

Article

Decoupled Topology Distance Distillation for Lightweight Smoke Detection in Aerial Remote Sensing Images

Dongyin Lai, Lin Liu, Juanxiu Liu, Jing Zhang, Xiaohui Du, Ruqian Hao and Xudong Wang *

School of Optoelectronic Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China; 202212050835@std.uestc.edu.cn (D.L.); ruqian_hao@uestc.edu.cn (R.H.)

* Correspondence: wang_xudong01@yeah.net

Abstract

Early aerial smoke detection is vital for wildfire response, but deploying accurate two-stage deep detectors on resource-limited Unmanned Aerial Vehicles (UAVs) remains computationally prohibitive. Moreover, under uniform supervision, standard knowledge distillation struggles on aerial smoke data, where foreground–background imbalance is severe and smoke boundaries are visually ambiguous. To resolve this, we propose the Decoupled Topology Distance Distillation (DeTD) framework to compress two-stage smoke detectors for real-time edge inference. DeTD features three key innovations. First, a decoupling module uses ground-truth-derived binary masks to isolate smoke and background features, mitigating distillation class imbalance. Second, a topology distance distillation module projects these decoupled features onto a unit hypersphere, employing a novel Symmetric Triplet Loss. This jointly optimizes the intra-class compactness and inter-class separability of both the foreground and background relational geometry between the teacher and student networks. Third, prediction-head soft-label distillation transfers class-conditional knowledge, synergistically complementing the intermediate-feature distillation. Comprehensive experiments on the D-Fire benchmark and a custom aerial UAV dataset yield mAP50 scores of 67.4% and 70.6%, respectively. DeTD consistently outperforms thirteen recent distillation baselines, and the lightweight student attains real-time-compatible inference, narrowing the accuracy–efficiency gap and indicating feasibility for deployment on resource-constrained UAV edge hardware.

Keywords: knowledge distillation; smoke detection; remote sensing; foreground–background decoupling; triplet loss; edge deployment; model compression



Academic Editors: Washington Franca-Rocha, António Vieira and Marcos Francos

Received: 6 May 2026

Revised: 27 June 2026

Accepted: 29 June 2026

Published: 6 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Wildfires represent one of the most destructive natural hazards worldwide, causing extensive ecological damage, significant human casualties, and substantial economic losses each year [1]. In recent years, the frequency and severity of wildfire events have increased markedly, driven in part by prolonged drought conditions and rising ambient temperatures associated with climate change. Early and accurate detection of smoke (typically the first visible indicator of ignition) is therefore indispensable for triggering swift suppression responses and mitigating downstream damage. UAVs have emerged as highly effective aerial monitoring platforms for large-area surveillance tasks [2,3], offering flexible deployment, on-demand coverage, and relatively low operational cost compared with manned aircraft or fixed-position camera arrays.

The combination of onboard deep learning inference with UAV platforms, commonly termed edge-AI, has opened new possibilities for autonomous wildfire monitoring [4]. By executing detection algorithms directly on the vehicle rather than relying on cloud-based processing, edge-AI systems eliminate round-trip communication latency and maintain operational capability even when wireless connectivity is unavailable [5]. However, realizing such capability in practice is challenging: modern high-accuracy object detectors, particularly two-stage anchor-based architectures such as Faster R-CNN [6], carry millions of parameters and require hundreds of giga-floating-point operations (GFLOPs) per inference pass, which far exceed the computational and energy budgets of embedded processors and mobile GPUs typically mounted on UAVs.

Existing approaches to mitigating this computational burden include network pruning [7], quantization [8], architecture search [9], and knowledge distillation (KD) [10,11]. Among these, knowledge distillation—the paradigm in which a compact student network is trained to mimic the behavior of a larger, more capable teacher network—has attracted sustained research interest because it imposes no architectural constraints on the student and incurs no inference-time overhead. For image classification, seminal work by Hinton et al. [10] demonstrated that soft probability outputs from a teacher encode rich inter-class relational information that substantially benefits student learning. Subsequent research has extended distillation to the object detection task by exploiting intermediate feature representations [12–14], attention maps [15,16], and output-level predictions, achieving compelling compression ratios while preserving detector accuracy.

Despite notable progress, applying knowledge distillation to smoke detection in aerial imagery introduces a set of domain-specific challenges that standard detection distillation methods do not adequately address.

- (i) **Foreground–background imbalance:** Smoke objects typically occupy a minority of pixels in any given image frame, particularly during high-altitude surveillance, where even large fire events appear as small, diffuse plumes. When distillation supervision is applied uniformly across the entire feature map, the overwhelming presence of background pixels dominates the gradient signal, starving the network of informative smoke-region supervision.
- (ii) **Boundary ambiguity:** Smoke lacks well-defined, rigid shape boundaries. Its appearance varies continuously as a function of wind, humidity, fuel type, and camera angle, making it visually similar to clouds, haze, and dusty terrain [17–19]. This ambiguity complicates feature-level distillation, where standard L2 regression losses aggressively penalize any discrepancy between teacher and student feature values—including discrepancies attributable to the inherent noise in boundary regions rather than to a true knowledge gap.
- (iii) **Feature magnitude inconsistency:** The teacher and student networks, differing in depth and capacity, develop feature distributions with substantially different magnitude ranges [20]. Direct feature imitation losses that do not account for this distributional mismatch introduce spurious optimization targets.

To overcome these limitations, we propose a Decoupled Topology Distance Distillation (DeTD) framework. Our principal insight is that decoupling foreground and background supervision, and then modeling their relational geometry on a unit hypersphere via a Symmetric Triplet Loss, provides a more faithful and noise-robust knowledge transfer signal than undifferentiated feature regression. The specific contributions of this work are summarized as follows:

- (1) We propose a Foreground–Background Decoupling module that constructs binary masks from ground-truth annotations to separately extract smoke-region and

background-region intermediate features, explicitly counteracting the class-imbalance bias that degrades existing distillation methods on smoke detection data.

- (2) We design a Decoupled Topology Distance Distillation (DeTD) module that aggregates decoupled features, projects them onto a high-dimensional unit hypersphere, and applies a Symmetric Triplet Loss to jointly optimize foreground-centric and background-centric distillation objectives. This enables the student network to learn discriminative smoke representations while retaining sensitivity to background context.
- (3) We incorporate temperature-scaled soft-label head distillation at the RoIHead classification branch, transferring class-conditional probability distributions from the teacher network to the student to further refine classification accuracy.
- (4) We conduct thorough experiments on the D-Fire public benchmark and a self-constructed aerial UAV smoke dataset, including comparisons against thirteen distillation baselines spanning 2015–2025, ablation studies, hyperparameter sensitivity analyses, and edge-deployment profiling.

2. Related Work

2.1. Smoke and Fire Detection in Aerial Remote Sensing

Early detection approaches for fire and smoke relied on scalar sensor arrays—photoelectric and ionization chambers mounted at fixed sites—whose restricted spatial coverage rendered them ineffective for open-environment monitoring [21]. The advent of vision-sensor-based systems brought considerably broader coverage, leading to decades of research on color-space segmentation, optical-flow motion analysis, and texture-based classifiers for fire and smoke recognition [22,23]. While computationally frugal, these conventional machine learning pipelines proved brittle in the presence of illumination changes, camera motion, and confounding objects whose color or texture resembles smoke—for instance, clouds, morning fog, and dust clouds.

Deep learning methods have largely supplanted handcrafted-feature approaches in recent years, with convolutional neural networks demonstrating strong generalization across diverse environmental conditions [24,25]. Classification-level architectures applied to the FLAME [26] and D-Fire [27] datasets achieve high image-level fire/no-fire accuracy; however, they offer no spatial localization of the fire source, limiting actionability for emergency responders. Object detection formulations—which predict bounding boxes around smoke or flame regions—address this limitation and have been applied to aerial imagery through YOLO-series variants [28,29], two-stage Faster R-CNN models [30], and attention-enhanced networks [31]. Segmentation-based methods push localization granularity to the pixel level [32,33], but incur high computational costs that are incompatible with edge deployment on UAV hardware with constrained power budgets. The FireViTNet [34] and FireFormer [35] architectures integrate vision transformer modules for improved long-range dependency modeling in fire scene understanding, while U3UNet [36] introduces a dual-branch U-shaped segmentation network for forest fire monitoring. Despite these advances, the challenge of deploying high-accuracy detection models on edge hardware at real-time frame rates remains largely unresolved.

Multi-modal sensing, which fuses visible-spectrum and thermal-infrared imagery, has been explored as a means of circumventing the visual ambiguity inherent in single-modality optical data [37,38]. Cross-modal knowledge distillation—training an optical-spectrum student with supervision from a thermal-spectrum teacher—represents an effective solution because it retains the advantage of thermal information during training while requiring only a single-modality sensor during deployment [39,40]. Nonetheless, cross-modal approaches

require paired training data and address a somewhat different problem than the intra-modal model compression scenario studied here.

2.2. Knowledge Distillation for Object Detection

The KD paradigm introduced by Hinton et al. [10] uses temperature-scaled soft class probabilities as training targets for a student, revealing the teacher's implicit inter-class similarity structure beyond what hard one-hot labels convey. This output-based distillation has since been extended with decoupled loss formulations [41] and logit standardization [42] to further improve classification transferability. Feature-based distillation methods go further by aligning intermediate representations: FitNet [43] performs layer-to-layer Euclidean feature regression; FGFI [12] argues that feature discrepancies near bounding-box locations carry more discriminative value and restrict distillation to such regions; FKD [14] applies joint spatial and channel attention masks; and FRS [17] decomposes the distillation target into focal regions and globally contextualized background.

The foreground–background imbalance problem in detection distillation has received dedicated attention. DeFeat [13] assigns differential weighting to foreground and background channels, recognizing that both contribute complementary information to detector performance. Relation-based distillation methods capture structural knowledge among samples: InsDist [44] constructs instance-aware imitation masks that encode similarity between class prototypes and pixel embeddings, additionally incorporating inter-instance relation distillation among foreground instances. ARSD [45] extends region distillation through a binarized spatial mask strategy tailored to the scale diversity characteristic of remote sensing images. AFD [46] combines spatial and channel attention into a compound feature mask to guide the student's focus in optical remote sensing scenes. The Two-Way [47] approach designs bidirectional assistant modules that simultaneously enrich student features and simplify teacher features, reducing the effective teacher–student capacity gap. The most recent DPGD framework [20] proposes dual prediction-guided distillation combining instance-wise and pixel-wise relation distillation with output distillation, achieving state-of-the-art results on multiple remote sensing detection benchmarks. Despite these advances, none of the above methods were designed with the specific structural properties of aerial smoke—diffuse boundaries, extreme foreground sparsity, and topological similarity to non-fire backgrounds—in mind, leaving meaningful performance headroom in this application domain.

2.3. Lightweight Models for Edge Deployment

Beyond distillation, complementary compression techniques include structured and unstructured pruning [7], post-training and quantization-aware training [8], and lightweight architecture design [9]. MobileNetV2 [48] and MobileViT [49] exemplify architecture-level efficiency through depth-wise separable convolutions and hybrid transformer-convolution blocks, respectively, achieving competitive accuracy at a fraction of the parameter count of standard ResNets [50]. Knowledge distillation is particularly attractive in conjunction with lightweight architectures because it can recover accuracy lost during architectural simplification without modifying the inference graph. Several recent works demonstrate that distilled lightweight detectors can achieve real-time performance on Jetson-class embedded processors and FPGA platforms while meeting the mAP requirements of practical wildfire monitoring systems [51,52]. The present study targets the Faster R-CNN [6] family with a ResNet-18 [50] student backbone, balancing detection accuracy with deployment feasibility on edge hardware.

In summary, three observations motivate the present work. First, although deep detectors have decisively outperformed handcrafted-feature pipelines for fire and smoke

recognition [24,25], their parameter and FLOP budgets remain incompatible with real-time inference on UAV edge hardware [51,52]. Second, knowledge distillation is an attractive route to compression because it leaves the student's inference graph unchanged [10,11], yet the prevailing detection-distillation methods—including the most recent remote-sensing frameworks ARSD [45], AFD [46], Two-Way [47], and DPGD [20]—apply their supervision over dense feature maps and were not designed for the specific statistics of aerial smoke. Third, aerial smoke simultaneously exhibits extreme foreground sparsity, diffuse low-contrast boundaries, and strong appearance similarity to non-fire backgrounds, a combination that exposes three weaknesses of uniform feature regression: imbalance-dominated gradients, boundary-noise over-penalization, and sensitivity to teacher–student feature-magnitude divergence. The research gap addressed here is therefore the absence of a distillation framework that (i) explicitly separates foreground and background supervision to neutralize the imbalance, and (ii) transfers a magnitude-invariant, relational representation rather than absolute feature values. DeTD is designed to fill precisely this gap.

3. Methodology

3.1. Overall Framework

The proposed DeTD framework employs a standard teacher–student distillation paradigm built on the two-stage Faster R-CNN detection architecture. As depicted in Figure 1, both teacher and student share the same image input during training, processing it through their respective backbone networks and neck layers. The teacher network—instantiated with a ResNet-101 backbone—is pretrained on the target dataset and kept frozen throughout student training. The student network—instantiated with a ResNet-18 backbone—is trained from scratch under the composite supervision of distillation losses and standard detection task losses.

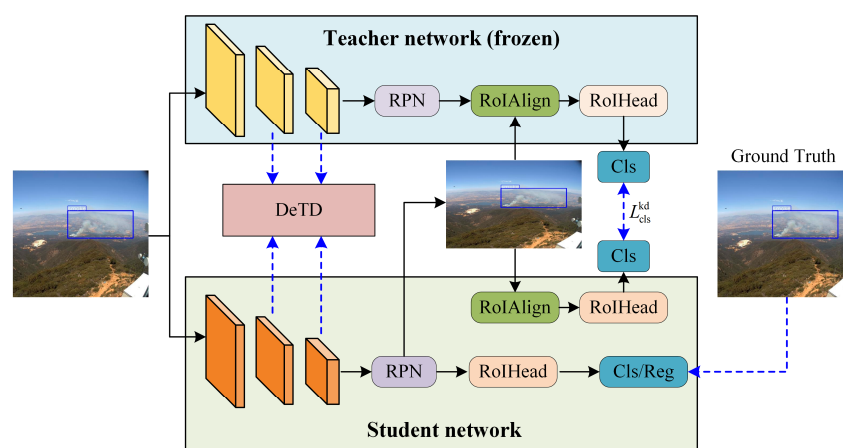


Figure 1. Overview of the proposed DeTD pipeline. The frozen teacher network (ResNet-101 Faster R-CNN) processes the input image through its backbone, RPN, and RoIAlign modules to yield intermediate features and classification outputs. The DeTD module receives backbone feature maps from both teacher and student branches, applies decoupled topology distance distillation, and passes a classification distillation loss to the student's prediction head. The student network (ResNet-18 Faster R-CNN) is trained jointly by the distillation losses and standard detection task losses (Cls/Reg and RPN).

The ResNet-101 teacher/ResNet-18 student pairing is chosen to balance two competing requirements. On the teacher side, ResNet-101 is a strong, widely adopted detection backbone whose accuracy provides an informative distillation target; on the student side, ResNet-18 is a compact backbone whose parameter and FLOP budget is compatible with

UAV onboard inference. The resulting capacity ratio ($\approx 2.1 \times$ in parameters) is a moderate, well-matched gap: the knowledge-distillation literature reports that an excessively strong teacher can be harder for a low-capacity student to mimic, so that the largest teacher does not necessarily yield the best student. We deliberately avoid this regime and empirically confirm in Section 4.5.5 that ResNet-101 is the most effective teacher for the ResNet-18 student among the backbones tested, and that ResNet-18 offers the best accuracy–efficiency balance among candidate students.

Distillation operates at two levels: (1) the intermediate feature level, where the DeTD module transfers structural knowledge from backbone feature maps via a decoupled Symmetric Triplet Loss; and (2) the prediction head level, where soft classification probabilities from the teacher’s RoIHead branch supervise the student’s corresponding branch via a Kullback–Leibler divergence loss. Detection task losses—comprising RoIHead classification and regression losses, and Region Proposal Network (RPN) classification and regression losses—provide task-specific supervision that grounds the student in the target detection objective. The overall training loss integrates all five components with empirically calibrated weighting coefficients.

Unlike methods that distil over dense Feature Pyramid Network (FPN) maps, DeTD first aggregates features into compact foreground and background descriptors. This lowers the dimensionality of the distillation target, mitigates the teacher–student feature-magnitude divergence, and concentrates the optimization on the foreground–background relational structure.

3.2. Foreground–Background Feature Decoupling

Mainstream two-stage detectors produce feature maps from the backbone that conflate foreground object regions with an often vast expanse of background pixels. In aerial smoke imagery, where smoke plumes may occupy only 3–15% of the image area depending on altitude and fire extent, undifferentiated feature distillation losses are overwhelmingly driven by background gradients [13]. The Foreground–Background Feature Decoupling module, illustrated in Figure 2, addresses this by constructing separate foreground and background feature sets from the intermediate backbone feature maps using ground-truth annotations.

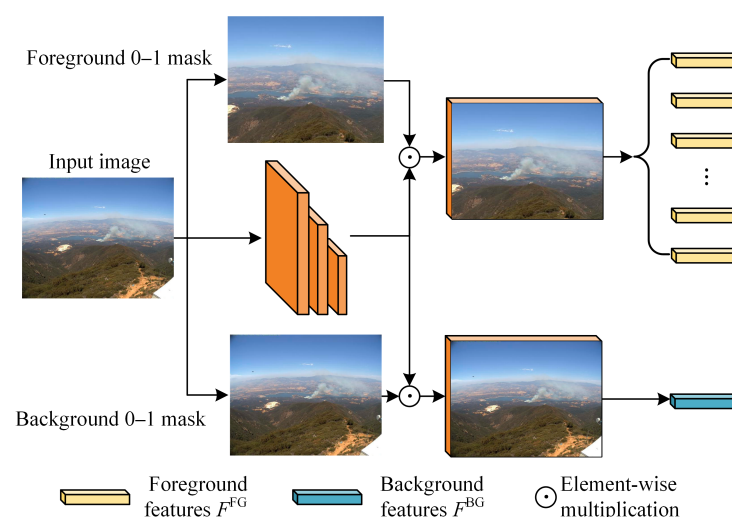


Figure 2. Foreground–background feature decoupling module. Binary foreground (0–1) and background (0–1) masks are generated from ground-truth annotations and applied elementwise to the student backbone feature maps. The resulting foreground feature set F^{FG} contains one descriptor per annotated object instance, while the background feature F^{BG} represents the non-object regions. Both feature sets are passed to the DeTD module for topology distance distillation.

Let $F \in \mathbb{R}^{C \times H \times W}$ denote an intermediate feature map from the backbone, where C is the number of channels and H, W are the spatial dimensions. Throughout, $\tilde{F}$ denotes masked or cropped feature tensors, F the pooled descriptor matrices derived from them, and $\tilde{F}$ the corresponding $\uparrow_2$ -normalized descriptors; subscripts s and t index the student and teacher networks, respectively. Owing to the backbone's downsampling factor S , each o -th ground-truth bounding box annotation $B_o = (x_1^{(o)}, y_1^{(o)}, x_2^{(o)}, y_2^{(o)})$ is rescaled to the feature-map coordinate system as $\tilde{B}_o = B_o/S$. The global foreground binary mask $M^{FG} \in \{0, 1\}^{H \times W}$ is then defined pointwise as:

$$M_{i,j}^{FG} = I \left[(i, j) \in \bigcup_{o=1}^{N_{obj}} \tilde{B}_o \right] \quad (1)$$

where $I[\cdot]$ is the indicator function that equals 1 when position (i, j) lies within any rescaled ground-truth box, and 0 otherwise. The complementary global background mask is defined as:

$$M_{i,j}^{BG} = 1 - M_{i,j}^{FG} \quad (2)$$

Applying these masks element-wise to the feature map separates it into foreground and background components. The foreground branch crops one feature region per ground-truth instance from the corresponding box, while the background branch retains the entire non-object region through the background mask:

$$\begin{aligned} \tilde{F}_o^{FG} &= \text{Crop}(F, \tilde{B}_o) \in \mathbb{R}^{C \times H_o \times W_o}, \quad o = 1, \dots, N_{obj} \\ \tilde{F}^{BG} &= M^{BG} \odot F \in \mathbb{R}^{C \times H \times W}, \quad \tilde{F}^{FG} = \left\{ \tilde{F}_o^{FG} \right\}_{o=1}^{N_{obj}} \end{aligned} \quad (3)$$

where N_{obj} is the number of ground-truth boxes in the image, $\text{Crop}(F, \tilde{B}_o)$ extracts the sub-tensor of F delimited by box $\tilde{B}_o$ (spatial size $H_o \times W_o$), and $\odot$ denotes element-wise multiplication in which the $H \times W$ mask is broadcast across all channels. The complementary mask M^{BG} from Equation (2) guarantees that the two branches partition the feature map without overlap. The full extraction flow is shown in Figure 2. Separating the feature space in this manner ensures that subsequent distillation supervision is calibrated to smoke regions and background regions independently, preventing the numerically dominant background from suppressing the informative foreground distillation signal.

3.3. Decoupled Topology Distance Distillation (DeTD)

Having obtained decoupled foreground and background features, the DeTD module—illustrated in Figure 3—reconstructs their mutual relational geometry and transfers it from teacher to student via a Symmetric Triplet Loss. The module operates in three sequential stages: feature aggregation, normalization, and triplet-based relational distillation.

3.3.1. Feature Aggregation

Each instance-level foreground tensor $\tilde{F}_o^{FG} \in \mathbb{R}^{C \times H_o \times W_o}$ is first reduced to a compact descriptor vector by global average pooling, $\text{AVG}(X) = \frac{1}{H_X W_X} \sum_{i,j} X_{:,i,j} \in \mathbb{R}^C$, for $X \in \mathbb{R}^{C \times H_X \times W_X}$. The N_{obj} pooled descriptors are then stacked column-wise along the instance dimension to form a unified foreground descriptor matrix:

$$F^{FG} = \left[\text{AVG}(\tilde{F}_1^{FG}) \parallel \text{AVG}(\tilde{F}_2^{FG}) \parallel \dots \parallel \text{AVG}(\tilde{F}_{N_{obj}}^{FG}) \right] \in \mathbb{R}^{C \times N_{obj}} \quad (4)$$

where $\parallel$ denotes concatenation along the instance dimension; the N_{obj} pooled descriptors thus form the columns of F^{FG} . For the background feature tensor $\tilde{F}^{BG}$, global average

pooling yields a single background descriptor vector, which is subsequently expanded along the instance dimension to match the foreground descriptor dimensionality:

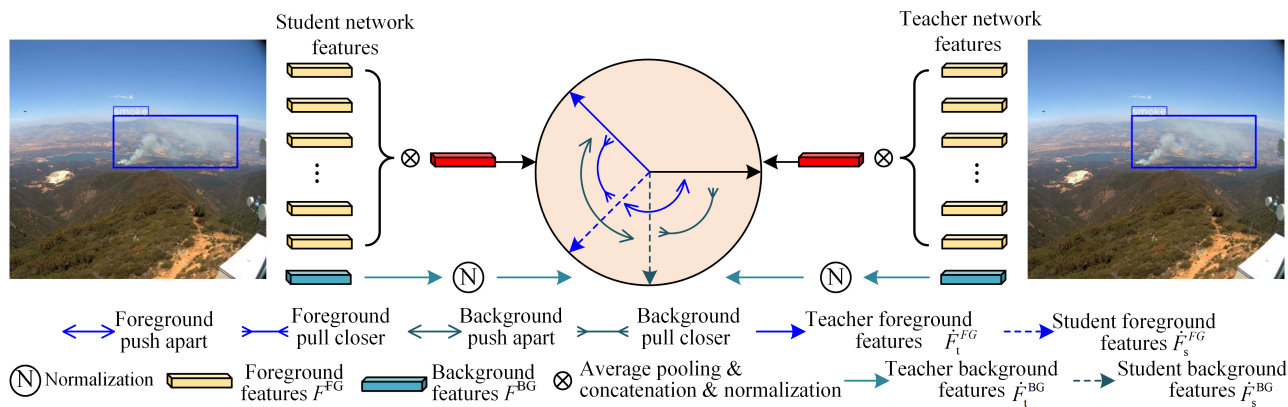


Figure 3. Decoupled topology distance distillation (DeTD) module. Foreground and background features from the student and teacher networks are aggregated, concatenated, and normalized onto a unit hypersphere. The Symmetric Triplet Loss operates in two directions: foreground-anchor triplets pull student foreground features toward teacher foreground features while pushing them away from background features (blue arrows); background-anchor triplets pull student background features toward teacher background features while pushing them away from foreground features (green arrows). N denotes L2 normalization.

$$F^{BG} = \text{Expand}\left(\text{AVG}\left(\tilde{F}^{BG}\right), N_{obj}\right) \in \mathbb{R}^{C \times N_{obj}} \quad (5)$$

where $\text{AVG}\left(\tilde{F}^{BG}\right) \in \mathbb{R}^C$ is the single background descriptor and $\text{Expand}(v, N)$ replicates the column vector v into N identical columns. Both F^{FG} and F^{BG} therefore share the shape $\mathbb{R}^{C \times N_{obj}}$.

3.3.2. Unit Hypersphere Normalization

Smoke targets exhibit pronounced visual overlap with background textures such as clouds, haze, and light-colored terrain surfaces. Distillation objectives that operate on raw feature magnitudes are sensitive to this overlap, as well as to the well-documented feature magnitude divergence between teacher and student networks. To obtain scale-invariant, topology-preserving descriptors, both foreground and background features are L2-normalized to the unit hypersphere along the channel dimension:

$$\dot{F}_{:,o}^{FG} = \frac{F_{:,o}^{FG}}{\|F_{:,o}^{FG}\|_2}, \dot{F}_{:,o}^{BG} = \frac{F_{:,o}^{BG}}{\|F_{:,o}^{BG}\|_2}, o = 1, \dots, N_{obj} \quad (6)$$

where $F_{:,o}^{FG} \in \mathbb{R}^C$ is the o -th column (the descriptor of instance o), and $\|\cdot\|_2$ is the Euclidean norm taken over the channel dimension; every column of $\dot{F}^{FG}$ and $\dot{F}^{BG}$ thus lies on the unit hypersphere $\mathbb{S}^{C-1}$. Unit normalization removes magnitude information while preserving directional (angular) relationships. This is particularly appropriate for distillation because what we seek to transfer is the relational geometry of the representation space—how smoke features relate to background features—rather than their absolute values, which are arbitrary consequences of initialization and optimization trajectory.

Projection onto the unit hypersphere is the design choice that makes the distilled signal magnitude-invariant: because teachers and students of different depths develop different feature norms, transferring only the angular (directional) configuration isolates the relational geometry we wish to preserve from the absolute scale we wish to ignore.

3.3.3. Symmetric Triplet Distillation Loss

With normalized foreground and background descriptors extracted from both the student network ($\dot{F}_s^{FG}, \dot{F}_s^{BG}$) and the teacher network ($\dot{F}_t^{FG}, \dot{F}_t^{BG}$), the DeTD module applies a symmetric triplet loss to jointly optimize the foreground–background relationship in the student’s representation space.

Triplet loss [53], widely used in metric learning and face recognition, trains a network by specifying an anchor sample, a positive sample (same class as the anchor), and a negative sample (different class), and penalizes configurations where the anchor–negative distance is not at least a margin m larger than the anchor–positive distance. Here, we construct two complementary triplets:

Foreground–anchor triplet: the student foreground descriptor serves as the anchor (x^a), the teacher foreground descriptor as the positive (x^p), and the teacher background descriptor as the negative (x^n). Intuitively, the student’s smoke representation should be pulled toward the teacher’s smoke representation while being pushed away from the background representation:

$$L_{tri}^{fg} = \frac{1}{N_{obj}} \sum_{o=1}^{N_{obj}} L_{TL} \left(\dot{F}_{s,o}^{FG}, \dot{F}_{t,o}^{FG}, \dot{F}_{t,o}^{BG} \right) \quad (7)$$

Background–anchor triplet: An analogous triplet is constructed with the student background descriptor as anchor, the teacher background descriptor as positive, and the teacher foreground descriptor as negative. This prevents the student from neglecting background representation quality—a crucial property for suppressing false alarms in smoke detection:

$$L_{tri}^{bg} = \frac{1}{N_{obj}} \sum_{o=1}^{N_{obj}} L_{TL} \left(\dot{F}_{s,o}^{BG}, \dot{F}_{t,o}^{BG}, \dot{F}_{t,o}^{FG} \right) \quad (8)$$

where the triplet loss function L_{TL} is defined as:

$$L_{TL}(x^a, x^p, x^n) = \left[\|x^a - x^p\|_2^2 - \|x^a - x^n\|_2^2 + m \right]_+ \quad (9)$$

where $[\cdot]_+ = \max(\cdot, 0)$ and m are the triplet margin hyperparameters. In Equations (7) and (8), each $\dot{F}$ term is the o -th instance column of the corresponding normalized descriptor matrix (subscripts s and t index the student and teacher networks), and the per-instance triplet loss is averaged over the N_{obj} foreground instances. We adopt a margin-based triplet rather than direct regression (MSE/cosine) because the objective is relational—what should be preserved is the separation between smoke and background, not the literal feature values; the margin m enforces this separation explicitly. The overall DeTD loss combines both triplets with weighting coefficients:

$$L_{tri} = \alpha_{fg} \cdot L_{tri}^{fg} + \alpha_{bg} \cdot L_{tri}^{bg} \quad (10)$$

The symmetry is deliberate: distilling the foreground triplet alone leaves the student’s background representation unconstrained and inflates the false-positive rate in cluttered aerial scenes, whereas the background–anchor triplet keeps this representation discriminative.

3.4. Prediction Head Knowledge Distillation

In addition to the intermediate-feature topology distillation, the student’s RoIHead classification branch receives soft supervision from the teacher’s RoIHead output. The teacher and student share the same region proposals—the student’s RPN proposal set—

ensuring that both branches process identical candidate regions and that classification differences are attributable solely to the capacity gap rather than differing spatial inputs.

Let $P_k^s \in \mathbb{R}^C$ and $P_k^t \in \mathbb{R}^C$ denote the classification score vectors for the k -th region proposal from the student and teacher networks, respectively, where C is the number of foreground classes plus background. Temperature-scaled softmax probabilities are computed as:

$$\dot{P}_{c,k}^{s/T} = \frac{\exp(P_{c,k}^s/T)}{\sum_{j=1}^C \exp(P_{j,k}^s/T)}, c = 1, 2, \dots, C \quad (11)$$

and analogously for the teacher's predictions $\dot{P}_{c,k}^{t/T}$.

The prediction head distillation loss is then defined using the Kullback–Leibler (KL) divergence, treating the teacher's soft output as the target distribution:

$$L_{cls}^{kd} = \frac{1}{K} \sum_{k=1}^K L_{KL}(\dot{P}_k^{t/T} \parallel \dot{P}_k^{s/T}) \quad (12)$$

where K is the number of region proposals, $L_{KL}(P \parallel Q) = \sum_{c=1}^C P_c \log(P_c/Q_c)$ is the Kullback–Leibler divergence, and T is the temperature parameter that controls the smoothness of the soft targets. Larger T values produce softer distributions that encode finer inter-class similarity structure, while small T values approach hard-label training.

3.5. Overall Loss Function

The total training objective integrates five loss components: the DeTD topology distillation loss L_{tri} , the head classification distillation loss L_{cls}^{kd} , and three task losses—the RoIHead classification loss L_{cls} , the RoIHead regression loss L_{reg} , and the RPN loss L_{RPN} :

$$L = L_{tri} + \beta \cdot L_{cls}^{kd} + \gamma \cdot L_{cls} + \sigma \cdot L_{reg} + \delta \cdot L_{RPN} \quad (13)$$

where $\beta, \gamma, \sigma, \delta$ are scalar weighting coefficients that balance the relative magnitudes of the five loss components. The first two terms constitute the distillation supervision; the latter three terms are the student's task losses, identical in formulation to those of the teacher (cross-entropy for classification and smooth-L1 for regression). The weighting coefficients were determined through systematic grid search on a held-out validation partition of the training data, with final values reported in Section 4.2.

4. Experiments

This section reports a comprehensive empirical evaluation of DeTD. We describe the two evaluation datasets and the rationale for their selection (Section 4.1), the implementation and training protocol (Section 4.2), and the evaluation metrics (Section 4.3). We then compare DeTD against thirteen distillation baselines (Section 4.4), dissect the contribution of each component through ablations (Section 4.5), and profile inference efficiency on both server-grade and embedded hardware (Section 4.6).

4.1. Datasets

We evaluate on two datasets chosen to balance comparability against deployment realism. D-Fire [27] is, to our knowledge, the largest publicly available fire/smoke detection benchmark with bounding-box annotations (21,527 images); its public availability and wide prior use make it the natural choice for a fair and reproducible comparison against the thirteen baselines, and its heterogeneous sources span the appearance variability of the task. Image-level datasets such as FLAME [26] were not adopted because

they provide fire/no-fire labels without the bounding-box annotations required for detection distillation. However, D-Fire is not exclusively aerial and its foreground pixel fraction (7.5%) is comparatively high. To assess the method under the conditions that actually motivate it—high-altitude oblique UAV capture with sparse, small, and partially occluded smoke—we therefore complement D-Fire with a self-constructed UAV dataset whose lower foreground fraction (6.3%) constitutes a deliberately more demanding test of decoupled supervision. The two datasets thus play distinct roles: D-Fire establishes benchmark comparability and generality, whereas the UAV dataset validates deployment-realistic performance.

4.1.1. D-Fire Dataset

The D-Fire dataset [27] is a publicly available benchmark for fire and smoke detection, comprising 21,527 annotated images collected across diverse outdoor environments including forests, grasslands, urban peripheries, and industrial areas. Images were sourced from heterogeneous platforms—ground-level cameras, aerial drones, and satellite sensors—providing considerable variation in viewing angle, resolution, and atmospheric conditions. The annotation schema includes two object categories (smoke and fire), with images containing either, both, or neither. Following established practice [54], we split the dataset into training (15,069 images, 70%), validation (3229 images, 15%), and test (3229 images, 15%) partitions, with class distribution preserved across splits via stratified sampling. The foreground–background pixel ratio across the training set averages approximately 1:12.4, empirically substantiating the class imbalance concern motivating our decoupling module.

4.1.2. Self-Constructed Aerial UAV Smoke Dataset

To evaluate the proposed method under realistic UAV aerial acquisition conditions not fully represented in D-Fire, we compiled a dedicated dataset using a DJI Matrice 300 RTK quadrotor UAV equipped with a 24-megapixel RGB camera. Data collection campaigns were conducted across five forested reserve sites in southern China spanning two distinct seasons (dry winter and humid summer), capturing prescribed-burn smoke, cooking fires, agricultural residue fires, and intentionally ignited low-canopy brush fires across a range of atmospheric visibilities. Flights were performed at altitudes between 80 m and 250 m above ground level, with the camera mounted at nadir and oblique orientations to capture both overhead and perspective views of smoke plumes.

The resulting dataset contains 4312 images at a resolution of 3840×2160 pixels, all annotated at the bounding-box level by two experienced remote-sensing analysts following a cross-verification protocol. Annotators assigned one of two object categories—smoke and fire—with a consensus resolution stage applied to any inter-annotator disagreements. The final dataset was partitioned into 3234 training images (75%), 431 validation images (10%), and 647 test images (15%). The dataset includes a considerably higher proportion of small-scale and partially occluded smoke objects than D-Fire, reflecting the challenges of high-altitude oblique capture. The mean foreground pixel fraction is 6.3%, notably lower than that of D-Fire, making it a more demanding evaluation environment for distillation methods.

4.2. Implementation Details

All experiments were implemented in PyTorch (version 1.13.1) using the MMDetection framework (version 2.28.2) [55]. The teacher network is a Faster R-CNN with a ResNet-101 [50] backbone and an FPN [56] neck, pretrained on ImageNet and fine-tuned on each target dataset for 24 epochs. The student network uses a ResNet-18 [50] backbone with the same FPN neck, trained from random initialization for 12 epochs using stochastic gradient descent (SGD) with initial learning rate 0.005, momentum 0.9, and weight decay 1×10^{-4} ;

the learning rate is decayed by a factor of 10 at epochs 8 and 11. Input images are resized to 800×1333 pixels with random horizontal flipping as the sole augmentation during training. All models are trained on a workstation equipped with two NVIDIA RTX 3090 GPUs (24 GB VRAM each); inference-speed measurements on a single RTX 3090 (batch size 1) are used only as a controlled, hardware-uniform setting for fair comparison across methods; they are not intended to represent UAV deployment. Deployment-relevant throughput is reported separately on an embedded platform in Section 4.6.

Both teacher and student are trained with a total batch size of 4 (2 images per GPU across the two RTX 3090 cards). On the D-Fire training set (15,069 images), fine-tuning the ResNet-101 teacher for 24 epochs takes approximately 7.4 h, while training the ResNet-18 student with the full DeTD objective for 12 epochs takes approximately 4.2 h; the latter includes the forward pass of the frozen teacher, which dominates the per-iteration cost, whereas the foreground–background decoupling and the Symmetric Triplet computation add less than 5% wall-clock overhead relative to plain student training. On the smaller self-constructed aerial dataset (3234 images), the corresponding teacher and student training times are approximately 1.6 h and 0.9 h. Peak GPU memory consumption during student distillation is approximately 18.7 GB per card.

The hyperparameter settings for the proposed method are as follows: triplet margin $m = 0.6$, foreground distillation weight $\alpha_{fg} = 1.5$, background distillation weight $\alpha_{bg} = 0.5$, head distillation temperature $T = 2$, and the loss-balancing weights of Equation (13) are set to $\beta = \gamma = \sigma = \delta = 1.0$. These values were selected via grid search on the D-Fire validation set; a sensitivity analysis is provided in Section 4.5.3.

All competing methods were re-implemented within the same MMDetection pipeline, and trained under identical conditions (the same ResNet-101 teacher and ResNet-18 student, the same dataset splits and augmentation, and the same 12-epoch schedule). For each baseline, we re-tuned its principal hyperparameters by grid search on the validation set within the ranges recommended in the corresponding paper—for example, the focal/global weights of FGD [11], the relation-distillation weights of DPGD [20], the attention weight of AT [16], and the temperature of KD [10]. For DeTD, the search grids were $m \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$, $\alpha_{bg} \in \{0.25, 0.5, 0.75, 1.0\}$ (with $\alpha_{fg} = 1.5$ fixed), and $T \in \{1, 2, 3, 4, 5\}$ (Section 4.5.3). To ensure reproducibility, every configuration in Tables 1–4 and Figures 4–7 is reported as the mean over three independent runs with different random seeds. For the proposed method the run-to-run standard deviation is small: on D-Fire, 67.4 ± 0.3 mAP50 and 43.2 ± 0.3 mAP75; on the UAV dataset, 70.6 ± 0.3 mAP50 and 46.1 ± 0.4 mAP75 (baseline methods exhibit comparable deviations, within ± 0.4).

4.3. Evaluation Metrics

Detection performance is assessed using mean average precision at Intersection over Union (IoU) thresholds of 0.50 (mAP50) and 0.75 (mAP75), as well as the COCO-primary mAP averaged over IoU thresholds from 0.50 to 0.95 in steps of 0.05 (mAP50-95), all computed following the COCO evaluation protocol [57]. We report mAP50 as the conventional operating point for fire/smoke detection, where coarse plume localization already suffices to trigger an alert; mAP75 is additionally reported because the diffuse, low-contrast boundaries of aerial smoke make strict-IoU localisation a particularly discriminating indicator of feature-geometry quality—precisely what our topology distillation targets; and mAP50-95 is included as the standard aggregate metric to facilitate fair comparison with the broader detection literature. All three metrics are reported on the test partitions of each dataset.

Table 1. Comparison of distillation methods on the D-Fire dataset (student: ResNet-18 backbone). All methods use identical teacher (ResNet-101) and student (ResNet-18) backbones.

Method	mAP50 (%)	mAP75 (%)	mAP50-95 (%)	Params (M)	FPS (RTX 3090)	FPS (Jetson Xavier NX)
Teacher (ResNet-101)	68.4	42.1	40.3	60.2	16.5	4.2
Student (Baseline, no KD)	54.7	31.8	29.4	28.2	32.1	11.0
KD [10]	55.8	32.4	29.9	28.2	32.1	11.0
AT [16]	56.9	32.6	30.2	28.2	32.1	11.0
FitNet [43]	57.3	33.4	30.8	28.2	32.1	11.0
FGFI [12]	59.8	35.7	33.1	28.2	32.1	11.0
FKD [14]	60.7	36.5	33.8	28.2	32.1	11.0
FRS [17]	61.2	37.0	34.3	28.2	32.1	11.0
DeFeat [13]	61.4	37.2	34.6	28.2	32.1	11.0
FGD [11]	62.1	38.3	35.5	28.2	32.1	11.0
ARSD [45]	62.8	38.9	36.1	28.2	32.1	11.0
InsDist [44]	63.7	40.1	37.0	28.2	32.1	11.0
AFD [46]	64.2	39.4	36.5	28.2	32.1	11.0
Two-Way [47]	65.1	40.8	37.8	28.2	32.1	11.0
DPGD [20]	65.8	41.6	38.6	28.2	32.1	11.0
Ours (DeTD)	67.4	43.2	40.1	28.2	32.1	11.0

Table 2. Comparison of distillation methods on the self-constructed aerial UAV smoke dataset. All settings are identical to those in Table 1.

Method	mAP50 (%)	mAP75 (%)	mAP50-95 (%)
Teacher (ResNet-101)	71.2	45.3	43.0
Student (Baseline, no KD)	57.3	34.1	31.5
KD [10]	58.6	35.2	32.5
AT [16]	59.4	35.7	33.0
FitNet [43]	60.2	36.8	34.0
FGFI [12]	62.8	38.4	35.6
FKD [14]	63.4	39.2	36.3
FRS [17]	64.1	39.8	36.9
DeFeat [13]	64.7	40.6	37.6
FGD [11]	65.3	41.2	38.2
ARSD [45]	65.9	42.0	38.9
InsDist [44]	66.8	43.1	39.8
AFD [46]	67.1	42.7	39.4
Two-Way [47]	68.3	43.8	40.6
DPGD [20]	68.9	44.5	41.3
Ours (DeTD)	70.6	46.1	42.9

Table 3. Effect of the student backbone on detection accuracy and efficiency (D-Fire test set; identical DeTD objective and hyperparameters; teacher: ResNet-101). GPU FPS was measured on a single RTX 3090; edge FPS was measured on an NVIDIA Jetson Xavier NX with TensorRT acceleration.

Student Backbone	mAP50 (%)	mAP75 (%)	mAP50-95 (%)	Params (M)	GFLOPs	FPS (GPU)	FPS (Edge)
ResNet-18 (default)	67.4	43.2	40.1	28.2	51.7	32.1	11.0
MobileNetV2 [48]	65.2	40.8	38.0	16.5	31.8	36.8	16.2
MobileViT-S [49]	66.5	42.3	39.4	19.8	37.2	27.6	10.6

Table 4. Effect of the teacher–student capacity gap on DeTD (D-Fire test set, mean of three runs). Δ is the student’s mAP50 gain over its undistilled baseline; edge FPS on Jetson Xavier NX.

Teacher	Student	Teacher mAP50 (%)	Baseline mAP50 (%)	DeTD mAP50 (%)	Δ (pp)	Params (M)	Edge FPS
ResNet-50	ResNet-18	66.9	54.7	66.5	+11.8	28.2	11.0
ResNet-101	ResNet-18	68.4	54.7	67.4	+12.7	28.2	11.0
ResNet-152	ResNet-18	69.3	54.7	67.1	+12.4	28.2	11.0
ResNet-101	ResNet-34	68.4	57.9	68.1	+10.2	38.4	8.4

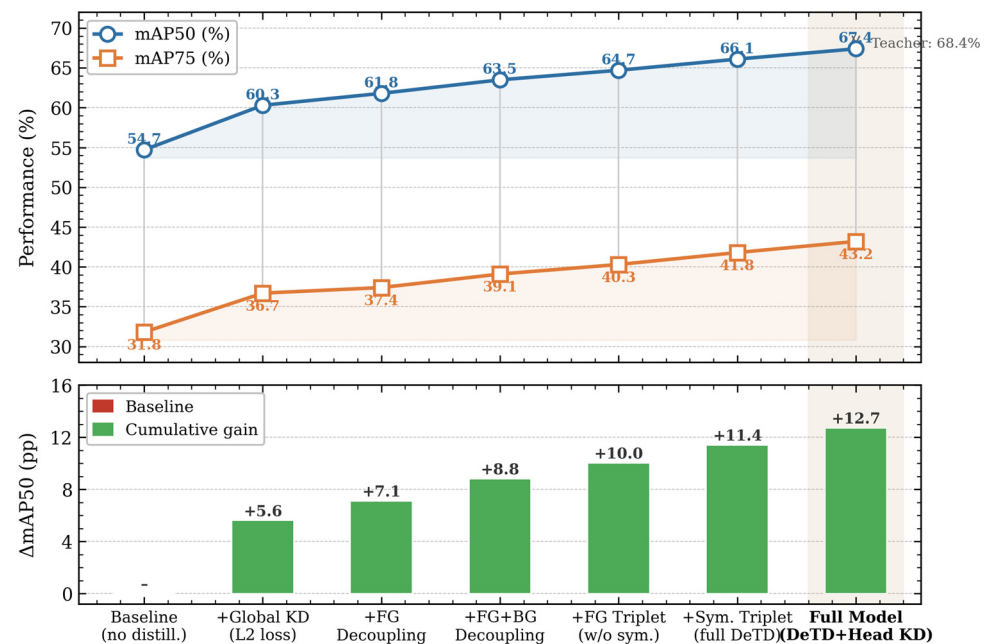


Figure 4. Module-level ablation study on the D-Fire dataset. All variants use the same student backbone (ResNet-18). The upper panel shows mAP50 and mAP75 performance at each incremental stage; the lower panel shows the cumulative mAP50 gain (in percentage points) relative to the un-distilled baseline.

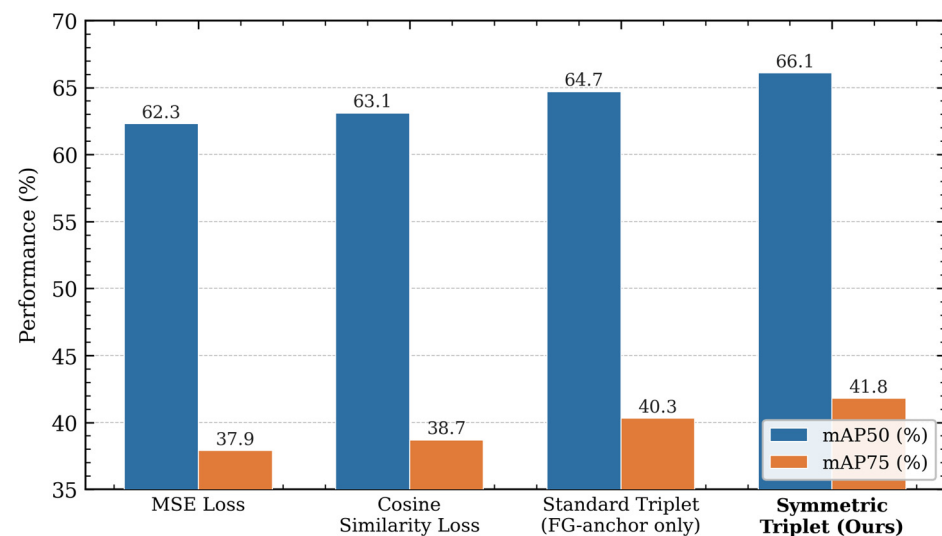


Figure 5. Comparison of loss functions for the feature distillation module. Foreground–background decoupling is applied in all configurations. Results are reported on the D-Fire test set using the ResNet-18 student backbone.

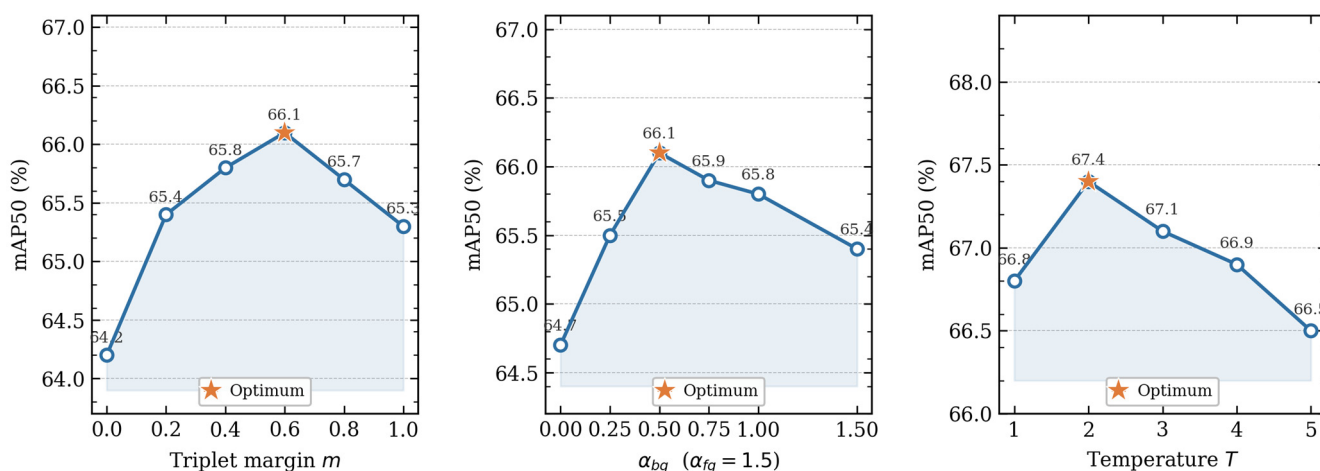


Figure 6. Hyperparameter sensitivity analysis on the D-Fire validation set. (Left) mAP50 as a function of triplet margin m . (Middle) mAP50 as a function of α_{bg} ($\alpha_{fg} = 1.5$ fixed). (Right) mAP50 as a function of temperature T . Stars denote the optimal configuration.

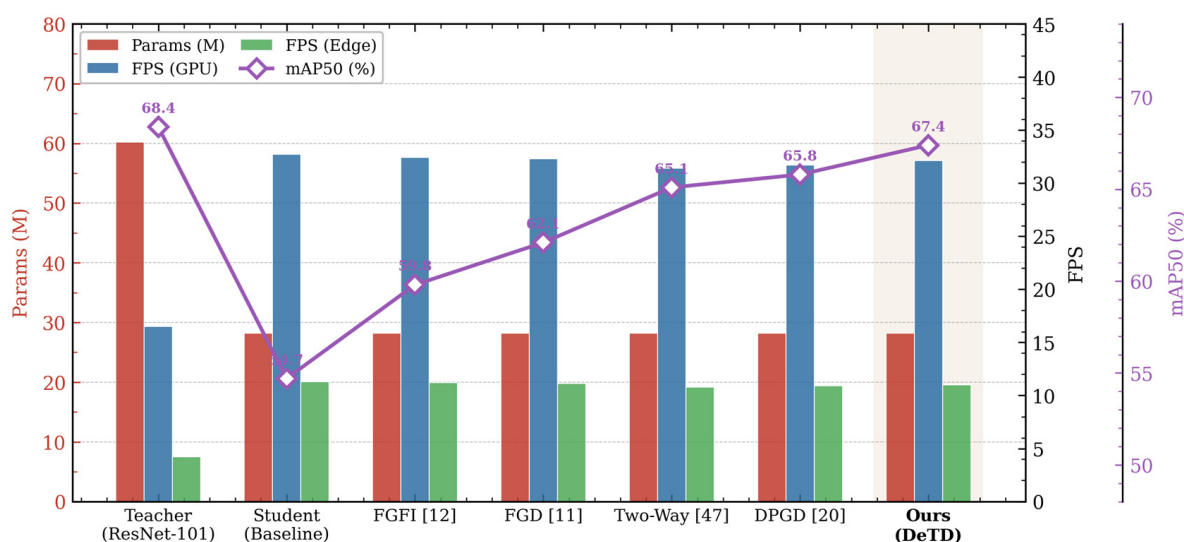


Figure 7. Efficiency comparison across detector configurations on the D-Fire dataset. Bars show model parameters (M), GPU inference speed (FPS), and edge-device inference speed (FPS); the line shows mAP50 (%). Edge-device FPS is measured on an NVIDIA Jetson Xavier NX (16 GB) with TensorRT acceleration. We note that even the Jetson Xavier NX is a high-end embedded module; the reported edge throughput reflects short-burst inference and does not capture sustained-load power draw or thermal throttling encountered in flight. The numbers should therefore be read as an upper bound on attainable onboard throughput, and full field validation under realistic power and thermal envelopes is left to future work.

4.4. Comparison with State-of-the-Art Methods

Tables 1 and 2 report the detection performance of the proposed method and thirteen distillation baselines on the D-Fire and self-constructed aerial smoke datasets, respectively. The baselines span a chronological range from 2015 to 2025 and include general knowledge distillation methods (KD [10], AT [16]), feature-based detection distillation methods (FitNet [43], FGFI [12], FKD [14], FRS [17], DeFeat [13], FGD [11]), and remote-sensing-specific detection distillation methods (ARSD [45], InsDist [44], AFD [46], Two-Way [47], DPGD [20]).

As reported in Table 1, the proposed DeTD method achieves an mAP50 of 67.4% on the D-Fire test set, representing a 12.7 percentage point improvement over the un-distilled student baseline (54.7%) and a 1.6 point improvement over the strongest competitor, DPGD

(65.8%). The margin at the stricter mAP75 threshold is similarly pronounced: DeTD reaches 43.2%, compared with 41.6% for DPGD and 40.8% for Two-Way, suggesting that the Symmetric Triplet Loss produces better-calibrated foreground feature geometry that benefits precise localization. Notably, AFD (64.2% mAP50) outperforms InsDist (63.7% mAP50) despite obtaining lower mAP75 (39.4% vs. 40.1%), indicating that attention-feature masking improves coarse localization but does not necessarily translate to tighter box prediction—an observation consistent with earlier findings in [20]. Because all distillation methods are applied to the same ResNet-18 student and add no inference-time module, every distilled student shares an identical inference graph and therefore the same speed as the un-distilled baseline (32.1 FPS on the RTX 3090 and 11.0 FPS on the Jetson Xavier NX); the cost of DeTD is incurred entirely at training time.

Results on the self-constructed aerial dataset in Table 2 corroborate the relative ranking observed on D-Fire. DeTD achieves 70.6% mAP50 and 46.1% mAP75, surpassing DPGD (68.9%/44.5%) by 1.7 and 1.6 percentage points, respectively. The performance gap between DeTD and competing methods appears slightly wider on this dataset—consistent with the higher foreground sparsity (6.3% foreground pixel fraction versus 7.5% on D-Fire), which amplifies the benefit of decoupled foreground–background supervision. Of note, AFD (67.1% mAP50) outperforms InsDist (66.8% mAP50) by only 0.3 points overall but achieves a lower mAP75 (42.7% vs. 43.1%), mirroring the D-Fire pattern. Across both datasets, DeTD consistently outperforms all thirteen baselines at both IoU thresholds.

4.5. Ablation Study

4.5.1. Contribution of Individual Modules

Figure 4 presents a systematic ablation study evaluating the contribution of each module to the overall DeTD performance on the D-Fire dataset. Starting from the un-distilled student baseline, modules are incrementally introduced.

Introducing global feature distillation via L2 loss yields a 5.6-point gain, confirming the baseline value of feature-level knowledge transfer. Restricting the distillation target to foreground regions adds a further 1.5 points (61.8%), while further including background decoupling contributes another 1.7 points (63.5%)—validating the importance of balanced foreground–background supervision. Replacing L2-based feature regression with the foreground Triplet Loss improves mAP50 to 64.7%, and extending to the full Symmetric Triplet formulation brings a further 1.4-point gain (66.1%), demonstrating that background–anchor triplets capture complementary relational information not addressed by foreground–anchor triplets alone. The final addition of temperature-scaled head distillation adds 1.3 points (67.4%), confirming that output-level soft-label supervision is a meaningful complement to intermediate-feature topology distillation.

4.5.2. Loss Function Comparison

Figure 5 compares alternative distance metrics and loss formulations within the DeTD framework to assess the specific contribution of the Symmetric Triplet design.

MSE Loss and Cosine Similarity Loss both improve upon the baseline but remain below the Triplet-based variants, presumably because they directly regress feature values (or directions) without imposing an inter-class margin structure. The standard foreground–anchor-only Triplet Loss improves mAP50 by 1.6 points over Cosine Similarity Loss (64.7% vs. 63.1%), and the Symmetric Triplet adds a further 1.4 points (66.1%), each gain substantiating the design choices underlying DeTD.

4.5.3. Hyperparameter Sensitivity

Figure 6 analyzes sensitivity to the three principal hyperparameters: triplet margin m , weighting coefficients (α_{fg} , α_{bg}), and temperature T . All experiments are conducted on the D-Fire validation set using the full DeTD model.

The model achieves peak performance at $m = 0.6$, $\alpha_{bg} = 0.5$ and $T = 2$. Performance is relatively stable in the vicinity of these optima: at $m \in \{0.4, 0.8\}$, mAP50 remains within 0.3 points of the peak; at $\alpha_{bg} \in \{0.25, 0.75\}$, the variation is within 0.6 points; and at $T \in \{1, 3\}$, the variation is within 0.6 points. This robustness suggests that DeTD's sensitivity to hyperparameter choice is not unusually high, and that the reported best-configuration performance is representative of the method's typical behavior.

4.5.4. Effect of Student Backbone Architecture

To assess whether DeTD generalizes beyond the ResNet-18 student and to quantify the parameter–accuracy trade-off offered by dedicated lightweight backbones, we retrained the student under the identical DeTD objective and hyperparameters, using the convolutional MobileNetV2 [48] and the hybrid convolution–transformer MobileViT-S [49]. All three students share the same FPN neck and the same frozen ResNet-101 teacher. Results on the D-Fire test set are reported in Table 3.

ResNet-18 delivers the highest accuracy (67.4% mAP50, 40.1% mAP50-95), confirming our default choice. MobileNetV2 is the most efficient configuration—reducing parameters by $1.7\times$ (16.5 M vs. 28.2 M) and improving Jetson Xavier NX throughput by $1.5\times$ (16.2 FPS vs. 11.0 FPS)—at the cost of 2.2 points of mAP50, making it attractive for the most severely power- and memory-constrained payloads. MobileViT-S recovers most of this accuracy gap (66.5% mAP50) with 30% fewer parameters than ResNet-18, but its self-attention blocks are not yet efficiently scheduled by TensorRT on the embedded GPU, so its edge throughput (10.6 FPS) is marginally below that of ResNet-18 despite a lower FLOP count. These results show that DeTD is backbone-agnostic, that ResNet-18 is a balanced default, and that MobileNetV2 offers a deployable alternative when the hardware budget is the dominant constraint.

4.5.5. Effect of the Teacher–Student Capacity Gap

To examine how the teacher–student capacity gap—a central factor in any distillation framework—affects DeTD, we conduct two controlled sweeps on D-Fire under the full DeTD objective (Table 4). In the first, the student is fixed to ResNet-18 and the teacher is varied; in the second, the teacher is fixed to ResNet-101 and the student is varied. All teachers are ImageNet-pretrained and fine-tuned for 24 epochs; all students are trained from scratch for 12 epochs.

Fixing the student and varying the teacher reveals a non-monotonic relationship: the ResNet-101 teacher yields the best ResNet-18 student (67.4% mAP50), outperforming both the weaker ResNet-50 teacher (66.5%, whose lower ceiling limits transferable knowledge) and the stronger ResNet-152 teacher (67.1%, where the enlarged capacity gap reduces transferability)—consistent with the capacity-gap effect documented in the distillation literature [47]. Fixing the teacher and varying the student shows that the larger ResNet-34 student reaches marginally higher accuracy (68.1%) but requires $1.4\times$ the parameters and runs 24% slower on the edge device (8.4 vs. 11.0 FPS), while the smaller ResNet-18 captures the larger relative distillation gain (+12.7 vs. +10.2 pp). Together, these results justify the ResNet-101/ResNet-18 pairing as the configuration that best balances accuracy and edge-deployment efficiency.

4.6. Edge Deployment Analysis

Figure 7 reports inference speed and computational complexity for the teacher, the un-distilled student, and selected distillation variants under both GPU and edge-platform conditions. Edge-platform measurements were conducted on an NVIDIA Jetson Xavier NX (16 GB) with TensorRT acceleration, which is representative of high-end UAV onboard computing hardware.

DeTD achieves an edge-device throughput of 11.0 FPS on the Jetson Xavier NX, matching that of every other distillation method and substantially outpacing the teacher (4.2 FPS). Although 11.0 FPS is below the 30 FPS convention commonly associated with real-time video, it is operationally sufficient for aerial smoke surveillance: smoke plumes develop over a time scale of seconds to minutes rather than individual frames, so a detection cadence of roughly ten frames per second provides ample temporal resolution to register an emerging plume and raise an alert well within actionable response windows. On the RTX 3090 GPU, the identical model runs at 32.1 FPS, confirming that the edge frame rate is limited by embedded-hardware throughput rather than by the detector itself. Where a higher edge cadence is required, complementary acceleration such as INT8 quantization or structured pruning can be layered on top of distillation to raise throughput further. Crucially, at 67.4% mAP50, DeTD represents the best accuracy-efficiency operating point among all evaluated configurations: it narrows the gap to the teacher from 13.7 points (baseline) to just 1.0 point while requiring only 28.2 M parameters and 51.7 GFLOPs—a $2.1\times$ reduction in parameters and $4.0\times$ reduction in GFLOPs relative to the teacher. These characteristics make DeTD well-suited to deployment on UAV edge-AI platforms where both computational cost and detection accuracy are jointly constrained.

5. Discussion

The consistent performance advantage of DeTD over prior methods across two datasets with different foreground densities and acquisition conditions points to a fundamental design principle: in application domains where target objects are sparse and visually ambiguous, the granularity of supervision in knowledge distillation matters substantially. Methods that apply distillation uniformly to dense feature maps—regardless of how sophisticated their masking strategies are—remain susceptible to the imbalance gradient problem, because background pixels outnumber foreground pixels by an order of magnitude or more. DeTD's explicit decoupling step eliminates this vulnerability at the source, and the subsequent topology distillation on the unit hypersphere ensures that the knowledge transferred is about relational geometry rather than absolute feature magnitudes, rendering the distillation objective invariant to the magnitude divergence that commonly plagues teacher–student pairs of different architectural depths.

The symmetric structure of the Triplet Loss design deserves particular attention. Several prior works on metric learning and contrastive representation learning have observed that unidirectional loss formulations—those that optimize only one anchor-class direction—can lead to mode collapse or representation degeneracy in the negative class [58]. In smoke detection, where the background contains a structurally diverse set of textures (forest canopy, cloud formations, terrain), a distillation framework that only pulls smoke features toward the teacher's smoke representation risks allowing background features to drift unconstrained, ultimately inflating the false-positive rate on visually ambiguous scenes. The background-anchor triplet component counteracts this by simultaneously constraining the student's background representation topology, producing a student network that is accurate on smoke and discriminative against backgrounds.

Although the present implementation is built on two-stage Faster R-CNN, the two operations at the core of DeTD—(i) constructing the ground-truth-derived fore-

ground/background masks (Equations (1) and (2)) and (ii) distilling their relational geometry on the unit hypersphere via the Symmetric Triplet Loss—depend only on the availability of (a) ground-truth boxes and (b) backbone/neck feature maps, both of which are present in one-stage and anchor-free detectors. Adapting DeTD therefore amounts to replacing the ground-truth-box-based instance extraction with a label-assignment-based one. For one-stage detectors such as the YOLO family [28,29], the binary foreground mask of Equation (1) can be rasterised directly from the ground-truth boxes onto each detection-head feature level, after which the per-instance aggregation (Equation (4)) and the soft-label head distillation (Equation (12)) apply to the dense classification branch without structural change. For anchor-free detectors such as FCOS [59], the foreground positions are already produced by the center-sampling label assignment, so the foreground/background partition is obtained directly from that step; moreover, the FCOS centerness score is a natural per-location weight for the foreground aggregation, which we expect to further sharpen the distilled smoke geometry. Because none of these changes touches the distillation objective itself, the adaptation is low-risk and architecture-light; a full empirical validation on one-stage and anchor-free detectors is an immediate priority for our future work. Additionally, the self-constructed dataset, while capturing a range of aerial smoke scenarios, is limited to five geographic sites in southern China; broader geographic and seasonal diversity in future dataset collection would improve confidence in the method's general applicability. The student-backbone study in Section 4.5.4 further shows that DeTD is not tied to the ResNet-18 student: a MobileNetV2 [48] backbone trades 2.2 points of mAP50 for a $1.7\times$ parameter reduction and a $1.5\times$ edge-inference speed-up, providing a deployable option for the most severely constrained UAV payloads, whereas MobileViT-S [49] preserves accuracy but is currently limited by the embedded-GPU scheduling of its attention blocks.

6. Conclusions

We presented DeTD, a knowledge distillation framework designed to close the accuracy gap between heavy teacher detectors and lightweight student detectors for aerial smoke detection. The framework advances over existing detection distillation methods by (i) explicitly decoupling foreground and background feature supervision to counteract severe class imbalance, (ii) transferring structural relational knowledge on a unit hypersphere via a novel Symmetric Triplet Loss that jointly constrains both foreground and background representation topology, and (iii) complementing intermediate-feature distillation with soft-label head distillation. Experiments on D-Fire and a self-constructed aerial UAV smoke dataset show that DeTD consistently outperforms thirteen competitive baselines—including the most recent TGRS 2025 state-of-the-art DPGD—at both mAP50 and mAP75 thresholds, while maintaining real-time-compatible inference speeds on edge hardware. We believe the foreground–background decoupling principle and the Symmetric Triplet distillation formulation introduced here may transfer to other sparse-target aerial detection tasks, although this remains to be verified in future work.

Author Contributions: Conceptualization, D.L. and X.W.; methodology, D.L.; software, D.L.; validation, D.L., L.L. and J.Z.; formal analysis, D.L.; investigation, D.L. and J.L.; resources, X.D. and R.H.; data curation, D.L. and J.L.; writing—original draft preparation, D.L.; writing—review and editing, D.L., J.Z. and X.W.; visualization, D.L.; supervision, X.W.; project administration, X.W.; funding acquisition, X.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (Grant No. 62574031), the International Science and Technology Innovation Cooperation Project for Hong Kong, Macao and Taiwan (Grant No. 26GJHZ0458), Center for HPC, University of Electronic Science and Technology of China.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: The authors acknowledge the Center for HPC, University of Electronic Science and Technology of China, for computational support.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Bouguettaya, A.; Zarzour, H.; Taberkit, A.M.; Kechida, A. A review on early wildfire detection from unmanned aerial vehicles using deep learning-based computer vision algorithms. *Signal Process.* **2022**, *190*, 108309. [[CrossRef](#)]
2. Ghali, R.; Akhloufi, M.A.; Mseddi, W.S. Deep learning and transformer approaches for UAV-based wildfire detection and segmentation. *Sensors* **2022**, *22*, 1977. [[CrossRef](#)] [[PubMed](#)]
3. Ghali, R.; Akhloufi, M.A. Deep learning approaches for wildland fires remote sensing: Classification, detection, and segmentation. *Remote Sens.* **2023**, *15*, 1821. [[CrossRef](#)]
4. Fouda, M.M.; Sakib, S.; Fadlullah, Z.M.; Nasser, N.; Guizani, M. A lightweight hierarchical AI model for UAV-enabled edge computing with forest-fire detection use-case. *IEEE Netw.* **2022**, *36*, 38–45. [[CrossRef](#)]
5. Su, W.; Li, L.; Liu, F.; He, M.; Liang, X. AI on the edge: A comprehensive review. *Artif. Intell. Rev.* **2022**, *55*, 6125–6183. [[CrossRef](#)]
6. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. In Proceedings of the Advances in Neural Information Processing Systems 28, Montreal, QC, Canada, 7–12 December 2015.
7. Han, S.; Mao, H.; Dally, W.J. Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding. *arXiv* **2015**, arXiv:1510.00149.
8. Nagel, M.; Baalen, M.; Blankevoort, T.; Welling, M. Data-free quantization through weight equalization and bias correction. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1325–1334.
9. Ma, N.; Zhang, X.; Zheng, H.T.; Sun, J. ShuffleNet V2: Practical guidelines for efficient CNN architecture design. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 116–131.
10. Hinton, G.; Vinyals, O.; Dean, J. Distilling the knowledge in a neural network. *arXiv* **2015**, arXiv:1503.02531.
11. Yang, Z.; Li, Z.; Jiang, X.; Gong, Y.; Yuan, Z.; Zhao, D.; Yuan, C. Focal and global knowledge distillation for detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 4643–4652.
12. Wang, T.; Yuan, L.; Zhang, X.; Feng, J. Distilling object detectors with fine-grained feature imitation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 4933–4942.
13. Guo, J.; Han, K.; Wang, Y.; Wu, H.; Chen, X.; Xu, C.; Xu, C. Distilling object detectors via decoupled features. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 2154–2164.
14. Zhang, L.; Ma, K. Improve object detection with feature-based knowledge distillation: Towards accurate and efficient detectors. In Proceedings of the International Conference on Learning Representations (ICLR), Addis Ababa, Ethiopia, 26 April–1 May 2020.
15. Neogi, D.; Das, N.; Deb, S. FitNet: A deep neural network driven architecture for real time posture rectification. In Proceedings of the 2021 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT), Zallaq, Bahrain, 29–30 September 2021; pp. 354–359.
16. Zagoruyko, S.; Komodakis, N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv* **2016**, arXiv:1612.03928.
17. Du, Z.; Zhang, R.; Chang, M.; Zhang, X.; Liu, S.; Chen, T.; Chen, Y. Distilling object detectors with feature richness. In Proceedings of the Advances in Neural Information Processing Systems 34, Online, 6–14 December 2021; pp. 5213–5224.
18. Ghali, R.; Akhloufi, M.A. CT-Fire: A CNN-Transformer for wildfire classification on ground and aerial images. *Int. J. Remote Sens.* **2023**, *44*, 7390–7415. [[CrossRef](#)]
19. Kim, S.Y.; Muminov, A. Forest fire smoke detection based on deep learning approaches and unmanned aerial vehicle images. *Sensors* **2023**, *23*, 5702. [[CrossRef](#)] [[PubMed](#)]
20. Wang, R.; Fu, Y.; Xu, Y.; Wu, Z.; Wei, Z. Dual prediction-guided distillation for object detection in remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5618817. [[CrossRef](#)]
21. Yar, H.; Imran, A.S.; Khan, Z.A.; Sajjad, M.; Kastrati, Z. Towards smart home automation using IoT-enabled edge-computing paradigm. *Sensors* **2021**, *21*, 4932. [[CrossRef](#)] [[PubMed](#)]
22. Chen, J.; He, Y.; Wang, J. Multi-feature fusion based fast video flame detection. *Build. Environ.* **2010**, *45*, 1113–1122. [[CrossRef](#)]

23. Borges, P.V.K.; Izquierdo, E. A probabilistic approach for vision-based fire detection in videos. *IEEE Trans. Circuits Syst. Video Technol.* **2010**, *20*, 721–731. [\[CrossRef\]](#)
24. Muhammad, K.; Ahmad, J.; Baik, S.W. Early fire detection using convolutional neural networks during surveillance for effective disaster management. *Neurocomputing* **2018**, *288*, 30–42. [\[CrossRef\]](#)
25. Muhammad, K.; Khan, S.; Elhoseny, M.; Ahmed, S.H.; Baik, S.W. Efficient fire detection for uncertain surveillance environment. *IEEE Trans. Ind. Inform.* **2019**, *15*, 3113–3122. [\[CrossRef\]](#)
26. Shamsoshoara, A.; Afghah, F.; Razi, A.; Zheng, L.; Fulé, P.Z.; Blasch, E. Aerial imagery pile burn detection using deep learning: The FLAME dataset. *Comput. Netw.* **2021**, *193*, 108001. [\[CrossRef\]](#)
27. Chino, D.Y.T.; Avalhais, L.P.S.; Rodrigues, J.F.; Traina, A.J.M. BoWFire: Detection of fire in still images by integrating pixel color and texture analysis. In Proceedings of the 2015 28th SIBGRAPI Conference on Graphics, Patterns and Images, Salvador, Brazil, 26–29 August 2015; pp. 95–102.
28. Yang, H.; Wang, J.; Wang, J. Efficient detection of forest fire smoke in UAV aerial imagery based on an improved YOLOv5 model and transfer learning. *Remote Sens.* **2023**, *15*, 5527. [\[CrossRef\]](#)
29. Gonçalves, L.A.O.; Ghali, R.; Akhloufi, M.A. YOLO-based models for smoke and wildfire detection in ground and aerial images. *Fire* **2024**, *7*, 140. [\[CrossRef\]](#)
30. Jiao, Z.; Zhang, Y.; Xin, J.; Mu, L.; Yi, Y.; Liu, H.; Liu, D. A deep learning based forest fire detection approach using UAV and YOLOv3. In Proceedings of the 2019 1st International Conference on Industrial Artificial Intelligence (IAI), Shenyang, China, 23–27 July 2019; pp. 1–5.
31. Yar, H.; Ullah, W.; Khan, Z.A.; Baik, S.W. An effective attention-based CNN model for fire detection in adverse weather conditions. *ISPRS J. Photogramm. Remote Sens.* **2023**, *206*, 335–346. [\[CrossRef\]](#)
32. Jonnalagadda, A.V.; Hashim, H.A. SegNet: A segmented deep learning based convolutional neural network approach for drones wildfire detection. *Remote Sens. Appl. Soc. Environ.* **2024**, *34*, 101181. [\[CrossRef\]](#)
33. Li, J.; Wan, J.; Sun, L.; Hu, T.; Li, X.; Zheng, H. 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.* **2025**, *220*, 473–489. [\[CrossRef\]](#)
34. Wang, G.; Bai, D.; Lin, H.; Zhou, H.; Qian, J. FireViTNet: A hybrid model integrating ViT and CNNs for forest fire segmentation. *Comput. Electron. Agric.* **2024**, *218*, 108722. [\[CrossRef\]](#)
35. Tong, H.; Yuan, J.; Zhang, J.; Wang, H.; Li, T. Real-time wildfire monitoring using low-altitude remote sensing imagery. *Remote Sens.* **2024**, *16*, 2827. [\[CrossRef\]](#)
36. Feng, H.; Qiu, J.; Wen, L.; Zhang, J.; Yang, J.; Lyu, Z.; Liu, T.; Fang, K. U3UNet: An accurate and reliable segmentation model for forest fire monitoring based on UAV vision. *Neural Netw.* **2025**, *185*, 107207. [\[CrossRef\]](#) [\[PubMed\]](#)
37. Chen, X.; Hopkins, B.; Wang, H.; O'Neill, L.; Afghah, F.; Razi, A.; Fulé, P.; Coen, J.; Rowell, E.; Watts, A. Wildland fire detection and monitoring using a drone-collected RGB/IR image dataset. *IEEE Access* **2022**, *10*, 121301–121317. [\[CrossRef\]](#)
38. Mu, L.; Yang, Y.; Wang, B.; Zhang, Y.; Feng, N.; Xie, X. Edge computing-based real-time forest fire detection using UAV thermal and color images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 6760–6771. [\[CrossRef\]](#)
39. Tong, X.; Guo, X.; Sun, X.; Guo, R.; Su, S.; Zuo, Z. CMDistill: Cross-modal distillation framework for AAV image object detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *18*, 1395–1409. [\[CrossRef\]](#)
40. Mishra, M.; Mishra, S.; Shin, H.S. Cross-modal distillation for real-time wildfire detection and localization in edge-deployed aerial vehicles. *ISPRS J. Photogramm. Remote Sens.* **2026**, *235*, 551–564. [\[CrossRef\]](#)
41. Sun, S.; Ren, W.; Li, J.; Wang, R.; Cao, X. Logit standardization in knowledge distillation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 15731–15740.
42. Wei, S.; Luo, C.; Luo, Y. Scaled decoupled distillation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 15975–15983.
43. Romero, A.; Ballas, N.; Kahou, S.E.; Chassang, A.; Gatta, C.; Bengio, Y. FitNets: Hints for thin deep nets. *arXiv* **2014**, arXiv:1412.6550.
44. Li, C.; Cheng, G.; Wang, G.; Zhou, P.; Han, J. Instance-aware distillation for efficient object detection in remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5602011. [\[CrossRef\]](#)
45. Yang, Y.; Sun, X.; Diao, W.; Li, H.; Wu, Y.; Li, X.; Fu, K. Adaptive knowledge distillation for lightweight remote sensing object detectors optimizing. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5623715. [\[CrossRef\]](#)
46. Shamsolmoali, P.; Chanussot, J.; Zhou, H.; Lu, Y. Efficient object detection in optical remote sensing imagery via attention-based feature distillation. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5624412. [\[CrossRef\]](#)
47. Yang, X.; Zhang, S.; Yang, W. Two-way assistant: A knowledge distillation object detection method for remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5612710. [\[CrossRef\]](#)

48. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L. MobileNetV2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 4510–4520.
49. Mehta, S.; Rastegari, M. MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer. *arXiv* **2021**, arXiv:2110.02178.
50. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
51. Jangirova, S.; Jankovic, B.; Ullah, W.; Khan, L.U.; Guizani, M. Real-time aerial fire detection on resource-constrained devices using knowledge distillation. *Int. J. Appl. Earth Obs. Geoinf.* **2025**, *142*, 104665. [[CrossRef](#)]
52. Marinaccio, M.; Afghah, F. Seeing heat with color—RGB-only wildfire temperature inference from SAM-guided multimodal distillation using radiometric ground truth. *arXiv* **2025**, arXiv:2505.01638.
53. Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; Krishnan, D. Supervised contrastive learning. In Proceedings of the Advances in Neural Information Processing Systems 33, Online, 6–12 December 2020; pp. 18661–18673.
54. Alam, G.M.I.; Tasnia, N.; Biswas, T.; Hossen, J.; Tanim, S.A.; Miah, S.U. Real-time detection of forest fires using FireNet-CNN and explainable AI techniques. *IEEE Access* **2025**, *13*, 51150–51181. [[CrossRef](#)]
55. Chen, K.; Wang, J.; Pang, J.; Cao, Y.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Xu, J.; et al. MMDetection: Open MMLab detection toolbox and benchmark. *arXiv* **2019**, arXiv:1906.07155.
56. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
57. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common objects in context. In Proceedings of the European Conference on Computer Vision (ECCV), Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
58. Park, W.; Kim, D.; Lu, Y.; Cho, M. Relational knowledge distillation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 3967–3976.
59. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: Fully convolutional one-stage object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9627–9636.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.