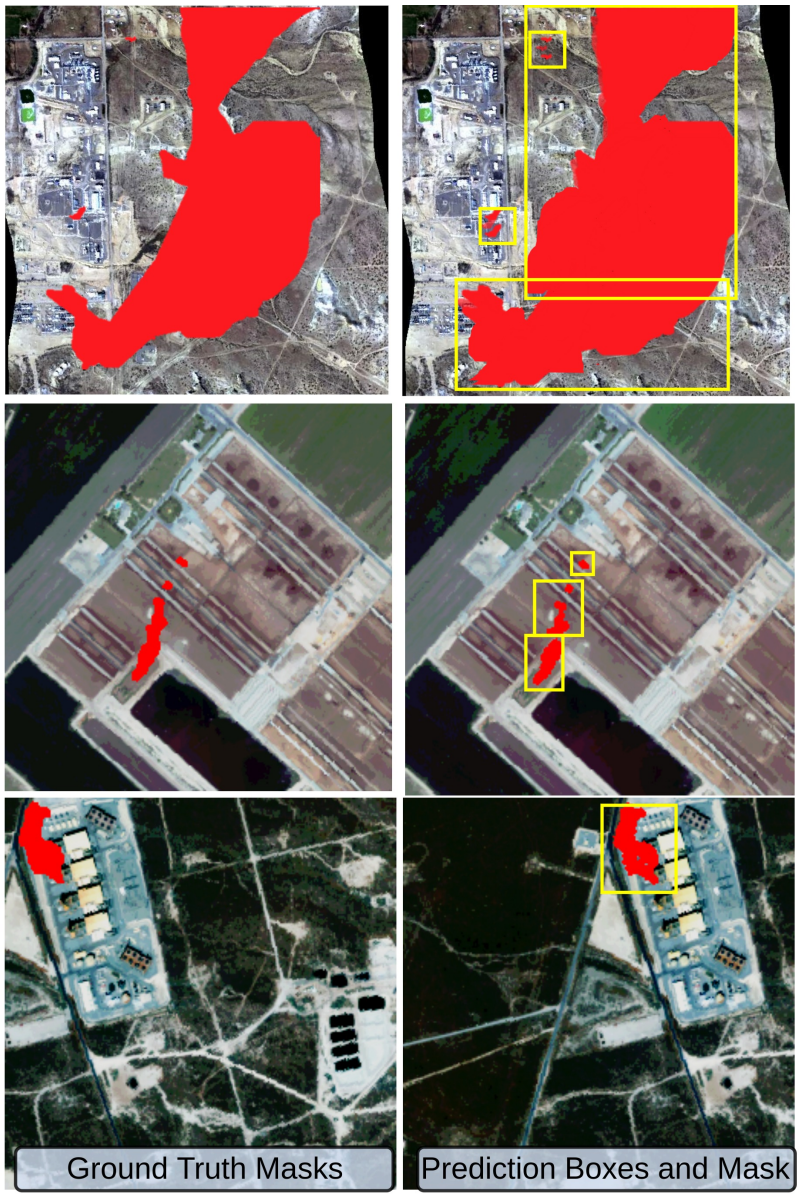

Figure 2.8: *Sample ground truths and predictions on MHS dataset. We show robustness of MethaneMapper predictions on different kind of ground terrain, rows 1 and 3 shows leakage at a refinery, row 2 shows leakage from pipeline in agricultural land, row 4 shows leakage from storage tank with concrete background.*

Geographic generalization: To assess the geographical generalization capabilities of

MM, we trained it on MHS data from all states except California and tested it on flightlines from California. We observed a slight drop of 0.04 mAP in detections. However, when trained on all data except Virginia, we noticed a significant drop of 0.09 mAP in detections. We attribute this to the fact that the land cover in Virginia is dense and moist vegetation, has a lower solar reflectance compared to the arid regions of California, Texas, and Nevada.

Temporal generalization: Testing MM on 2015 after training on data from 2016-2022 showed no performance drop.

Unbalanced test set: MM's performance dropped by 0.05 mAP on an unbalanced test set with only 10% positive samples (CH_4) and 90% negative samples (no-CH_4). This highlights the challenges in CH_4 detection. Future work will address this issue.

2.9 Conclusion

This paper presents MethaneMapper – a hyperspectral Transformer for methane plume detection. It utilize spectral and spatial correlations using a spectral feature generator and a query refiner, to accurately delineate the CH_4 plumes. Additionally, we curated a large-scale dataset for the task, a first of its kind, which is made available to all researchers via web platform BisQue [6]. The proposed MethaneMapper significantly improves upon the current methods in terms of detection and localization accuracy, as our extensive experiments demonstrate. Future work will extend the model to global monitoring [61] using multispectral satellite imaging data.

Chapter 3

Methane SatelliteMapper

In chapter 2 we discussed methane emission monitoring from aerial imagery. But aerial imagery has several limitations, such as limited coverage, and limited frequency of revisits at the same location. To address these limitation, we present the development of SatelliteMapper for processing multispectral data from space borne sensor on satellites. Our primary focus is on multispectral data from Sentinel-2 and LandSat-8 satellites. The initial exploration content of this chapter is published in a workshop in Neural Information Processing Systems (NeurIPS) 2022 [62] and 2024 conference.

Recent studies have shown that imagery from the multi-spectral instrument on the Sentinel-2 satellite is capable of detecting and estimating large methane emissions. However, most of the current methods rely on temporal relations between a ratio of shortwave-infrared spectra and assume relatively constant ground conditions, and availability of ground information on when there was no methane emission on site. To address such limitations we propose a guided query-based transformer neural network architecture, that will detect and quantify methane emissions without dependence on temporal information. The guided query aspect of our architecture is driven by a Spectral Linear Filter (SLF) approach, also discussed in this paper. Our network uses all 12 spectral

channels of Sentinel-2 imagery to estimate ground terrain and detect methane emissions. No dependence on temporal data makes it more robust to changing ground and terrain conditions and more computationally efficient as it reduces the need to process historical time-series imagery to compute a single date emissions analysis.

3.1 Introduction

The increases in atmospheric CH_4 have prompted governments to enact regulations and action plans such as the 'U.S. Methane Emissions Reduction Action Plan' in 2021 and the 'Global Methane Pledge Energy Pathway' in 2022 to curb CH_4 emission [63,64]. Accurately identifying and tracking the contribution of various sources to the methane budget will be paramount to enforce these regulations.

Given the strong potential of satellite-based instruments for data collection at high-frequency (multiple times a month) on global scales and even remote and hard-to-access regions, recent research has depicted the potential of deploying methane emissions analysis on public, global-mapping, multi-spectral instruments like the ESA Sentinel-2 mission [65,66,66–68]. With two polar-orbiting, sun-synchronous satellites, the Multispectral Instrument (MSI) onboard the Sentinel-2 satellites measures the reflected radiance from Earth in multiple bands covering various areas of the electromagnetic spectrum [44]. Among these bands, band 11 ($\sim 1500nm - 1660nm$) and band 12 ($\sim 2090nm - 2290nm$) are able to capture methane's SWIR (Short Wave InfraRed) absorption features at a spatial resolution of $20m^2$, leading to a large breadth of work and studies on using Sentinel-2 data to detect and quantify methane emissions [65, 69–71].

Most previous Sentinel-2-based methane analysis approaches use similar approaches to the methane column retrieval method in [65], building large parts of signal exploitation on an analysis of temporal deviation between times of excessive methane concentrations

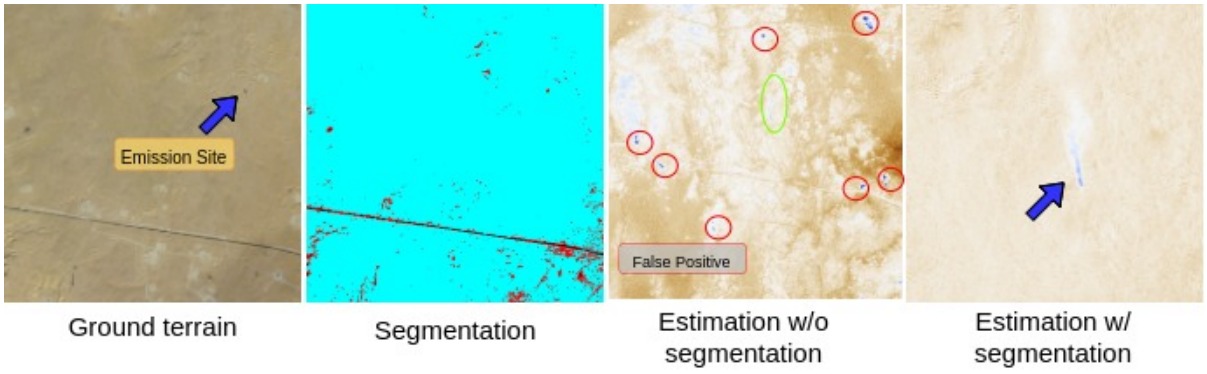

Figure 3.1: *Qualitative visualization of SFG intermediate steps and final estimation.*

in the atmosphere and times without, merged with ratios between methane-sensitive and less-methane-sensitive bands. While this method, and variations of it [71] have revolutionized capabilities of detecting methane emissions with public satellite data, the strong dependency on time-series analysis of spectral reflectance data expose the approach to risky assumptions on (a) knowing when emissions did not exist and (b) temporal albedo stability of the background - that the albedo of a certain area stays constant over time. Consequently, these assumptions lead to high amounts of false positives, especially in areas with heterogeneous, temporally deviating land cover [65, 69].

To overcome shortcomings of time-series based methane analytics methods, we propose a deeper exploitation of signals from other non-methane-sensitive spectral bands of Sentinel-2 multi-spectral data. We propose two approaches to explore this: first is Beer-Lambert law to model the drop in intensity of light as it passes through a medium [72], and second is matched filtering [9,15,18,73] as discussed in the previous chapter. Building on the recent advancements in signal processing techniques and the successful application of Machine Learning models for methane emissions analysis [9], we propose a two-step methodology: (1) the generation of potential methane candidates using a Spectral Feature Generator (SFG), and (2) the integration of SFG into a Transformer-based Convolutional Neural Network architecture [45, 74], as illustrated in Figure 3.2. Using the full spectral

response captured by Sentinel-2 instruments, we expect the SFG to support Signal to Noise separation by classifying confusing and hard-to-detect land cover types, artifacts and temporal deviations, such as water bodies, dark green vegetation, calcite, and white painted roofs which are never considered in current band/channel ratio methods. Using these classes, SFG computes statistical properties for each class separately for the whitening of background pixels.

3.2 Related Work

Over the past decade, several satellites such as GOSAT [75], OCO-2, TROPOMI [76] have been launched to detect and quantify greenhouse gas (GHG) emissions from fossil fuel activities, enabling continuous monitoring of carbon dioxide and methane levels. The Sentinel-5P (TROPOMI) mission [76], for example, provides hyperspectral imagery in the shortwave infrared (SWIR) spectrum, where methane (CH_4) strongly absorbs light. This satellite offers daily measurements of CH_4 column mole fractions across the globe, although at a relatively low spatial resolution of $5 - 7$ km [68, 77, 78]. This resolution is sufficient for identifying large emissions and regional anomalies, but it falls short in detecting smaller emissions ($\leq 25\ tCH_4/hr$) or pinpointing emissions to specific facilities in densely populated oil and gas regions [79].

To address this limitation, high spatial-resolution hyperspectral satellite imagery from instruments like PRISMA [66] and GHGSat-C [80] has been utilized, offering much lower emission detection thresholds. PRISMA and GHGSat-C can detect emissions as small as 0.2 tCH_4/hr and 0.1 tCH_4/hr, respectively. However, their tasking nature and relatively small fields of view limit their viability for persistent global monitoring. Airborne campaigns, such as those using AVIRIS [81], provide even better spatial resolution and lower detection limits, down to 0.01 tCH_4/hr. Other systems, like Scientific Aviation's

in-situ measurements, Kairos's passive imaging system [82], and Bridger Photonics' active system [83], offer detection thresholds as low as 0.005 tCH_4/hr, 0.01 tCH_4/hr, and 0.002 tCH_4/hr, respectively, depending on wind speed. Yet, similar to high-resolution satellite imagery, these airborne campaigns also suffer from limited spatial coverage.

Standard retrieval algorithms estimate vertical column concentrations of atmospheric methane by fitting a radiative transfer model to remotely sensed SWIR spectra. These algorithms typically analyze highly resolved spectra with full-width at half-maximum ranging from 0.1 to 10 nm and involving tens to thousands of spectral samples [66, 70, 84]. This high spectral resolution allows for the joint optimization of methane, other trace gases, and surface albedo from a single observation. However, methane column concentrations can, in principle, be retrieved with just two spectral measurements: one with methane absorption and one without. This can be achieved within a single spectral band by comparing observations of the same scene with and without a methane plume, or by using two adjacent spectral bands that differ in their methane absorption properties but are close enough to have similar surface and aerosol reflectance properties. Techniques like these have previously been employed to retrieve methane column concentrations using ground-based [85] and airborne [14, 86] remote sensing instruments. For instance, in this work the Spectral Linear Filter utilized Sentinel-2 bands 11 and 12 to demonstrate the retrieval of potential methane candidates by exploiting their differences in methane absorption properties.

The Sentinel-2 mission, although not specifically designed for methane detection, provides persistent multi-spectral imagery in the SWIR range with a revisit time of two to ten days. By leveraging its bands that are sensitive to methane, it is possible to detect and quantify large CH_4 emissions. As shown by Varon et al. [65], combining the two SWIR bands affected by methane increases the contrast of the plumes, and using a reference image taken at a different time without a methane anomaly can further enhance

this contrast.

In this context, our proposed SatelliteMapper advances methane plume detection by integrating both spectral and spatial correlations, providing a more robust and accurate delineation of CH_4 plumes. SatelliteMapper's approach effectively mitigates the issues of false positives and enhances the reliability of methane detection in complex environments, addressing the limitations of previous methods while utilizing the strengths of advanced spectral analysis techniques.

3.3 Approach

The proposed approach is a transformer [87] based neural network architecture with a SLF guidance. The input to the network is B1-B12 bands from Sentinel-2 Level 1C data [44]. The output is a segmentation mask that is used with a radiative transfer model for methane emissions analysis. The overall architecture (Figure **??**) presents 2 feature extraction blocks (ResNet [47]) as shown in Figure **??**, that will extract useful features from the input, the RGB channels of the image and a stack of B1-B12 bands from Sentinel-2 Level 1C data [44]. While the singled-out RGB image will provide information about land cover (e.g. Urban areas), the full B1-B12 stack provides additional land cover feature extraction (e.g. water bodies) while also capturing information about methane presence. Extracted features will be projected in a common subspace via a MLP [88] and passed on to the transformer encoder network along with positional information of each pixel in the image as shown in Figure 3.2. The output attention map [87] from the transformer encoder along with project features are passed onto the transformer decoder. The decoder network uses SFG to generate a query of the potential methane emission sites. The SFG is discussed in more detail Section 3.3.2.

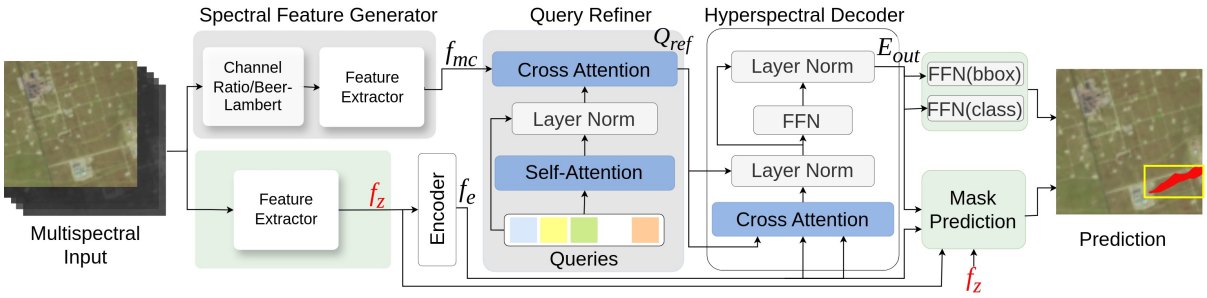

Figure 3.2: *Overview of Methane SatelliteMapper (MSM) architecture. Given a multispectral image, our RGB (B1, B2, B3) is passed to Feature Extractor which is a CNN backbone (ResNet) to extract features. Parallely, all channels (B1-B12) of the multispectral input as passed to Spectral Feature Generator (SFG). The SFG module generates methane candidates features. Next these candidates are sent to Query Refiner (QR) to refine queries. Then these queries decoded using encoded feature from Transformer Encoder. Finally each decoded query is used to predict a plume mask via Mask Prediction and, bounding box and class via FFNs (Feed Forward Network*

3.3.1 Dataset

We will be training and testing the proposed network on a mix of large-eddy based methane plume simulation data (synthetic data) [89] and single-blind release, human-labelled data [90]. The synthetic data includes images that contain simulated methane emissions on different types of background terrain. Each image is be a $10km \times 10km$ tile with 12 channels at different spatial resolutions per pixel. Next to the simulation data, we propose model validation to happen on manually selected emissions data from controlled ground releases [90]. The labels used in training corresponding to each multi-spectral image will be a binary segmentation mask (methane, no-methane) with the same spatial dimension as the input image. Along with that, we have a concentration mask, representing the concentration of methane per pixel in the mol/m^2.

3.3.2 Spectral Feature Generator (SFG)

The SFG is designed to generator potential methane candidates. These candidates can be generated either using the Beer-Lambert model or using Sentinel Enhanced Matched Filter (SEMF). Boths of these approaches are discussed in details below:

Beer-Lambert (Channel Ratio): We use a simple absorption model to characterize the attenuation due to the presence of methane. The Beer-Lambert law states that for a light source with intensity I_0 and a wavelength λ.

$$I = I_0 e^{-\Sigma_{i=0}^{N} A_i(\lambda) l_i} \tag{3.1}$$

where the light goes through N gases defined by their absorption $A_i(\lambda)$ and equivalent optical path length l_i defined as the product of the actual optical path and the concentration of the i^{th} gas. In our case, the N gases correspond to the atmosphere and I_0 is the sunlight in the SWIR spectrum. We can also reasonably assume I_0 to be constant for all wavelengths λ in each band respectively. Taking into account that the sensor of a satellite integrates over a band of wavelengths described by a sensitivity function s, the intensity of the light seen by a space-borne sensor becomes

$$I = I_0 \int s(\lambda)\alpha(\lambda) e^{-\gamma \Sigma_{i=0}^{N} A_i(\lambda) l_i} d\lambda \tag{3.2}$$

where the two passes through the atmosphere are taken into account in γ (which is a function of both the sun azimuth angle and the satellite view angle). The reflection coefficient of the ground is represented in the formula by the surface albedo $\alpha(\lambda)$.

In the presence of a methane emission, characterized by l_{leak}, the intensity of the light seen by the sensors becomes

$$I_{leak} = I_0 \int s(\lambda)\alpha(\lambda)e^{-\gamma\Sigma_{i=0}^{N} A_i(\lambda)l_i}e^{-A_{CH_4}(\lambda)l_{leak}}d\lambda \tag{3.3}$$

Supposing that we have both the exact same observation with and without a methane emission, it becomes very easy to detect the emission. Now to estimate the methane path length enhancement, we exploit the correlation between SWIR bands, similar to multiple-band-multiple-pass (MBMP) [65]. We take ratio of I_{leak} (represents the target date of leakage) and I (reference day when there was no leak).

Sentinel Enhanced Matched Filter (SEMF): $SEMF$ model is effective when the spectral resolution is relatively higher ($\geq$ 20 bands) then two bands of Sentinel-2 satellite, such as data from multispectral data from TROPOMI, PRISMA satellites. $SEMF$ is inspired by a deterministic linear match-filtering approach of finding CH_4 [9,15]. The linear approach is taking a n-dimensional (number of spectral channels) feature α, and apply as a dot product to each pixel (n-dimension) in the multi-spectral image to generate a scalar output per pixel. The α vector is "matched filter" [9, 15], making the process of finding the best-fitting α critical for signature exploitation in the ground terrain distribution at hand.

In ideal instances when there is no background (i.e. all white ground terrain) and just CH_4 gas present, the α is just the scaled version of the CH_4 signature ($\mathbf{t}$). However in real-world scenarios with spatially varying ground terrain this is not the case. For example, water has strong absorption of solar radiations, therefore the methane on such backgrounds has very weak visibility [91]. On the other hand, bare soil, rocks, etc have lower absorption, and the methane present in such background has strong visibility. An understanding of ground terrain and underlying albedo properties (especially in the methane sensitive spectral ranges) is critical to improve Signal to Noise ratios in our Sentinel-2 data. To account for spatial albedo differences in real-world scenes, we did a

land cover classification as shown in Figure 3.1 and use that land cover information to build our $SEMF$. The final SEMF used in our architecture is:

$$\hat{\alpha_k}(\mathbf{r}_i) = \frac{(\mathbf{r}_i - \mu_k)^T \mathbf{Cov}_k^{-1} \epsilon \mathbf{t}}{\sqrt{\epsilon \mathbf{t}^T \mathbf{Cov}_k^{-1} \epsilon \mathbf{t}}} \; \forall \; i \in k, \tag{3.4}$$

$$SEMF(\mathbf{r}_i) = \frac{(\mathbf{r}_i - \mu)^T \mathbf{Cov}^{-1} \mathbf{t}}{\sqrt{\epsilon \mathbf{t}^T \mathbf{Cov}^{-1} \mathbf{t}}} \tag{3.5}$$

where $\hat{\alpha_k}(\mathbf{r}_i)$ is the estimated methane column enhancement, $\mathbf{r}_i$ is the captured radiance at i^{th} pixel in the multispectral image, μ_k & $\mathbf{Cov}^{-1}$ are the mean and the inverse of covariance matrix for k^{th} class and ϵ represent the chemical properties of CH_4. An example of $\hat{\alpha_k}(\mathbf{r}_i)$ estimations is shown in column-4 of Figure 3.1. Our approach is simple and effective, it can be implemented with basic python code. Details about $SEMF$ can be found in the Appendix A at the end of the thesis.

3.4 Training and Inference

Training Process: The training process for the proposed transformer-based neural network architecture involves a specialized fine-tuning approach utilizing knowledge distillation from the original MethaneMapper model [10]. Knowledge distillation allows the transfer of learned representations from a pre-trained model to the new model, enhancing its ability to detect methane emissions with high accuracy. In this setup, the transformer encoder and decoder, which are critical components of the model, are kept frozen during training. This strategy ensures that the high-level representations learned by the original MethaneMapper model are retained and not altered during the fine-tuning process.

The focus of the fine-tuning is on the feature extractor, backbone, query refiner, and feed-forward network (FFN) layers. These components are adjusted to better suit the specific dataset and task at hand, ensuring that the model can accurately identify methane plumes from satellite imagery.

The training is conducted on a curated dataset containing 1200 methane plumes with corresponding ground truth locations [10, 13, 65, 72]. This dataset has been meticulously compiled from various reputable sources, including papers published in leading journals and conferences. By using this well-documented dataset, the model is fine-tuned to capture subtle variations in methane emissions, improving its detection capabilities.

The objective function used during training includes a combination of segmentation loss and a specialized loss function tailored to enhance methane plume detection. This loss function encourages the model to prioritize the identification of methane emissions while maintaining accuracy in land cover classification. The model is trained using a stochastic gradient descent (SGD) optimizer with a learning rate scheduler to gradually reduce the learning rate as training progresses, ensuring convergence to an optimal solution.

Inference Process: During inference, the trained model takes Sentinel-2 Level 1C data as input, specifically utilizing the B1-B12 bands to generate segmentation masks that highlight potential methane emission sites. The RGB channels of the imagery provide critical information about land cover, aiding in the contextual understanding of the detected emissions. These segmentation masks are then used in conjunction with a radiative transfer model to analyze methane emissions.

3.5 Results

In this section, we present the qualitative results from a pilot test conducted in collaboration with an Oil and Gas operator based in Los Angeles as shown in Fig. 3.3.

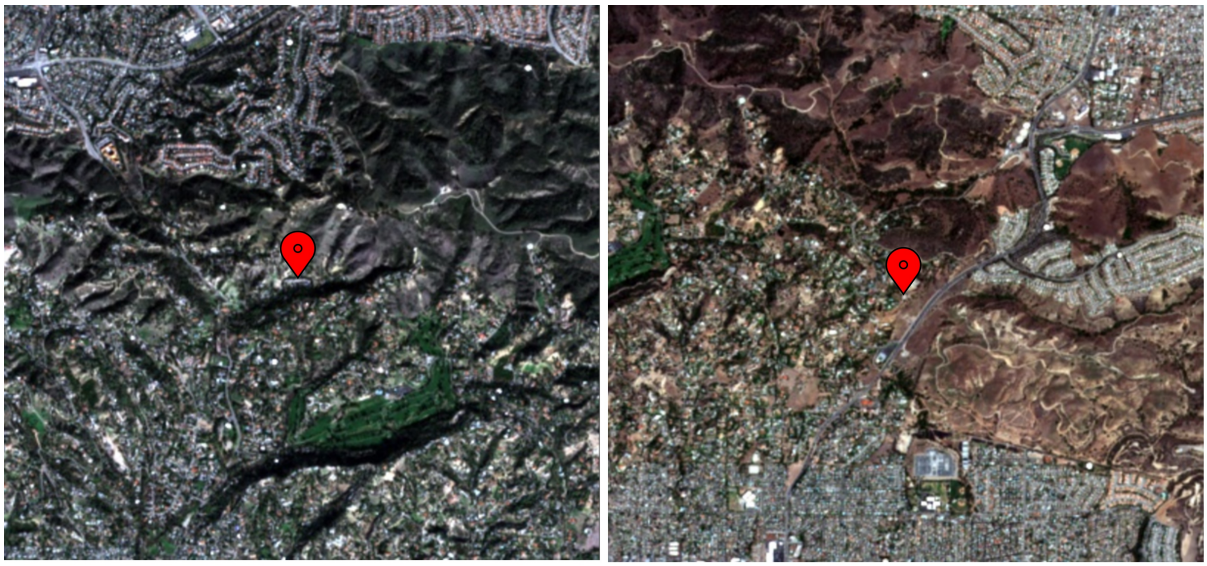

Figure 3.3: *Sites of Oil and Gas operator in Los Angeles where pilot was conducted. The images (10km × 10km in size) represent the RGB channels from Sentinel-2 satellite. The red pin represents the location of interest*

For the test, we focused on two sites of interest, using Sentinel-2 satellite data collected between 2016 and 2022. Our analysis on these sites revealed 11 instances of methane emissions at Site 1, while no emissions were detected at Site 2. Few sample visualizations are shown in Fig. 3.4. The relatively low number of detection instances can be attributed to the stringent regulations in Los Angeles, which likely result in fewer large emissions exceeding the 1000 kg/hr detection threshold of the Sentinel-2 satellite.

It's important to note that there may have been additional emission events that were not detected by our model. These undetected instances likely involved smaller concentrations of methane that were below the sensitivity threshold of the satellite sensor.

Despite this, the preliminary results are encouraging and highlight the potential for further refinement and validation of our model in future studies.

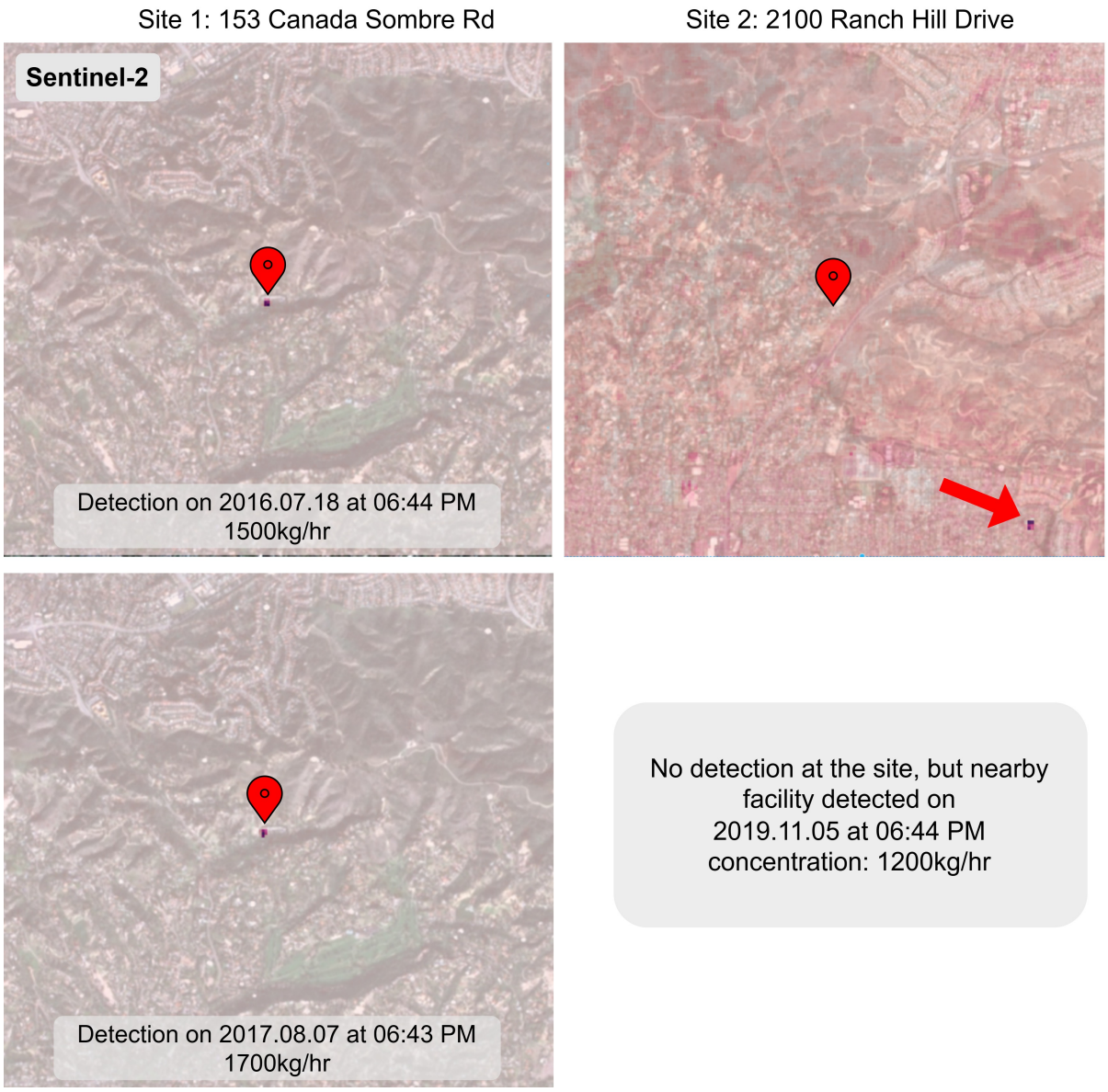

Figure 3.4: *Detection of methane emissions at the sites of Oil and Gas operator where pilot was conducted. The images shows an overlay of concentration mask on the RGB imagery. The red pin represents the location of interest. Total instances of detections were 11, we are only showing two here for site 1 and one for site 2. Site 2 shows an unknown detection in a neighboring area of almost 1200kg/hr concentration*

3.6 Conclusion

This paper presents Methane SatelliteMapper – a hyperspectral Transformer specifically designed for methane plume detection. Methane SatelliteMapper effectively leverages both spectral and spatial correlations through an advanced spectral feature generator and a query refiner to accurately delineate CH_4 plumes. Building on the success of our existing model, MethaneMapper, we demonstrate its effectiveness on satellite data by distilling its weights and fine-tuning Methane SatelliteMapper on a focused dataset of approximately 1,200 samples. Additionally, we present promising results from an ongoing pilot test conducted over six years of data from an oil and gas site in Los Angeles. Future work will focus on enhancing this model by collecting additional ground truth data samples and adapting the MethaneHotSpot dataset for use with Sentinel-2 satellite imagery. These efforts aim to further refine Methane SatelliteMapper's capabilities and expand its utility for global methane monitoring.

Chapter 4

WildlifeMapper

As discussed in the previous 2 chapters, guided feature extraction places a key role detection performance of the large vision models. In this chapter we talk about the application of our large vision models for detecting and identifying animals in aerial imagery. Specifically, We introduce WildlifeMapper(WM) , a model designed to identify animals in large open grasslands, where the animals occupy less than 0.01% of the total image pixels. It addresses the limitations of traditional, labor-intensive wildlife population assessments that are central to advancing environmental conservation efforts worldwide. While a number of methods exist to automate this process, they are often limited in their ability to generalize to different species or landscapes due to the dominance of homogeneous backgrounds and/or poorly captured local image structures. WM introduces two novel modules that help to capture the local structure, and the context of the objects of interest to accurately localize and identify them, achieving a state-of-the-art (SOTA) detection rate of 0.56 mAP. Further, we introduce a large aerial imagery dataset with more than 11k Images and 28k annotations verified by domain experts. WM also achieves SOTA performance on three other publicly available aerial survey datasets collected across four different countries, improving mAP by 42%. Source code and trained models are available

at Github [1]. The results of this chapter are published in Computer Vision and Pattern Recognition (CVPR) [92] 2024 conference.

4.1 Introduction

This paper introduces WildlifeMapper (WM) - an automated and scalable method for counting wildlife in aerial imagery. Aerial wildlife surveys are recognized as a cornerstone of modern conservation biology. By facilitating large-scale biological monitoring in remote landscapes, this technique has underpinned the ability to track changes in the abundance and distribution of wildlife across open landscapes for decades. However, traditional survey approaches often rely on manual observers to identify, count, and validate species of interest. This labor-intensive process can be time-consuming and error-prone, which potential to limits the utility of final results [93–95].

Automated approaches offer a promising alternative for efficient and accurate detection of wildlife in aerial survey images. Recent work, for example, illustrates how artificial intelligence is used to count a variety of species from the air, including antelope in grasslands [96], whales in the ocean [97], and seals on the beach [98]. When combined with advancements in low-cost, high-resolution imaging platforms (e.g., UAVs), these case studies underscore the potential for such data to significantly reduce the effort and cost of traditional wildlife census methods. However, the majority of these techniques struggle to generalize to new species or landscapes due to the dominance of homogeneous backgrounds and poorly captured local structures [99–102].

WM overcomes these limitations by adapting a novel application of the segment anything transformer model [3]. This model combines high frequency component correlations and spatial correlations in the image data to generate a map of potential locations

[1] *https://github.com/UCSB-VRL/WildlifeMapper*

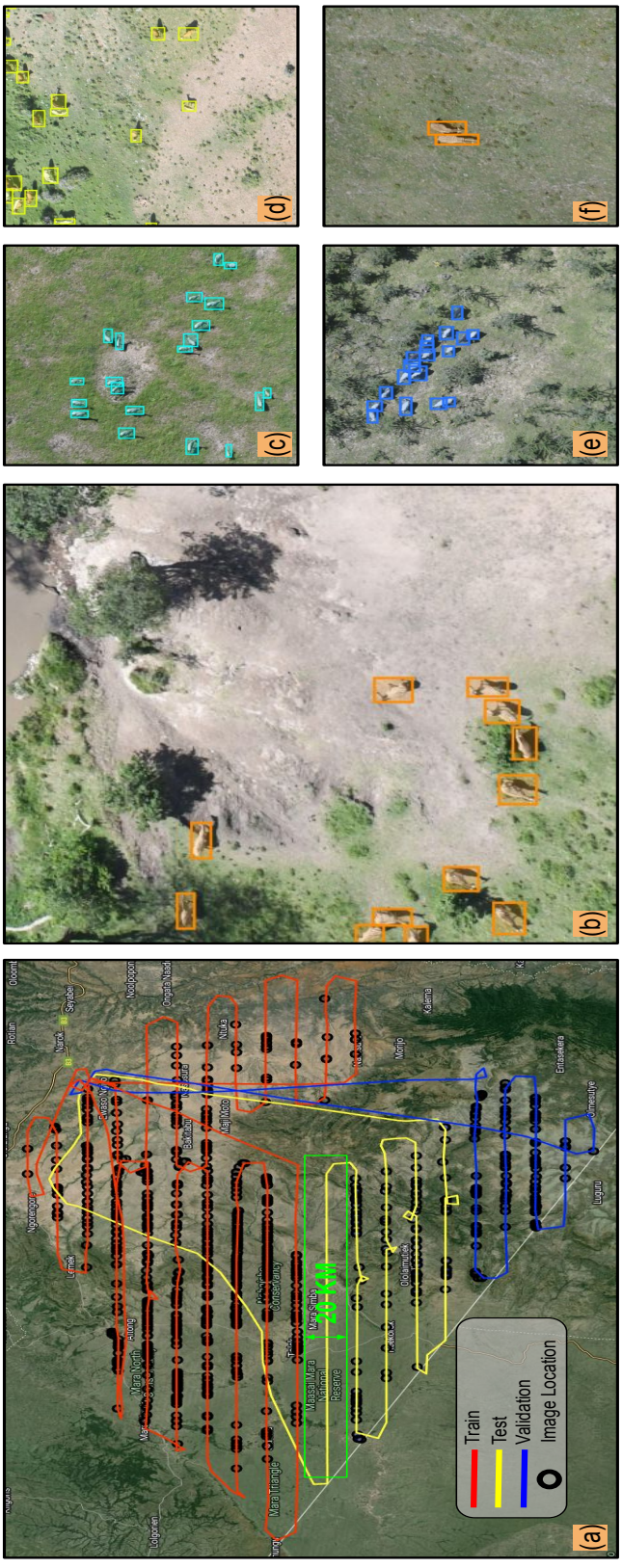

Figure 4.1: *Summary of Mara-Wildlife dataset. (a) Satellite view indicating the four flight trajectories, each represented in a different color. (b, c, d, e, f, g) Annotations of (b) zebra, (c) hartebeest, (d) cattle, (e) shoats (sheep and goats), and (f) zebra. These are the zoomed in verion of aerial images. The pixel footprint of object of interest is* $\leq 0.001\%$ *of the image. Best viewed in color.*

of objects of interest (i.e., wildlife, livestock). In addition, we address the challenge of identifying multiple species from a relatively small footprint in these images.

To demonstrate the WM analysis workflow, we provide a case study example across the Masai Mara Ecosystem in southwestern Kenya. Renowned for its rich biological diversity, the abundance of large mammals (such as buffalo (*Syncerus caffer*), giraffe (*Giraffa tippelskirchi*), and wildebeest (*Connochaetes taurinus*)) have declined precipitously over the past few decades [103]. Our analysis incorporated $11,151$ images of size 8400×5500 collected from a digital camera affixed to the bellyport of a Partenavia P68 airplane during Systematic Reconnaissance Flight (SRF) surveys. Part of these images were annotated by trained observers with $28,146$ annotations of 21 species of large mammals ($\geq 15kg$), providing an unprecedented opportunity to develop species detection models across a complex, heterogeneous environment. The dataset was systematically verified by trained observers as described in Section 4.3. Our contributions can be summarized as follows:

1. A novel, single-stage end-to-end approach for animal detection. The modules, a *High Frequency Feature Generator*, a *Feature Refiner*, and a *Query Refiner*, work together to improve upon the traditional methods of object detection in aerial imagery and enable generalizability across different habitats. The high frequency features reduce dependence on dominant backgrounds/landscapes.

2. An input patch embedding layer that is specifically designed to capture contextual information to help in identifying individual animal species.

3. The release of a new benchmark dataset via the data owner (Kenya's Wildlife Research and Training Institute - WRTI) once all approvals are in place. An international user community is already engaged in further enhancing these data and using the WM through the BisQue [6] platform.

Dataset	*# of annot. images*	*# of annot. tiles*	*# of species*	*# of annotation*	*Image size*	*GSD (cm)*	*Location*
Virunga	739	30069	6	5664	6000x4000	2.4	*DRC*
Garamba	158	6429	6	1611	6000x4000	2.0	*DRC*
AED	2067	69387	1	15581	5500x3600	2.4-13.0	*Bostwana, Namibia, South Africa*
Mara-Wildlife	**1012**	**77966**	**21**	**28146**	**8256x5504**	**1.45**	***Masai Mara National Reserve***

Table 4.1: *Comparison of Mara-Wildlife dataset with other publicly available dataset. Mara-Wildlife dataset has* $\times 3$ *more unique species than the total of all other datasets. Each image is significantly larger and higher ground resolution making* $77k$ *unique images of size* 1024×1024 *with* 21 *different animal species. GSD: ground sampling distance; DRC: Democratic Republic of Congo.*

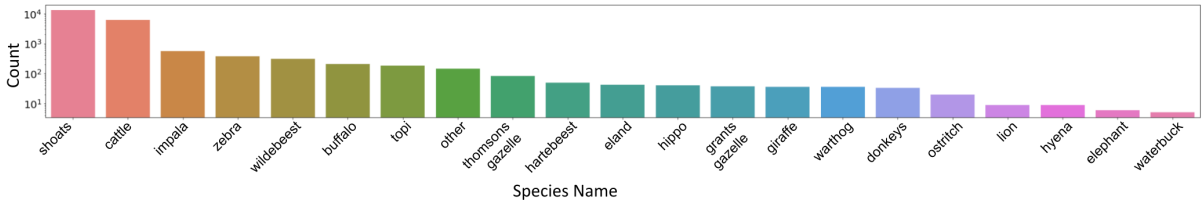

Figure 4.2: *Distribution of* ($\geq 15kg$) *mammals identified in digital imagery collected across the Masai Mara Ecosystem, Kenya.*

4.2 Related Works

Manual Methods: Aerial surveys using Front- and Rear-Seat Observers (FSO and RSO, respectively) are commonly used to inventory wildlife populations across open landscapes [104]. However, several important biases can impact these counts, including the experience level and fatigue of the human observers [105].

Deep Learning Methods: To address these issues, researchers have begun incorporating digital cameras on piloted aircraft and UAVs [101, 106, 107].

This minimizes the influence of observer bias while increasing transparency and reproducibility of results. For example, [108] replaced RSOs with an oblique camera mount system minimized the influence of observer bias while producing comparable estimates of large mammals under partial canopy cover. Similarly, [96] used a nadir mounted camera

to improve the accuracy and efficiency of manual counts of large antelope in an open grassland ecosystem.

However, detecting animals in the wild from aerial imagery poses many challenges. For example, most publicly available datasets for aerial object detection are focused on identifying relatively distinct features such as buildings, roads, vehicles, and other man-made structures [109–111]. Animals, tend to blend in with their surroundings [112], can be occluded by trees, exhibit considerable variation in color and pattern, or have behavioral adaptations that make them difficult to detect [99, 100, 113, 114].

[113] proposed a solution for this problem involving a two-branch CNN model based on AlexNet to perform animal recognition and localization. [99] evaluated three state-of-the-art object detection algorithms, including Faster-RCNN, Libra-RCNN, and RetinaNet on six African wild mammals. All three algorithms, however, showed poor performance in animal detection when animals were grouped closely together in herds. [102] adopted a segmentation approach, employing a UNet model to detect livestock from drone imagery. [114] uses a comparable model to analyze high-resolution satellite imagery, producing segmentation masks of wildebeest-sized animals, which are subsequently utilized for detection and counting.

Transformers: WM adopts a transformer architecture based on past success in modeling different types of aerial imagery. Examples include incorporating multispectral imagery for change detection [115], landcover classification [116], greenhouse gas (GHG) emission detection [10, 62] and RGB aerial imagery for object detection [117–119].

The most effective applications of transformer-based models have been tailored for standard object detection tasks [40, 120–123]. These works leverage the self-attention to model dependencies among the patches in an end-to-end fashion, unlike CNN-based models [124, 125]. However, when directly applied to aerial imagery, these models cannot effectively exploit the local structures as they divide the image into a sequence of

patches. This limits the detection of small-scale objects in a homogeneous and dominant background.

Dataset: The existing publicly available animal aerial imagery datasets are listed in Table 4.1. The Virunga dataset [99] was collected in Virunga National Park, Democratic Republic of Congo (DRC). This dataset contains 897 annotated images of 6 animal species. The Garamba dataset was collected in Garamba National Park, DRC and contains 7034 images. Only 158 images have been annotated, containing 7 animal species. The aerial elephant dataset (AED) [126] was collected across a mosaic of woodland, open shrubland, and grassland habitats in Botswana, Namibia, and South Africa. Only a single species (i.e., elephant) was targeted during SRF surveys. See Table 4.1 for details.

4.3 Mara-Wildlife Dataset

The Mara-Wildlife dataset is a distinctive dataset that captures the essence of the Masai Mara ecosystem through a compilation of 77966 images of size 1024 x 1024. This habitat is heterogeneous, including woodland, shrubland, and grassland vegetation with 21 unique animal species.

4.3.1 Image Collection

Flightline Details: Data collection was in collaboration with the Smithsonian National Zoo and Conservation Biology Institute (SNZCBI), Kenya's Wildlife Research and Training Institute (WRTI), the Kenya Wildlife Trust (KWT), and the Directorate of Resource Surveys and Remote Sensing (DRSRS). In March 2022, we fitted a Partenavia P68 with a Nikon D850 digital camera and collected high resolution (8256×5504) digital images. During data acquisition, the aircraft adhered to a predetermined flight trajectory, depicted in Figure 4.1, at 400 ft above ground level (agl). This trajectory was optimized to

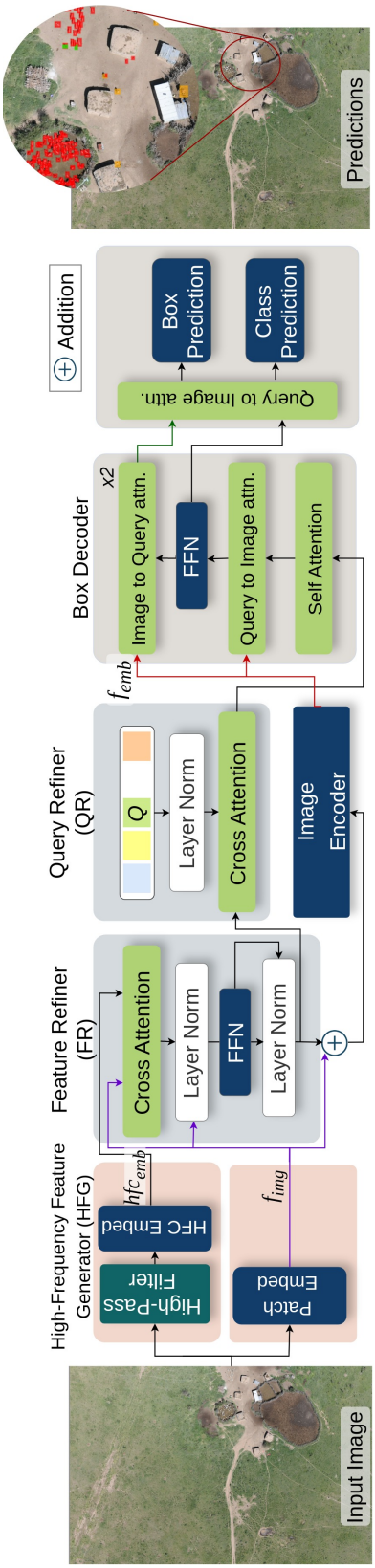

Figure 4.3: *Overview of WildlifeMapper (WM) architecture. Given an input image of size* $1024 \times 1024 \times 3$*, the High-Frequency Feature Generator (HFG) module generates information about potential location of object of interest. The Feature Refiner (FR) takes these potential location along with contextual features from Patch Embed layer and sent output to Image Encoder. In parallel, the Query Refiner (QR) incorporates the output of FR to refine learnable queries. Finally these queries are decoded using encoded features from Image Encoder and predict bounding box and class.*

encompass open grassland areas across the Masai Mara ecosystem, including the Masai Mara National Reserve, 22 adjacent private conservancies, and unprotected peripheral areas.

The aerial survey was conducted during a wet season period when the Serengeti migratory population of wildebeest have moved southward to locate more suitable forage in Tanzania. Thus, the survey primarily captured resident species, including wildebeest, zebra, topi, hartebeest, giraffe, and other large ($\geq$ 15 kg) antelope. Data collection was conducted in the early mornings (prior to 10:00 EAT) and late afternoons (after 15:00 EAT) when lighting conditions were optimal and animals are most active.

The geographical positioning of each image was acquired through a GPS system that recorded the plane's altitude, speed, and geographical coordinates. These data were then synchronized with the image using the image capture timestamps, enabling us to determine the geographic location of the centroid of each image. The camera was placed in the bellyport of the airplane, capturing a nadir view of the landscape every two seconds along the flight path.

4.3.2 Image Annotation

Initial bounding box annotations (21796) generated using AIDE platform [127] were exported in CSV format. These were then imported into BisQue [6] for further validation and correction. Finally, these annotations underwent validation by a single trained observer specializing in ecology, resulting in 28146 annotations in total.

4.3.3 Dataset Statistics

The Mara-Wildlife dataset showcases a detailed assemblage of wildlife, inclusive of 21 distinct species classes. The dataset is composed of approximately $77,966$ tiled images,

derived from 1,012 original rasters (Table. 4.1) The meticulous process of annotation has culminated in labeling 28,146 animals. Species identified include domestic cattle (*Bos taurus*), white-bearded wildebeest (*Connochaetes taurinus*), topi (*Damaliscus lunatus*), shoats (domesticated sheep and goats), kongoni (*Alcelaphus buselaphus*), waterbuck (*Kobus ellipsiprymnus*), impala (*Aepyceros melampus*), Grant's gazelle (*Nanger granti*), Thomson's gazelle (*Eudorcas thomsonii*), Cape buffalo (*Syncerus caffer*), zebra (*Equus quagga*), ostrich (*Struthio camelus*), Masai giraffe (*Giraffa tippelskirchi*), warthog (*Phacochoerus africanus*), eland (*Taurotragus oryx*), donkey (*Equus africanus*), hyena (*Crocuta crocuta*), hippopotomus (*Hippopotamus amphibius*), lion (*Panthera leo*), and elephant (*Loxodonta africana).*

These represent just a few of the many potential applications. We believe the Mara-Wildlife dataset, with its distinctive combination of rich imagery and detailed metadata, will stand as a foundational resource for both ecological studies and computer vision research, ushering in innovations and novel solutions.

In summary, Mara-wildlife dataset is a comprehensive dataset, that aids both foundational research and advanced studies in wildlife recognition using computer vision. More details about the dataset are in Appendix B.

4.4 WildlifeMapper Architecture

4.4.1 Technical Overview

WildlifeMapper's architecture is inspired by the success of the Segment Anything Model (SAM) [3], created to segment small/large (all sizes) of objects. Referring to Fig. 4.3, WM contains the following main components: (i) A patch embedding layer designed to capture long-range context, (ii) a *High-Frequency Feature Generator* (**HFG**),

(iii) a *Feature Refiner* (**FR**) followed by ViT based image encoder [3, 10, 87], and (iv) a *Query Refiner* (**QR**) module followed by a box decoder module. The input image is first processed through two separate branches, the patch embedding layer which captures long-range context [128] and **HFG** which suppresses all the low-frequency components in the image and generates a feature embedding. The **HFG** (Sec. 4.4.3) exploits prior knowledge that aerial images from areas such as forests, grasslands, and shrublands have a homogeneous and dominant background representing the dominant low frequency image content. Fig. 4.4 shows that on suppressing the lower frequencies, the object of interest is easy to locate.

The **FR** (Sec. 4.4.4) takes the embeddings from each of the two branches and generates a high quality embedding that contains information about potential locations of animals and captures the local context. The **QR** modules refines a set of learnable queries using the location information from **FR** module. These refined queries are passed to the box decoder. The box decoder takes the refined queries and encoded features image encoder to generate the final detection box and class of the object.

4.4.2 Patch Embed

The patch embeddidng layer utilizes a larger kernel convolution with an increasing dilation rate. This design rapidly expands the receptive field, allowing explicit extraction of features rich in contextual information. This approach is particularly beneficial for aerial imagery, where the small sized object makes classification based on appearance alone challenging. Contextual information thus becomes crucial for the accurate recognition of these objects.

4.4.3 High-frequency Feature Generator (HFG)

Along with patch embedding, the input image is processed in parallel by the **HFG** module to generate features with information about the location of the animal or cluster as shown in Fig. 4.4. The **HFG** module is inspired from the limitation of ViT models [87]. ViT models face challenges in efficiently utilizing local structures. They segment an image into patches and apply self-attention to model relationships, but this approach often falls short in capturing detailed local features [129, 130].

Local features in images are closely linked to high-frequency components [131, 132]. We hypothesize that suppressing low-frequency components can mitigate the influence of a dominant homogeneous background. To test this, we performed a discrete Fourier Transform (DFT) on the images, filtering out the low-frequency components before reconstructing the images, as shown in Fig. 4.4.

For a given input image $I \in \mathbb{R}^{H \times W \times C}$, where C is channel dimension, we compute Discrete Fourier Transform (DFT) of I. In next step we suppress the low frequency components with a controlling parameter and construct the image I with inverse DFT ($IDFT$).

$$I' = IDFT[hpf(DFT(I)] \tag{4.1}$$

$$F(u, v) = DFT(I) \tag{4.2}$$

$$I' = IDFT[hpf(F(u, v))] \tag{4.3}$$

where hpf is a high pass filter. Then we reduce the dimension of the reconstructed image I' via an embedding layer to generate embedding hfc_{emb} and pass them to the **FR** module. See supplementary materials for more details.

4.4.4 Feature Refiner (FR)

Next, the features from the patch embedding layer (f_{img}) and **HFG** (hfc_{emb}) are fed to the **FR** module. The f_{img} are refined with the hfc_{emb} via cross-attention mechanism. The **FR** is a simple module with cross-attention and linear layers [87]. The output contains information about the potential location of the object and the long-range context.

Following the standard architecture of SAM [3], we pass **FR** output to our ViT based image encoder supplemented with learnable positional embeddings p. The encoded feature map is f_{emb}:

$$f_{emb} = \text{ViT}\ [\mathbf{FR}(f_{img},\ hfc_{emb}),\ \ p] \tag{4.4}$$

Query Refiner (QR): The **QR** follows a transformer decoder like architecture and takes as input a set of 100 learnable queries $Q \in \mathbb{R}^{100\times d}$ and output of **FR** module. Here $d = 256$ is same as channel dimension of f_{emb} from image encoder. The **FR** output refines the Q via a cross-attention mechanism. The refined queries narrows the search space for box decoder module to accurately locate and identify object of interest. [**?**, 10].

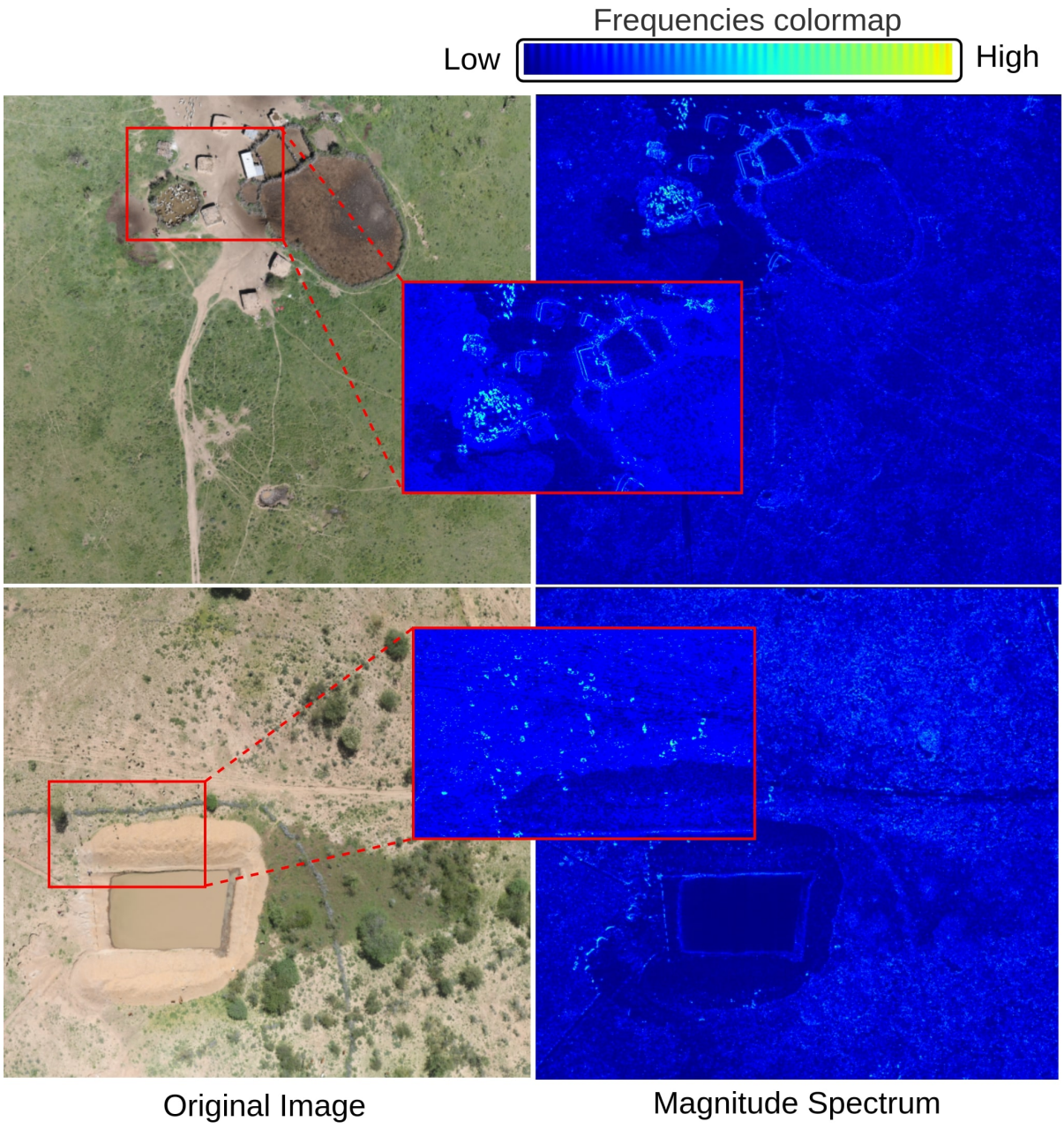

Figure 4.4: *The sample output visualization from the High-Frequency Feature Generator (**HFG**) module. The illustration shows the effectiveness of the module in suppressing the homogeneous and dominant background, while highlighting objects of interest (i.e., animals). The top image shows bomas, natural structures constructed to contain livestock, and paths that have been suppressed. Animals, however, are clearly identified, especially inside the boma. The bottom image shows a water body (a dam created for livestock) that has been suppressed by the module. Animals can again be highlighted throughout the image.*

4.4.5 Box Decoder

Next, the refined queries are sent to the box decoder module. We concatenate these queries with the f_{emb} and pass box decoder's self attention layer as qf_{emb}. This is inspired from the idea of *class_token* used by Vaswani et.al. [87] to make decoding process memory efficient. The transformer network takes 4 input variables, those are *position embeddings of queries and image embeddings*, qf_{emb}, and f_{emb}. Our transformer model uses two-way attention inspired from [3, 40, 133] and our box decoder uses self-attention and cross-attention in two directions (queries-to-image embedding and vice-versa) to update all embeddings. We keep the box decoder very light weight (two blocks). The top 100 (equal to number of queries) indexes from the output of the final block is passed to two separate MLP blocks to regress the output bounding box prediction and class of the predicted box.

4.4.6 Training and Inference

We've trained WM using a single-stage, end-to-end approach to determine bounding boxes and classify them. The loss strategy we applied for WM is akin to what's used in DETR [40]. Initially, we perform bipartite matching to align our model's predictions with the actual bounding box data. Then, we proceed to compute the loss for these matched pairs. To achieve the best possible match between our predictions and the real data, we use the Hungarian algorithm [134]. Once matched, each prediction is paired with its respective ground truth. We then measure the l_1 (L1 distance) and $GIoU$ loss for the bounding box and the cross entropy loss for the classification [40].

Inference: The inference pipeline is straightforward and similar to training code. During inference, we first filter the detections at 50% threshold and then use non-maximal suppression to remove any overlapping boxes.

4.5 Experiments

Train-val-test split: There is no data leak in train-val-test split. Fig. 4.1 shows the flight paths with location of each image represented by a circle. Each color coded flight path represents the train-val-test image set. No images taken while the airplane was cutting across the transects plus turn-around points as can be seen there are no circles. We created a spatial disjoint of $20km$ distance between the transects as shown in Fig. 4.1 and achieved consistent performance ($0.56mAP$).

Evaluation Metrics: We report our performance on multiple metrics. Following the protocols of standard object detection, we report the performance in mean average-precision (mAP) for detection and mean intersection-over-union (mIOU) for localization of animals. We also report a commonly used metric by the ecologists on the team, the class-wise mean absolute error (MAE) indicating the counting accuracy of each species:

$$\text{MAE} = \frac{1}{I}\sum_{i=1}^{I}\sum_{c=1}^{C}\left|\hat{n}_{i,c} - n_{i,c}\right|, \tag{4.5}$$

Where I is the number of images, C is the number of classes, $\hat{n}_{i,c}$ and $n_{i,c}$ are the predicted and ground truth counts for class c in image i.

Implementation Details: Each image taken from the drone is $8256 \times 5506 \times 3$. We create tiles for each image in the spatial domain, with the size of $1024 \times 1024 \times 3$ with 25% of overlap. The Patch Embed layer uses a single CNN layer with a large kernel of size 16×16 with stride 16. In the parallel branch, the *High-Frequency Feature Generator*, we use DFT to compute the Fourier transform, the mask is a binary disk with the radius set to 128. The HFC Embed layer uses 3 CNN layers with ReLU activation with a kernel of size 3×3 and a global average pool at the end. The Feature Refiner (**FR**) module consists of one cross attention layer with 1 linear layer. The image encoder is a pre-trained ViT

model [3] with 24 transformer layers and 16 heads. The Query Refiner (**QR**) module takes in 100 queries each of channel dimension 256, those are cross attended with hfc_{emb} output. The box decoder contains 3 layers of two-way attention with 8 heads. We train WM with AdamW optimizer [135] setting the learning rate to 10^{-4} for the **FR**, **QR** and box decoder with a weight decay to 10^{-4}. We set the learning rate for the Patch Embed and HFC Embed layer to 10^{-5}. We load the image encoder with pre-trained weights from segment anything [3] and keep it frozen.

Data Augmentation: In the train and test datasets, we incorporated an equal number of images without any objects to assess the model's robustness against empty background images. We applied multiple data augmentation techniques, including HSV (hue, saturation, and value) (10%), rotation (5%), translation (10%), affine transformation (20%), scale (10%), shear (5%) and mosaic (70%) augmentation [125]. Mosaic augmentation is proven to be the most effective, with an improvement of 0.07 mAP

Hard Negative Mining: After training for 100 epochs, we take all the False Positives (FP) predictions having IOU$\leq$ 0.10 with ground truth box and mark them as background class. Then fine-tuning for 20 epochs improved the performance of FP reduction for detecting rocks, dead tress or other artifacts on ground as animal.

4.6 Results

4.6.1 Performance Comparison

We trained WM separately on Mara-Wildlife dataset and Virunga-Garamba-AED dataset for comparison with existing works, see Table 4.2 for a summary of the results on all of the tested datasets. The mAP values are compared for IoU of 0.50-0.95 and 0.50. We also provide the average counting error in animal counting per image. We trained

	Methods	#epochs	mAP	mAP50	Counting Error
	Mara Wildlife Dataset				
1	Faster-Rcnn [124]	100	0.24	0.58	2.59
2	DETR [40]	200	0.22	0.57	2.75
3	Co-DETR-R50 [136]	100	0.27	0.66	2.72
4	Co-DETR-swingL [136]	100	0.28	0.65	2.60
5	Yolo v5 [137]	100	0.30	0.67	2.12
6	Yolo v8 [138]	100	0.27	0.61	3.97
7	LSKNet [118]	100	0.29	0.66	-
8	DroneDetect [139]	100	0.18	0.48	-
9	**WildlifeMapper**	120	**0.56**	**0.79**	**1.9**
	Virunga-Garamba-AED Datasets				
1	Faster-Rcnn	120	0.34	0.65	0.27
2	DETR	200	0.30	0.62	0.45
3	Yolo v5	100	0.48	0.78	0.12
4	Yolo v8	100	0.48	0.77	0.42
5	WildlifeMapper	80	**0.68**	**0.85**	**0.11**

Table 4.2: *Comparison with baseline models. The top section shows performance on species detection on Mara-wildlife dataset and low section shows performance on the mixed dataset from Virunga-Garamba-AED dataset. The overall detection accuracy is generally higher in Virunga-Garamba-AED dataset because there are only 6 species and the terrain is quite similar in all images.*

	HFG	FR	QR	mAP
WM	✗	✗	*	0.46
	✓	✓	*	0.54
	✓	✗	✓	0.49
	✓	✓	✓	**0.56**

Table 4.3: *HFG module effectiveness in refining the image features and queries. "✗" represents not used, "✓" represents used and "*" represents that random queries are used but there was not refining with HFG features.*

all the baseline models with the default set of conditions on Mara-Wildlife dataset. We merged the Virunga, Garamba, and AED datasets and created the train-val-test split according to [99]. The combination of these datasets contains 6 unique animal species and diversity of landscapes such as woodland, savannahs, open shrubland, and grasslands across multiple countries – Democratic Republic of Congo (DRC), Botswana, Namibia, and South Africa. **WM outperforms all current SOTA methods by a significant margin as shown in Table 4.2.** We note that in [99] the authors did not make the code base or trained model public, hence we could not verify the results. We implemented these methods from the original public repositories and trained according to the training strategy detailed in [99]. We attribute the model poor learning performance due to salient features of the homogeneous background being learned more than the object of interest. The detection of the object of interest is then dependent on the landscape properties instead of object properties. Hence when used on a slight variations of landscapes for the same object, the models struggle to detect. This limitation is specifically addressed in the WM, where the **HFG** modules suppresses the background and highlights the object of interest.

Qualitative results: Fig. 4.5 shows the quality of detection by WM in different scenarios. Those include, detection when animal is partially visible under a tree, or a big clustering. WM makes correct predictions in varying scenarios. Some examples are shown in the Appendix B.

4.6.2 Ablation Studies

We performed all ablation experiments on Mara-Wildlife (MW) dataset and validate the design choices.

High-frequency Feature Generator Module: In Table 4.3, we show the effectiveness of the **HFG** module. We experimented with **HFG**'s output in 3 ways: first, we passed **HFG**'s output to Feature Refiner (**FG**) module. It leads to significant improvement in detection by 0.09 mAP over the baseline. This demonstrate that providing potential location candidates features to image encoder module produce better embeddings. Second, we pass the **HFG**'s output to Query Refiner (QR) module only. This leads to an improvement of 0.03 mAP over baseline. This shows the effectiveness of guiding queries with location candidates features. In the third case, we passed the **HFG**'s output to both **FR** and **QR** modules and achieved an improvement of 0.11 mAP over the baseline. We hypothesize that this reduces the dominance of features from homogeneous and dominant background in aerial imagery. The also observed this while testing WM across flightlines different types of terrain such as green grasslands, dry grasslands, and forest areas.

Feature Refiner Module: We tested hfc_{emb} and f_{img} merging by 3 ways: addition, concatenation and cross-attention. Cross-attention is most effective, because with addition and concatenation, the hfc_{emb} get lost, while cross-attention generates better embeddings giving attention to potential location candidates.

Kernel Size: We observed that a larger kernel size of 31×31 results in reduced misclassification. For example, a *topi* or *warthog* cannot be found inside a *boma* because only domestic species are kept in *bomas*. So context helps in making the right class detection. We observed an improvement of 0.02 mAP. Experiments were done in 3 kernel sizes. **1.**

7×7: 0.55 mAP, **2.** 16×16: 0.558 mAP; **3.** 32×32: 0.57 mAP.

Query Refiner Module: We experimented with only providing random queries and guiding the queries with **HFG** module output. We merged them with direct addition or concatenation and cross-attention. With cross-attention, we observed an improvement in performance of 0.03mAP.

Geographic generalization: To test the geographic generalizability of WM across different terrains, we trained WM only on images from Kenya; and tested on images from Democratic Republic of Congo, Botswana and Namibia. The test was done on 4 common species which were present in both the ecosystems. WM achieved a detection performance of 0.48 mAP. This shows the adaptability of WM across varying landscapes.

Domain generalization: We train-test WM on a different domain, a bird species tern [140], commonly found on/near water bodies. Live in huge clusters. WM achieved the accuracy of 0.71 mAP. Showing adaptability of WM across different domains.

Failure Cases: The detections from WM are inaccurate when animals are clustered in shadows, such as when animals are located inside bomas and the sun angle makes a strong shadow on the enclosure. These are some of the difficult cases shown in Fig. 4.6. Other cases of false positives are the small rocks or trees that sometimes resemble animals. Some examples are shown in the Appendix B.

4.7 Conclusion

This paper presents WildlifeMapper (WM) - a transformer based approach for the detection of animals of varying densities and sizes across natural backgrounds. The WM utilizes a high frequency features generator, feature refiner, and query refiner to

accurately locate and classify 8 animal species. WM stands to significantly improve the efficiency and accuracy of wildlife monitoring and conservation efforts. Future work will extend this model and dataset to a larger number of species and habitats.

Community Adoption The practical use of WildlifeMapper (WM) extends beyond theoretical and computational success and is operationalized through BisQue [6]. WM is made available to users through a series of training modules that demonstrate how to *(i)* upload digital imagery, *(ii)* create annotations, *(iii)* apply and/or improve existing models, *(iv)* evaluate model fit, *(v)* improve annotations and re-fit models, and *(vi)* generate summary statistics. In the future, we envision WM to be of great value to ecologists, wildlife managers, and government officials, providing accurate information about the state of wildlife populations in near real-time, facilitating decision-making processes, and improving the conservation of ecosystems globally.

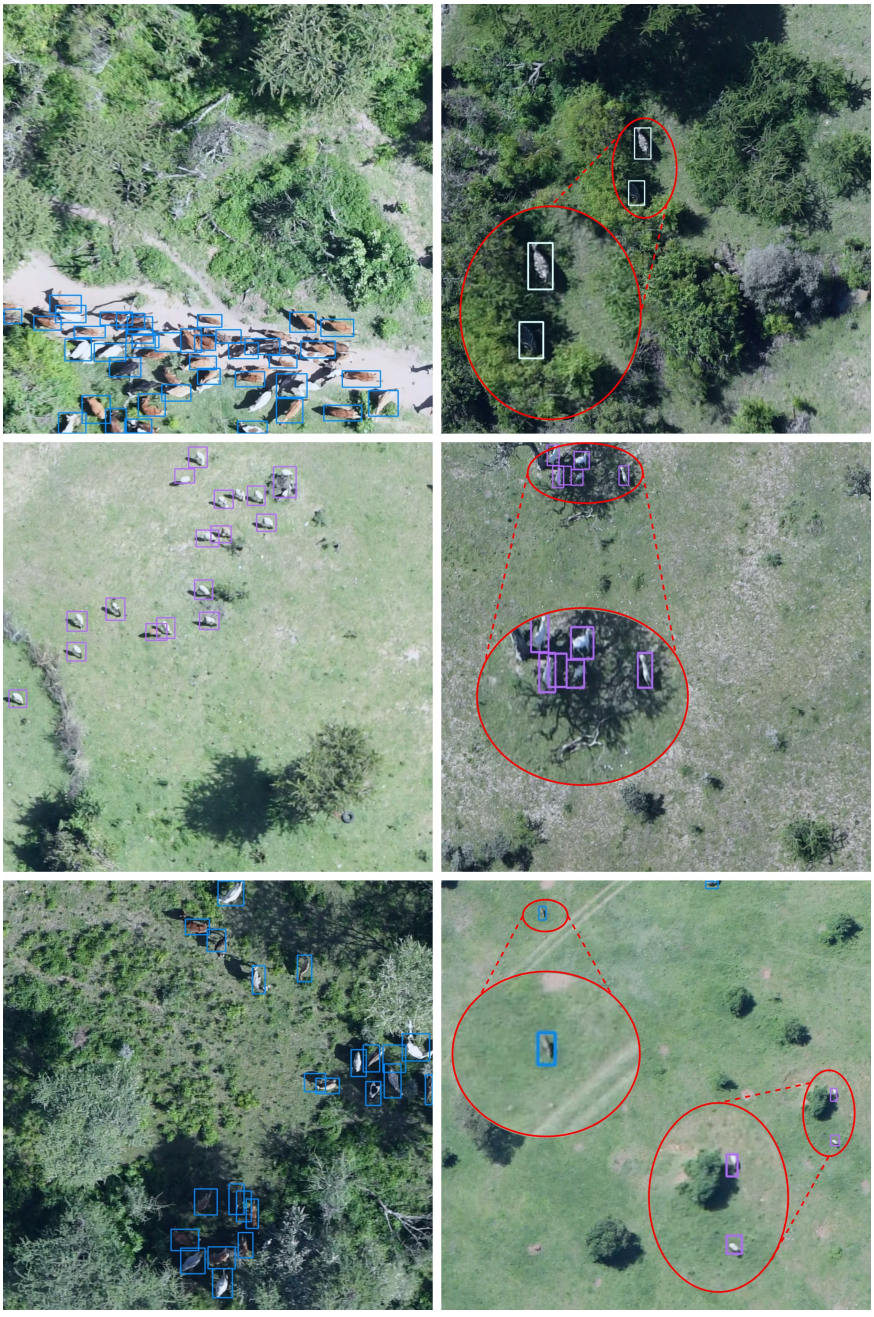

Figure 4.5: *Qualitative results. The top row highlights examples of crowded and partially occluded scenes. Row-1, Column-1 & Row-3, Column-3 shows examples where animals are obstructed by shadows. The zoomed-in box in Row-3, Column-3 shows a zebra partially occluded. The bounding box color is coded according to class names: shoats-"hot pink", cattle-"deep sky blue", zebra-"light yellow".*

Figure 4.6: *Failure cases. Left shows an example where animals are occluded by shadows and are difficult to detect. Right shows an example of rock detected as an impala, emphasizing the difficulty in differentiating objects in the image from animals of interest.*

Chapter 5

StressNet

This chapter explores the biomedical applications of our large vision models. In this chapter we explored the key role of feature guiding in videos. The video data contains thermal imagery of the human face. We explored feature guiding in temporal domain using which we built StressNet. StressNet processes this data to estimate physiological signals such as Electro Cardio Graph (ECG), Impedance Cardio Graph (ICG), and Initial Systolic Time Interval (ISTI). These signals are used to determine whether a subject is experiencing physical stress. The results of this chapter has been published in Winter Conference on Applications of Computer Vision (WACV) [48] 2021.

Precise measurement of physiological signals is critical for the effective monitoring of human vital signs. Recent developments in computer vision have demonstrated that signals such as pulse rate and respiration rate can be extracted from digital video of humans, increasing the possibility of contact-less monitoring. This paper presents a novel approach to obtaining physiological signals and classifying stress states from thermal video. The proposed network–"StressNet"–features a hybrid emission representation model that models the direct emission and absorption of heat by the skin and underlying blood vessels. This results in an information-rich feature representation of the face,

which is used by spatio-temporal network for reconstructing the ISTI (Initial Systolic Time Interval : a measure of change in cardiac sympathetic activity that is considered to be a quantitative index of stress in humans). The reconstructed ISTI signal is fed into a stress-detection model to detect and classify the individual's stress state (i.e. stress or no stress). A detailed evaluation demonstrates that StressNet achieves estimated the ISTI signal with 95% accuracy and detect stress with average precision of 0.842. The source code is available on Github[1]

5.1 Introduction

As the world has come to a standstill due to a deadly pandemic [141], the need for non-contact, non-invasive health monitoring systems has become imperative. Remote photoplethysmography (rPPG) provides a way to measure physiological signals remotely without attaching sensors, requiring only video recorded with a high-resolution camera to measure the physiological signals of human health. Much of the recent research in the area of rPPG [142] has focused on leveraging modern computer vision based systems [143–146] to monitor human vitals such as heart rate and breathing rate. More recent work has expanded these methods to detecting more complex human physiological signals and using them to classify stress states [144–146].

Whereas all recent datasets for rPPG only collect electrocardiogram (ECG) as the cardiovascular ground truth signal, here we recorded both ECG and impedance cardiography (ICG). ICG is a noninvasive technology measuring total electrical conductivity of the thorax. It is the measure of change in impedance due to blood flow. With these two signals, we have the ability to estimate more accurate quantifiers of cardiac sympathetic activity [147]. Two common metrics are pre-ejection time (PEP) and initial systolic time

[1] *https://github.com/UCSB-VRL/StressNet-Detecting-stress-from-thermal-videos*

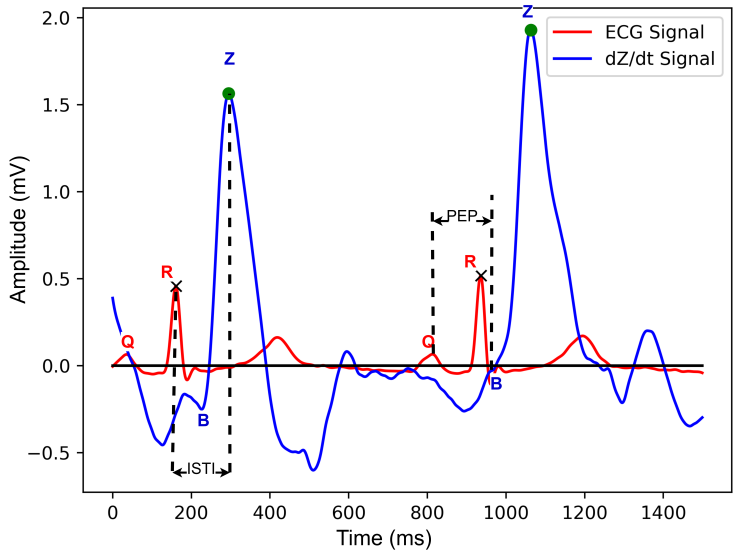

Figure 5.1: Example of ECG and $\partial Z/\partial t$ waveforms computed from the present data. $\partial Z/\partial t$ represents the change in impedance recorded by ICG (Z) signal with time. After each ECG peak value there exists an $\partial Z/\partial t$ peak value. The time difference between these two values is known as the initial systolic time interval (ISTI).

interval (ISTI).

PEP is the strongest cue for cardiac sympathetic activity. It is defined as the interval from the onset of left ventricular depolarization, reflected by the Q-wave onset in the ECG, to the opening of the aortic valve, reflected by the B-point in the $\partial Z/\partial t$ (derivative of ICG or Z) signal [148,149] as can be seen in Figure 5.1. Unfortunately, measuring PEP from ECG and $\partial Z/\partial t$ signals is quite difficult as the Q and B points that define PEP are subtle and very difficult to pinpoint [150,151]. Accuracy of methods to estimate PEP are low and precision differs highly among studies [150, 152]. Instead, ISTI can be used as a reliable index of cardiac sympathetic activity [147]. ISTI is a straightforward calculation defined as the time difference between the consecutive peaks of ECG and $\partial Z/\partial t$. ISTI is considered a strong index of myocardial contractility [151,153] and numerous efforts have shown that ISTI can be used to analyze different physiological phenomena e.g. stress,

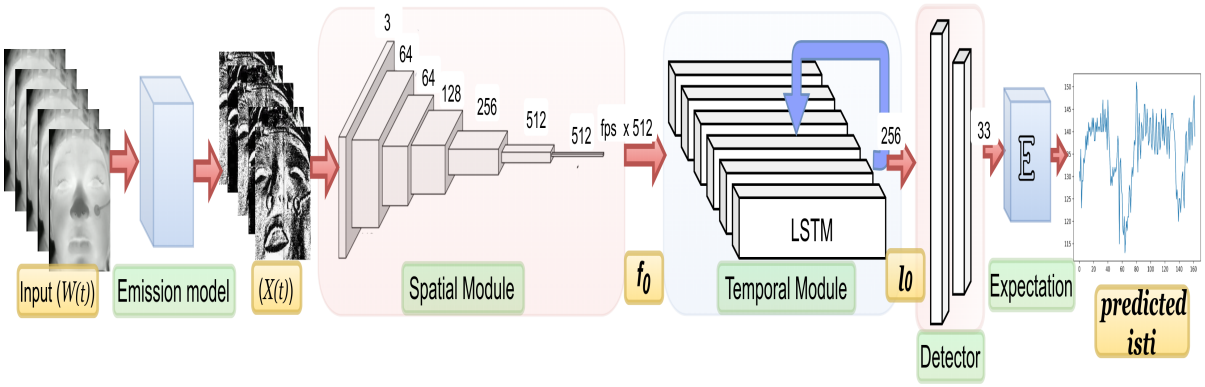

Figure 5.2: Model Architecture. Green boxes are the different modules of the model. Yellow boxes are the variables throughout the model. The Emission model processes the raw input data which is then fed into spatial and temporal modules. The Detector network predicts ISTI value for each of the frames from the output of these modules. This ISTI signal is used as input in our stress detection network.

blood pressure [153, 153–155].

Here we introduce StressNet, a non-contact based approach to estimating ISTI. To the best of our knowledge this approach is the first of it's kind. StressNet leverages the ISTI signal to classify whether a person is experiencing stress or not. To estimate the ISTI signal, a spatial-temporal deep neural network has been developed along with an emission representation model. Other physiological signals like heart rate (HR) or heart rate variability (HRV) cannot measure the changes in contractility, which are influenced by sympathetic, but not by parasympathetic activity, in humans [156].

Recently a number of studies have applied deep learning methods to the detection of HR or HRV from face videos [143, 144, 157, 158]. Most of these methods either fail to correctly identify the peak information in ECG or do not properly exploit the temporal relations in the face videos [144]. Recent work by [144] has developed a spatial-temporal deep network to measure rPPG signals such as heart rate variability (HRV) and average heart rate (AHR). Although these measurements are important, we show that in our experimental setup, the ground-truth ISTI signals allow for more accurate classification

of stress state than AHR or HRV.

In addition, thermal images mitigate some privacy concerns because the true likeness of the face is not being stored unlike RGB based models [159].

StressNet is an end-to-end spatial-temporal network that estimates ISTI signal and attempts to classify stress states based on thermal video recordings of the human face. An extensive analysis of the detailed dataset developed for this work has shown correlation between the estimated signal and ground truth. The effectiveness of this predicted ISTI signal is further validated by the model's ability to accurately classify an individual's stress state.

Technical Contributions:

- An emission representation module is proposed that can be applied to infrared videos to model variations in emitted radiation due to motion of blood and head movements.
- A spatial temporal deep neural network is developed to estimate ISTI.
- A simple classifier is then trained to estimate the stress level from the computed ISTI signal. To the best of our knowledge this is the first attempt to directly estimate ISTI and stress from thermal video.

5.2 Related Works

ISTI has been proposed as an effective, quantitative measure of psychological and physiological stress [160–164]. Measurement of ISTI requires both the ECG and $\partial Z/\partial t$ signals. Heart rate variability has also been used in several studies to index psychological and physiological stress [143, 144, 157, 158]. Different camera modalities have also been

used, namely, infrared, visible RGB, and five-channel multi-spectral [143, 144, 146, 165, 166]. A distinction is also made between research that takes place under laboratory and real-world settings. In the latter, environmental variables can complicate the detection and/or estimation task.

While no other works have included ISTI estimation or the ICG signal in their frameworks, the common video and ECG inputs lend themselves to similar network designs.

Several works for estimation of heart rate rely solely upon registration and classical signal processing techniques. For example, work from [167] registered a region of interest on the face, took the mean of the green channel, and passed that signal through a bandpass filter to estimate the heartbeat signal. At its time of publication in 2014, it achieved state-of-the-art performance on the MAHNOB-HCI dataset with a mean-squared error of 7.62 bpm [167].

Several studies have investigated heart rate variability estimation, using a variety of sensor types [145, 146, 168, 169].

The first end-to-end trainable neural network for rPPG was DeepPhys [143]. It replaced the classical face detection methods with a deep learning attention mechanism. Temporal frame differences are fed to the model, in addition to the current frame, to allow the network to learn motion compensation.

A more recent model built on DeepPhys is PhysNet [144]. This work incorporated a recurrent neural network (RNN), specifically long short term memory (LSTM) over the temporal domain. For tasks such as heart rate detection and pulse detection, modest gains were observed over DeepPhys. The addition of the LSTM also allowed the network to be trained on the task of atrial fibrillation detection.

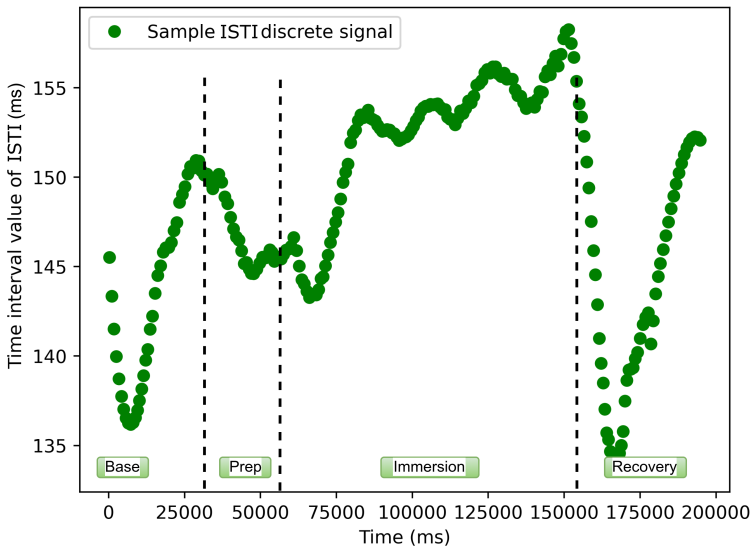

Figure 5.3: Discrete ISTI values are plotted against the peak positions of the ECG signal for a single participant. The "Base", "Prep", "Immersion" and "Recovery" labels refer to different phases of our stress induction protocol, whereby participants immerse their feet in either ice-water ("stress" condition) or lukewarm water ("no-stress" condition). The data shown were randomly selected from the "no-stress" condition. See section 5.4.1 for a detailed description of the protocol.

5.3 Method

Using raw thermal videos, our emission representation model generates the input for the spatial-temporal network. This network, along with the detection network predicts the ISTI signal from the raw input thermal videos. Our proposed model architecture is shown in Figure 5.2.

5.3.1 Generating ISTI signal

Electrocardiography (ECG) and Impedance cardiography's (ICG or Z) derivative ($\partial Z/\partial t$) act as the gold-standard physiological signals. ISTI is defined as the interval from the onset of left ventricular depolarization, reflected by the Q-wave onset in the

ECG, to the peak blood flow volume through aortic valve, reflected by the Z-point in the $\partial Z/\partial t$ signal. This time interval is computed from each peak of ECG and corresponding $\partial Z/\partial t$ peak. The discrete time interval value of ISTI is plotted at corresponding ECG peak positions as shown in Figure 5.3 and then interpolated with cubic interpolation to form a continuous signal. In Figure 5.3, the x-axis represents time (ms) while y-axis values represent the ISTI value (ms) for a particular ECG peak at that time of the video. The interpolated continuous ISTI signal is used as the ground truth for ISTI prediction.

5.3.2 Emission Representation Model

According to [142], RGB video based physiological signal measurement involves modeling the reflection of external light by skin tissue and blood vessels underneath. However, in the case of thermal videos, the radiation received by the camera involves direct emissions from skin tissue and blood vessels, absorption of radiation from surrounding objects, and absorption of radiation by atmosphere [170, 171]. Here, we build our learning model based on Shafer's dichromatic reflection model (DRM) [142] as it provides a basic idea to structure our problem of modeling emissions and absorption. We can define the radiation received by the camera at each pixel location (x, y) in the image as a function of time:

$$\boldsymbol{W}^{x,y}(t) = E^{x,y}_{ems}(t) + E^{x,y}_{abs}(t) + E^{x,y}_{atm}(t) \tag{5.1}$$

where $\boldsymbol{W}(t)$ is an energy vector (we drop the (x, y) pixel location index in the following for simplicity.) $E_{ems}(t)$ is the total emissions from skin tissue and blood vessels; $E_{abs}(t)$ is absorption of radiations by skin tissue and blood vessels; $E_{atm}(t)$ is the absorption of radiation by atmosphere. In current experimental setup the person is in a closed environment and 3ft from thermal camera, therefore the atmospheric absorption is negligible. According to [172], human skin behaves as a black-body radiator, therefore the reflections

are close to zero and emission is almost equivalent to absorption.

This implies that the only variation in energy comes from the head motion and from blood flow underneath skin. If we decompose the $E_{ems}(t)$ and $E_{abs}(t)$ into stationary and time-dependent components:

$$E_{ems}(t) = E_o \,.\, (\epsilon_s + \epsilon_b \,.\, f_1(m(t),\, p(t)) \tag{5.2}$$

where E_o is the energy emitted by a black body at constant temperature, it is modulated by two components: ϵ_s, is the emissivity of skin and ϵ_b, is the emissivity of blood. $f_1(m(t),\, p(t))$ represents the variations observed by thermal camera; [142, 143] $m(t)$ denotes all non-physiological variations like head rotations and facial expressions; $p(t)$ is the blood volume pulse (BVP). In a perfect black body, emissivity is equal to absorbtivity, therefore the absorbed energy is:

$$E_{abs}(t) = E_{ab}(t) \,.\, (\epsilon_s + \epsilon_b \,.\, p(t)) \tag{5.3}$$

where E_{ab} is the energy absorbed that changes with surrounding objects and their positions with respect to skin tissue.

$$E_{ab}(t) = E_o \,.\, (1 + f_2(m(t),\, p(t))) \tag{5.4}$$

where $f_2(m(t),\, p(t))$ represents the variation observed by the skin tissue. After substituting (5.4), (5.3), (5.2) in equation (5.1) and fusing constants; then neglecting the product of f_1 and f_2 as it is generally complex non-linear functions. Neglecting product of varying

terms, we get an approximate $\boldsymbol{W}(t)$ as :

$$\begin{aligned}\boldsymbol{W}(t) \approx K + E_o \,.\, \epsilon_b \,.\, (p(t) + f_1(m(t),\, p(t))) \\ + E_o \,.\, \epsilon_s \,.\, f_2(m(t)\,, p(t))\end{aligned} \tag{5.5}$$

where K is $2E_o \,.\, \epsilon_s$. We can get rid of this constant by taking first order derivative in the temporal domain.

$$\begin{aligned}\boldsymbol{W}'(t) = p'(t) \,.\, E_o \,.\, (\epsilon_b + \epsilon_b \,.\, \frac{\partial f_1}{\partial p} + \epsilon_s \,.\, \frac{\partial f_2}{\partial p}) \\ + m'(t) \,.\, E_o \,.\, (\epsilon_b \,.\, \frac{\partial f_1}{\partial m} + \epsilon_s \,.\, \frac{\partial f_2}{\partial m})\end{aligned} \tag{5.6}$$

This representation encompasses all the factors contributing to variations in radiation due to blood and face motion captured by the camera. Thus, we can suppress all possible non-necessary elements from data recorded by the camera. We use log non-linearity on each pixel to suppress any outlier in each image and separate the E_o, as its spatial distribution does not contribute to the physiological signal. The non-linearity looks as follows.

$$\boldsymbol{X}(t) = sign(\,\boldsymbol{W}'(t)) \,.\, \log(1 + \mod \boldsymbol{W}'(t)) \tag{5.7}$$

To remove high frequency components, we do a Gaussian filtering with $\sigma = 3$ in the spatial domains, and $\sigma = 4$ in the temporal domain. This filtered $\boldsymbol{X}(t)$ is the input to our deep learning model.

5.3.3 Deep Learning Model

Spatial-Temporal Network: Spatial-Temporal networks are highly successful in action detection and recognition tasks [49, 173, 174]. More recently, such networks have been used to process multispectral signals [9, 175, 176] The input to our spatial-temporal

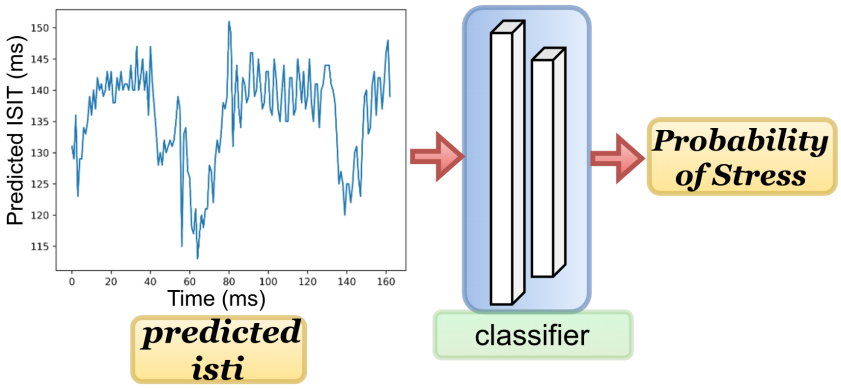

Figure 5.4: Stress detection network. Estimated ISTI signal is directly fed into the classifier network to predict the probability that the subject is under stress.

network is the stacked features from the emission representation model, which are then fed to a backbone network (e.g. resnet-50 [47]). This backbone network serves as a feature extractor. We mainly tested with object detection networks without the classification blocks as backbone networks.

Weights of these backbone networks are initialized with ImageNet pretrained values so that they can converge quickly on thermal videos. Global average pooling operation follows by the backbone block.

$$\mathbf{f}_0 = GAP(B\left(\boldsymbol{X}(t)\right) \tag{5.8}$$

where $B(\,.\,)$ stands for the backbone network, GAP is global average pooling operation, $\boldsymbol{X}(t)$ is from equation[5.7] (all processed frames stacked horizontally) and $\mathbf{f}_o$ is the output feature vector.

The backbone network is followed by long short term memory (LSTM) [177, 178] network, which captures the temporal contextual connection information from the extracted spatial features. LSTM [179] units include a 'memory cell' that capture long range temporal context. A set of gates is used to control the flow of information which

in turn helps the LSTM network learn temporal relations among the input features. The extracted feature vector from the backbone network is fed to the LSTM network.

$$\mathbf{l}_o = L_{STM}\left(\mathbf{f}_0\right) \tag{5.9}$$

where $\mathbf{l}_o$ is the feature output from LSTM network, $L_{STM}(\,.\,)$ stands for LSTM network and f_0 is the extracted feature vector from the backbone network.

Detection Network: Instead of directly predicting the continuous value of the ISTI signal from the output of LSTM network, we have divided the whole range [0,1] of ISTI values in n number of bins following [180]. To obtain the exact value of the ISTI signal ($\widehat{\mathbf{ISTI}}$) from each frame the expectation of the probability is taken for over all bins ($\widehat{\mathbf{isti}}_{bins}$),

$$\widehat{\mathbf{isti}}_{bins} = D\left(\mathbf{l}_o\right);\ \ \widehat{\mathbf{ISTI}} = E\left(\widehat{\mathbf{isti}}_{bins}\right) \tag{5.10}$$

where D stands for detection network which consists of fully connected layers, $\widehat{\mathbf{isti}}_{bins}$ is the probability of each bin, E is the Expected value, $\widehat{\mathbf{ISTI}}$ is the predicted ISTI value of each frame. This two stage approach makes our network more robust.

The predicted $\widehat{\mathbf{ISTI}}$ signal is fed to stress detection network which consists of fully connected layers, see Figure 5.4. The output of this network is probability of stress for the subject whose ISTI signal is estimated by our spatial-temporal network.

5.3.4 Multi Loss Approach

Previous works which predicted heart rate, breathing rate, or blood volume pulse mostly use mean squared error (MSE) loss. Another approach bins the regression output,

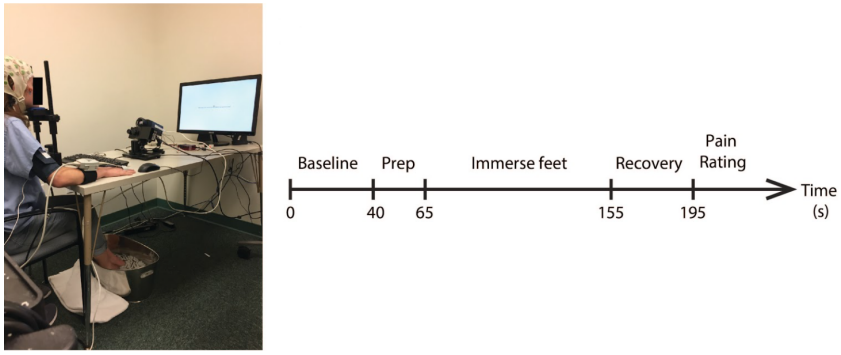

Figure 5.5: CPT/WPT Setup and Protocol. An example of a fully instrumented participant is shown. Participants followed instructions for the protocol presented on a computer monitor. After the baseline period the participant is instructed to position both feet on the edge of the bucket and prepare for immersion (prep). They then immerse the feet for 90s, then withdraw the feet and rest them on a towel for a 40 s recovery period.

and modifies the network output layer to be a multi-class classification. This method provides more stability to outliers than MSE, but its accuracy is limited by the number of bins. So for the ISTI signal prediction model, we use the multi loss approach used by [180]. This type of loss is a combination of two components: a binned ISTI classification and an ISTI regression loss.

$$\begin{aligned} \mathbf{L}\left(\Theta\right) = BCE\left(\widehat{\mathbf{isti}}_{bins},\ \mathbf{isti}_{bins}\right) \ + \\ \alpha\ .\ MSE(\widehat{\mathbf{ISTI}},\ \mathbf{ISTI}) \end{aligned} \tag{5.11}$$

For the stress detection network only binary cross entropy (BCE) is used as loss function.

5.4 Experiments

5.4.1 Dataset

42 healthy adults (22 males, mean age 20.35 years) were recruited as part of the Biomarkers of Stress States (BOSS) study run at UC Santa Barbara, designed to investigate how different types of stress impact human brain, physiology and behavior. Participants were considered ineligible if any of the following criteria applied: heart condition or joint issues, recent surgeries that would inhibit movement, BMI > 30, currently taking blood pressure medication or any psychostimulants or antidepressants. Informed consent was provided at the beginning of each session, and all procedures were approved by Western IRB and The U.S. Army Human Research Protection Office, and conformed to UC Santa Barbara Human Subjects Committee policies.

Participants attended the lab for five sessions on five separate days as part of the BOSS protocol. For collection of impedance cardiography (ICG), 8 electrodes were placed on the torso and neck, two on each side of the neck and two on each side of the torso. For electrocardiogram (ECG), 2 electrodes were placed on the chest, one under the right collarbone. For videos, thermal camera (Model A655sc, Flir Systems, Wilsonville, OR, USA),was positioned ∼65 cm from the participant's face and set to record at 640 $\times$ 240 pixels and 15 Hz frame rate. A large metal bucket was then positioned in front of the participant's feet. In the Cold Pressor Test (CPT) session, the bucket was filled with ice water ($\sim$ 0.5 $^\circ$ C), whereas the in the control session (Warm Pressor Test; WPT), the bucket was filled with lukewarm water ($\sim$ 34 $^\circ$ C). In each session, participants were required to immerse their feet in the water five times for 90 s, following the test protocol outlined in Figure 5. The CPT is popular method for inducing acute stress in humans in the laboratory. It causes pain and a multiplex of physiological responses e.g. elevated heart rate and blood pressure and increased circulating levels of epinephrine and nore-

pinephrine [181, 182]. The WPT was devised as an "active" control task, designed such that participants engaged in exactly the same protocol as with the CPT, but without the discomfort of cold-water immersion. This ensured that any psychological or physiological effects induced by engaging in the protocol and immersing the feet in water, were controlled for. Each of the five CPT/WPT immersions were separated by $\sim$ 25 minutes. Between immersions, participants completed tests designed to measure performance across a range of cognitive domains (these data are not reported in this paper). Session order was counterbalanced between participants.

Nine participants' data were excluded due to technical failures (the thermal imaging camera failed to record one or more sessions). Thirty-three participants' data were used for modeling. This sample is similar in size to existing public data sets of a similar nature [183, 184].

5.4.2 Evaluation Metrics

Performance metrics for evaluating ISTI prediction are Mean Squared Error (MSE) and Pearson's correlation coefficient (R). For stress detection, average precision (AP) is used as the validation metric.

Mean Squared Error is a model evaluation metric used for regression tasks. The main reason for using MSE as evaluation metric is that the precise value of predicted ISTI signal is important.

Pearson Correlation coefficients are used in statistics to measure how strong a relationship is between two signals. It is defined as covariance of the two signals divided by the product of their standard deviations. Pearson correlation is also used here as an extra validator on the predicted ISTI signal, signifying that the shape of predicted curve

also corresponds well with the ground-truth.

$$\rho_{X,Y} = \frac{cov(\ X,\ Y)}{\sigma_X\ .\ \sigma_Y}$$

Average Precision (AP) is the most commonly used evaluation metric for object detection tasks [185]. It estimates the area under the curve of precision and recall plot. Precision measures how many predictions are correct. Recall calculates the correctly predicted portion of the ground truth values.

5.4.3 Implementation Details

In experiments, the effectiveness of the spatio-temporal network is evaluated. The dataset is split as follows: 80% training, 10% validation and 10% testing set. The input video frames are cropped to 360 $\times$ 240 to remove the lateral blank areas before being fed to our emission representation model.

For backbone model, experiments with different architectures of resnet were performed, those are resnet18, resnet34, resnet50, resnet101. In the final model resnet50 is used as feature extractor. The output of resnet50 is average-pooled instead of max-pooling operation. The reason for that is removing a less important feature from important feature (max-pool operation) can reduce the signal-to-noise ratio in physiological measurement, so average pooling is used to keep even the less important feature vector information. Before feeding to temporal network, the average-pooled feature vector is reshaped so that each input sequence to LSTM consists of 1 second of time information. The reasoning was that since the peaks of ECG and $\partial Z/\partial t$ signal occurs almost once per second, the LSTM network will better captures the relation between adjacent peaks. For the temporal network, we experimented with a number of LSTM layers (2-8), 6 LSTM layers are best suited for capturing the temporal contextual information. Hidden

Name of the Method	PC Coefficient	MSE
Baseline	0.170	103.829
DeepPhys [143]	0.575	47.530
I3D [186] + Detection Network	0.84	5.227
StressNet	**0.843**	**5.845**

Table 5.1: StressNet's performance in predicting ISTI signal. The performance is measured on Pearson-Correlation Coefficient(PC Coefficient) and mean square error. Our model clearly outperforms the existing methods by a good margin.

unit size is kept at half the feature vector length from the spatial network (resnet50), $hidden\_unit\_size$ is $256 \times frame\_rate$, ensuring that the number of memory cells is sufficient enough to transfer information from previous LSTM cell to next. The number of fully connected layers following LSTM is two, with ReLU added as non-linearity. The output of the final fully connected layer is 33 bins output. 33 bins is an empirical value.

The emission representation model works online in the pipeline and is loaded on the same machine on which deep learning model is trained. Each video is approximately of size (frames $\times$ H $\times$ W) 2500 $\times$ 640 $\times$ 240 with 16bit depth information per pixel. Due to memory constraints on the GPU, $batch\_size$ is kept at 500 frames. The learning rate for resnet50 is started at 0.001, for LSTM and FC layers at 0.01, which reduces after every 10 epochs by a factor of 0.1. Stochastic Gradient Descent is used as optimizer for the network.

5.5 Results

The proposed method is evaluated in two main criteria. First we evaluated the quality of our predicted ISTI signal, then we tested the effectiveness of the predicted ISTI signal in detecting stress.

Predicting ISTI Signal: For the first part, as mentioned in the evaluation metrics section, the model performance is evaluated on Mean Squared Error (MSE) and Pearson

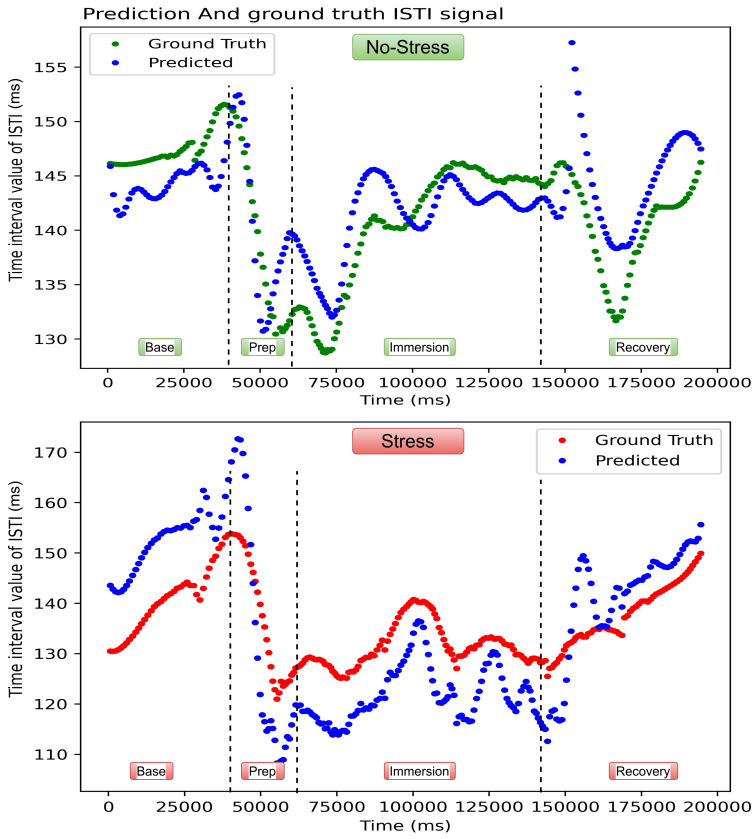

Figure 5.6: Quality of our predicted ISTI signal in stress and no-stress conditions. Data shown are examples from a single participant's data (selected at random). The "Base", "Prep", "Immersion" and "Recovery" labels refer to the different phases of the CPT/WPT procedure.

Correlation coefficient (PC Coefficient). In Table 5.1 our model's performance can be seen compared to the other methods. Our model outperforms the other methods in both of the evaluation metrics with a good margin. As shown in Figure 5.6, our model agrees well with the ground truth signal in both stress and no-stress cases.

Since no work has been done on detecting the ISTI signal before, to validate our model we have implemented DeepPhys [143]. As can be seen in Table 5.1 our implementation of DeepPhys model [143] did not perform well in detecting the ISTI signal. This poor performance mostly stems from two main reasons. First, DeepPhys model is designed

to predict periodic physiological signals and since ISTI is non-periodic in nature, loss in DeepPhys does not suit this particular task. Second, the skin reflection model in [143] does not expand properly for modeling the infrared radiation. For baseline methods, ECG signal is extracted from the face using simple statistical filtering methods. According to [187–189] temperature changes in the tip of the nose and forehead can index different stress states, so for our baseline approach we tracked these regions and then band-pass filtered to extract the signal. This signal is quite noisy which contributes to our baseline's poor performance.

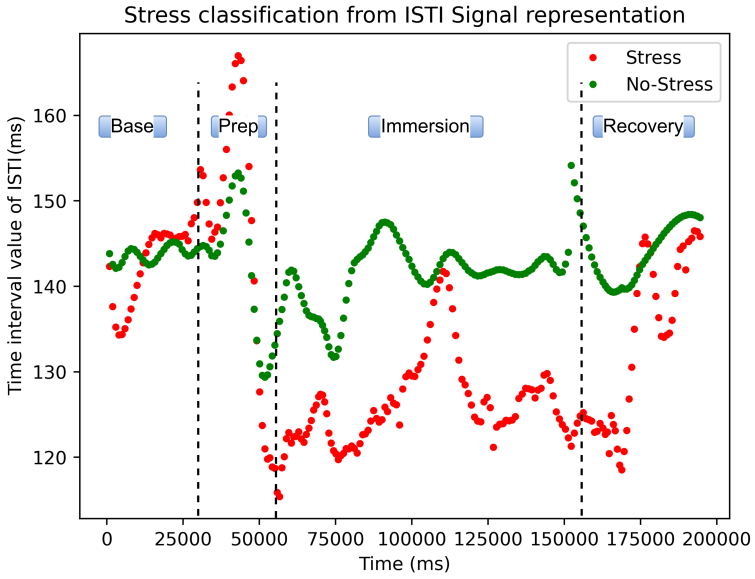

Figure 5.7: Importance of ISTI signal in detecting stress. Ground truth ISTI data from a single participant (randomly selected) are shown. Clearly, ISTI signal in the stress condition is different from the ISTI signal in no-stress condition. The "Base", "Prep", "Immersion" and "Recovery" labels refer to the different phases of the CPT/WPT procedure.

Detecting Stress: For the second part of stress detection, we evaluate whether the ISTI signal provides a robust index for stress detection.

An example of the ISTI response to CPT/WPT in a single participant is shown in

Input Signal	AP (Average Precision)
Heart Rate (HR)	0.753
Heart Rate Variability(HRV)	0.814
ISTI (Ground Truth Signal)	0.902
ISTI (StressNet Predicted)	**0.842**

Table 5.2: StressNet can classify stress state with greater AP using contact-less ISTI estimates when compared to other commonly used contact-less signal estimates (HR and HRV).

Figure 5.7. Here, we observe a clear distinction in ISTI in anticipation of cold- vs. warm-water immersion (i.e. during the prep period) as well as during immersion and recovery.

To evaluate the predictive validity of the ISTI signal, we compare it to heart rate (HR) and heart rate variability (HRV) by entering these alternative signals into our model. We compare ISTI with HR and HRV because these measures are considered to be reliable indices of stress [190] and have been used in many stress classification studies [146, 191]. Here, we compute them from the ground truth ECG signal. HR is computed by counting number of beats in a sliding window approach with window size 15 (seconds) and stride 1 (seconds). For HRV, time between R peaks is recorded over a defined time interval (15 seconds) and then HRV is computed according to the Root Mean Square of Successive Differences (RMSSD) method [192].

In Table 5.2 we can see how our predicted ISTI signal is better in detecting stress state than HR (12% higher AP) and HRV (4 % higher AP). Also, higher AP with the ground truth ISTI signal confirms that ISTI is the most reliable index of stress state in the context of our dataset.

5.5.1 Ablation Study

Analysis of Emission Representation Model: The overall architecture of StressNet consists of three main models: the emission representation model, the spatial-

Name of the Backbone	PC Coefficient	MSE
vgg19 [193]	0.605	33.164
resnet18 [47]	0.749	15.095
resnet34 [47]	0.815	6.223
resnet50 [47]	**0.843**	**5.845**
resnet101 [47]	0.779	14.373

Table 5.3: Comparison of different backbones' performance. In the task of estimating ISTI signal resnet50 is better than all other backbones.

temporal model and the detection model. To evaluate how each model affects the overall performance, we evaluated the spatial-temporal model with and without the emission representation model. The fully pre-trained network was tested without the emission representation model and we observed a 1.119 increase in the mean squared error in predicting the ISTI signal. The best results for ISTI signal prediction as mentioned in table 5.1 are obtained using all three models mentioned above.

Analysis with Backbone CNNs: The spatial-temporal model is evaluated with all the ResNet models [47]. We also tested with VGG19 [193] as our backbone. The performance comparison is shown in table 5.3.

Analysis with Breathing signal: The breathing signal is captured by tracking the area under the nostrils for changes in temperature. The computed time series signal is passed through band-pass filter with low and high cutoff frequencies of 0.1 Hz and 0.85 Hz, respectively. This breathing signal is also used as an input to our stress detection model and the predictions from this model are multiplied with the predicted ISTI signal input. This process boosts the stress detection results by 0.1774 AP. This shows how ISTI can be combined with other physiological signals to detect stress.

Limitations of the Model: Despite being instructed to stay still, participants occasionally made large head movements and/or obscured their face with a hand (see Figure 5.8). There were also occasions where the ECG/ICG signal was noisy due to

movement or bad electrode connections. In these instances the model fails to detect ISTI.

Different Spatial Temporal Network: To validate the effectiveness of spatial temporal networks in detecting ISTI signal, we implemented I3D [186] architecture, a 3D convolution based spatial-temporal network proposed for action recognition. We replaced the classification branch in I3D with our detection network. The performance is similar to StressNet's performance.

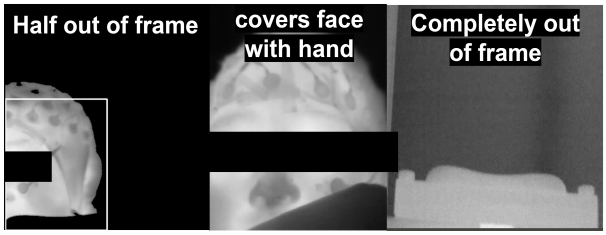

Figure 5.8: Example StressNet failure cases. Network performance is impaired when the face is outside the video frame or obscured.

5.6 Conclusion

Here we present a novel method for the estimation of ISTI from thermal video and provide evidence suggesting ISTI is a better index for stress classification than HRV or HR. Overall, our method is more accurate than existing methods when performing binary stress classification on thermal video data.

Our model achieved state-of-the-art performance, and performance could potentially be boosted even further by using different spatial-temporal models. The most successful backbone model used only spatial data from each frame independently, compared to the I3D network [186] that employed simultaneous processing of both spatial and temporal information. However, to test this we require larger dataset, that would allow for improved pre-trained initialization of the spatial-temporal backbones and better transfer

learning performance.

This work has several limitations. First, it is unclear whether StressNet's performance can generalize to the classification of different forms of stress e.g. social stress, physical and mental fatigue. Second, it is possible that exposure to lukewarm-water in the control condition may have induced eustress (beneficial stress), meaning that StressNet is actually classifying distress vs. eustress, not distress vs. neutral states, and this may impact performance. Third, the data used to test StressNet were collected under controlled laboratory conditions, so it is unclear how performance may be impacted in real world use case scenarios that may by subject to increased atmospheric noise and movement artifacts. Further testing with a diverse range of datasets collected under different stress conditions and scenarios is required to determine the efficacy and generalizability of StressNet in the real world.

Chapter 6

Discussions and Future Work

6.1 Conclusions

This thesis has focused on addressing the challenges inherent in building and training large vision models for scientific applications, where data is often complex, scarce, and requires domain-specific knowledge for effective interpretation. The work presented here has explored the integration of advanced signal processing techniques with domain-specific knowledge to enhance the performance of vision models across various problem contexts, including remote sensing, biomedical applications, and ecological monitoring. By developing novel methods for detection, segmentation, and analysis, this research has significantly contributed to the advancement of large vision models, making them more applicable and efficient in specialized scientific domains.

The primary contributions of this thesis include the development of advanced methodologies for methane detection, as discussed in Chapters 2 and 3. In these chapters, we introduced MethaneMapper, a powerful tool for detecting methane emissions using both hyperspectral and multispectral data. By addressing the challenges of training large vision models with limited data, MethaneMapper integrates domain-specific knowledge,

such as the chemical properties of methane, into the model, enabling accurate detection across multiple spatial scales. This work not only improves the efficiency and accuracy of methane detection but also sets a foundation for applying similar approaches to other environmental monitoring tasks.

In Chapters 4 and 5, we extended the principles developed in MethaneMapper to solve problems in ecology and biomedical field. In Chapter 4, we introduced WildlifeMapper, a model designed to detect and identify animals in large open grasslands using aerial imagery. This work significantly enhances wildlife monitoring by addressing the challenges of sparse and noisy data, ensuring accurate detection even in complex environmental conditions. In Chapter 5, we presented StressNet, a model for estimating physiological signals from multispectral facial imagery to detect stress. This non-invasive approach has proven effective in reconstructing signals like ECG and ICG, offering a valuable tool for healthcare and personalized medicine. These applications demonstrate the versatility and impact of the developed models, showcasing their potential across diverse scientific domains.

6.2 Future Work

The pressing global challenges of our time, such as climate change, biodiversity loss, and public health crises, demand innovative, interdisciplinary solutions. The convergence of different scientific domains can provide new insights and methodologies that are essential to tackling these complex issues. An interesting pathway in future could be, how can we detect traces of gas from aerial imagery or space borne imagery. One way to do that could be integrating plant sciences with advanced computer vision techniques to develop innovative new methods for global pollution assessment and mitigation.

6.2.1 Expanding the Scope of Pollution Monitoring

Pollution monitoring is a critical component in addressing climate change and environmental degradation. Traditional methods typically rely on high-resolution data from satellites, aerial imagery, or costly ground-based sensors, each designed to detect specific types of pollutants [9, 10, 65, 72]. However, these methods come with significant limitations, including the high cost of sensor deployment [13, 44] and the need for different sensors to detect different pollutants. Furthermore, these methods often fail to capture the intricate interactions between pollutants and the natural environment. To overcome these challenges, we can use plants as natural bioindicators of pollution and model the spectral changes in plants due to certain type of pollutants.

Plants are ubiquitous, and exhibit specific, measurable responses to various pollutants which can be measured cost effectively. These responses include changes in growth patterns, leaf morphology, pigmentation, and overall health, making plants effective and sensitive indicators of environmental pollution. By leveraging plants as the "eyes and ears" of pollution monitoring, we can develop a more comprehensive, scalable, and cost-effective system for detecting environmental stressors on a global scale.

6.2.2 Research Objectives and Methodology

The future research can focus on two primary objectives:

1. **Identification of Bioindicator Species:** A good first step would be to identify plant species that exhibit distinct physiological and morphological responses to specific pollutants. We can begin by studying species known for their sensitivity to air pollution, such as mosses, lichens, and ferns. Mosses and lichens are particularly sensitive to sulfur dioxide (SO_2) and heavy metals, while ferns like the sensitive fern (Onoclea sensibilis) and the bracken fern (Pteridium aquilinum) are known to

Figure 6.1: *Depiction of visible morphological and pigment change in plants due to pollution. It can be seen that the green color of the leaves turning white and brown shows pigment change in the plant due to presence of certain pollutants around, in this specific case, this is due to presence of SO_2 in the atmosphere as shown in Red boxes.*

respond to ozone (O_3) and nitrogen dioxide (NO_2). By establishing a correlation between these species' spectral signatures and the concentration of pollutants, we can create a detailed map of environmental stress across different regions. This phase will involve extensive fieldwork, laboratory analysis, and collaboration with

botanists and environmental scientists to accurately document the responses of these bioindicator species to various pollutants [194, 195]. Understanding these relationships will be critical to developing a robust model that can be generalized across different ecosystems and pollutant types.

2. **Development of Advanced Multimodal AI models:** Once the bioindicator species are identified, the next step will be to develop deep learning models capable of detecting subtle changes in plant health from hyperspectral and multispectral imagery. Hyperspectral imagery captures a wide range of spectral bands [10, 62, 196], allowing for the detection of minute variations in reflectance that are indicative of stress. The initial approach will involve using deterministic methods, such as matched filtering and band ratio techniques to analyze the spectral signatures of healthy plants versus those exposed to pollutants.

   Building on this foundational analysis, I will employ transformer networks [87], known for their ability to model long-range dependencies and relationships in data to detect and interpret these spectral variations. These networks will be trained on high-resolution data collected from ground-based sensors and drones, with the goal of identifying pollution-induced stress at a fine-grained level. The training process will leverage semi-supervised learning and contrastive learning frameworks to enhance the model's ability to distinguish between healthy and stressed plants under varying environmental conditions.

   To capture pollution-induced stress effectively, my research will focus on several key indicators, including changes in leaf morphology and pigmentation. For example, pollution can cause leaves to become smaller, thicker, or develop abnormal shapes, and it can also lead to changes in their chlorophyll content. These morphological and pigmentary changes are strong indicators of pollution exposure, particularly

for pollutants like nitrogen dioxide, particulate matter, sulfur dioxide, and ozone. Monitoring these changes through multispectral sensors operating in the visible and near-infrared wavelengths (400 nm - 1100 nm) will provide a reliable method for early detection of environmental stress.

6.2.3 Scaling to Global Applications

Once the deep learning models are validated on high-resolution, ground-based data, the research will focus on scaling these methods to work with satellite imagery. This scaling process will involve fine-tuning the models to adapt to the lower resolution of satellite data while maintaining accuracy in detecting pollution-induced stress. One of the key challenges in this phase will be the adaptation of models trained on detailed, high-resolution images to the broader, less detailed context of satellite imagery. This will require innovative techniques, such as knowledge distillation, where insights gained from high-resolution models are transferred to models designed for lower-resolution data.

The use of satellite imagery will enable large-scale, continuous monitoring of bioindicator species across diverse geographical areas, providing a global perspective on pollution levels and their impacts on ecosystems. By leveraging my expertise in computer vision, particularly in the detection of small objects in high-dimensional data, I aim to optimize these models for accuracy and efficiency, ensuring that they can be effectively applied in real-world scenarios.

6.3 Long-term Impact and Vision

The long-term impact of this research lies in its potential to revolutionize how we monitor and respond to environmental pollution. By utilizing bioindicator species in conjunction with advanced imaging and deep learning techniques, this approach offers a

scalable, cost-effective solution that addresses many of the limitations of current pollution monitoring methods. The findings from this research will provide valuable data for policymakers, environmental scientists, and conservationists, aiding in the development of informed strategies to mitigate pollution and protect both human health and the environment.

Moreover, this research aligns with the broader goals of sustainable development, contributing to the creation of more resilient ecosystems and communities. The integration of plant sciences with computer vision not only advances our understanding of how plants respond to environmental stress but also opens new avenues for interdisciplinary research and innovation.

Appendix A

MethaneMapper Appendix

A.1 Introduction

In this section, we provide will all the details about the data collection and annotations creation process. We also provide with the complete derivation of Spectral Linear Filter (**SLF**) along with a pseudo implementation of **SLF** algorithm. Next in the document we provide some more qualitative examples of success and failure cases of MethaneMapper. Towards the end of the document we provide graph plots about training convergence of all the ablation experiments with Spectral Feature Generator (SFG) and Query Refiner (QR) module.

A.1.1 Dataset

AVIRIG-NG

AVIRIS-NG [13] is an acronym for the *Airborne Visible InfraRed Imaging Spectrometer - Next Generation* developed by Jet Propulsion Laboratory (JPL) in 2009. JPL conducted thousands of flight lines recording data with AVIRIS-NG instrument in last 7 years. On the AVIRIS-NG instrument an array of total 598 sensors in push-broom order captures an unortho-rectified data-cube of spatial dimension $\sim 23k \times 598$, where each sensor records a spectral wavelengths ranging from $380nm - 2510nm$ [197] making a dimension of 432 channels. It has 34^o field of view with a 1 mrad instantaneous field of view the generates spatial resolution of $1 - 8m$ based on altitude. This data is then rectified using a geometric lookup table and the resulting data cube is of size $\sim 23k \times \sim 1.5k \times 432$. The data is provided in Band Interleaved by Line (BIL) ordering. BIL ordering signifies the 3D matrix is indexed first by image row, then by channel, and then by the image column [2]. One can find details about the naming convention and the type of data each files contain in "README.txt" file in each flightline folder. The data can be loaded into a *numpy* array easily using python libraries. All data is orthorectified.

Annotations

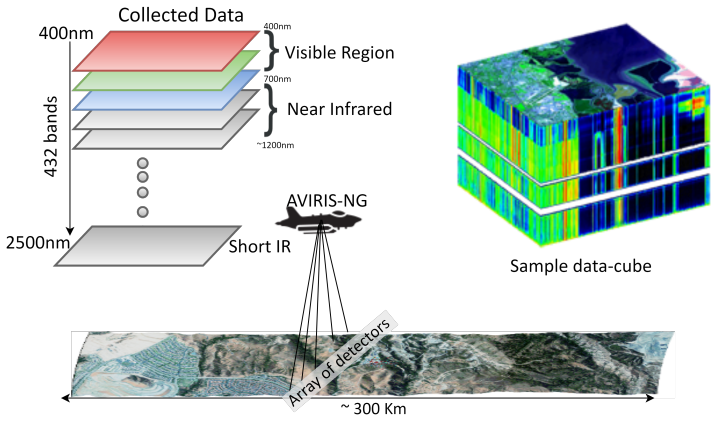

Figure A.1: *Depiction of data collection process. Each flightline is* $\sim$ 300 *kms long. An array of 598 sensors records data at 1.5m/pixel spatial resolution. All flightlines are ortho-corrected. Each data-cube is of dimension* $\sim 23k \times \sim 1.5k \times 432$.

Transformation and Ortho-correction. First step is to read the annotation GeoTiff patch of size 150×150 of a methane concentration mask and convert its Coordinate Reference System (CRT) to AVIRIS-NG flightlines' CRT (EPSG 4326). Next, we use the corresponding AVIRIS-NG flightlines' geometric lookup table and unortho-corrected geographic pixel location to generate ortho-corrected geographic pixel location data of the flightline. Next, we find the flightline's geographic indices that are closest to the geographic indexes of the methane concentration mask (annotation GeoTiff). Finally, we use these corresponding pixels to compute a homography transform matrix that maps the methane concentration mask (annotation GeoTiff) to the AVIRIS-NG flightline's spatial dimensions. We repeat this process for each plume in the flightline in order to generate the CH_4 concentration map for the entire flightline.

Resolution matching. To match the resolution of transformed annotation GeoTiff patch to AVIRIS-NG flightline, we use nearest-neighbor resampling. A pixel from the transformed annotation GeoTiff patch may be repeated multiple times in the CH_4 concentration map for the entire flightline.

Annotation Style. The *Point Source* and *Diffused Source* are coded following the same standard as JPL-CH4-detection-V1.0 [2] dataset. The 3-channels have values in [0-255] range.

- Red (255,0,0): plume, believed to be associated with a *Point Source*
- Blue (0,0,255): plume, believed to be associated with a *Diffuse Source*
- Black (0,0,0): no plume (or unlabeled)

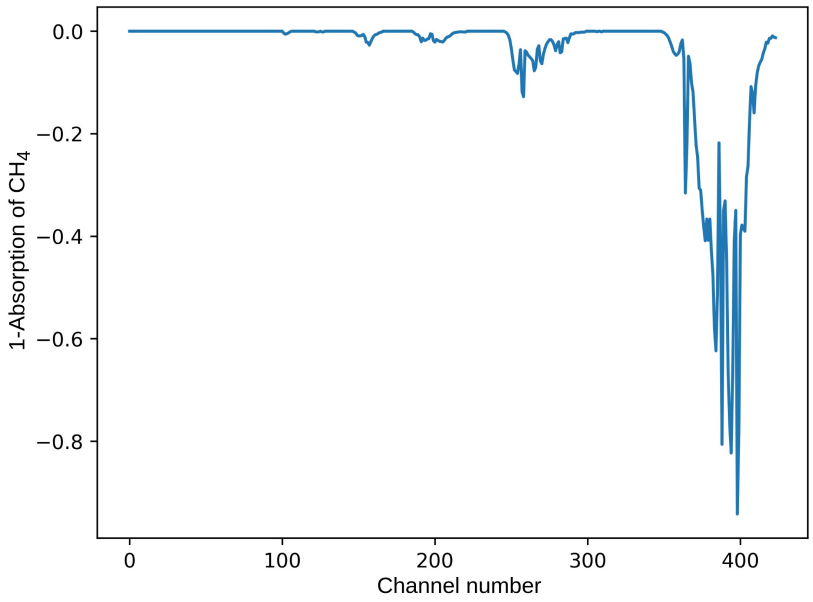

Figure A.2: *Spectral absorption pattern of CH_4 gas. The x-axis show the channel number ranging from 0-400 corresponding to wavelength range ($400nm - 2500nm$). It is obtained from the public repository HITRAN [1].*

We kept our annotation style consistent with JPL-CH4-detection-V1.0 benchmark dataset [2] so that both JPL-CH4-detection-V1.0 and MHS datasets can be merged seamlessly.

A.1.2 Spectral Linear Filter(SFL)

Traditional Matched Filter

Passive hyperspectral imaging sensors captures spectral radiances values from N_0 ($N_0 = 432$) spectral channels corresponding to wavelengths ranging from $400nm - 2500nm$ as shown in Fig. A.1 with sample data-cube. The complete hyperspectral image is represented as $\mathbf{x} \in \mathbb{R}^{H_0 \times W_0 \times N_0}$ where $H_0, W_0 \& N_0$ are height, width and number of channels respectively. In this hyperspectral data, we are looking for a very weak signature of interest hidden in background. In this case the signature of interest is CH_4 and the background is ground terrain. CH_4 shows strong absorption patterns around $2100nm - 2500nm$ wavelength.

The most common linear approach for finding CH_4 candidates is taking a N_0-dimension (same as number of spectral channels) vector α, and apply as a dot product to each pixel (N_0-dimension) in the hyperspectral image to generate a scalar output per pixel. This operation is supposed to reduce or remove the ground terrain, sensor noise and amplifies CH_4 signature. The α vector used here is called as "matched filter". Therefore computing

right α is very critical for generating better candidates of CH_4 emission. It is dependent on absorption pattern of CH_4 and on the distribution of the ground terrain. To model α, let $\mathbf{r}_i \in \mathbb{R}^n$ be a i^{th} pixel from the hyperspectral image representing the ground terrain pixel and sensor noise, and $\mathbf{t}$ be the CH_4 absorption pattern [1]. This is modeled as the additive perturbation as shown below:

$$\mathbf{x}_i = \mathbf{r}_i + \mathbf{t}, \tag{A.1}$$

where $\mathbf{x}_i$ is the spectrum when CH_4 is present. The CH_4 absorption pattern $\mathbf{t}$ represents the change in radiance units of the background caused by adding a unit mixing ratio length of CH_4 absorption [9, 38]. Figure A.2 shows the spectral absorption pattern of CH_4 per channel. In the ideal scenario where only CH_4 gas is present in signal (i.e. all white background), the matched filter output is $\alpha^T\mathbf{t}$. In case there is no gas and just ground terrain and sensor noise, the matched filter output is $\alpha^T\mathbf{r}_i$. The variance (Var) of $\alpha^T\mathbf{r}_i$ for latter is represented by:

$$Var(\alpha^T\mathbf{r}_i) = \langle(\alpha^T\mathbf{r}_i - \alpha^T\boldsymbol{\mu})^2\rangle = \alpha^T\mathbf{Cov}\alpha, \tag{A.2}$$

where $\mathbf{Cov}$ and μ are covariance and mean respectively computed for $\mathbf{r}_i$. Inspired from [9, 38] we define the Methane-to-Ground terrain Ratio (MGR) is:

$$\text{MGR} = \frac{|\alpha^T\mathbf{t}|^2}{\alpha^T\mathbf{Cov}\alpha}, \tag{A.3}$$

We can see that the magnitude of α does not affect MGR. According to [9, 38, 198], the MGR can be maximized subject to constraints(zero mean and $\alpha^T\mathbf{K}\alpha$ constraint to 1). The matched filter α is then represented by:

$$\alpha = \frac{\mathbf{Cov}^{-1}\mathbf{t}}{\sqrt{\mathbf{t}^T\mathbf{Cov}^{-1}\mathbf{t}}}. \tag{A.4}$$

In ideal instances when there is no background (i.e. all white background) and just CH_4 gas present. The matched filter in equation A.5 is directly proportional to $\mathbf{t}$. This is just the target signature ($\mathbf{t}$) itself scaled so that the filtered output has variance of one. The methane enhancement per pixel can be computed as follows:

$$\hat{\alpha}(\mathbf{x}_i) = \frac{(\mathbf{x}_i - \mu)^T\mathbf{Cov}^{-1}\mathbf{t}}{\sqrt{\mathbf{t}^T\mathbf{Cov}^{-1}\mathbf{t}}}, \tag{A.5}$$

where $\hat{\alpha}(\mathbf{x}_i)$ is the per pixel estimation of methane, on other words, column enhancement of methane. The covariance matrix ($\mathbf{Cov}$) used is not known as *prior* and is estimated from data. It is computed as outer product of the mean subtracted radiance over all the pixels. In other words, the traditional matched filter from equation A.5 computes the covariance ($\mathbf{Cov}$) of ground terrain with an underlying assumption that in all elements have similar absorption pattern. Same covariance matrix ($\mathbf{Cov}$) matrix is used to whitens

the varying ground terrain and amplify the CH_4 present. But in realistic scenarios, the ground terrain is varying, the type of terrain changes frequently, there is water bodies, bare soil, vegetation, dense vegetation, building structures in cities, roads etc in a single image. For example, water have a strong absorption of solar radiations, therefore the methane on such backgrounds have a very weak visibility. Similarly, wet fields dense vegetation have similar behaviour. On the other hand, bare soil, rocks, etc have lower absorption, the methane present on such background have strong visibility. A simple and single approximation of the covariance (**Cov**) of ground distribution can not provide the right and effective estimate of methane enhancement. To tackle this limitation, we developed an spectral linear Filter (SLF) that does land cover classification and segmentation and reduces the noise as discussed in the next sections.

Landcover Classification and Segmentation

In this section, we improve upon the limitations mentioned in the previous section. We start with taking hyperspectral bands from visible spectrum ($400nm - 700nm$) and near-mid infrared region ($800nm - 1350nm$). We recreated the RGB representation of the ground terrain by a weighted normal distribution for each color band. Same is done for near infrared region. Next we take a simple, very effective and efficient approach for doing landcover classification and segmentation. We compute the Normalized Difference Vegetation Index (NDVI) [53,199] and Normalized Difference Water Index (NDWI) [54]. NDVI quantifies vegetation by measuring the difference between near-infrared (which vegetation strongly reflects) and red light (which vegetation absorbs) [53]. It ranges from -1 to $+1$. It is a very effective index and has been used in literature for more than 4 decades. [54] created NDWI and used it to highlight open water features in a satellite image, allowing a water body to "stand out" against the soil and vegetation. It is calculated using the GREEN-NIR (visible green and near-infrared) and ranges from -1 to $+1$. Its primary use today is to detect and monitor slight changes in water content of the water bodies.

$$ndvi = \frac{NIR - R}{NIR + R}; \;\; ndwi = \frac{NIR - MIR}{NIR + MIR} \tag{A.6}$$

where NIR is near infrared region normalized around $880nm$, MIR is mid infrared normalized around $1240nm$ and R is red, normalized around $660nm$. We take advantage of these indexes and create segmentation maps for different types of vegetation, water bodies, bare soil, rocks, mountains, city/urban areas, roads etc. We take the classification thresholds from $[52, 200]$. For simplification, we also tested by splitting the scale -1 to $+1$ in 20 classes, each with a range of $< 0.1 >$. We obtained comparable results as compared to using classification ranges from $[52, 200]$. This simple, effective and efficient approach gives three fold boost to our spectral linear filter CH_4 candidates estimation.

Covariance (cov) per class

We take the segmented image from previous step, we will call segmented image as segmentation mask for simplicity now onward. In practice we have 20 classes, each with a segmentation mask. We merged two or more adjacent classes into one if the number of pixels in that class is less 10000 . The Number of pixels in each class is kept higher to ensure that while computing the covariance (**Cov**) matrix, the methane signal does not have any or have negligible effect. It is okay to merge adjacent classes into one because they have almost similar radiance/reflectance, for example, light vegetation and normal vegetation have similar reflectance, etc. For each class we compute a separate mean and covariance matrix. The covariance $\mathbf{Cov}_k$ of k^{th} class is computed as:

$$\mathbf{Cov}_k = \frac{1}{N}\sum_{i=1}^{i=j}(\mathbf{x}_i - \mu_k)(\mathbf{x}_i - \mu_k)^T \ \forall \ j \in k, \tag{A.7}$$

where N is the number of pixels (> 10000) in k^{th} class and μ_k is the mean of k^{th} class. For each class we compute the mean μ_k, covariance matrix $\mathbf{Cov}_k$. While iterating through each pixel of hyperspectral image, we check to which class k the pixel $\mathbf{x}_i$ belongs to and use those pre-computed values. The final Spectral Linear Fitler (**SLF**) is shown as below:

$$\mathbf{SLF}(\mathbf{x}_i) = \frac{(\mathbf{x}_i - \mu_k)^T \mathbf{Cov}_k^{-1} t}{\sqrt{t^T \mathbf{Cov}_k^{-1} t}} \ \forall \ (i) \in \text{class } k \tag{A.8}$$

where $\mathbf{Cov}^{-1}$ is the inverse of covariance matrix. Next to suppress the sensor noise, we exploit the simple method of tracking each sensor. Each sensor have different physical properties, that can influence the data captured by it. We track each individual sensor in the flight line. Since the data is rectified, the data from each sensor does not belong to single column, instead it is spread randomly across all the columns. This is dependent on the flight path and the movement in the airplane while moving. We used simple data-structure algorithms like depth first search. Tracked each boundary pixels and assigned them to single sensor. We used data from 10-15 adjacent sensor at one time, normalize it and then compute the covariance matrix in previous step with segmentation mask. Our approach is very simple and straight forward.

The algorithm A.1.2 shows the pseudo code for our Spectral Linear Filter (**SLF**).

Training policy

We trained MethaneMapper in two styles, (i) pre-training the bounding box and class detection first and then freezing the pre-trained model parameters and training only the mask prediction layer; and (ii) trained whole pipeline end-to-end and achieved similar performance on both the cases.

Data: *MHS dataset*
Result: *CH_4 concentration map*
initialization;
for mhs *in* MHS **do**
    1. create memory map mhs;
    2. seg_mask = compute segmentation mask;
    for *mask in seg_mask* **do**
        data.**append**(mhs[mask])
        if (len(data) $<$ 100000): **continue**
        Cov, μ = compute_stats(data);
    end
    3. sensor_array = individual sensors;
    for *arrays in sensor_array* **do**
        data = mhs[arrays]
        for x_i *in data* **do**
            k = seg_mask[i];
            $\mathbf{SLF}(\mathbf{x}_i) = \frac{(\mathbf{x}_i - \mu_k)^T \mathbf{Cov}_k^{-1} \mathbf{t}}{\sqrt{\mathbf{t}^T \mathbf{Cov}_k^{-1} \mathbf{t}}}$
        end
    end
    $\mathbf{SLF}(\mathbf{x}_i)\ \forall\ classes\ and\ i \in mhs$
end

Algorithm 1: Enhanced Matched Filter

Qualitative Results

In this section we show few more qualitative examples of CH_4 plume mask prediction and few cases where MM failed to detect any CH_4 gas emission.Figure. A.3 shows the CH_4 detections in different types of background terrain and different types of emission source.

Figure A.4 shows some examples of missed CH_4 plume detections. We observed that going back to dataset samples and checking the timelines, these flightlines were recorded during the evening time. We believe that this might be because of evening time, the reflectance from the ground terrain is very weak and small. Hence we believe there is minimum absorption of reflected solar radiation by CH_4 gas present in the atmosphere and the plume goes undetected.

A.1.3 Ablations Studies

Attention Type: We also explored different attention mechanisms to encode and decode information. We replaced only the attention layers with deformable-attention [121] in the our architecture that resulted in a drop of 0.1 mAP in the baseline model.

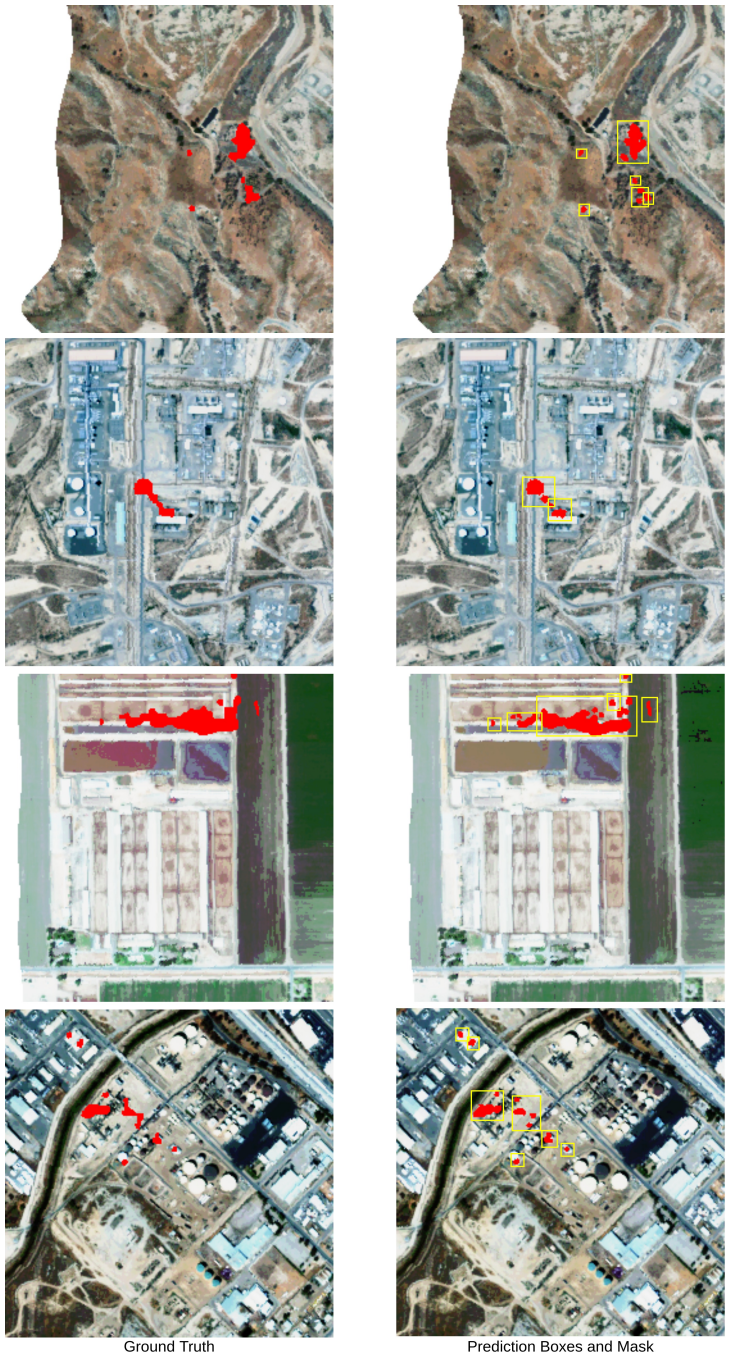

Figure A.3: *Sample ground truths and predictions on MHS dataset. We are showing different type of terrains and CH_4 predictions on them. The type of emission source in all samples varies too.*

Implementation details

The whole network is trained with AdamW [135] optimizer, batch size of 12, with initial learning rate for backbones set to 10^{-4} and for transformer the learning rate is set

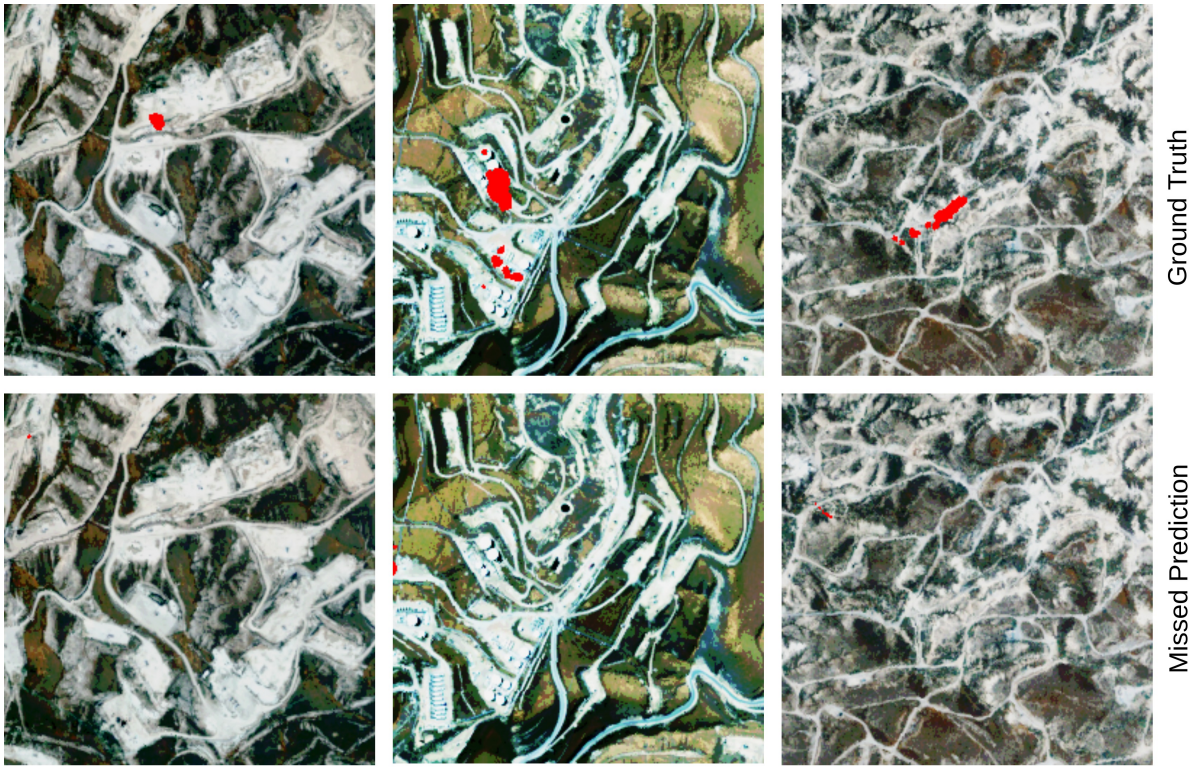

Figure A.4: *Samples where MM fails to detect the CH_4 plume. We observed that these samples were recorded during the evening time and hence reflectance from the ground terrain is very weak. Therefore the absorption of reflected solar radiations by CH_4 is very low and hence the emissions goes undetected.*

to 10^{-5} with a weight decay of 10^{-4}. The learning rate for mask prediction module is set to 10^{-4}. The learning rate is dropped at every 150 epochs, we train for 300 epochs. The baseline model is trained on 2 V100 GPUs.

Appendix B

WildlifeMapper Appendix

B.1 Introduction

We provide all the additional information related to WildlifeMapper (WM) and the Mara-Wildlife (MW) dataset here. We also provide qualitative examples.

	Methods	**mAP**	**Counting Error**
	Mara Wildlife Dataset		
1	*Faster-Rcnn*	0.24	2.59
2	*DETR*	0.22	2.75
3	*Co-DETR-R50*	0.27	2.72
4	*Co-DETR-swingL*	0.28	2.60
5	*Yolo v5*	0.30	2.12
6	*Yolo v8*	0.27	3.97
7	*LSKNet*	0.29	-
8	*DroneDetect*	0.18	-
9	***WildlifeMapper***	**0.56**	**1.9**

Table B.1: *Comparison with baseline models. The top section shows performance on species detection on Mara-wildlife dataset and low section shows performance on the mixed dataset from Virunga-Garamba-AED dataset. The overall detection accuracy is generally higher in Virunga-Garamba-AED dataset because there are only 6 species and the terrain is quite similar in all images.*

B.1.1 Method

High Frequency Feature Generator (HFG): This section covers the detailed derivation and implementation information of our **HFG** module. The input image is processed in parallel by the **HFG** module to generate features with information about the location of the animal or cluster. The **HFG** module is inspired from the limitation of ViT mod-

els [87]. ViT models face challenges in efficiently utilizing local structures. They segment an image into patches and apply self-attention to model relationships, but this approach often falls short in capturing detailed local features [129, 130].

Research indicates that local features in images are closely linked to high-frequency components [131, 132]. We hypothesize that suppressing low-frequency components can mitigate the influence of a dominant homogeneous background. To test this, we performed a discrete Fourier Transform (DFT) on the images, filtering out the low-frequency components before reconstructing the images..

For a given input image $I \in \mathbb{R}^{H \times W \times C}$, where C is channel dimension, we compute Discrete Fourier Transform (DFT) of I. In next step we suppress the low frequency components with a controlling parameter and construct the image I with inverse ($IDFT$) to get back image I'. The DFT is computed as:

$$F(u,v) = \sum_{x=0}^{H-1} \sum_{y=0}^{W-1} I(x,y) \cdot e^{-j2\pi(ux/H + vy/W)} \tag{B.1}$$

where $F(u,v)$ is the magnitude spectrum, u and v are the frequency coordinates, and j is the imaginary unit. Next, we shift the lower frequency components to center of frequency spectrum as:

$$F'(u,v) = F\left((u + \frac{H}{2}) \mod H, (v + \frac{W}{2}) \mod W\right), \tag{B.2}$$

where *mod* is modulus operation. Next we mask the lower frequencies with a controlling parameter r. The modified Fourier transform G with mask M is defined as:

$$G(u,v) = M(u,v) \odot F'(u,v) \tag{B.3}$$

$$\text{where } M = \begin{cases} \blacksquare & \text{if} (u - H/2)^2 + (v - W/2)^2 \leq r^2 \\ 0 & \text{otherwise} \end{cases} \tag{B.4}$$

Given the modified Fourier transform G, the reconstructed image I' is given by:

$$I'(x,y) = \sum_{u=0}^{H-1} \sum_{v=0}^{W-1} G(u,v) \cdot e^{j2\pi(ux/H + vy/W)} \tag{B.5}$$

Next we reduce the dimension of the reconstructed image $I'(x,y)$ via an embedding layer to generate embedding hfc_{emb} and pass them to the **FR** module

B.1.2 Implementation Details

Each image taken from the drone is $8256 \times 5506 \times 3$. We create tiles for each image in the spatial domain, with the size of $1024 \times 1024 \times 3$ with 25% of overlap. The Patch Embed layer uses a single CNN layer with a large kernel of size 16×16 with stride 16. In the parallel branch, the *High-Frequency Feature Generator*, we use DFT to compute the Fourier transform, the mask is a binary disk with the radius set to 128. The HFC Embed layer uses 3 CNN layers with ReLU activation with a kernel of size 3×3 and a global average pool at the end. The Feature Refiner (**FR**) module consists of one cross attention layer with 1 linear layer. The image encoder is a pre-trained ViT model [3] with 24 transformer layers and 16 heads. The Query Refiner (**QR**) module takes in 100 queries each of channel dimension 256, those are cross attended with hfc_{emb} output. The box decoder contains 3 layers of two-way attention with 8 heads. We train WM with AdamW optimizer [135] setting the learning rate to 10^{-4} for the **FR**, **QR** and box decoder with a weight decay to 10^{-4}. We set the learning rate for the Patch Embed and HFC Embed layer to 10^{-5}. We load the image encoder with pre-trained weights from segment anything [3] and keep it frozen.

B.1.3 Mara-Wildlife Dataset:

Flight path details The chosen flight path prioritized vast open grasslands, as they frequently serve as habitats and transit routes for larger fauna. The survey was conducted in March. March is typically a rainy month when grasses are green. The Serengeti migratory herd of wildebeest have already moved south from the region. Data collection was typically scheduled during the early mornings or late afternoons. These times are when animals are most active, avoiding the midday sun. Although the evening presents challenges due to diminished sunlight, the majority of the data was acquired between 7AM and 10AM local time to ensure optimal lighting conditions.

Camera Settings and Specifications: We mounted a NIKON D850 camera to the bellyport of the airplane. The camera was placed in a NADIR view and configured with an intervalometer to collect an image every two seconds along flight transects.

B.1.4 Rich Metadata for Computer Vision Benchmarks:

Each raster included in the dataset is accompanied with detailed metadata, the timestamp of the image capture, and other camera EXIF (Exchangeable Image File Format) information such as focal length, FNumber, ISO, and ExposureTime. We collected latitude and longitude and elevation information from the GPS log of the pilot and merged this information using the camera data and time. If released with the data, these metadata properties enrich the dataset's ecological value and unlock the potential for a myriad of computational applications.

While the primary intent of our analysis was to provide an estimate of the abundance of large mammals across the Masai Mara ecosystem, the dataset's comprehensive nature presents opportunities that extend beyond wildlife studies. These include: *Sun Angle Prediction*, *Image Registration*, *GPS Estimation*, *Elevation Prediction*.

- *Sun Angle Prediction:* Given the timestamp and known location of each image capture, the dataset could be employed to develop models that predict the sun's angle based on the image content. Such applications can benefit fields ranging from photovoltaic systems to architectural planning.
- *Image Registration:* The dataset provides a platform for researchers to work on algorithms that align or 'register' multiple images of the same region, even if taken from varying angles or times. Such tasks find relevance in areas like medical imaging and satellite image analysis.
- *GPS Estimation:* The precise latitudinal and longitudinal coordinates embedded in the metadata allow for the creation of models that predict the GPS location of specific objects or even individual pixels using only the image content. This potential extends the bounds of localization models in the realm of computer vision.
- *Elevation Prediction:* The dataset's rich elevation data provides an avenue to train models that can estimate the altitude at which an image was taken, based purely on visual cues. Such models can have vast applications, from aviation to drone technology.

These represent just a few of the many potential applications. We believe the Mara-Wildlife dataset has the potential to be a foundational resource for both ecological studies and computer vision research, ushering in innovations and novel solutions.

B.1.5 Results

In this section, we present more qualitative results of detection of WildlifeMapper on the Mara-Wildlife dataset. Good detection samples are shown in Fig. B.1 and the failure cases are shown in Fig.

Good cases: Each column in Fig. B.1 shows detections of different types of animals. Column-1 shows large animals: $cattle, buffalo$; Column-2 shows detection of small animals: $warthog, topi$. Column-3 shows detection of animals hidden or occluded in Row 1 & 2, Row 3 & 4 show examples of *other* categories (i.e., lion).

Failure cases: Fig. B.2 shows examples of where WildlifeMapper struggled to make an accurate detection. Each image shows a unique scenario where the detection was either missed or misclassified or confused from the contextual information. For example in Column-1, Row-1, the dry wooden log is detected as an object and misclassified as $shoat(sheep\ or\ goat)$ since *shoats* almost always occur in a group. Hence, the additional object identified was mislabeled a *shoat*. ***Figures are on next page.***

Figure B.1: *Good cases. Each column shows different category of detection. Column-1 shows large animals: cattle, buffalo; Column-2 shows detection of small animals (warthog, topi), Column-3 shows detection of animals hidden or occluded.*

Figure B.2: *Failure cases. The animals hiding in the shade are difficult to detect. Additional examples of misclassification also provided.*

Bibliography

[1] I. Gordon, L. Rothman, R. Hargreaves, R. Hashemi, E. Karlovets, F. Skinner, E. Conway, C. Hill, R. Kochanov, Y. Tan, *et. al.*, *The hitran2020 molecular spectroscopic database*, *Journal of quantitative spectroscopy and radiative transfer* **277** (2022) 107949.

[2] D. R. Thompson, A. Karpatne, I. Ebert-Uphoff, C. Frankenberg, A. K. Thorpe, B. D. Bue, and R. O. Green, *Isgeo dataset jpl-ch4-detection-2017-v1. 0: A benchmark for methane source detection from imaging spectrometer data*, .

[3] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, *et. al.*, *Segment anything*, *arXiv preprint arXiv:2304.02643* (2023).

[4] B. Xiao, H. Wu, W. Xu, X. Dai, H. Hu, Y. Lu, M. Zeng, C. Liu, and L. Yuan, *Florence-2: Advancing a unified representation for a variety of vision tasks*, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4818–4829, 2024.

[5] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, *et. al.*, *Gpt-4 technical report*, *arXiv preprint arXiv:2303.08774* (2023).

[6] K. Kvilekval, D. Fedorov, B. Obara, A. Singh, and B. Manjunath, *Bisque: a platform for bioimage analysis and management*, Jan., 2023.

[7] J. Lelieveld, P. J. Crutzen, and F. J. Dentener, *Changing concentration, lifetime and climate forcing of atmospheric methane*, *Tellus B* **50** (1998), no. 2 128–150.

[8] P. Ciais, C. Sabine, G. Bala, L. Bopp, V. Brovkin, J. Canadell, A. Chhabra, R. DeFries, J. Galloway, M. Heimann, *et. al.*, *Carbon and other biogeochemical cycles*, in *Climate change 2013: the physical science basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, pp. 465–570. Cambridge University Press, 2014.

[9] S. Kumar, C. Torres, O. Ulutan, A. Ayasse, D. Roberts, and B. Manjunath, *Deep remote sensing methods for methane detection in overhead hyperspectral imagery*,

in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1776–1785, 2020.

[10] S. Kumar, I. Arevalo, A. Iftekhar, and B. Manjunath, *Methanemapper: Spectral absorption aware hyperspectral transformer for methane detection*, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17609–17618, 2023.

[11] P. IEA, *Methane from oil & gas*, 2020.

[12] S. Kirschke, P. Bousquet, P. Ciais, M. Saunois, J. G. Canadell, E. J. Dlugokencky, P. Bergamaschi, D. Bergmann, D. R. Blake, L. Bruhwiler, *et. al.*, *Three decades of global methane sources and sinks*, *Nature geoscience* **6** (2013), no. 10 813.

[13] C. I. o. T. Jet Propulsion Laboratory, *Airborne visible infrared imaging spectrometer - next generation (aviris-ng) overview*, 2009.

[14] D. A. Roberts, E. S. Bradley, R. Cheung, I. Leifer, P. E. Dennison, and J. S. Margolis, *Mapping methane emissions from a marine geological seep source using imaging spectrometry*, *Remote Sensing of Environment* **114** (2010), no. 3 592–606.

[15] D. Thompson, I. Leifer, H. Bovensmann, M. Eastwood, M. Fladeland, C. Frankenberg, K. Gerilowski, R. Green, S. Kratwurst, T. Krings, *et. al.*, *Real-time remote detection and measurement for airborne imaging spectroscopy: a case study with methane*, *Atmospheric Measurement Techniques* **8** (2015), no. 10 4383–4397.

[16] A. K. Thorpe, D. A. Roberts, E. S. Bradley, C. C. Funk, P. E. Dennison, and I. Leifer, *High resolution mapping of methane emissions from marine and terrestrial sources using a cluster-tuned matched filter technique and imaging spectrometry*, *Remote Sensing of Environment* **134** (2013) 305–318.

[17] A. K. Thorpe, C. Frankenberg, D. R. Thompson, R. M. Duren, A. D. Aubrey, B. D. Bue, R. O. Green, K. Gerilowski, T. Krings, J. Borchardt, *et. al.*, *Airborne doas retrievals of methane, carbon dioxide, and water vapor concentrations at high spatial resolution: application to aviris-ng*, *Atmospheric Measurement Techniques* **10** (2017), no. 10 3833.

[18] C. Frankenberg, A. K. Thorpe, D. R. Thompson, G. Hulley, E. A. Kort, N. Vance, J. Borchardt, T. Krings, K. Gerilowski, C. Sweeney, *et. al.*, *Airborne methane remote measurements reveal heavy-tail flux distribution in four corners region*, *Proceedings of the national academy of sciences* **113** (2016), no. 35 9734–9739.

[19] C. Frankenberg, U. Platt, and T. Wagner, *Iterative maximum a posteriori (imap)-doas for retrieval of strongly absorbing trace gases: Model studies for ch 4*

and co 2 retrieval from near infrared spectra of sciamachy onboard envisat, *Atmospheric Chemistry and Physics* **5** (2005), no. 1 9–22.

[20] S. Jongaramrungruang, A. K. Thorpe, G. Matheou, and C. Frankenberg, *Methanet–an ai-driven approach to quantifying methane point-source emission from high-resolution 2-d plume imagery*, *Remote Sensing of Environment* **269** (2022) 112809.

[21] A. K. Ayasse, A. K. Thorpe, D. A. Roberts, C. C. Funk, P. E. Dennison, C. Frankenberg, A. Steffke, and A. D. Aubrey, *Evaluating the effects of surface properties on methane retrievals using a synthetic airborne visible/infrared imaging spectrometer next generation (aviris-ng) image*, *Remote Sensing of Environment* **215** (2018) 386–397.

[22] U. B. Gewali, S. T. Monteiro, and E. Saber, *Machine learning based hyperspectral image analysis: a survey*, *arXiv preprint arXiv:1802.08701* (2018).

[23] Q. Cheng, P. K. Varshney, and M. K. Arora, *Logistic regression for feature selection and soft classification of remote sensing data*, *IEEE Geoscience and Remote Sensing Letters* **3** (2006), no. 4 491–494.

[24] M. Khodadadzadeh, J. Li, A. Plaza, and J. M. Bioucas-Dias, *A subspace-based multinomial logistic regression for hyperspectral image classification*, *IEEE Geoscience and Remote Sensing Letters* **11** (2014), no. 12 2105–2109.

[25] G. Camps-Valls, *Support vector machines in remote sensing: the tricks of the trade*, in *Image and Signal Processing for Remote Sensing XVII*, vol. 8180, p. 81800B, International Society for Optics and Photonics, 2011.

[26] D. M. Tax and R. P. Duin, *Support vector data description*, *Machine learning* **54** (2004), no. 1 45–66.

[27] W. Sakla, A. Chan, J. Ji, and A. Sakla, *An svdd-based algorithm for target detection in hyperspectral imagery*, *IEEE Geoscience and Remote Sensing Letters* **8** (2010), no. 2 384–388.

[28] C. Shah, P. K. Varshney, and M. Arora, *Ica mixture model algorithm for unsupervised classification of remote sensing imagery*, *International Journal of Remote Sensing* **28** (2007), no. 8 1711–1731.

[29] Y. Chen, X. Zhao, and X. Jia, *Spectral–spatial classification of hyperspectral data based on deep belief network*, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **8** (2015), no. 6 2381–2392.

[30] S. T. Monteiro, Y. Minekawa, Y. Kosugi, T. Akazawa, and K. Oda, *Prediction of sweetness and amino acid content in soybean crops from hyperspectral imagery*, *ISPRS Journal of Photogrammetry and Remote Sensing* **62** (2007), no. 1 2–12.

[31] J. Xia, P. Du, X. He, and J. Chanussot, *Hyperspectral remote sensing image classification based on rotation forest*, *IEEE Geoscience and Remote Sensing Letters* **11** (2013), no. 1 239–243.

[32] G.-B. Huang, D. H. Wang, and Y. Lan, *Extreme learning machines: a survey*, *International journal of machine learning and cybernetics* **2** (2011), no. 2 107–122.

[33] Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu, *Deep learning-based classification of hyperspectral data*, *IEEE Journal of Selected topics in applied earth observations and remote sensing* **7** (2014), no. 6 2094–2107.

[34] W. Zhao and S. Du, *Spectral–spatial feature extraction for hyperspectral image classification: A dimension reduction and deep learning approach*, *IEEE Transactions on Geoscience and Remote Sensing* **54** (2016), no. 8 4544–4554.

[35] W. Zhao, Z. Guo, J. Yue, X. Zhang, and L. Luo, *On combining multiscale deep learning features for the classification of hyperspectral remote sensing imagery*, *International Journal of Remote Sensing* **36** (2015), no. 13 3368–3379.

[36] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, *Deep feature extraction and classification of hyperspectral images based on convolutional neural networks*, *IEEE Transactions on Geoscience and Remote Sensing* **54** (2016), no. 10 6232–6251.

[37] C. Frankenberg, U. Platt, and T. Wagner, *Iterative maximum a posteriori (imap)-doas for retrieval of strongly absorbing trace gases: Model studies for ch 4 and co 2 retrieval from near infrared spectra of sciamachy onboard envisat*, *Atmospheric Chemistry and Physics Discussions* **4** (2004), no. 5 6067–6106.

[38] C. C. Funk, J. Theiler, D. A. Roberts, and C. C. Borel, *Clustering to improve matched filter detection of weak gas plumes in hyperspectral thermal imagery*, *IEEE transactions on geoscience and remote sensing* **39** (2001), no. 7 1410–1420.

[39] D. R. Thompson, A. Karpatne, I. Ebert-Uphoff, C. Frankenberg, A. K. Thorpe, B. D. Bue, and R. O. Green, *Is-geo dataset jpl-ch4-detection-2017-v1. 0: A benchmark for methane source detection from imaging spectrometer data*, .

[40] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, *End-to-end object detection with transformers*, in *European conference on computer vision*, pp. 213–229, Springer, 2020.

[41] E. K. David, H. Joseph, S. Megs, and P. A. Gregory, *Global airborne observatory visible to infrared imaging spectrometer report*, 2020.

[42] I. CARBON MAPPER, *Carbon mapper methane emission exploration*, 2022.

[43] META-STANFORD, *Methane emissions technology alliance (meta), stanford natural gas initiative*, 2022.

[44] D. Phiri, M. Simwanda, S. Salekin, V. R. Nyirenda, Y. Murayama, and M. Ranagalage, *Sentinel-2 data for land cover/use mapping: a review*, *Remote Sensing* **12** (2020), no. 14 2291.

[45] A. Iftekhar, S. Kumar, R. A. McEver, S. You, and B. Manjunath, *Gtnet: Guided transformer network for detecting human-object interactions*, *arXiv preprint arXiv:2108.00596* (2021).

[46] A. Iftekhar, H. Chen, K. Kundu, X. Li, J. Tighe, and D. Modolo, *What to look at and where: Semantic and spatial refined transformer for detecting human-object interactions*, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5353–5363, 2022.

[47] K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition*, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[48] S. Kumar, A. Iftekhar, M. Goebel, T. Bullock, M. H. MacLean, M. B. Miller, T. Santander, B. Giesbrecht, S. T. Grafton, and B. Manjunath, *Stressnet: detecting stress in thermal videos*, in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 999–1009, 2021.

[49] O. Ulutan, A. Iftekhar, and B. S. Manjunath, *Vsgnet: Spatial attention network for detecting human object interactions using graph convolutions*, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13617–13626, 2020.

[50] S. Kumar, A. Iftekhar, E. Prashnani, and B. Manjunath, *Locl: Learning object-attribute composition using localization*, *arXiv preprint arXiv:2210.03780* (2022).

[51] G. Turin, *An introduction to matched filters*, *IRE transactions on Information theory* **6** (1960), no. 3 311–329.

[52] F. A. S. US Dept. of Agriculture, *Normalized difference vegetation index (ndvi)*, 1969.

[53] MGISGeography, *Ndvi (normalized difference vegetation index)*, 1979.

[54] B.-C. Gao, *Ndwi—a normalized difference water index for remote sensing of vegetation liquid water from space*, *Remote sensing of environment* **58** (1996), no. 3 257–266.

[55] K. He, G. Gkioxari, P. Dollár, and R. Girshick, *Mask r-cnn*, in *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

[56] C. A. R. BOARD, *Green house gas inventory by california air resource board*, 2022.

[57] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, *Microsoft coco: Common objects in context*, in *European conference on computer vision*, pp. 740–755, Springer, 2014.

[58] K. He, X. Zhang, S. Ren, and J. Sun, *Delving deep into rectifiers: Surpassing human-level performance on imagenet classification*, in *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

[59] D. Hong, Z. Han, J. Yao, L. Gao, B. Zhang, A. Plaza, and J. Chanussot, *Spectralformer: Rethinking hyperspectral image classification with transformers*, *IEEE Transactions on Geoscience and Remote Sensing* **60** (2021) 1–15.

[60] Y. Xiong, R. Liao, H. Zhao, R. Hu, M. Bai, E. Yumer, and R. Urtasun, *Upsnet: A unified panoptic segmentation network*, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8818–8826, 2019.

[61] S. Kumar, W. Kingwill, O. Earth, R. Mouton, W. Adamczyk, R. Huppertz, and E. Sherwin, *Guided transformer network for detecting methane emissions in sentinel-2 satellite imagery*, .

[62] S. Kumar, W. Kingwill, R. Mouton, W. Adamczyk, R. Huppertz, and E. D. Sherwin, *Guided transformer network for detecting methane emissions in sentinel-2 satellite imagery*, in *NeurIPS 2022 Workshop on Tackling Climate Change with Machine Learning*, 2022.

[63] M. Crowell, *President biden announces new methane emissions reduction strategy*, 2022.

[64] U. D. of State, *U.s.-eu joint press release on the global methane pledge energy pathway*, 2022.

[65] D. J. Varon, D. Jervis, J. McKeever, I. Spence, D. Gains, and D. J. Jacob, *High-frequency monitoring of anomalous methane point sources with multispectral sentinel-2 satellite observations*, *Atmospheric Measurement Techniques* **14** (2021), no. 4 2771–2785.

[66] D. H. Cusworth, D. J. Jacob, D. J. Varon, C. Chan Miller, X. Liu, K. Chance, A. K. Thorpe, R. M. Duren, C. E. Miller, D. R. Thompson, *et. al.*, *Potential of next-generation imaging spectrometers to detect and quantify methane point sources from space*, *Atmospheric Measurement Techniques* **12** (2019), no. 10 5655–5668.

[67] T. Hamazaki, Y. Kaneko, A. Kuze, and K. Kondo, *Fourier transform spectrometer for greenhouse gases observing satellite (gosat)*, in *Enabling sensor and platform technologies for spaceborne remote sensing*, vol. 5659, pp. 73–80, SPIE, 2005.

[68] S. Pandey, R. Gautam, S. Houweling, H. D. Van Der Gon, P. Sadavarte, T. Borsdorff, O. Hasekamp, J. Landgraf, P. Tol, T. Van Kempen, *et. al.*, *Satellite observations reveal extreme methane leakage from a natural gas well blowout*, *Proceedings of the National Academy of Sciences* **116** (2019), no. 52 26376–26381.

[69] Z. Zhang, E. D. Sherwin, D. J. Varon, and A. R. Brandt, *Detecting and quantifying methane emissions from oil and gas production: algorithm development with ground-truth calibration based on sentinel-2 satellite imagery*, *EGUsphere* (2022) 1–23.

[70] D. J. Jacob, A. J. Turner, J. D. Maasakkers, J. Sheng, K. Sun, X. Liu, K. Chance, I. Aben, J. McKeever, and C. Frankenberg, *Satellite observations of atmospheric methane and their value for quantifying methane emissions*, *Atmospheric Chemistry and Physics* **16** (2016), no. 22 14371–14396.

[71] T. Ehret, A. D. Truchis, M. Mazzolini, J.-M. Morel, A. d'Aspremont, T. Lauvaux, R. M. Duren, D. H. Cusworth, and G. Facciolo, *Global tracking and quantification of oil and gas methane emissions from recurrent sentinel-2 imagery.*, *Environmental science & technology* (2022).

[72] T. Ehret, A. De Truchis, M. Mazzolini, J.-M. Morel, A. d'Aspremont, T. Lauvaux, R. Duren, D. Cusworth, and G. Facciolo, *Global tracking and quantification of oil and gas methane emissions from recurrent sentinel-2 imagery*, *Environmental science & technology* **56** (2022), no. 14 10517–10529.

[73] D. Thompson, A. Thorpe, C. Frankenberg, R. Green, R. Duren, A. Hollstein, L. Guanter, E. Middleton, L. Ong, and S. Ungar, *Orbital measurement of the aliso canyon ch4 super-emitter*, *Geophys Res Lett* **43** (2016) 6571–6578.

[74] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et. al.*, *An image is worth 16x16 words: Transformers for image recognition at scale*, *arXiv preprint arXiv:2010.11929* (2020).

[75] T. Yokota, Y. Yoshida, N. Eguchi, Y. Ota, T. Tanaka, H. Watanabe, and S. Maksyutov, *Global concentrations of co2 and ch4 retrieved from gosat: First preliminary results*, *Sola* **5** (2009) 160–163.

[76] J. P. Veefkind, I. Aben, K. McMullan, H. Förster, J. De Vries, G. Otter, J. Claas, H. Eskes, J. De Haan, Q. Kleipool, *et. al.*, *Tropomi on the esa sentinel-5 precursor: A gmes mission for global observations of the atmospheric composition*

for climate, air quality and ozone layer applications, *Remote sensing of environment* **120** (2012) 70–83.

[77] O. Schneising, M. Buchwitz, M. Reuter, S. Vanselow, H. Bovensmann, and J. P. Burrows, *Remote sensing of methane leakage from natural gas and petroleum systems revisited*, *Atmospheric Chemistry and Physics* **20** (2020), no. 15 9169–9182.

[78] J. Barré, I. Aben, A. Agustí-Panareda, G. Balsamo, N. Bousserez, P. Dueben, R. Engelen, A. Inness, A. Lorente, J. McNorton, *et. al.*, *Systematic detection of local ch 4 anomalies by combining satellite measurements with high-resolution forecasts*, *Atmospheric Chemistry and Physics* **21** (2021), no. 6 5117–5136.

[79] T. Lauvaux, C. Giron, M. Mazzolini, A. d'Aspremont, R. Duren, D. Cusworth, D. Shindell, and P. Ciais, *Global assessment of oil and gas methane ultra-emitters*, *Science* **375** (2022), no. 6580 557–561.

[80] J. McKeever, H. Deglint, D. Gains, D. Jervis, J. MacLean, A. Ramier, W. Shaw, M. Strupler, E. Tarrant, D. Varon, *et. al.*, *First methane sensing results from ghgsat's commercial constellation*, in *17th International Workshop on Greenhouse Gas Measurements from Space, June*, pp. 14–17, 2021.

[81] R. D. M. Scafutto, H. van der Werff, W. H. Bakker, F. van der Meer, and C. R. de Souza Filho, *An evaluation of airborne swir imaging spectrometers for ch4 mapping: Implications of band positioning, spectral sampling and noise*, *International Journal of Applied Earth Observation and Geoinformation* **94** (2021) 102233.

[82] S. Conley, I. Faloona, S. Mehrotra, M. Suard, D. H. Lenschow, C. Sweeney, S. Herndon, S. Schwietzke, G. Pétron, J. Pifer, *et. al.*, *Application of gauss's theorem to quantify localized surface emissions from airborne measurements of wind and trace gases*, *Atmospheric Measurement Techniques* **10** (2017), no. 9 3345–3358.

[83] E. D. Sherwin, Y. Chen, A. P. Ravikumar, and A. R. Brandt, *Single-blind test of airplane-based hyperspectral methane detection via controlled releases*, *Elem Sci Anth* **9** (2021), no. 1 00063.

[84] G. J. Komar, J. Wang, and T. Kimura, *Enabling sensor and platform technologies for spaceborne remote sensing*, *Enabling Sensor and Platform Technologies for Spaceborne Remote Sensing* **5659** (2005).

[85] F. Innocenti, R. Robinson, T. Gardiner, A. Finlayson, and A. Connor, *Differential absorption lidar (dial) measurements of landfill methane emissions*, *Remote sensing* **9** (2017), no. 9 953.

[86] I. Leifer, D. Roberts, J. Margolis, and F. Kinnaman, *In situ sensing of methane emissions from natural marine hydrocarbon seeps: A potential remote sensing technology*, *Earth and Planetary Science Letters* **245** (2006), no. 3-4 509–522.

[87] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, *Attention is all you need*, *Advances in neural information processing systems* **30** (2017).

[88] H. Taud and J. Mas, *Multilayer perceptron (mlp)*, in *Geomatic approaches for modeling land change scenarios*, pp. 451–455. Springer, 2018.

[89] A. Ražnjević, C. van Heerwaarden, B. van Stratum, A. Hensen, I. Velzeboer, P. van den Bulk, and M. Krol, *Interpretation of field observations of point-source methane plume using observation-driven large-eddy simulations*, *Atmospheric Chemistry and Physics* **22** (2022), no. 10 6489–6505.

[90] E. D. Sherwin, J. S. Rutherford, Y. Chen, S. Aminfard, E. A. Kort, R. B. Jackson, and A. R. Brandt, *Single-blind validation of space-based point-source methane emissions detection and quantification*, .

[91] I. Irakulis-Loitxate, L. Guanter, Y.-N. Liu, D. J. Varon, J. D. Maasakkers, Y. Zhang, A. Chulakadabba, S. C. Wofsy, A. K. Thorpe, R. M. Duren, *et. al.*, *Satellite-based survey of extreme methane emissions in the permian basin*, *Science Advances* **7** (2021), no. 27 eabf4507.

[92] S. Kumar, B. Zhang, C. Gudavalli, C. Levenson, L. Hughey, J. A. Stabach, I. Amoke, G. Ojwang, J. Mukeka, S. Mwiu, *et. al.*, *Wildlifemapper: Aerial image analysis for multi-species detection and identification*, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12594–12604, 2024.

[93] R. Lamprey, F. Pope, S. Ngene, M. Norton-Griffiths, H. Frederick, B. Okita-Ouma, and I. Douglas-Hamilton, *Comparing an automated high-definition oblique camera system to rear-seat-observers in a wildlife survey in tsavo, kenya: taking multi-species aerial counts to the next level*, *Biological Conservation* **241** (2020) 108243.

[94] M. J. S. Peel and M. Stalmans, *The systematic reconnaissance flight (srf) as a tool in assessing the ecological impact of a rural development programme in an extensive area of the lowveld of south africa*, *African Journal of Ecology* **37** (1999), no. 4 449–456.

[95] W. K. Ottichilo, J. Grunblatt, M. Y. Said, and P. W. Wargute, *Wildlife and livestock population trends in the kenya rangeland*, *Wildlife conservation by sustainable use* (2000) 203–218.

[96] C. J. Torney, D. J. Lloyd-Jones, M. Chevallier, D. C. Moyer, H. T. Maliti, M. Mwita, E. M. Kohi, and G. C. Hopcraft, *A comparison of deep learning and citizen science techniques for counting wildlife in aerial survey images*, *Methods in Ecology and Evolution* **10** (2019), no. 6 779–787.

[97] E. Guirado, S. Tabik, M. L. Rivas, D. Alcaraz-Segura, and F. Herrera, *Whale counting in satellite and aerial images with deep learning*, *Scientific reports* **9** (2019), no. 1 14259.

[98] A. Seymour, J. Dale, M. Hammill, P. Halpin, and D. Johnston, *Automated detection and enumeration of marine wildlife using unmanned aircraft systems (uas) and thermal imagery*, *Scientific reports* **7** (2017), no. 1 45127.

[99] A. Delplanque, S. Foucher, P. Lejeune, J. Linchant, and J. Théau, *Multispecies detection and identification of african mammals in aerial imagery using convolutional neural networks*, *Remote Sensing in Ecology and Conservation* **8** (2022), no. 2 166–179.

[100] T. Petso, R. S. Jamisola Jr, D. Mpoeleng, E. Bennitt, and W. Mmereki, *Automatic animal identification from drone camera based on point pattern analysis of herd behaviour*, *Ecological Informatics* **66** (2021) 101485.

[101] B. Kellenberger, D. Marcos, and D. Tuia, *Detecting mammals in uav images: Best practices to address a substantially imbalanced dataset with deep learning*, *Remote Sensing of Environment* **216** (Oct, 2018) 139–153.

[102] L. Han, P. Tao, and R. R. Martin, *Livestock detection in aerial images using a fully convolutional network*, *Computational Visual Media* **5** (2019) 221–228.

[103] J. O. Ogutu, H.-P. Piepho, M. Y. Said, G. O. Ojwang, L. W. Njino, S. C. Kifugo, and P. W. Wargute, *Extreme wildlife declines and concurrent increase in livestock numbers in kenya: What are the causes?*, *PloS one* **11** (2016), no. 9 e0163249.

[104] W. K. Ottichilo, J. Grunblatt, M. Y. Said, and P. W. Wargute, *Wildlife and Livestock Population Trends in the Kenya Rangeland*, pp. 203–218. Springer Netherlands, Dordrecht, 2000.

[105] R. Lamprey, F. Pope, S. Ngene, M. Norton-Griffiths, H. Frederick, B. Okita-Ouma, and I. Douglas-Hamilton, *Comparing an automated high-definition oblique camera system to rear-seat-observers in a wildlife survey in tsavo, kenya: Taking multi-species aerial counts to the next level*, *Biological Conservation* **241** (2020) 108243.

[106] C. Vermeulen, P. Lejeune, J. Lisein, P. Sawadogo, and P. Bouché, *Unmanned aerial survey of elephants*, *PloS one* **8** (2013), no. 2 e54700.

[107] J. Linchant, J. Lisein, J. Semeki, P. Lejeune, and C. Vermeulen, *Are unmanned aircraft systems (uas s) the future of wildlife monitoring? a review of accomplishments and challenges*, *Mammal Review* **45** (2015), no. 4 239–252.

[108] R. Lamprey, D. Ochanda, R. Brett, C. Tumwesigye, and I. Douglas-Hamilton, *Cameras replace human observers in multi-species aerial counts in murchison falls, uganda*, *Remote Sensing in Ecology and Conservation* **6** (2020), no. 4 529–545.

[109] P. Zhu, L. Wen, X. Bian, H. Ling, and Q. Hu, *Vision meets drones: A challenge*, *arXiv preprint arXiv:1804.07437* (2018).

[110] J. Ding, N. Xue, G.-S. Xia, X. Bai, W. Yang, M. Y. Yang, S. Belongie, J. Luo, M. Datcu, M. Pelillo, *et. al.*, *Object detection in aerial images: A large-scale benchmark and challenges*, *IEEE transactions on pattern analysis and machine intelligence* **44** (2021), no. 11 7778–7796.

[111] M.-R. Hsieh, Y.-L. Lin, and W. H. Hsu, *Drone-based object counting by spatially regularized regional proposal networks*, in *The IEEE International Conference on Computer Vision (ICCV)*, IEEE, 2017.

[112] T. Hollings, M. Burgman, M. van Andel, M. Gilbert, T. Robinson, and A. Robinson, *How do you find the green sheep? a critical review of the use of remotely sensed imagery to detect and count animals*, *Methods in Ecology and Evolution* **9** (2018), no. 4 881–892.

[113] B. Kellenberger, M. Volpi, and D. Tuia, *Fast animal detection in uav images using convolutional neural networks*, in *2017 IEEE international geoscience and remote sensing symposium (IGARSS)*, pp. 866–869, IEEE, 2017.

[114] Z. Wu, C. Zhang, X. Gu, I. Duporge, L. F. Hughey, J. A. Stabach, A. K. Skidmore, J. G. C. Hopcraft, S. J. Lee, P. M. Atkinson, *et. al.*, *Deep learning enables satellite-based monitoring of large populations of terrestrial mammals across heterogeneous landscape*, *Nature communications* **14** (2023), no. 1 3072.

[115] H. Chen, Z. Qi, and Z. Shi, *Remote sensing image change detection with transformers*, *IEEE Transactions on Geoscience and Remote Sensing* **60** (2021) 1–14.

[116] J. Jakubik et. al., *Prithvi-100M*, Aug., 2023.

[117] X. Xu, Z. Feng, C. Cao, M. Li, J. Wu, Z. Wu, Y. Shang, and S. Ye, *An improved swin transformer-based model for remote sensing object detection and instance segmentation*, *Remote Sensing* **13** (2021), no. 23 4779.

[118] Y. Li, Q. Hou, Z. Zheng, M.-M. Cheng, J. Yang, and X. Li, *Large selective kernel network for remote sensing object detection*, *arXiv preprint arXiv:2303.09030* (2023).

[119] J. Zhu, X. Chen, H. Zhang, Z. Tan, S. Wang, and H. Ma, *Transformer based remote sensing object detection with enhanced multispectral feature extraction*, *IEEE Geoscience and Remote Sensing Letters* (2023).

[120] Q. Chen, J. Wang, C. Han, S. Zhang, Z. Li, X. Chen, J. Chen, X. Wang, S. Han, G. Zhang, *et. al.*, *Group detr v2: Strong object detector with encoder-decoder pretraining*, *arXiv preprint arXiv:2211.03594* (2022).

[121] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, *Deformable detr: Deformable transformers for end-to-end object detection*, *arXiv preprint arXiv:2010.04159* (2020).

[122] D. Meng, X. Chen, Z. Fan, G. Zeng, H. Li, Y. Yuan, L. Sun, and J. Wang, *Conditional detr for fast training convergence*, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3651–3660, 2021.

[123] B. Roh, J. Shin, W. Shin, and S. Kim, *Sparse detr: Efficient end-to-end object detection with learnable sparsity*, *arXiv preprint arXiv:2111.14330* (2021).

[124] S. Ren, K. He, R. Girshick, and J. Sun, *Faster r-cnn: Towards real-time object detection with region proposal networks*, in *Advances in neural information processing systems*, pp. 91–99, 2015.

[125] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, *Yolov4: Optimal speed and accuracy of object detection*, *arXiv preprint arXiv:2004.10934* (2020).

[126] J. Naude and D. Joubert, *The aerial elephant dataset: A new public benchmark for aerial object detection.*, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 48–55, 2019.

[127] B. Kellenberger, D. Tuia, and D. Morris, *Aide: Accelerating image-based ecological surveys with interactive machine learning*, *Methods in Ecology and Evolution* **11** (2020), no. 12 1716–1727.

[128] X. Ding, X. Zhang, J. Han, and G. Ding, *Scaling up your kernels to 31x31: Revisiting large kernel design in cnns*, in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11963–11975, 2022.

[129] N. Park and S. Kim, *How do vision transformers work?*, *arXiv preprint arXiv:2202.06709* (2022).

[130] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z.-H. Jiang, F. E. Tay, J. Feng, and S. Yan, *Tokens-to-token vit: Training vision transformers from scratch on imagenet*, in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 558–567, 2021.

[131] F. W. Campbell and J. G. Robson, *Application of fourier analysis to the visibility of gratings*, *The Journal of physiology* **197** (1968), no. 3 551.

[132] W. Sweldens, *The lifting scheme: A construction of second generation wavelets*, *SIAM journal on mathematical analysis* **29** (1998), no. 2 511–546.

[133] B. Cheng, A. Schwing, and A. Kirillov, *Per-pixel classification is not all you need for semantic segmentation*, *Advances in Neural Information Processing Systems* **34** (2021) 17864–17875.

[134] H. W. Kuhn, *The hungarian method for the assignment problem*, *Naval research logistics quarterly* **2** (1955), no. 1-2 83–97.

[135] I. Loshchilov and F. Hutter, *Decoupled weight decay regularization*, *arXiv preprint arXiv:1711.05101* (2017).

[136] Z. Zong, G. Song, and Y. Liu, *Detrs with collaborative hybrid assignments training*, in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6748–6758, 2023.

[137] G. J. et. al., *ultralytics/yolov5: v3.1 - Bug Fixes and Performance Improvements*, Oct., 2020.

[138] G. Jocher, A. Chaurasia, and J. Qiu, *Ultralytics YOLO*, Jan., 2023.

[139] Y. Yu, X. Sun, and Q. Cheng, *Expert teacher based on foundation image segmentation model for object detection in aerial images*, *Scientific Reports* **13** (2023), no. 1 21964.

[140] B. G. Weinstein, L. Garner, V. R. Saccomanno, A. Steinkraus, A. Ortega, K. Brush, G. Yenni, A. E. McKellar, R. Converse, C. D. Lippitt, *et. al.*, *A general deep learning model for bird detection in high-resolution airborne imagery*, *Ecological Applications* **32** (2022), no. 8 e2694.

[141] M. Roser, H. Ritchie, E. Ortiz-Ospina, and J. Hasell, *Coronavirus pandemic (covid-19)*, *Our World in Data* (2020).

[142] W. Wang, A. C. den Brinker, S. Stuijk, and G. de Haan, *Algorithmic principles of remote ppg*, *IEEE Transactions on Biomedical Engineering* **64** (2016), no. 7 1479–1491.

[143] W. Chen and D. McDuff, *Deepphys: Video-based physiological measurement using convolutional attention networks*, in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 349–365, 2018.

[144] Z. Yu, X. Li, and G. Zhao, *Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks*, in *Proc. BMVC*, 2019.

[145] D. McDuff, S. Gontarek, and R. Picard, *Remote measurement of cognitive stress via heart rate variability*, in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 2957–2960, IEEE, 2014.

[146] F. Bousefsaf, C. Maaoui, and A. Pruski, *Remote assessment of the heart rate variability to detect mental stress*, in *2013 7th International Conference on Pervasive Computing Technologies for Healthcare and Workshops*, pp. 348–351, IEEE, 2013.

[147] P. J. Silvia, *Rz interval as an impedance cardiography measure of effort-related cardiac sympathetic activity paul j. silvia, ashley n. mchone, zuzana mironovová, kari m. eddington, kelly l. harper university of north carolina at greensboro sarah h. sperry, thomas r. kwapil*, .

[148] H. Riese, P. F. Groot, M. van den Berg, N. H. Kupper, E. H. Magnee, E. J. Rohaan, T. G. Vrijkotte, G. Willemsen, and E. J. de Geus, *Large-scale ensemble averaging of ambulatory impedance cardiograms*, *Behavior Research Methods, Instruments, & Computers* **35** (2003), no. 3 467–477.

[149] G. H. Willemsen, E. J. DeGeus, C. H. Klaver, L. J. VanDoornen, and D. Carrofl, *Ambulatory monitoring of the impedance cardiogram*, *Psychophysiology* **33** (1996), no. 2 184–193.

[150] M. D. Seery, C. L. Kondrak, L. Streamer, T. Saltsman, and V. M. Lamarche, *Preejection period can be calculated using r peak instead of q*, *Psychophysiology* **53** (2016), no. 8 1232–1240.

[151] M. A. Van Eijnatten, M. J. Van Rijssel, R. J. Peters, R. M. Verdaasdonk, and J. H. Meijer, *Comparison of cardiac time intervals between echocardiography and impedance cardiography at various heart rates*, *Journal of electrical bioimpedance* **5** (2014), no. 1 2–8.

[152] D. L. Lozano, G. Norman, D. Knox, B. L. Wood, B. D. Miller, C. F. Emery, and G. G. Berntson, *Where to b in dz/dt*, *Psychophysiology* **44** (2007), no. 1 113–119.

[153] J. H. Meijer, S. Boesveldt, E. Elbertse, and H. Berendse, *Method to measure autonomic control of cardiac function using time interval parameters from impedance cardiography*, *Physiological measurement* **29** (2008), no. 6 S383.

[154] S. W. Wilde, D. S. Miles, R. J. Durbin, M. N. Sawka, A. G. Suryaprasad, R. W. Gotshall, and R. M. Glaser, *Evaluation of myocardial performance during wheelchair ergometer exercise.*, *American journal of physical medicine* **60** (1981), no. 6 277–291.

[155] R. van Lien, N. M. Schutte, J. H. Meijer, and E. J. de Geus, *Estimated preejection period (pep) based on the detection of the r-wave and dz/dt-min peaks does not adequately reflect the actual pep across a wide range of laboratory and ambulatory conditions*, *International Journal of Psychophysiology* **87** (2013), no. 1 60–69.

[156] D. B. Newlin and R. W. Levenson, *Pre-ejection period: Measuring beta-adrenergic influences upon the heart*, *Psychophysiology* **16** (1979), no. 6 546–552.

[157] G.-S. Hsu, A. Ambikapathi, and M.-S. Chen, *Deep learning with time-frequency representation for pulse estimation from facial videos*, in *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pp. 383–389, IEEE, 2017.

[158] X. Niu, H. Han, S. Shan, and X. Chen, *Synrhythm: Learning a deep heart rate estimator from general to specific*, in *2018 24th International Conference on Pattern Recognition (ICPR)*, pp. 3580–3585, IEEE, 2018.

[159] J. Gunther and N. E. Ruben, *Remote heart rate estimation*, Dec. 26, 2017. US Patent 9,852,507.

[160] R. M. Kelsey, *Beta-adrenergic cardiovascular reactivity and adaptation to stress: The cardiac pre-ejection period as an index of effort.*, .

[161] K. P. Prasad and D. B. Anuradha, *Detection of abnormalities in fetal electrocardiogram*, *International Journal of Applied Engineering Research* **12** (2017), no. 1 2017.

[162] M. Forouzanfar, F. C. Baker, I. M. Colrain, A. Goldstone, and M. de Zambotti, *Automatic analysis of pre-ejection period during sleep using impedance cardiogram*, *Psychophysiology* **56** (2019), no. 7 e13355.

[163] J. B. Hinnant, L. Elmore-Staton, and M. El-Sheikh, *Developmental trajectories of respiratory sinus arrhythmia and preejection period in middle childhood*, *Developmental psychobiology* **53** (2011), no. 1 59–68.

[164] S. L. Brenner and T. P. Beauchaine, *Pre-ejection period reactivity and psychiatric comorbidity prospectively predict substance use initiation among middle-schoolers: A pilot study*, *Psychophysiology* **48** (2011), no. 11 1588–1596.

[165] C. P. Bara, M. Papakostas, and R. Mihalcea, *A deep learning approach towards multimodal stress detection*, in *Proceedings of the AAAI-20 Workshop on Affective Content Analysis, New York, USA, AAAI*, 2020.

[166] Z. Yu, X. Li, X. Niu, J. Shi, and G. Zhao, *Autohr: A strong end-to-end baseline for remote heart rate measurement with neural searching*, *arXiv preprint arXiv:2004.12292* (2020).

[167] X. Li, J. Chen, G. Zhao, and M. Pietikainen, *Remote heart rate measurement from face videos under realistic situations*, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4264–4271, 2014.

[168] H. Monkaresi, N. Bosch, R. A. Calvo, and S. K. D'Mello, *Automated detection of engagement using video-based estimation of facial expressions and heart rate*, *IEEE Transactions on Affective Computing* **8** (2016), no. 1 15–28.

[169] T. Blöcher, J. Schneider, M. Schinle, and W. Stork, *An online ppgi approach for camera based heart rate monitoring using beat-to-beat detection*, in *2017 IEEE Sensors Applications Symposium (SAS)*, pp. 1–6, IEEE, 2017.

[170] F. J. Sanchez-Marin, S. Calixto-Carrera, and C. Villaseñor-Mora, *Novel approach to assess the emissivity of the human skin*, *Journal of Biomedical Optics* **14** (2009), no. 2 024006.

[171] K. Ammer, *Temperature effects of thermotherapy determined by infrared measurements*, *Physica medica* **20** (2004) 64–66.

[172] J. D. Hardy *et. al.*, *The radiation of heat from the human body: Iii. the human skin as a black-body radiator*, *The Journal of clinical investigation* **13** (1934), no. 4 615–620.

[173] O. Ulutan, S. Rallapalli, M. Srivatsa, C. Torres, and B. Manjunath, *Actor conditioned attention maps for video action detection*, in *The IEEE Winter Conference on Applications of Computer Vision*, pp. 527–536, 2020.

[174] L. Wang, Y. Qiao, and X. Tang, *Video action detection with relational dynamic-poselets*, in *European conference on computer vision*, pp. 565–580, Springer, 2014.

[175] H. Zhang, M. Gong, P. Zhang, L. Su, and J. Shi, *Feature-level change detection using deep representation and feature change analysis for multispectral imagery*, *IEEE Geoscience and Remote Sensing Letters* **13** (2016), no. 11 1666–1670.

[176] H. Teffahi, H. Yao, S. Chaib, and N. Belabid, *A novel spectral-spatial classification technique for multispectral images using extended multi-attribute profiles and sparse autoencoder*, *Remote Sensing Letters* **10** (2019), no. 1 30–38.

[177] M. Sundermeyer, R. Schlüter, and H. Ney, *Lstm neural networks for language modeling*, in *Thirteenth annual conference of the international speech communication association*, 2012.

[178] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, *Lstm: A search space odyssey*, *IEEE transactions on neural networks and learning systems* **28** (2016), no. 10 2222–2232.

[179] M. Schuster and K. K. Paliwal, *Bidirectional recurrent neural networks*, *IEEE transactions on Signal Processing* **45** (1997), no. 11 2673–2681.

[180] N. Ruiz, E. Chong, and J. M. Rehg, *Fine-grained head pose estimation without keypoints*, in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 2074–2083, 2018.

[181] X. Zhang, P. Bachmann, T. M. Schilling, E. Naumann, H. Schächinger, and M. F. Larra, *Emotional stress regulation: The role of relative frontal alpha asymmetry in shaping the stress response*, *Biological psychology* **138** (2018) 231–239.

[182] P. Bachmann, X. Zhang, M. F. Larra, D. Rebeck, K. Schönbein, K. P. Koch, and H. Schächinger, *Validation of an automated bilateral feet cold pressor test*, *International Journal of Psychophysiology* **124** (2018) 62–70.

[183] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, *A multimodal database for affect recognition and implicit tagging*, *IEEE transactions on affective computing* **3** (2011), no. 1 42–55.

[184] M. Jaiswal, C.-P. Bara, Y. Luo, M. Burzo, R. Mihalcea, and E. M. Provost, *Muse: a multimodal dataset of stressed emotion*, in *Proceedings of The 12th Language Resources and Evaluation Conference*, pp. 1499–1510, 2020.

[185] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, *The pascal visual object classes (voc) challenge*, *International journal of computer vision* **88** (2010), no. 2 303–338.

[186] J. Carreira and A. Zisserman, *Quo vadis, action recognition? a new model and the kinetics dataset*, in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6299–6308, 2017.

[187] V. Engert, A. Merla, J. A. Grant, D. Cardone, A. Tusche, and T. Singer, *Exploring the use of thermal infrared imaging in human stress research*, *PloS one* **9** (2014), no. 3 e90782.

[188] H. Genno, K. Ishikawa, O. Kanbara, M. Kikumoto, Y. Fujiwara, R. Suzuki, and M. Osumi, *Using facial skin temperature to objectively evaluate sensations*, *International Journal of Industrial Ergonomics* **19** (1997), no. 2 161–171.

[189] H. Veltman and W. Vos, *Facial temperature as a measure of operator state*, *Foundations of augmented cognition* **293** (2005).

[190] C. Spellenberg, P. Heusser, A. Büssing, A. Savelsbergh, and D. Cysarz, *Binary symbolic dynamics analysis to detect stress-associated changes of nonstationary heart rate variability*, *Scientific Reports* **10** (2020), no. 1 1–10.

[191] U. Pluntke, S. Gerke, A. Sridhar, J. Weiss, and B. Michel, *Evaluation and classification of physical and psychological stress in firefighters using heart rate variability*, in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 2207–2212, IEEE, 2019.

[192] B. Farnsworth, "Heart rate variability – how to analyze ecg data." `https://imotions.com/blog/heart-rate-variability/`, 2019.

[193] K. Simonyan and A. Zisserman, *Very deep convolutional networks for large-scale image recognition*, *arXiv preprint arXiv:1409.1556* (2014).

[194] M. Agrawal, *Effects of air pollution on agriculture: an issue of national concern*, *Natl Acad Sci Lett* **28** (2005), no. 3/4 93–106.

[195] M. H. Unsworth and D. Ormrod, *Effects of gaseous air pollution in agriculture and horticulture*. No. 32. Butterworth-Heinemann, 2013.

[196] R. Gu, *Methane gas emission detection using deep learning and hyperspectral imagery*, in *2021 IEEE 3rd International Conference on Frontiers Technology of Information and Computer (ICFTIC)*, pp. 36–44, IEEE, 2021.

[197] L. Hamlin, R. O. Green, P. Mouroulis, M. Eastwood, D. Wilson, M. Dudik, and C. Paine, *Imaging spectrometer science measurements for terrestrial ecology: Aviris and new developments*, in *2011 Aerospace Conference*, pp. 1–7, 2011.

[198] J. Theiler, B. R. Foy, and A. M. Fraser, *Beyond the adaptive matched filter: nonlinear detectors for weak signals in high-dimensional clutter*, in *Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XIII*, vol. 6565, p. 656503, International Society for Optics and Photonics, 2007.

[199] N. Pettorelli, *The normalized difference vegetation index*. Oxford University Press, 2013.

[200] F. Kriegler, W. Malila, R. Nalepka, and W. Richardson, *Preprocessing transformations and their effects on multispectral recognition*, *Remote sensing of environment, VI* (1969) 97.