

---

# RELIABLE SMOKE DETECTION VIA OPTICAL FLOW-GUIDED FEATURE FUSION AND TRANSFORMER-BASED UNCERTAINTY MODELING

---

Nitish Kumar Mahala<sup>1\*</sup>

Muzammil Khan<sup>2</sup>

Pushpendra Kumar<sup>1</sup>

<sup>1</sup>Department of Mathematics, Bioinformatics and Computer Applications,  
Maulana Azad National Institute of Technology Bhopal, India

<sup>2</sup>The Robotics and Mechatronics Group,  
University of Twente, The Netherlands

223130002@stu.manit.ac.in, m.khan@utwente.nl, pkumarfma@manit.ac.in

August 21, 2025

## ABSTRACT

Fire outbreaks pose critical threats to human life and infrastructure, necessitating high-fidelity early-warning systems that detect combustion precursors such as smoke. However, smoke plumes exhibit complex spatiotemporal dynamics influenced by illumination variability, flow kinematics, and environmental noise, undermining the reliability of traditional detectors. To address these challenges without the logistical complexity of multi-sensor arrays, we propose an information-fusion framework by integrating smoke feature representations extracted from monocular imagery. Specifically, a Two-Phase Uncertainty-Aware Shifted Windows Transformer for robust and reliable smoke detection, leveraging a novel smoke segmentation dataset, constructed via optical flow-based motion encoding, is proposed. The optical flow estimation is performed with a four-color-theorem-inspired dual-phase level-set fractional-order variational model, which preserves motion discontinuities. The resulting color-encoded optical flow maps are fused with appearance cues via a Gaussian Mixture Model to generate binary segmentation masks of the smoke regions. These fused representations are fed into the novel Shifted-Windows Transformer, which is augmented with a multi-scale uncertainty estimation head and trained under a two-phase learning regimen. First learning phase optimizes smoke detection accuracy, while during the second phase, the model learns to estimate plausibility confidence in its predictions by jointly modeling aleatoric and epistemic uncertainties. Extensive experiments using multiple evaluation metrics and comparative analysis with state-of-the-art approaches demonstrate superior generalization and robustness, offering a reliable solution for early fire detection in surveillance, industrial safety, and autonomous monitoring applications.

**Keywords** Fractional-order variational model · Gaussian mixture fusion · Plausibility confidence · Shifted windows transformer · Smoke detection.

## 1 Introduction

Smoke detection is a pivotal task in computer vision and image processing, underpinning critical applications in fire safety, industrial automation, and surveillance. As the earliest visible manifestation of combustion, smoke enables

---

\*Preprint Notice - This work has been submitted to *Fire Safety Journal* for possible publication. This is the author's original manuscript and has not undergone peer review. Subsequent versions of this manuscript may differ from this one. The final published version, if accepted, will be available via Fire Safety Journal. Corresponding author: 223130002@stu.manit.ac.in

proactive hazard mitigation, thereby reducing risks to human life and infrastructure. Conventional approaches [1] employ physical sensors such as photoelectric, ionization, carbon monoxide, and thermal detectors, which monitor particle concentrations or temperature variations. While these methods perform reliably in enclosed environments, they suffer from limited spatial coverage and an inability to capture the spatiotemporal evolution of smoke plumes in open or large-scale settings.

Given these constraints, vision-based techniques have emerged to exploit the rich spatial and temporal information in video streams. Yet, smoke detection remains a non-trivial challenge due to the non-deterministic dynamics of plumes, which deform under varying illumination, turbulent airflow, and heterogeneous combustion sources. Moreover, smoke shares spectral and textural properties with fog, clouds, and dust, reducing the discriminative power of conventional detection models. Recent studies [2, 3] demonstrate that fusing complementary motion and appearance cues from disparate sensor modalities can mitigate these limitations. Building on this insight, the principal contribution of the proposed study is a feature-level fusion framework, which integrates the following components:

### 1.1 Key contributions

1. **Four-Color-theorem-inspired Dual-phase Level-set Fractional Order Variational (FCDLe-FOV) model**, incorporating a  $L_1$ -norm [4] data fidelity term and Marchaud fractional derivative [5] regularization, to accurately capture smoke motion from monocular RGB image sequences, while preserving discontinuities under non-stationary illumination.
2. **Efficient discretization scheme for FCDLe-FOV model** by integrating Legendre-Fenchel transform-based [6] primal-dual algorithm with Grünwald-Letnikov (GL) fractional derivative [5], which ensures stable convergence.
3. **Gaussian Mixture Model (GMM)-based [7] probabilistic fusion of color-encoded motion maps with RGB image features** to generate precise smoke motion segmentation masks and create a publicly available dataset <sup>2</sup>.
4. **Two-Phase Uncertainty-Aware Shifted-windows Transformer (TP-UAST) model** that leverages the RGB images and their corresponding motion-encoded smoke masks. It is trained in two phases: Phase I optimizes detection accuracy, while Phase II jointly models aleatoric and epistemic uncertainties [8] to generate calibrated confidence estimates.
5. **Comprehensive experimental validation** using various evaluation metrics including accuracy, precision, recall, and F1-score [9]. Uncertainty analysis via Expected Calibration Error (ECE) [8] and reliability diagrams, present confusion matrices, uncertainty histograms, uncertainty vs. error, uncertainty by class, and plausibility analysis and confidence, as well as comparisons against state-of-the-art (SOTA) smoke detection methods.

### 1.2 Related Work

Recent advances in computer vision and deep learning have significantly propelled fire safety research. However, developing robust early-warning systems remains challenging due to dynamic environments. Existing smoke detection methods, such as image-based, video-based, and optical flow information fusion-based, lack integrated feature fusion and uncertainty modeling, limiting their effectiveness in real-world hazard prevention applications.

#### 1.2.1 Image-based techniques

Image-based smoke detection methods predominantly focus on extracting static spatial features such as texture, shape, color, motion patterns, and flicker signatures [10]. Traditional approaches utilize handcrafted feature extraction pipelines to distinguish smoke regions from background clutter [11]. With the advent of deep learning, convolutional architectures such as the multi-scale dual separable convolutional neural network (CNN) proposed by Huo et al. [10], which incorporates a CSPDarknet53 backbone with spatial pyramid pooling, have enhanced multi-scale feature representation capabilities. Similarly, Ke et al. [12] leveraged convolutional layers coupled with batch normalization to improve classification robustness, while Li et al. [13] explored object detection-based CNN models, including Faster R-CNN, R-FCN, SSD, and YOLOv3, for smoke identification tasks.

Although these models achieve considerable performance gains, they are fundamentally constrained by their reliance on single-frame analysis, treating smoke as a purely spatial phenomenon. This absence of temporal modeling or motion analysis leads to high susceptibility to background variations and dynamic scene noise, thereby limiting their applicability in complex real-world scenarios where multi-source information, such as motion and appearance, must be jointly considered.

---

<sup>2</sup><https://www.kaggle.com/datasets/nitishkumarmahala/motion-features-and-appearance-cues-datasets>

### 1.2.2 Video-based techniques

Video-based smoke detection methods exploit temporal consistency to enhance robustness. Lin et al. [14] combined 3D-CNN and R-CNN architectures to jointly encode both motion and appearance cues. Transformer-based frameworks have since advanced the field, including CNN-ViT hybrids [15], dual-branch structures [16], and Swin Transformer variant [17] that improve multi-scale feature extraction and global context modeling. Mardani et al. [18] demonstrated fire localization using transformer-based segmentation.

Despite these advancements, video-based methods typically require computationally intensive architectures and large-scale annotated datasets. Moreover, they primarily focus on feature extraction without explicitly integrating uncertainty modeling or lightweight feature fusion strategies, limiting their scalability for real-time hazard detection in dynamic environments.

### 1.2.3 Optical flow information fusion-based techniques

The dynamic motion patterns of smoke can be effectively captured using optical flow estimation [19, 20]. Variational formulations [21, 22] are commonly employed to solve for the optical flow field representing pixel displacements across frames. Early works applied optical flow for smoke segmentation [19] and forest fire prediction [20], demonstrating its effectiveness for fluid-like objects. Khondaker et al. [23] utilized a fire chromatic model and optical flow for robust fire segmentation, employing Mivia and Zenodo datasets for classification tasks. However, existing methods typically rely on integer-order derivatives, which assume motion continuity and struggle with capturing discontinuities inherent in turbulent or flickering smoke regions. To address these limitations, recent works [5, 24] introduced fractional-order derivative (FOD) formulations that offered greater flexibility in capturing complex smoke motion dynamics.

Building upon advancements in FOD-based optical flow approaches [24], Khan et al. [9] proposed a CNN-based fusion framework that combines dynamical features extracted via FOD-based variational models with static appearance cues such as color, shape, and texture. Further advancements embedded active contour-based level-set segmentation [5] into the FOD framework, enabling precise boundary evolution critical for accurate smoke region delineation. Chunyu et al. [25] proposed an early video smoke detection method based on the fusion of color and motion information. Wu et al. [26] introduced an information fusion framework that integrates dense optical flow with CNN-extracted spatial features to enhance detection robustness. More recently, Kikuta et al. [27] developed a method that fuses optical flow variance with HSV color characteristics for daytime smoke detection.

Despite these advancements, existing fusion approaches lack robustness and transparency by not quantifying the predictive uncertainty, which is an essential requirement for high-reliability fire monitoring applications. These limitations underscore the need for an uncertainty-aware, robust, and generalizable information fusion framework capable of effectively handling diverse environmental conditions while ensuring reliable decision-making.

## 2 Methodology

This study presents a robust and transparent framework for early fire prediction through accurate and reliable smoke detection by integrating fractional-order variational optical flow estimation with a novel TP-UAST model. The proposed framework comprises three primary components: fractional-order motion encoding, probabilistic motion segmentation, and uncertainty-aware classification. Optical flow estimation is formulated as a dual-phase level-set driven variational model, termed FCDLe-FOV model. The objective function incorporates an  $\mathcal{L}_1$ -norm-based data fidelity term [4] to enhance robustness against outliers and a regularization term based on the Marchaud fractional derivative [5], parameterized by a smoothing coefficient  $\lambda$ , to preserve motion discontinuities and complex flow boundaries. To solve the variational formulation efficiently, a Legendre-Fenchel transform-based primal-dual optimization algorithm [6] is employed, combined with GL discretization [5], ensuring stable convergence while accurately capturing topological complexities such as triple junctions. The estimated optical flow fields are subsequently color-encoded to visualize motion dynamics and segmented using the GMM [7]. The GMM leverages probabilistic priors to differentiate smoke-induced motion from background dynamics, producing binary segmentation masks. These masks are then paired with the corresponding RGB frames, facilitating the construction of a high-fidelity smoke segmentation dataset that captures both motion and appearance features. For robust smoke classification, the TP-UAST model is introduced which leverages RGB images and their corresponding segmented color maps. TP-UAST has an architecture incorporating a hierarchical shifted-window self-attention mechanism to capture multi-scale spatial dependencies characteristic of smoke evolution patterns. To enhance reliability, this model is augmented with a multi-scale uncertainty estimation head that jointly models aleatoric uncertainty [8], arising from observational noise, and epistemic uncertainty [8], associated with model capacity limitations. The network is optimized through a two-phase learning regimen: Phase I targets smoke

detection accuracy, while Phase II focuses on predictive uncertainty modeling to improve decision reliability. The complete methodology of the proposed fusion framework is illustrated in Fig. 1.

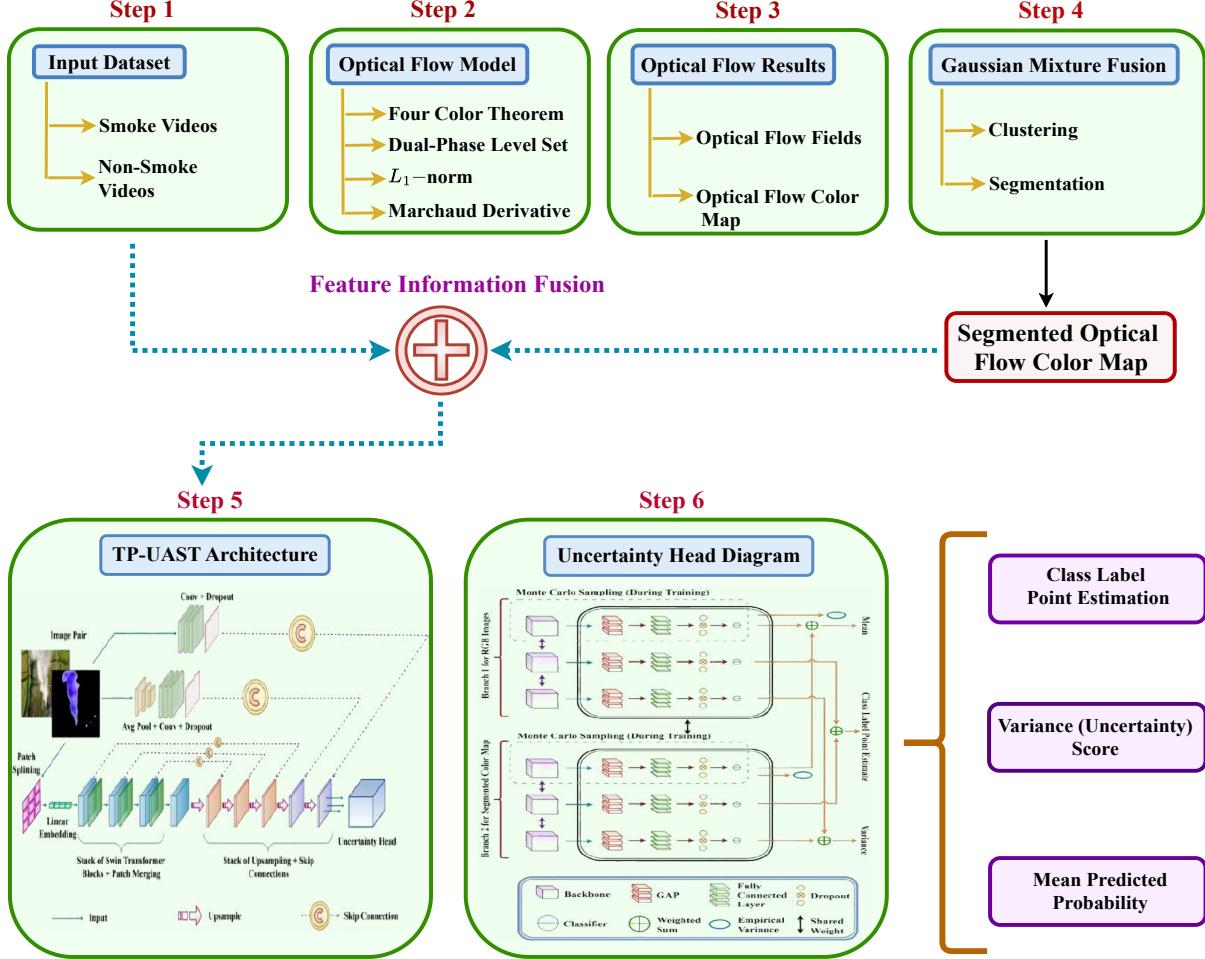


Figure 1: Overview of proposed fusion framework.

## 2.1 Formulation of FCDLe-FOV optical flow model

Optical flow estimation is a fundamental task in computer vision. The objective of optical flow estimation is to examine motion patterns across successive image frames, thereby providing dynamic information about the objects in the scene. Most of the optical flow estimation methods [5, 24] rely on minimizing the continuous energy functionals. Therefore, the proposed FCDLe-FOV model is given as,

$$\mathcal{N}(\mathbf{Z}) = \int_{\Omega} \left[ \lambda |\mathbb{I}_t + \nabla \mathbb{I}^T \mathbf{Z}| + \|D^\alpha \mathbf{Z}^T\|_F + \nu \sum_{a=1}^2 \{\|\delta_{\mathbf{Z},a}^d \nabla \varkappa_{\mathbf{Z},a}\|_{C1}\} \right] d\mathbf{X} \quad (1)$$

where,  $\nabla \mathbb{I} = (\mathbb{I}_x, \mathbb{I}_y)^T$ ,  $\delta_{\mathbf{Z},a} = \delta(\varkappa_{\mathbf{Z},a}) = (\delta(\varkappa_{u,a}), \delta(\varkappa_{v,a}))^T$ ,  $\varkappa_{\mathbf{Z},a} = (\varkappa_{u,a}, \varkappa_{v,a})^T$ ,  $\|\cdot\|_{C1}$  is the sum of  $L_1$ -norm of columns in a matrix,  $\|\cdot\|_F$  is the Frobenius norm,  $\mathbf{X} = (x, y)$ , and  $\mathbf{Z} = (u, v)^T$  represents the optical flow components. Also, superscript  $d$  denotes the diagonal form in a matrix. Here,  $D^\alpha \mathbf{Z} = (D_x^\alpha \mathbf{Z}, D_y^\alpha \mathbf{Z})^T$  represents the Marchaud fractional derivative of order  $\alpha \in (0, 1)$  and  $\lambda$ ,  $\delta$ , and  $\varkappa_{\mathbf{Z},a}$  are the smoothing parameter, Dirac's delta function, and level curves associated with the optical flow fields  $\mathbf{Z}$ . The positive parameter  $\nu$  attracts the contour to the boundary.

Now, in order to minimize the variational functional in (1), it is decomposed based on the algorithm of Chambolle [6]. Hence, the proposed variational functional in (1) is split into two parts and introduces the auxiliary variables  $\hat{\mathbf{Z}} = (\hat{u}, \hat{v})$

that approximate  $\mathbf{Z}$  for a sufficiently small parameter  $\theta$ , which is given as

$$\mathcal{N}_\theta(\hat{\mathbf{Z}}) = \int_{\Omega} \left[ \lambda \mathbb{I}_t + \nabla \mathbb{I}^T \mathbf{Z} | + \frac{1}{2\theta} \|\hat{\mathbf{Z}} - \mathbf{Z}\|_2^2 \right] d\mathbf{X} \quad (2)$$

$$\mathcal{N}_\theta(\mathbf{Z}) = \int_{\Omega} \left[ \frac{1}{2\theta} \|\hat{\mathbf{Z}} - \mathbf{Z}\|_2^2 + \|D^\alpha \mathbf{Z}^T\|_F + \nu \left\{ \sum_{a=1}^2 \|\delta_{\mathbf{Z},a}^d \nabla \varkappa_{\mathbf{Z},a}\|_{C1} \right\} \right] d\mathbf{X} \quad (3)$$

These convex minimization problems can be solved by using the alternating approaches as suggested by Bardeji et al. [4], where at each iteration, either  $\hat{\mathbf{Z}}$  or  $\mathbf{Z}$  is updated. Specifically, the procedure is as follows: first, treat  $\mathbf{Z}$  as constant and solve for  $\hat{\mathbf{Z}}$  by minimizing  $\mathcal{N}_\theta(\hat{\mathbf{Z}})$ , then, treat  $\hat{\mathbf{Z}}$  as constant and solve for  $\mathbf{Z}$  by minimizing  $\mathcal{N}_\theta(\mathbf{Z})$ . Thus, in order to solve the expression in (2), we use the primal-dual algorithm by employing the Legendre-Fenchel transform technique [6] and the concept of calculus of variation [5].

### 2.1.1 Primal-dual formulation by Legendre-Fenchel transform

According to the Legendre-Fenchel transform definition [6], we have

$$\text{if } \phi(p) = |p|, \phi^*(d) = \begin{cases} 0 & , |d| \leq 1 \\ \infty & , |d| > 1 \end{cases}, \text{ then, we get}$$

$$|p| = \phi(p) = \sup_{|d| \leq 1} \left\{ p \cdot d - \phi^*(p) \right\} = \sup_{|d| \leq 1} pd$$

where,  $p = \mathbb{I}_t + \nabla \mathbb{I}^T \mathbf{Z}$ . Thus, the formulation of primal-dual algorithm is given as

$$\mathcal{N}_\theta(\hat{\mathbf{Z}}) = \sup_{|d| \leq 1} \int_{\Omega} \left[ \lambda (\mathbb{I}_t + \nabla \mathbb{I}^T \mathbf{Z}) d + \frac{1}{2\theta} \|\hat{\mathbf{Z}} - \mathbf{Z}\|_2^2 \right] d\mathbf{X} \quad (4)$$

Now, in order to minimize the expression (4) using the Euler-Lagrange equation, we get the following system of equations as

$$\lambda d \nabla I + \frac{1}{\theta} (\hat{\mathbf{Z}} - \mathbf{Z}) = 0 \quad (5)$$

Now, using the equation (5) into equation (4), we obtain

$$T(d) = \int_{\Omega} \left[ \lambda (\mathbb{I}_t + \nabla \mathbb{I}^T \mathbf{Z}) d - \frac{\theta}{2} \lambda^2 d^2 (\nabla \mathbb{I}^T \nabla \mathbb{I}) \right] d\mathbf{X} \quad (6)$$

In order to compute the Frechet derivative of the expression in (6), consider the functional  $T(d)$  at  $d = \bar{d} + \epsilon \psi$ , then, the Frechet derivative of  $T(d)$  is given by

$$\frac{d}{d\tau} T(\bar{d}) = \lambda (\mathbb{I}_t + \nabla \mathbb{I}^T \mathbf{Z}) - \lambda^2 \theta (\nabla \mathbb{I}^T \nabla \mathbb{I}) \bar{d} \quad (7)$$

Now, using the projected gradient ascent technique in equation (7), we get

$$\bar{d}_{temp}^{k+1} = \bar{d}^k + \left[ \lambda (\mathbb{I}_t + \nabla \mathbb{I}^T \mathbf{Z}) - \lambda^2 \theta (\nabla \mathbb{I}^T \nabla \mathbb{I}) \bar{d}^k \right] \quad (8)$$

$$\bar{d}^{k+1} = \begin{cases} \bar{d}_{temp}^{k+1}, & \text{if } |\bar{d}_{temp}^{k+1}| \leq 1 \\ \pm 1, & \text{otherwise} \end{cases} \quad (9)$$

The equations (5), (8), and (9) are solved alternatively to obtain  $\hat{\mathbf{Z}}$ . The minimization of expression in (3) is solved by again decomposing the expression into two parts, and  $\hat{\mathbf{Z}}$  is taken as a constant.

$$\mathcal{N}_\theta(\mathbf{Z}, \varkappa_{\mathbf{Z},a}) = \int_{\Omega} \left[ \frac{1}{2\theta} \|\hat{\mathbf{Z}} - \mathbf{Z}\|_2^2 + \|D^\alpha \mathbf{Z}^T\|_F + \nu \sum_{a=1}^2 (\delta_{\mathbf{Z},a} \|\nabla \varkappa_{\mathbf{Z},a}\|_{C1}) \right] d\mathbf{X} \quad (10)$$

The connection between the four regions and the function  $\mathbf{Z}$  can be established by proposing the four functions such as  $\mathbf{Z}^{++}$ ,  $\mathbf{Z}^{+-}$ ,  $\mathbf{Z}^{-+}$ , and  $\mathbf{Z}^{--}$  according to the Four-Color theorem [28]. These functions effectively restrict  $\mathbf{Z}$  to each of the four regions, as outlined below

$$\mathbf{Z}(x, y) = \mathbf{Z}^{ij}$$

$$\text{where } ij = \begin{cases} ++, & \text{if } (r, s, n\Delta\tau) |\varkappa_{u,1,r,s}^{(n)} > 0 \text{ on } \varkappa_{u,2,r,s}^{(n)} > 0 \\ +- , & \text{if } (r, s, n\Delta\tau) |\varkappa_{u,1,r,s}^{(n)} > 0 \text{ on } \varkappa_{u,2,r,s}^{(n)} < 0 \\ -+, & \text{if } (r, s, n\Delta\tau) |\varkappa_{u,1,r,s}^{(n)} < 0 \text{ on } \varkappa_{u,2,r,s}^{(n)} > 0 \\ --, & \text{if } (r, s, n\Delta\tau) |\varkappa_{u,1,r,s}^{(n)} < 0 \text{ on } \varkappa_{u,2,r,s}^{(n)} < 0 \end{cases}$$

The relation between  $\mathbf{Z}$ , four functions  $\mathbf{Z}^{++}$ ,  $\mathbf{Z}^{+-}$ ,  $\mathbf{Z}^{-+}$ , and  $\mathbf{Z}^{--}$ , and the level set functions  $\varkappa_1$  and  $\varkappa_2$  can again be expressed using the function  $\mathcal{H}$  as

$$\mathbf{Z} = \mathbf{Z}^{++}\mathcal{H}_{\mathbf{Z}_1} + \mathbf{Z}^{+-}\mathcal{H}_{\mathbf{Z}_2} + \mathbf{Z}^{-+}\mathcal{H}_{\mathbf{Z}_3} + \mathbf{Z}^{--}\mathcal{H}_{\mathbf{Z}_4}$$

where,  $\mathcal{H}_{\mathbf{Z}_1} = (\mathcal{H}_{u,1}\mathcal{H}_{u,2}, \mathcal{H}_{v,1}\mathcal{H}_{v,2})^T$ ,  $\mathcal{H}_{\mathbf{Z}_2} = (\mathcal{H}_{u,1}(I - \mathcal{H}_{u,2}), \mathcal{H}_{v,1}(I - \mathcal{H}_{v,2}))^T$ ,  $\mathcal{H}_{\mathbf{Z}_3} = ((I - \mathcal{H}_{u,1})\mathcal{H}_{u,2}, (I - \mathcal{H}_{v,1})\mathcal{H}_{v,2})^T$ ,  $\mathcal{H}_{\mathbf{Z}_4} = ((I - \mathcal{H}_{u,1})(I - \mathcal{H}_{u,2}), (I - \mathcal{H}_{v,1})(I - \mathcal{H}_{v,2}))^T$ . Thus, in accordance with the Vese-Chan model [28], the expression (10) can be written as

$$\mathcal{N}(\mathbf{Z}^{ij}, \varkappa_{\mathbf{Z},a}) = \int_{\Omega} \sum_{i,j} \left( \frac{1}{2\theta} \|\hat{\mathbf{Z}} - \mathbf{Z}^{ij}\|_2^2 + \|D^\alpha \mathbf{Z}^{ij}\|_F \right) \mathcal{H}_{\varkappa_{\mathbf{Z},a}}^i \mathcal{H}_{\varkappa_{\mathbf{Z},a}}^j + \nu \left\{ \sum_{a=1}^2 (\delta_{\mathbf{Z},a} \|\nabla \varkappa_{\mathbf{Z},a}\|_{C1}) \right\} d\mathbf{X} \quad (11)$$

where  $i, j \in \{+, -\}$ , and  $\mathcal{H}_{\varkappa_{\mathbf{Z},a}}^+ = \mathcal{H}_{\varkappa_{\mathbf{Z},a}}$ ,  $\mathcal{H}_{\varkappa_{\mathbf{Z},a}}^- = 1 - \mathcal{H}_{\varkappa_{\mathbf{Z},a}}$  for  $a = 1, 2$ . So, the Euler-Lagrange equations derived from minimizing the variational functional (11) using the calculus of variations are given as

$$\frac{\partial \varkappa_{\mathbf{Z},a}}{\partial \tau} = -\delta_{\mathbf{Z},a} \left[ \frac{1}{2\theta} \sum_{i,j} j \cdot [(\hat{\mathbf{Z}} - \mathbf{Z}^{ij})^2 + \|D^\alpha \mathbf{Z}^{ij}\|_F] - \nu \nabla \cdot \left( \frac{\nabla \varkappa_{\mathbf{Z},a}}{|\nabla \varkappa_{\mathbf{Z},a}|} \right) \right] \quad (12)$$

$$\mathbf{Z}^{ij} = \hat{\mathbf{Z}} - 2\theta \{ (D_-^\alpha D_+^\alpha)^T e \} \mathbf{Z}^{ij} \text{ over } (x, y, \tau) \quad (13)$$

where,  $\partial\tau$  denotes the artificial time step in which the level set functions evolve. These equations allow to compute  $\mathbf{Z}^{ij}$  over the auxiliary flow field  $\hat{\mathbf{Z}}$ .

## 2.2 Numerical discretization and solution

### 2.2.1 Dual-phase level set discretization scheme

We approximate Heaviside's unit step function and Dirac's delta function for numerical implementation purposes as

$$\mathcal{H}(\varkappa) \approx \mathcal{H}_\epsilon(\varkappa) = \frac{1}{2} \left[ 1 + \frac{2}{\pi} \arctan \left( \frac{\varkappa}{\epsilon} \right) \right] \quad \text{and} \quad \delta(\varkappa) \approx \delta_\epsilon(\varkappa) = \frac{1}{\pi} \frac{\epsilon}{\epsilon^2 + \varkappa^2}$$

Let  $\varkappa_{\mathbf{Z},1,r,s}^0$  and  $\varkappa_{\mathbf{Z},2,r,s}^0$  represent the initial approximations of the level surfaces, and  $\varkappa_{\mathbf{Z},1,r,s}^{(n)}$  and  $\varkappa_{\mathbf{Z},2,r,s}^{(n)}$  denote the  $n^{th}$  iterate approximations of the level surfaces. Then, the two-phase level set discretization is performed according to the theory of Vese and Chan [28] is given as

$$\frac{\partial \varkappa_{\mathbf{Z},a}}{\partial \tau} \approx \frac{\varkappa_{\mathbf{Z},a,r,s}^{(n+1)} - \varkappa_{\mathbf{Z},a,r,s}^{(n)}}{\Delta\tau} \quad (14)$$

$$\begin{aligned} \nabla \cdot \left( \frac{\nabla \varkappa_{\mathbf{Z},a}}{|\nabla \varkappa_{\mathbf{Z},a}|} \right) = & C_1 \left( \varkappa_{\mathbf{Z},a,r+1,s}^{(n)} - \varkappa_{\mathbf{Z},a,r,s}^{(n+1)} \right) + C_2 \left( \varkappa_{\mathbf{Z},a,r-1,s}^{(n)} - \varkappa_{\mathbf{Z},a,r,s}^{(n+1)} \right) + C_3 \left( \varkappa_{\mathbf{Z},a,r,s+1}^{(n)} - \varkappa_{\mathbf{Z},a,r,s}^{(n+1)} \right) \\ & + C_4 \left( \varkappa_{\mathbf{Z},a,r,s-1}^{(n)} - \varkappa_{\mathbf{Z},a,r,s}^{(n+1)} \right) \end{aligned} \quad (15)$$

$$\text{where, } C_1 = \left\{ \left( \frac{\varkappa_{\mathbf{Z},k,r+1,s}^{(n)} - \varkappa_{\mathbf{Z},k,r,s}^{(n)}}{h} \right)^2 + \left( \frac{\varkappa_{\mathbf{Z},k,r,s+1}^{(n)} - \varkappa_{\mathbf{Z},k,r,s-1}^{(n)}}{2h} \right)^2 \right\}^{-\frac{1}{2}},$$

$$C_2 = \left\{ \left( \frac{\varkappa_{\mathbf{Z},k,r,s}^{(n)} - \varkappa_{\mathbf{Z},k,r-1,s}^{(n)}}{h} \right)^2 + \left( \frac{\varkappa_{\mathbf{Z},k,r-1,s+1}^{(n)} - \varkappa_{\mathbf{Z},k,r-1,s-1}^{(n)}}{2h} \right)^2 \right\}^{-\frac{1}{2}}$$

$$C_3 = \left\{ \left( \frac{\varkappa_{\mathbf{Z},k,r+1,s}^{(n)} - \varkappa_{\mathbf{Z},k,r-1,s}^{(n)}}{h} \right)^2 + \left( \frac{\varkappa_{\mathbf{Z},k,r,s+1}^{(n)} - \varkappa_{\mathbf{Z},k,r,s}^{(n)}}{2h} \right)^2 \right\}^{-\frac{1}{2}},$$

$$C_4 = \left\{ \left( \frac{\varkappa_{\mathbf{Z},k,r+1,s}^{(n)} - \varkappa_{\mathbf{Z},k,r-1,s}^{(n)}}{h} \right)^2 + \left( \frac{\varkappa_{\mathbf{Z},k,r,s}^{(n)} - \varkappa_{\mathbf{Z},k,r,s-1}^{(n)}}{2h} \right)^2 \right\}^{-\frac{1}{2}}$$

### 2.2.2 Fractional derivative discretization scheme

Let the dimensions of the optical flow components  $\mathbf{Z} = (\mathbf{u}, \mathbf{v})^T$  be same as those of the reference image of size  $m \times n$ . Now, by using the GL derivative [5], we discretize the optical flow fields  $\mathbf{Z}^{ij}$  as

$$\{(D_-^\alpha D_+^\alpha) e\} \mathbf{Z}^{ij} \approx \sum_{(\bar{r}, \bar{s}) \in \xi(r, s)} w_{q_{\bar{r}\bar{s}}}^\alpha \{\mathbf{Z}^{ij} - \bar{\mathbf{Z}}^{ij}\} \quad (16)$$

Here, the set  $\xi$  represents all the pixels in the neighborhood of the pixel location  $(r, s)$  in both the  $x$  and  $y$  directions, where  $q_{\bar{r}\bar{s}} = \max[|\bar{r} - r|, |\bar{s} - s|]$ .

On using the equations (14), (15), and (16), we get the following discretized system of equations for optical flow  $\mathbf{Z}$  as

$$\mathbf{Z}_{r,s}^{(n+1),ij} = \mathfrak{R}^{-1} \left[ \hat{\mathbf{Z}}_{r,s}^{(n)} + 2\theta \sum_{(\bar{r}, \bar{s}) \in \xi(r, s)} w_{q_{\bar{r}\bar{s}}}^\alpha \mathbf{Z}_{\bar{r}, \bar{s}}^{(n),ij} \right] \quad (17)$$

here,  $\mathbf{Z}^{ij}(\bar{r}, \bar{s}) = \mathbf{Z}_{\bar{r}, \bar{s}}^{ij}$ , and  $\mathfrak{R} = 1 + 2\theta \sum_{(\bar{r}, \bar{s}) \in \xi(r, s)} w_{q_{\bar{r}\bar{s}}}$ .

$$\begin{aligned} \varkappa_{\mathbf{Z}, a, r, s}^{(n+1)} &= \frac{1}{C_a^{(n)}} \left[ \varkappa_{\mathbf{Z}, a, r, s}^{(n)} + \gamma_a^{(n)} \left\{ C_1^n \varkappa_{\mathbf{Z}, a, r+1, s}^{(n)} + C_2^n \varkappa_{\mathbf{Z}, a, r-1, s}^{(n)} + C_3^n \varkappa_{\mathbf{Z}, a, r, s+1}^{(n)} + C_4^n \varkappa_{\mathbf{Z}, a, r, s-1}^{(n)} \right\} \right. \\ &\quad \left. - \frac{\gamma_a^{(n)}}{2} \left\{ \frac{1}{2\theta} \left\{ \sum_i i (\hat{\mathbf{Z}}^{(n)} - \mathbf{Z}^{(n),i+})^d (\hat{\mathbf{Z}}^{(n)} - \mathbf{Z}^{(n),i+}) \right\} + (h^{-\alpha})^2 \right. \right. \\ &\quad \left. \left. \left\{ \sum_i i \text{diag} \left( \sum_{q=0}^W w_q^{(\alpha)} E_{\mathbf{x}}^{-q} \mathbf{Z}^{i+, T} \right)^T \left( \sum_{q=0}^W w_q^{(\alpha)} E_{\mathbf{x}}^{-q} \mathbf{Z}^{p+, T} \right) \right\} \right\} \mathcal{H}(\varkappa_{\mathbf{Z}, 1}^{(n)}) \right. \\ &\quad \left. - \frac{\gamma_a^{(n)}}{2} \left\{ \frac{1}{2\theta} \left\{ \sum_p p (\hat{\mathbf{Z}}^{(n)} - \mathbf{Z}^{(n),p-})^d (\hat{\mathbf{Z}}^{(n)} - \mathbf{Z}^{(n),p-}) \right\} + (h^{-\alpha})^2 \right. \right. \\ &\quad \left. \left. \left\{ \sum_i i \text{diag} \left( \sum_{q=0}^W w_q^{(\alpha)} E_{\mathbf{x}}^{-q} \mathbf{Z}^{i-, T} \right)^T \left( \sum_{q=0}^W w_q^{(\alpha)} E_{\mathbf{x}}^{-q} \mathbf{Z}^{p-, T} \right) \right\} \right\} (2 - \mathcal{H}(\varkappa_{\mathbf{Z}, 2}^{(n)})) \right] \end{aligned} \quad (18)$$

where,  $\gamma_a^{(n)} = \frac{\Delta\tau}{h^2} \delta_\epsilon(\varkappa_a^{(n)}(r, s))$  and  $C_a^{(n)} = 1 + \gamma_a^{(n)} (C_1 + C_2 + C_3 + C_4)$  for  $a = 1$  and  $2$ . Hence, these equations produce the optical flow fields. In this work, the color maps represent the estimated optical flow for smoke as well as non-smoke images, which are further processed to create the segmented binary mask using GMM technique.

### 2.3 Stability and convergence analysis

#### 2.3.1 Stability analysis

In order to analyze the stability of the discretized system of equations (17). The fourier decomposition of  $\mathbf{Z}_{r,s}^{(n),ij}$  can be expressed as  $\mathbf{Z}_{r,s}^{(n),ij} = \mathbf{Z}^{(n),ij} e^{i(kr+ls)}$ . Then, the expression in (17) becomes

$$\begin{aligned} \mathbf{Z}_{r,s}^{(n+1),ij} &= \mathfrak{R}^{-1} \left[ \mathbf{Z}_{r,s}^{(n),ij} + 2\theta \sum_{(\bar{r}, \bar{s}) \in \xi(r, s)} w_{q_{\bar{r}\bar{s}}}^\alpha \mathbf{Z}_{\bar{r}, \bar{s}}^{(n),ij} e^{i(k\bar{r}+l\bar{s})} \right] \\ &= \mathfrak{R}^{-1} \left[ 1 + 2\theta \sum_{(\bar{r}, \bar{s}) \in \xi(r, s)} w_{q_{\bar{r}\bar{s}}}^\alpha e^{i(k\bar{r}+l\bar{s})} \right] \mathbf{Z}_{r,s}^{(n),ij} \\ &= G(k, l) \mathbf{Z}_{r,s}^{(n),ij} \end{aligned} \quad (19)$$

here,  $G(k, l)$  represents the amplification factor of the expression in (17), and  $k$  and  $l$  are the wavenumbers. Hence, the amplification factor is written as

$$G(k, l) = \mathfrak{R}^{-1} \left[ 1 + 2\theta \sum_{(\bar{r}, \bar{s}) \in \xi(r, s)} w_{q_{\bar{r}\bar{s}}}^\alpha e^{i(k\bar{r}+l\bar{s})} \right]$$

Now,

$$\begin{aligned} |G(k, l)| &= \left| \Re^{-1} \left[ 1 + 2\theta \sum_{(\bar{r}, \bar{s}) \in \xi(r, s)} w_{q_{\bar{r}\bar{s}}}^\alpha e^{i(k\bar{r}+l\bar{s})} \right] \right| \\ &\leq \left| \Re^{-1} \left[ \left| 1 + 2\theta \sum_{(\bar{r}, \bar{s}) \in \xi(r, s)} w_{q_{\bar{r}\bar{s}}}^\alpha e^{i(k\bar{r}+l\bar{s})} \right| \right] \right| \end{aligned} \quad (20)$$

Since  $|e^{i(k\bar{r}+l\bar{s})}| = 1$ , then we get

$$\left| \sum_{(\bar{r}, \bar{s}) \in \xi(r, s)} w_{q_{\bar{r}\bar{s}}}^\alpha e^{i(k\bar{r}+l\bar{s})} \right| \leq \sum_{(\bar{r}, \bar{s}) \in \xi(r, s)} |w_{q_{\bar{r}\bar{s}}}^\alpha| \quad (21)$$

Therefore, according to theorem [29], we obtain

$$|\Re^{-1}|[1 + 2\theta \sum_{(\bar{r}, \bar{s}) \in \xi(r, s)} |w_{q_{\bar{r}\bar{s}}}^\alpha|] \leq 1 \quad (22)$$

On simplifying the expression in (22), we have

$$1 + 2\theta \sum_{(\bar{r}, \bar{s}) \in \xi(r, s)} |w_{q_{\bar{r}\bar{s}}}^\alpha| \leq \Re \quad (23)$$

Thus, the proposed FCDLe-FOV model is stable, when  $\sum_{(\bar{r}, \bar{s}) \in \xi(r, s)} |w_{q_{\bar{r}\bar{s}}}^\alpha| \leq \frac{\Re-1}{2\theta}$ .

### 2.3.2 Convergence analysis

Let the sequence  $\{\mathbf{Z}_{r,s}^{(n),ij}\}$  has an error term  $e_{r,s}^{(n),ij}$  such that  $\lim_{n \rightarrow \infty} e_{r,s}^{(n),ij} = 0$  with  $|G(k, l)| \leq 1$ , then,

$$e_{r,s}^{(n+1),ij} = G(k, l)e_{r,s}^{(n),ij} \text{ as given in expression (19).}$$

Now,

$$\left| \mathbf{Z}_{r,s}^{(n+1),ij} - \mathbf{Z}_{r,s}^{(n),ij} \right| = \left| G(k, l)e_{r,s}^{(n),ij} - e_{r,s}^{(n),ij} \right| = e_{r,s}^{(n),ij} |G(k, l) - 1|$$

Since  $\lim_{n \rightarrow \infty} e_{r,s}^{(n),ij} = 0$  and  $|G(k, l) - 1|$  is bounded, therefore

$$\lim_{n \rightarrow \infty} \left| \mathbf{Z}_{r,s}^{(n+1),ij} - \mathbf{Z}_{r,s}^{(n),ij} \right| = 0$$

Hence,  $\{\mathbf{Z}_{r,s}^{(n),ij}\}$  is convergent.

## 2.4 GMM-based binary mask for information fusion

The segmentation pipeline employed to produce the binary mask is shown in Fig. 2. To identify the smoke region of interest (RoI), the system utilizes a dense optical flow color map that encodes motion information from the flow field. However, the spatiotemporal complexity of smoke motion leads to uneven and dispersed pixel intensity values in the color map, influenced by diverse motion dynamics and background noise. To address this, a GMM [7] technique is used to segment the motion map. GMM is a probabilistic clustering method that represents data as a combination of multiple gaussian distributions, allowing it to effectively handle complex intensity patterns. This segmentation facilitates the extraction of motion features, which are then used in the subsequent step.

## 3 Architecture of proposed TP-UAST model

The architecture of the proposed TP-UAST model is illustrated in Fig. 3. It integrates an uncertainty-aware prediction head into a shared Swin Transformer backbone [30], leveraging its hierarchical, window-based self-attention [31] to effectively capture multi-scale spatial features. TP-UAST processes two parallel input branches, one for the RGB image and one for its corresponding segmented color map, each passed independently through the same backbone and decoder modules. To enable reliable uncertainty estimation for the predicted class labels, the model jointly captures epistemic and aleatoric uncertainties. The epistemic uncertainty is optimized via Monte Carlo (MC) sampling with random input transformations, including horizontal flipping and rotations, while ensuring consistency between paired branches by using the same random seed. Aleatoric uncertainty is captured through a dedicated uncertainty-aware prediction head that regresses the predicted mean and variance along with the point prediction. This design enhances both discriminative accuracy and prediction confidence under challenging smoke detection conditions. The core architecture following the input step is as follows:

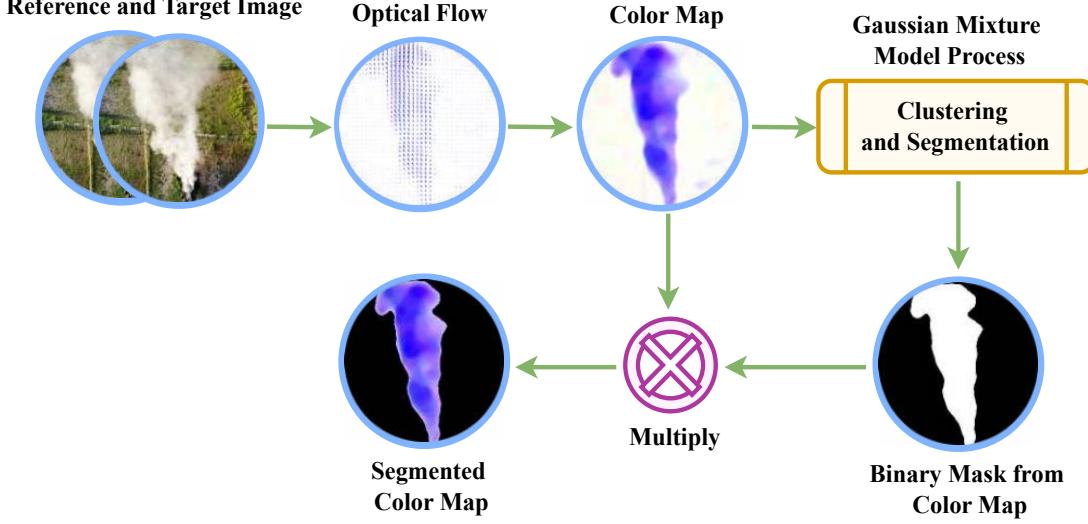


Figure 2: Extraction of smoke motion features using optical flow color maps.

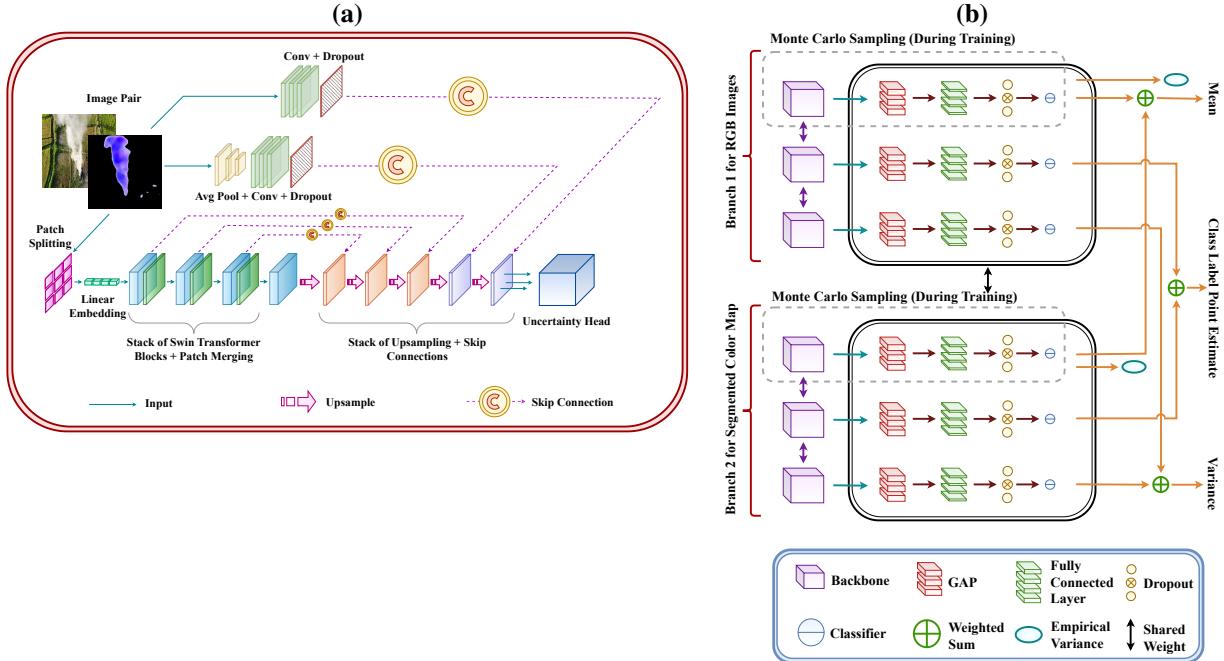


Figure 3: (a) The Swin Transformer backbone; (b) The uncertainty-aware prediction head.

- Patch splitting and linear embedding:** The TP-UAST model pipeline begins with the patch splitting layer [32], which partitions the input sample into non-overlapping patches of size  $4 \times 4$  pixels. Each patch is subsequently linearly embedded [33] into a feature vector of dimension 96, resulting in a sequence of embeddings. This operation effectively reduces the spatial resolution of the input by a factor of 4, while preserving critical spatial and structural information for downstream processing.
- Swin Transformer blocks:** The network comprises four hierarchical stages to refine the feature representations at progressively larger scales, with each stage consisting of multiple Swin Transformer blocks. At the end of each stage, except the final one, patch merging layers [30] are employed to downsample the feature maps by a factor of two while doubling the number of feature channels. In the initial stage, two Swin Transformer blocks operate with four attention heads and a window size of  $7 \times 7$ , processing features at a dimension of 96. The second stage also includes two blocks with four attention heads but doubles the feature dimension to

192 while maintaining the same window size. Stage three expands to four blocks with eight attention heads, further increasing the feature dimension to 384, allowing the model to capture more intricate spatial patterns. Finally, the fourth stage concludes the refinement process with two blocks and eight attention heads, pushing the feature dimension to 768. This structure progressively reduces the spatial resolution and increases the feature dimensionality, allowing the model to focus on increasingly abstract and high-level features.

- **Decoder blocks and skip connections:** Following the transformer blocks, TP-UAST incorporates five sequential decoder blocks to restore the original spatial dimensions of the input, while integrating feature information from earlier stages of the network. Each decoder block begins by upsampling the input feature map through bilinear interpolation, doubling the spatial dimensions. The upsampled feature map is then concatenated with the corresponding feature map from the encoder through skip connections, ensuring that both high-level semantic features and low-level spatial details are preserved. In each decoder block, after concatenation, the feature map is further processed through a series of two Conv layers, each followed by a ReLU activation. It refines the feature representations post-upsampling. The decoder blocks are applied sequentially, beginning with the first block that upsamples the feature maps from Swin Transformer block 4 and combines them with those from block 3. The subsequent blocks progressively upsample the feature maps and merge them with the outputs from Swin Transformer blocks 2 and 1, respectively. The fourth and fifth decoder blocks perform additional upsampling, combining the feature maps with the original convolved input features to ensure the preservation of fine spatial details.

### 3.1 Uncertainty-aware prediction head

The uncertainty-aware prediction head operates on the final feature maps extracted from each input branch, which comprises three parallel sub-heads based on a shared layer design. Each sub-head first applies global average pooling (GAP) to reduce spatial dimensions, followed by a fully connected layer with dropout regularization and a final classification layer. The primary head outputs the predicted class score, while two auxiliary heads predict the mean and variance of the predicted probability distribution corresponding to the input sample, respectively. During training, the empirical mean and standard deviation computed via MC sampling serve as target signals for the mean and variance prediction sub-heads, respectively.

## 4 Two-Phase training procedure

The TP-UAST model is trained on pairs of RGB images and their corresponding segmented color maps, resized to  $256 \times 256$  pixels, using a two-phase curriculum. This approach is designed to decouple point and mean prediction from variance learning for the class of the input sample. In phase I, only the primary classification and mean prediction sub-heads, along with the backbone weights, are trained. The variance sub-head remains frozen in this phase. A composite loss is optimized that balances the binary cross-entropy on both the point prediction and the mean prediction. The Adam optimizer is used with an initial learning rate of  $1 \times 10^{-4}$ , decayed by a polynomial schedule. Training continues until the epoch-averaged loss falls below a threshold value of 0.2, at which point phase I completes. During phase II, all weights of the model except those of the variance sub-head are frozen. In this phase, the  $L_1$  loss is computed between the output  $\sigma^2$  of the variance sub-head and the empirical variance  $\hat{\sigma}_{MC}^2$  computed via MC sampling. This targeted regression teaches the variance head to match the observed spread of predictions.

## 5 Experimental results and discussion

### 5.1 Performance evaluation metrics

The performance of the FCDLe-FOV model is assessed using three key metrics: Average Angular Error (AAE), which measures the angular deviation between estimated and ground truth flow vectors; Average Endpoint Error (AEPE), which calculates the Euclidean distance between the predicted and ground truth flow vectors; and Average Error Normal to the Gradient (AENG), which evaluates the accuracy of flow estimation in directions perpendicular to image gradients, particularly around object boundaries [24]. Moreover, a Structural Similarity Index Measure (SSIM), which quantifies the structural similarity between the optical flow color maps of the reference and noisy images, is employed for robustness evaluation.

The TP-UAST framework is evaluated across four complementary dimensions: Discriminative Performance, via accuracy, precision, recall and  $F_1$ -score to quantify classification correctness and positive-class sensitivity; Calibration, via reliability diagram to ensure predicted probabilities align with empirical frequencies; Predictive Uncertainty Quantification, where histograms of per-sample predictive standard deviations assess uncertainty distribution, scatter

plots of uncertainty versus absolute error to validate error-uncertainty correlation, and class-stratified boxplots to detect systematic biases; and Plausibility Analysis, with empirical histograms of Z-score  $Z = (\ell - \mu)/\sigma$  and plausibility confidences  $C = \exp(-\frac{1}{2}Z^2)$ , which exhibit a  $Z$  and a  $C$  value corresponding to each input's posterior predictive distribution  $\ell \sim \mathcal{N}(\mu, \sigma^2)$ . Here  $\ell$ ,  $\mu$  and  $\sigma^2$  denote the class label point estimate, predicted probability, and uncertainty score, respectively.

## 5.2 Experimental discussion

The experiments conducted in this study have been performed using MATLAB R2023a and WSL Ubuntu 22.04, running on a Windows 11 system equipped with an NVIDIA GeForce RTX 4080 Laptop GPU. The experimental design consists of three main components: the FCDLe-FOV model, used for estimating robust smoke motion features; a GMM technique, for segmenting the smoke ROI; and the TP-UAST deep learning model, trained on fusion datasets. Furthermore, the plausibility of the TP-UAST model is thoroughly validated through quantitative and qualitative analyses.

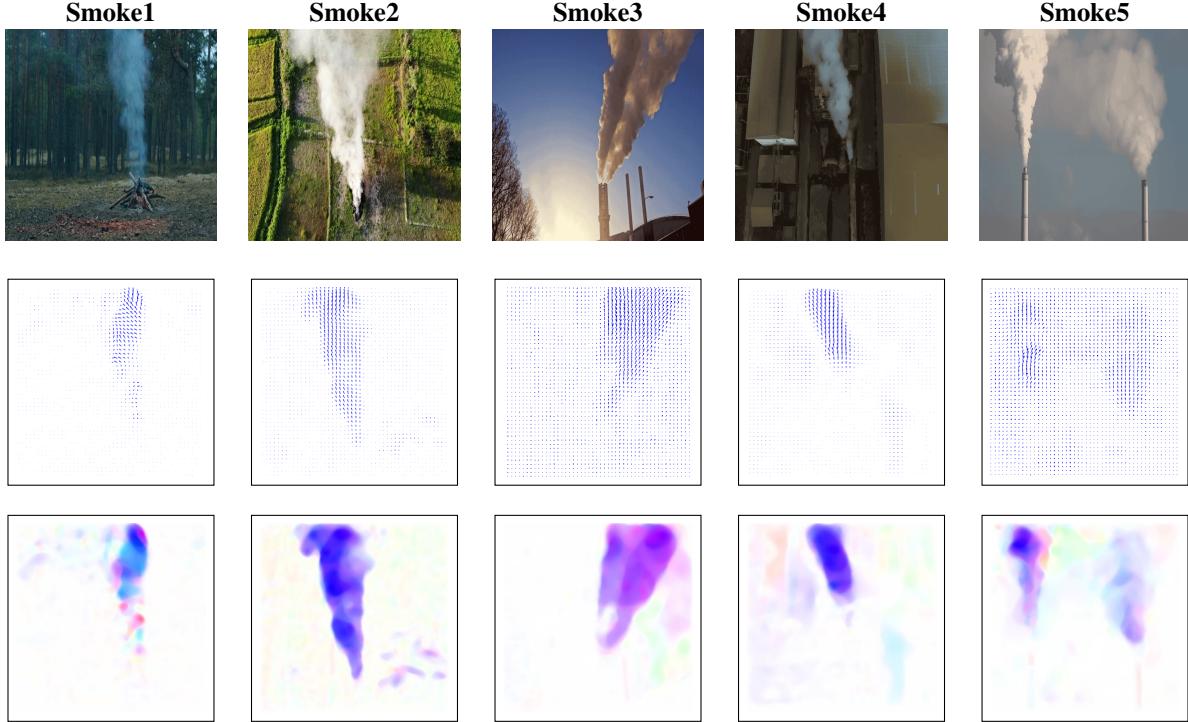


Figure 4: Reference images (first row), vector plots (second row), and optical flow color maps (third row) for smoke videos.

The first experiment presents the optical flow estimation using the FCDLe-FOV model with parameters  $\alpha = 0.5$ ,  $\lambda = 225$ ,  $\theta = 0.001$ , and  $\nu = 1000$ , and 100 iterations. The datasets used in the experiment include scenes like forests, roads, industrial areas, and crowds, captured under both stationary and non-stationary illumination conditions. The experimental results for smoke and non-smoke images are shown in Figs. 4 and 5, with vector plots and color maps illustrated in the second and third rows, respectively. According to the concept introduced by Muller et al. [19], smoke predominantly moves upward, with the blue channel of the color maps showing higher sensitivity to this motion, as shown in Fig. 6. In the non-smoke class, objects often do not exhibit upward motion patterns, as seen in the blue channel of Fig. 6. Moreover, the optical flow maps effectively preserve motion boundaries, supporting the use of dual-phase level set segmentation combined with fractional-order derivatives. Furthermore, dense optical flow fields maintain motion edges, and smoke motion regions are clearly visible, confirming the reliability of the FCDLe-FOV model.

The second experiment consists of two stages to validate the effectiveness of the FCDLe-FOV model. In the initial stage, a quantitative evaluation is conducted using two Middlebury datasets and one Sintel dataset, selected for their inclusion of ground truth data, which is lacking in real-world smoke and non-smoke datasets. The model's accuracy is assessed using AAE, AEPE, and AENG metrics and compared against existing models such as NFVLS [5], FS-FOV [34],

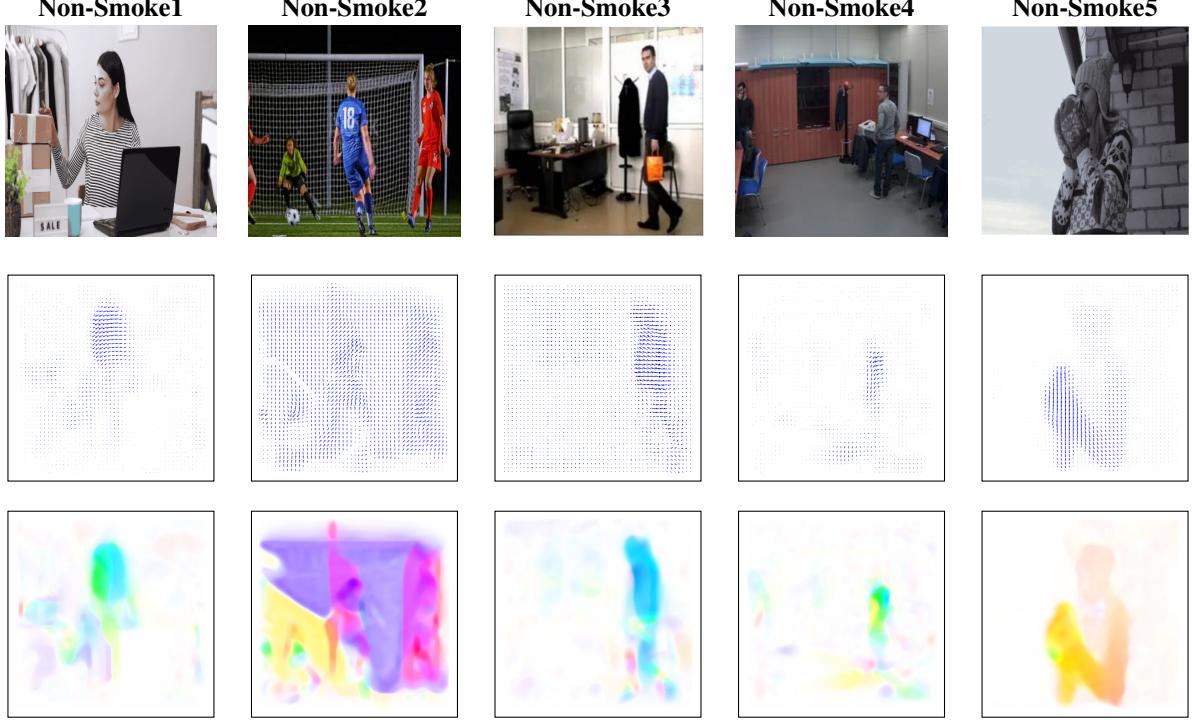


Figure 5: Reference images (first row), vector plots (second row), and optical flow color maps (third row) for non-smoke videos.

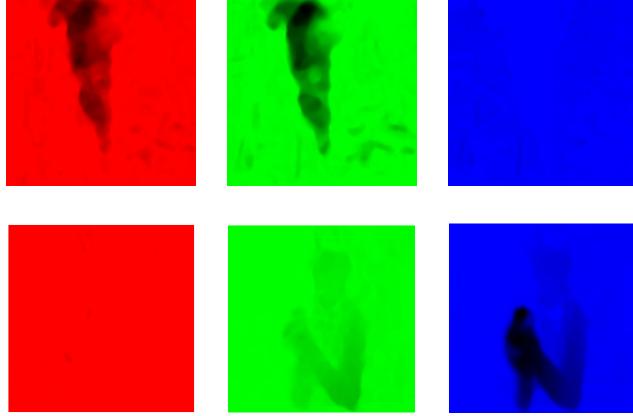


Figure 6: First, second, and third columns correspond to the red, green, and blue channels of the smoke and non-smoke images, respectively.

NLW [22], and HS-NE [21]. As shown in Table 1, FCDLe-FOV model outperforms all compared models across these metrics. The second stage presents a qualitative comparison, as illustrated in Figs. 7 and 8, which show optical flow color maps for two smoke scenes (Smoke2 and Smoke3) and two non-smoke scenes (Non-Smoke3 and Non-Smoke5), respectively. Together, these visual representations of color maps demonstrate that the FCDLe-FOV model generates more accurate results and preserves motion boundaries more effectively than other models. Therefore, the results validate the effectiveness of the FCDLe-FOV model in handling complex smoke and non-smoke scenarios.

The primary objective of the third experiment is to evaluate the robustness of the FCDLe-FOV model in the presence of Salt-and-pepper, Gaussian, and Poisson noise. These noise types are randomly introduced into four image sequences, comprising two smoke scenarios (Smoke1 and Smoke4) and two non-smoke scenarios (Non-Smoke1 and Non-Smoke4). Both Salt-and-pepper noise and Gaussian noise are added with a mean of zero and a standard deviation of 0.01, while

Table 1: Performance comparison of FCDLe-FOV model with other optical flow models.

Datasets	Middlebury1			Middlebury2			Sintel1		
Models	AAE	AEPE	AENG	AAE	AEPE	AENG	AAE	AEPE	AENG
FCDLe-FOV	0.173	0.877	1.875	0.176	2.330	2.461	0.101	0.393	1.079
NFVLs [5]	0.211	1.159	3.397	0.231	3.669	5.363	0.114	0.516	1.783
FS-FOV [34]	0.212	0.997	2.762	0.242	3.407	4.784	0.126	0.542	1.803
NLW [22]	0.378	2.248	5.329	0.605	5.812	8.471	0.274	1.184	3.919
HS-NE [21]	0.782	3.051	6.835	0.451	2.023	3.488	0.451	2.023	3.488

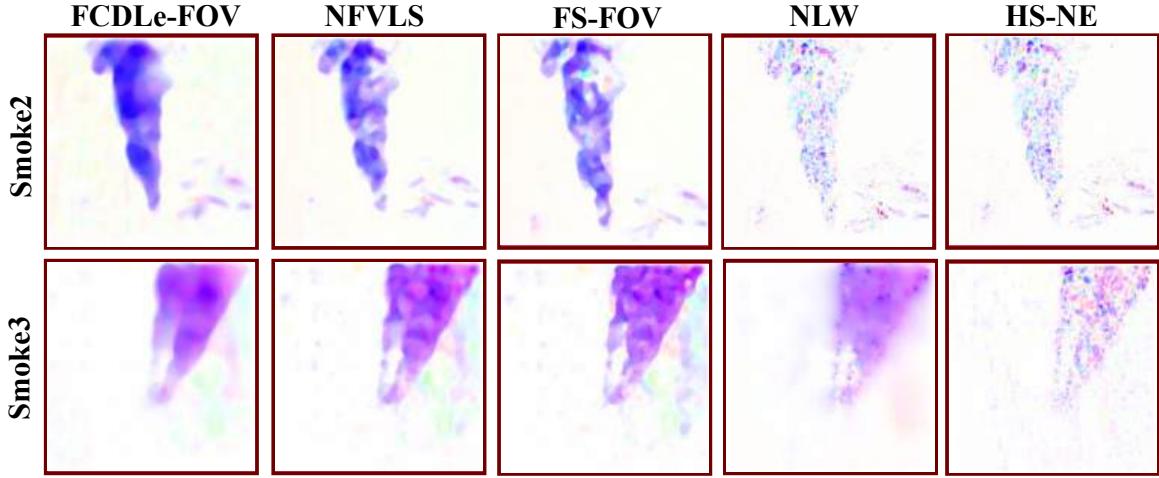


Figure 7: Qualitative comparison of FCDLe-FOV model and SOTA models on smoke datasets.

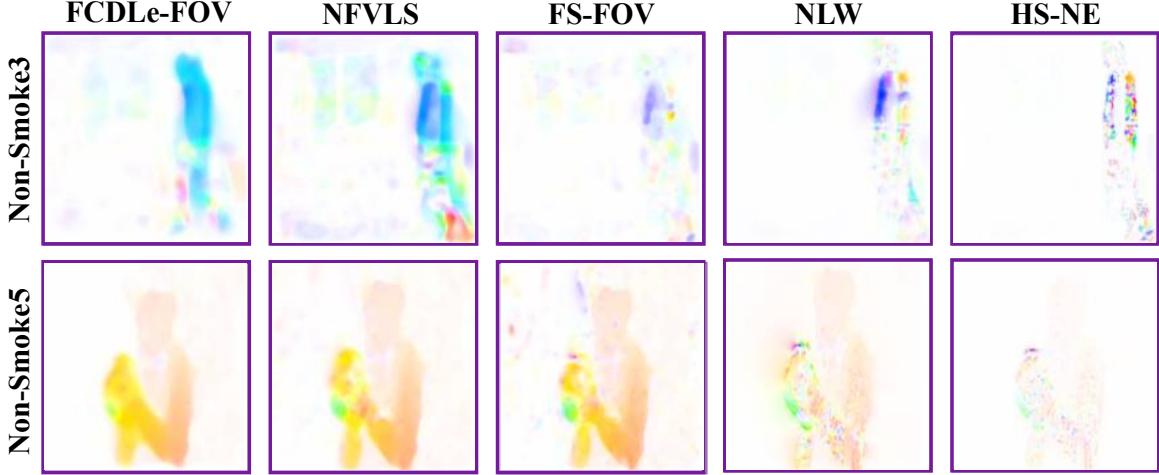


Figure 8: Qualitative comparison of FCDLe-FOV model and SOTA models on non-smoke datasets.

Poisson noise is inherently present in the datasets rather than artificially introduced. The robustness of the FCDLe-FOV model is assessed using the SSIM score, with results in Table 2 demonstrating that it maintains high SSIM values, confirming its effectiveness in handling various noise types while preserving structural details in optical flow color maps.

The fourth experiment aims to effectively highlight smoke-active regions by generating a binary mask that suppresses background features such as vehicles, trees, the sky, and human motion. This binary mask is derived from the estimated color maps using the GMM technique, which effectively distinguishes between smoke and non-smoke regions based

on their statistical distributions. Subsequently, the resulting binary mask is multiplied by the optical flow color maps to produce segmented color maps. In Fig. 9, the first and fourth rows depict the binary masks of the smoke and non-smoke images, respectively, while the second and fifth rows show the segmented optical flow color maps for the corresponding smoke and non-smoke image frames. These are based on the reference images in Figs. 4 and 5 for the smoke and non-smoke images, respectively. The second and third rows of this figure compare the results of the GMM technique with the method proposed by Khan et al. [9] for smoke segmentation, while the fifth and sixth rows provide the corresponding comparison for non-smoke segmentation. This comparison reveals that the proposed binary masks using GMM technique accurately segment the smoke regions by capturing their distinct intensity and texture characteristics. In the case of the non-smoke class, the model often segments background objects separately, as reflected in the binary masks shown in the fourth row of Fig. 9. However, the segmented optical flow color maps indicate that interference from background motion is highly suppressed. This confirms the effectiveness of the proposed GMM technique in isolating the smoke-related motion while minimizing interference from background movements and leading to more precise segmentation of smoke regions.

Optical flow-based motion information captures the dynamic and complex behavior of smoke, while reference images provide complementary spatial information. By fusing segmented optical flow maps with reference images, the system effectively leverages both motion and appearance cues to enhance the detection accuracy of the proposed TP-UAST model.

The fifth experiment analyzes the learning curves of the TP-UAST model in terms of accuracy, loss, and loss in accuracy across two phases, as shown in Fig. 10. The vertical dashed line at epoch 11 marks the point of convergence, after which changes are minimal. In Fig. 10(a), during Phase I, the accuracy curve rises rapidly toward 1.0, indicating the rapid convergence of the TP-UAST model. In phase II, the curve remains stable. Similar convergence behavior is demonstrated by the loss and loss in accuracy curves shown in Fig. 10(b) and Fig. 10(c), respectively. Thus, these curves demonstrate the TP-UAST's efficiency and stability.

The sixth experiment presents Table 3 for a comparison of TP-UAST against leading smoke-detection algorithms. Remarkably, TP-UAST achieves a satisfactory score of 1.00 across accuracy, precision, recall and F1-score, yielding a balanced classifier with zero false positives and false negatives on the validation set. By contrast, the best competing method MFFNet [35] attains 0.98 accuracy and 0.94 precision, while Safarov et al. [17] achieve 0.96 precision and 0.97 recall. Yang et al. [36] and the energy-efficient model [37] both report balanced performance metrics in the range of 0.94–0.95. Moreover, YOLOv2 [38] attains 0.96 accuracy and balances precision against recall, while classic deep networks such as ResNet50 [39] and VGG16 [40] suffer from trade-offs between recall and precision. Other models like TFNet [41], YOLOv8s [42], and HPO-YOLOv5 [43] perform moderately. Methods such as DBN [20] and the lightweight architecture of MobileNet [44] further underperform in either precision or recall. These results underscore the efficacy of the TP-UAST model and the fused optical-flow and appearance representation. Additionally, TP-UAST's discriminatory performance is also illustrated by the confusion matrix as shown in Fig. 11(a).

The seventh experiment is described by the Fig. 11(b), which shows the TP-UAST model's reliability diagram on the test set. Here, the mean of the predicted probabilities, denoted by blue markers, are plotted against the empirical fraction of positives. The number of bins taken is ten. The calibration curve lies essentially very close to the ideal diagonal (green dashed line) over the entire  $[0, 1]$  range, yielding an ECE approximately equal to 0. This near-perfect alignment demonstrates that TP-UAST's predicted probability estimates faithfully reflect true outcome frequencies. Such strong calibration is critical, since it allows end-users to set actionable probability thresholds with known reliability, which is 0.5 in the proposed study.

In the eighth experiment, Fig. 12(a) presents the distribution of per-sample predictive uncertainty  $\sigma_p$  in probability space. The histogram is sharply peaked at low values  $< 0.005$ , with a long tail extending to  $\approx 0.06$ , indicating that most predictions are made with high confidence while a minority of cases exhibit elevated uncertainty. Fig. 12(b) overlays each sample's uncertainty against its absolute error  $|PL - TL|$ . A clear positive trend emerges: points with  $\sigma_p > 0.02$  are disproportionately associated with larger errors, whereas low-uncertainty predictions cluster near zero error. This correlation validates  $\sigma_p$  as an effective proxy for identifying potentially misclassified or ambiguous inputs. Fig. 12(c) shows class-stratified boxplots of  $\sigma_p$  for the "Smoke" and "No Smoke" classes. Both classes exhibit a median uncertainty around 0.002, but the "Smoke" distribution has a heavier upper quartile and more extreme outliers, reflecting

Table 2: SSIM scores of the FCDLe-FOV model on smoke and non-smoke images under different noise conditions.

Datasets	Smoke Datasets		Non-Smoke Datasets	
Noises	Smoke1	Smoke4	Non-Smoke1	Non-Smoke4
Gaussian noise	0.87	0.91	0.97	0.95
Poisson noise	0.95	0.97	0.98	0.98
Salt-and-pepper noise	0.97	10.98	0.99	0.99

greater aleatoric variability when detecting smoke plumes. Together, these analyses demonstrate that TP-UAST's uncertainty estimates are well-calibrated, informative of error, and sensitive to class-dependent difficulty.

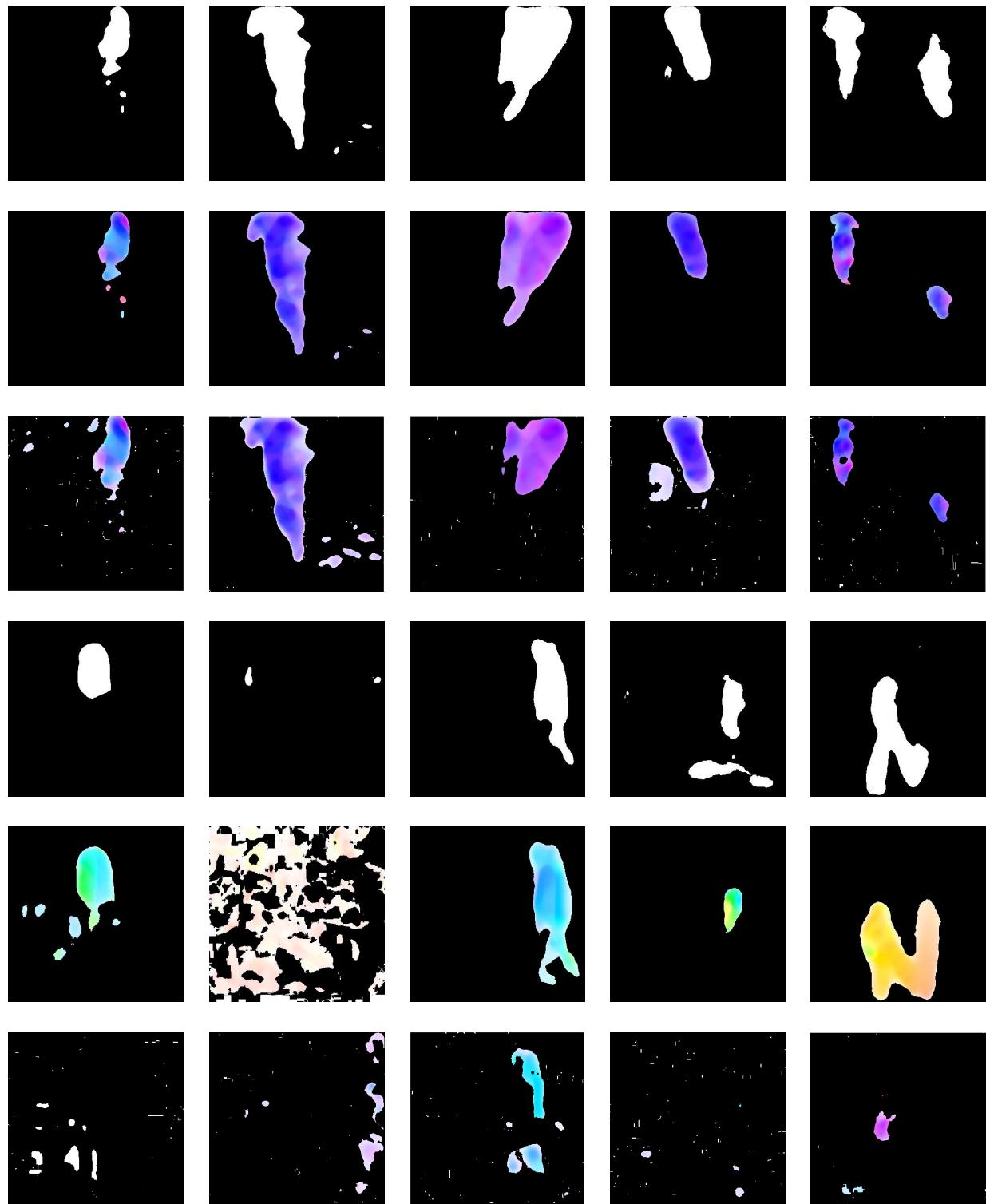


Figure 9: Binary mask (first and fourth row), segmented color map (second and fifth row), and comparison of GMM technique with existing method [9] (third and sixth row) for smoke and non-smoke images.

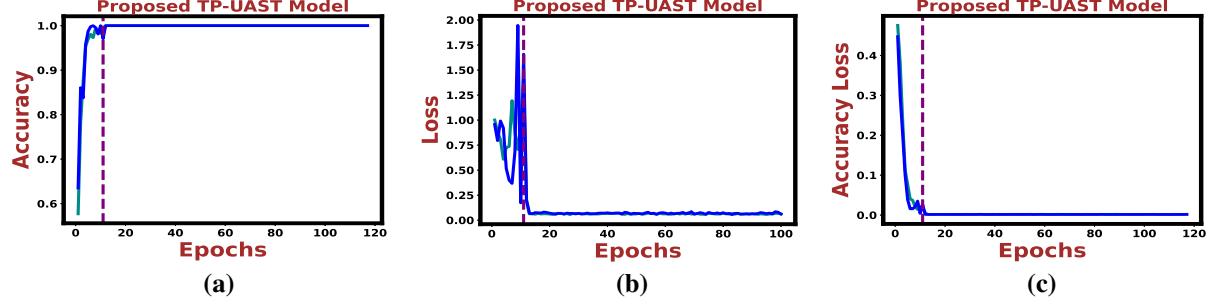


Figure 10: Learning curves in terms of (a) accuracy, (b) loss, and (c) loss in accuracy.

Table 3: A comparison of proposed TP-UAST model against SOTA models.

Algorithm	Accuracy	Precision	Recall	F1-Score
<b>TP-UAST model</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>
HPO-YOLOv5 [43]	—	0.93	0.86	0.89
MFFNet [35]	0.98	0.94	—	—
TFNet [41]	—	0.82	0.75	0.78
YOLOv8s [42]	—	0.91	0.85	0.88
Safarov et al. [17]	—	0.96	0.97	—
Yang et al. [36]	—	0.94	0.93	—
Energy efficient model [37]	0.95	0.95	0.95	0.95
DLBN [26]	0.89	0.98	0.81	0.89
YOLOv2 [38]	0.96	0.97	0.95	0.97
DBN [20]	0.94	0.93	0.96	0.94
MobileNet model [44]	0.88	0.81	0.99	0.89
ResNet50 model [39]	0.89	0.81	1.00	0.89
ResNet101 model [39]	0.95	0.95	0.93	0.94
GoogleNet model [45]	0.89	0.85	0.97	0.90
VGG16 model [40]	0.81	0.72	1.00	0.84

The ninth experiment is exhibited in the Fig. 13(a), which plots the empirical distribution of per-sample Z-scores  $Z = (\ell - \mu)/\sigma_p$ , obtained from the posterior predictive draws  $\ell \sim \mathcal{N}(\mu, \sigma^2)$ . The histogram is tightly centered around  $Z = 0$  with approximately 68% of values in  $|Z| \leq 1$  and 95% in  $|Z| \leq 2$ . Tail events beyond  $|Z| > 2$  remain scarce, indicating that the model's predicted  $\sigma_p$  accurately captures the scatter of its own logits. Figure 13(b) shows the corresponding distribution of plausibility confidences  $C = \exp(-\frac{1}{2}Z^2)$ , peaked near  $C = 1$  with a smooth decay toward 0.2. Over 80% of samples achieve  $C > 0.8$ , confirming that most predictions lie well within their predicted uncertainty bounds, while the low-confidence tail correctly flags rare, less-plausible draws. Together, these results validate that TP-UAST's two-phase uncertainty learning yields self-consistent posterior predictive distributions, enabling reliable per-sample confidence estimates. Accordingly,  $C$  is partitioned into four operational tiers: High, Moderate, Low, and Very Low Confidence, facilitating threshold-based decision-making.

The final experiment illustrates test-case outputs from the proposed TP-UAST model, as shown in Fig. 14, reporting the predicted probability (PP) and its associated probability confidence (PC) score for each image in the smoke and non-smoke datasets. Across eight smoke examples, the network reliably identifies smoke. For instance, Smoke1, which contains a small campfire plume, yields  $PP=0.9993$  with  $PC=0.68$ , indicating high confidence ( $C_H$ ). Likewise, the dense emissions in Smoke2, Smoke3, and Smoke6-Smoke8 produce equally high confidence scores. By contrast, the more diffuse, cloud-like plumes in Smoke5 and Smoke9 return  $PC = 0.10$  (moderate confidence ( $C_M$ )) and  $PC = 0.04$  (low confidence ( $C_L$ )), respectively, underscoring the difficulty of discriminating thin, wispy smoke against complex terrain or sky-like backgrounds.

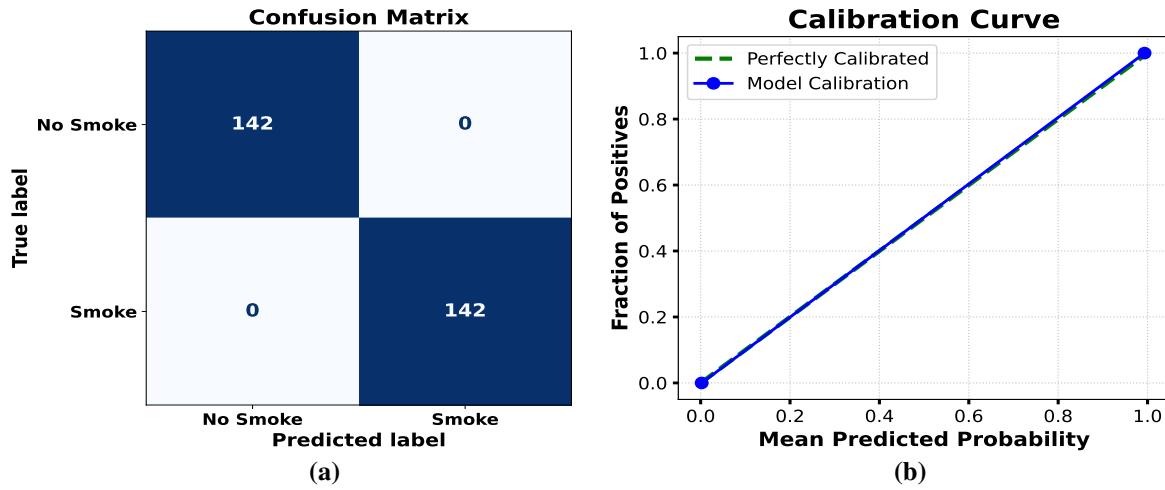


Figure 11: Demonstrating (a) Confusion matrix and (b) reliability diagram for TP-UAST model.

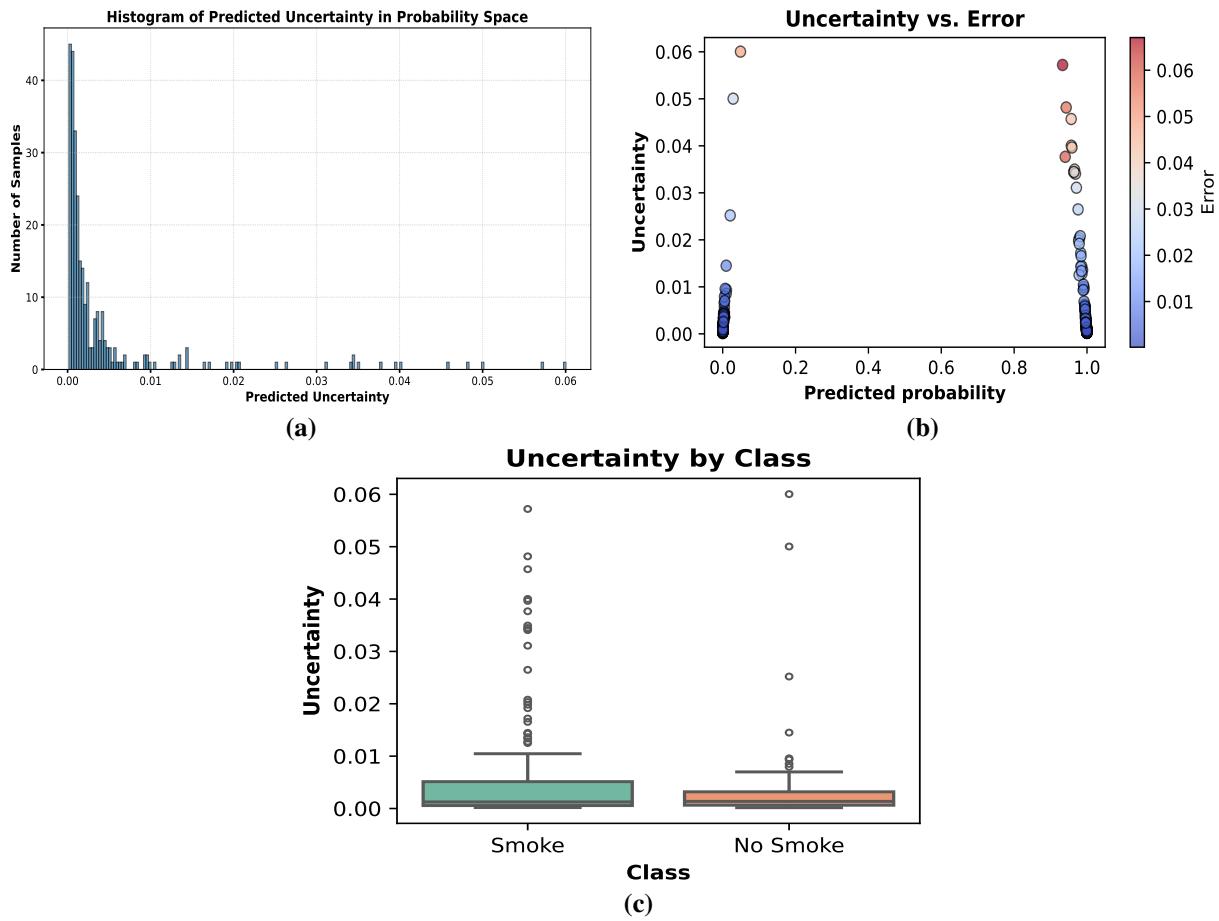


Figure 12: (a) Histogram of predicted uncertainty in probability space, (b) uncertainty vs. error plot, and (c) uncertainty by class for TP-UAST model.

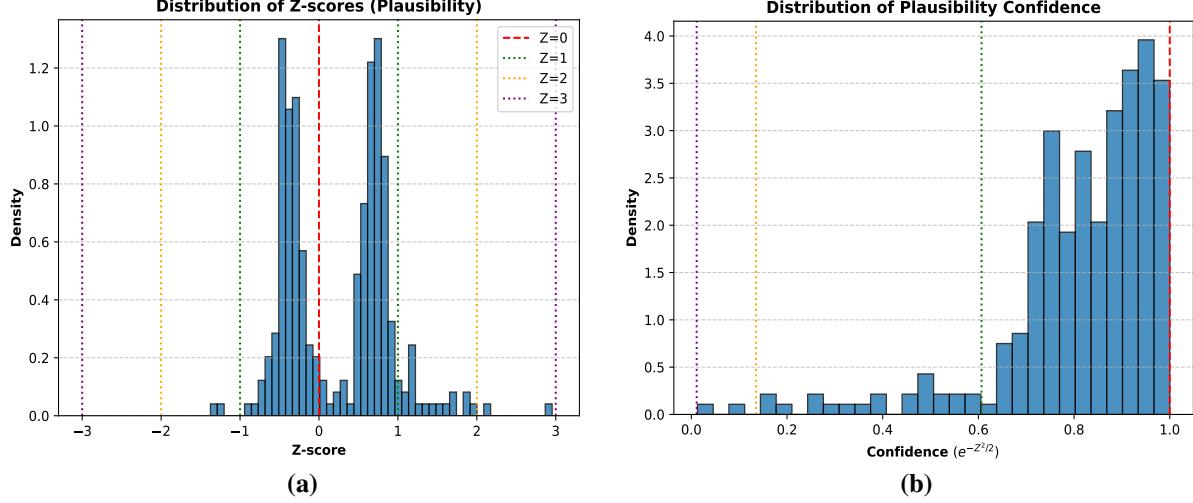


Figure 13: (a) Distribution of Z-scores (plausibility) and distribution of plausibility confidence for TP-UAST model.

In the non-smoke dataset, all eight samples yield significantly low PP values, confirming strong negative discrimination. For instance, the indoor office scene in Non-Smoke3, the exhaled vapor sample of Non-Smoke5, the firefighter ascending a ladder in Non-Smoke6, the distant hillside view of Non-Smoke7, and the outdoor sports scene in Non-Smoke11 demonstrate high confidence associated PP values. While, challenging low-light scenarios, such as a wooded cabin in a dark forest setting of Non-Smoke8 and overlapping foliage edges in Non-Smoke9, produce  $PP = 0.0006$ ,  $PC = 0.09$  ( $C_M$ ) and  $PP = 0.0030$ ,  $PC = 0.05$  ( $C_L$ ). Cloud formations in the top-left and top-right regions of Non-Smoke10 yield  $PP = 0.0002$ ,  $PC = 0.01$ , returning a very low confidence ( $C_{VL}$ ) in the corresponding PP values. These results demonstrate that by jointly predicting  $PP$  and  $PC$ , the model offers transparent insight into each decision, thereby ensuring reliable performance across diverse real-world environments.

## 6 Conclusion

In this work, a unified framework is introduced for smoke detection that fuses motion and appearance cues via an optical-flow-driven segmentation pipeline and a novel TP-UAST model. The proposed FCDLe-FOV model produces high-fidelity optical flow maps under variable illumination while preserving motion boundaries, and a Gaussian Mixture Model generates precise smoke region masks for training. The TP-UAST architecture leverages a dual-branch Swin Transformer backbone augmented with dedicated heads for point prediction, mean estimation, and variance regression. A two-phase curriculum decouples classification accuracy from uncertainty learning, yielding a classifier that achieves excellent detection metrics on the test set and produces well-calibrated confidence scores. Extensive quantitative and qualitative experiments demonstrate the robustness, reliability, and interpretability of the proposed approach across diverse scenarios. By explicitly modeling both aleatoric and epistemic uncertainties, TP-UAST not only identifies smoke with high precision and recall, but also quantifies its own confidence, enabling risk-aware decision thresholds in safety-critical applications. Future work will focus on real-time deployment of TP-UAST on embedded platforms with limited compute resources. Additionally, we plan to extend the framework to multi-class hazard detection (e.g., fire, steam, dust) to broaden its applicability in industrial and environmental monitoring.

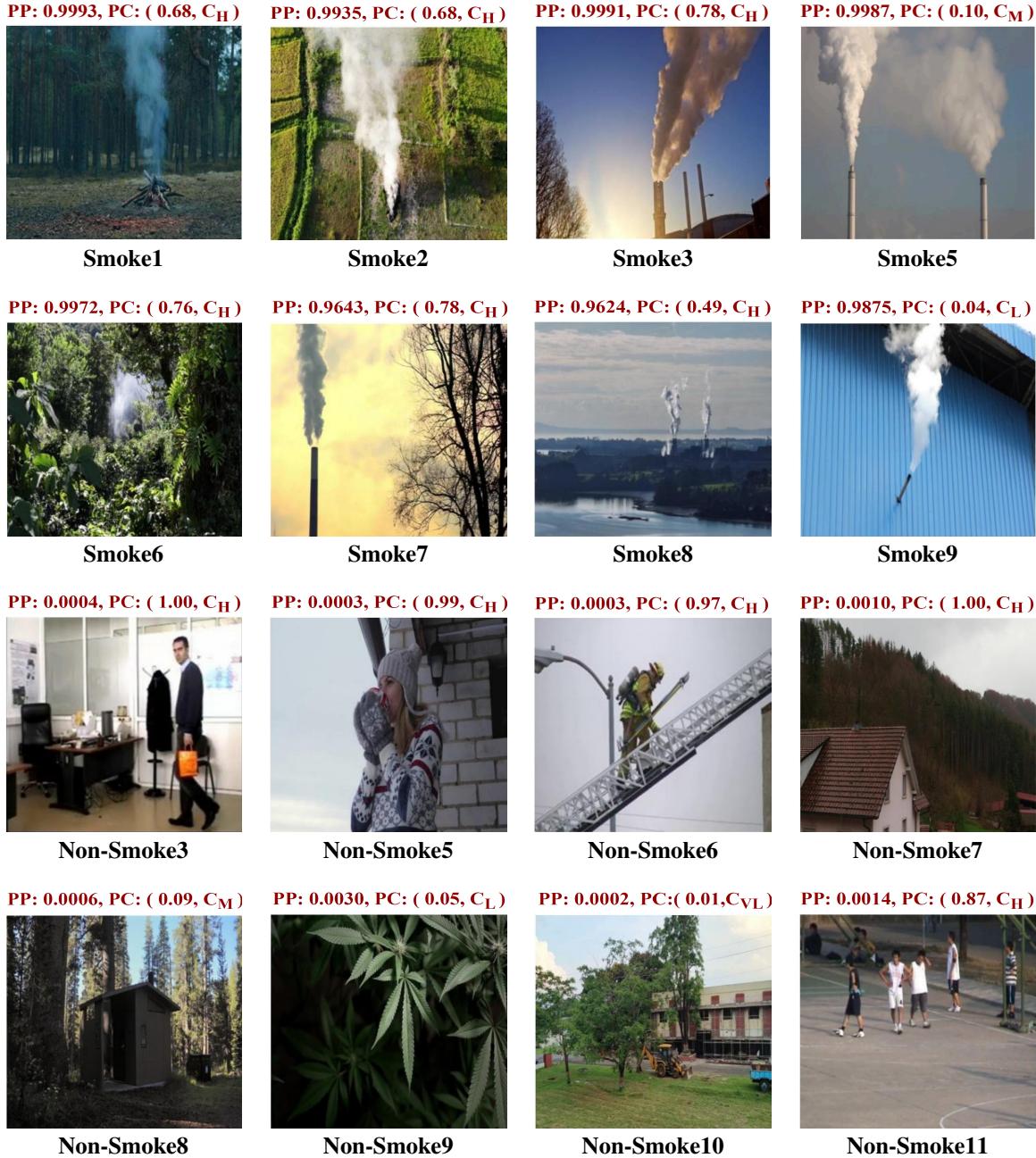


Figure 14: Results demonstrating the predicted probability ( $PP$ ) and associated probability confidence ( $PC$ ) scores corresponding to smoke samples (first and second row) and non-smoke samples (third and fourth row) from the TP-UAST model.

## References

- [1] Fengmei Cui. Deployment and integration of smart sensors with iot devices detecting fire disasters in huge forest environment. *Computer Communications*, 150:818–827, 2020.
- [2] Pingshan Liu, Pingchuan Xiang, and Dianjie Lu. A new multi-sensor fire detection method based on lstm networks with environmental information fusion. *Neural Computing and Applications*, 35(36):25275–25289, 2023.
- [3] Peixian Jin, Pingle Cheng, Xiaodong Liu, and Ying Huang. From smoke to fire: A forest fire early warning and risk assessment model fusing multimodal data. *Engineering Applications of Artificial Intelligence*, 152:110848, 2025.
- [4] Somayeh Gh Bardeji, Isabel N Figueiredo, and Ercília Sousa. Optical flow with fractional order regularization: variational model and solution method. *Applied Numerical Mathematics*, 114:188–200, 2017.
- [5] Khan Muzammil and Kumar Pushpendra. A level set based fractional order variational model for motion estimation in application oriented spectrum. *Expert Systems with Applications*, 219:119628, 2023.
- [6] Antonin Chambolle. An algorithm for total variation minimization and applications. *Journal of Mathematical Imaging and Vision*, 20:89–97, 2004.
- [7] R Farnoush and PAK B ZAR. Image segmentation using gaussian mixture model. 2008.
- [8] Prerak Mody, Nicolas F Chaves-de Plaza, Chinmay Rao, Eleftheria Astrenidou, Mischa de Ridder, Nienke Hoekstra, Klaus Hildebrandt, and Marius Staring. Improving uncertainty-error correspondence in deep bayesian medical image segmentation. *arXiv preprint arXiv:2409.03470*, 2024.
- [9] Muzammil Khan, Pushpendra Kumar, and Nitish Kumar Mahala. Cnn-based fire prediction using fractional order optical flow and smoke features. In *Applications of Optimization and Machine Learning in Image Processing and IoT*, pages 156–180. Chapman and Hall/CRC, 2023.
- [10] Yinuo Huo, Qixing Zhang, Yang Jia, Dongcai Liu, Jinfu Guan, Gaohua Lin, and Yongming Zhang. A deep separable convolutional neural network for multiscale image-based smoke detection. *Fire Technology*, pages 1–24, 2022.
- [11] Jerome Vicente and Philippe Guillemant. An image processing technique for automatically detecting forest fire. *International Journal of Thermal Sciences*, 41(12):1113–1120, 2002.
- [12] Ke Gu, Zhifang Xia, Junfei Qiao, and Weisi Lin. Deep dual-channel neural network for image-based smoke detection. *IEEE Transactions on Multimedia*, 22(2):311–323, 2019.
- [13] Pu Li and Wangda Zhao. Image fire detection algorithms based on convolutional neural networks. *Case Studies in Thermal Engineering*, 19:100625, 2020.
- [14] Gaohua Lin, Yongming Zhang, Gao Xu, and Qixing Zhang. Smoke detection on video sequences using 3d convolutional neural networks. *Fire Technology*, 55:1827–1847, 2019.
- [15] Guangtao Cheng, Yancong Zhou, Shan Gao, Yingyu Li, and Hao Yu. Convolution-enhanced vision transformer network for smoke recognition. *Fire Technology*, 59(2):925–948, 2023.
- [16] Huajun Song and Yulin Chen. Video smoke detection method based on cell root–branch structure. *Signal, Image and Video Processing*, pages 1–9, 2024.
- [17] Furkat Safarov, Shakhnoza Muksimova, Misirov Kamoliddin, and Young Im Cho. Fire and smoke detection in complex environments. *Fire*, 7(11):389, 2024.
- [18] Konstantina Mardani, Nicholas Vretos, and Petros Daras. Transformer-based fire detection in videos. *Sensors*, 23(6):3035, 2023.
- [19] Martin Mueller, Peter Karasev, Ivan Kolesov, and Allen Tannenbaum. Optical flow estimation for flame detection in videos. *IEEE Transactions on Image Processing*, 22(7):2786–2797, 2013.
- [20] Arun Singh Pundir and Balasubramanian Raman. Deep belief network for smoke detection. *Fire technology*, 53:1943–1960, 2017.
- [21] Pushpendra Kumar and Sanjeev Kumar. A modified variational functional for estimating dense and discontinuity preserving optical flow in various spectrum. *AEU-International Journal of Electronics and Communications*, 70(3):289–300, 2016.
- [22] Zhenghua Huang and Aimin Pan. Non-local weighted regularization for optical flow estimation. *Optik*, 208:164069, 2020.

- [23] Arnisha Khondaker, Arman Khandaker, and Jia Uddin. Computer vision-based early fire detection using enhanced chromatic segmentation and optical flow analysis technique. *International Arab Journal of Information Technology*, 17(6):947–953, 2020.
- [24] Muzammil Khan, Nitish Kumar Mahala, and Pushpendra Kumar. Caputo derivative based nonlinear fractional order variational model for motion estimation in various application oriented spectrum. *Sādhanā*, 49(1):1–28, 2024.
- [25] Yu Chunyu, Fang Jun, Wang Jinjun, and Zhang Yongming. Video fire smoke detection using motion and color features. *Fire technology*, 46:651–663, 2010.
- [26] Yuanlu Wu, Minghao Chen, Yan Wo, and Guoqiang Han. Video smoke detection base on dense optical flow and convolutional neural network. *Multimedia Tools and Applications*, 80:35887–35901, 2021.
- [27] Kazutaka Kikuta, Ken T Murata, and Yuki Murakami. A daytime smoke detection method based on variances of optical flow and characteristics of hsv color on footage from outdoor camera in urban city. *Fire Technology*, 60(3):1427–1452, 2024.
- [28] Luminita A Vese and Tony F Chan. A multiphase level set framework for image segmentation using the mumford and shah model. *International journal of computer vision*, 50:271–293, 2002.
- [29] Arun Govind Neelan. Von neumann stability analysis for multi-level multi-step methods. *arXiv preprint arXiv:2310.08274*, 2023.
- [30] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.
- [31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *corr abs/1706.03762*, 2017.
- [32] Haitao Dong, Chengjun Chen, Jinlei Wang, Feixiang Shen, and Yong Pang. Vit-saps: Detail-aware transformer for mechanical assembly semantic segmentation. *IEEE Access*, 11:41467–41479, 2023.
- [33] Zobeir Raisi, Mohamed A Naiel, Paul Fieguth, Steven Wardell, and John Zelek. 2d positional embedding-based transformer for scene text recognition. *Journal of Computational Vision and Imaging Systems*, 6(1):1–4, 2020.
- [34] Jin Lu, Hua Yang, Qinghu Zhang, and Zhouping Yin. A field-segmentation-based variational optical flow method for piv measurements of nonuniform flows. *Experiments in Fluids*, 60:1–17, 2019.
- [35] Hongying Liu, Fuquan Zhang, Yiqing Xu, Junling Wang, Hong Lu, Wei Wei, and Jun Zhu. Tfnet: Transformer-based multi-scale feature fusion forest fire image detection network. *Fire*, 8(2):59, 2025.
- [36] Huanyu Yang, Jun Wang, and Jiacun Wang. Efficient detection of forest fire smoke in uav aerial imagery based on an improved yolov5 model and transfer learning. *Remote Sensing*, 15(23):5527, 2023.
- [37] Jefferson Silva Almeida, Chenxi Huang, Fabrício Gonzalez Nogueira, Surbhi Bhatia, and Victor Hugo C de Albuquerque. Edgefiresmoke: A novel lightweight cnn model for real-time video fire–smoke detection. *IEEE Transactions on Industrial Informatics*, 18(11):7889–7898, 2022.
- [38] Sergio Saponara, Abdussalam Elhanashi, and Alessio Gagliardi. Real-time video fire/smoke detection based on cnn in antifire surveillance systems. *Journal of Real-Time Image Processing*, 18:889–900, 2021.
- [39] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [40] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [41] Yupeng Wang, Yongli Wang, Zaki Ahmad Khan, Anqi Huang, and Jianghui Sang. Multi-level feature fusion networks for smoke recognition in remote sensing imagery. *Neural Networks*, 184:107112, 2025.
- [42] Derui Kong, Yinfeng Li, and Manzhen Duan. Fire and smoke real-time detection algorithm for coal mines based on improved yolov8s. *Plos one*, 19(4):e0300502, 2024.
- [43] Md Shafak Shahriar Sozol, M Rubaiyat Hossain Mondal, and Achmad Husni Thamrin. Indoor fire and smoke detection based on optimized yolov5. *PLoS One*, 20(4):e0322052, 2025.
- [44] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.
- [45] Pedro Ballester and Ricardo Araujo. On the performance of googlenet and alexnet applied to sketches. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.