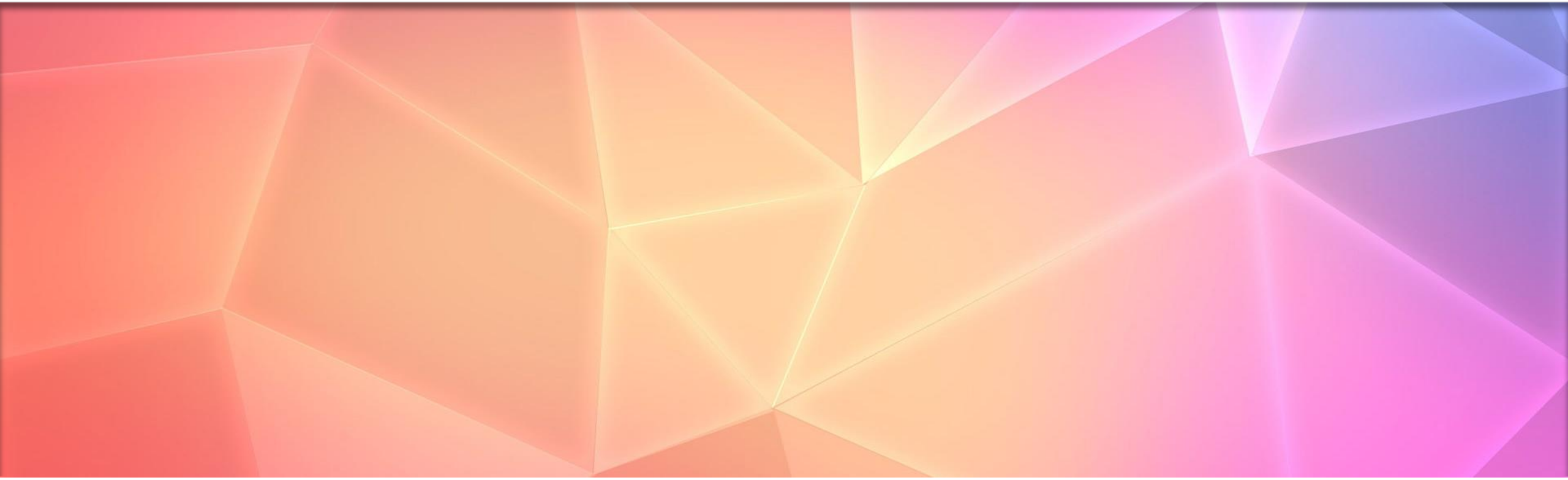
Universität St.Gallen



Deep Learning (Fall 2023)

# Proposal: VQA in Remote Sensing

ViTalVisions by Damian Falk, Kaan Aydin

# Our Team

## Damian Falk

## Kaan Aydin

Universität St.Gallen

# Our proposal

- Introduction
- Current Research
- Our Approach
- Appendix
  - Key Papers

# Introduction

## Problem Statement

Advancements in sensing technology have led to a **large volume of visual info** being acquired and processed in the Remote Sensing (RS) domain

RS data has **immense potential for various tasks** (e.g., land cover classification, object detection) which could benefit numerous applications

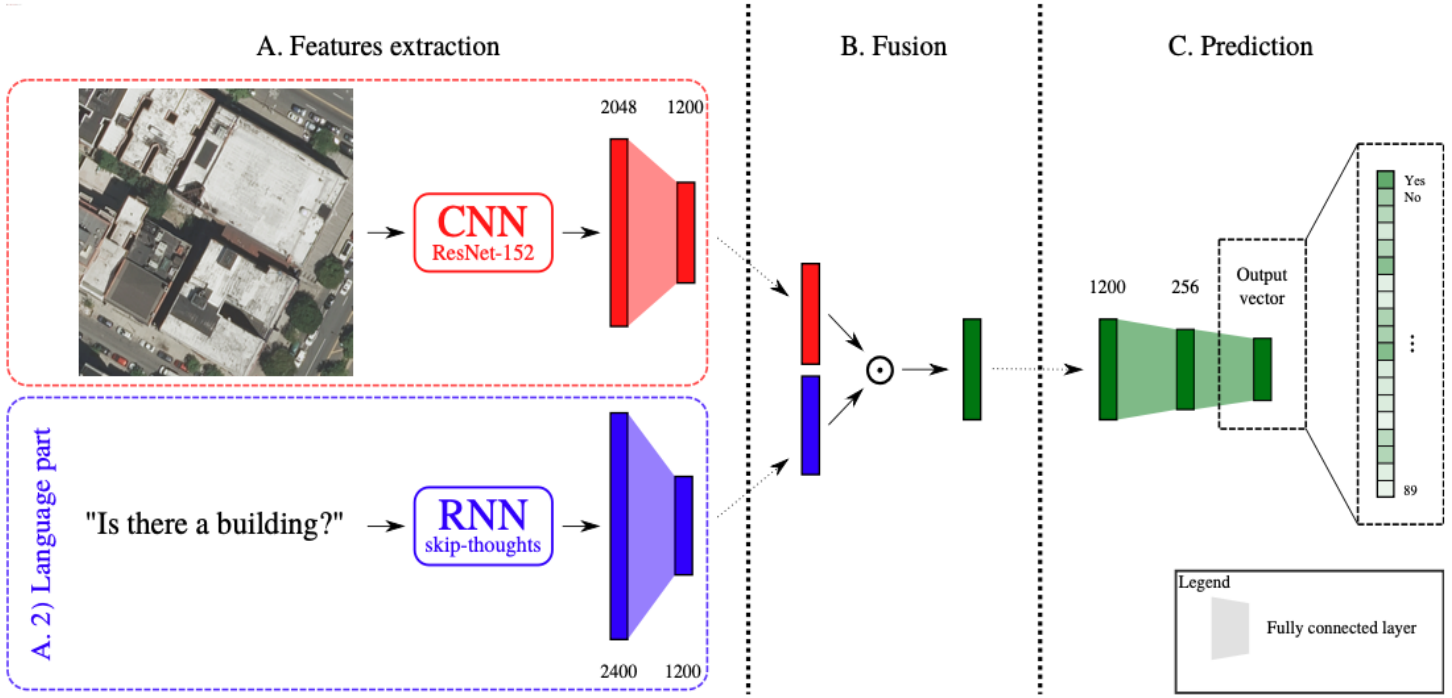
Current methodologies are **largely static and task-specific** and require specialized expert knowledge – hindering generic and easy access to the information

## Proposed Solution: RSVQA

- Remote Sensing Visual Question Answering System (RSVQA) that allows for **info extraction through queries formulated in natural language**
- VQA models enable **answering of free-form and open-ended questions about RS** - providing a more intuitive solution to data extraction / analysis
- This approach can potentially **address classical problems** such as:
  - Specific object detection (e.g., "Is there a thatched roof in this image?")
  - More complex tasks involving relations between different objects (e.g., "Is there a building on the right of the river?")

# Current Research (1/3) – First Approaches

## Framework of the VQA model (Lobry, Marcos, Murray & Tuia, 2020)




## Description

### A. Feature Extraction

- **Visual:** ResNet-152 model, pretrained on ImageNet
- **Language:** Skip-thoughts, pre-trained on the BookCorpus dataset
- Fully-connected layers to reduce dimensions to 1200

### B. Fusion

- Point-wise multiplication after tanh function to each element
- Fixed operation – e2e training ensures comparability of features

### C. Prediction

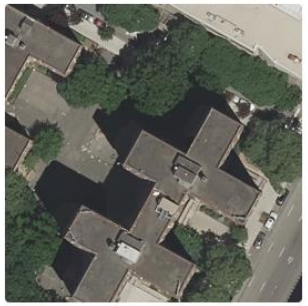
- Projection of 1200 dimensional vector to answer space by using MLP
- Formulation of problem as a classification task

Lobry, S., Demir, B., & Tuia, D. (2021). RSVQA Meets Bigearthnet: A New, Large-Scale, Visual Question Answering Dataset for Remote Sensing. In 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS (pp. 1218-1221).

# Current Research (2/3) - Dataset

## High Resolution Dataset

- Based on USGS' high resolution orthorectified images
- Questions and answers derived from OSM



| Question | Ground Truth |
|---|---|
| Is there a commercial building? | no |
| What is the area covered by commercial buildings? | 0m2 |
| Is a road present? | yes |
| What is the amount of residential areas? | 1 |
| What is the area covered by medium residential buildings? | 0m2 |

**10569** images

**955664** questions

## Low Resolution Dataset

- Based on Sentinel-2 images
- Questions and answers derived from OSM.

| Question | Ground Truth |
|---|---|
| Is it a rural or an urban area | rural |
| Is there a commercial building? | no |
| Is a building present in the image? | yes |
| What is the number of circular roads? | 0 |

**772** images

**77232** questions

Lobry, S., Demir, B., & Tuia, D. (2021). RSVQA Meets Bigearthnet: A New, Large-Scale, Visual Question Answering Dataset for Remote Sensing. In 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS (pp. 1218-1221).

# Current Research (3/3) – Selected Past Results

## Results from two papers

**Based on the HR dataset**

| Class | RSVQA | CrossAtt | InfoMax | InfoMax + CrossAtt |
|---|---|---|---|---|
| Count | 52.34 | 55.56 | 55.34 | 56.22 |
| Presence | 92.54 | 93.48 | 93.16 | 94.12 |
| Comparison | 75.34 | 88.51 | 88.23 | 88.29 |
| Area | 85.32 | 82.61 | 83.96 | 84.12 |
| OA | 78.23 | 78.48 | 78.5 | 79.11 |
| AA | 76.38 | 80.04 | 80.16 | 80.63 |

| Class | RSVQA | CrossAtt | InfoMax | InfoMax + CrossAtt |
|---|---|---|---|---|
| Count | 48.44 | 50.39 | 49.92 | 51.22 |
| Presence | 90.04 | 88.75 | 90.46 | 91.34 |
| Comparison | 76.33 | 77.23 | 81.91 | 80.21 |
| Area | 65.81 | 65.77 | 71.96 | 70.32 |
| OA | 72.45 | 70.56 | 72.95 | 73.87 |
| AA | 70.15 | 70.54 | 73.27 | 73.56 |

**Based on the LR dataset**

| Class | RSVQA | CrossAtt | InfoMax | InfoMax + CrossAtt |
|---|---|---|---|---|
| Count | 65.38 | 72.21 | 70.56 | 73.12 |
| Presence | 84.78 | 91.67 | 90.12 | 92.86 |
| Comparison | 78.44 | 93.08 | 88.95 | 93.01 |
| Rural/Urban | 89.67 | 85 | 85.12 | 88.12 |
| OA | 79.56 | 85.46 | 84.13 | 85.98 |
| AA | 78.34 | 85.49 | 83.68 | 86.77 |

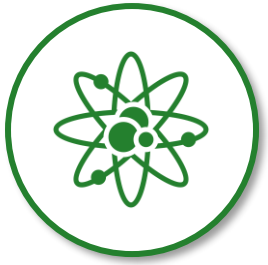
- Another paper (Songara et. al, 2023) implemented a similar approach, but used CrossAtt and InfoMax for fusing
- By using the same dataset, they were able to increase the performance on multiple classes
- Counting task is still considered “difficult”

Songara, J., Pande, S., Choudhury, S., Banerjee, B., & Velmurugan, R. (2023). Visual Question Answering in Remote Sensing with Cross-Attention and Multimodal Information Bottleneck. arXiv preprint arXiv:2306.14264.

# Our Approach (1/3) – Our ideas

## Employing alternative vision feature extraction methods

- Proposed approaches score low primarily in counting tasks
- Also, to our knowledge, no paper has implemented a vision transformer or SSL methods for improved feature extraction



## Exploration of diverse Data Fusion Approaches

- There are different approaches to fuse modality representations, e.g.,
  - Element-wise multiplication
  - Cross-Attention (with Information Maximization)
  - Multimodal Compact Bi-Linear Pooling (MCB)
  - ….

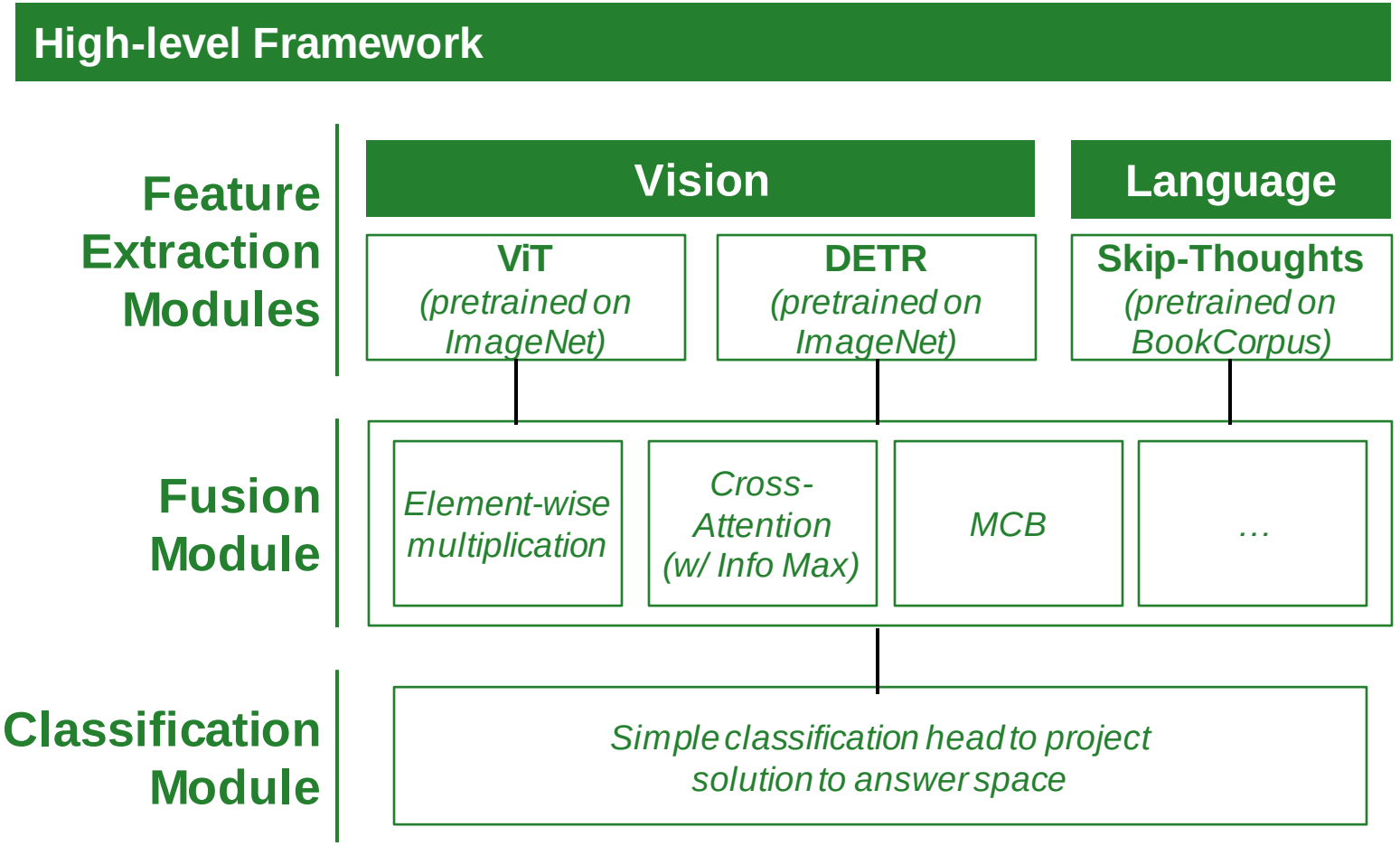
Focus of our project based on time & data availability

## VQA Leveraging Time-Series / Video Features

- To our knowledge, no research has combined time-series data with VQA system for RS
- We believe that this would enable a new set of questions, e.g.,
  - How has deforestation changed in the last 5 years?
  - How many buildings have been built in the last 3 years?

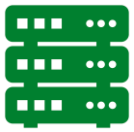
Out of scope as no dataset available –possible MA thesis?

# Our Approach (2/3) – High-level Suggestion




- Combined use of ViT and DETR to enable better vision feature extraction
- Use of SSL to improve vision feature extraction
- Use of different fusion techniques to enable improved interaction between modalities
- No (or limited) changes to the language feature extraction module and classification to ensure comparability

# Our Approach (3/3) – Risks



## Computational limitations

ViTs are rather compute intensive compared to small CNNs – We except to be able to train locally / via Colab Pro

## Performance Concerns

We might not outperform standard scores / pre-training might not lead to better downstream performance

## Limited Experience

of the team with multi-model models and multi-modal training setups

## Overhead

More data (than what is available) might be required for a transformer to outperform

## Risk of Obsolescence / Competitors

LLMs might be soon to be able to this for all images / other researchers might employ a similar approach

We identified several moderate risks, but no severe risks for this project

# Appendix

Universität St.Gallen

# Key Papers (1/2)


- Chappuis, C., Lobry, S., Kellenberger, B., Le Saux, B., & Tuia, D. (2021). How to find a good image-text embedding for remote sensing visual question answering? arXiv preprint arXiv:2109.11848.
- Hackel, L., Clasen, K. N., Ravanbakhsh, M., & Demir, B. (2023). LiT-4-RSVQA: Lightweight Transformer-based Visual Question Answering in Remote Sensing. arXiv preprint arXiv:2306.00758.
- Lobry, S., Demir, B., & Tuia, D. (2021). RSVQA Meets Bigearthnet: A New, Large-Scale, Visual Question Answering Dataset for Remote Sensing. In 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS (pp. 1218-1221).
- Lobry, S., Marcos, D., Murray, J., & Tuia, D. (2020). RSVQA: Visual Question Answering for Remote Sensing Data. IEEE Transactions on Geoscience and Remote Sensing, 58(12), 8555-8566.
- Siebert, T., Clasen, K. N., Ravanbakhsh, M., & Demir, B. (2022). Multi-Modal Fusion Transformer for Visual Question Answering in Remote Sensing. arXiv preprint arXiv:2210.04510.
- Songara, J., Pande, S., Choudhury, S., Banerjee, B., & Velmurugan, R. (2023). Visual Question Answering in Remote Sensing with Cross-Attention and Multimodal Information Bottleneck. arXiv preprint arXiv:2306.14264.

Universität St.Gallen

# Key Papers (2/2)


- Srivastava, Y., Murali, V., Dubey, S. R., & Mukherjee, S. (2021). Visual question answering using deep learning: A survey and performance analysis. In Computer Vision and Image Processing: 5th International Conference, CVIP 2020, Prayagraj, India, December 4-6, 2020, Revised Selected Papers, Part II 5 (pp. 75-86). Springer Singapore.
- Yuan, Z., Mou, L., Wang, Q., & Zhu, X. X. (2022). From easy to hard: Learning language-guided curriculum for visual question answering on remote sensing data. IEEE Transactions on Geoscience and Remote Sensing, 60, 1-11.
- Yuan, Z., Mou, L., & Zhu, X. X. (2023). Overcoming Language Bias in Remote Sensing Visual Question Answering via Adversarial Training. arXiv preprint arXiv:2306.00483.
- Yuan, Z., Mou, L., & Zhu, X. X. (2023, May). Multilingual augmentation for robust visual question answering in remote sensing images. In 2023 Joint Urban Remote Sensing Event (JURSE) (pp. 1-4). IEEE.
- Zhang, Z., Jiao, L., Li, L., Liu, X., Chen, P., Liu, F., ... & Guo, Z. (2023). A spatial hierarchical reasoning network for remote sensing visual question answering. IEEE Transactions on Geoscience and Remote Sensing, 61, 1-15.


# Questions?