Edge Vision for Wildfires: A Comprehensive Investigation

Haolin Du *

School of Mathematics and Physics, Beijing Institute of Technology, Zhuhai, China

* Corresponding Author Email: d2316182165@gmail.com

Abstract. Wildfire monitoring needs fast, reliable, and scalable systems that can work in different terrains and lighting conditions. Traditional methods such as satellites, towers, and patrols cover large areas but often have delays, errors, and problems with smoke or obstacles. This review introduces recent computer vision progress for early wildfire detection and presents it as a practical pipeline focused on real-time use, stability, and easy operation. The first part reviews image-based screening and localization. Lightweight CNN models can quickly filter fire-related scenes and are improved through continual learning. The Learning-Without-Forgetting (LwF) method keeps new models consistent with old ones without saving past data, preventing forgetting. This helps maintain accuracy for small fires and allows frequent updates that follow seasonal or sensor changes. For localization, anchor-free detectors such as YOLO are efficient, combining good accuracy and speed. Training uses data augmentation and hard-negative samples like fog or clouds, and quantized models run on UAVs or tower devices. Temporal models using short video windows help reduce false alarms by learning fire motion patterns. The second part focuses on pixel-level mapping and system design. Multimodal fusion of RGB and infrared data improves segmentation under smoke or low light. Efficiency is achieved through model pruning, quantization, and adaptive inference. Evaluation includes not only accuracy but also detection speed, alert stability, and energy use, connecting perception with fire prediction and emergency planning.

Keywords: Early wildfire detection, computer vision, UAV/drone monitoring, one-stage object detection (YOLO).

1. Introduction

Wildfires are among the most damaging natural hazards, driving profound ecological degradation, biodiversity loss, and socio-economic disruption. Annual burn areas measured in millions of hectares release vast quantities of greenhouse gases, degrade air quality across regions, and endanger communities at the wildland-urban interface, Traditional monitoring methods - satellite constellations, lookout towers, and human patrols - have provided foundational capability for decades, yet they struggle with latency, localization precision, terrain occlusions, and the practical constraints of staffing and cost in remote landscapes. These constraints have driven a steady move to automated, camera-based analytics that run in real time, scale to vast areas, and lessen human workload while boosting reliability [1].

In recent years, deep-learning-based computer vision has progressed from prototypes to pipelines credible for real-world deployment. State-of-the-art surveys highlight both strengths and limits: convolutional backbones extract highly discriminative features; modern detectors provide rapid localization; and multimodal inputs (e.g., thermal/infrared) enhance robustness under challenging illumination - yet advances remain constrained by limited and imbalanced datasets, domain shift across regions and sensors, and edge-side power/compute budgets [2]. Within this context, transfer learning is especially effective on small to moderate wildfire datasets; for example, a VGG-19 model initialized from large-scale image pretraining achieves near-human-level accuracy on fire/non-fire classification, confirming that generic visual features transfer well to this task [3].

Accurate localization and low detection latency are both essential for early warning. In response, one-stage detectors optimized for aerial and fixed-camera imagery have achieved competitive mAP while maintaining real-time throughput suitable for on-board inference. An improved variant of a modern detector, augmented with multi-scale feature aggregation and an efficient pooling block, reached real-time throughput on UAV video while increasing precision relative to baseline

configurations - an encouraging step toward persistent patrol with autonomous platforms [4]. At the same time, image-understanding modules must be efficient, and streamlined convolutional architectures with carefully designed separable operations and residual pathways have demonstrated that high recall and precision can be achieved without prohibitive compute overhead, a key consideration for embedded deployments in the field [5].

Perhaps the most consistent empirical results are that multimodal fusion helps under smoke, haze, and low-light conditions. A drone-collected dataset combining RGB and infrared channels models by a notable margin in macro-F1, particularly in scenes with thin smoke or partial occlusion. This suggests that fusing complementary modalities is not merely beneficial but often necessary when operating across day-night cycles and varied atmospheric conditions [6]. Collectively, these advances point to a pragmatic paradigm: pair mobile sensing (UAVs, tower networks) with vision models that are accurate, efficient, and robust to environmental variability: emphasize localization and time-to-detect; and leverage multimodal views wherever feasible. While open challenges remain in cross-region generalization, annotation efficiency, and edge constraints, the evidence indicates that vision-based systems can meaningfully augment and, in some contexts, surpass legacy surveillance approaches in both speed and reliability.

2. Model

Wildfire perception benefits from a pipeline perspective: a lightweight classifier screens frames, a detector localizes smoke/flame in real time, temporal modules exploit motion cues, segmentation and fusion produce pixel-level products, and the system adapts continuously while meeting edge constraints. The subsections that follow summarize core components and design choices, each anchored once to a representative, recent source.

2.1. CNN-Based Classification with Continual Learning

A front-end classifier provides rapid screening before heavier inference. Architectures such as Xception strike a favorable balance between accuracy and latency on embedded devices. However, long-running deployments face appearance drift driven by seasonality, sensor changes, and terrain variability. Learning Without Forgetting (LwF) offers a practical mechanism for incremental updates without storing or replaying historical data: the updated model is guided to preserve the earlier model's responses while learning from fresh samples, reducing catastrophic forgetting and preserving legacy competence. In wildfire monitoring, an Xception-style backbone updated via LwF can maintain recognition of previously seen smoke/flame signatures while adapting to new camera viewpoints or aerosol regimes [7].

2.2. Real-Time Object Detection on Edge Devices

After screening, the pipeline must localize smoke and flame with minimal delay. One-stage designs exemplified by YOLOv8 are well suited to UAVs and tower nodes because they employ an anchor-free head, decoupled classification/regression branches, and streamlined feature aggregation that together simplify training, improve small-object sensitivity, and sustain throughput when compiled to accelerated runtimes. In practice, a robust wildfire detector pairs multi-scale augmentation with hard-negative mining (cloud edges, fog banks, industrial plumes) and is exported as an INT8/FP16 engine for on-board execution under tight power and thermal budgets [8].

2.3. Temporal and Video Modeling for Motion-Aware Reasoning

Frame-wise detectors are vulnerable to transient artifacts. Temporal medeling treats time as a first-class signal: slow-fast architectures learn complementary scales by combining a low-rate semantic pathway with a high-rate motion pathway. An efficient slow-fast variant demonstrated how to couple these streams without prohibitive compute, a pattern directly transferable to wildfire video, where the slow branch stabilizes background and horizon while the fast branch tracks turbulent eddies and rising

plumes. Appending a short-window temporal head behind the detector converts per-frame scores into video-level decisions, reducing flicker and cutting transient false positives [9].

2.4. Video-Language Interfaces for Operator-Centric Analytics

Operations benefit from interpretable summaries and interactive queries (e.g., "Is the plume height increasing?"). Video-language models integrate visual, audio, and text streams via a spatio-temporal connector, enabling captioning, question answering, and semantically grounded alerts for analysts supervising multiple feeds. Deployed server-side, this layer enriches situational awareness without burdening on-board inference and creates an auditable trail of explanations to support incident review and training [10].

2.5. Semi-Supervised Temporal Encoders for Label Efficiency

Dense frame-level annotations are costly. Semi-supervised temporal encoders pretrain to enforce frame-to-frame consistency, learning motion-aware invariances that are resilient to illumination shifts, compression artifacts, and camera jitter. Validated on dynamic vision tasks, this strategy transfers naturally to smoke evolution, where subtle inter-frame changes carry decisive evidence of combustion. Leveraging abundant unlabeled tower/UAV video, such encoders reduce reliance on exhaustive labeling while improving stability in borderline cases [11].

2.6. Segmentation and Multimodal Fusion for Pixel-Level Products

Incident response often requires pixel-level maps for perimeter estimation, rate-of-spread modeling, and GIS overlays. While U-shaped CNNs remain popular, the multimodal foundation paradigm unifies RGB, infrared, and contextual text/metadata in a transformer backbone that can attend across modalities to produce calibrated masks. Two practical fusion regimes dominate: late fusion (independent encoders with feature concatenation) for simplicity and robustness, and cross-attention fusion when the model must learn to rely on thermal cues (low light, partial occlusion) versus color/texture cues (clear daylight). Conditional triggering-running segmentation only after a confident detection - saves energy, and light temporal smoothing suppresses mask flicker for stable overlays [12].

2.7. Sustainable Updating and Edg-Oriented Optimization

Long-running systems should improve continuously without downtime. Incremental updates with LwF keep "muscle memory" intact while absorbing new footage; semi-supervised temporal signals provide pseudo-labels to expand training sets with minimal human effort; and structured pruning, distillation, and quantization preserve small-object sensitivity while meeting edge latency and power limits. Dynamic inference policies skip heavy branches when confidence is high, while temporal voting enforces persistence before escalating alerts or streaming high-bitrate clips. When server-side multimodal components are available, retrieval-augmented prompts surface similar historical snippets for fast comparison and operator trust, without adding weight to the on-board loop.

3. Discussion

3.1. Challenges

The most reliable wildfire perception systems are staged pipelines, not monolithic models. Each component confers distinct strengths and weaknesses: the classifier efficiently filters bulk video but lacks geometry; the detector supplies bounding-box localization but can be jittery on small, transient targets; temporal heads stabilize decisions at the cost of modest latencyl segmentation and multimodal fusion yield rich pixel-level products but increase complexity and energy; and continual updating maintains accuracy across seasons and sensors if executed without catastrophic forgetting. A tiered policy-screen everywhere, localizes when needed, and escalates to temporal reasoning and

segmentation only for credible events - keeps false alarms manageable and preserves energy on UAVs and towers.

A pivotal inflection is treating time as a signal on par with space. Motion patterns - plume rise, drift direction, and persistence - separate genuine combustion from nuisances such as backlit clouds, stram vents, and distant dust. Appending a short temporal head to a fast detector converts frame decisions into video decisions, suppressing transient false positives while often improving time-to-detect under the same false-alarm budget. Choosing the window length is an engineering trade-off: a few seconds on fixed cameras typically balances stability and delay; on moving UAVs subject to platform motion, roughly one second often suffices. In parallel, fusion of RGB and infrared strengthens robustness across dusk, haze, and partial occlusion. Late fusion is simple and dependable; cross-attention fusion lets the model learn when to trust thermal cues versus color and texture. For operations, pixel masks should be lightly smoothed over time to avoid flicker in GIS overlays and displays, and escalation thresholds should incorporate simple persistence rules (e.g., consecutive-frame agreement) to convert noisy frame-level scores into stable video-level alerts.

Label efficiency and longevity are operational necessities. Semi-supervised temporal pretraining exploits abundant unlabeled video to learn motion-aware invariances, reducing dependence on densely labeled datasets and curbing single-frame overfitting (a common cause of false alarms). Incremental updates using continual-learning objectives keep models aligned with changing conditions while preserving competence on legacy scenarios; in practice, small, frequent micro-updates are less risky and more effective than rare, large-scale retrains. Robustness further depends on curating a local hard-negative bank reflecting failure modes in the region of interest (cloud edges, fog banks, industrial plumes) and on domain randomization that mimics haze, compression, color cast, lens flare, and motion blur encountered in the field.

Hardware limits on the edge largely determine what a wildfire system can do. Detectors should target accelerated back ends, accept post-training quantization, and still deliver at least 20 FPS within tight power and thermal envelopes. Use dynamic inference: when confidence is high, bypass expensive branches; when confidence drops, fall back to the full stack and, only then, upload high-bitrate clips. Decision thresholds should be elastic, switching from conservative “patrol” setting to a more sensitive “response” mode once an event looks credible. Qualification must include day-night transitions, very small long-range targets, platform vibration, and unstable links so that failure degrades gracefully. Equally, people matter: saliency cues, calibrated masks, and short rationales cut cognitive load and speed triage; if uncertainty remains, the system should request context (wind fields, planned burns, industrial activity) to support better decisions and traceability.

Evaluation has to reflect field use rather than lab ideals. Besides accuracy, report event-level time-to-detection under a fixed false-alarm budget; alert persistence (the share that lasts $\geq$K seconds); small-object sensitivity by distance/altitude bands; energy per frame on the target device; and drift robustness before and after updates. For fusion pipelines, run ablations (RGB-only, IR-only, fused) and audit decision latency for every conditional branch to separate true algorithmic gains from pipeline effects. With this framing, detection becomes the front door to prediction: stable masks, motion cues, and contextual layers allow estimation of spread direction and speed, identification of exposed assets, and rapid translation of perception into actions for incident commanders. Ultimately, success is measured by earlier, more reliable alerts, lower false-alarm rates that sustain operator trust, and tight integration with communications, GIS, and operations centers so that timely detections become effective responses.

3.2. Future prospects

Next-generation wildfire perception will hinge on three converging directions. First, representation learning that adapts with minimal labels and without forgetting. Generic image-recognition research points toward pipelines that combine large-scale pretraining with continual updates, lowering data demands while preserving prior competence during domain shifts (season, camera, terrain) [13]. For fire monitoring, this implies self/semi-supervised video pretraining, non-forgetting updates (e.g.,

LwF), and possibly federated optimization so edge nodes learn from distributed streams without centralizing raw footage [13].

Second, video-native and multimodal modeling at scale, Underwater ecology shows that robust ecological classification benefits from explicit spatio-temporal reasoning, fusion across sensors, and embedded/edge execution under difficult visibility and illumination [14]. The same principles transfer to smoke/ flame perception: short-window temporal modules (slow-fast style) to turn fleeting cues into stable event decisions; RGB-IR fusion to sustain performance at dusk, through haze, and under partial occlusion; and cross-platform data integration across towers, UAVs, and satellites to bolster coverage and resilience [14]. These choices align with mainstream vision trends - CNN/Transformer backbones that learn discriminative features at scale and outperform hand-crafted pipelines on complex imagery [13, 14].

Third, pipeline-level engineering for real time, reliability, and auditability. Industrial vision underscores the value of end-to-end, stepwise pipelines-from calibration and sensing to detection, tracking, and quality assessment-with metrics tied to process outcomes, not just accuracy [15]. Wildfire systems should mirror this discipline: edge-oriented optimization (quantization/pruning/distillation) to maintain $\geq$20 FPS within power/thermal envelopes; dynamic inference that skips heavy branches at high confidence; and field-oriented evaluation- event-level time-to-detection at fixed false-alarm budgets, alert persistence, small-object sensitivity by distance/altitude, energy per frame, and pre/post-update drift resilience. Hard-negative banks (cloud edges, fog, industrial plumes), domain randomization, and targeted synthetic data will harden models against failure modes, while operator-facing outputs (saliency overlays, calibrated masks, concise rationales) turn timely perception into accountable action. Together, these steps move the field from promising pilots to dependable, at-scale surveillance [13,15].

4. Conclusion

This review reframes early wildfire perception as a challenge in pipeline engineering rather than a quest for a single, all-powerful model. The most dependable systems stage computation in layers that match operations: a lightweight CNN filters large video flows; an anchor-free, decoupled one-stage detector localizes smoke and flame on edge hardware; a short-window temporal head converts noisy frame scores into video-level decisions by exploiting plume rise, drift, and persistence; and, when warranted, segmentation with RGB-IR fusion yields pixel-wise products that plug directly into GIS and command workflows. This division of labor reduces false alarms, shortens time-to-detection, and preserves scarce computing and power on UAVs and tower nodes.

Sustained performance across seasons, sensors, and terrain relies on learning strategies that adapt without forgetting. Learning-Without-Forgetting enables rolling micro-updates that retain legacy competence while absorbing new conditions, and semi-supervised temporal encoders mine frame-to-frame consistency in unlabeled streams to stabilize borderline cases. On the engineering side, edge-oriented optimization, quantization, pruning, distillation, dynamic inference that skips heavy branches at high confidence, and temporal voting that requires persistence before escalation keep throughput at or above real time under thermal and power ceilings. Targeted hard-negative mining (cloud edges, fog banks, industrial plumes) and domain randomization (haze, compression, color cast, motion blur) further harden the stack.

Success should be measured by the way systems are used in the field. Beyond accuracy, report event-level time-to-detection at a fixed false-alarm budget, alert stability over seconds, small-object sensitivity versus distance/altitude, energy per frame on target hardware, and drify resilience before and after updates. For fusion stacks, RGB-only/IR-only/fused ablations and decision-latency audits for conditional branches separate genuine algorithmic gains from pipeline artifacts. Human factors are equally important: saliency overlays, calibrated masks, and concise rationales reduce operator load; when uncertainty persists, systems should request context (wind, planned burns, industrial activity) to convert detections into informed decisions.

Taking together, the evidence outlines an operational roadmap that links sensing with actuation. When outputs are temporally consistent and spatially calibrated using persistent mask, motion-aware features, and environmental context the same pipeline that notices the first wisp of smoke can also infer likely growth vectors and rates, highlight exposed communities and infrastructure, and deliver concise, actionable briefs to incident command. The next gains are chiefly engineering: establish shared multimodal benchmarks, run cross-region and cross-season stress tests, and tightly couple vision services with GIS and dispatch workflows. With these pieces in place, wildfire surveillance can move from promising pilots to dependable, at-scale operations that issue faster, more trustworthy alerts where they matter most.

References

[1] Bouguettaya A, Zarzour H, Taberkit A M, Kechida A. A view on early wildfire detection from unmanned aerial vehicles using deep learning-based computer vision algorithms. Ecological Informatics, 2022, 69: 108309.

[2] Ramos L, Casas E, Bendek E, Romero C, Rivas-Echeverria F. Computer vision for wildfire detection: A critical brief review. Multimedia Tools and Applications, 2024, 83: 83427 - 83470.

[3] Khan S, Alotaibi A, Alqazzaz A, Baslem A. DeepFire: A novel dataset and deep transfer-learning benchmark for forest fire. Mobile Information Systems, 2022.

[4] Mukhiddinov M, Abdusalomov AB, Cho J. A wildfire smoke detection system using unmanned aerial vehicle images based on the optimized YOLOv5. Sensors. 2022 Dec 1; 22 (23): 9384.

[5] Seydi ST, Saeidi V, Kalantar B, Ueda N, Halin AA. Fire-Net: A Deep Learning Framework for Active Forest Fire Detection. Journal of Sensors. 2022; 2022 (1): 8044390.

[6] 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 Nov 17; 10: 121301 - 17.

[7] Sathishkumar VE, Cho J, Subramanian M, Naren OS. Forest fire and smoke detection using deep learning-based learning without forgetting. Fire ecology. 2023 Feb 17; 19 (1): 9.

[8] Li D, Yu Z, Bao J, Yuan X, Ji Q, Wang M, Zhang K, Yin Y. Lightweight real-time object detection for coal mine underground unmanned vehicles based on improved YOLOv8. InFourth International Conference on Machine Vision, Automatic Identification, and Detection (MVAID 2025) 2025 Sep 19 (Vol. 13793, pp. 188 - 194). SPIE.

[9] Zhang B, Sarhan MH, Goel B, Petculescu S, Ghanem A. SF-TMN: Slow Fast temporal modeling network for surgical phase recognition. International Journal of Computer Assisted Radiology and Surgery. 2024 May; 19 (5): 871 - 80.

[10] Cheng Z, Leng S, Zhang H, Xin Y, Li X, Chen G, Zhu Y, Zhang W, Luo Z, Zhao D, Bing L. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. arXiv preprint arXiv: 2406.07476. 2024 Jun 11.

[11] Yu W, Liu L, Lu J. Exploring facial expression recognition through semi-supervised pre-training and temporal modeling. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2024.

[12] Ameen N, Hosany S, Tarhini A. Consumer interaction with cutting-edge technologies: Implications for future research. Computers in Human Behavior. 2021 Jul 1; 120: 106761.

[13] Shrivastava A, Kumar V, Maurya JP. Cutting-Edge Image Recognition Leveraging Deep Learning and Machine Learning for Enhanced Accuracy. In2024 International Conference on Artificial Intelligence and Quantum Computation-Based Sensor Application (ICAIQSA) 2024 Dec 20 (pp. 1 - 6). IEEE.

[14] Saleh A, Sheaves M, Rahimi Azghadi M. Computer vision and deep learning for fish classification in underwater habitats: A survey. Fish and Fisheries. 2022 Jul; 23 (4): 977 - 99.

[15] Eren B, Demir MH, Mistikoglu S. Recent developments in computer vision and artificial intelligence aided intelligent robotic welding applications. The International Journal of Advanced Manufacturing Technology. 2023 Jun; 126 (11): 4763 - 809.