ISPRS Journal of Photogrammetry and Remote Sensing 235 (2026) 551–564

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

Cross-modal distillation for real-time wildfire detection and localization in edge-deployed aerial vehicles

Medhavi Mishra [a], Sumit Mishra [b], Hyo-Sang Shin [a,c,*]

[a] CCS Graduate School of Mobility, Korea Advanced Institute of Science and Technology, Daejeon, South Korea
[b] Robotics Program, Korea Advanced Institute of Science and Technology, Daejeon, South Korea
[c] Centre for AI, Robotics and Space, FEAS, Cranfield University, Cranfield, United Kingdom

ARTICLE INFO

Keywords:
Cross-modal knowledge distillation
Wildfire detection
Edge-AI AAVs
Patch-level fire localization

ABSTRACT

Wildfire detection and localization in aerial imagery is critical for rapid response and damage mitigation. Autonomous aerial vehicles (AAVs) enable large area monitoring but face real-time processing challenges due to limited onboard computational and sensor resources. This work introduces a cross-modal knowledge distillation framework for edge-deployed AAVs. A teacher network trained only on thermal images transfers semantic and spatial representations to an optical image based student network when trained in an offline fashion using thermal and optical image pairs. During deployment, the student uses only optical images, thus reducing reliance on multi-sensor payloads while maintaining high detection accuracy. The student model incorporates dual classification heads: an image-level head for fire-free vs. fire-impacted scenes, and a patch-level head for flame vs. no-flame discrimination. This patch-level strategy provides effective fire localization while avoiding the computational overhead of segmentation, making it practical for resource-constrained deployment. Evaluated on aerial wildfire dataset, the framework achieves 90.97% patch-level accuracy, with false alarm and missed detection rates of 8.82% and 14.78%, respectively. The lightweight student model requires only 2.99 GFLOPS with inference time of 0.004s and generates patch-level probability heatmaps for fire region localization. Unlike conventional unimodal systems, this approach leverages thermal-to-optical knowledge transfer to deliver high accuracy, low latency, and precise localization under edge-computing constraints. The code and dataset will be released at https://github.com/medh132/cmkd.

1. Introduction

Wildfire is a growing global threat, causing catastrophic environmental damage and significant loss of life and property. The timely and accurate detection of these events is paramount for effective mitigation and response. To this end, edge-AI is an emerging technology that offers a transformative solution (Su et al., 2022). By deploying machine learning (ML) and deep learning (DL) models directly on edge devices like autonomous aerial vehicles (AAVs) or unmanned aerial vehicles (UAVs), edge-AI eliminates the latency of cloud-based processing (Mahdi and Mahmood, 2022). This on-device intelligence enables real-time data analysis and instant anomaly detection, allowing AAVs to identify and respond to wildfires with unprecedented speed. This capability is crucial for triggering autonomous actions, such as deploying fire-suppression measures, directly in the field (Fouda et al., 2022). This paradigm shift from centralized cloud computing to decentralized, on-device intelligence is key to creating self-sufficient systems for wildfire management in remote and challenging environments.

Though AAVs enable large-area monitoring, they face real-time processing challenges due to limited onboard computational and sensor resources. These devices operate under severe power and compute constraints, lacking the robust hardware (e.g., high-performance GPUs) and extensive power sources available in cloud data centers. This necessitates the use of highly optimized, lightweight ML/DL models that can perform inference with minimal computational overhead. Energy efficiency is also a paramount concern, as every computational cycle directly impacts the AAV's finite battery life. Hence, such models must be designed not only for accuracy but also for speed and minimal computational complexity to maximize operational endurance, a critical factor for long-duration missions like wildfire localization (Fouda et al., 2022). Therefore, the core challenge lies in developing a methodology that balances the need for accurate, real-time perception with the stringent resource limitations of the edge platform (Su et al., 2022).

Traditional computer vision approaches, including pixel-level segmentation and object detection, encounter significant challenges in

* Corresponding author at: Centre for AI, Robotics and Space, FEAS, Cranfield University, Cranfield, United Kingdom.
E-mail address: hyosangshin@kaist.ac.kr (H.-S. Shin).

https://doi.org/10.1016/j.isprsjprs.2026.03.019
Received 30 October 2025; Received in revised form 6 February 2026; Accepted 14 March 2026
Available online 24 March 2026
0924-2716/© 2026 The Authors. Published by Elsevier B.V. on behalf of International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS). This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

wildfire detection (Ghali and Akhloufi, 2023b). Pixel-wise segmentation is computationally expensive and less suited to the highly dynamic, amorphous nature of fire (Feng et al., 2025; Ribeiro et al., 2023; Li et al., 2025). Bounding-box–based object detection similarly struggles with the large variability in fire size and shape, while region proposal mechanisms (Ren et al., 2015) further increase computational cost during both training and inference (Setyanto et al., 2023). These limitations underscore the need for more efficient detection strategies, particularly for deployment in resource-constrained environments.

Additionally, in real-time wildfire monitoring aerial platforms due to limited onboard computation and payload capacity, multi-sensor systems prove to be costly and often impractical. Hence, thermal-guided distillation addresses this challenge by leveraging thermal information during training to enhance optical-only models at inference, significantly reducing hardware and computational overhead while maintaining competitive performance (Bouguettaya et al., 2022).

In this regard, a patch-level classification strategy leveraging cross-modal knowledge distillation offers a more pragmatic solution. This method segments the image into fixed-size patches and classifies each patch as either containing flame or not. This approach drastically reduces computational overhead, as the model performs a simpler classification task rather than a complex localization or masking. Hence, by avoiding the need to precisely define the irregular boundaries of the fire, patch-level detection provides a more robust, resource-efficient, and practical solution for real-time wildfire localization.

Further, wildfire localization in aerial imagery presents unique challenges compared to terrestrial sensing and remains relatively underexplored. Despite offering large-scale situational awareness, aerial detection is hindered by lower spatial resolution, occlusions, illumination artifacts, and the scarcity of standardized, densely annotated datasets (Ramos et al., 2024). In light of the challenges posed by limited on-board sensors and the difficulties of aerial imagery, the primary research objective of this work is to develop a lightweight computer vision model for real-time wildfire detection. This involves creating a model with low latency and computational complexity, specifically compatible for edge deployment. The key contributions of this work are as follows:

1. A patch-level wildfire localization framework for fine-grained predictions in aerial imagery is proposed; the patch-level design is specifically adopted to enable a lightweight model for resource-constrained Edge-AI deployment on AAV/UAV platforms.
2. A comprehensive training strategy is introduced in which cross-modal knowledge distillation is employed, incorporating contrastive loss with text anchors to enhance feature learning. During inference, only optical images are utilized, keeping the process independent of text and thermal modalities.
3. A patch-level annotated aerial image dataset using FLAME2 is created for wildfire-flame localization.

2. Related works

2.1. Aerial wildfire detection

Wildfire localization in aerial imagery presents challenges that differ fundamentally from terrestrial sensing. While ground-based systems benefit from close-range, high-resolution observations, their coverage is inherently limited. In contrast, aerial platforms enable large-scale monitoring of fire extent and progression across remote and inaccessible regions (Ghali and Akhloufi, 2023b). However, this advantage comes at the cost of reduced spatial resolution and increased visual ambiguity. Aerial wildfire detection is further complicated by canopy occlusion, sun-induced artifacts, and the scarcity of standardized datasets with dense and reliable annotations (Bouguettaya et al., 2022; Ghali and Akhloufi, 2023b; Ramos et al., 2024).

Deep learning–based approaches dominate recent advances in wildfire detection, with most methods adopting either semantic segmentation or object detection paradigms. Segmentation-based techniques aim to identify fire-affected regions at the pixel level (Jonnalagadda and Hashim, 2024; Wang et al., 2024; Feng et al., 2025), offering fine-grained spatial detail. Despite their accuracy, such methods are computationally expensive and less suited to aerial wildfire monitoring, where fire boundaries are highly dynamic, fragmented, and irregular. Object detection frameworks, on the other hand, localize fire regions using bounding boxes (Jiao et al., 2019; Yang et al., 2023; Gonçalves et al., 2024), providing a more efficient and operationally viable solution for aerial platforms. Their practicality is further enhanced when combined with model compression and optimization strategies such as pruning and knowledge distillation (Mu et al., 2025; Hinton et al., 2015; X. Tong et al., 2024; Su et al., 2022; Setyanto et al., 2023), which are critical for deployment on resource-constrained airborne systems.

Beyond localization, several studies frame wildfire detection as a binary image-level classification task, distinguishing between fire and no-fire scenes (Ghali and Akhloufi, 2023a,b) or smoke and no-smoke conditions (Kim and Muminov, 2023). While computationally efficient, these methods lack spatial localization, limiting their usefulness for real-time wildfire response and decision-making (Yar et al., 2023). To overcome this limitation in terrestrial imagery, hierarchical frameworks combining global image-level classification with patch-level localization have been proposed (Zhang et al., 2016). However, transferring such strategies to aerial imagery remains challenging. In aerial views, flames frequently appear at sub-pixel scales or are partially obscured by vegetation canopy (Bouguettaya et al., 2022; Ghali and Akhloufi, 2023b), and smoke distributions are heavily influenced by atmospheric conditions, often decoupling smoke presence from precise fire locations (Kim and Muminov, 2023). Additionally, aerial datasets suffer from severe patch-level class imbalance, as fire regions occupy only a small proportion of the image area (Bouguettaya et al., 2022). This imbalance significantly degrades classifier performance and underscores the need for specialized learning strategies tailored to aerial wildfire detection.

2.2. Multi-modal sensing and cross-modal knowledge distillation

The limited computational resources and payload capacity of aerial platforms make real-time wildfire monitoring with multiple sensors costly, technically challenging, and frequently infeasible. Recent studies have shown that fusing spatial, temporal, and spectral information from multi-type and multi-source data can significantly enhance early wildfire detection performance (Zhang et al., 2023, 2025). However, such approaches typically rely on extensive sensing infrastructure and substantial computational resources, which limit their applicability to edge-deployed UAV/AAV systems.

Consequently, most existing wildfire detection methods adopt single-modality sensing, predominantly relying on optical imagery (Alam et al., 2025; Ghali and Akhloufi, 2023a; Shamsoshoara et al., 2021; Ghali et al., 2022) or thermal data (Chen et al., 2022; Hopkins et al., 2024; Rajoli et al., 2024). Optical imagery is effective for capturing visible flames and smoke, but it degrades under dense smoke, low illumination, or canopy occlusion. In contrast, thermal imagery enables robust hotspot detection, but it is susceptible to false positives caused by non-fire heat sources. To leverage the complementary strengths of these modalities, several multi-sensor approaches have been explored, including sequential transfer learning (Zachariadis et al., 2023), early and late fusion strategies (Chen et al., 2022; Mu et al., 2025), and the incorporation of additional modalities such as synthetic aperture radar (Abdusalomov et al., 2024). Despite their effectiveness, these methods require multi-modal sensing during inference, increasing payload requirements, system complexity, and computational cost, thereby limiting their suitability for aerial edge deployment.

To overcome these limitations, recent works have explored knowledge distillation as a more efficient alternative for resource-constrained platforms by transferring knowledge from a more informative teacher model to a lightweight student network (Hinton et al., 2015). In cross-modal distillation, a student operating on optical imagery is guided by a teacher trained on thermal data, enabling the student to implicitly learn thermal cues without requiring thermal sensors at inference time (Marinaccio and Afghah, 2025; X. Tong et al., 2024). Building on this paradigm, several studies adopt teacher–student learning frameworks that leverage multi-modal information during training while maintaining unimodal inference (Zhou et al., 2023; Feng et al., 2023; Huo et al., 2024). Recent approaches have also explored vision-language models like CLIP as teachers to distill knowledge into lightweight CNN students, leveraging CLIP's rich semantic representations to improve performance (Jang et al., 2025). In the context of wildfire detection, thermal-guided distillation has demonstrated that an optical-only student model can achieve performance comparable to multi-sensor systems while substantially reducing system complexity and resource demands (Bouguettaya et al., 2022).

Despite its promise, cross-modal knowledge distillation remains challenging due to the semantic gap between modalities and their heterogeneous feature representations, which can limit the effectiveness of direct distillation strategies (X. Tong et al., 2024). Moreover, most existing cross-modal distillation methods focus on segmentation or object detection tasks, which are often computationally intensive and less suitable for real-time aerial edge deployment. Motivated by these challenges, this work proposes a lightweight wildfire detection framework that leverages cross-modal knowledge distillation to enhance unimodal model performance while remaining practical for real-time UAV/AAV applications.

3. Methodology

3.1. Dataset description

FLAME2 dataset consists of daytime aerial images of wildfire including visible flames, smoke plumes, burning vegetation, and fire fronts captured in outdoor forest and grassland environments. It consists of images under three classes: no-fire and no-smoke, fire with smoke, and fire with no smoke. Following the standard strategy prevalent in the literature while using FLAME2 dataset (Chen et al., 2022) to mitigate redundancy present in the original video-frames based FLAME2 dataset, we curated a representative subset of 1800 images for our study.

For patch-level annotations, preliminary labels obtained using a heuristic method (Rajoli et al., 2024) were meticulously reviewed and corrected as described in Algorithm 1. Another dataset, FLAME3 CV-subset (Hopkins et al., 2024) consisting of daytime aerial images for fire and no-fire classification was used for testing the trained model. For cross-dataset testing, 100 randomly selected optical images from the FLAME3 CV-subset were curated and a similar patch label annotation process was followed to obtain ground-truth.

3.2. Overall framework

The curated FLAME2 dataset was first used to train a teacher model on thermal imagery with a three-class classification scheme: class0 (no fire, no smoke), class1 (fire with smoke), and class2 (fire with no smoke). The dataset is split such that 80% is training and 20% is the test dataset. The teacher network comprises of contrastive language–image pre-training, CLIP network (`clip-vit-base-patch32`) (Radford et al., 2021) as backbone with its last two layers as trainable, and MLP layers for image classification, as shown in Fig. 1. The teacher model was trained for 30 epochs using the Adam optimizer with a learning rate of 0.0001, employing the cross-entropy loss function. Once trained, the teacher network is frozen while training the student network.

Algorithm 1 Patch-Level Annotation Verification Procedure

Require: RGB images $Img_{[R,G,B]}$, thermal heatmaps Img_{IR}, initial labels $L_{heuristic}$ from FlameSeg(Rajoli et al., 2024), smoke coverage threshold $\tau_{smoke} = 0.5$, thermal intensity threshold $\tau_{thermal}$
Ensure: Verified patch-level labels $L_{verified}$
Stage 1: Independent Human Verification
for each patch p in dataset **do**
  Display $Img_{[R,G,B]}(p)$ and $Img_{IR}(p)$ to Annotator A
  Annotator A assigns label based on criteria:

$$L_A(p) \leftarrow \begin{cases} \text{Smoke (no-flame)}, & \text{coverage_area} > \tau_{smoke} \\ \text{Background (no-flame)}, & \text{coverage_area} \le \tau_{smoke} \text{ and } Img_{IR}(p) < \tau_{thermal} \\ \text{Fire (flame)}, & \text{flame visible or } Img_{IR}(p) \ge \tau_{thermal} \end{cases}$$

  Independently, Annotator B assigns $L_B(p)$ using same criteria
end for
Stage 2: Conflict Resolution
for each patch p where $L_A(p) \neq L_B(p)$ **do**
  Joint review by both annotators with $Img_{[R,G,B]}(p)$ and $Img_{IR}(p)$
  Apply labeling criteria consistently and reach consensus: $L_{consensus}(p)$
end for
Stage 3: Final Label Assignment
for each patch p **do**

$$L_{verified}(p) \leftarrow \begin{cases} L_A(p), & \text{if } L_A(p) = L_B(p) \\ L_{consensus}(p), & \text{if } L_A(p) \neq L_B(p) \end{cases}$$

end for
Result: 46.53% of labels corrected from $L_{heuristic}$
return $L_{verified}$

Labeling Criteria: Smoke: > 50% area obscured, no flame; Background: ≤ 50% smoke, no flame, $Img_{IR} < \tau_{thermal}$; Fire: Visible flame or thermal confirmation $Img_{IR} \ge \tau_{thermal}$

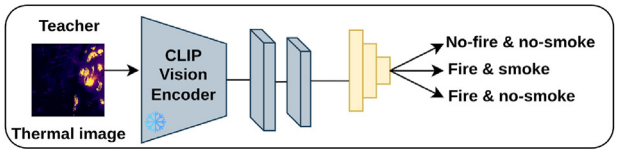


Fig. 1. Teacher model with frozen CLIP backbone and last two trainable layers. It is trained for a classification task using thermal images via cross entropy loss.

For the subsequent student model training, the class labels were redefined for a two-headed network architecture. The image head was trained for a binary classification task where class0 was designated as 'fire-free', and the previous class1 and class2 were combined into a single 'fire-impacted' class, also designated as class1. Concurrently, a patch head was trained to classify image patches. For this head, patches containing contextual cues like forest backgrounds or thick smoke were labeled as class0 ('no-flame'), while patches that contained actual flame pixels were labeled as class1 ('flame'). This dual-head approach was designed to explore how models learn to distinguish between direct fire features and broader environmental indicators.

The student network comprises of a fine-tunable convolutional neural network (CNN) backbone with two additional modules, namely deformable convolutional network (DCN) (Dai et al., 2017) and convolutional block attention module (CBAM) (Woo et al., 2018). The student network includes two classification heads: image-head for image level classification and patch-head for patch level flame localization. During training of the student network, a paired thermal and optical image is used. The thermal image is input into the frozen teacher network, and the corresponding optical image is input into the trainable student network. The student network leverages the distilled features from the thermal counterpart in the teacher network while processing optical images. The overall architecture of our proposed model is shown in Fig. 2.

3.2.1. Deformable Convolutional Networks (DCN)

DCNs are a powerful extension of standard convolutional neural networks (CNNs) that address the limitation of fixed-geometry kernels (Dai

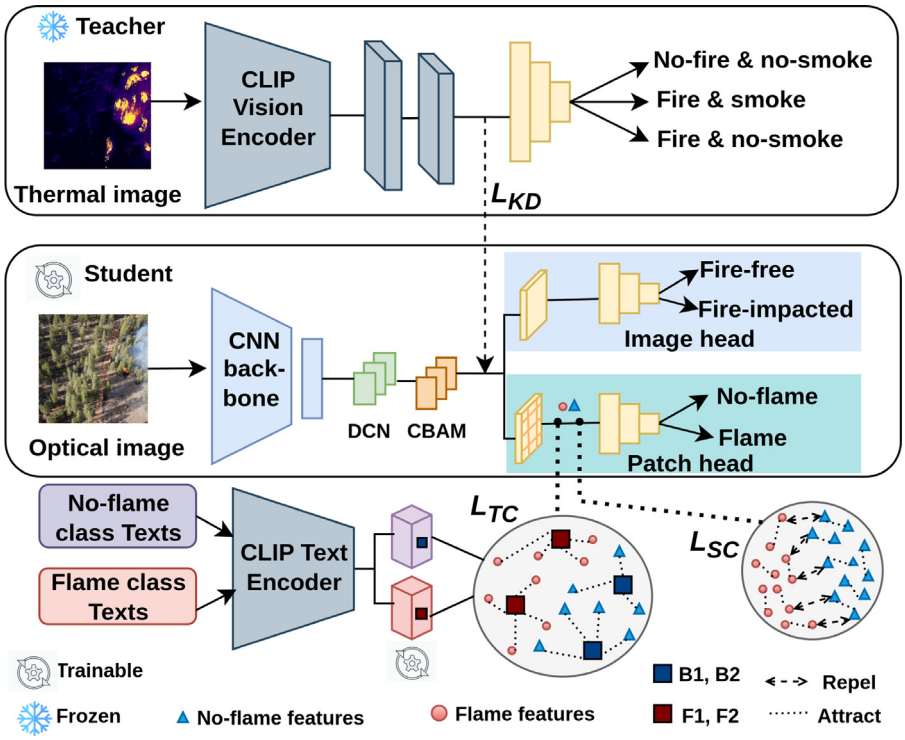


Fig. 2. Overall architecture for cross-modal knowledge distillation setup for wildfire detection and localization. Here, DCN is deformable convolutional network; CBAM is convolutional block attention module; B1 and B2 are two text-anchors for no-flame/background class; F1 and F2 are two text-anchors for the flame class. L_{KD}, L_{TC}, L_{SC} are knowledge distillation loss, text-based contrastive loss and supervised contrastive loss respectively given by Eqs. (13), (15) and (18).

et al., 2017). Unlike traditional convolutions which sample a rigid, predefined grid of pixels (e.g., a 3x3 square), DCN introduce learnable offsets to the sampling locations. These offsets are dynamically predicted by the network itself, based on the input features. This allows the convolution to adaptively change its field of view to match the varying geometry and scale of objects within an image. Therefore, DCN helps the model to learn "how" to look.

This ability to model complex spatial transformations with a flexible receptive field significantly boosts accuracy in tasks like classification (Zhu et al., 2018) and object detection (Cao et al., 2020). Importantly, this is accomplished with only a marginal computational overhead, as the additional parameters for learning the offsets are minimal. The core advantage of DCN is its ability to learn these transformations directly from the data without the need for extensive data augmentation or complex hand-crafted features, making the network more robust and generalizable (Burgos Madrigal et al., 2024).

Given an input feature map $\mathbf{F}$, a regular convolutional grid $\mathcal{R}$, and learnable offsets $\Delta\mathbf{p}_k$, deformable convolution is defined as

$$y(\mathbf{p}) = \sum_{k\in\mathcal{R}} w_k\, \mathbf{F}(\mathbf{p}+\mathbf{p}_k+\Delta\mathbf{p}_k), \qquad \Delta\mathbf{p}_k = f_{\text{off}}(\mathbf{F}), \tag{1}$$

where the offsets are dynamically predicted from the input features, enabling adaptive receptive fields.

3.2.2. Convolutional Block Attention Module (CBAM)

CBAM is a lightweight attention mechanism that enhances a neural network's feature representation by applying attention to both the channel and spatial dimensions of a feature map (Woo et al., 2018). This sequential dual-attention process allows the network to refine features by learning to focus on "what" is important (channels) and "where" it is important (spatial locations).

The module first applies a channel attention module to determine the most meaningful feature channels by compressing the spatial information and using a shared multi-layer perceptron (MLP) to generate an attention vector. This vector is then used to re-weight the channels. Next, a spatial attention module is applied that aggregates channel-wise information to create a spatial attention map. This map identifies the precise locations of key features. By combining these two attention maps, CBAM effectively induces the network to attend to the most relevant information, boosting accuracy with minimal computational overhead (Mishra et al., 2023; Jiang and Yin, 2023).

Given an intermediate feature map $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$, CBAM refines features via sequential channel and spatial attention:

$$\mathbf{F}' = \mathbf{M}_s\big(\mathbf{M}_c(\mathbf{F})\odot\mathbf{F}\big)\odot\big(\mathbf{M}_c(\mathbf{F})\odot\mathbf{F}\big), \tag{2}$$

where $\odot$ denotes element-wise multiplication.

The channel attention map $\mathbf{M}_c$ is computed as

$$\mathbf{M}_c(\mathbf{F}) = \sigma(\text{MLP}(\text{AvgPool}(\mathbf{F})) + \text{MLP}(\text{MaxPool}(\mathbf{F}))), \tag{3}$$

and the spatial attention map $\mathbf{M}_s$ is defined by

$$\mathbf{M}_s(\mathbf{F}) = \sigma\big(f^{7\times7}([\text{Avg}_c(\mathbf{F});\text{Max}_c(\mathbf{F})])\big), \tag{4}$$

with $\sigma(\cdot)$ denoting the sigmoid function.

3.3. Loss function and training strategy

3.3.1. Knowledge distillation loss ($\mathcal{L}_{KD}$)

Spatial feature distillation: In cross-modal knowledge distillation from a thermal modality based teacher to an optical modality based student, there exists a significant modality gap. Thermal and optical sensors capture fundamentally different physical properties of a scene (heat vs. reflected light). This makes direct feature imitation with losses like L2 ineffective, as it incorrectly penalizes differences in absolute feature values which are inherently divergent (X. Tong et al., 2024). To address this, pearson correlation coefficient (ρ_{hw}) based spatial feature distillation loss ($\mathcal{L}_{\text{fea}}$) is used, as given by (5). This loss function focuses on maximizing the linear correlation between teacher $T \in \mathbb{R}^{(C\times H\times W)}$ and student $S \in \mathbb{R}^{(C\times H\times W)}$ feature vectors across channels at each spatial location. Knowledge is transferred using the features

from the teacher model after the CLIP backbone, and is transferred to the student model at the point following its CBAM module. For both points in the teacher and student models, the feature shape is the same (i.e. 768 × 7 × 7) for applying the knowledge transfer loss. By leveraging the scale and shift-invariant properties of ρ_{hw}, the student distills the teacher's high-level structural patterns and relational knowledge, rather than attempting to replicate modality-specific values (X. Tong et al., 2024). This approach fosters a more robust and generalized student representation by efficiently handling the domain shift between thermal and optical data.

$$\mathcal{L}_{\text{fea}} = 1 - \left(\frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} \rho_{hw} \right) \tag{5}$$

where,

$$\rho_{hw} = \frac{\sum_{c=1}^{C} \hat{T}_{c,h,w} \cdot \hat{S}_{c,h,w}}{\sqrt{\sum_{c=1}^{C} \hat{T}_{c,h,w}^2} \cdot \sqrt{\sum_{c=1}^{C} \hat{S}_{c,h,w}^2}} \tag{6}$$

$$\hat{T}_{c,h,w} = (T_{c,h,w} - \mu_{T_{hw}}) \tag{7}$$

$$\hat{S}_{c,h,w} = (S_{c,h,w} - \mu_{S_{hw}}) \tag{8}$$

$$\mu_{T_{hw}} = \frac{1}{C} \sum_{c=1}^{C} T_{c,h,w} \tag{9}$$

$$\mu_{S_{hw}} = \frac{1}{C} \sum_{c=1}^{C} S_{c,h,w} \tag{10}$$

Semantic-features relation distillation: The objective of semantic feature relation distillation is to transfer the teacher's understanding of global context by preserving the relational structure among high-level semantic features. The modality gap prevents effective direct feature comparison. To bridge this gap, the semantic-features relation distillation method constructs affinity matrices for both the teacher and student models. These matrices encode the pairwise similarity between all spatial feature vectors, typically using cosine similarity (X. Tong et al., 2024). Such matrices create a structural fingerprint of the scene. An L1 loss ($\| \cdot \|_1$) is then employed to minimize the element-wise difference between the teacher's and student's affinity matrices. Thus, ($\mathcal{L}_{\text{rel}}$) given by (11) aids in supervising the student to accurately learn the modality-agnostic relational cues embedded in the teacher's deepest features.

$$\mathcal{L}_{\text{rel}} = \frac{1}{K^2} \sum_{i=1}^{K} \sum_{j=1}^{K} \| A_{i,j}^{t} - A_{i,j}^{s} \|_1 \tag{11}$$

where,

$$A_{i,j} = \frac{f_i^T f_j}{\|f_i\|_2 \cdot \|f_j\|_2} \tag{12}$$

and K represents the total number of spatial feature vectors, which is equal to $H \times W$ of the feature map $\mathbb{R}^{(C \times H \times W)}$. Hence, combining spatial feature distillation and semantic-features relation distillation forms knowledge distillation loss $\mathcal{L}_{\text{KD}}$, which is given by (13).

$$\mathcal{L}_{\text{KD}} = a_1 \cdot \mathcal{L}_{\text{fea}} + a_2 \cdot \mathcal{L}_{\text{rel}} \tag{13}$$

where, a_1 and a_2 are respective weights.

3.3.2. Contrastive loss ($\mathcal{L}_{Con}$)

Text anchor-based contrastive loss ($\mathcal{L}_{\text{TC}}$): It aligns patch-level visual features with semantic representations through text-anchor encodings. Since CLIP text features provide distinctive concept representations, leveraging them to enhance the semantic representations of student features is a promising approach (Shipard et al., 2025). Around 30 texts were generated for each class, as listed in Table 1. Text-embeddings were extracted using the CLIP text encoder, followed by K-means clustering to obtain two text-anchors per class as shown in Fig. 3.

Table 1
Text prompts used for positive and negative classes during model training.

No-flame/background class texts
forest floor with fallen leaves; tree trunk bark texture; dry grassland terrain; rocky mountain surface; concrete pavement surface; metal industrial panel; asphalt road texture; brick wall pattern; geometric pattern background; minimalist white surface; textured fabric material; smooth glass panel; sandy beach surface; muddy wet ground; snowy winter landscape; desert sand dunes; dense gray smoke plume; wispy white smoke drift; thick black smoke haze; smoldering smoke residue; foggy misty atmosphere; clear blue sky background; cloudy overcast conditions; dark shadowy interior; water surface reflection; wet pavement after rain; wooden plank surface; tile floor pattern; marble stone texture
Flame class texts
wildfire flame burning; campfire ember glow; forest fire spreading; bonfire orange flame; wood fire crackling; grass fire burning; brush fire ignition; outdoor fire pit; flame front advancing; flickering flame tongue; glowing red embers; bright yellow flame; dancing fire light; smoldering wood coals; brush pile ignited; dry vegetation ablaze; forest undergrowth burning; campsite fire burning; backyard fire bowl; beach bonfire flame; cabin fireplace glow; white hot flame center; incandescent fire glow; radiant heat emission; thermal infrared hotspot; combustion zone burning; ignition point flare; outdoor torch flame; gas grill burner; propane flame heat

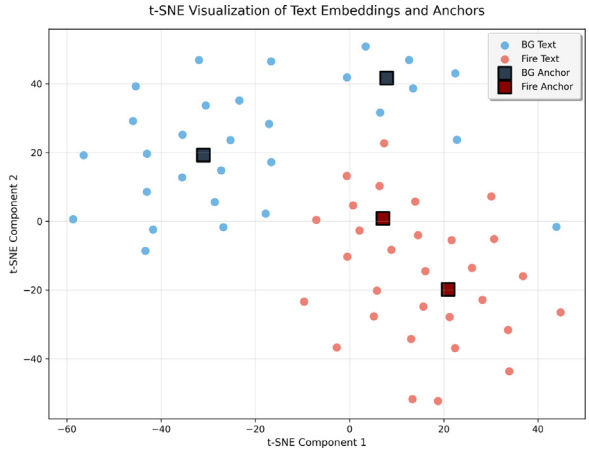


Fig. 3. CLIP based text embeddings for each class text and text-anchors after applying K-means clustering.

The framework employs multiple learnable anchors per class to address key limitations of conventional approaches. Firstly, single anchors cannot adequately represent intra-class visual diversity. By representing a single class with multiple anchors, the model can learn a more robust and multifaceted semantic embedding space. Secondly, static anchors lack the adaptability needed for domain-specific specialization. During the proposed model's training process, the text-anchors are fine-tuned for better alignment with the embedding space of the student model. This learnable nature permits the semantic representations to adapt and specialize from a CLIP domain space to the specific visual embedding space of the student model.

Unlike traditional contrastive learning that includes both attraction and repulsion terms, the formulation of $\mathcal{L}_{\text{TC}}$ focuses solely on attracting relevant features to their corresponding text-anchors (Mishra et al., 2025). Such formulation focuses on gradually guiding features toward semantic alignment rather than hard separation, as defined by (14).

$$\mathcal{L}_{\text{CL}}(S, m) = -\frac{1}{2} \sum_{k=1}^{2} \log \left(\frac{\exp(s_{i,k}/\tau)}{\exp(s_{i,1}/\tau) + \exp(s_{i,2}/\tau)} \right) \tag{14}$$

where S is the similarity between features and anchors ($k \in \{1, 2\}$), m is class label and τ is the temperature. As discussed, text-anchors are initialized for each class with two k-mean cluster centers of pre-trained CLIP encodings to establish semantic grounding, then made learnable

during training. The loss $\mathcal{L}_{TC}$ is constructed using $\mathcal{L}_{CL}(S, m)$ for two classes in our case. It enhances the contextual and semantic understanding of visual data while improving the discriminative performance of the model. Hence, $\mathcal{L}_{TC}$ is given by (15), with hyperparameter α and β as weighting factor with values 8 and 0.2 respectively.

$$\mathcal{L}_{TC} = \alpha \mathcal{L}_{CL}(S^+, m^+) + \beta \mathcal{L}_{CL}(S^-, m^-) \tag{15}$$

where,

$$S^+ = f_i \cdot [a^+]_k^T, \qquad S^- = f_i \cdot [a^-]_k^T \tag{16}$$

$$m^+ = y, \qquad m^- = 1 - y \tag{17}$$

where, y is ground truth vector with index i corresponding to patches: 1 for patch label of 'flame' patch and 0 for 'no-flame' patch. The $\mathcal{L}_{TC}$ loss is implemented only for 'fire-impacted' images, which contain both 'flame' and 'no-flame' patches. Here, $f_i \in \mathbb{R}^{N\times 512}$ is patch-level visual feature, $a^+ \in \mathbb{R}^{2\times 512}$ and $a^- \in \mathbb{R}^{2\times 512}$ are positive-class and negative-class text-anchor encodings respectively.

Supervised contrastive loss ($\mathcal{L}_{SC}$): This loss aims to structure model's embedding space by utilizing class labels. Its core mechanism is to enforce intra-class compactness by pulling feature representations of same-labeled samples together, while simultaneously promoting inter-class separability by pushing representations of different-labeled samples apart (Khosla et al., 2020). It compels the model to learn highly discriminative features, leading to more robust and generalizable models (Shyam et al., 2022; Stival et al., 2025). In contexts with significant class imbalance, such as detecting small 'flame' regions within a larger image, balanced sampling is a crucial component. The dataset balanced $\mathcal{L}_{SC}$ calculation strategy for batch B, number of patch in height and width as H, W, and feature dimension D, is shown in Algorithm 2, ensures that both 'dominant' and 'rare' classes contribute equally to the loss. This prevents the model from overlooking the features of the underrepresented class.

Algorithm 2 Patch-level Supervised Contrastive Loss

Require: Patch features $\mathcal{F} \in \mathbb{R}^{B\times H\times W\times D}$, patch labels $y_p \in \{0,1\}^{B\times H\times W}$, image labels $Y_i \in \{0,1\}^B$, temperature τ
Ensure: Supervised contrastive loss $\mathcal{L}_{SC}$
1: Filter images containing fire-impacted target class: $\mathcal{F}' \leftarrow \{\mathcal{F}_k \mid Y_{i,k} = 1\}$
2: Flatten filtered image spatial dimensions: $F \in \mathbb{R}^{N\times D}, y \in \{0,1\}^N$
3: Identify positive and negative patch indices: $I_{pos} = \{j \mid y_j = 1\}$, $I_{neg} = \{j \mid y_j = 0\}$
4: For keeping the balance to avoid dataset bias, find optimal number of samples required: $k \leftarrow \min(|I_{pos}|, |I_{neg}|)$
5: Sample balanced set: $S \leftarrow \text{RandomSample}(I_{pos}, k) \cup \text{RandomSample}(I_{neg}, k)$
6: Normalize features: $z_j \leftarrow F_j / \|F_j\|_2$ for all $j \in S$
7: Compute loss:

$$\mathcal{L}_{SC} = \sum_{i\in S} \frac{-1}{|P(i)|} \sum_{p\in P(i)} \log \frac{\exp(z_i \cdot z_p/\tau)}{\sum_{j\in S\setminus\{i\}} \exp(z_i \cdot z_j/\tau)} \tag{18}$$

where $P(i) = \{p \in S \setminus \{i\} \mid y_p = y_i\}$
8: **return** $\mathcal{L}_{SC}$

Hence, overall contrastive loss $\mathcal{L}_{Con}$ is defined by combining $\mathcal{L}_{TC}$ and $\mathcal{L}_{SC}$, as given by (19), where b_1 and b_2 are respective weights.

$$\mathcal{L}_{Con} = b_1 \cdot \mathcal{L}_{TC} + b_2 \cdot \mathcal{L}_{SC} \tag{19}$$

3.3.3. Classification loss ($\mathcal{L}_{Cls}$)

Image head classification loss ($\mathcal{L}_{CE}$): The image head of the proposed model classifies input images into 'fire-free' and 'fire-impacted' class. During training, cross entropy loss is used, as given by (20). In the image head, feature map $(B, 768, 7, 7)$ is reduced to $(B, 768)$ using adaptive average pooling, which is then passed through MLP layers yielding an output of $(B, 2)$ to produce image-level predictions, where B is batch size.

$$\mathcal{L}_{CE} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \tag{20}$$

where, y_i is ground truth label (one-hot encoded), $\hat{y}_i$ is predicted probability for class i and C is number of classes.

Patch head classification loss($\mathcal{L}_{FL}$): The patch head classifies image-patches into 'no-flame' (negative) and 'flame' (positive) classes. This effectively trains the model to localize flame in an image, using focal loss function (Lin et al., 2017). Focal loss inherently addresses class imbalance between dominant and rare class samples. Dominant-class samples are typically easy to classify, whereas rare-class samples are harder due to limited representation and the dominance of other class samples in the loss and gradient computation. This imbalance is particularly pronounced at the patch level, where dividing an image into 49 patches typically yields only 2–3 'flame' patches, with the remainder corresponding to background or non-flame regions. Such disproportion skews the training process, as the loss and gradients become dominated by abundant non-flame patches, reducing the model's sensitivity to flame regions. By design, focal loss mitigates this imbalance while amplifying the learning signal from hard, informative examples, and is employed here as defined in (21).

In the patch head, feature map $(B, 768, 7, 7)$ is reshaped and reduced to $(B\times 7\times 7, 512)$ and passed through MLP layers, yielding an output of $(B\times 7\times 7, 2)$, which is then reshaped to $(B, 7, 7, 2)$ to produce patch-level predictions.

$$\mathcal{L}_{FL}(p_t) = -\alpha_t (1 - p_t)^\gamma \log(p_t), \tag{21}$$

$$p_t = \begin{cases} p & \text{if the sample is positive,} \\ 1 - p & \text{if the sample is negative.} \end{cases} \tag{22}$$

where p is prediction probability of model, α_t is a weighting factor for class imbalance; $\gamma \geq 0$ is a focusing parameter, where a higher value of γ places more focus on hard examples. So, the classification loss combining image-head and patch-head is given by $\mathcal{L}_{Cls} = c_1 \cdot \mathcal{L}_{CE} + c_2 \cdot \mathcal{L}_{FL}$, where c_1 and c_2 are respective weights.

3.3.4. Penalty loss ($\mathcal{L}_P$)

Missed flame-detection penalty($\mathcal{L}_{PL}$): In fire detection tasks, the consequences of missed detections (false negatives) are far more severe than occasional false alarms (false positives). Standard loss functions such as cross-entropy treat all errors equally, which may lead to models biased toward overall accuracy while under-penalizing false negatives. To address this imbalance, a penalty loss is introduced. The loss function assigns higher weights to false negative errors, ensuring that missed fire detections contribute more strongly to the optimization objective. This design makes the model sensitive to recall, thereby balancing the tradeoff between overall accuracy and the critical need to minimize undetected fire events. For this purpose $\mathcal{L}_{PL}$ is proposed as

$$\mathcal{L}_{PL} = 100 \times (1 - R_P) \tag{23}$$

$$R_P = \frac{P_{TP}}{P_{TP} + P_{FN} + \epsilon} \tag{24}$$

Here, $P_{TP} = p_{flame} \cdot y_{flame}$, $P_{FN} = (1 - p_{flame}) \cdot y_{flame}$. Where, p_{flame} is the predicted fire probability of a given patch, y_{flame} is the corresponding ground-truth patch label, and ϵ is a small constant for numerical stability.

An overall penalty loss $\mathcal{L}_P$ combined with sparsity component is given by (25). The sparsity component is implemented via L1 regularization on model network weights W_j, encouraging the network to learn compact, informative activations.

$$\mathcal{L}_P = d_1 \cdot \mathcal{L}_{PL} + d_2 \cdot \sum_j |W_j| \tag{25}$$

where, d_1 and d_2 are respective weights.

Thus, the total loss or the overall objective function for the optimization of the proposed model is given by (26).

$$\mathcal{L}_T = \mathcal{L}_{KD} + \mathcal{L}_{Con} + \mathcal{L}_{Cls} + \mathcal{L}_P \tag{26}$$

Table 2
Experimental setting and hyperparameter values used for training.

Parameter	Value
Image resolution	224 × 224
Patch size & #patches	32 × 32, 49
Batch size	64 (32 +ve, 32 −ve)
Epochs	250
Learning rate	0.001, StepLR γ=0.95/10 steps
Optimizer	AdamW
$[a_1, a_2, b_1, b_2, c_1, c_2, d_1, d_2]$	[275, 740, 40, 20, 25, 100, 80, 2e-6]
Temperature τ (L_{TC}, L_{SC})	(0.08, 0.08)
Focal loss (α, γ)	[0.2, 0.8], 2.6

Table 3
Architecture Ablation Study on St_DenseNet121.

Student model configuration	Acc_{img}	Acc_{patch}	$Sens_{patch}$	FNR	FPR
Backbone+CBAM	100	77.4	84.12	8.65	23.12
Backbone+DCN	75	62.47	62.9	36.64	37.57
Backbone+DCN+CBAM	100	90.97	88.2	14.78	8.82

4. Experimental results and analysis

4.1. Implementation details

For student model CNN backbone, four types of networks were tested, referred as: `St_ResNet50` (He et al., 2016), `St_EfficiencyNetB0` (Tan and Le, 2019), `St_MobileNetV2` (Sandler et al., 2018) and `St_DenseNet121` (Huang et al., 2017). Collectively, they are referred hereby as `St_CNN`. The experimental setting details and parameter values are provided in Table 2. Since different loss components exhibit varying magnitudes, weight factors are applied to normalize them for balanced optimization, which is a standard practice in multi-objective deep learning. The loss weights were determined through a systematic grid search within magnitude-comparable ranges. In the training phase, we used images from two different modalities (RGB and IR/optical and thermal) to train the student and teacher models. In the inference phase, only a single modality (RGB) is used for testing. All methods were implemented in PyTorch and evaluated on a workstation with an Intel Core i9-14900K CPU (up to 6.0 GHz) and an NVIDIA GeForce RTX 5070Ti GPU.

4.2. Evaluation metrics

For the image head, we use standard metrics: accuracy, precision, recall, and F1-score (Eqs. (23)–(26)). For the patch head, we report accuracy and sensitivity as overall metrics (Eqs. 23 and 25). In addition, for flame-class detection, we include false negative rate (FNR) and false positive rate (FPR) as class-specific metrics (Eqs. 27 and 28). As discussed earlier, the patch data are highly imbalanced, with background or no-flame class patches dominating. Due to such imbalanced data, accuracy alone is misleading because a model biased toward background prediction may still achieve high scores. FNR and FPR address this limitation. A high FNR indicates missed flame detections, which is a critical safety risk. A high FPR leads to false alarms, causing alert fatigue. Together, these metrics provide a more reliable evaluation of patch-level performance by measuring both sensitivity to true flames and robustness against false alerts.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{27}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{28}$$

$$\text{Recall/Sensitivity} = \frac{TP}{TP + FN} \tag{29}$$

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{30}$$

$$\text{FNR} = \frac{FN_{flame}}{TP_{flame} + FN_{flame}} \tag{31}$$

$$\text{FPR} = \frac{FP_{flame}}{FP_{flame} + TN_{flame}} \tag{32}$$

where, TP is true positives, TN is true negative, FP is false positives and FN is false negatives.

The computational efficiency of a model for edge-AI deployment is commonly evaluated using the trainable parameter count (in millions), inference time (in seconds), and computational complexity (in GFLOPS). Among these, GFLOPS measures the total floating-point operations per forward pass, providing a hardware-agnostic estimate of workload. It is crucial for assessing whether a model can run efficiently on resource-constrained devices, enabling real-time, energy-efficient inference that is independent of specific hardware.

4.3. Quantitative results

Teacher model is trained on thermal images only and is compared with other baseline models. The comparison in Table 4 highlights that the `CLIP_Teacher` achieves near-perfect accuracy (99.16%), markedly higher than conventional CNNs (e.g., LeNet5, ResNet18) and specialized flame detection models such as FLAME (Chen et al., 2022). This performance gain can be attributed to CLIP's large-scale multimodal pretraining, which enables it to learn transferable representations and adapt effectively even to thermal imagery. However, CLIP's heavy computational footprint makes it unsuitable for real-time deployment in resource-constrained fire monitoring systems. To bridge this gap, our framework leverages CLIP as a teacher model, transferring its discriminative power to a lightweight student network. This strategy ensures that the advantages of robust representation learning are retained, thereby preserving high accuracy while ensuring compatibility with edge-AI platforms.

The inference performance on optical images demonstrates the effectiveness of our proposed model trained via cross-modal knowledge distillation from thermal images. Compared to single-modal RGB-based models, the results (Table 5) show that at the image level, our `St_CNN` based framework evaluated with each backbone individually achieves superior accuracy and sensitivity. Further, to assess the statistical consistency of the `St_DenseNet121` backbone, four test sets of 360 images were constructed with 'fire-free' to 'fire-impacted' ratios of (0.5, 0.5), (0.6, 0.4), (0.7, 0.3), and (0.8, 0.2). For each set, inference was repeated 10 times with randomly sampled images from the FLAME2 dataset, excluding those used in training or primary testing. Across all 40 experiments, consistent results with perfect metrics were obtained as shown in Table 5. This highlights the relevance of cross-modal knowledge distillation, where leveraging complementary thermal representations enables the optical (student) model to capture richer feature semantics than conventional single-modal learning.

Following this perfect performance on FLAME2 dataset, we conducted additional experiments to specifically assess the generalization capability of the proposed model's image-level head. As shown in Table 6, strong generalization is observed on the BurnedAreaUAV dataset after resolution alignment, whereas a noticeable performance drop occurs on FLAME3 due to increased scene diversity and coarser GCR. This indicates that image-level generalization remains sensitive to variations in scene context and ground coverage resolution.

Table 7 reports the patch-level inference performance of various student CNN backbones trained on optical images using full test set. The results demonstrate that the proposed student configurations (`St_CNN`) consistently achieve strong performance, with `St_DenseNet121` and `St_MobileNetV2` outperforming other backbones by delivering high accuracy (90.97% and 91.75%) coupled with superior sensitivity (88.2% and 87.39%), while maintaining relatively low FNR and FPR. In contrast, FireNet-CNN, adapted for patch-level classification, achieves very high accuracy (96.92%) but suffers from extremely poor sensitivity (24.21%), missing nearly 75% of flame instances, thereby limiting its practical applicability. This discrepancy highlights that accuracy alone is insufficient for evaluating fire detection models, as sensitivity

Table 4
Comparison for thermal image modality-based methods on the FLAME2 dataset. For the proposed method, teacher training was done on thermal images.

Model/Metrics (%)	Accuracy	Precision	Recall	F1-score
Logistic[a]	92.43	92.94	92.43	92.61
LeNet5[a]	92.15	92.19	92.79	92.30
VGG16[a]	97.38	97.57	97.26	97.35
MobileNetV2[a]	97.38	97.65	97.43	97.51
ResNet18[a]	96.26	96.97	96.27	96.54
FLAME (Chen et al., 2022)	85.79	87.47	86.91	86.81
CLIP_Teacher (ours)	**99.16**	**99.20**	**99.17**	**99.16**

[a] Baseline models trained and evaluated as in Chen et al. (2022).

Table 5
Comparison of image level metrics for different models with the proposed student model trained on optical images. Image head evaluated with each backbone under `St_CNN` individually.

Model/Metrics (%)	Dataset type	Accuracy	Precision	Recall	F1-score
Logistic[a]	aerial/FLAME2	90.37	90.99	89.48	89.57
LeNet5[a]	aerial/FLAME2	95.33	95.86	95.12	95.39
VGG16[a]	aerial/FLAME2	99.91	99.90	99.93	99.92
MobileNetV2[a]	aerial/FLAME2	99.35	99.33	99.42	99.36
ResNet18[a]	aerial/FLAME2	96.26	96.97	96.27	96.54
FLAME (Chen et al., 2022)	aerial/FLAME2	95.86	95.18	94.38	94.53
CT-Fire (Ghali and Akhloufi, 2023a)	terrestrial/custom	99.62	–	–	99.61
CT-Fire (Ghali and Akhloufi, 2023a)	aerial/FLAME	87.77	–	–	82.35
FireNet-CNN (Alam et al., 2025)	terrestrial/custom A	97.96	98.22	98.41	98.31
FireNet-CNN (Alam et al., 2025)	terrestrial/custom B	99.05	99.41	98.28	98.84
FireNet-CNN (Alam et al., 2025)	aerial/FLAME2	47.00	43.03	50.25	44.07
Ensemble Model (Ghali et al., 2022)	aerial/FLAME	85.12	–	–	84.77
St_CNN (ours)	aerial/FLAME2	**100**	**100**	**100**	**100**

[a] Baseline models trained and evaluated as in Chen et al. (2022).

Table 6
Image-level performance comparison across diverse scenes and ground coverage resolutions (GCR), evaluated on 50 wildfire and 50 no-fire scene images.

Dataset	GCR with respect to FLAME2	Accuracy	Precision	Recall	F1-score
FLAME2	1	100	100	100	100
BurnedAreaUAV (Ribeiro et al., 2023)#	1	97	97.17	97	97
FLAME3	Mixed resolution and scenes	81	86.23	81	80.29
FLAME3 CV subset (scene)	4.3	79	82.79	79	78.38
FLAME3 Nadir subset (scene)	4.3	80	80.19	80	79.97

As only burn site images are available in this dataset, 50 images are randomly selected and cropped to match GCR of FLAME2, and 50 no-fire images were taken from FLAME2 test dataset to make balanced test data.

Table 7
Comparison of patch level metrics for different models with the patch head results of the proposed student model, evaluated for different CNN backbones.

Student CNN backbones	Dataset type	Accuracy ↑	Sensitivity ↑	FNR ↓	FPR ↓
Zhang et al. (2016)	terrestrial/custom	90	39.2	–	2.3
FireNet-CNN (Alam et al., 2025)#	aerial/FLAME2	96.92	24.21	75.79	3.6
St_ResNet50	aerial/FLAME2	80.69	83.7	13.05	19.54
St_EfficiencyNetB0	aerial/FLAME2	90.6	85.66	19.65	9.02
St_MobileNetV2	aerial/FLAME2	91.75	87.39	17.3	7.92
St_DenseNet121	aerial/FLAME2	90.97	**88.2**	14.78	8.82

Model adapted for patch-level classification; ↑ indicates higher is better; ↓ indicates lower is better.

and error rates directly impact reliability in safety-critical scenarios. Overall, the proposed cross-modal student training framework yields a balanced trade-off, achieving competitive accuracy while substantially reducing false negatives and false positives across different backbones, with `St_DenseNet121` and `St_MobileNetV2` emerging as the most effective architectures. To further assess statistical consistency of the model (`St_DenseNet121`) for patch level classification, we ran the model 5 times for inference on different sets of randomly sampled 100 images from the FLAME2 test set. The mean and standard deviation of patch-head results (in %) are reported as follows- accuracy: 90.98 ± 0.74, sensitivity: 88.47 ± 0.87, FNR: 14.22 ± 1.64 and FPR: 8.82 ± 0.75.

The patch-level receiver operating characteristic (ROC) achieves a high area under curve (AUC) of 0.9487, indicating strong discriminative performance between flame and no-flame patches, as shown in Fig. 4(a). However, the precision–recall curve shows an AUC of 0.5857, as shown in Fig. 4(b). Although PR-AUC substantially exceeds random chance (0.0361), it reflects challenges imposed by severe class imbalance. This indicates room for improvement in fine-grained flame detection under challenging conditions, including background clutter and class imbalance.

To evaluate the generalizability of the proposed model, we tested it on 100 randomly selected optical image samples from the FLAME3 CV-subset scenes (Hopkins et al., 2024). These images were cropped at different ground coverage resolutions (GCR) relative to the FLAME2

Table 8
Image-level performance of the FLAME3 CV-subset scene for different crop resolutions relative to FLAME2.

Crops	GCR wrt FLAME2	Accuracy	Precision	Recall	F1-score
640 × 512	4.3	79.00	82.79	79.00	78.38
443 × 443	2.5	88.00	90.32	88.00	87.82
313 × 313	1.3	**90.00**	**91.67**	**90.00**	**89.90**
254 × 254 (FLAME2)	1	100	100	100	100

Table 9
Patch-level performance of the FLAME3 CV-subset scene for different crop resolutions relative to FLAME2.

Crops	GCR wrt FLAME2	Accuracy ↑	Sensitivity ↑	FNR ↓	FPR ↓
640 × 512	4.3	93.24	81.51	30.83	6.15
443 × 443	2.5	94.90	92.22	10.50	**4.98**
313 × 313	1.3	**94.80**	**93.34**	**8.19**	5.13
254 × 254 (FLAME2)	1	90.98	88.47	14.22	8.82

↑ indicates higher is better; ↓ indicates lower is better.

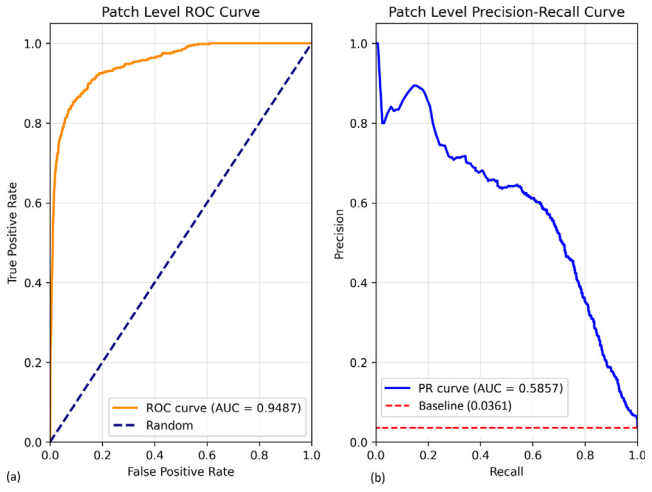


Fig. 4. ROC-AUC and PR-AUC curves for patch level classification in patch-head. (a) The patch-level ROC (AUC = 0.9487) shows strong discriminative ability. (b) The PR curve (AUC = 0.5857) remains well above the chance-level baseline (0.0361, corresponding to the ratio of positive to total samples). The relatively lower PR-AUC underscores the challenge of sustaining high precision in the presence of class imbalance.

dataset, resulting in three GCR ratios: 4.3 (original resolution of 640 × 512), 2.5, and 1.3. While cropping, it was ensured that flame-patches are present in fire class images. Accordingly, the patch-level ground truth annotation was done. Image-level and patch-level performances are reported in Tables 8 and 9, respectively.

At the image level, performance consistently improves as the GCR approaches that of FLAME2, with accuracy increasing from 79.0% at a GCR of 4.3 to 90.0% at 1.3. This trend indicates that closer alignment with the training GCR positively impacts classification performance, with the best FLAME3 results observed at 1.3 times GCR. In contrast, patch-level results exhibit stronger generalization across GCRs. While performance improves as the GCR decreases from 4.3 to 2.5, the best localization performance is achieved at a GCR of 1.3, surpassing even the FLAME2 baseline across all patch-level metrics. Specifically, higher sensitivity and lower false negative rates at GCR 1.3 indicate more robust fire localization. These results demonstrate that the patch head generalizes better to resolution variations and is less sensitive to domain shifts compared to the image-level head, reinforcing its effectiveness for fine-grained fire localization tasks.

The comparative analysis in Table 10 indicates that patch classification is an effective strategy for flame localization for edge-deployed UAVs/AAVs. Unlike object detection and segmentation models, which often have large parameter counts and higher computational complexity, the proposed patch classification networks (`St_DenseNet121` and `St_MobileNetV2`) achieve fine-grained localization with substantially lower resource demands. With parameter sizes of 11.14 million and 15.42 million and inference times ranging from 0.0018 to 0.0040 s, these models demonstrate strong computational efficiency. In contrast, detection models such as YOLOv7x require approximately 71 million parameters, and segmentation models like FireViTNet incur heavy computational loads (110 GFLOPs), making them less suited for edge deployment. The lightweight design of the proposed model ensures seamless compatibility with edge-AI platforms, enabling real-time operation in resource-constrained environments. These results highlight the practicality of patch-level approaches, offering a balanced trade-off between localization performance and deployability in real-world fire detection systems.

4.4. Ablation study

Conventional models designed for terrestrial fire detection, such as FireNet-CNN (Alam et al., 2025), have limitations in generalizing to aerial imagery, while others like CT-Fire (Ghali and Akhloufi, 2023a) demonstrate degraded performance, as evidenced in Table 5. To address the unique scale variations and resolution challenges inherent in aerial images, the proposed model integrates DCN and CBAM. Furthermore, the ablation results using densenet backbone for the student model (`St_DenseNet121`) in Table 3 highlight that although DCN and CBAM individually contribute performance gains, their joint integration achieves the most consistent and balanced improvements across all evaluation metrics. Using only CBAM improves image-level accuracy but leaves patch-level accuracy and sensitivity relatively low. In contrast, DCN alone significantly degrades performance, indicating that spatial adaptability without attention is insufficient. However, when DCN and CBAM are integrated, the model achieves the highest patch-level accuracy (90.97%) and strong sensitivity (88.2%) with notably reduced error rates (FNR 14.78%, FPR 8.82%). This highlights that DCN and CBAM act as complementary components. Together, they play a critical role in enhancing the discriminative ability and overall performance of the model.

Table 11 demonstrates that each loss function contributes uniquely to the student model's (`St_DenseNet121`) performance. Removing $\mathcal{L}_{KD}$ causes a drastic drop in patch-level accuracy (39.93%) and a surge in FPR (62.29%), highlighting the critical role of knowledge distillation in transferring teacher supervision. Excluding $\mathcal{L}_{Con}$ lowers patch-level accuracy (75.82%) and increases FPR, showing that contrastive loss is essential for feature discrimination. Without $\mathcal{L}_{P}$, patch accuracy remains high (97.44%) but sensitivity drops severely

Table 10
Comparison of models on edge-AI relevant performance metrics.

Model	Task	Trainable Params (M)	Inference Time (s)	GFLOPS	Hardware (GPU)
YOLOv5S-Fire (Wang et al., 2023)	Object detection	4.27	0.011	~16.5	GeForce RTX 2080
Modified YoloV7x (Gonçalves et al., 2024)	Object detection	~71.3	–	188	Tesla V100 SXM2
YOLOX-L (50% pruned) (Mu et al., 2025)	Object detection	–	0.012	17.62	GeForce RTX 4090
FireViTNet (Wang et al., 2024)	Segmentation	15.54	0.012	110	GeForce RTX 4060
FireFormer (H. Tong et al., 2024)	Segmentation	11.66	0.026	42.01	GTX 3060
SAM-TIFF U-Net (ResNet-152)# (Marinaccio and Afghah, 2025)	Segmentation	69.36	0.0163	157.16	GeForce RTX 5070Ti
SAM-TIFF DeepLabV3+ (ResNet-101) # (Marinaccio and Afghah, 2025)	Segmentation	58.77	0.0088	64.97	GeForce RTX 5070Ti
FireNet-CNN (Alam et al., 2025)	Image classification	2.75	0.95	2.085	Tesla P100
Ensemble Model (Ghali et al., 2022)	Image classification	–	0.018	–	GeForce RTX 2080Ti
St_MobileNetV2 (ours)	Patch classification	11.14	0.0018	1.41	GeForce RTX 5070Ti
St_DenseNet121 (ours)	Patch classification	15.42	0.0040	2.99	GeForce RTX 5070Ti

Performed on local hardware to estimate metrics using same input size 512X640 as used for this model in Marinaccio and Afghah (2025).

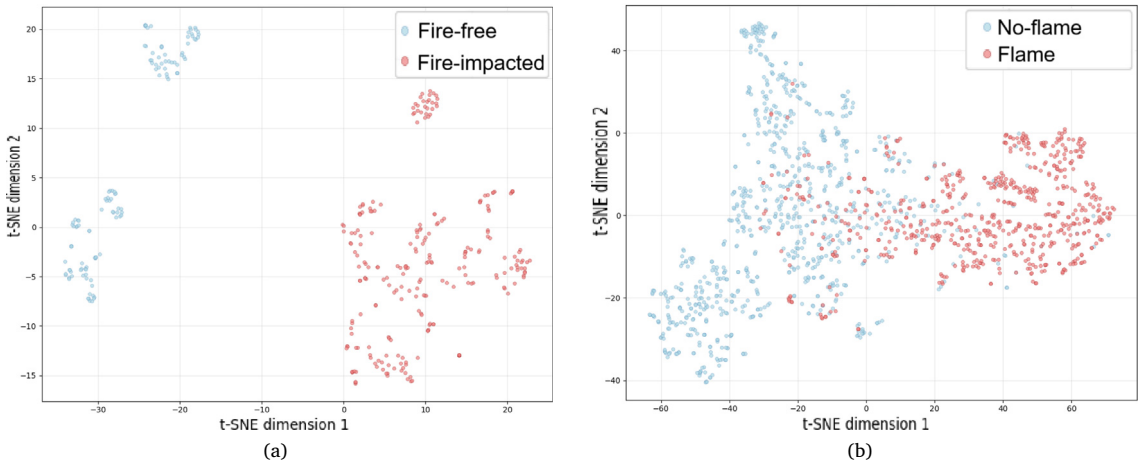


Fig. 5. t-SNE visualization of feature-embeddings (a) Image head (all test set images)(b) Patch head (randomly sampled 1500 patches from test set for better representation).

Table 11
Loss Function Ablation Study on St_DenseNet121.

Cost Function	Acc_{img}	Acc_{patch}	$Sens_{patch}$	FNR	FPR
w/o $\mathcal{L}_{KD}$	100	39.93	68.46	0.79	62.29
w/o $\mathcal{L}_{Con}$	100	75.82	82.92	9.43	24.74
w/o $\mathcal{L}_{P}$	100	97.44	78.09	42.77	1.06
w/o $\mathcal{L}_{TC}$	100	78.48	84.37	9.28	21.98
with all loss	100	90.97	88.20	14.78	8.82

Table 12
Performance comparison of different text-anchor settings in patch-head.

Text-anchors(%)	Accuracy	Recall	FNR	FPR
No anchors	78.48	84.37	9.28	21.98
Fixed anchors	86.58	86.00	14.62	13.38
Fine-tunable anchors	90.97	88.20	14.78	8.82

(78.09%) with a sharp rise in FNR (42.77%), indicating that patch-level supervision prevents excessive false negatives. Similarly, excluding $\mathcal{L}_{TC}$ reduces both accuracy and sensitivity, confirming the relevance of text-anchor alignment for inducing contextual awareness in the patch head. Furthermore, an ablation study of text anchors corroborates the effectiveness of using fine-tunable text anchors. The conditions in Table 12 are defined as follows: 'no anchors' refers to excluding the loss $\mathcal{L}_{TC}$ during training, 'fixed anchors' utilize static text anchor encodings, and 'fine-tunable anchors' employ trainable weights for the two text anchors per class. The best trade-off is achieved when all losses are combined while keeping text-anchors as trainable, reinforcing that each loss term plays a complementary role in balancing accuracy, sensitivity, and error rates.

5. Discussion

5.1. Qualitative analysis

The t-SNE (van der Maaten and Hinton, 2008) visualizations in Fig. 5 illustrate the separability of learned embeddings at both the image and patch levels. At the image level, embeddings of fire and non-fire classes form compact and clearly separable clusters, indicating that the model effectively captures global discriminative features for reliable fire scene classification. At the patch level, embeddings for flame and non-flame regions also exhibit distinct clustering, though with partial overlap. This overlap aligns with the accuracy observed at the patch level of 90.97% and reflects the inherent challenges of local discrimination due to similarities in smoke, illumination and background. Overall, the results confirm that the proposed framework learns highly discriminative global embeddings for image-level fire detection while preserving meaningful local feature separability to support flame localization.

Gradient-weighted class activation map or Grad-CAM (Selvaraju et al., 2017) is a heatmap-based visualization technique used to identify the spatial regions that contribute to a model's class decision. The Grad-CAM analysis on the image head provides critical insight into the model's decision-making process, as shown in Fig. 6. It reveals that trained model does not rely on the most intuitive feature which should be the flame pixels, for fire detection. Instead, the model has most likely learned to classify images based on contextual cues such as the presence of smoke and the type of surrounding environment while avoiding the cues of the flame pixels. As flame pixels are consistently allocated low

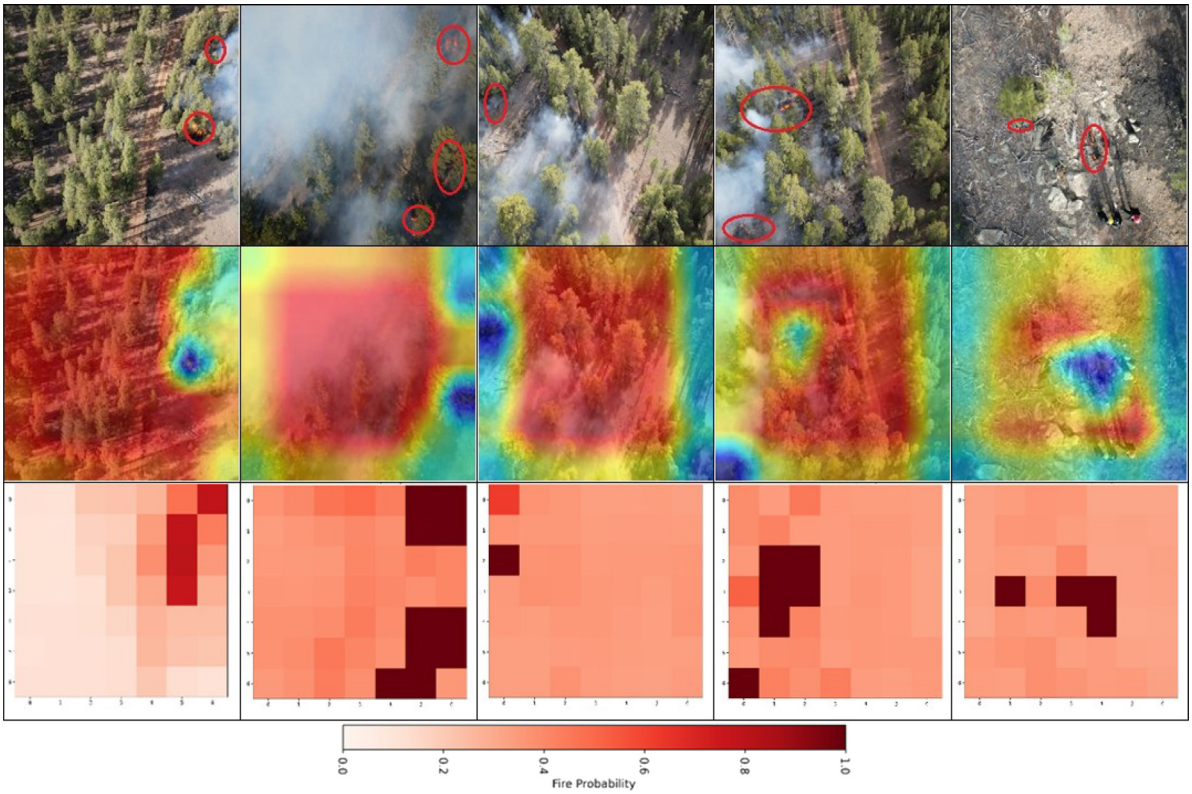


Fig. 6. Row#1 consists of sample optical fire-impacted images with highlighted flame-regions. Row#2 shows respective class activation maps for the image head, where intensity of red color denotes the most active pixels contributing to the 'fire-impacted' class and deep blue color corresponds to less activated pixels. Row#3 illustrates the respective prediction-based heatmaps of the patch head, thereby localizing the fire.

activation i.e. blue region, and hence model understanding optimizes in a way that if there is no low-activation (blue region), then it is fire-free; else the image is fire-impacted. Additionally, fire-free images containing clouds or dust were analyzed using Grad-CAM with the fire class and no-fire class as the target to evaluate robustness against misleading contextual cues. In no-fire images, class activation maps (CAMs) exhibit weak, diffuse activations for the fire class (target = 1), while the no-fire class (target = 0) shows strong, localized responses on clouds and terrain. This contrast indicates that contextual cues such as vegetation, atmospheric conditions, and terrain textures alone do not trigger false fire predictions. Instead, the model relies on a specific combination of spatial contextual patterns which are jointly evaluated by the downstream MLP layers to determine the final class.

This behavior can be attributed to the varied scales of fire and the dominant presence of a uniform forest background or widespread smoke in aerial images, which likely make the small flame regions less visually salient. Consequently, the model's attention is drawn to these more extensive and consistent patterns, explaining why the Grad-CAM heatmaps highlight smoke and foliage while showing low activation (blue regions) over the actual flames. This finding further underscores the importance of patch head in our proposed model, wherein an image is dissected into patches for closer analysis and flame localization. Corroborating this, the probability heatmaps derived from patch head predictions, as shown in Fig. 6 (Row#3), illustrate the effective localization of even the small scale and faintly visible flames at the patch level.

Finally, we emphasize that the proposed method and loss design are primarily optimized for patch-level localization, where class imbalance is significantly more severe than at the image level. The patch head and image head operate in parallel; consequently, strong and generalized patch-level performance does not necessarily imply equivalent generalization at the image level. The image-level loss is included as an auxiliary objective to support patch-level learning and convergence. Therefore, weaker image-level generalization is acceptable and does not detract from the model's main objective of accurate and robust fire localization, particularly given that high image-level accuracy does not always correspond to correct visual cues, as evidenced by the Grad-CAM analysis (Fig. 6).

5.2. Comparative study and future work

As shown in Table 13, previous approaches such as FLAME and ShRe-Xception achieved high accuracy through dual-modality fusion but lacked spatial localization, as they operated at image-level without flame localization. YOLOX-L improved detection through fused dual-modal images, yet remained dependent on paired sensors during edge deployment for inference. FlameFinder addressed localization but its reliance on heuristic annotations inflated performance, while human-expert annotations revealed substantial performance degradation.

In contrast, our proposed model integrates patch-level flame localization with expert-verified annotations, ensuring reliable and interpretable predictions. More importantly, the use of knowledge distillation (thermal → optical) transfers thermal discriminative cues into the optical domain, enabling deployment with only a single-modality sensor. This strategy reduces hardware requirements while maintaining competitive accuracy, thereby addressing both the localization and practicality gaps observed in previous works. Consequently, the model presents a viable solution for real-world fire monitoring systems, where accurate localization and minimal sensor payload are critical. Regarding spatial resolution, the 7×7 patch grid represents a deliberate design choice balancing localization granularity with edge-computing constraints. This configuration, supported by findings that

Table 13
Comparison of flame detection models with different modalities, training strategies, and inference setups.

Model	Training input	Strategy	Inference input	Key performance metrics (%)	Comments
FLAME (Chen et al., 2022)	Paired optical + thermal	Early fusion	Paired optical + thermal	accuracy: 95.70, f1-score: 95.67	Needs dual-modality sensors, no flame localization
FLAME (Chen et al., 2022)	Paired optical + thermal	Late fusion	Paired optical + thermal	accuracy: 98.60, f1-score: 98.63	Needs dual-modality sensors, no flame localization
ShRe-Xception (Zachariadis et al., 2023)	Sequential optical & thermal	Transfer learning	Optical	accuracy: 90.1	No flame localization
ShRe-Xception (Zachariadis et al., 2023)	Sequential optical & thermal	Transfer learning	Thermal	accuracy: 99.31	No flame localization
YOLOX-L (50% pruned) (Mu et al., 2025)	Paired optical + thermal	Fused dual-modal image data (IR:RGB::0.4:0.6)	Paired optical + thermal	mAP: 93.61	Needs dual-modality sensors
FlameFinder (Rajoli et al., 2024)	Thermal image (complete) + masked thermal image	Attentive deep metric learning	Thermal	accuracy: 82.9	Model performance relies on heuristic annotation (skewing accuracy to 97.8%). With human expert annotations, accuracy is 82.9% for patch-level flame localization.
Our proposed model	Paired optical + thermal	Knowledge distillation (thermal → optical)	Optical	accuracy: 90.97, FNR: 14.78, FPR: 8.82	1. Annotations by human-expert 2. Patch-level flame localization 3. Needs only single-modality sensor

larger kernels (≥7×7) improve convolutional performance (Hou et al., 2024), provides approximately 50 m^2 ground coverage per patch at operational altitudes as that of FLAME2. Geographic fire localization is accomplished via direct georeferencing, combining UAV GNSS/altitude data with IMU orientation and camera parameters to project patch coordinates onto real-world positions with accuracy determined by the platform's navigation system.

We further analyzed the effect of smoke and texture ambiguity on patch-level fire detection for FLAME2 dataset. It is observed that smoke remains the dominant source of false alarms, with a smoke-to-fire false positive rate of 16.32%, compared to 4.66% for background objects with similar textures. This indicates that while thermal-guided training mitigates missed detections, smoke-induced visual ambiguity continues to be the primary challenge. Moreover, while the proposed model demonstrates strong capability in detecting small-scale and even faintly visible flame regions, its generalization is influenced by the characteristics of the training data. The training dataset, though carefully curated, exhibits limited diversity in terms of geographical regions, seasonal variations, and weather conditions, which may restrict exposure to more complex wildfire scenarios. Although fire is more distinguishable in low-light conditions, further efforts are required to improve model generalization across varying weather and light conditions in future work. Additionally, flame annotations are currently generated through manual labeling. Future work will mitigate this limitation by investigating self-supervised learning strategies that reduce dependence on explicit flame annotations.

6. Conclusion

In this study, we developed a lightweight deep learning–based computer vision model for wildfire detection from an aerial view. The proposed model is tailored for resource-constrained UAV/AAV platforms. It incorporates dedicated modules to guide the model in identifying how and where to attend within aerial imagery and what visual cues to prioritize, thereby enabling effective wildfire detection and flame localization in complex forest and wildland environments. To further strengthen performance, we adopted a cross-modal knowledge distillation framework, where knowledge is transferred from a thermal teacher to the optical student network. This strategy not only reduces the dependence on additional sensing modalities and edge computing resources, but also improves the detection accuracy and robustness of the proposed model, achieving 90.97% patch-level classification accuracy, with false alarm and missed detection rates of 8.82% and 14.78%, respectively. Thus, it enhances the localization of the fire patch. The lightweight model operates with only 2.99 GFLOPS and a 0.004s inference time, demonstrating suitability for real-time edge deployment. These results highlight the framework's capability to deliver high accuracy, low-latency fire detection with precise localization, confirming its potential for practical wildfire monitoring and rapid response applications.

CRediT authorship contribution statement

Medhavi Mishra: Writing – original draft, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Sumit Mishra:** Writing – original draft, Methodology. **Hyo-Sang Shin:** Writing – review & editing, Supervision, Resources.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Abdusalomov, A., Umirzakova, S., Shukhratovich, M.B., Mukhiddinov, M., Kakhorov, A., Buriboev, A., Jeon, H.S., 2024. Drone-based wildfire detection with multi-sensor integration. Remote. Sens. 16 (24), 4651.

Alam, G.M.I., Tasnia, N., Biswas, T., Hossen, M.J., Tanim, S.A., Miah, M.S.U., 2025. Real-time detection of forest fires using FireNet-CNN and explainable AI techniques. IEEE Access.

Bouguettaya, A., Zarzour, H., Taberkit, A.M., Kechida, A., 2022. A review on early wildfire detection from unmanned aerial vehicles using deep learning-based computer vision algorithms. Signal Process. 190, 108309.

Burgos Madrigal, A., Romero Bautista, V., Díaz Hernández, R., Altamirano Robles, L., 2024. Evaluation of deformable convolution: An investigation in image and video classification. Mathematics 12 (16), 2448.

Cao, D., Chen, Z., Gao, L., 2020. An improved object detection algorithm based on multi-scaled and deformable convolutional neural networks. Human-Centric Comput. Inf. Sci. 10 (1), 14.

Chen, X., Hopkins, B., Wang, H., O'Neill, L., Afghah, F., Razi, A., Fulé, P., Coen, J., Rowell, E., Watts, A., 2022. Wildland fire detection and monitoring using a drone-collected RGB/IR image dataset. IEEE Access 10, 121301–121317.

Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y., 2017. Deformable convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision. ICCV, pp. 764–773.

Feng, Z., Guo, Y., Sun, Y., 2023. CEKD: Cross-modal edge-privileged knowledge distillation for semantic scene understanding using only thermal images. IEEE Robot. Autom. Lett. 8 (4), 2205–2212.

Feng, H., Qiu, J., Wen, L., Zhang, J., Yang, J., Lyu, Z., Liu, T., Fang, K., 2025. U3UNet: An accurate and reliable segmentation model for forest fire monitoring based on UAV vision. Neural Netw. 185, 107207.

Fouda, M.M., Sakib, S., Fadlullah, Z.M., Nasser, N., Guizani, M., 2022. A lightweight hierarchical AI model for UAV-enabled edge computing with forest-fire detection use-case. IEEE Netw. 36 (6), 38–45.

Ghali, R., Akhloufi, M.A., 2023a. CT-fire: A CNN-transformer for wildfire classification on ground and aerial images. Int. J. Remote Sens. 44 (23), 7390–7415.

Ghali, R., Akhloufi, M.A., 2023b. Deep learning approaches for wildland fires remote sensing: Classification, detection, and segmentation. Remote. Sens. 15 (7), 1821.

Ghali, R., Akhloufi, M.A., Mseddi, W.S., 2022. Deep learning and transformer approaches for UAV-based wildfire detection and segmentation. Sensors 22 (5), 1977.

Gonçalves, L.A.O., Ghali, R., Akhloufi, M.A., 2024. YOLO-based models for smoke and wildfire detection in ground and aerial images. Fire 7 (4), 140.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 770–778.

Hinton, G., Vinyals, O., Dean, J., 2015. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

Hopkins, B., O'Neill, L., Marinaccio, M., Rowell, E., Parsons, R., Flanary, S., Nazim, I., Seielstad, C., Afghah, F., 2024. FLAME 3 dataset: Unleashing the power of radiometric thermal UAV imagery for wildfire management. arXiv preprint arXiv:2412.02831.

Hou, Q., Lu, C.-Z., Cheng, M.-M., Feng, J., 2024. Conv2former: A simple transformer-style convnet for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. 46 (12), 8274–8283.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 4700–4708.

Huo, F., Xu, W., Guo, J., Wang, H., Guo, S., 2024. C2KD: Bridging the modality gap for cross-modal knowledge distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 16006–16015.

Jang, J., Ma, C., Lee, B., 2025. VL2lite: Task-specific knowledge distillation from large vision-language models to lightweight networks. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 30073–30083.

Jiang, M., Yin, S., 2023. Facial expression recognition based on convolutional block attention module and multi-feature fusion. Int. J. Comput. Vis. Robot. 13 (1), 21–37.

Jiao, Z., Zhang, Y., Xin, J., Mu, L., Yi, Y., Liu, H., Liu, D., 2019. A deep learning based forest fire detection approach using UAV and YOLOv3. In: Proceedings of the 1st International Conference on Industrial Artificial Intelligence. IAI, pp. 1–5.

Jonnalagadda, A.V., Hashim, H.A., 2024. SegNet: A segmented deep learning based convolutional neural network approach for drones wildfire detection. Remote. Sens. Appl.: Soc. Environ. 34, 101181.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D., 2020. Supervised contrastive learning. In: Advances in Neural Information Processing Systems, vol. 33, pp. 18661–18673.

Kim, S.-Y., Muminov, A., 2023. Forest fire smoke detection based on deep learning approaches and unmanned aerial vehicle images. Sensors 23 (12), 5702.

Li, J., Wan, J., Sun, L., Hu, T., Li, X., Zheng, H., 2025. Intelligent segmentation of wildfire region and interpretation of fire front in visible light images from the viewpoint of an unmanned aerial vehicle (UAV). ISPRS J. Photogramm. Remote Sens. 220, 473–489.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision. ICCV, pp. 2980–2988.

van der Maaten, L., Hinton, G., 2008. Visualizing data using t-SNE. J. Mach. Learn. Res. 9, 2579–2605.

Mahdi, A.S., Mahmood, S.A., 2022. An edge computing environment for early wildfire detection. Ann. Emerg. Technol. Comput. (AETiC) 6 (3), 56–68.

Marinaccio, M., Afghah, F., 2025. Seeing heat with color–RGB-only wildfire temperature inference from SAM-guided multimodal distillation using radiometric ground truth. arXiv preprint arXiv:2505.01638.

Mishra, S., Mishra, M., Shyam, P., Har, D., 2025. TIME-VAD: Text-informed magnitude enhancement feature learning for vehicle accident detection and anticipation. IEEE Trans. Intell. Transp. Syst..

Mishra, S., Rajendran, P.K., Vecchietti, L.F., Har, D., 2023. Sensing accident-prone features in urban scenes for proactive driving and accident prevention. IEEE Trans. Intell. Transp. Syst. 24 (9), 9401–9414.

Mu, L., Yang, Y., Wang, B., Zhang, Y., Feng, N., Xie, X., 2025. Edge computing-based real-time forest fire detection using UAV thermal and color images. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens..

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., 2021. Learning transferable visual models from natural language supervision. In: Proceedings of the International Conference on Machine Learning. ICML, pp. 8748–8763.

Rajoli, H., Khoshdel, S., Afghah, F., Ma, X., 2024. FlameFinder: Illuminating obscured fire through smoke with attentive deep metric learning. IEEE Trans. Geosci. Remote Sens..

Ramos, L., Casas, E., Bendek, E., Romero, C., Rivas-Echeverría, F., 2024. Computer vision for wildfire detection: A critical brief review. Multimedia Tools Appl. 83 (35), 83427–83470.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems (NeurIPS). pp. 91–99.

Ribeiro, T.F.R., Silva, F., Moreira, J., de C. Costa, R.L., 2023. Burned area semantic segmentation: A novel dataset and evaluation using convolutional networks. ISPRS J. Photogramm. Remote Sens. 202, 565–580.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.-C., 2018. MobileNetV2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 4510–4519.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision. ICCV, pp. 618–626.

Setyanto, A., Sasongko, T.B., Fikri, M.A., Kim, I.K., 2023. Near-edge computing aware object detection: A review. IEEE Access 12, 2989–3011.

Shamsoshoara, A., Afghah, F., Razi, A., Zheng, L., Fulé, P.Z., Blasch, E., 2021. Aerial imagery pile burn detection using deep learning: The flame dataset. Comput. Netw. 193, 108001.

Shipard, J., Wiliem, A., Nguyen, K.T., Xiang, W., Fookes, C., 2025. Zoom-shot: Fast, efficient and unsupervised zero-shot knowledge transfer from CLIP to vision encoders. Pattern Recognit. 112323.

Shyam, P., Mishra, S., Yoon, K.-J., Kim, K.-S., 2022. Infra sim-to-real: An efficient baseline and dataset for infrastructure based online object detection and tracking using domain adaptation. In: Proceedings of the IEEE Intelligent Vehicles Symposium. IV, pp. 1393–1399.

Stival, L., da Silva Torres, R., Pedrini, H., 2025. Semantically-aware contrastive learning for multispectral remote sensing images. ISPRS J. Photogramm. Remote Sens. 223, 173–187.

Su, W., Li, L., Liu, F., He, M., Liang, X., 2022. AI on the edge: A comprehensive review. Artif. Intell. Rev. 55 (8), 6125–6183.

Tan, M., Le, Q., 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. In: Proceedings of the International Conference on Machine Learning. ICML, pp. 6105–6114.

Tong, X., Guo, X., Sun, X., Guo, R., Su, S., Zuo, Z., 2024. Cmdistill: Cross-modal distillation framework for UAV image object detection. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens..

Tong, H., Yuan, J., Zhang, J., Wang, H., Li, T., 2024. Real-time wildfire monitoring using low-altitude remote sensing imagery. Remote. Sens. 16 (15), 2827.

Wang, H., Bai, Y., Li, H., 2023. YOLOv5s-fire: A lightweight model for flame detection. In: Proceedings of the 6th International Conference on Image and Graphics Processing. pp. 51–57.

Wang, G., Bai, D., Lin, H., Zhou, H., Qian, J., 2024. FireViTNet: A hybrid model integrating ViT and CNNs for forest fire segmentation. Comput. Electron. Agric. 218, 108722.

Woo, S., Park, J., Lee, J.-Y., Kweon, I.S., 2018. CBAM: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 3–19.

Yang, H., Wang, J., Wang, J., 2023. Efficient detection of forest fire smoke in UAV aerial imagery based on an improved YOLOv5 model and transfer learning. Remote. Sens. 15 (23), 5527.

Yar, H., Ullah, W., Khan, Z.A., Baik, S.W., 2023. An effective attention-based CNN model for fire detection in adverse weather conditions. ISPRS J. Photogramm. Remote Sens. 206, 335–346.

Zachariadis, E., Antonakakis, M., Zervakis, M., 2023. WildForest fire detection with shre-xception network on aerial optical and infrared images. In: Proceedings of the IEEE International Conference on Imaging Systems and Techniques. IST, pp. 1–6.

Zhang, Q., Xu, J., Xu, L., Guo, H., 2016. Deep convolutional neural networks for forest fire detection. In: Proceedings of the International Forum on Management, Education and Information Technology Applications. pp. 568–575.

Zhang, Q., Zhu, J., Dong, Y., Zhao, E., Song, M., Yuan, Q., 2025. 10-Minute forest early wildfire detection: Fusing multi-type and multi-source information via recursive transformer. Neurocomputing 616, 128963.

Zhang, Q., Zhu, J., Huang, Y., Yuan, Q., Zhang, L., 2023. Beyond being wise after the event: Combining spatial, temporal and spectral information for himawari-8 early-stage wildfire detection. Int. J. Appl. Earth Obs. Geoinf. 124, 103506.

Zhou, S., Liu, W., Hu, C., Zhou, S., Ma, C., 2023. Unidistill: A universal cross-modality knowledge distillation framework for 3D object detection in bird's-eye view. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 5116–5125.

Zhu, J., Fang, L., Ghamisi, P., 2018. Deformable convolutional neural networks for hyperspectral image classification. IEEE Geosci. Remote. Sens. Lett. 15 (8), 1254–1258.



Medhavi Mishra received B.Tech. in Production and Industrial Engineering from Motilal Nehru National Institute Of Technology, India, in 2016, and M.S. degree in Mobility from Korea Advanced Institute of Science and Technology, South Korea, in 2025, where she is currently pursuing a Ph.D. degree. Earlier, she worked with HighRadius Technologies Pvt. Ltd. as a Techno-functional Consultant. Her research interests include multi-modal perception, edge AI for resource-constrained autonomous systems, vision-language foundation models, knowledge distillation and transfer in deep learning.



Sumit Mishra received B.Tech. in Electronics and Communication Engineering from Dr. A.P.J. Abdul Kalam Technical University, India, in 2016, and M.S. degree from the Robotics department, Korea Advanced Institute of Science and Technology, South Korea, in 2023, where he is currently pursuing a Ph.D. degree. Earlier, he worked with Learnogether Technologies Pvt. Ltd. as a Research Consultant. His research interests include vision-language models, multimodal foundation models, cross-modal alignment by deep learning, explainable AI for safety-critical systems, and computer vision for intelligent transportation systems.



Hyo-Sang Shin received his B.Sc. in aerospace engineering from Pusan National University in 2004, his M.Sc. in aerospace engineering from KAIST in 2006, and his Ph.D. in cooperative missile guidance from Cranfield University in 2010. He is a Professor at KAIST and an Adjunct Professor in GNC at Cranfield University. At KAIST, he leads the Autonomous and Intelligent Systems Group, focusing on advanced/urban air mobility , unmanned traffic management, sensor fusion, and data-driven guidance and control. He is also active in the global research community, serving on technical, program, and editorial committees.



https://dspace.org/

Cross-modal distillation for real-time wildfire detection and localization in edge-deployed aerial vehicles

Mishra, Medhavi; Mishra, Sumit; Shin, Hyo-Sang

2026-05
Elsevier
https://doi.org/10.1016/j.isprsjprs.2026.03.019
Mishra M, Mishra S, Shin H-S. (2026) Cross-modal distillation for real-time wildfire detection and localization in edge-deployed aerial vehicles. ISPRS Journal of Photogrammetry and Remote Sensing, Volume 235, May 2026, pp. 551-564

Downloaded from a DSpace repository. https://dspace.org/