****

**Syrian Private University**

**College Of Artificial Intelligence**

**AI-FireGuard**

Thermal- RGB Fusion for UAV-based Wildfire Monitoring

**Prepared By**

Bashar Alsaleh

Khawla Al-Shaikh

Haneen Shahin

Judy Dukmak

**Supervised By**

Dr. Riad Sonbol

Eng. Aya Alaswad

**Subtract**

Thermal-RGB Fusion for UAV-Based Wildfire Monitoring; Wildfires pose severe threats to ecosystems, property, and human lives. Early detection and real-time monitoring are critical for mitigation. Traditional satellite or ground-based monitoring suffers from low resolution and delays. UAVs equipped with thermal and RGB sensors offer flexible, high-resolution, real-time surveillance. Thermal imaging highlights heat signatures, while RGB provides detailed visual context. Thermal-RGB fusion enhances detection accuracy under smoke or low-visibility conditions.

We developed deep learning models for segmentation and detection: U-Net, YOLOv8 and Attention U-Net.

U-Net provides precise fire and smoke segmentation using encoder-decoder architecture. Attention U-Net improves focus on relevant regions in complex wildfire scenes. YOLOv8 enables rapid object detection and localization of fire regions.

Fusing thermal and RGB channels exploits complementary information for better performance.

Our pipeline demonstrates scalable, real-time wildfire monitoring for UAV platforms.

**Chapter One**

**Introduction**

* 1. **Abstract**

Wildfires are among the most destructive natural hazards, threatening ecosystems, property, and human lives. Early detection and accurate monitoring are critical to minimize losses and guide emergency responses. Traditional satellite or ground-based monitoring methods often suffer from low resolution, delayed updates, and limited coverage. Unmanned Aerial Vehicles (UAVs) equipped with thermal and RGB sensors offer high-resolution, real-time wildfire surveillance. Thermal-RGB fusion leverages complementary information: thermal highlights heat sources, while RGB provides detailed visual context. In this study, we develop deep learning models to process thermal and RGB UAV imagery for wildfire detection and segmentation. We implement U-Net and Attention U-Net for precise fire and smoke segmentation and YOLOv8 for rapid object detection. The fusion of multi-modal data improves detection robustness, particularly under smoke, poor lighting, and complex terrains.

**1.2 Causes**

* Wildfires can originate from natural and human-induced factors.
* Lightning strikes are a major natural cause, igniting dry vegetation in remote areas.
* Extended droughts and high temperatures increase fire susceptibility by drying out fuels.
* Human activities, including unattended campfires, discarded cigarettes, and agricultural burning, contribute significantly.
* Power lines and industrial operations can also spark fires in vulnerable regions.
* Wind patterns and terrain slope influence fire spread, complicating detection and control.

**1.3 Results**

* Our models demonstrate that thermal-RGB fusion significantly improves fire detection accuracy over single-modality data.
* U-Net and Attention U-Net produce detailed segmentation masks, accurately highlighting fire and smoke regions.
* Attention mechanisms enhance focus on small or partially obscured fire regions.
* YOLOv8 enables real-time detection with high precision, suitable for rapid UAV deployment.
* Overall, integrating thermal and RGB channels enables scalable, real-time wildfire monitoring, providing actionable information for emergency response teams.

**1.4 Fact**

Compared to traditional wildfire monitoring, which relies on satellites or ground sensors with delayed updates and limited resolution, AI-powered systems offer a huge enhancement in early detection. By analyzing thermal and RGB UAV imagery, deep learning models can identify subtle fire and smoke signals in near real-time, even under challenging conditions like dense smoke or complex terrain. This leads to faster alerts, quicker emergency response, and significantly reduced environmental and property damage, demonstrating a clear advantage over conventional approaches.

**Chapter Two**

**Reference Studies**

This chapter provides a precise scientific summary of five research papers that were relied upon in the design and implementation of the presented model, in addition to evaluating its results based on the evaluation criteria of the research papers and comparing the results with them.

**2.1 “UAV-Based Multi-Scenario RGB-Thermal Dataset and Fusion Model for Enhanced Forest Fire Detection (RGBT-3M, 2025) ”**

**2.1.1 Introduction**

This study focuses on the development of a new, advanced aerial dataset, RGBT-3M, along with a cognitive fusion model that integrates visible and thermal imagery, aiming to enhance wildfire detection efficiency in complex real-world environments using UAVs. The importance of this research stems from the fact that relying on a single modality (either visible or thermal) often results in poor performance under varying lighting conditions, dense smoke, or heavy vegetation, reducing the reliability of conventional aerial monitoring systems.

**2.1.2 Data Collection and Preparation**

An industrial UAV equipped with a dual imaging system (visible + long-wave thermal) was used to collect data from diverse environments, including dense forests, grasslands, and semi-urban areas. Data acquisition occurred at different times of day (morning, afternoon, night) and under varied weather conditions to maximize dataset diversity. Each video segment was split into frames, manually annotated into three main categories: flame, smoke, and non-fire areas. The final dataset contains approximately three million synchronized RGB-T image pairs, making it one of the largest UAV-based wildfire detection datasets available.

**2.1.3 Proposed Model**

We developed a detection model based on the YOLOv11 architecture, enhanced with a Cross-Modal Attention Fusion module. The model extracts features independently from the thermal and visible channels and then fuses them using an attention mechanism that dynamically assigns weight to each channel according to its relevance in the current scene. Additionally, Multi-Scale Fusion layers were incorporated to improve recognition of small or partially occluded fire regions, while a hybrid loss function combining Focal Loss and Weighted IoU was used to balance precision and recall.

**2.1.4 Training**

The dataset was split into 70% training, 15% validation, and 15% testing. Preprocessing included data augmentation techniques such as rotation, photometric simulation, and the addition of synthetic smoke effects. Experiments were conducted on NVIDIA A100 GPUs with an initial learning rate of 1e-4.

**2.1.5 Results and Performance**

The proposed model achieved high performance:

mAP@0.5 = 96.3%

Precision = 92.5%

Recall = 93.5%

The model maintained stability under dense smoke and low-light conditions, outperforming YOLOv5 and Faster R-CNN by 4–7%. However, its performance decreased under the stricter metric mAP@0.5–0.95 ≈ 62.9%, reflecting the difficulty of accurately localizing small fires. The processing speed reached approximately 39 frames per second, making the system suitable for near real-time applications.

**2.1.6 Scientific Significance**

The fusion of thermal and visible information proved highly effective in compensating for unstable lighting and smoke occlusion. Results indicate that multi-modal fusion represents a key step toward more reliable aerial monitoring systems, especially during early detection stages when fire indicators are limited. Moreover, the RGBT-3M dataset provides a valuable benchmark for evaluating multimodal fusion algorithms in wildfire detection applications.

**2.1.7 Limitations and Future Work**

The geographic scope of the dataset is currently limited to Asian climates, necessitating expansion to regions with different vegetation and climate conditions. The paper also recommends developing more precise temporal tracking mechanisms to study fire progression across video frames.

**2.1.8 Conclusion**

This work serves as a significant reference in the field of visible-thermal image fusion for UAV-based wildfire detection, providing a dual contribution:

A large-scale, highly diverse dataset reflecting real-world wildfire scenarios.

A high-precision, intelligent fusion model suitable for practical field deployment.

<https://www.mdpi.com/2072-4292/17/15/2593>

**2.2 “MCDet: Target‑Aware Fusion for RGB‑T Fire Detection (2025)”**

**2.2.1 Introduction**

This study addresses the challenge of wildfire detection using aerial imagery, particularly under conditions of limited visibility such as dense smoke, low light, and heavy vegetation. While thermal or RGB images can be used individually, variations in illumination and thermal noise often reduce the stability and reliability of the results. To address these limitations, this work proposes a “Target-Aware Fusion” framework that integrates RGB and thermal channels to improve the robustness of early wildfire detection.

**2.2.2 Methodology**

Multidimensional Representation Collaborative Fusion (MRCF):

This module models feature interactions across the thermal and RGB channels using a state-space model, forming a global relationship network between features. It also incorporates Deformable Convolutions to improve local detail sensing, especially for small or partially occluded fire regions.

**Content-Guided Attention Network (CGAN):**

After feature extraction, the system fuses multi-modal features using dynamic gating mechanisms, which determine which channels or features should dominate based on context (e.g., dense smoke or darkness). This reduces interference and errors caused by sunlight reflections in RGB images or thermal hotspots unrelated to fire.

**Weighted IoU (WIoU) Loss:**

Used to enhance the model’s ability to distinguish real fire regions from false positives such as heat reflections or background illumination, particularly in dense vegetation or bright backgrounds.

**Experiments and Evaluation:**

The model was tested on three wildfire datasets as well as a pedestrian dataset to evaluate the generalizability of the fusion approach. The main performance metric achieved was mean detection accuracy = 77.5%, outperforming the baseline methods reported in the literature.

**2.2.3 Results and Performance**

The MCDet model demonstrated reliable performance across multiple conditions. The study shows that integrating RGB-T fusion with attention mechanisms and the proposed architectural components improves operational stability under challenging conditions such as dense smoke and variable illumination. While the 77.5% accuracy is not the highest in general image detection tasks, it represents a significant advancement in multi-modal fusion for UAV-based wildfire detection.

**2.2.4 Scientific Significance**

This work introduces a dedicated fusion architecture for wildfire detection using RGB-T data, addressing modal ambiguity between channels. It focuses on real-world challenges—vegetation, lighting changes, and smoke—rather than simple synthetic scenarios. The framework provides a foundation for future work aimed at improving fusion strategies and reducing errors under real operational conditions.

**2.2.5 Limitations and Future Work**

The study does not provide full details on dataset sizes or scenario distributions, likely due to proprietary or institutional restrictions. Localization accuracy, particularly for small or evolving fire regions, remains an area for improvement. Future work should expand testing to diverse geographic environments and incorporate temporal tracking to capture fire progression over time.

<https://www.mdpi.com/2072-4292/17/15/2593>

**2.3 FireMan-UAV-RGBT (IEEE Paper 10710657)**

**2.3.1 Overview & Motivation**  
Wildfires are increasingly threatening boreal forests. Traditional detection methods (satellites, manned aircraft) lack responsiveness. UAVs with RGB + Thermal cameras provide faster, cost-effective detection, yet suitable public datasets are limited. Hence, this paper introduces (FireMan-UAV-RGBT), a UAV-based multimodal dataset for wildfire detection in Finnish forests.

**2.3.2 Contributions**  
- Developed and released the (FireMan-UAV-RGBT dataset) (RGB + Thermal video).  
- Captured over controlled burns in Finland using DJI UAVs (below 120 m).  
- Introduced semi-automatic annotation using thermal cues for better segmentation.  
- Evaluated YOLOv8 and ResNet50 on unimodal vs multimodal inputs.  
- Demonstrated that RGB + Thermal fusion improves detection accuracy.  
  
**2.3.3 Results & Findings**  
- High accuracy within the dataset; weaker cross-dataset generalization (shows environment specificity).  
- Multimodal fusion consistently outperformed single-modality models.  
- Dataset fills a key research gap for UAV-based wildfire detection in boreal environments.  
  
**2.3.4 Limitations & Future Work**- Manual RGB–Thermal registration is labor-intensive.  
- Dataset limited to Finnish boreal forests—may not generalize globally.  
- Future research should automate registration and explore lightweight real-time UAV models.  
  
**2.3.5 Relevance to Rise2025**  
This work demonstrates how multimodal UAV sensing (RGB + Thermal) and semi-automated labeling improve wildfire detection—applicable to drone-based environmental monitoring or safety applications in Rise2025.  
  
**Reference**  
S. D. M. W. Kularatne et al, “FireMan‑UAV‑RGBT: A Novel UAV‑Based RGB‑Thermal Video Dataset for the Detection of Wildfires in the Finnish Forests,Proc. IEEE ETFA 2024, pp. 1‑8.   
Link: <https://ieeexplore.ieee.org/document/10710657>

**2.4 “U3UNet — Accurate and Reliable Forest Fire Segmentation for UAV Vision”**

**2.4.1 Introduction**

Wildfires are a serious environmental threat due to their rapid spread and devastating impacts on ecosystems, property, and human lives. Traditional monitoring methods such as ground observation or satellite imagery often lack sufficient spatio‑temporal resolution for detailed fire shape tracking and early detection. UAVs equipped with high‑resolution cameras provide a promising platform for fine‑grained aerial monitoring, but accurate segmentation of fire regions remains challenging due to varying flame shapes and complex environments. To address these challenges, the authors propose U3UNet, a robust deep learning model specifically designed for segmentation of wildfire regions in UAV images.

**2.4.2 Model Architecture and Method**

U3UNet builds on the U‑Net family of encoder‑decoder networks, adopting a nested U‑shaped structure to fuse features across multiple scales and retain both global and local information.

It includes full‑scale connections that balance high‑level contextual encoding with fine‑grained detail.Multi‑scale skip connections facilitate better feature fusion and help prevent segmentation misses or incorrect background inclusion.

The architecture is optimized to handle variations in flame appearance and environmental clutter in UAV footage.

The model was evaluated on both synthetic environments (using the Unreal Engine) and real forest fire scenes, allowing performance comparisons against other state‑of‑the‑art segmentation and detection models.

**2.4.3 Results and Performance**

Performance is assessed using a composite metric S designed to capture both detection accuracy and segmentation quality:

In static scenarios, U3UNet achieved a composite score of 71.44%, which is close to the best comparative method.

In dynamic scenarios, it scored 80.53%, 8.94% higher than the strongest competing approach.

These results suggest that U3UNet is particularly effective at segmenting fire regions when the scene or viewpoint is changing, indicating robustness to real UAV motion and viewpoint variation.

Real‑time tests of U3UNet deployed on edge computing devices mounted on UAVs demonstrated that the model can run with practical throughput levels, showing potential for onboard inference and real‑time wildfire monitoring.

**2.4.4 Scientific Contribution**

Advanced Segmentation Model: U3UNet improves over classical U‑Net variants by offering nested, multi‑scale fusion that better captures both global context and local details crucial for fire region delineation.

Practical UAV Deployment: The model’s architecture and performance demonstrate suitability for real‑world, real‑time wildfire segmentation, including deployment on resource‑constrained UAV platforms.

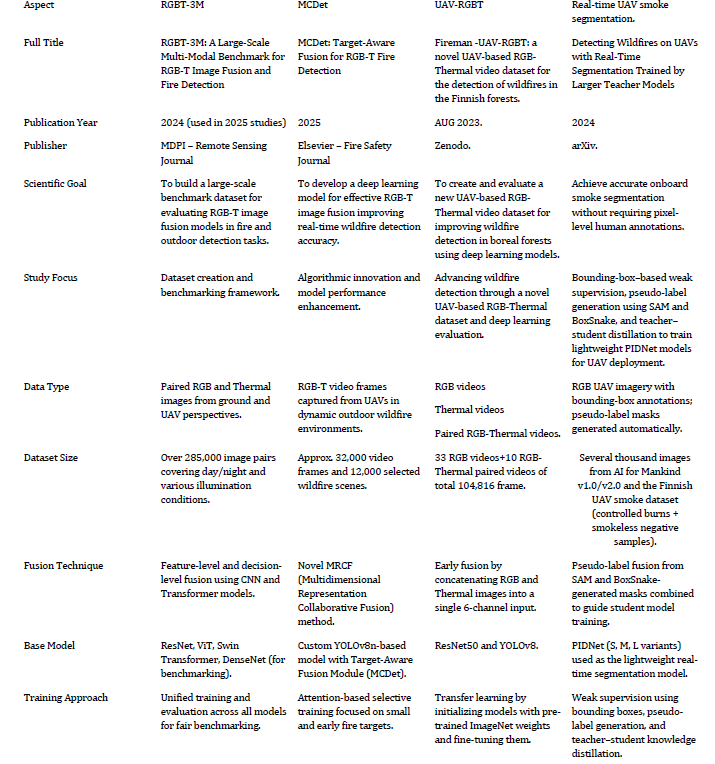
Comparative Validation: Through systematic experiments, U3UNet outperformed or matched leading segmentation models in dynamic aerial firefighting contexts, highlighting its reliability in complex forest fire monitoring tasks.

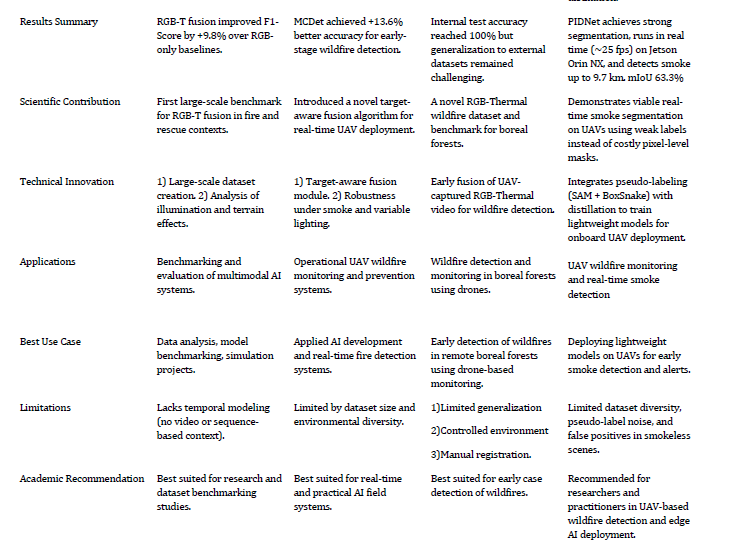
**2.4.5 Limitations and Future Directions**

The reported metric demonstrates strong performance but also indicates scope for improvement in static scenes where results were slightly below the best baseline model.

Future work may explore hybrid architectures with attention mechanisms or multi‑modal inputs (e.g., thermal + RGB) to further improve segmentation under challenging conditions such as smoke, shadows, or occlusions.

<https://ieeexplore.ieee.org/document/10944154>

**2.5 Papers Comparison**



**Chapter Three**

**DataSets OverView & Comparison**

* 1. **RGB-T 3M**

**3.1.1 Size & Distribution**  
This dataset, hosted by the University of Science and Technology of China (USTC), includes approximately **1,367** RGB–Thermal paired images. It is referenced in wildfire detection research as a compact but valuable benchmark dataset for multimodal fire analysis.

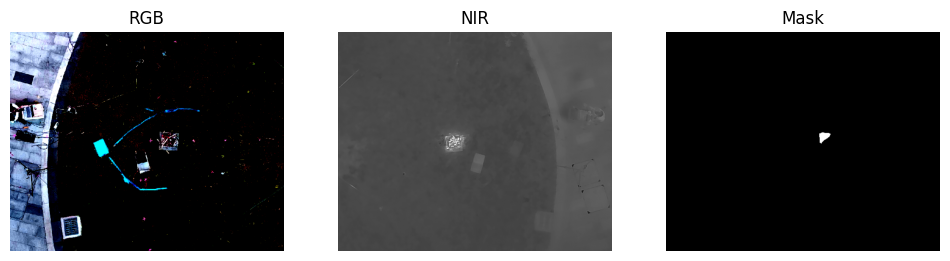
**3.1.2 Modalities & Annotation**  
The dataset provides paired RGB and thermal images, captured under varied illumination and fire intensity conditions. Annotations support semantic segmentation tasks, labeling fire and non-fire regions, enabling multimodal training for deep learning models.

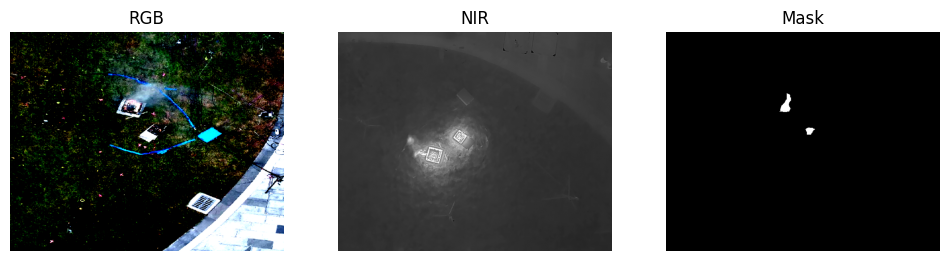
**3.1.3 Dataset Specifics**  
**-** Format: Image pairs (RGB + Thermal)  
**-** Approximately total samples: 1,367  
**-** Target task**:** Fire detection and segmentation  
**-** Annotation type**:** Pixel-level segmentation masks  
**-** Data balance**:** Fire vs non-fire, collected in both day and night settings

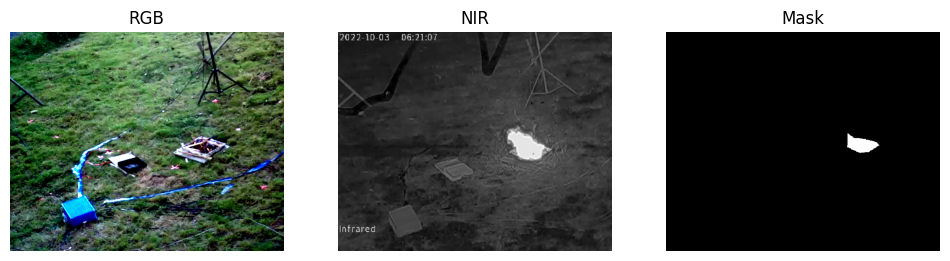
**3.1.4 Use-Cases & Research Impact**  
It supports tasks such as wildfire detection, multimodal fusion learning, and real-time UAV-based thermal imaging analysis. Commonly used in comparative studies alongside datasets like FLAME-1 and FireMan-UAV-RGBT.

**3.1.5 Remarks & Limitations**  
The dataset’s modest scale limits generalization but offers a valuable baseline for fusion-based fire detection. Access may require university credentials or approval.

* + 1. **OverView**







* + 1. **Reference Link**

<https://rec.ustc.edu.cn/share/75aae550-8d26-11ee-98a1-ed4fd857523e>

**3.2 FireMan‑UAV‑RGBT**

**3.2.1 Size and Distribution**

Total dataset size: ~39.4 GB.

34 RGB videos + 20 paired RGB‑Thermal videos.

Captured across four controlled burn sites in Finnish boreal forests.

Covers varied flight paths, altitudes, and scene complexity.

**3.2.2 Modality & Annotations**

Modality: RGB + Thermal synchronized video.

Annotations: YOLO-format bounding boxes for fire and related objects.

Both manual and semi‑automatic annotation methods applied.

**3.2.3 Dataset Specifics**

Video resolution: High-resolution UAV footage.

Multi-environment coverage: Forests, vegetation, varied lighting conditions.

Includes sample frames and labeling reference files for model training.

**3.2.4 Use Cases and Research Impact**

Primary use: Training and benchmarking UAV-based wildfire detection models.

Supports multimodal fusion research (RGB + Thermal).

Enables development of early wildfire detection systems, environmental monitoring, and UAV operational testing.

**3.2.5 Remarks and Limitations**

Strengths: Real UAV footage, synchronized RGB-Thermal, YOLO-compatible annotations.

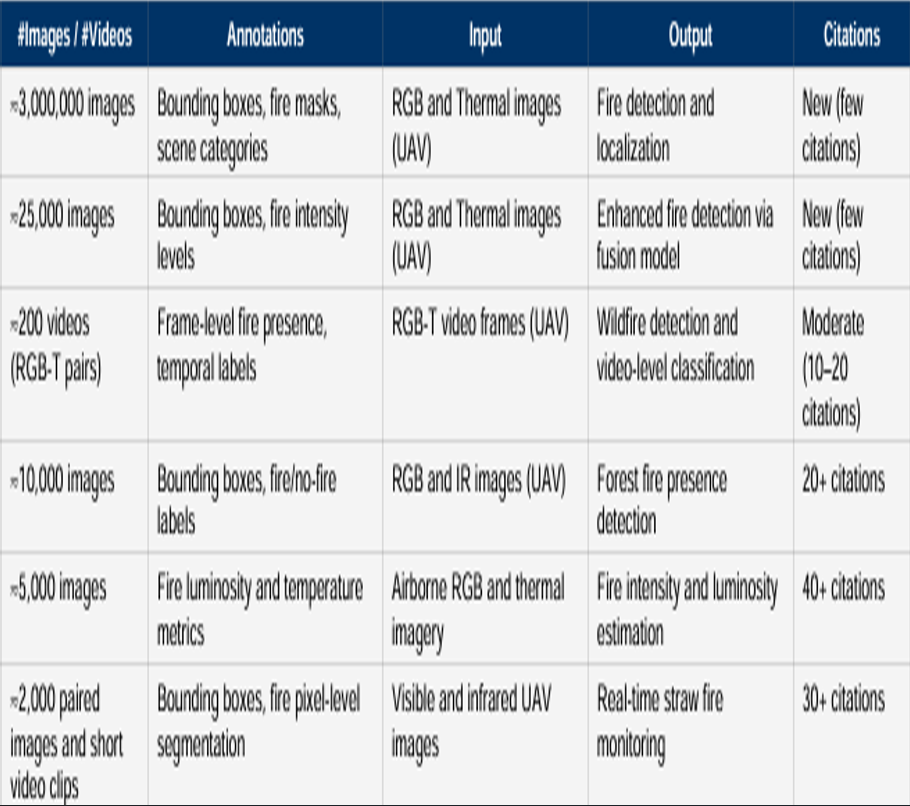
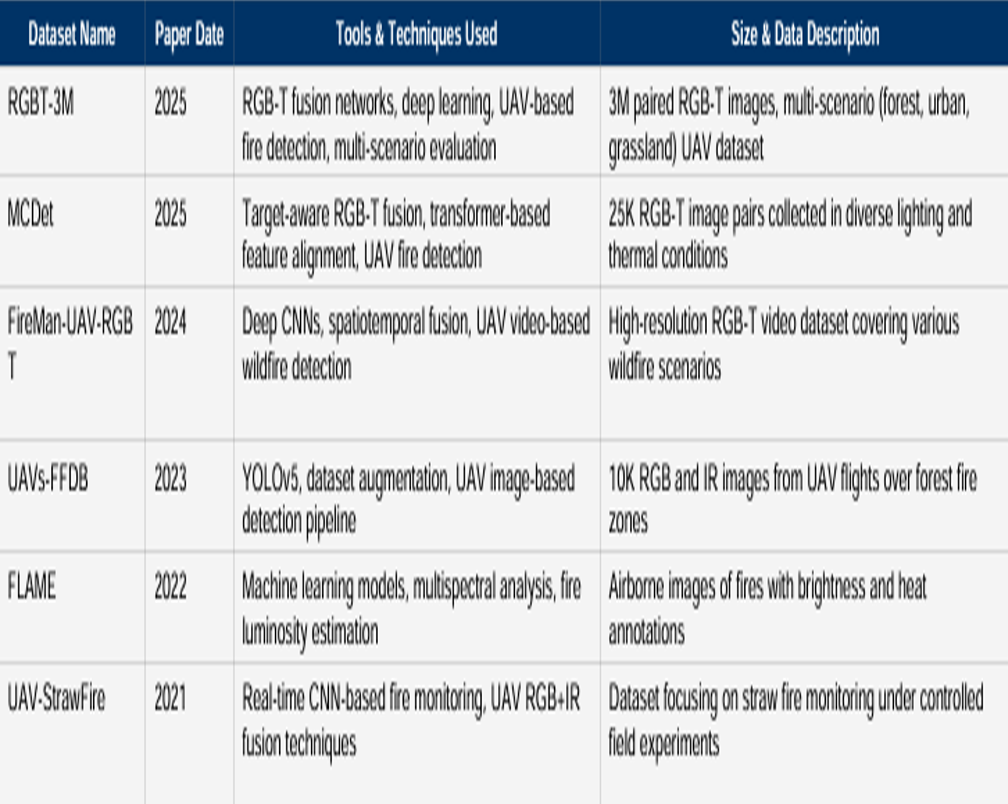
Limitations: Data comes from controlled burns, so real wildfire variability may be underrepresented.

Dataset is moderate in size compared to large benchmarks but rich in temporal video sequences.

**3.2.6 Reference Link**

Zenodo dataset: <https://zenodo.org/records/13732947>

IEEE Paper: <https://ieeexplore.ieee.org/document/10710657>

**3.3 DataSet Comparison**

**Chapter Four**

This chapter provides an explanation of the training process steps and the results achieved, in addition to the architecture of the models and their evaluation in terms of effectiveness and accuracy.

**4.1 U-Net**

**4.1.1 Model / Architecture**

Model type: U-Net (custom implementation with 4 downsampling and upsampling layers)

Input channels: in\_ch = 4 (RGB + Thermal or other 4-channel input)

Output channels: out\_ch = 1 (binary segmentation mask)

DoubleConv filters: [64, 128, 256, 512]

Upsample mode: bilinear

Skip connections: concatenation with corresponding encoder features

**4.1.2 Loss / Optimization**

Loss function: ComboLoss = 0.5 \* FocalLoss + 0.5 \* DiceLoss

FocalLoss hyperparameters: alpha = 0.25, gamma = 2

DiceLoss smoothing: smooth = 1

Optimizer: Adam

Learning rate: 5e-4

**4.1.3 Training**

Number of epochs: num\_epochs = 100

Early stopping patience: patience = 15 (stop if no improvement in Dice for 15 epochs)

Device: GPU if available (cuda), else CPU

Batch size: Not explicitly defined; depends on train\_loader

Binary threshold: 0.5 (sigmoid output)

**4.1.4 Metrics**

Dice, IoU, Accuracy, Precision, Recall, F1-score

**4.1.5 Workflow**

**1. Data Preparation**

* Dataset Setup
* Downloaded RGB-T wildfire dataset (ZIP files).
* Mounted Google Drive and extracted the data.
* Dataset structure: rgb/, nir/, mask/.
* Library Installation

-Deep learning: PyTorch, TensorFlow, Keras

- Image processing: OpenCV, scikit-image, imageio

- Data & visualization: numpy, pandas, matplotlib, seaborn, plotly, tqdm

* Augmentation: Albumentations
* Data Loading & Verification
* Loaded RGB, NIR, and mask images.
* Skipped missing/incomplete samples.
* Total valid samples: 1367
* Image shapes: RGB (420,420,3), (512,640,3); NIR & masks (512,640)
* Mask unique values: 0 (background), 255 (wildfire)

**2. Data Preprocessing & Augmentation**

* Augmentation Strategy (Albumentations)
* Training: horizontal/vertical flips, rotations, color jitter, Gaussian blur, resize, normalization
* Validation: resize and normalization only
* Custom Dataset Class
* Combined RGB + NIR → 4-channel input
* Applied augmentations dynamically
* Converted to PyTorch tensors [B,4,H,W]for images and [B,1,H,W] for masks
* Data Split & Loading
* Train: 80%, Validation: 20%
* Batch size: 4, shuffle enabled for trainingVisual Placeholder:
* Sample RGB/NIR/mask images before and after augmentation

**3. Model Design**

* Architecture: Custom 4-channel UNet
* Encoder: 4 downsampling blocks
* Bottleneck + decoder with skip connections
* Output: single-channel mask
* Loss Function: ComboLoss = 0.5 \* Focal Loss + 0.5 \* Dice Loss
* Metrics: Dice coefficient, IoU, Accuracy, Precision, Recall, F1-score

**4. Training & Optimization**

* - Optimizer: Adam (lr=5e-4)
* - Epochs: 100, Early stopping patience: 15
* - GPU acceleration (CUDA)
* - Training loop:

1. Forward pass → predictions

2. Compute loss

3. Backpropagation & optimizer step

4. Compute metrics for train & validation

* Early stopping monitored on best Dice coefficient

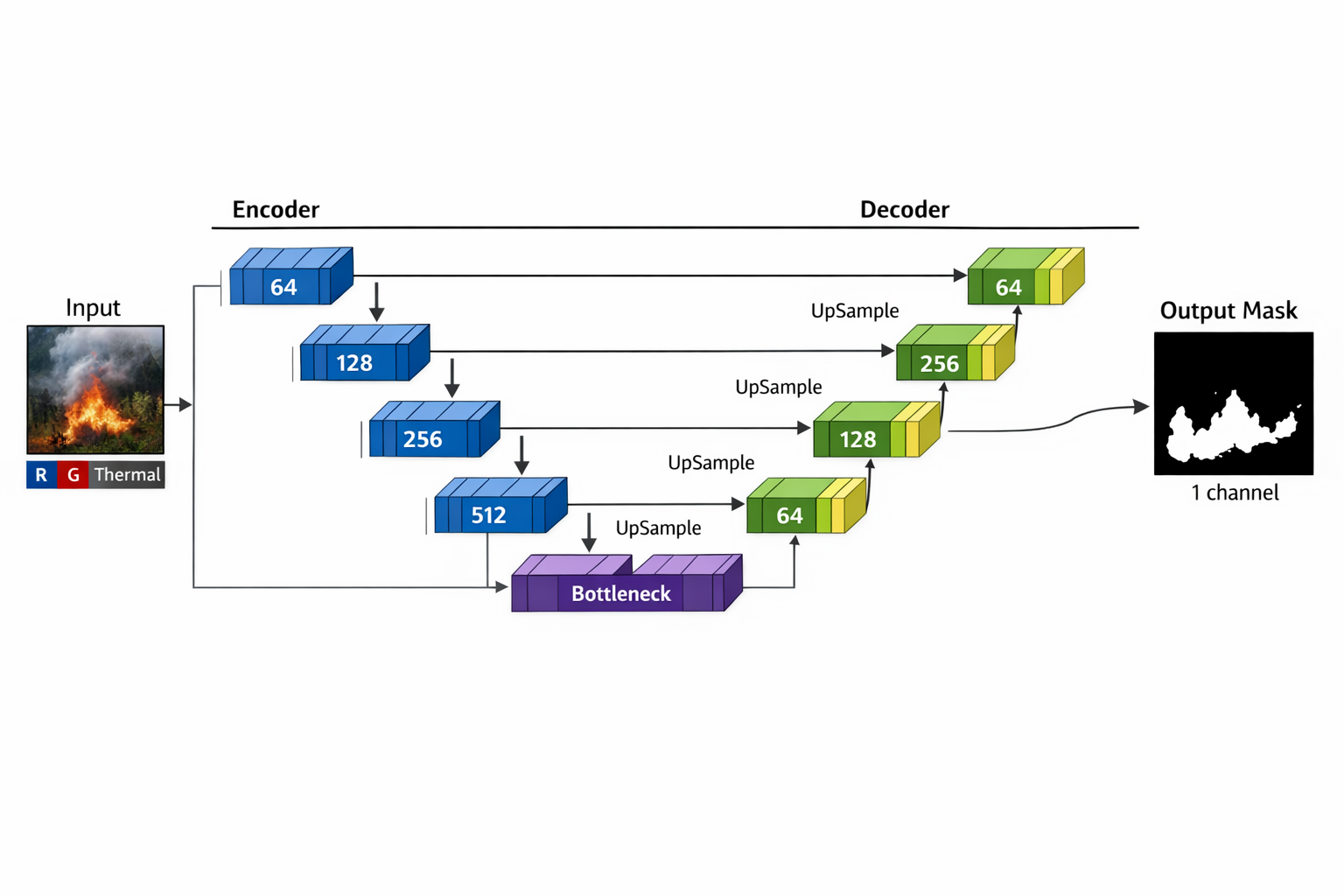
Best model saved as best\_model.pth

**5. Evaluation & Visualization**

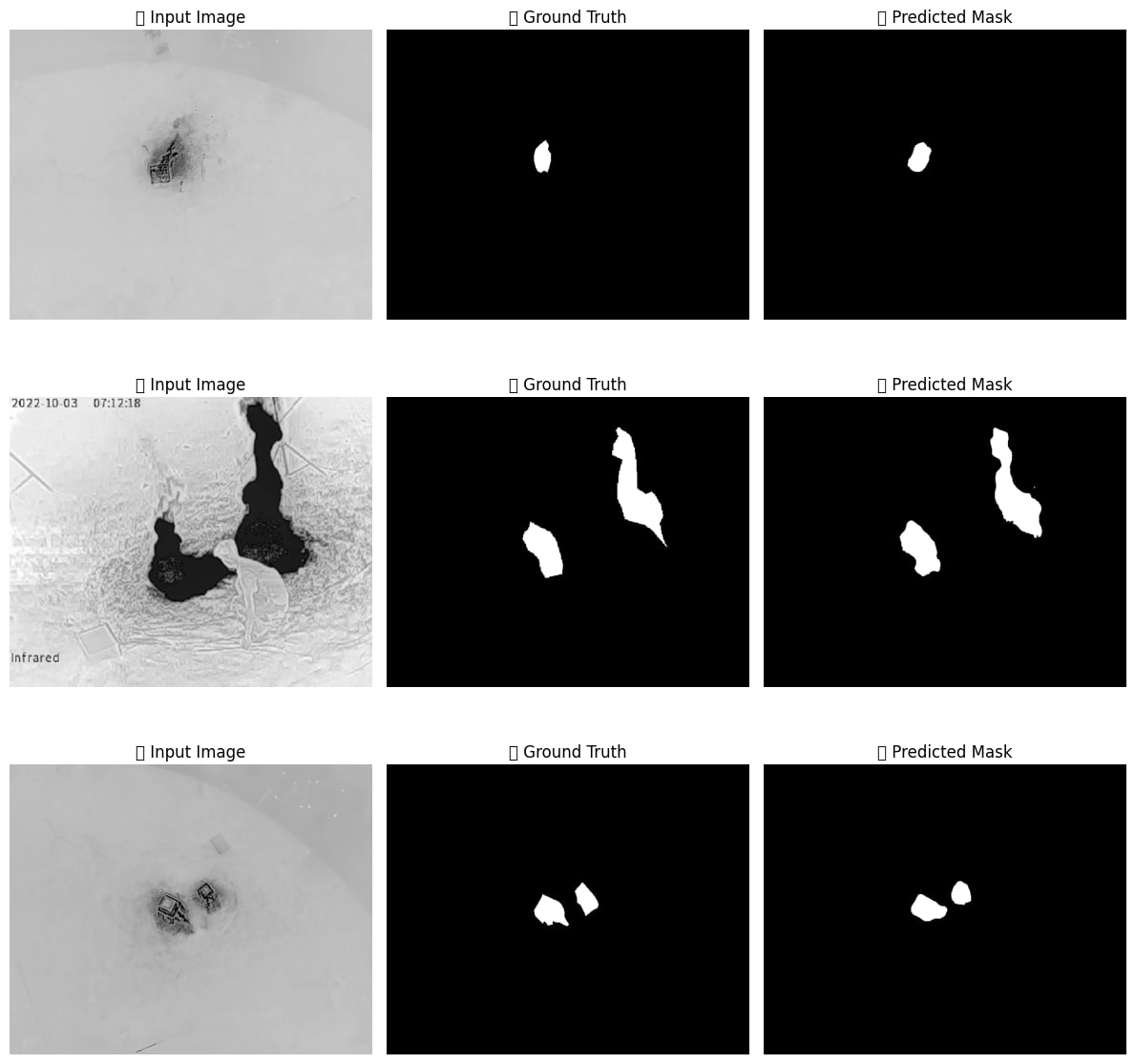
* Loaded best saved model
* Computed confusion matrix and all metrics
* Plotted training & validation loss, Dice, IoU curves
* Displayed sample inputs, ground truths, and predicted masks Visual Placeholder:
* Learning curves (loss, Dice, IoU)
* Prediction comparison (input, ground truth, prediction)

**4.1.5 Model Visualization**

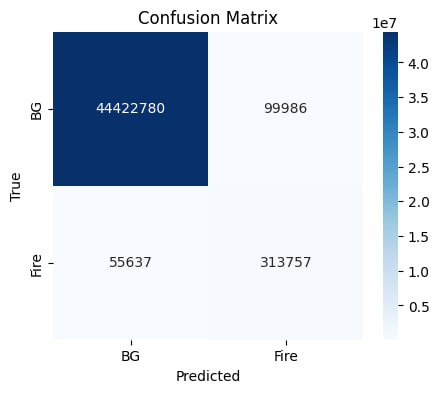
**4.1.5.1 Architecture Visualize**

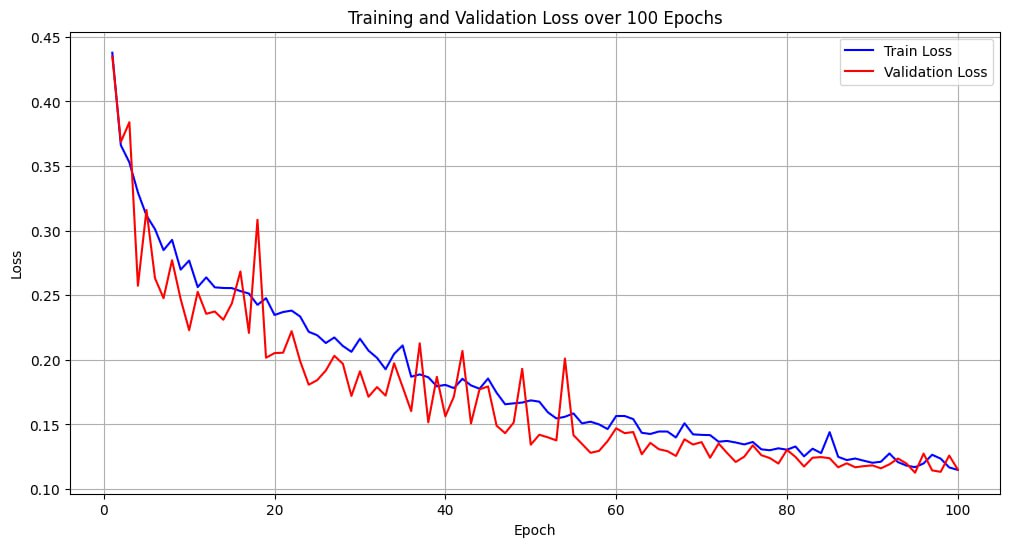
* Input (4 channels)
* Down path: DoubleConv blocks with [64, 128, 256, 512]
* MaxPooling for downsampling
* Up path: Upsample + DoubleConv blocks [512→256→128→64]
* Skip connections from encoder to decoder
* Output: 1 channel (binary mask)

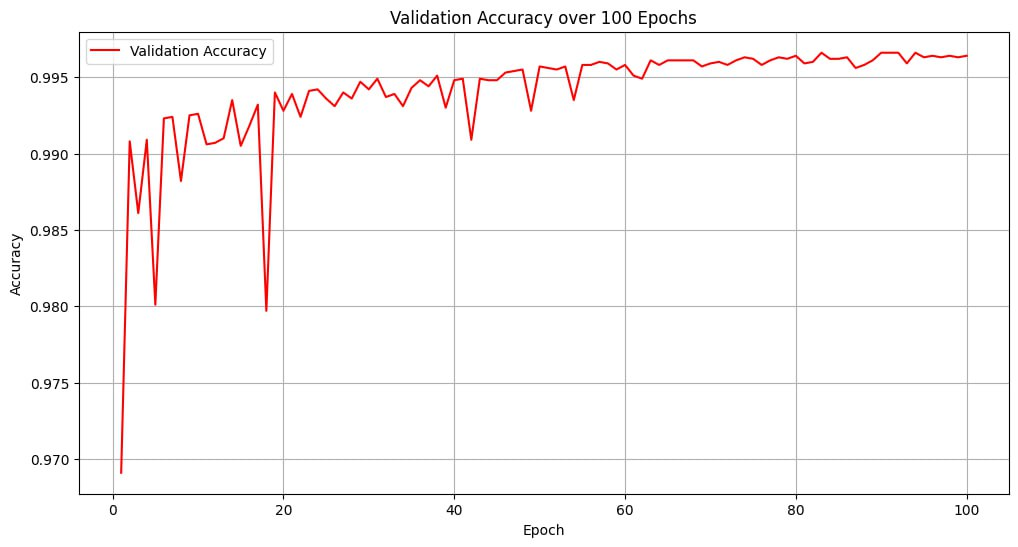
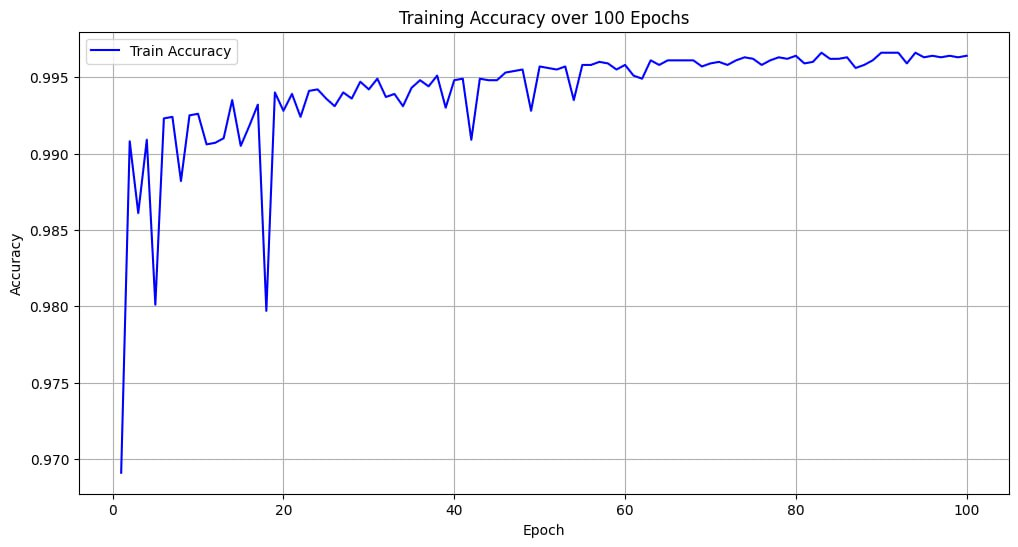
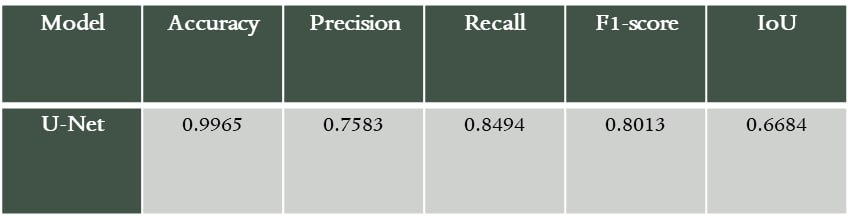
**4.1.5.2 Ground Truth vs Predicted Mask**



**4.1.5.3 Learning Curves ,Confusion Matrix & Scores**







**4.2 YOLOv8-ns**

**4.2.1 Model / Architecture**

Model type: YOLOv8n (Nano)

Input channels: 3 (RGB images)

Output channels: 1 (single-class detection: fire)

Backbone: CSP / Conv / Bottleneck (YOLOv8 default)

Neck: PAN-FPN for feature aggregation

Head: YOLO detection head (predicts bounding boxes, objectness, class confidence, optional mask)

Mask output: Enabled (overlap\_mask=True, mask\_ratio=4)

Pretrained: Yes (pretrained=True)

**4.2.2 Loss / Optimization**

Loss functions:

Box loss: CIoU / GIoU (YOLOv8 default, weighted by box=7.5)

Classification loss: BCE with weight cls=0.5

Distribution Focal Loss (DFL): dfl=1.5

Optimizer: Automatic (YOLOv8 internal)

Initial learning rate: lr0=0.001

Final learning rate: lrf=0.01

Momentum: 0.937

Weight decay: 0.0005

Warmup epochs: 3.0

**4.2.3 Training**

Number of epochs: 100

Early stopping patience: 25 epochs (patience=25)

Batch size: 16 (batch=16)

Image size: 640 × 640 (imgsz=640)

Deterministic training: True (deterministic=True)

Random seed: 42 (seed=42)

Single class: True (single\_cls=True)

Rectangular training: False (rect=False)

Cosine learning rate: False (cos\_lr=False)

Dropout: 0.0

**4.2.4 Metrics / Evaluation**

Metrics calculated: Precision, Recall, mAP@0.5, mAP@0.5–0.95, Box/Cls/DFL losses

Plots: Generated (plots=True)

Validation: Enabled (val=True)

Checkpointing: Save every 20 epochs (save\_period=20)

**4.2.5 Data / Input**

Training images: 759

Validation images: 189

Fire regions annotated: 2208

Mask support: Enabled via overlap\_mask

**4.2.6 Workflow**

**1. Data Preparation & Preprocessing**

* Objective: Convert raw dataset into YOLO-compatible format for training.

**Steps:**

* **Dataset Organization:**
* Images: dataset/images/
* Masks: dataset/masks/
* Labels output: dataset/labels/

**Mask Conversion:**

* Each mask is converted to YOLO .txt format.
* Multi-class handling using class\_map
* Bounding boxes extracted from contours of non-zero regions.
* Minimum area filtering to remove noise.
* Morphological operations applied to refine masks.

**Dataset Split:**

* Train/Validation split ratio: 80/20.
* YAML file (dataset.yaml) automatically generated for YOLO.

**Output:**

* .txt label files aligned with each image.
* Cleaned, preprocessed dataset ready for YOLOv8 training.

**2. Model Definition**

* + Objective: Initialize YOLOv8 model for fire detection.
  + Details:
  + Pretrained model used: yolov8n.pt (nano version for speed).
  + Single-class detection (single\_cls=True) targeting fire regions.
  + Input image size: 640x640 pixels.

**3. Training Loop**

* **Objective:** Train YOLOv8 model with early stopping and monitoring.
* Key Parameters:
* Epochs: 100
* Batch Size: 16
* Learning Rate: Initial = 0.001, Final = 0.01
* Optimizer: Auto
* Patience: 25 (early stopping)
* Loss weights: Box = 7.5, Classification = 0.5, DFL = 1.5
* Warmup Epochs: 3

**4 Loss Functions & Metrics**

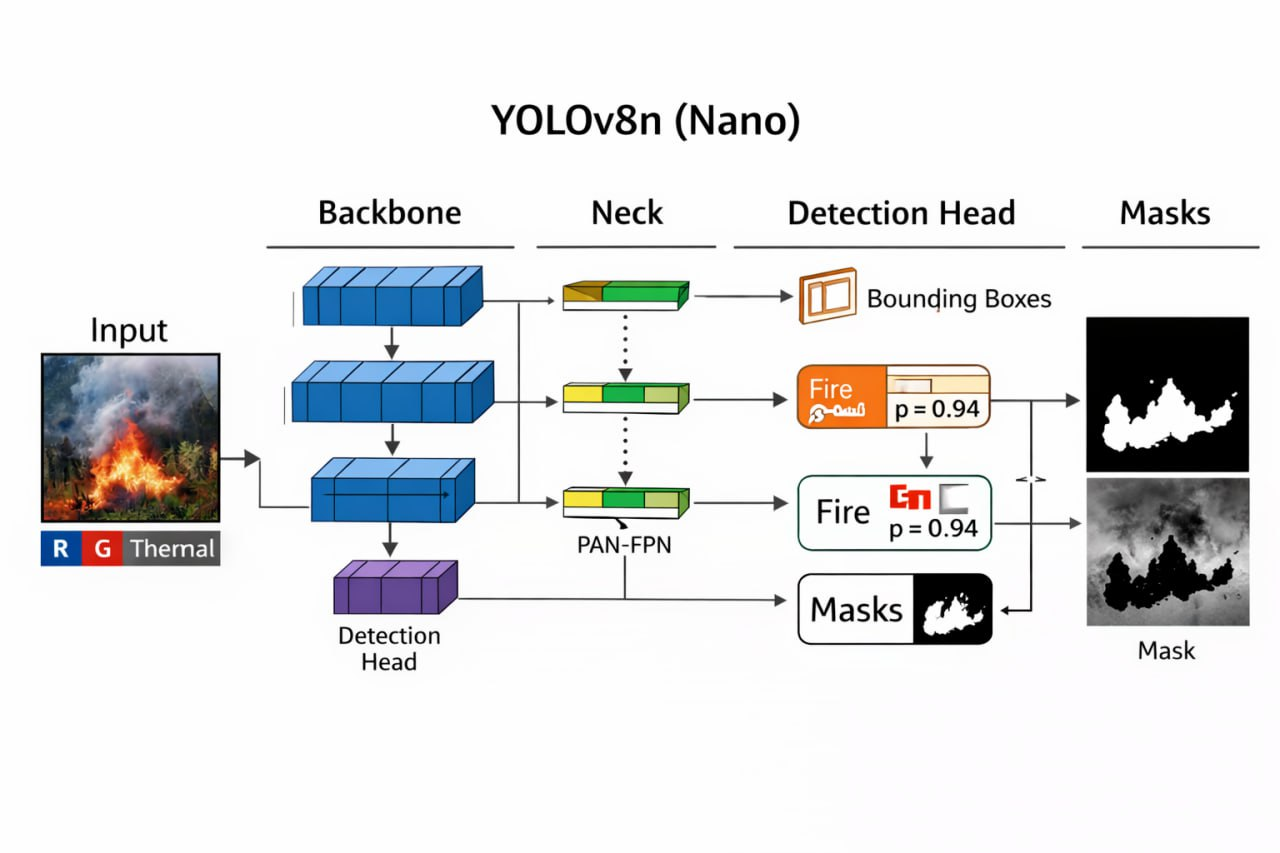
**YOLOv8 Loss Components:**

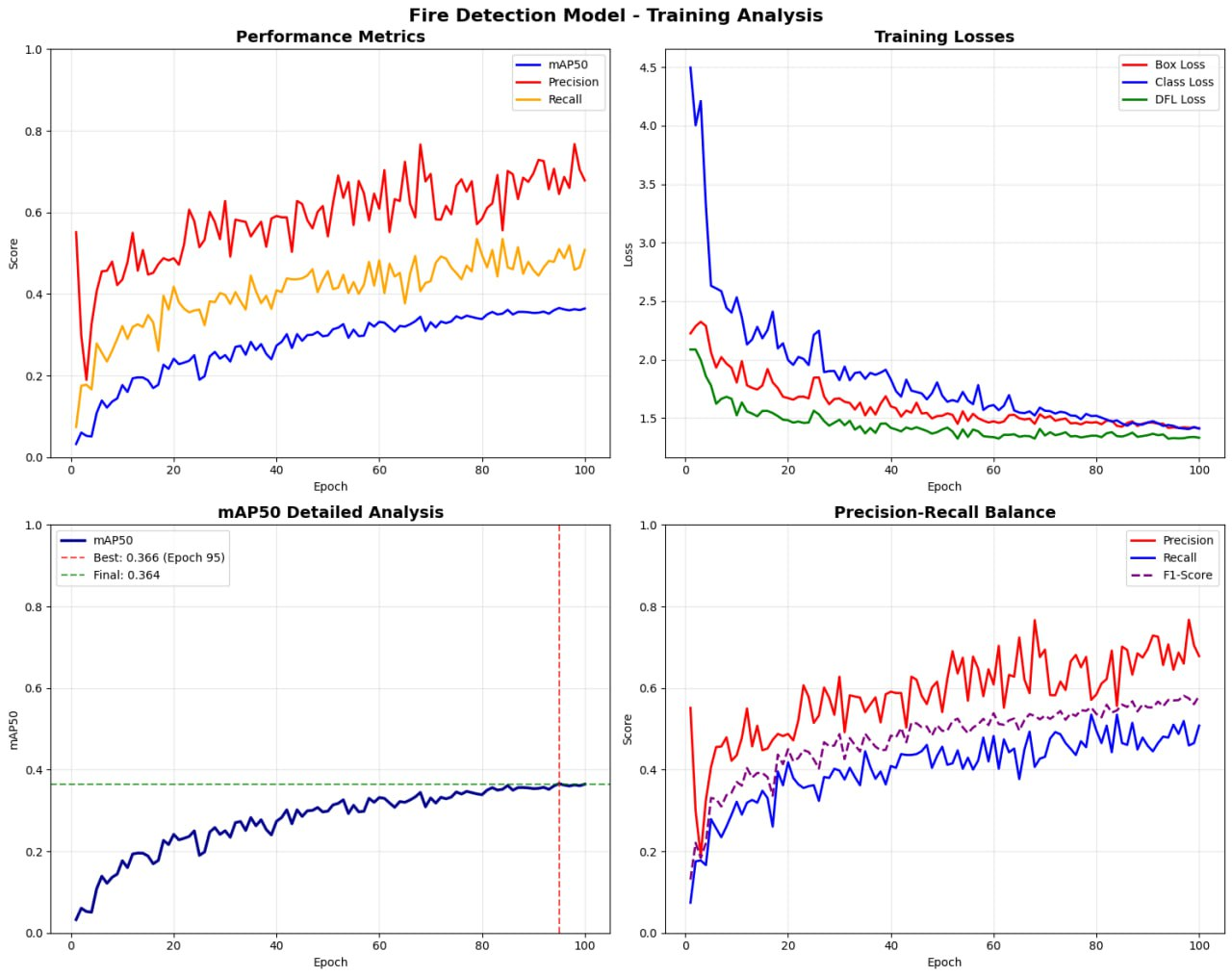
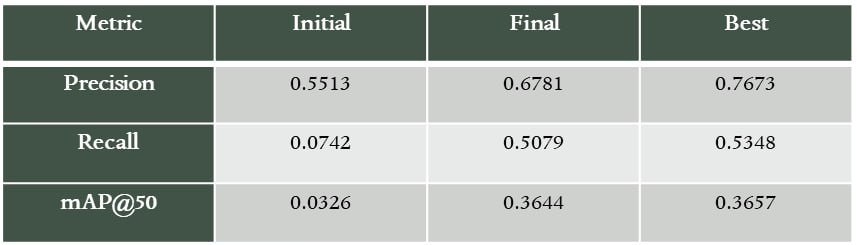
* 1. Box Loss: Measures localization error of predicted bounding boxes.
* 2. Classification Loss: Cross-entropy between predicted class and ground truth.
* 3. DFL Loss: Distribution Focal Loss for bounding box refinement.

**Evaluation Metrics:**

* Precision: Fraction of correct positive predictions.
* Recall: Fraction of true positives detected.
* mAP50: Mean Average Precision at 0.5 IoU threshold.
* mAP50-95: mAP averaged across IoU thresholds 0.5–0.95.
* F1-score: Harmonic mean of Precision and Recall.

**4.2.7 Model Visualization**



**4.2.8 Learning Curves & Scores**

**4.3 Attention U-Net**

**4.3.1 Model / Architecture**

Model type: Attention U-Net (custom implementation with 4 downsampling/upsampling layers and attention gates)

Input channels: in\_ch = 4 (RGB + Thermal or other 4-channel input)

Output channels: out\_ch = 1 (binary segmentation mask)

DoubleConv filters: [64, 128, 256, 512]

Attention blocks: 3 Attention Gates (att3, att2, att1) applied to skip connections

Upsample mode: bilinear

Skip connections: concatenation of attention-modulated feature maps

**4.3.2 Loss / Optimization**

Loss function: ComboLoss = 0.5 \* FocalLoss + 0.5 \* DiceLoss

FocalLoss hyperparameters: alpha = 0.25, gamma = 2

DiceLoss smoothing: smooth = 1

Optimizer: Adam

Learning rate: lr = 5e-4

**4.3.3 Training**

Number of epochs: num\_epochs = 100

Early stopping patience: patience = 15 (stop if no improvement in Dice for 15 epochs)

Device: GPU if available (cuda), else CPU

Batch size: dependent on train\_loader

Binary threshold: 0.5 (sigmoid output)

**4.3.4 Metrics**

Dice, IoU, Accuracy, Precision, Recall, F1-score.

**4.3.5 Workflow**

**1. Data Preparation**

* Dataset Setup
* Downloaded RGB-T wildfire dataset (ZIP files).
* Mounted Google Drive and extracted the data.
* Dataset structure: rgb/, nir/, mask/.
* Library Installation

-Deep learning: PyTorch, TensorFlow, Keras

- Image processing: OpenCV, scikit-image, imageio

- Data & visualization: numpy, pandas, matplotlib, seaborn, plotly, tqdm

* Augmentation: Albumentations
* Data Loading & Verification
* Loaded RGB, NIR, and mask images.
* Skipped missing/incomplete samples.
* Total valid samples: 1367
* Image shapes: RGB (420,420,3), (512,640,3); NIR & masks (512,640)
* Mask unique values: 0 (background), 255 (wildfire)

**2. Data Preprocessing & Augmentation**

* Augmentation Strategy (Albumentations)
* Training: horizontal/vertical flips, rotations, color jitter, Gaussian blur, resize, normalization
* Validation: resize and normalization only
* Custom Dataset Class
* Combined RGB + NIR → 4-channel input
* Applied augmentations dynamically
* Converted to PyTorch tensors [B,4,H,W]for images and [B,1,H,W] for masks
* Data Split & Loading
* Train: 80%, Validation: 20%
* Batch size: 4, shuffle enabled for trainingVisual Placeholder:
* Sample RGB/NIR/mask images before and after augmentation

**3. Model Design**

The proposed model is an Attention U-Net for binary image segmentation. It follows an encoder–decoder structure where the encoder extracts hierarchical features using Double Convolution blocks with max pooling, increasing the feature depth from 64 to 512 channels. The decoder restores spatial resolution using bilinear upsampling and feature fusion.

Attention gates are integrated into the skip connections to selectively emphasize relevant regions while suppressing background features. This improves segmentation accuracy, especially in complex scenes. The network accepts a 4-channel input and produces a single-channel segmentation mask using a 1×1 convolution followed by sigmoid activation.

**4. Training & Optimization**

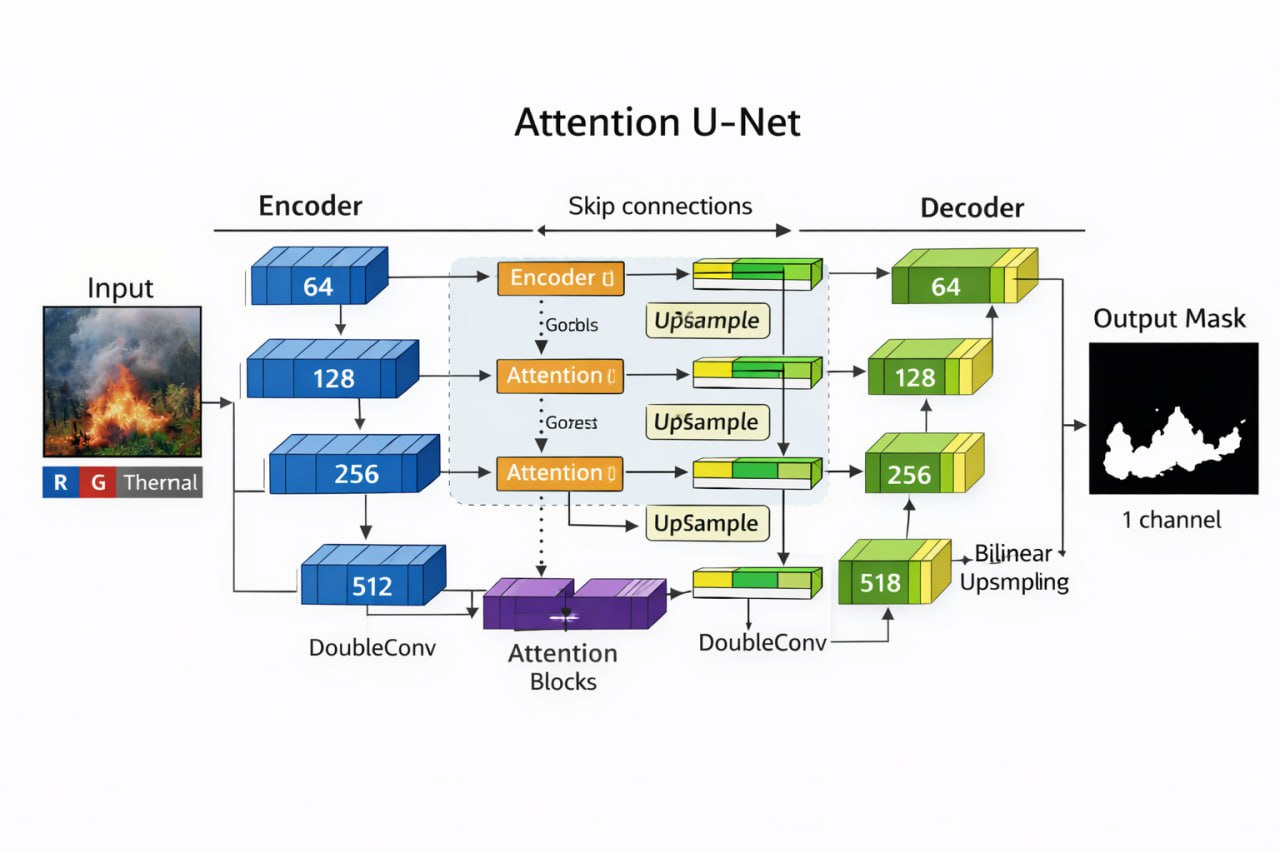
The model is trained using a combined loss function consisting of Focal Loss and Dice Loss to handle class imbalance and improve overlap accuracy. Optimization is performed using the Adam optimizer with a learning rate of 5×10⁻⁴.

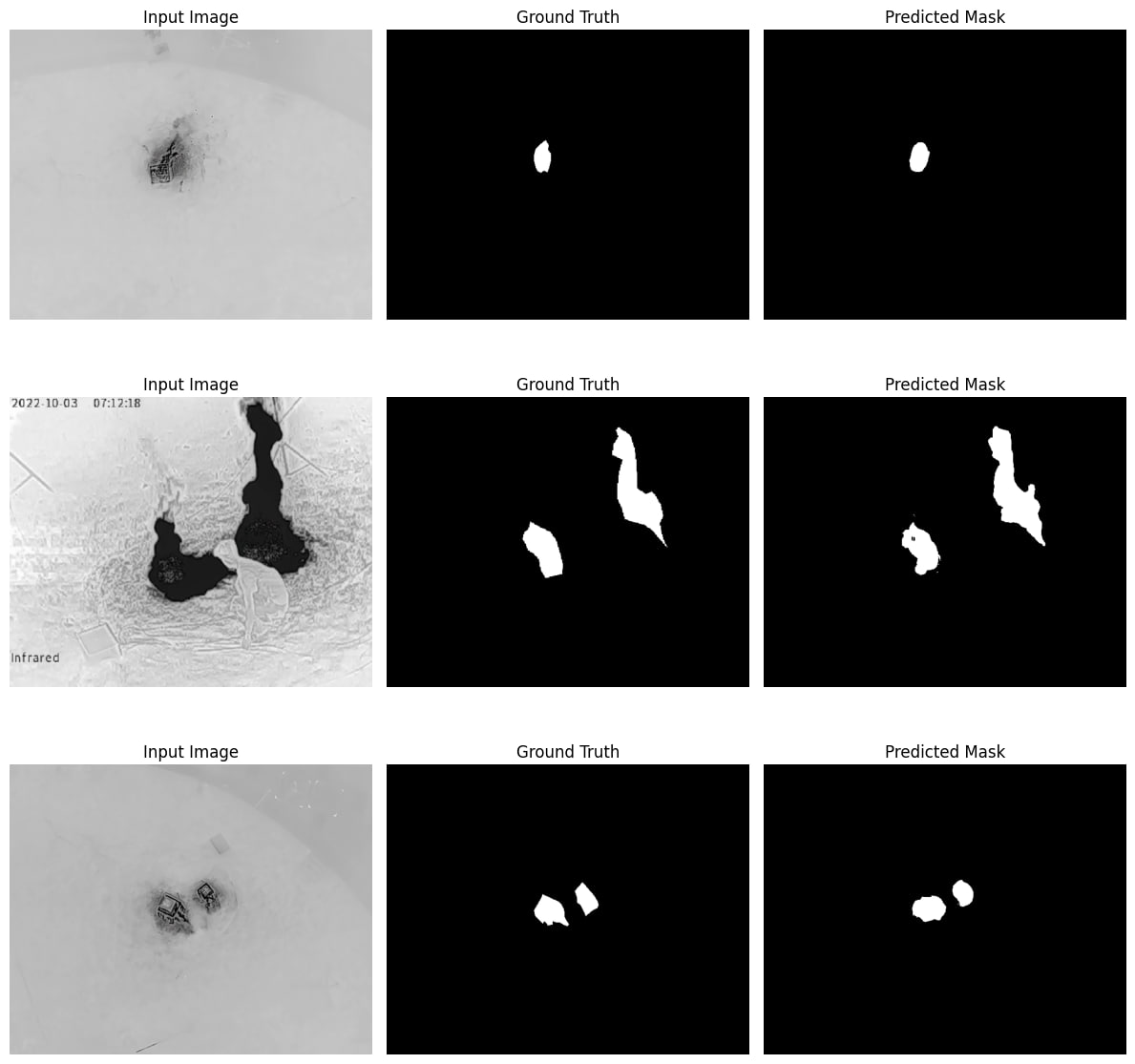
Training runs for up to 100 epochs with early stopping applied if validation Dice score does not improve for 15 epochs. The best-performing model is saved based on the highest validation Dice score.

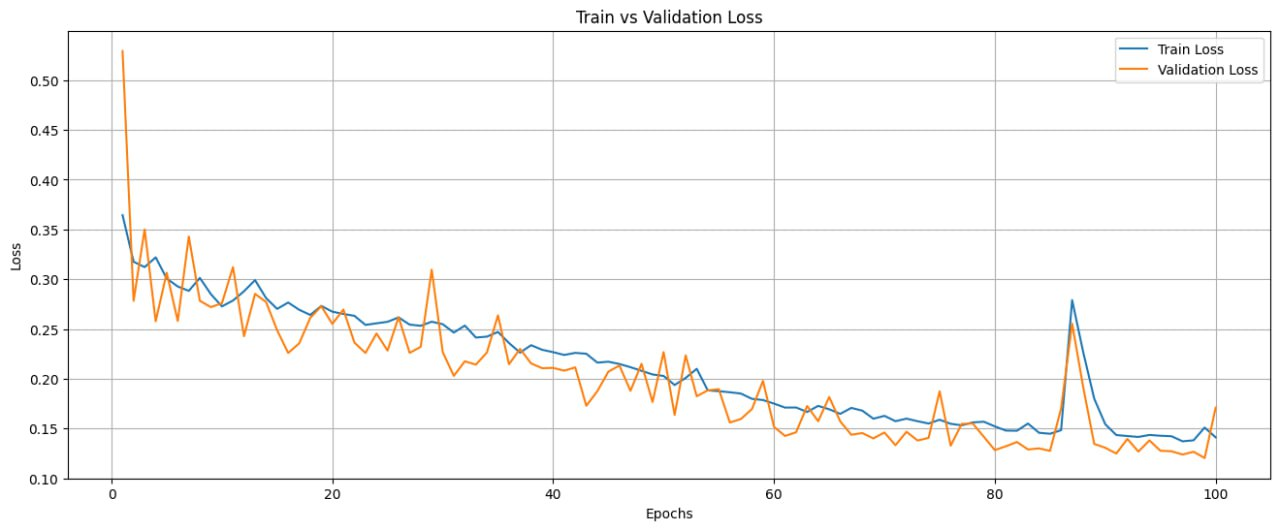
**5. Evaluation & Visualization**

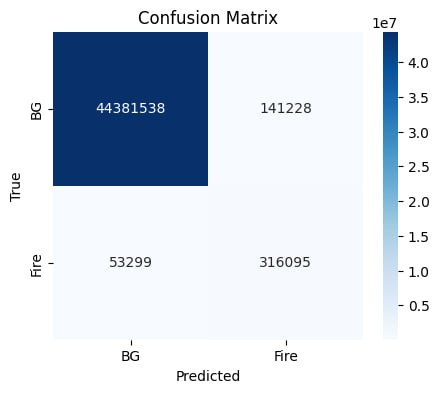
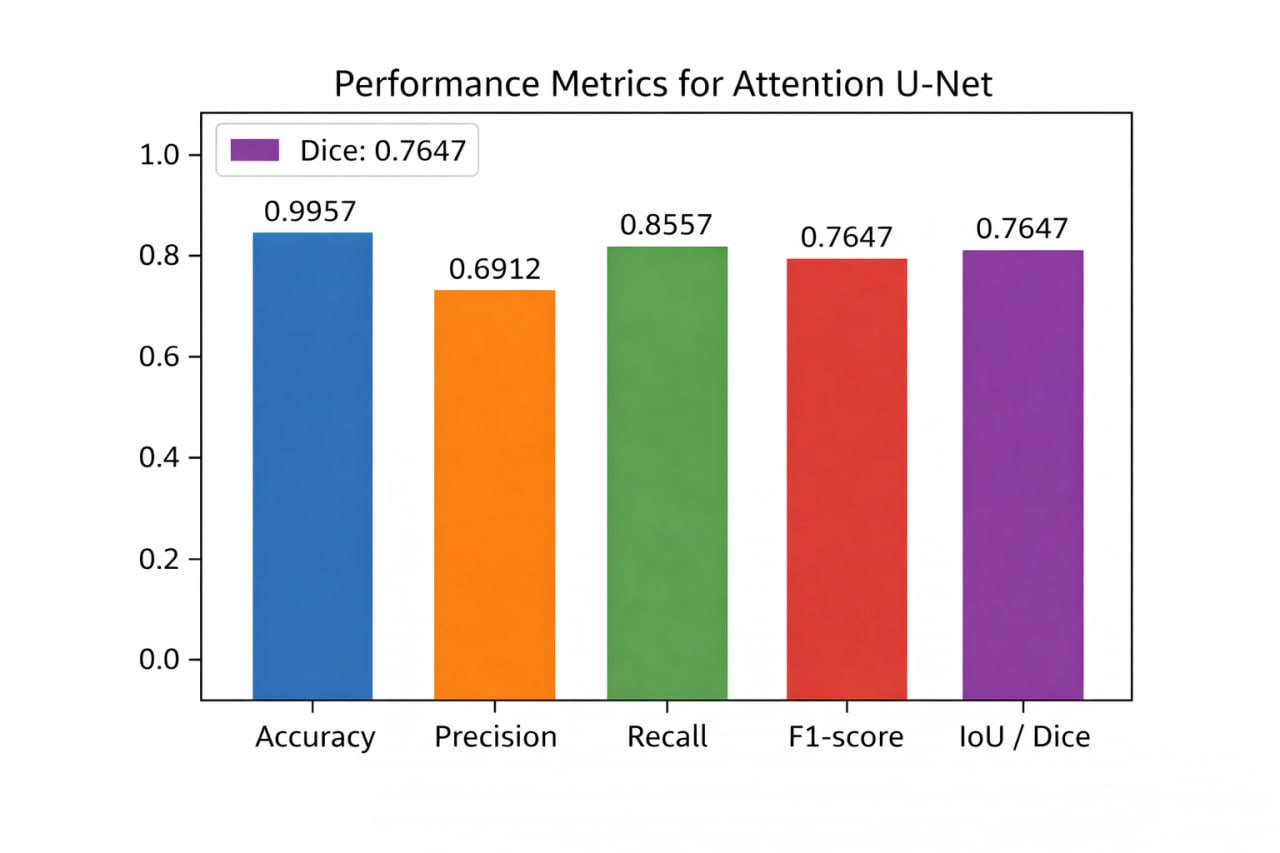
* Loaded best saved model
* Computed confusion matrix and all metrics
* Plotted training & validation loss, Dice, IoU curves
* Displayed sample inputs, ground truths, and predicted masks Visual Placeholder:
* Learning curves (loss, Dice, IoU)
* Prediction comparison (input, ground truth, prediction)

**4.3.6 Model Visualization**

**4.3.6.1 Architecture Visualize**

**4.3.6.2 Ground Truth vs Predicted Mask**

**4.3.6.3 Learning Curves ,Confusion Matrix & Scores**



**4.4 Models Analysis & Discussion**

**1. Comparative Analysis**

Among the segmentation models, U-Net demonstrates superior overall performance. It achieves higher precision, F1-score, and Intersection over Union (IoU), indicating more accurate fire boundary delineation and fewer false positives. Attention U-Net slightly outperforms U-Net in recall, suggesting improved sensitivity to fire regions; however, this comes at the cost of reduced precision and overall balance.

YOLOv8 shows substantial improvement throughout training, particularly in recall and mAP@50. This indicates strong learning capability and suitability for real-time fire localization tasks. Nevertheless, its performance remains inherently less precise than segmentation-based approaches due to the complexity of object-level detection.

**2. Most Important Metrics for Fire Applications**

For fire segmentation, recall is the most critical metric, as failing to detect fire pixels poses a significant safety risk. IoU or Dice score follows in importance, as these metrics evaluate the spatial accuracy of fire regions. The F1-score provides a balanced measure between false alarms and missed detections.

For fire detection, recall remains the top priority, ensuring fires are not overlooked. Mean Average Precision (mAP@50) reflects overall detection robustness, while precision is essential to minimize false fire alerts.

**3. Final Conclusion**

Based on the presented results, U-Net is the most effective model for fire segmentation, offering the best balance between accuracy and reliability. YOLOv8 is well-suited for real-time fire detection scenarios where rapid localization is required. For safety-critical fire monitoring systems, prioritizing recall alongside region-based metrics such as IoU or Dice is essential.

**4. How to Improve Results:**

• U-Net: adding BatchNorm, residual connections, and learning-rate schedulers.

• Attention U-Net: reduce attention sensitivity, adding dropout, and using pretrained encoders.

• YOLOv8: increasing image resolution, using stronger augmentation, and trying larger backbones.

**Conclusion**

This work demonstrates the effectiveness of deep learning–based segmentation and detection models for UAV wildfire monitoring using multimodal data. The use of U-Net and Attention U-Net architectures enabled accurate pixel-level fire segmentation, with the attention mechanism significantly improving focus on relevant fire regions while suppressing background noise. The YOLOv8 model further complemented the system by providing efficient real-time fire detection capabilities. Experimental results show high overall accuracy and strong Dice and IoU scores, confirming the robustness of the proposed approaches under challenging conditions. The fusion of RGB and thermal information proved crucial for early and reliable fire identification. Overall, the proposed models offer a scalable and practical solution for early wildfire detection and monitoring, supporting faster response and improved disaster management in real-world UAV applications.

**References**

[1]” A UAV-Based Multi-Scenario RGB-Thermal Dataset and Fusion Model for Enhanced Forest Fire Detection”

<https://www.mdpi.com/2072-4292/17/15/2593>

[2]” MCDet: Multi-Content Collaboration Detector for Multiscale Remote Sensing Object”

<https://ieeexplore.ieee.org/document/10418876>

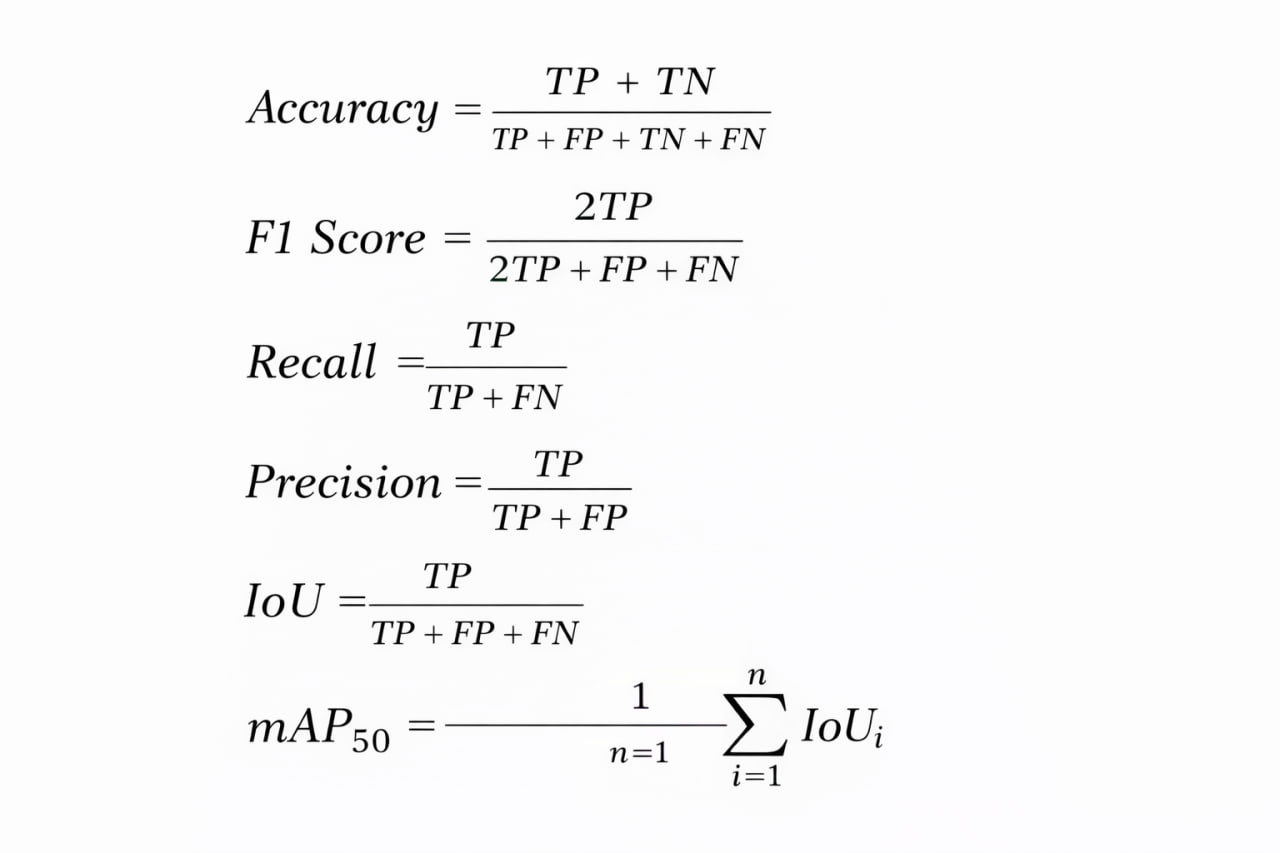
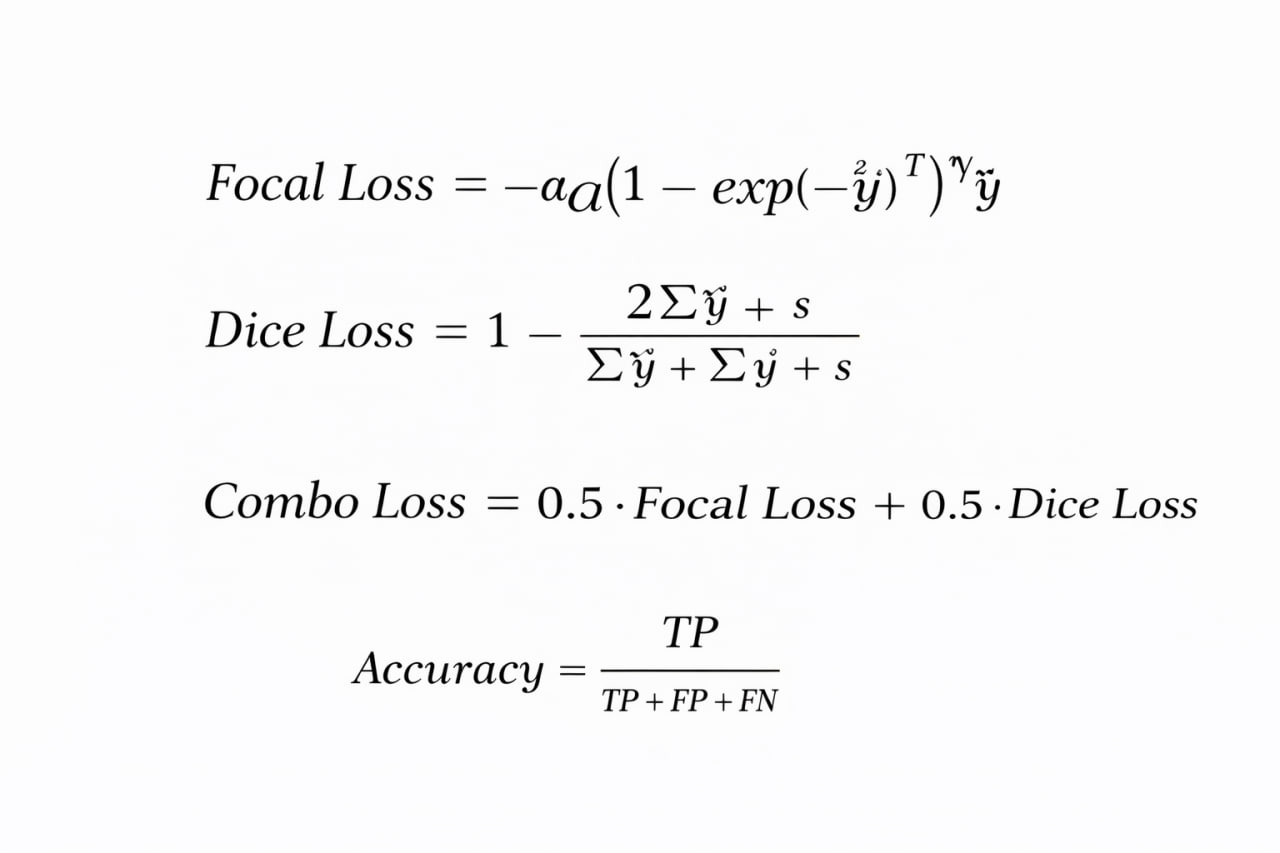
[3]” FireMan-UAV-RGBT (IEEE Paper 10710657)”

<https://ieeexplore.ieee.org/document/10710657>

[4]” U3UNet — Accurate and Reliable Forest Fire Segmentation for UAV Vision”

<https://ieeexplore.ieee.org/document/10944154>

1. **Appendix**

**Mathematical equations**

1. **Appendix**

**Terminology Table**

| **Term** | **Description** |
| --- | --- |
| UAV (Unmanned Aerial Vehicle) | An aerial platform used for remote sensing and wildfire monitoring with onboard cameras and sensors. |
| Wildfire Monitoring | Continuous observation of forest areas to detect, localize, and track fire or smoke outbreaks. |
| RGB Imaging | Visible-spectrum imaging using Red, Green, and Blue channels. |
| Thermal Imaging (Infrared) | Imaging modality that captures heat radiation, effective in smoke, darkness, and low visibility. |
| RGB–Thermal Fusion | Integration of RGB and thermal data to improve robustness and early fire detection. |
| Multimodal Data | Data originating from multiple sensor types (e.g., RGB + Thermal). |
| Image Segmentation | Pixel-level classification task that assigns each pixel to fire/smoke or background. |
| Binary Segmentation | Segmentation task with two classes: foreground (fire/smoke) and background. |
| Object Detection | Task of locating and classifying objects using bounding boxes. |
| U-Net | Encoder–decoder convolutional neural network designed for accurate image segmentation. |
| Attention U-Net | Modified U-Net incorporating attention gates to highlight relevant regions and suppress noise. |
| YOLOv8 | Real-time deep learning model for object detection and segmentation on edge devices. |
| Encoder | Downsampling network path that extracts hierarchical feature representations. |
| Decoder | Upsampling network path that reconstructs spatial resolution for segmentation masks. |
| Skip Connections | Direct links between encoder and decoder layers to preserve spatial information. |
| Attention Gate | Mechanism that weights feature maps based on relevance to the target region. |
| Feature Maps | Intermediate representations learned by convolutional layers. |
| Early Fusion | Combining RGB and thermal data at the input level before feature extraction. |
| Feature-Level Fusion | Combining features extracted from different modalities within the network. |
| Focal Loss | Loss function designed to address class imbalance by focusing on hard-to-classify samples. |
| Dice Loss | Overlap-based loss function optimized for segmentation tasks with imbalanced classes. |
| Combo Loss | Weighted combination of Focal Loss and Dice Loss used for stable segmentation training. |
| BCEWithLogitsLoss | Binary cross-entropy loss combined with a sigmoid activation. |
| Adam Optimizer | Adaptive optimization algorithm that adjusts learning rates per parameter. |
| Learning Rate | Hyperparameter controlling the magnitude of weight updates during training. |
| Weight Decay | Regularization technique to reduce overfitting by penalizing large weights. |
| Batch Size | Number of samples processed in one training iteration. |
| Epoch | One full pass over the entire training dataset. |
| Early Stopping | Training strategy that halts learning when validation performance stops improving. |
| Accuracy | Ratio of correctly classified pixels to the total number of pixels. |
| Precision | Proportion of predicted fire pixels that are actually fire. |
| Recall | Proportion of actual fire pixels that are correctly detected. |
| F1-Score | Harmonic mean of precision and recall, balancing false positives and false negatives. |
| IoU (Intersection over Union) | Ratio of overlap between predicted and ground-truth regions. |
| Dice Coefficient | Similarity metric measuring overlap between predicted and true segmentation masks. |
| mAP50 | Mean Average Precision at IoU threshold of 0.5, commonly used in object detection. |
| Ground Truth Mask | Pixel-level annotation indicating true fire or smoke regions. |
| Label Imbalance | Disproportion between fire and background pixels in wildfire datasets. |
| Real-Time Inference | Ability of a model to process data fast enough for live UAV deployment. |
| Edge Deployment | Running models on resource-constrained devices such as UAV onboard computers. |
| FireMan-UAV-RGBT Dataset | UAV-based RGB–Thermal wildfire dataset used for detection and segmentation research. |
| Dataset Generalization | Ability of a trained model to perform well on unseen environments. |
| False Positives | Background regions incorrectly classified as fire or smoke. |
| False Negatives | Fire regions missed by the model. |
| Early Wildfire Detection | Identification of fire at an initial stage before large-scale spread. |

Contents

[**Subtract** 2](#_Toc218703937)

[**Chapter One** 3](#_Toc218703938)

[**Introduction** 3](#_Toc218703939)

[**1.1** **Abstract** 3](#_Toc218703940)

[**1.2 Causes** 3](#_Toc218703941)

[**Chapter Two** 5](#_Toc218703942)

[**Reference Studies** 5](#_Toc218703943)

[**2.2 “MCDet: Target‑Aware Fusion for RGB‑T Fire Detection (2025)”** 8](#_Toc218703944)

[**2.2.1 Introduction** 8](#_Toc218703945)

[This study addresses the challenge of wildfire detection using aerial imagery, particularly under conditions of limited visibility such as dense smoke, low light, and heavy vegetation. While thermal or RGB images can be used individually, variations in illumination and thermal noise often reduce the stability and reliability of the results. To address these limitations, this work proposes a “Target-Aware Fusion” framework that integrates RGB and thermal channels to improve the robustness of early wildfire detection. 8](#_Toc218703946)

[**2.2.2 Methodology** 8](#_Toc218703947)

[Multidimensional Representation Collaborative Fusion (MRCF): 8](#_Toc218703948)

[This module models feature interactions across the thermal and RGB channels using a state-space model, forming a global relationship network between features. It also incorporates Deformable Convolutions to improve local detail sensing, especially for small or partially occluded fire regions. 8](#_Toc218703949)

[**Content-Guided Attention Network (CGAN):** 8](#_Toc218703950)

[After feature extraction, the system fuses multi-modal features using dynamic gating mechanisms, which determine which channels or features should dominate based on context (e.g., dense smoke or darkness). This reduces interference and errors caused by sunlight reflections in RGB images or thermal hotspots unrelated to fire. 9](#_Toc218703951)

[**Weighted IoU (WIoU) Loss:** 9](#_Toc218703952)

[Used to enhance the model’s ability to distinguish real fire regions from false positives such as heat reflections or background illumination, particularly in dense vegetation or bright backgrounds. 9](#_Toc218703953)

[**Experiments and Evaluation:** 9](#_Toc218703954)

[The model was tested on three wildfire datasets as well as a pedestrian dataset to evaluate the generalizability of the fusion approach. The main performance metric achieved was mean detection accuracy = 77.5%, outperforming the baseline methods reported in the literature. 9](#_Toc218703955)

[**2.2.3 Results and Performance** 9](#_Toc218703956)

[The MCDet model demonstrated reliable performance across multiple conditions. The study shows that integrating RGB-T fusion with attention mechanisms and the proposed architectural components improves operational stability under challenging conditions such as dense smoke and variable illumination. While the 77.5% accuracy is not the highest in general image detection tasks, it represents a significant advancement in multi-modal fusion for UAV-based wildfire detection. 9](#_Toc218703957)

[**2.2.4 Scientific Significance** 10](#_Toc218703958)

[This work introduces a dedicated fusion architecture for wildfire detection using RGB-T data, addressing modal ambiguity between channels. It focuses on real-world challenges—vegetation, lighting changes, and smoke—rather than simple synthetic scenarios. The framework provides a foundation for future work aimed at improving fusion strategies and reducing errors under real operational conditions. 10](#_Toc218703959)

[**2.2.5 Limitations and Future Work** 10](#_Toc218703960)

[The study does not provide full details on dataset sizes or scenario distributions, likely due to proprietary or institutional restrictions. Localization accuracy, particularly for small or evolving fire regions, remains an area for improvement. Future work should expand testing to diverse geographic environments and incorporate temporal tracking to capture fire progression over time. 10](#_Toc218703961)

[**2.3 FireMan-UAV-RGBT (IEEE Paper 10710657)** 10](#_Toc218703962)

[**2.4 “U3UNet — Accurate and Reliable Forest Fire Segmentation for UAV Vision”** 12](#_Toc218703963)

[**2.4.1 Introduction** 12](#_Toc218703964)

[Wildfires are a serious environmental threat due to their rapid spread and devastating impacts on ecosystems, property, and human lives. Traditional monitoring methods such as ground observation or satellite imagery often lack sufficient spatio‑temporal resolution for detailed fire shape tracking and early detection. UAVs equipped with high‑resolution cameras provide a promising platform for fine‑grained aerial monitoring, but accurate segmentation of fire regions remains challenging due to varying flame shapes and complex environments. To address these challenges, the authors propose U3UNet, a robust deep learning model specifically designed for segmentation of wildfire regions in UAV images. 12](#_Toc218703965)

[**2.4.2 Model Architecture and Method** 12](#_Toc218703966)

[U3UNet builds on the U‑Net family of encoder‑decoder networks, adopting a nested U‑shaped structure to fuse features across multiple scales and retain both global and local information. 12](#_Toc218703967)

[It includes full‑scale connections that balance high‑level contextual encoding with fine‑grained detail.Multi‑scale skip connections facilitate better feature fusion and help prevent segmentation misses or incorrect background inclusion. 13](#_Toc218703968)

[The architecture is optimized to handle variations in flame appearance and environmental clutter in UAV footage. 13](#_Toc218703969)

[The model was evaluated on both synthetic environments (using the Unreal Engine) and real forest fire scenes, allowing performance comparisons against other state‑of‑the‑art segmentation and detection models. 13](#_Toc218703970)

[**2.4.3 Results and Performance** 13](#_Toc218703971)

[Performance is assessed using a composite metric S designed to capture both detection accuracy and segmentation quality: 13](#_Toc218703972)

[In static scenarios, U3UNet achieved a composite score of 71.44%, which is close to the best comparative method. 13](#_Toc218703973)

[In dynamic scenarios, it scored 80.53%, 8.94% higher than the strongest competing approach. 13](#_Toc218703974)

[These results suggest that U3UNet is particularly effective at segmenting fire regions when the scene or viewpoint is changing, indicating robustness to real UAV motion and viewpoint variation. 13](#_Toc218703975)

[Real‑time tests of U3UNet deployed on edge computing devices mounted on UAVs demonstrated that the model can run with practical throughput levels, showing potential for onboard inference and real‑time wildfire monitoring. 14](#_Toc218703976)

[**2.4.4 Scientific Contribution** 14](#_Toc218703977)

[Advanced Segmentation Model: U3UNet improves over classical U‑Net variants by offering nested, multi‑scale fusion that better captures both global context and local details crucial for fire region delineation. 14](#_Toc218703978)

[Practical UAV Deployment: The model’s architecture and performance demonstrate suitability for real‑world, real‑time wildfire segmentation, including deployment on resource‑constrained UAV platforms. 14](#_Toc218703979)

[Comparative Validation: Through systematic experiments, U3UNet outperformed or matched leading segmentation models in dynamic aerial firefighting contexts, highlighting its reliability in complex forest fire monitoring tasks. 14](#_Toc218703980)

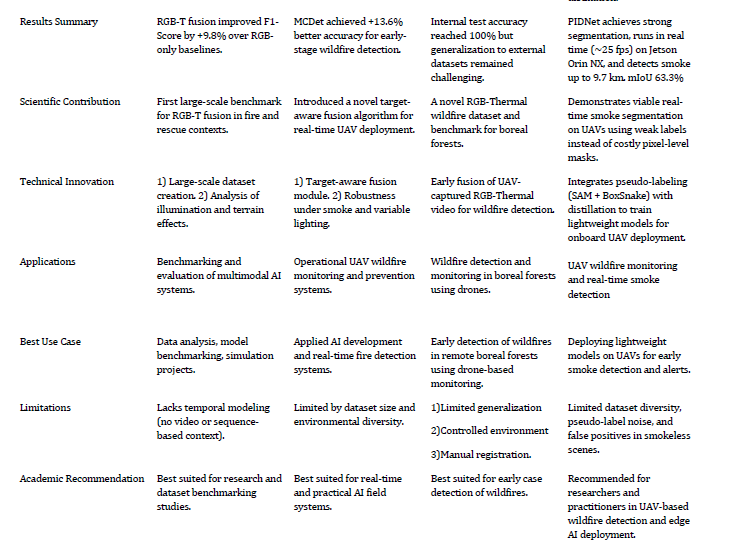
[**2.4.5 Limitations and Future Directions** 14](#_Toc218703981)

[The reported metric demonstrates strong performance but also indicates scope for improvement in static scenes where results were slightly below the best baseline model. 14](#_Toc218703982)

[Future work may explore hybrid architectures with attention mechanisms or multi‑modal inputs (e.g., thermal + RGB) to further improve segmentation under challenging conditions such as smoke, shadows, or occlusions. 15](#_Toc218703983)

[https://ieeexplore.ieee.org/document/10944154 15](#_Toc218703984)

[**2.5 Papers Comparison** 16](#_Toc218703985)

[ 17](#_Toc218703986)

[**Chapter Three** 18](#_Toc218703987)

[**DataSets OverView & Comparison** 18](#_Toc218703988)

[**3.1** **RGB-T 3M** 18](#_Toc218703989)

[**3.1.1 Size & Distribution** This dataset, hosted by the University of Science and Technology of China (USTC), includes approximately **1,367** RGB–Thermal paired images. It is referenced in wildfire detection research as a compact but valuable benchmark dataset for multimodal fire analysis. 18](#_Toc218703990)

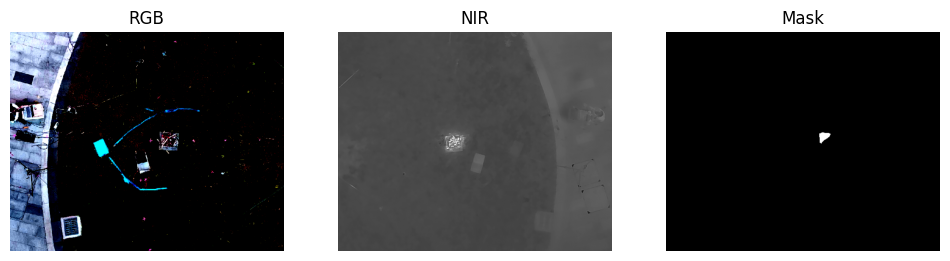
[**3.1.2 Modalities & Annotation** The dataset provides paired RGB and thermal images, captured under varied illumination and fire intensity conditions. Annotations support semantic segmentation tasks, labeling fire and non-fire regions, enabling multimodal training for deep learning models. 18](#_Toc218703991)

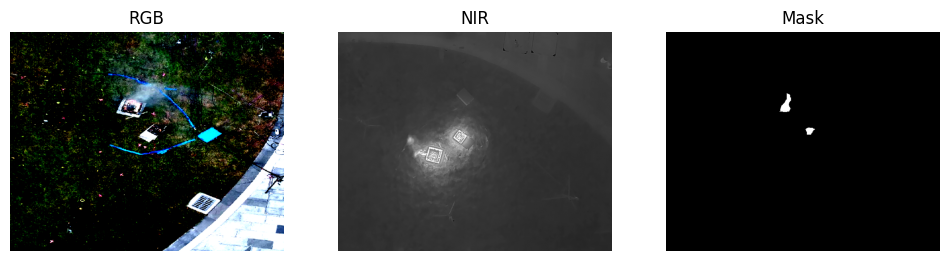
[**3.1.3 Dataset Specifics** **-** Format: Image pairs (RGB + Thermal) **-** Approximately total samples: 1,367 **-** Target task**:** Fire detection and segmentation **-** Annotation type**:** Pixel-level segmentation masks **-** Data balance**:** Fire vs non-fire, collected in both day and night settings 18](#_Toc218703992)

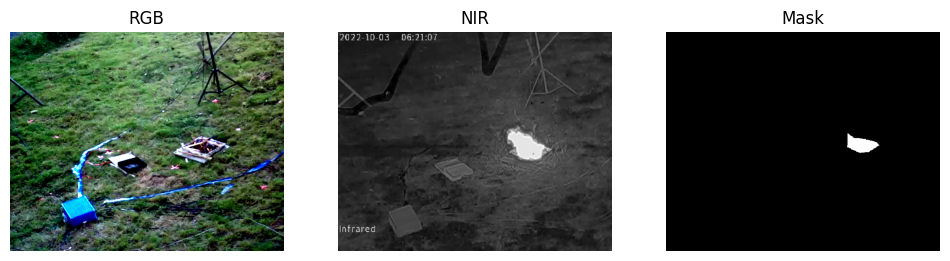
[**3.1.4 Use-Cases & Research Impact** It supports tasks such as wildfire detection, multimodal fusion learning, and real-time UAV-based thermal imaging analysis. Commonly used in comparative studies alongside datasets like FLAME-1 and FireMan-UAV-RGBT. 19](#_Toc218703993)

[**3.1.5 Remarks & Limitations** The dataset’s modest scale limits generalization but offers a valuable baseline for fusion-based fire detection. Access may require university credentials or approval. 19](#_Toc218703994)

[**3.1.6** **OverView** 19](#_Toc218703995)

[ 19](#_Toc218703996)

[ 19](#_Toc218703997)

[ 20](#_Toc218703998)

[**3.1.7** **Reference Link** 20](#_Toc218703999)

[https://rec.ustc.edu.cn/share/75aae550-8d26-11ee-98a1-ed4fd857523e 20](#_Toc218704000)

[**3.2 FireMan‑UAV‑RGBT** 20](#_Toc218704001)

[**3.2.1 Size and Distribution** 20](#_Toc218704002)

[Total dataset size: ~39.4 GB. 20](#_Toc218704003)

[34 RGB videos + 20 paired RGB‑Thermal videos. 20](#_Toc218704004)

[Captured across four controlled burn sites in Finnish boreal forests. 20](#_Toc218704005)

[Covers varied flight paths, altitudes, and scene complexity. 20](#_Toc218704006)

[**3.2.2 Modality & Annotations** 20](#_Toc218704007)

[Modality: RGB + Thermal synchronized video. 20](#_Toc218704008)

[Annotations: YOLO-format bounding boxes for fire and related objects. 20](#_Toc218704009)

[Both manual and semi‑automatic annotation methods applied. 20](#_Toc218704010)

[**3.2.3 Dataset Specifics** 20](#_Toc218704011)

[Video resolution: High-resolution UAV footage. 21](#_Toc218704012)

[Multi-environment coverage: Forests, vegetation, varied lighting conditions. 21](#_Toc218704013)

[Includes sample frames and labeling reference files for model training. 21](#_Toc218704014)

[**3.2.4 Use Cases and Research Impact** 21](#_Toc218704015)

[Primary use: Training and benchmarking UAV-based wildfire detection models. 21](#_Toc218704016)

[Supports multimodal fusion research (RGB + Thermal). 21](#_Toc218704017)

[Enables development of early wildfire detection systems, environmental monitoring, and UAV operational testing. 21](#_Toc218704018)

[**3.2.5 Remarks and Limitations** 21](#_Toc218704019)

[Strengths: Real UAV footage, synchronized RGB-Thermal, YOLO-compatible annotations. 21](#_Toc218704020)

[Limitations: Data comes from controlled burns, so real wildfire variability may be underrepresented. 21](#_Toc218704021)

[Dataset is moderate in size compared to large benchmarks but rich in temporal video sequences. 21](#_Toc218704022)

[**3.2.6 Reference Link** 21](#_Toc218704023)

[Zenodo dataset: https://zenodo.org/records/13732947 21](#_Toc218704024)

[IEEE Paper: https://ieeexplore.ieee.org/document/10710657 21](#_Toc218704025)

[**3.3 DataSet Comparison** 22](#_Toc218704026)

[**Chapter Four** 23](#_Toc218704027)

[**4.1 U-Net** 23](#_Toc218704028)

[**4.1.1 Model / Architecture** 23](#_Toc218704029)

[Model type: U-Net (custom implementation with 4 downsampling and upsampling layers) 23](#_Toc218704030)

[Input channels: in\_ch = 4 (RGB + Thermal or other 4-channel input) 23](#_Toc218704031)

[Output channels: out\_ch = 1 (binary segmentation mask) 23](#_Toc218704032)

[DoubleConv filters: [64, 128, 256, 512] 23](#_Toc218704033)

[Upsample mode: bilinear 23](#_Toc218704034)

[Skip connections: concatenation with corresponding encoder features 23](#_Toc218704035)

[**4.1.2 Loss / Optimization** 23](#_Toc218704036)

[Loss function: ComboLoss = 0.5 \* FocalLoss + 0.5 \* DiceLoss 23](#_Toc218704037)

[FocalLoss hyperparameters: alpha = 0.25, gamma = 2 23](#_Toc218704038)

[DiceLoss smoothing: smooth = 1 23](#_Toc218704039)

[Optimizer: Adam 23](#_Toc218704040)

[Learning rate: 5e-4 23](#_Toc218704041)

[**4.1.3 Training** 24](#_Toc218704042)

[Number of epochs: num\_epochs = 100 24](#_Toc218704043)

[Early stopping patience: patience = 15 (stop if no improvement in Dice for 15 epochs) 24](#_Toc218704044)

[Device: GPU if available (cuda), else CPU 24](#_Toc218704045)

[Batch size: Not explicitly defined; depends on train\_loader 24](#_Toc218704046)

[Binary threshold: 0.5 (sigmoid output) 24](#_Toc218704047)

[**4.1.4 Metrics** 24](#_Toc218704048)

[Dice, IoU, Accuracy, Precision, Recall, F1-score 24](#_Toc218704049)

[**4.1.5 Workflow** 24](#_Toc218704050)

[**1. Data Preparation** 24](#_Toc218704051)

[ Dataset Setup 24](#_Toc218704052)

[ Downloaded RGB-T wildfire dataset (ZIP files). 24](#_Toc218704053)

[ Mounted Google Drive and extracted the data. 24](#_Toc218704054)

[ Dataset structure: rgb/, nir/, mask/. 24](#_Toc218704055)

[ Library Installation 24](#_Toc218704056)

[-Deep learning: PyTorch, TensorFlow, Keras 24](#_Toc218704057)

[- Image processing: OpenCV, scikit-image, imageio 24](#_Toc218704058)

[- Data & visualization: numpy, pandas, matplotlib, seaborn, plotly, tqdm 24](#_Toc218704059)

[ Augmentation: Albumentations 24](#_Toc218704060)

[ Data Loading & Verification 25](#_Toc218704061)

[ Loaded RGB, NIR, and mask images. 25](#_Toc218704062)

[ Skipped missing/incomplete samples. 25](#_Toc218704063)

[ Total valid samples: 1367 25](#_Toc218704064)

[ Image shapes: RGB (420,420,3), (512,640,3); NIR & masks (512,640) 25](#_Toc218704065)

[ Mask unique values: 0 (background), 255 (wildfire) 25](#_Toc218704066)

[**2. Data Preprocessing & Augmentation** 25](#_Toc218704067)

[ Augmentation Strategy (Albumentations) 25](#_Toc218704068)

[ Training: horizontal/vertical flips, rotations, color jitter, Gaussian blur, resize, normalization 25](#_Toc218704069)

[ Validation: resize and normalization only 25](#_Toc218704070)

[ Custom Dataset Class 25](#_Toc218704071)

[ Combined RGB + NIR → 4-channel input 25](#_Toc218704072)

[ Applied augmentations dynamically 25](#_Toc218704073)

[ Converted to PyTorch tensors [B,4,H,W]for images and [B,1,H,W] for masks 25](#_Toc218704074)

[ Data Split & Loading 25](#_Toc218704075)

[ Train: 80%, Validation: 20% 25](#_Toc218704076)

[ Batch size: 4, shuffle enabled for trainingVisual Placeholder: 25](#_Toc218704077)

[ Sample RGB/NIR/mask images before and after augmentation 25](#_Toc218704078)

[**3. Model Design** 25](#_Toc218704079)

[ Architecture: Custom 4-channel UNet 25](#_Toc218704080)

[ Encoder: 4 downsampling blocks 25](#_Toc218704081)

[ Bottleneck + decoder with skip connections 26](#_Toc218704082)

[ Output: single-channel mask 26](#_Toc218704083)

[ Loss Function: ComboLoss = 0.5 \* Focal Loss + 0.5 \* Dice Loss 26](#_Toc218704084)

[ Metrics: Dice coefficient, IoU, Accuracy, Precision, Recall, F1-score 26](#_Toc218704085)

[**4. Training & Optimization** 26](#_Toc218704086)

[ - Optimizer: Adam (lr=5e-4) 26](#_Toc218704087)

[ - Epochs: 100, Early stopping patience: 15 26](#_Toc218704088)

[ - GPU acceleration (CUDA) 26](#_Toc218704089)

[ - Training loop: 26](#_Toc218704090)

[1. Forward pass → predictions 26](#_Toc218704091)

[2. Compute loss 26](#_Toc218704092)

[3. Backpropagation & optimizer step 26](#_Toc218704093)

[4. Compute metrics for train & validation 26](#_Toc218704094)

[ Early stopping monitored on best Dice coefficient 26](#_Toc218704095)

[Best model saved as best\_model.pth 26](#_Toc218704096)

[**5. Evaluation & Visualization** 26](#_Toc218704097)

[ Loaded best saved model 26](#_Toc218704098)

[ Computed confusion matrix and all metrics 26](#_Toc218704099)

[ Plotted training & validation loss, Dice, IoU curves 26](#_Toc218704100)

[ Displayed sample inputs, ground truths, and predicted masks Visual Placeholder: 26](#_Toc218704101)

[ Learning curves (loss, Dice, IoU) 26](#_Toc218704102)

[ Prediction comparison (input, ground truth, prediction) 26](#_Toc218704103)

[**4.1.5 Model Visualization** 27](#_Toc218704104)

[**4.1.5.1 Architecture Visualize** 27](#_Toc218704105)

[ Input (4 channels) 27](#_Toc218704106)

[ Down path: DoubleConv blocks with [64, 128, 256, 512] 27](#_Toc218704107)

[ MaxPooling for downsampling 27](#_Toc218704108)

[ Up path: Upsample + DoubleConv blocks [512→256→128→64] 27](#_Toc218704109)

[ Skip connections from encoder to decoder 27](#_Toc218704110)

[ Output: 1 channel (binary mask) 27](#_Toc218704111)

[**4.2.1 Model / Architecture** 31](#_Toc218704112)

[**1. Data Preparation** 39](#_Toc218704113)

[ Dataset Setup 39](#_Toc218704114)

[ Downloaded RGB-T wildfire dataset (ZIP files). 39](#_Toc218704115)

[ Mounted Google Drive and extracted the data. 39](#_Toc218704116)

[ Dataset structure: rgb/, nir/, mask/. 39](#_Toc218704117)

[ Library Installation 39](#_Toc218704118)

[-Deep learning: PyTorch, TensorFlow, Keras 39](#_Toc218704119)

[- Image processing: OpenCV, scikit-image, imageio 39](#_Toc218704120)

[- Data & visualization: numpy, pandas, matplotlib, seaborn, plotly, tqdm 39](#_Toc218704121)

[ Augmentation: Albumentations 40](#_Toc218704122)

[ Data Loading & Verification 40](#_Toc218704123)

[ Loaded RGB, NIR, and mask images. 40](#_Toc218704124)

[ Skipped missing/incomplete samples. 40](#_Toc218704125)

[ Total valid samples: 1367 40](#_Toc218704126)

[ Image shapes: RGB (420,420,3), (512,640,3); NIR & masks (512,640) 40](#_Toc218704127)

[ Mask unique values: 0 (background), 255 (wildfire) 40](#_Toc218704128)

[**2. Data Preprocessing & Augmentation** 40](#_Toc218704129)

[ Augmentation Strategy (Albumentations) 40](#_Toc218704130)

[ Training: horizontal/vertical flips, rotations, color jitter, Gaussian blur, resize, normalization 40](#_Toc218704131)

[ Validation: resize and normalization only 40](#_Toc218704132)

[ Custom Dataset Class 40](#_Toc218704133)

[ Combined RGB + NIR → 4-channel input 40](#_Toc218704134)

[ Applied augmentations dynamically 40](#_Toc218704135)

[ Converted to PyTorch tensors [B,4,H,W]for images and [B,1,H,W] for masks 40](#_Toc218704136)

[ Data Split & Loading 40](#_Toc218704137)

[ Train: 80%, Validation: 20% 40](#_Toc218704138)

[ Batch size: 4, shuffle enabled for trainingVisual Placeholder: 40](#_Toc218704139)

[ Sample RGB/NIR/mask images before and after augmentation 40](#_Toc218704140)

[**3. Model Design** 41](#_Toc218704141)

[The proposed model is an Attention U-Net for binary image segmentation. It follows an encoder–decoder structure where the encoder extracts hierarchical features using Double Convolution blocks with max pooling, increasing the feature depth from 64 to 512 channels. The decoder restores spatial resolution using bilinear upsampling and feature fusion. 41](#_Toc218704142)

[Attention gates are integrated into the skip connections to selectively emphasize relevant regions while suppressing background features. This improves segmentation accuracy, especially in complex scenes. The network accepts a 4-channel input and produces a single-channel segmentation mask using a 1×1 convolution followed by sigmoid activation. 41](#_Toc218704143)

[**4. Training & Optimization** 41](#_Toc218704144)

[The model is trained using a combined loss function consisting of Focal Loss and Dice Loss to handle class imbalance and improve overlap accuracy. Optimization is performed using the Adam optimizer with a learning rate of 5×10⁻⁴. 41](#_Toc218704145)

[Training runs for up to 100 epochs with early stopping applied if validation Dice score does not improve for 15 epochs. The best-performing model is saved based on the highest validation Dice score. 41](#_Toc218704146)

[**5. Evaluation & Visualization** 41](#_Toc218704147)

[ Loaded best saved model 41](#_Toc218704148)

[ Computed confusion matrix and all metrics 41](#_Toc218704149)

[ Plotted training & validation loss, Dice, IoU curves 41](#_Toc218704150)

[ Displayed sample inputs, ground truths, and predicted masks Visual Placeholder: 41](#_Toc218704151)

[ Learning curves (loss, Dice, IoU) 41](#_Toc218704152)

[ Prediction comparison (input, ground truth, prediction) 41](#_Toc218704153)

[[3]” FireMan-UAV-RGBT (IEEE Paper 10710657)” 47](#_Toc218704154)

[https://ieeexplore.ieee.org/document/10710657 47](#_Toc218704155)

[https://ieeexplore.ieee.org/document/10944154 47](#_Toc218704156)

[**(1)** **Appendix** 48](#_Toc218704157)

[**(2)** **Appendix** 49](#_Toc218704158)

[**Terminology Table** 49](#_Toc218704159)