Scalable Automated Video Labeling for Early Wildfire Smoke Detection with Fast-Then-Precise Two-Stage Inference

Srikantnag A. Nagaraja[1], Chang Zhao[2], and Imre Bartos[3,*]

[1]Department of Electrical and Computer Engineering, University of Florida, Gainesville, Florida, USA
[2]Agronomy Department, UF/IFAS, Gainesville, Florida, USA
[3]Department of Physics, University of Florida, Gainesville, Florida, USA
[*]Corresponding Author: Imre Bartos (email: imrebartos@ufl.edu)

Abstract*—Early wildfire response depends on detecting the first faint appearance of smoke while maintaining low false-alarm rates across diverse cameras, lighting conditions, and environments. A central barrier to progress is the lack of scalable, reliable supervision for subtle early-stage smoke plumes, which makes models brittle under real-world domain shift. We address this challenge by introducing a scalable automated video labeling pipeline based on SAM2 mask propagation, including reverse-frame processing, that enables consistent annotation of early smoke emergence from long time-series camera data. Segmentation masks are converted into tight bounding boxes with targeted human validation to remove cloud artifacts, producing a large and diverse training set spanning fixed-view and zoom-capable wildfire camera networks. Building on this dataset, we design a fast-then-precise two-stage smoke detection system that mirrors operational alerting logic. A high-recall early-warning stage based on RT-DETR prioritizes rapid detection, while a high-precision confirmation stage using YOLOv11 stabilizes alerts and suppresses false positives. The system is evaluated on a strict temporally held-out benchmark consisting of all available FIgLib ignition sequences from 2023 to 2025, which were excluded from training. On this real-world dataset, the early-warning stage achieves high recall (0.94) and detects smoke in 7.0 ± 6.3 minutes on average, while the confirmation stage reaches high precision (0.95) with no false positives observed across the full 2023–2025 evaluation set. These results demonstrate that scalable video labeling combined with complementary two-stage inference enables reliable early wildfire smoke detection under realistic operating conditions.

***Index Terms*—Wildfire early detection system, Forest, Deep Learning, SAM2 Applications, Wildfire Prevention, Wildfire Smoke Detection, Object Detection, Transfer Learning, Deployable Smoke Detection Model, AlertCalifornia.**

I. Introduction

The global frequency of wildfires has risen sharply in recent decades, with recent events occurring in unexpected regions such as Europe and U.S. national parks. In 2021 alone, 58,985 wildfires burned 7.1 million acres, compared to 18,229 fires consuming 1.3 million acres in 1983 a 223% increase [1], [2]. In California, the 2020 wildfires released more than 91 million metric tons of CO_2, approximately 25% of the state's annual fossil fuel emissions [3]. While wildfires can provide certain ecological benefits, effective management is critical to minimize emissions, property damage, and loss of life. Early detection remains central to mitigating these impacts [4].

Because most wildfires begin with visible smoke, smoke serves as the primary indicator for early detection. Camera networks are specifically designed to detect smoke and issue timely alerts, leveraging its upward, dispersive motion as a distinguishing feature [5]. Earlier research relied on handcrafted features such as color, motion, texture, and shape, which were then classified using algorithms like SVMs, k-nearest neighbors, or ensemble methods. While effective in controlled environments, these approaches struggled with real-world variability in lighting, background complexity, and smoke density [5].

Deep learning approaches now dominate wildfire smoke detection, automatically learning discriminative features and outperforming traditional pipelines. CNNs such as AlexNet and subsequent deeper architectures improved accuracy, while frameworks including Faster R-CNN, DeepLabV3+, GANs, and 3D CNNs enhanced spatial localization and temporal modeling. Hybrid pipelines that combine handcrafted cues with real-time detectors like YOLO or lightweight backbones (e.g., MobileNetV2) have further advanced smoke detection for surveillance and edge devices [5].

The effectiveness of deep learning depends heavily on access to large, diverse, and representative datasets [6]. Although data augmentation is commonly used to increase diversity, excessive or poorly designed transformations risk introducing semantic drift for example, color jitter may cause clouds to resemble smoke. Such artifacts can mislead models into learning shortcuts rather than meaningful patterns, thereby undermining generalization [6].

Labels remain essential for supervised learning, as they provide the ground truth needed for training. However, manual annotation is costly and time-consuming. Automated labeling methods have emerged as scalable alternatives, offering improved speed, consistency, and reduced human bias, thereby supporting more efficient and reliable model training [7], [8], [9], [10].

Recent advances such as SAM2 [11], trained on the large-scale SA-V dataset, offer strong generalization and an interactive prompt-based framework. In this work, we employ SAM2 not for segmentation itself, but to accelerate the annotation process by generating object masks that are automatically converted into bounding boxes around smoke. This reduces manual effort and ensures consistent annotations. Comparable applications in medical imaging, remote sensing, and agriculture further highlight SAM-based methods' versatility [12], [13], [14].

For dataset development, we utilized open-source wildfire ignition videos from AlertCalifornia [15] and the HPWREN Fire Ignition Image Library (FIgLib) [16], which we annotated with the assistance of SAM2 [11]. Combined with existing labeled datasets such as D-Fire [17] and Pyro-SDIS [18], this effort enabled us to compile a large-scale, diverse smoke detection dataset designed to support robust model training.

In real-world scenarios, wildfire data is inherently stochastic, shaped by unpredictable environmental and observational factors. Prior studies, such as Guede-Fernández et al. [19], evaluated early detection performance on 16 daytime FIgLib videos (2016–2018) by reporting the time of first smoke identification. More recently, Park and Lee [20] extended this evaluation to 68 additional FIgLib sequences [16], though they did not explicitly clarify whether training and testing data were fully separated, since both phases relied on FIgLib [16]. In contrast, our study emphasizes the stochastic nature of real-world conditions by employing all available FIgLib ignition sequences from 2023–2025 including both daytime and nighttime videos that were strictly excluded from training to assess early detection performance. Additionally, we analyze the impact of different input resolutions (480×480, 640×640, and 960×960 pixels) and present both quantitative and qualitative evaluation metrics to identify the optimal model configuration and trade-offs for various edge-computing scenarios.

The main objectives of this research are as follows:

- Build a large, diverse smoke dataset with minimal manual labeling by using SAM2 video propagation (including reverse-frame labeling) to capture the earliest visible plumes.
- Design and evaluate a two-stage detector where RT-DETR triggers fast, high-recall alerts and YOLOv11 provides high-precision confirmation to suppress false alarms.
- Quantify real-world early-detection performance on a strict held-out benchmark (FIgLib 2023–2025), reporting detection-time distributions and key failure modes across lighting and weather conditions.

II. Data and Methodology

A. Dataset

In our study, we used the Pyro-SDIS dataset as the first of two baseline resources for the detection of wildfire smoke. Developed in collaboration with the French Fire and Rescue Services (SDIS) and the Pyronear association, it contains 33,636 images, including 28,103 with smoke and 31,975 annotated smoke instances, captured by Pyronear cameras and annotated by volunteers. The data set is formatted for compatibility with the Ultralytics YOLO framework, making it well suited for training and evaluating detection models [18].

As the second baseline resource, we use the D-Fire dataset, an image collection of fire and smoke occurrences designed for machine learning and object detection. It contains more than 21,000 images, including 5,867 with only smoke and 4,658 with fire and smoke, with a total of 11,865 annotated smoke bounding boxes in YOLO format [17].

As the third resource, we have used the HPWREN Fire Ignition image Library (FIgLib), maintained by UCSD's High Performance Wireless Research and Education Network (HPWREN). FIgLib contains more than 400 fire ignition sequences captured by fixed field-of-view cameras throughout California, with full-resolution MP4 time-lapse videos and JPEG images spanning 40 minutes before and after ignition events. This large archive provides a unique temporal perspective for studying early plume development. In our work, we labeled the 2016–2021 sequences with the help of SAM2 [11] and a YOLOv11x [21] model trained on the two baseline datasets described above and also performed visual validation of the generated annotations to ensure that misclassified clouds were excluded [16].

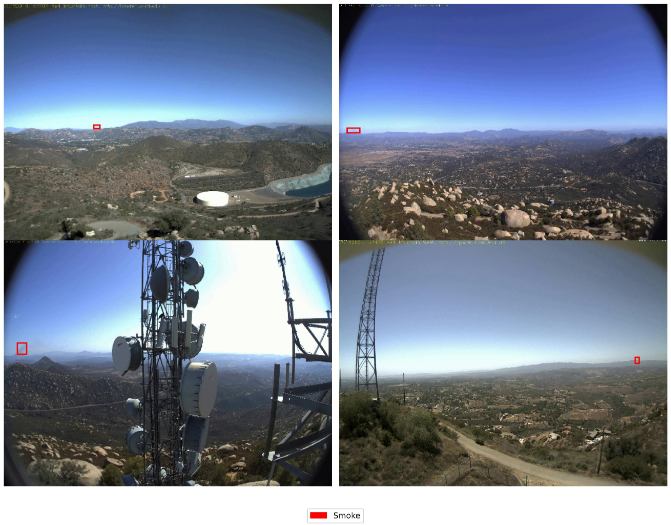

Fig. 1: Sample of SAM-2 labeled FIgLib Dataset [16].

As the fourth resource, we have used the AlertCalifornia dataset, based on the University of California San Diego's statewide monitoring program. AlertCalifornia manages more than 1,190 cameras and sensor arrays throughout California (as of July 2025) and provides open source data for natural hazard monitoring, including wildfires. We obtained relevant wildfire-related camera sequences from YouTube using the yt-dlp library [22] and labeled smoke occurrences with the help of SAM2 [11] and a YOLOv11x [21] model [23] trained on baseline data sets. To improve quality, the generated labels were visually validated, and videos with mislabeled clouds were eliminated. Since AlertCalifornia cameras include zooming functionality, they offer more diverse perspectives compared to the fixed-field HPWREN system.

To ensure a comprehensive and diverse wildfire smoke detection resource, we integrated the Pyro-SDIS, D-Fire, FIgLib (2016–2021), and AlertCalifornia datasets. This combined collection encompasses a broad array of environmental scenarios and optical setups, ranging from fixed-view surveillance to dynamic zoom-enabled systems. The resulting dataset was partitioned into training (80%), validation (15%), and testing (5%) subsets. Furthermore, we evaluated real-world performance using the FIgLib (2022–2025) dataset. Table I provides a detailed breakdown of the total frames and smoke labels across all sources.

B. SAM2 Preliminaries

SAM2 [11] is a promptable visual segmentation model designed for both image and video tasks. The model takes as input a sequence of frames

$$X = x_tt = 1^T[11] \tag{1}$$

and, optionally, a corresponding set of prompts

$$P = P_tt = 1^T.[11] \tag{2}$$

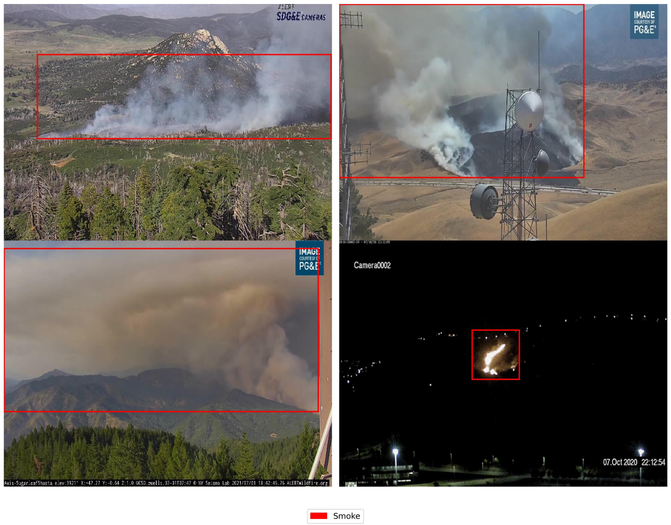

Fig. 2: Sample of SAM2 labeled AlertCalifornia Dataset [15].

TABLE I: Datasets used for wildfire smoke detection

Dataset	No. of Frames	Smoke Frames	Smoke Boxes
D-Fire [17]	21,527	10,525	11,865
Pyro-SDIS [18]	29,537	24,004	27,876
FIgLib (2016-2021) [16]	13,767	3,094	3,094
AlertCalifornia [15]	105,742	85,350	85,350
Final Combined	170,573	122,973	128,185
FIgLib (2023-2025) [16] Real World Testing	11,664	NA*	NA*

Note: N.A = Smoke frames and Smoke Boxes are not applicable for testing dataset.

Its objective is to predict a sequence of segmentation masks

$$Y = y_{t}{}_{t=1}^{T}, [11] \tag{3}$$

where each mask y_t corresponds to frame x_t.

a) Image encoding.: Each frame x_t is first processed by the image encoder E_{img}, producing a feature embedding:

$$F_t = E_{\text{img}}(x_t).[11] \tag{4}$$

b) Prompt encoding.: User prompts P_t, when provided, are processed by the prompt encoder E_{prompt}, generating prompt embeddings:

$$Q_t = E_{\text{prompt}}(P_t).[11] \tag{5}$$

c) Memory mechanism.: To incorporate temporal context, the model maintains a memory bank M_t that stores embeddings from the K most recent frames preceding time t:

$$M_t = \{E_i \mid i \in \{\max(j,0)\},\ j = t-K-1,\dots,t-1\}.[11] \tag{6}$$

d) Memory attention.: The current frame embedding F_t, the memory bank M_t, and the prompt embedding Q_t are integrated via a memory attention mechanism $A(\cdot)$:

$$H_t = A(F_t, M_t, Q_t).[11] \tag{7}$$

e) Mask decoding.: Finally, the mask decoder $D(\cdot)$ uses the attended features to predict the segmentation mask:

$$y_t = D(H_t) = D!\left(A(F_t, M_t, Q_t)\right), \quad t = 1, \dots, T.[11] \tag{8}$$

In our workflow, we employed SAM2 to automatically generate YOLO-compatible labels from *FIgLib* [16] and *AlertCalifornia* [15] video streams. Following mask propagation, these segmentation masks were converted into YOLO bounding boxes to serve as training labels. This was achieved by extracting the four spatial extrema $(x_{min}, y_{min}, x_{max}, y_{max})$ of each mask to define the tightest possible axis-aligned rectangle enclosing the detected object.

C. Object Detection Models

a) Study Context: We use YOLOv11 and RT-DETR in our smoke detection study to balance accuracy, speed, and scalability. YOLOv11 provides a wide range of model sizes for deployment, while RT-DETR offers NMS-free inference with competitive accuracy.

b) YOLOv11 for Object Detection: YOLOv11 [21] continues the single-stage detection paradigm by predicting bounding boxes and class probabilities in a unified framework. The detection loss combines localization, objectness, and classification terms:

$$\mathcal{L} = \lambda_{\text{box}}\mathcal{L}\text{IoU} + \lambda\text{obj}\mathcal{L}\text{obj} + \lambda\text{cls}\mathcal{L}\text{cls}, [21] \tag{9}$$

where $\mathcal{L}$IoU is the box regression loss, $\mathcal{L}_{\text{obj}}$ is binary cross-entropy on objectness, and $\mathcal{L}_{\text{cls}}$ is the classification loss. During inference, Non-Maximum Suppression (NMS) removes duplicates by discarding boxes b_i with:

$$\text{IoU}(b_i, b) = \frac{|b_i \cap b|}{|b_i \cup b|} > \tau, [21] \tag{10}$$

where b is the top-scoring box and τ is an IoU threshold.

c) RT-DETR for Object Detection: RT-DETR [24] is a real-time Transformer-based detector that removes the need for NMS. It predicts a fixed set of objects using bipartite matching, with a loss defined as:

$$\mathcal{L} = \lambda_{\text{cls}}\mathcal{L}\text{cls} + \lambda\text{box}\mathcal{L}\text{box}, [24] \tag{11}$$

where $\mathcal{L}$cls is focal loss and $\mathcal{L}_{\text{box}}$ combines ℓ_1 and GIoU losses.

D. Labeling AlertCalifornia Dataset

We trained our base model, YOLOv11x [21], using the D-Fire [17] and Pyronear Pyro-SDIS [18] datasets (further details about the base model are provided in Appendix IV-A). To extend labeling, we employed the AlertCalifornia dataset [15], which was divided into segments of 3,000 frames. These frames were processed with the SAM2 model [11] for segmentation propagation, using prompts generated by YOLOv11x [21]. Following the propagation phase, segmentation masks from SAM2 were converted into YOLO-compliant labels. This was achieved by extracting the four spatial extrema of each mask to generate axis-aligned bounding boxes automatically. Each SAM2-labeled video was manually reviewed to ensure quality and to exclude false positives caused by clouds or other non-smoke artifacts. All experiments were conducted on the University of Florida's HiPerGator 4.0 system [25], utilizing

NVIDIA DGX B200 GPUs [26]. Examples of the resulting labeling process are shown in Figures 3 and 4.

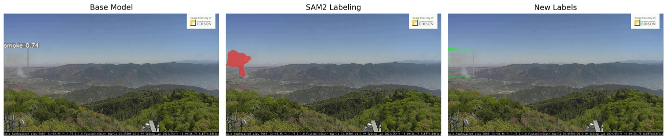

Fig. 3: Labeling workflow: detection by YOLOv11x (left), segmentation propagation with SAM2 (middle), and conversion into refined YOLO bounding boxes (right).

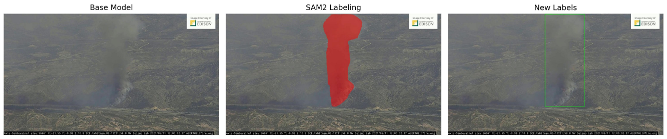

Fig. 4: Example where base YOLOv11x failed to detect smoke, but SAM2, leveraging memory attention, successfully segmented the smoke plume. The segmentation was subsequently converted into YOLO labels.

SAM2 demonstrated the ability to generate accurate smoke labels even under nighttime conditions. Its memory attention mechanism, combined with training on a large-scale dataset, enabled reliable segmentation of smoke regardless of lighting. Figures 5 and 6 illustrate examples where SAM2 successfully labeled nighttime smoke that the base model failed to detect.

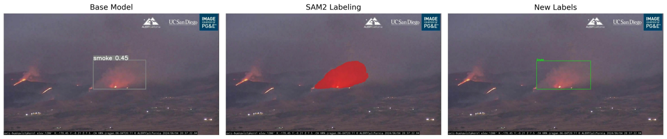

Fig. 5: Transition from daylight to evening: both the base model and SAM2 identified smoke, with SAM2 providing refined segmentation that propagated reliably across frames.

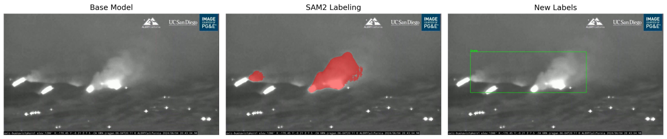

Fig. 6: Nighttime scenario: while the base model did not identify smoke, SAM2 successfully segmented the plume and propagated the labels through low-light conditions.

E. Labeling FigLib Dataset

The HPWREN Fire Ignition Library (FigLib) [16] provides image sequences where each frame corresponds to a one-minute interval, capturing the temporal progression of wildland fire ignitions. Each video spans 40 minutes prior to and 40 minutes following the initial appearance of visible smoke, creating a complete 80-minute window around ignition events. This time-series structure makes the dataset highly valuable for studying the dynamics of early fire development, but also introduces challenges for labeling: smoke may emerge gradually, fluctuate in intensity, or be partially obscured across consecutive frames, making consistent annotation difficult. The combination of dense temporal coverage and subtle visual changes highlights both the strength and complexity of using FigLib for neural network training and evaluation.

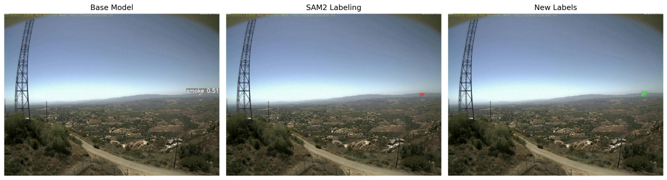

Fig. 7: Example where the base model first detects visible smoke in the Figlib [16] dataset.

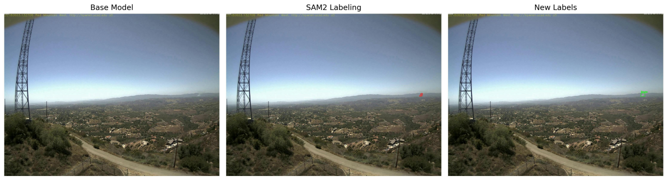

Fig. 8: Using the reverse-frame strategy, SAM2 extends the labeling one frame earlier (–1 minute), capturing smoke emergence before the base model.

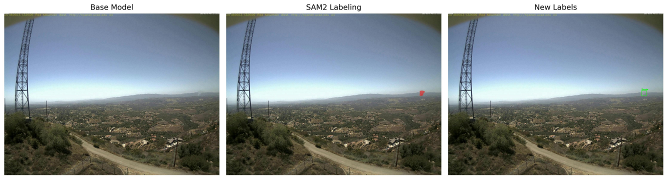

Fig. 9: SAM2 further propagates the labeling to an earlier frame (–2 minutes), highlighting the importance of this strategy for early wildfire smoke detection.

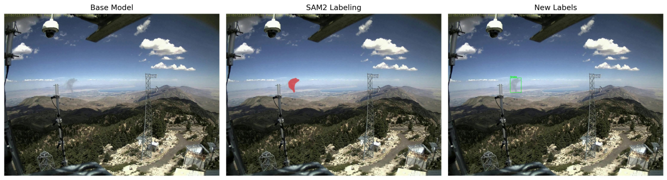

Fig. 10: Example where stacking multiple videos allowed SAM2 to detect smoke that the base model missed, improving labeling for early wildfire smoke detection.

We trained another YOLOv11x [21] model as a base detector using the D-Fire [17], Pyronear Pyro-SDIS [18], and the labeled AlertCalifornia dataset [15], and then applied it to assist in labeling FigLib. Although the FIgLib [16] archive spans from 2016 through 2025, we limited our labeling effort to videos from 2016 to 2021.

Unlike labeling the AlertCalifornia dataset [15], annotating FIgLib [16] proved more challenging due to its time-series structure and subtle smoke emergence. To address this, we employed a reverse-frame labeling strategy. Specifically, each video was reversed and then stacked by year, after which we applied the SAM2 [11] model for labeling. This approach allowed us to capture smoke 2–3 frames prior to ignition, since SAM2's [11] memory attention could propagate detections backward in time: if the base model detected smoke at frame m, SAM2 [11] was able to extend the labeling to frames m–1, m–2, . . . , m–k because of reverse frame stratergy. Figures 7, 8, and 9 illustrate this effect. Furthermore, by stacking multiple reversed videos, we achieved consistent labeling even in cases where the base model failed entirely, as demonstrated in Figure 10. This reverse-frame and stacking strategy is especially significant for early detection of wildfire smoke, as it enables labeling of smoke emergence before ignition is clearly visible, a crucial step toward developing models that can provide earlier warnings in real-world deployments while also reducing human bias in the annotation process.

F. *Real-time Conditions Testing Dataset*

A meaningful assessment of a deep learning model requires evaluating its performance beyond a random hold-out split. While a conventional 5% test set helps verify basic generalization, it does not fully capture the complexities of real-time wildfire conditions. Our training set integrated multiple heterogeneous sources including D-Fire [17], Pyronear Pyro-SDIS [18], AlertCalifornia [15], and FigLib [16] sequences from 2016 to 2021 yet testing solely on randomized partitions remains insufficient for measuring temporal robustness and early detection capability.

We utilized sequences from 2023–2025 in the FigLib [16] dataset to facilitate temporally independent testing. By using this time-series independent data, we can more accurately assess how the model generalizes to future environmental conditions. These sequences contain dense time-series ignition events, enabling us to measure both overall detection accuracy and the model's ability to pinpoint the first visible appearance of smoke in previously unseen footage. Because this dataset differs from the training split in environmental conditions, acquisition dates, illumination patterns, and camera configurations, it provides a challenging and realistic benchmark for assessing robustness and early-detection performance under operational, real-world conditions.

III. Results and Discussion

A. *Detection Model Training and Evaluation*

For model development, the datasets summarized in Table I were split into 80% training, 15% validation, and 5% testing. All models were trained for 300 epochs with a batch size of 64 and an image size of 640 by 640 pixels. A cosine learning rate scheduler was applied with an initial learning rate of 0.0001 and a final learning rate of 0.00001. The warmup period was set to 2 epochs. The Adam optimizer was used. Training was carried out on NVIDIA B200 GPUs through the HiPerGator 4 [25] supercomputing facility at the University of Florida. Default augmentation settings provided by Ultralytics were employed, as augmentation was not explicitly disabled.

To ensure a fair comparison, a range of YOLOv11 [21] architectures—from the lightweight YOLOv11n to the larger YOLOv11x—were trained. For RT-DETR [24], both RT-DETR-l and RT-DETR-x variants were evaluated. Additionally, the RT-DETR-l model was trained with higher input resolutions of 960 by 960 and 480 by 480 pixels to examine the influence of image size on false-positive behavior and early-detection capability, as discussed further in Subsection C.

The evaluation metrics on the test dataset are summarized in Table II. Among the YOLO models, YOLOv11l [21] achieved the highest precision (0.9503), indicating fewer false positives, while YOLOv11x [21] obtained the best mAP50-95 score (0.8801), reflecting robust overall detection performance. The relatively high mAP50-95 values across YOLO variants can be attributed in part to effective Non-Maximum Suppression (NMS) [21], which improves multi-object localization. On the other hand, RT-DETR-l [24] achieved the highest recall (0.9373), a key advantage for early wildfire smoke detection. High recall ensures fewer false negatives, which is especially critical in this application, since missing an ignition event can be far more costly than a false alarm. These differences in performance trends are further explored in Subsection B through real-world evaluation.

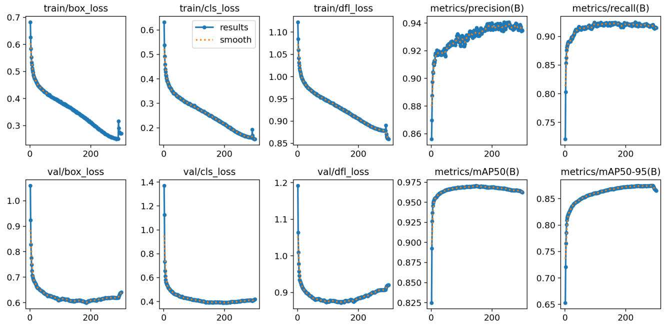

Fig. 11: Training and validation performance for the YOLOv11x model. The first three columns show convergence of Box, Classification, and Distribution Focal loss for both training (top) and validation (bottom) sets. The final two columns display primary accuracy metrics: Precision, Recall, mAP@50, and mAP@50–95. Performance stabilizes after approximately 120 epochs, reaching a precision of 0.94 and mAP@50 of 0.97, indicating robust generalization across diverse wildfire smoke conditions.

The training and validation profiles in Figure 11 and Figure 12 reveal a clear performance distinction between the baseline YOLOv11x model, trained solely on the Pyro-SDIS and D-Fire datasets, and the extended YOLOv11x model trained on a substantially larger and more heterogeneous dataset summarized in Table I. The baseline exhibits higher overall losses, slower convergence, and early metric saturation, with box, classification, and distribution focal losses stabilizing around 0.6, 0.5, and 1.2, respectively. In contrast, the proposed model demonstrates smooth, monotonic convergence across all loss components, reaching significantly lower final values of approximately 0.28, 0.20, and 0.85 without any late-epoch rebounds, indicating stable optimization and reduced overfitting. The validation curves mirror this improvement, maintaining a consistent downward trajectory throughout 300 epochs, while the

TABLE II: Evaluation of different models on the test dataset.

Model	Image Size	Epochs	Precision	Recall	mAP50	mAP50-95
Yolo v11(n) [21]	640	300	0.9336	0.9103	0.9669	0.8529
Yolo v11(l) [21]	640	300	**0.9503**	0.9210	0.9701	0.8784
Yolo v11(x) [21]	640	300	0.9429	0.9225	0.9674	**0.8801**
RT DETR(l) [24]	640	300	0.9461	**0.9373**	**0.9711**	0.8556
RT DETR(l) [24]	480	300	0.9374	0.9303	0.9657	0.8430
RT DETR(l) [24]	960	300	0.9440	0.9365	0.9670	0.8546

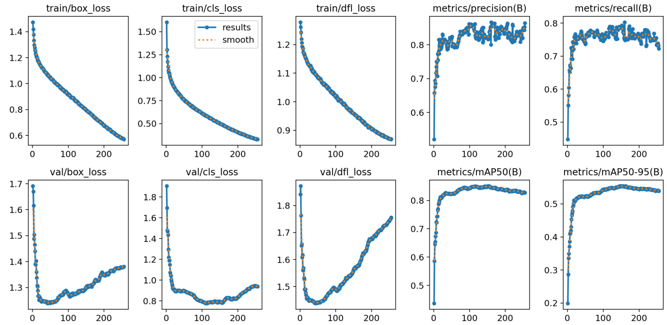

Fig. 12: Training and validation curves of the baseline YOLOv11x model. The first three columns display Box, Classification, and Distribution Focal loss; unlike the proposed model, the validation losses (bottom row) show clear divergence after epoch 100. The final two columns track Precision, Recall, and mAP metrics, which saturate at lower thresholds (mAP@50–95 of 0.52). This behavior reflects limited generalization and a higher tendency toward overfitting on the Pyro-SDIS + D-Fire dataset compared to the extended multi-source dataset.

baseline begins to diverge after roughly 120 epochs. Metric trends further highlight the superiority of the extended model: precision and recall stabilize at 0.94 and 0.91, compared to 0.82 and 0.78 for the baseline, evidencing stronger calibration and higher prediction confidence. Similarly, the extended model achieves mAP@50 = 0.97 and mAP@50–95 = 0.86, surpassing the baseline's 0.80 and 0.52, reflecting substantial gains in both detection accuracy and localization consistency. These results demonstrate that the inclusion of large-scale, diverse, and multi-source annotations decisively enhances model robustness by capturing broader feature distributions across varying atmospheric, geometric, and illumination conditions, resulting in faster convergence, improved generalization, and more reliable early smoke plume detection under real-world wildfire scenarios.

B. Real World Performance

Detection time is defined as the first frame after the dataset-provided ignition reference at which a model produces a smoke detection above threshold.

For real-time evaluation, we first applied our trained YOLOv11x and RT-DETR-L models to 15 daytime fire sequences previously analyzed by Guede-Fernández et al. [19]. This was done to enable direct comparison with established benchmarks in the literature.

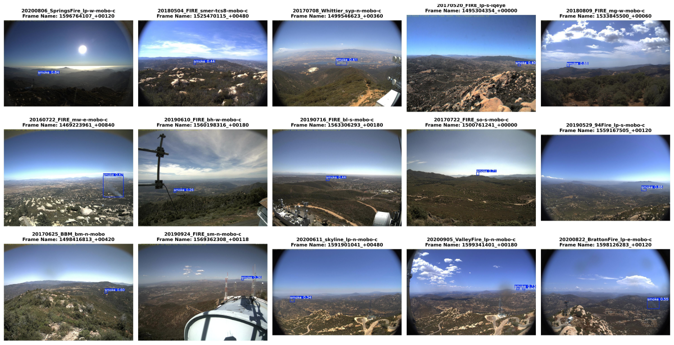

Fig. 13: Example frames of wildfire smoke early detection using our RT-DETR-L model on 15 sequences from the FIgLib dataset [16]. The samples correspond to the events listed in Table III, showing that the model successfully identifies smoke in diverse conditions and viewpoints, as also listed in Guede-Fernández et al. [19].

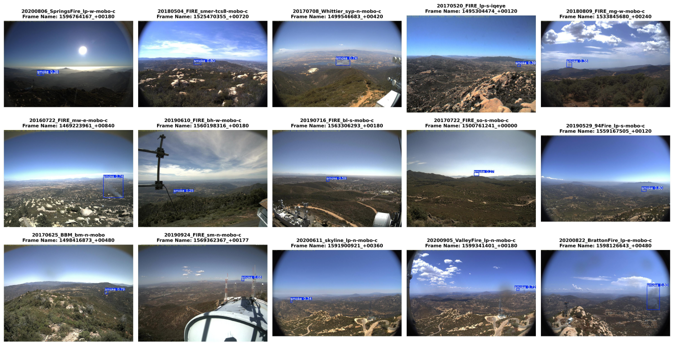

Fig. 14: Example frames of wildfire smoke early detection using our Yolov11-x model on 15 sequences from the FIgLib dataset [16]. The samples correspond to the events listed in Table III, showing that the model successfully identifies smoke in diverse conditions and viewpoints, as also listed in Guede-Fernández et al. [19].

Both models successfully identified smoke in all sequences, with RT-DETR-L achieving earlier detections on average compared to the approaches reported in prior work (Table III). In particular, RT-DETR-L detected smoke plumes within an average of 3.9 minutes from ignition, with a standard deviation of 3.6 minutes. Figures 13 and 14 illustrate representative detections from both models across the 15 sequences. While these results provide a standardized reference for comparison, our real-time testing is conducted using the following methodology, which evaluates performance under more

TABLE III: Daytime Fire Detection Time of smoke sequences extracted from the FIgLib [16] Database. We compare our model with the results reported by Guede-Fernández et al. [19].

Video Name	**Time Elapsed (min)**				
	FRCNN 5000 CYC 3AE [19]	FRCNN 5000 WUP 3AE [19]	RetinaNet 3500 WUP 1CE [19]	YOLOv11-x (Ours)	RT-DETR-L (Ours)
20190529_94Fire_lp-s-mobo-c	3	3	3	**2**	**2**
20190610_FIRE_bh-w-mobo-c	6	5	N.D	**3**	**3**
20190716_FIRE_bl-s-mobo-c	18	18	N.D	**3**	**3**
20190924_FIRE_sm-n-mobo-c	17	7	N.D	3	**2**
20200611_skyline_lp-n-mobo-c	5	**4**	N.D	6	8
20200806_SpringsFire_lp-w-mobo-c	8	**1**	37	3	2
20200822_BrattonFire_lp-e-mobo-c	**2**	5	N.D	8	**2**
20200905_ValleyFire_lp-n-mobo-c	4	**3**	N.D	**3**	**3**
20160722_FIRE_mw-e-mobo-c	**3**	5	N.D	14	14
20170520_FIRE_lp-s-iqeye	8	2	N.D	2	**0**
20170625_BBM_bm-n-mobo	23	21	25	8	**7**
20170708_Whittier_syp-n-mobo-c	4	5	**6**	7	**6**
20170722_FIRE_so-s-mobo-c	6	13	27	**0**	**0**
20180504_FIRE_smer-tcs8-mobo-c	7	9	16	12	**6**
20180809_FIRE_mg-w-mobo-c	6	2	N.D	4	**1**
Mean ± sd	8.0 ± 6.0	6.9 ± 5.8	19.0 ± 12.0	5.2 ± 3.8	**3.9 ± 3.6**

Note: N.D = No Smoke detected by the model. Time Elapsed is after ignition.

TABLE IV: Evaluation of our Yolo v11x and RT DETR l on FIgLib [16] 2023-2025 Dataset.

Year	Model	True Positives	False Positives	No Detections	Total Videos	Detection Time(mean ± sd)
2023	Yolo v11(x) [21]	26	**0**	10	36	11.77 ± 10.68
2023	RT DETR(l) [24]	**27**	1	8	36	**6.78± 4.53**
2024	Yolo v11(x) [21]	**54**	**0**	16	70	9.36 ± 7.84
2024	RT DETR(l) [24]	53	2	15	70	**6.61 ± 6.42**
2025	Yolo v11(x) [21]	**31**	**0**	7	38	12.57±9.36
2025	RT DETR(l) [24]	**31**	3	4	38	**7.73 ± 7.42**
Yolo v11(x) overall		**111**	**0**	**33**	**144**	**10.82 ± 9.32**
RT DETR(l) overall		**111**	**6**	**27**	**144**	**6.98 ± 6.31**

Note: True Positives, False Positives, and No Detection are all numbers of videos in FIgLib Dataset [16] for that particular year. Detection time in minutes after ignition.

diverse and operationally realistic conditions.

To assess performance under more realistic conditions, we further evaluated our models on FIgLib [16] sequences from 2023 to 2025. These videos were not part of the training set, ensuring an unbiased assessment. Unlike the benchmark evaluation, this testing encompassed all available sequences in FIgLib [16], including both daytime and nighttime scenarios. This design aligns with our training dataset, which also incorporates nighttime labels as discussed in Subsection II-D, and therefore provides a more comprehensive assessment of real-world applicability across varying illumination conditions.

Across the 144 sequences spanning 2023–2025, both models successfully detected smoke in 111 cases, reflecting strong overall robustness under real-world conditions. YOLOv11x followed a more conservative detection strategy, recording zero false positives across all years while retaining nearly the same number of true positives as RT-DETR-L. This precision-oriented behavior is reflected in its mean detection time of 10.82 ± 9.32 10.82±9.32 minutes (Table IV).

RT-DETR-L, by contrast, prioritized rapid identification and consistently delivered earlier detections, achieving an average of 6.98 ± 6.31 6.98±6.31 minutes (Table IV), corresponding to roughly a 35% reduction in latency compared to YOLOv11x. Notably, RT-DETR-L maintained mean detection times below 10 minutes in every year from 2023 to 2025, whereas YOLOv11x exceeded this threshold in two of the three years (2023 and 2025). RT-DETR-L also demonstrated the highest recall among all evaluated models (Table II), consistently capturing a larger fraction of ignition events. Although this approach introduced a small number of false positives, the rate remained low and stable across the dataset, preserving its practical reliability.

Together, these results emphasize how the two models complement one another: RT-DETR-L offers fast, recall-driven early detection, while YOLOv11x provides a highly reliable, precision-focused confirmation step. This balance aligns with the composite strategy outlined in this work, enabling both timely alerts and stable refinement. As shown in Figure 15, both models effectively handled diverse real-world challenges including nighttime conditions, variable illumination, lightning activity, and thin moving clouds—underscoring the generalizability of the overall system.

TABLE V: Evaluation of our RT-DETR(l) model on the FIgLib [16] 2023–2025 dataset based on different input image sizes.

Year	Model	Input Image Size	True Positives	False Positives	No Detections	Total Videos	Detection Time (mean ± sd)
2023	RT DETR(l) [24]	640	27	1	8	36	6.78 ± 4.53
2023	RT DETR(l) [24]	480	**26**	1	9	36	**6.52** ± **4.36**
2023	RT DETR(l) [24]	960	**28**	1	7	36	7.86 ± 6.17
2024	RT DETR(l) [24]	640	53	2	15	70	**6.61** ± **6.42**
2024	RT DETR(l) [24]	480	52	6	12	70	7.40 ± 7.49
2024	RT DETR(l) [24]	960	**57**	3	10	70	7.33 ± 6.25
2025	RT DETR(l) [24]	640	31	3	4	38	**7.73** ± **7.42**
2025	RT DETR(l) [24]	480	28	4	6	38	8.15 ± 9.3
2025	RT DETR(l) [24]	960	31	3	4	38	9 ± 8.2
RT DETR(l) overall		**640**	**111**	**6**	**27**	**144**	**6.98** ± **6.31**
RT DETR(l) overall		**480**	**106**	**11**	**27**	**144**	**7.41** ± **7.44**
RT DETR(l) overall		**960**	**116**	**7**	**21**	**144**	**8.03** ± **6.89**

Note: True Positives, False Positives, and No Detections are all numbers of videos in the FIgLib Dataset [16] for that particular year. Detection time in minutes after ignition.

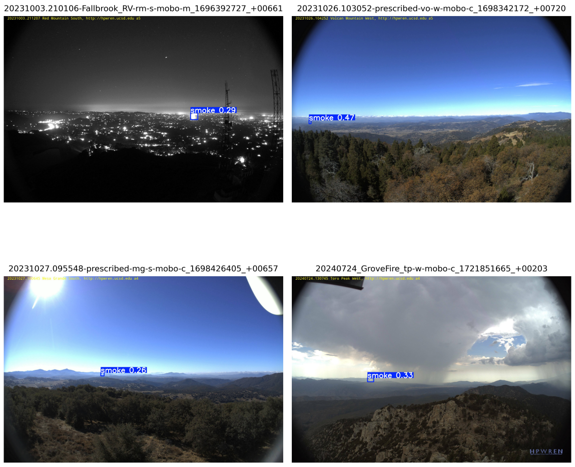
20231003.210106-Fallbrook_RV-rm-s-mobo-m_1696392727_+00661
smoke 0.29
20231026.103052-prescribed-vo-w-mobo-c_1698342172_+00720
smoke 0.47
20231027.095548-prescribed-mg-s-mobo-c_1698426405_+00657
smoke 0.26
20240724_GroveFire_tp-w-mobo-c_1721851665_+00203
smoke 0.33
HPWREN

Fig. 15: Examples of smoke plume detection under diverse conditions. Top-left: nighttime detection; Top-right: detection with thin clouds below the camera height; Bottom-left: detection under bright sunlight; Bottom-right: detection of a lightning-ignited fire plume.

TABLE VI: Analysis of the 33 videos where YOLOv11x failed to detect smoke (as referenced in Table IV).

Condition	Number of Videos
Night time	14
Foggy conditions	7
Clear daylight	7
Heavy cloud	3
Low light	2
Total	**33**

Upon closer examination, we observed that out of the 33 failed FIgLib videos reported in Table VI, 14 corresponded to ignitions occurring at night. Additional failures were observed under foggy conditions (7 cases), clear daylight scenarios (7 cases), as well as during rain, heavy cloud cover, and other low-light situations. Figure 16 illustrates representative examples of these challenging scenarios where detection was unsuccessful.

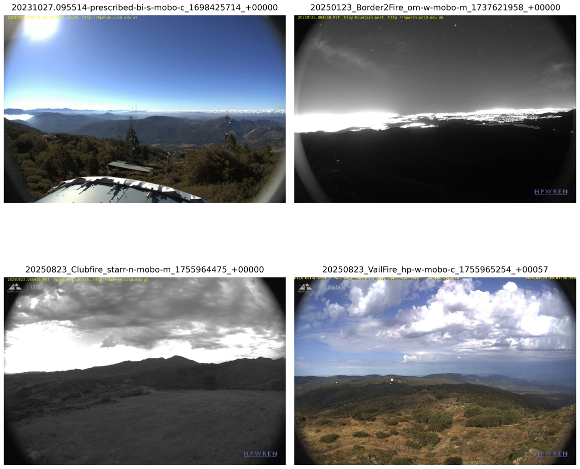
20231027.095514-prescribed-bi-s-mobo-c_1698425714_+00000
20250123_Border2Fire_om-w-mobo-m_1737621958_+00000
20250823_Clubfire_starr-n-mobo-m_1755964475_+00000
20250823_VailFire_hp-w-mobo-c_1755965254_+00057
HPWREN

Fig. 16: Example conditions under which the YOLOv11x model failed to detect smoke. Top-left: foggy condition; top-right: nighttime condition; bottom-left: low-light condition; bottom-right: raining condition.

These findings highlight specific limitations of the current models, particularly in handling nighttime and low-light events. Although we have already incorporated nighttime labels from the Alert California dataset into our training process, further dedicated labeling and model adaptation will be necessary to improve robustness under these conditions. Future improvements can be achieved by expanding the use of nighttime and low-light smoke or fire labels and by leveraging advanced approaches such as SAM2's [11] memory attention mechanism and continuous video segmentation, which offer promising pathways for capturing subtle spatiotemporal cues. Combined with our initial experiments on nighttime labeling, as described in Subsection II-D, these strategies provide a foundation for building more resilient and adaptable wildfire smoke detection systems.

C. Performance based on Image Size

We varied the input image size of the RT-DETR(l) model to analyze how resolution influences both detection performance and processing time. As summarized in Table IV, increasing the input resolution from 640×640 to 960×960 pixels led to a slight improvement in true positives (116 vs. 111 across 2023–2025), accompanied by a marginal rise in false positives. This higher resolution enhanced detection sensitivity but increased the mean detection time by roughly one minute and the standard deviation by 0.5 minutes, indicating a trade-off between accuracy and efficiency.

It is important to note that the current evaluation relies on open-source static datasets rather than continuous live camera feeds. Consequently, metrics such as precision and recall provide a controlled but limited view of real-world performance. In future deployments, when continuous camera access becomes available, additional operational metrics such as the estimated number of alerts per day per 100 cameras or false-alert rates under varying environmental conditions can be incorporated. These qualitative and time-based indicators will provide a more realistic assessment of model reliability and practical usability for real-time wildfire monitoring systems.

IV. Conclusion

The primary contribution of this work is a scalable automated video labeling pipeline that enables reliable supervision of early wildfire smoke, paired with a complementary two-stage detection system optimized for rapid alerts and low false-alarm rates.

We presented a deployable wildfire smoke detection framework that couples SAM2-assisted automated labeling with a composite two-stage detection strategy built from RT-DETR and YOLOv11. By using forward-propagation labeling and reverse-frame propagation for FIgLib sequences, we transformed the annotation bottleneck into a scalable pipeline capable of generating diverse supervision across fixed and zoomable cameras, variable lighting, and challenging environmental conditions.

Within this framework, RT-DETR's high recall and rapid inference enable fast early decisions, while YOLOv11's high precision provides a refined confirmation stage that suppresses false alarms. Extensive evaluations on unseen FIgLib footage (2023–2025) demonstrate that this combined approach consistently detects early smoke while preserving robustness and generalizability across real-world scenarios.

While RT-DETR and YOLOv11 are used here, the framework is detector-agnostic and can incorporate any high-recall trigger and high-precision confirmation model.

Overall, this work delivers a practical path from raw camera streams to trustworthy early wildfire alerts: an automated labeling engine, a generalizable dataset, and a complementary two-stage detection pipeline that balances early detection with reliable refinement. Together, these components move the field closer to an "any camera, any time" capability for real-world wildfire monitoring.

Data Availability

The data is available on request to authors.

Appendix

A. Base Model

The YOLOv11x [21] and RT-DETR-l [24] base models were trained on a unified dataset constructed by merging the D-Fire and Pyro-SDIS collections. The combined dataset was partitioned into 80% for training, 15% for validation, and 5% for testing. Training was conducted for 300 epochs with a batch size of 64 and an input resolution of 640 pixels. A cosine-annealing learning-rate schedule was employed, decreasing from 0.0001 to 0.00001 with a two-epoch warm-up period, and the Adam optimizer was used for model optimization.

TABLE VII: Evaluation of base models on the test dataset.

Model	Precision	Recall	mAP@50	mAP@50:95
YOLOv11x [21]	0.8362	0.8187	0.8554	0.5450
RT-DETR-l [24]	0.8518	0.7807	0.8300	0.5100

The trained YOLOv11x model was subsequently employed, in conjunction with SAM2, to automatically annotate the Alert California dataset as described in Section II-D. The quantitative evaluation of the YOLOv11x and RT-DETR-l base models on the test set is summarized in Table VII.

B. Performance of base model

As shown in Table VIII, the YOLOv11x and RT-DETR-L base models trained solely on the D-Fire and Pyro-SDIS datasets exhibit a higher number of false positives compared to our model trained on a more diverse dataset (Table IV). For instance, the RT-DETR-L model produced 37 false positives even before ignition. Although it successfully detected 96 fires at the moment of ignition in real-time, such a high false-positive rate would increase the need for human oversight and reduce the overall reliability of the system.

TABLE VIII: Evaluation of base models on the FIgLib [16] 2023–2025 dataset.

Year	Model	TP	FP	ND	Total	DT
2023	YV11x (BM) [21]	28	4	4	36	10.38 ± 9.04
2023	RTD-l (BM) [24]	30	6	2	36	5.11 ± 7.34
2024	YV11x (BM) [21]	49	8	13	70	7.07 ± 6.51
2024	RTD-l (BM) [24]	43	20	7	70	5.96 ± 6.23
2025	YV11x (BM) [21]	27	6	5	38	8.44 ± 7.70
2025	RTD-l (BM) [24]	23	11	4	38	5.43 ± 7.41
YV11x	(overall)	**104**	**18**	**22**	**144**	**8.63 ± 7.75**
RTD-l	(overall)	**96**	**37**	**13**	**144**	**5.57 ± 6.82**

Note: TP = True Positives, FP = False Positives, ND = No Detections, DT = Detection Time (mean ± sd). YV11x = YOLOv11x, RTD-l = RT-DETR-l, BM = Base Model trained only on open-source Pyro-SDIS and D-Fire datasets. Detection time is reported in minutes after ignition.

References

1 National Interagency Fire Center, "National fire news," https://www.nifc.gov/fire-information/nfn, [Online; accessed 20-August-2025].

2 National Interagency Fire center, "Wildfires and acres," https://www.nifc.gov/fire-information/statistics/wildfires, [Online; accessed 20-August-2025].

3 Alberts, E. C., ""off the chart": Co2 from california fires dwarf state's fossil fuel emissions," https://news.mongabay.com/2020/09/off-the-chart-co2-from-california-fires-dwarf-states-fossil-fuel-emissions/, 2020, [Online; accessed 20-August-2025].

4 Zhang, A. and Zhang, A. S., "Real-time wildfire detection and alerting with a novel machine learning approach," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 13, no. 8, 2022. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2022.0130801

5 Shi, J., Wang, W., Gao, Y., and Yu, N., "Optimal placement and intelligent smoke detection algorithm for wildfire-monitoring cameras," *IEEE Access*, vol. 8, pp. 72 326–72 339, 2020.

6 Mumuni, A. and Mumuni, F., "Data augmentation: A comprehensive survey of modern approaches," *Array*, vol. 16, p. 100258, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2590005622000911

7 Dietterich, T. G., “Machine learning for sequential data: A review,” in *Structural, Syntactic, and Statistical Pattern Recognition (SSPR/SPR 2002)*, ser. Lecture Notes in Computer Science, vol. 2396. Springer, 2002, pp. 15–30.

8 Ng, A. Y. and Jordan, M. I., “On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 14, 2001, pp. 841–848.

9 Sebastiani, F., “Machine learning in automated text categorization,” *ACM Computing Surveys*, vol. 34, no. 1, pp. 1–47, 2002.

10 Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L., “Imagenet large scale visual recognition challenge,” *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

11 Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K. V., Carion, N., Wu, C.-Y., Girshick, R., Dollár, P., and Feichtenhofer, C., “Sam 2: Segment anything in images and videos,” *arXiv preprint*, 2025.

12 Zhang, Y., Shen, Z., and Jiao, R., “Segment anything model for medical image segmentation: Current applications and future directions,” *Computers in Biology and Medicine*, vol. 171, p. 108238, 2024.

13 Parulekar, B., Singh, N., and Ramiya, A. M., “Evaluation of segment anything model (sam) for automated labelling in machine learning classification of uav geospatial data,” *Earth Science Informatics*, vol. 17, pp. 4407–4418, 2024.

14 Kovačević, V., Pejak, B., and Marko, O., “Enhancing machine learning crop classification models through sam-based field delineation based on satellite imagery,” in *2024 12th International Conference on Agro-Geoinformatics (Agro-Geoinformatics)*. IEEE, 2024, pp. 1–4.

15 ALERTCalifornia Program, “Alertcalifornia: Statewide wildfire and natural hazard monitoring network,” https://alertcalifornia.ucsd.edu/, 2025, university of California, San Diego. Accessed July 2025.

16 Braun, H.-W., “Hpwren fire ignition image library (figlib),” https://www.hpwren.ucsd.edu/HPWREN-FIgLib-Data/, 2020, high Performance Wireless Research and Education Network (HPWREN), University of California, San Diego. Last updated: December 18, 2024.

17 Borges de Venâncio, P. V. A., Lisboa, A. C., and Barbosa, A. V., “An automatic fire detection system based on deep convolutional neural networks for low-power, resource-constrained devices,” *Neural Computing and Applications*, 2022.

18 Team, P., “Pyro-sdis dataset,” 2024. [Online]. Available: https://huggingface.co/pyronear/pyro-sdis

19 Guede-Fernández, F., Martins, L., de Almeida, R. V., Gamboa, H., and Vieira, P., “A deep learning based object identification system for forest fire detection,” *Fire*, vol. 4, no. 4, p. 75, 2021, † These authors contributed equally to this work. * Corresponding author: Pedro Vieira.

20 Park, G. and et al., “Wildfire smoke detection enhanced by image augmentation with stylegan2-ada for YOLOv8 and RT-DETR models,” *Remote Sensing*, vol. 17, no. 5, p. 1203, 2025.

21 Khanam, R. and Hussain, M., “Yolov11: An overview of the key architectural enhancements,” *arXiv preprint arXiv:2410.17725*, 2024.

22 yt-dlp developers, “yt-dlp: A feature-rich command-line audio/video downloader,” https://github.com/yt-dlp/yt-dlp, 2025, accessed September 2025.

23 Wang, C.-Y. and Liao, H.-Y. M., “YOLOv9: Learning what you want to learn using programmable gradient information,” 2024.

24 Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., and Chen, J., “DETRs Beat YOLOs on Real-time Object Detection,” *arXiv preprint arXiv:2304.08069*, 2023.

25 University of Florida Research Computing, “Hipergator – research computing, university of florida,” https://www.rc.ufl.edu/hipergator/, 2025, accessed September 6, 2025.

26 NVIDIA Corporation, “NVIDIA dgx b200 datasheet,” https://resources.nvidia.com/en-us-dgx-systems/dgx-b200-datasheet, 2025, accessed September 6, 2025.