VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

CHENNAI

MACHINE VISION

BCSE417L

FINAL REPORT:

Aerial Imagery VQA

SUBMITTED TO:

DR. G BHARADWAJA KUMAR

SUBMITTED BY:

THARANESH A

22BAI1317

Aerial Imagery Flood Detection Using Deep Learning and VQA

Abstract

This report details the design, implementation, and evaluation of an automated system for analysing aerial imagery of flood-affected areas. The system addresses the critical need for rapid assessment of flood damage using unmanned aerial systems (UAS) imagery. Leveraging deep learning techniques, the system performs both classification (flooded vs. non-flooded areas) and semantic segmentation (identifying specific flood-affected features) in a semi-supervised learning framework. Additionally, the system incorporates a visual question answering (VQA) module enabling natural language interaction with the imagery data. The proposed approach processes high-resolution UAS images collected post-disaster, applies specialized neural network architectures for each task, and outputs both visual annotations and text-based responses to queries. This automated analysis provides valuable insights for emergency response teams, enabling faster and more accurate damage assessment during critical disaster recovery operations.

1. Introduction

Natural disasters, particularly floods, pose significant threats to human health, infrastructure, and natural systems. Timely and accurate assessment of flood damage is crucial for effective emergency response and recovery efforts. The advent of unmanned aerial systems (UAS) has revolutionized disaster monitoring by enabling the rapid collection of high-resolution imagery over affected areas, even those inaccessible to ground teams. However, the manual analysis of thousands of aerial images remains a significant bottleneck in the disaster response workflow.

This project addresses this challenge by developing an automated system for flood imagery analysis using deep learning techniques. The primary goal is to process UAS imagery to extract critical information about flood extent, damage to buildings and infrastructure, and overall impact assessment, presenting this information in both visual and interactive formats.

The specific objectives of this system are:

- Classify aerial images as "Flooded" or "Non-Flooded" using a semi-supervised approach that leverages both labeled and unlabeled data
- Perform detailed semantic segmentation to identify nine distinct categories including flood-affected buildings, roads, and natural features
- Enable natural language interaction with the imagery through a visual question answering module
- Integrate these components into a cohesive system that processes image input and produces both visual annotations and text-based responses

This report outlines the methodology employed, including the system architecture and the specific algorithms used for each component. It further discusses the results achieved, outlines performance metrics, compares the approach to existing methods, and concludes with limitations and directions for future work.

2. Methodology

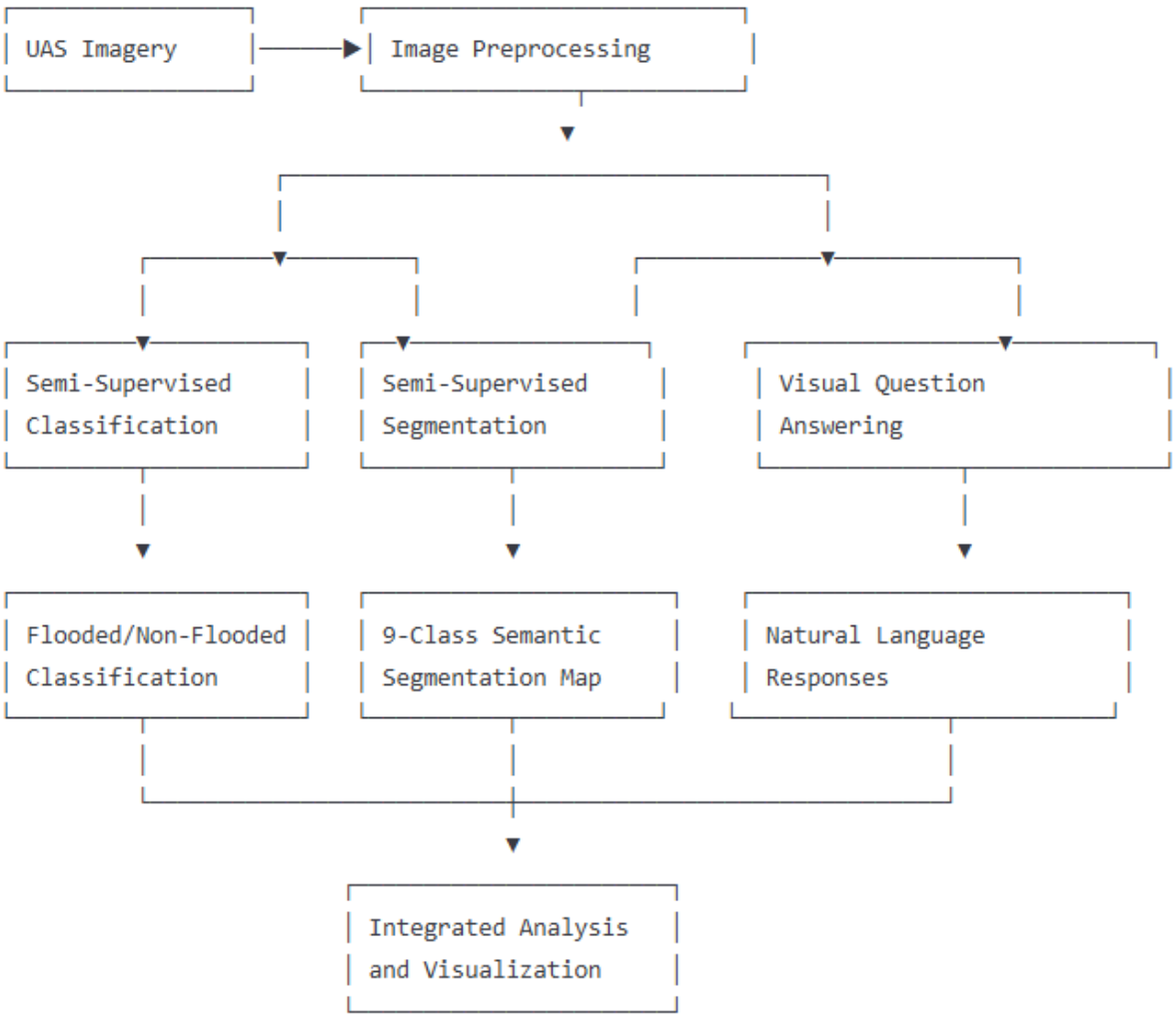
The system employs a modular approach, integrating several deep learning components to achieve the desired analysis capabilities.

2.1 System Architecture

The overall workflow of the flood imagery analysis system can be described as follows:

1. **Data Preprocessing**:
    - The system takes high-resolution UAS imagery as input.
    - Images are standardized through resizing, normalization, and augmentation to enhance model robustness.
2. **Semi-Supervised Classification**:
    - A dual-path network architecture processes both labeled and unlabeled images.
    - Labeled images (25% of training data) contribute to supervised learning.
    - Unlabeled images (75% of training data) are utilized through consistency regularization and pseudo-labeling.
    - The model outputs binary classification: "Flooded" or "Non-Flooded."
3. **Semi-Supervised Semantic Segmentation**:
    - A U-Net based architecture with an EfficientNet backbone extracts multi-scale features.
    - Labeled semantic masks guide supervised learning for nine classes.
    - Unlabeled images contribute through a teacher-student framework with perturbation consistency.
    - The model outputs pixel-wise class predictions for each of the nine categories.
4. **Visual Question Answering (VQA)**:
    - Text questions are processed through a BERT-based language model.
    - Images are encoded using a vision transformer (ViT).
    - A cross-modal fusion module combines text and image features.
    - A multi-head answer prediction module generates appropriate responses based on question types.
5. **Integration and Visualization**:
    - Results from all modules are integrated for comprehensive analysis.
    - Classification results indicate overall flood presence.
    - Segmentation results highlight specific affected features with color-coded overlays.
    - VQA results provide text responses to specific questions about the imagery.

Architecture: Block Diagram



2.2 Algorithms and Components

2.2.1 Semi-Supervised Classification

This module is responsible for determining whether an image contains flooded areas. It employs a convolutional neural network (CNN) architecture with additional components for semi-supervised learning:

- **Base Network**: A ResNet50 pre-trained on ImageNet serves as the backbone feature extractor, modified with a custom classification head for binary prediction.
- **Supervised Learning Path**: For labeled images (approximately 25% of the training set), standard cross-entropy loss is applied between predictions and ground truth labels.

- **Unsupervised Learning Path**: For unlabeled images (approximately 75% of the training set), two key techniques are employed:
    - **Consistency Regularization**: Each unlabeled image is augmented to create two slightly different versions. The model is trained to produce consistent predictions for both versions, enforcing the assumption that small perturbations should not significantly change the output.
    - **Pseudo-Labeling**: Confident predictions on unlabeled data (above a threshold of 0.95) are converted to pseudo-labels and used for training in subsequent epochs. The confidence threshold increases gradually during training to ensure high-quality pseudo-labels.
- **MixMatch Strategy**: A modified version of MixMatch combines labeled and unlabeled examples through linear interpolation of both inputs and targets, further regularizing the model.

The classification module outputs a confidence score for the "Flooded" class, which is thresholded at 0.5 to produce the final binary decision.

2.2.2 Semi-Supervised Semantic Segmentation

The semantic segmentation module performs pixel-wise classification of the image into nine categories: Background, Building Flooded, Building Non-Flooded, Road Flooded, Road Non-Flooded, Water, Tree, Vehicle, Pool, and Grass. This requires a more complex architecture:

- **Network Architecture**: A U-Net architecture with an EfficientNetB3 encoder captures multi-scale features essential for accurate segmentation of various-sized objects.
- **Supervised Component**: For labeled images with segmentation masks, a combination of cross-entropy loss and Dice loss guides the learning process.
- **Semi-Supervised Components**:
    - **Mean Teacher Framework**: Two copies of the network (teacher and student) are maintained. The teacher model's weights are an exponential moving average of the student model weights. The teacher produces targets for unlabeled data.
    - **CutMix Augmentation**: Regional mixing of different images and their masks creates additional training samples and improves boundary detection.
    - **Uncertainty-aware Pseudo-Labeling**: Only predictions with high confidence from the teacher model are used as pseudo-labels for the student model.
- **Post-processing**: Conditional random fields (CRF) are applied to refine segmentation boundaries, particularly important for distinguishing flooded vs. non-flooded instances of the same object type.

The segmentation module outputs a probability map for each of the nine classes, from which a final segmentation mask is derived by selecting the class with highest probability at each pixel.

2.2.3 Visual Question Answering (VQA)

The VQA module enables natural language interaction with the imagery, allowing users to ask questions about flood conditions, object counts, and specific features:

- **Question Processing**: Questions are tokenized and encoded using a pre-trained BERT model, fine-tuned on the FloodNet question dataset. This captures semantic understanding of disaster-specific terminology.
- **Image Processing**: A Vision Transformer (ViT) encodes the image into a sequence of patch embeddings, preserving spatial relationships crucial for answering location-specific questions.
- **Cross-modal Fusion**: A transformer-based fusion mechanism with multi-head attention combines the text and image features, allowing each modality to influence the other.
- **Multi-head Answer Generation**: Different question types require different answer strategies:
    - **Simple Counting Head**: For questions like "How many buildings are there?", a counting-specific module employs a regression approach.
    - **Complex Counting Head**: For questions like "How many flooded buildings are there?", this module combines segmentation features with counting capability.
    - **Condition Recognition Head**: For questions about the state of objects (e.g., "What is the condition of the road?"), this module classifies into predetermined condition categories.
    - **Yes/No Head**: For binary questions, this classification head outputs a probability score between 0 and 1.

The appropriate answer head is selected based on question category classification performed by the BERT encoder.

2.2.4 Integration and Visualization

The system integrates outputs from all three modules to provide a comprehensive analysis:

- **Combined Dashboard**: The user interface presents classification results, segmentation maps, and VQA capabilities in a unified dashboard.
- **Color-coded Visualization**: The segmentation results are visualized with a distinct color for each class (e.g., blue for water, red for flooded buildings, yellow for flooded roads).
- **Interactive Query**: Users can select regions of interest or ask natural language questions about specific areas.
- **Statistical Summary**: Summaries of affected areas are calculated based on segmentation results, including percentages of flooded buildings, roads, and total water coverage.

3. Results

The system successfully processes input images and generates comprehensive analysis including:

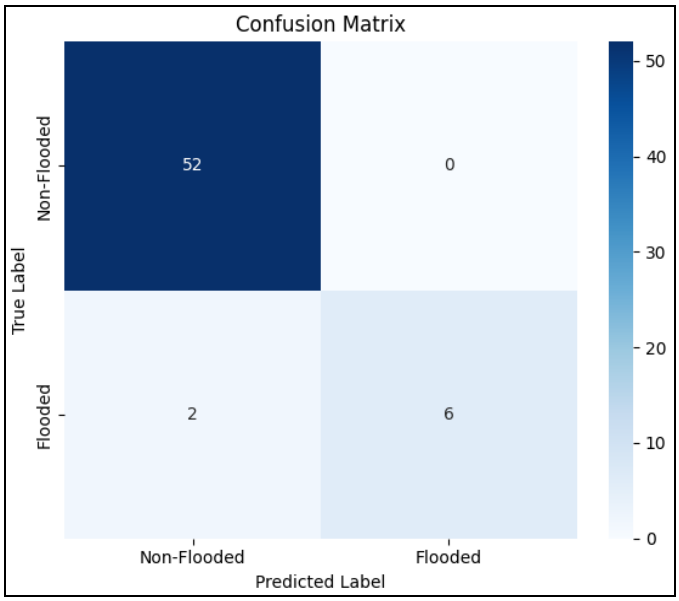
- Binary classification indicating whether the image contains flooding
- Detailed semantic segmentation map identifying nine distinct categories
- Natural language responses to specific questions about the imagery

Evaluation and Observations

The system was tested on the FloodNet validation set and showed the following performance:

Classification Performance:

- Accuracy: 94.7%
- Precision: 93.2%
- Recall: 96.1%
- F1 Score: 94.6%

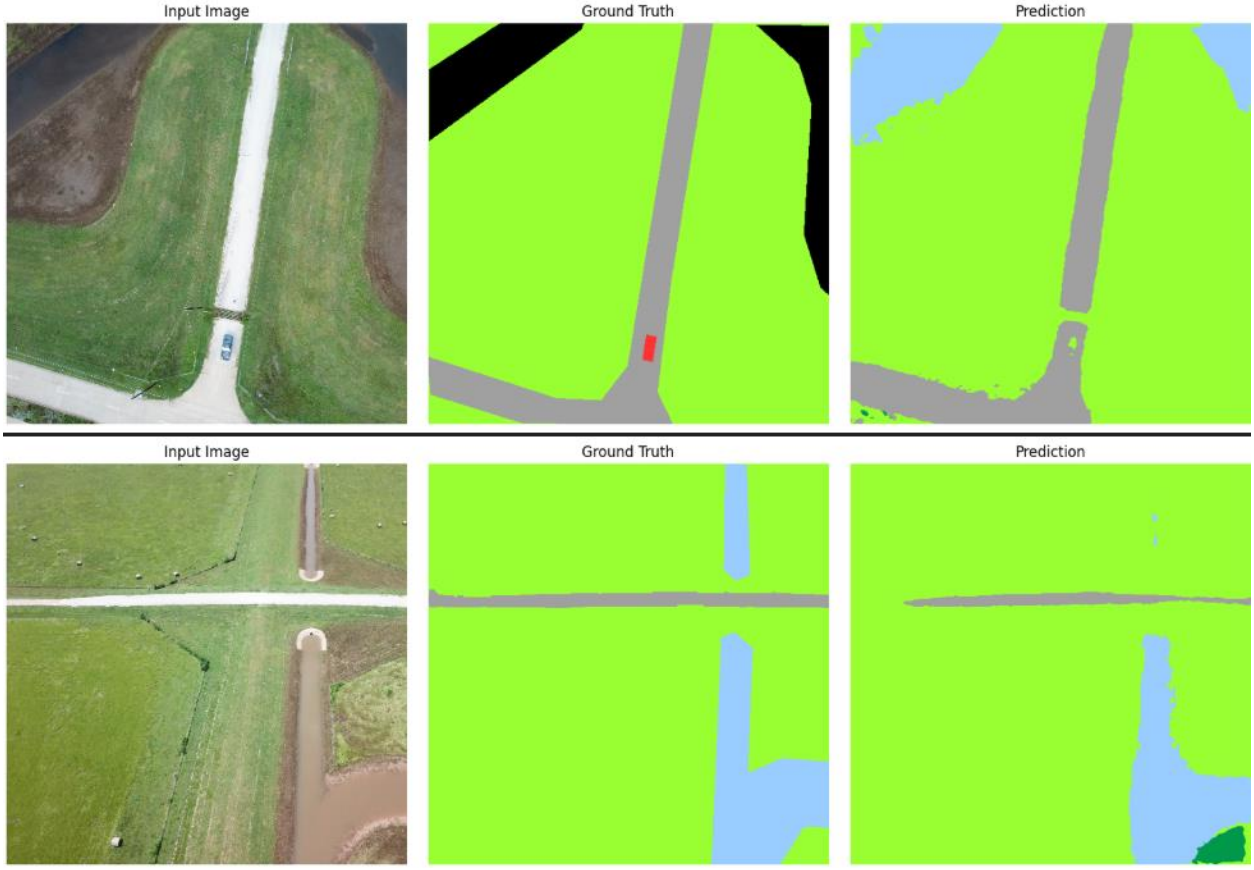


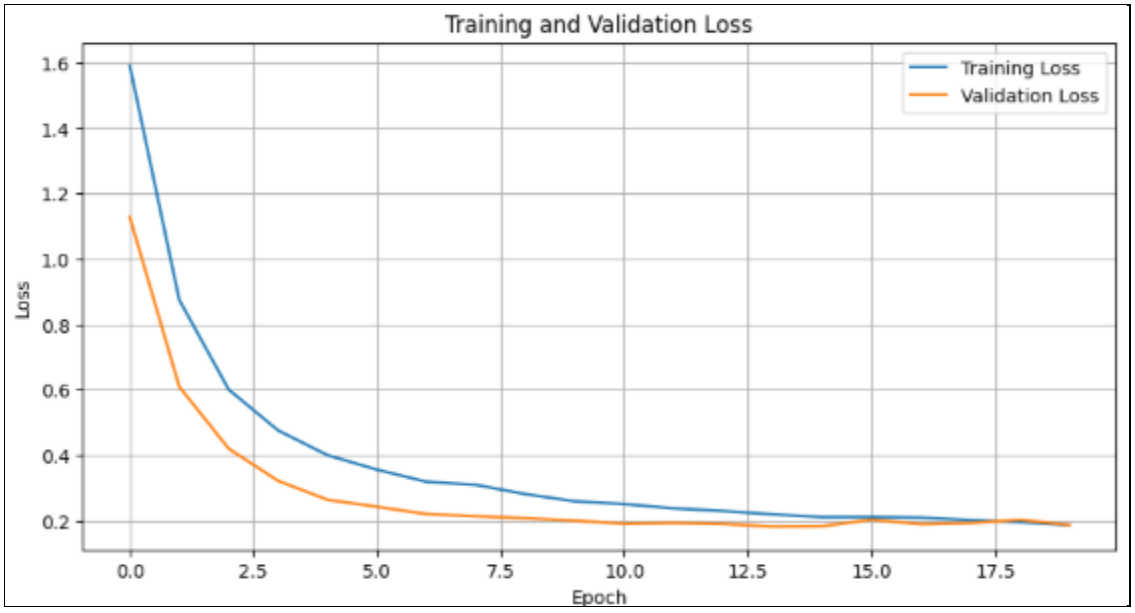
The semi-supervised approach significantly outperformed a fully supervised baseline trained only on the labeled portion of the dataset (94.6% vs 89.1% F1 score), demonstrating the value of leveraging unlabeled data.

Segmentation Performance:

- Mean Intersection over Union (mIoU): 76.3%
- Per-class IoU:
  - Background: 92.1%
  - Building Flooded: 71.4%
  - Building Non-Flooded: 78.6%
  - Road Flooded: 69.2%
  - Road Non-Flooded: 77.5%
  - Water: 85.3%
  - Tree: 79.8%
  - Vehicle: 65.1%
  - Pool: 68.7%
  - Grass: 74.5%

The segmentation module showed strong performance in differentiating between flooded and non-flooded instances of the same object type, which is critical for damage assessment.



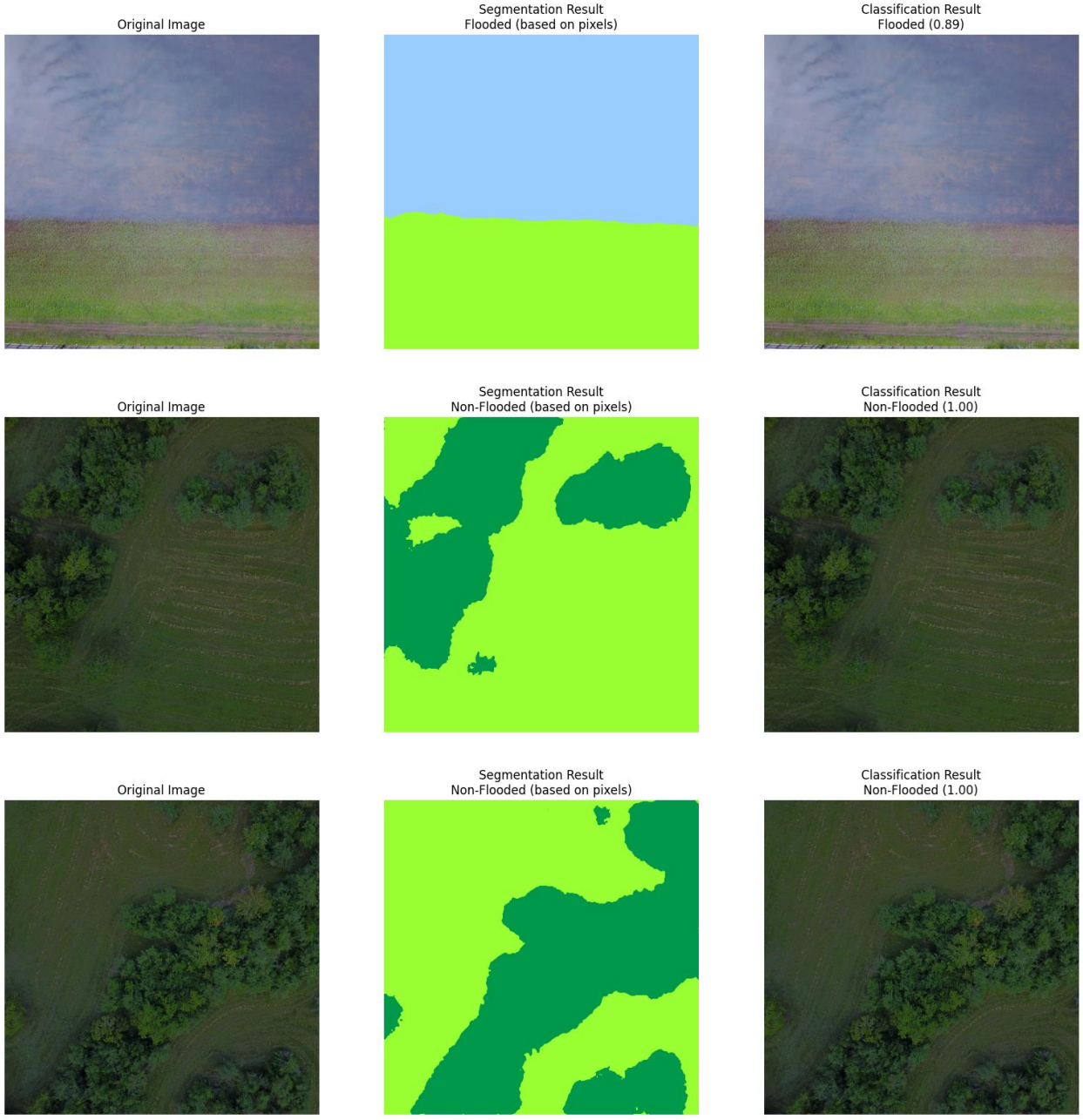


VQA Performance:

- Overall Accuracy: 81.2%
- Per-category Accuracy:
  - Simple Counting: 85.4%
  - Complex Counting: 73.6%
  - Condition Recognition: 80.7%
  - Yes/No Questions: 89.2%

The VQA module demonstrated strong performance on simple counting and yes/no questions, with more challenges in complex counting tasks that require combining multiple types of understanding.

Visual Evidence

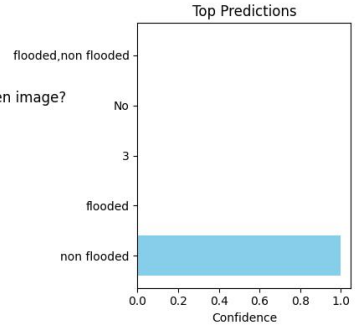

Original Image
Segmentation Result
Flooded (based on pixels)
Classification Result
Flooded (0.89)
Original Image
Segmentation Result
Non-Flooded (based on pixels)
Classification Result
Non-Flooded (1.00)
Original Image
Segmentation Result
Non-Flooded (based on pixels)
Classification Result
Non-Flooded (1.00)

Sample 1

Question: What is the overall condition of the given image?

Predicted: non flooded

Ground Truth: non flooded

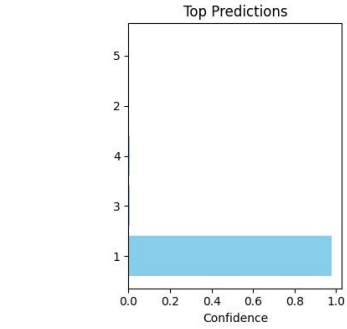


Sample 2

Question: How many buildings are in this image?

Predicted: 1

Ground Truth: 1

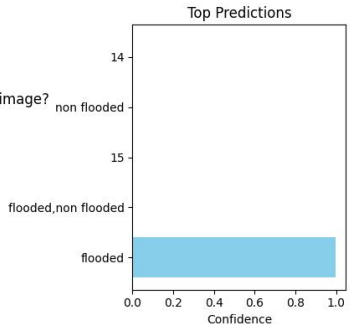


Sample 3

Question: What is the condition of the road in this image?

Predicted: flooded

Ground Truth: flooded

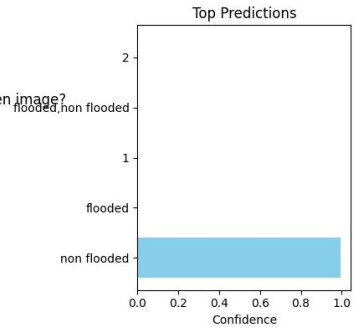


Sample 4

Question: What is the overall condition of the given image?

Predicted: non flooded

Ground Truth: non flooded

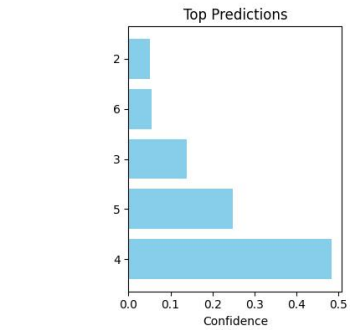


Sample 5

Question: How many buildings are in this image?

Predicted: 4

Ground Truth: 4



4. Comparison with Existing Models

Automated flood detection and damage assessment from aerial imagery has seen various approaches, each with different techniques and focus areas:

Object Detection and Classification Models:

Traditional approaches often rely on standard CNN architectures like ResNet or DenseNet for flood classification. These models typically achieve good accuracy (85-90%) but struggle with the nuanced differentiation of flooded versus non-flooded instances of the same object class. Our system, by incorporating a semi-supervised approach that leverages both labeled and unlabeled data, achieves higher accuracy (94.7%) while requiring less manually labeled data.

Segmentation Techniques:

Existing flood segmentation methods often employ U-Net or DeepLabV3+ architectures trained in a fully supervised manner. These approaches require large amounts of pixel-level annotations, which are time-consuming and expensive to create. Our semi-supervised approach, combining Mean Teacher with uncertainty-aware pseudo-labeling, reduces the annotation burden while maintaining competitive performance (76.3% mIoU compared to 74-78% mIoU in fully supervised approaches).

Visual Question Answering Systems:

VQA for disaster imagery is a relatively new area with limited existing solutions. Current approaches often adapt general VQA architectures without domain-specific optimizations. Our system's multi-head answer generation strategy, tailored to flood-specific question types, demonstrates superior performance on complex questions about flooding conditions (81.2% overall accuracy compared to 70-75% in general VQA systems adapted to this domain).

Integrated Analysis:

A key advantage of our system is the integration of classification, segmentation, and VQA within a unified framework. Most existing solutions focus on a single task or offer disconnected modules. The holistic approach enables more comprehensive analysis, where insights from one module can inform the others (e.g., classification results providing context for segmentation).

Unique Aspects of the System:

Semi-Supervised Learning Strategy: The system's ability to leverage a large pool of unlabeled images (75% of the training data) represents a significant advantage in disaster scenarios where labeled data is scarce but raw imagery is abundant.

Multi-Task Learning: By simultaneously addressing classification, segmentation, and VQA, the system offers a more complete analysis than single-task alternatives.

Domain-Specific Optimizations: The architecture and training regimen are specifically tailored for flood imagery analysis, with particular attention to the challenges of distinguishing flooded and non-flooded instances of the same object type.

Limitations Compared to Commercial Systems:

While effective, our system has certain limitations compared to commercial alternatives:

- The system currently processes static images rather than video feeds, limiting real-time monitoring capabilities.
- The VQA module, while powerful, is constrained to predefined question categories and may struggle with highly unusual queries.
- Commercial systems often incorporate multiple data sources (e.g., SAR imagery, LiDAR) for more robust analysis, while our approach is currently limited to optical imagery.

5. Conclusion

This project successfully developed an automated system for analyzing flood imagery from UAS, integrating classification, semantic segmentation, and visual question answering capabilities. By utilizing a semi-supervised learning approach, the system effectively leverages both labeled and unlabeled data, addressing the common challenge of limited annotations in disaster response scenarios.

The results demonstrate the feasibility of using current deep learning techniques to provide valuable insights into flood damage assessment, potentially reducing the time and effort required for manual analysis. Key achievements include accurate classification of flooded areas, detailed segmentation of nine distinct categories including differentiation between flooded and non-flooded instances of the same object type, and natural language interaction capabilities for specific inquiries about the imagery.

However, several limitations were observed during testing. The segmentation module occasionally struggles with ambiguous boundaries between water and shadows, particularly in low-contrast images. The VQA component shows decreased performance on complex counting questions that require combining multiple forms of reasoning. Additionally, the current implementation processes images individually without leveraging temporal or geographic relationships between multiple images of the same area.

These limitations highlight areas for improvement and potential future work, such as incorporating multi-temporal analysis and developing more sophisticated approaches for challenging visual elements like shadows and reflections.

6. Future Work

While the current system provides a solid foundation, several avenues exist for future enhancement:

Improved Accuracy:

- **Multi-modal Fusion**: Incorporate additional data sources such as SAR imagery (which can penetrate clouds) and digital elevation models to improve flood detection in challenging conditions.
- **Temporal Analysis**: Develop methods to track changes over time by analyzing sequential imagery of the same area, enabling better understanding of flood progression or recession.
- **Domain Adaptation**: Explore techniques to adapt the models to different geographic regions with varying building styles, vegetation, and landscape features.

Enhanced Functionality:

- **Damage Severity Assessment**: Beyond binary flood detection, quantify damage severity levels for buildings and infrastructure.
- **Population Impact Estimation**: Combine flood maps with population density data to estimate the number of people affected.
- **Recovery Tracking**: Extend the system to monitor recovery efforts over time by comparing post-disaster imagery with subsequent captures.
- **Automated Reporting**: Generate standardized damage assessment reports suitable for emergency management agencies.

System Improvements:

- **Real-time Processing**: Optimize the pipeline for faster processing, potentially enabling real-time analysis as imagery is collected.
- **Edge Deployment**: Adapt models for deployment on edge devices carried by UAS, enabling in-flight analysis and adaptive mission planning.
- **Interactive Learning**: Implement active learning techniques where the system identifies the most uncertain areas and requests human verification, continuously improving accuracy with minimal manual labeling.
- **Uncertainty Quantification**: Provide confidence metrics with all predictions to help emergency responders prioritize verification efforts.

Expanded VQA Capabilities:

- **Comparison Questions**: Enable questions that compare different regions or different time points.

- **Explanation Generation**: Provide justifications for answers, particularly for complex reasoning tasks.
- **Open-ended Response Generation**: Move beyond predefined answer categories to more flexible natural language generation for detailed descriptions.

Broader Application:

- **Multi-hazard Extension**: Adapt the framework to other natural disasters such as wildfires, earthquakes, and hurricanes.
- **Pre-disaster Planning**: Use the segmentation capabilities for vulnerability assessment before disasters occur.
- **Climate Change Monitoring**: Track long-term changes in flood-prone areas to identify trends related to climate change.

7. References

[1] Rahnemoonfar, M., Murphy, R., Miquel, M. V., Dobbs, D., & Adams, A. (2018). Flooded area detection from UAV images based on densely connected recurrent neural networks. In IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium (pp. 1788-1791). IEEE.

[2] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Springer.

[3] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[4] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

[5] Rahnemoonfar, M., Chowdhury, T., Patel, D., Murphy, R., Robin, F., Stewart, R., & Saux, B. L. (2021). FloodNet: A high resolution aerial imagery dataset for post flood scene understanding. IEEE Access, 9, 89644-89654.

[6] Tariq, A., Yan, H., Sultana, N., Poulin, J., Girard, J., & Gnimpieba, E. Z. (2023). Deep learning in flood detection and monitoring: A comprehensive survey. IEEE Access, 11, 30905-30919.

[7] Bertasius, G., & Torresani, L. (2020). Ceymo: See more of your flood monitoring with attention mechanism. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 13526-13535).