



Estimating flame volume from fire tests using machine learning

Samuel Tistelgren Holappa^a, Gustav Schelin^a, Johan Anderson^{b,*} 

^a University of Borås, S-501 90, Borås, Sweden

^b RISE, Brinellgatan 4, 504 62, Borås, Sweden

ARTICLE INFO

Keywords:

Machine learning
Flame volume
Semantic segmentation
Fire tests
Computer vision

ABSTRACT

Flame volume is a fundamental descriptor of fire dynamics and is closely related to heat release rate and fire spread potential. Traditional methods for assessing fire size are indirect, relying on measurements of mass loss rate, oxygen consumption calorimetry or simplified geometrical assumptions. With recent advances in computer vision and artificial intelligence (AI), new opportunities arise for quantifying flame geometry directly from visual data. In this study, the semantic segmentation models U-Net, ENet, Fast-SCNN, and two custom DeepLab variants, were trained on datasets from small-scale calibration tests as well as full-scale façade fire experiments. The models were evaluated in terms of segmentation accuracy, inference speed, and generalization ability. Results show that DeepLab-B0 achieved the highest segmentation accuracy, while Fast-SCNN provided the best trade-off between speed and precision. The findings suggest that AI-based segmentation combined with tailored geometric modeling can provide reliable estimates of flame volume, with clear trade-offs depending on whether accuracy or speed is prioritized.

1. Introduction

Artificial Intelligence (AI) has evolved rapidly since its inception in the 1950s [1], when early systems focused on problems that could be fully described by formal rules, such as board games like chess. These early approaches were limited to domains where clear logic could be encoded. Many real-world tasks, however, such as medical diagnosis, cannot be captured easily by explicit rules. The emergence of machine learning (ML), a subfield of AI, addressed this limitation by enabling algorithms to learn patterns directly from data and produce probabilistic predictions. ML models are trained on labeled datasets to infer relationships between input parameters and outcomes, allowing them to handle complex, nonlinear phenomena that traditional rule-based methods struggle to model [2].

Fires exemplify this complexity, where the underlying physical processes span a wide range of temporal and spatial scales, some of them remain only partially understood. Consequently, semi-empirical models are common, and there is considerable potential for ML to contribute to this field. Properly trained and validated ML models can act as surrogate models for complex experiments or simulations, providing fast predictions within the bounds of their training data [3]. While extrapolation beyond the training domain remains challenging, combining

specialized sub-models can yield accurate predictions for complex fire scenarios. Artificial neural networks often outperform traditional regression when nonlinear relationships dominate.

AI and, more specifically, ML have become increasingly relevant tools within fire science. Historically, fire phenomena have been studied using experimental methods and computational fluid dynamics (CFD) simulations, see e.g. Ref. [4]. While these approaches provide valuable insights, they are resource-intensive and often lack the flexibility needed for real-time applications. By contrast, ML enables the development of surrogate models that can be learned directly from empirical or simulated data, offering computationally efficient and adaptable solutions. Application of ML in fire science span both engineering practice and fundamental research. In fire safety engineering, ML can automate assessments when large datasets or image collections are available. In research, ML supports deeper analysis of the governing mechanisms in combustion, enabling the development of improved tests and assessment methods. Notable applications include laminar flame characterization [5], forest fire prediction [6,7] and fire dynamics modeling [3,8–10]. While these studies highlight promising avenues, AI in fire science remains an emerging area, and expert domain knowledge is essential to ensure that ML models are applied appropriately.

Despite this progress, relatively little work has focused on the

This article is part of a special issue entitled: FISJ_IAFSS 2026 published in Fire Safety Journal.

* Corresponding author.

E-mail addresses: samuel.holappa@gmail.com (S.T. Holappa), gustav.schelin10@gmail.com (G. Schelin), johan.anderson@ri.se (J. Anderson).

<https://doi.org/10.1016/j.firesaf.2026.104732>

Received 23 September 2025; Received in revised form 8 March 2026; Accepted 13 March 2026

Available online 14 March 2026

0379-7112/© 2026 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

quantification of flame geometry from visual data. Flame height, area, and especially volume, are parameters of direct importance for understanding fire behavior [11–13]. Flame volume, in particular, provides a proxy for energy release and spread potential, but is rarely measured in fire tests due to methodological challenges. Traditionally, assessments rely on indirect measures or simplified geometrical assumptions. However, advances in deep learning and semantic segmentation now make it possible to analyze flames at the pixel level in video frames, enabling systematic and scalable quantification.

Video-based fire analysis has been an active research area for more than two decades. Early work primarily focused on flame and smoke detection using handcrafted features and rule-based classifiers, as comprehensively reviewed in Ref. [14]. More recently, deep-learning approaches have significantly improved robustness and real-time capability, with convolutional neural networks increasingly used for fire segmentation and detection tasks [15,16]. While these studies demonstrate strong performance for fire recognition and localization, they typically stop at binary detection or pixel-wise classification and do not address the extraction of physically meaningful fire descriptors such as flame volume.

This study investigates whether ML-based semantic segmentation [11], combined with geometric reconstruction methods, can provide reliable estimates of flame geometry [17]. Building on datasets from façade fire tests, several segmentation models are trained and evaluated. Their outputs are then used as input to different geometric volume approximation techniques using the Shape from Silhouette method. The objective is to assess both accuracy and feasibility, weighing the trade-offs between precision and computational efficiency. This methodology should be seen as a complementary tool to traditional visual methods sometimes used.

The novelty of the present work does not lie in proposing a new semantic segmentation architecture per se, but in demonstrating how modern deep-learning-based image segmentation can be systematically adapted, validated, and applied to full-scale fire experiments conducted under standardized façade test conditions. In contrast to most prior work, which focuses on fire detection or classification, this study targets quantitative geometric descriptors of the flame, with particular emphasis on flame volume. By training and benchmarking multiple segmentation models on video data from controlled façade fire tests, the study provides insight into the trade-offs between segmentation accuracy, inference speed, and robustness when applied to fire testing environments rather than generic image datasets. The contribution should be viewed as a methodological demonstration and benchmark rather than a finalized operational tool. The results are intended to inform both fire researchers and testing laboratories about the feasibility and limitations of extracting flame geometry from standard video recordings, and to serve as a reference point for future development of automated post-processing tools for fire tests.

2. Methodology

The primary dataset used in this study was obtained from façade fire tests conducted within the European project 'Finalization of the European approach to assess the fire performance of façades', [18]. The aim of the project was to develop a methodology to assess the fire performance of façades for medium and large fire exposure. The façade fire experiments were conducted as part of the experimental Round Robin within the European Commission-funded project. The testing programme comprised both medium-scale exposures (similar to the DIN 4102-20 façade fire test method [19]) and large-scale exposures (similar to the BS 8414 façade fire test [20]), based on a standardized combustion chamber with a wood crib fuel source and a defined secondary opening. Tests were performed at multiple laboratories to evaluate repeatability and inter-laboratory variability. While camera models and lighting conditions varied between laboratories, camera positioning and viewing geometry were kept comparable to ensure consistency. In

addition to conventional instrumentation, all tests were documented using RGB video recordings from fixed camera positions. The cameras were primarily intended for visual documentation and qualitative assessment; however, their stable positioning and repeatable test geometry also make them suitable for post-test quantitative analysis. Across the participating laboratories, camera (GoPro Hero 10 in most cases which have CMOS sensors) resolutions ranged from 1920×1080 to 2560×1440 pixels, with frame rates between 25 and 30 frames per second. However, all captured video streams were scaled to 1920×1080 for consistency in comparisons. The recorded videos capture the temporal evolution of the flame plume along the façade under controlled and well-characterized exposure conditions. In the present work, selected video sequences from these façade tests were used to investigate the feasibility of extracting flame geometry using image-based methods. Preprocessing included lens distortion correction (Gyroflow [21]), frame extraction, and data augmentation techniques such as rotations, scaling, and brightness variations. The standardized nature of the test configuration, combined with the availability of repeat tests across laboratories, provides a unique dataset for evaluating the robustness of automated flame segmentation and subsequent geometric reconstruction. While the façade experiments were not originally designed for image-based flame measurement, they offer a realistic and relevant environment for assessing the potential of such techniques in fire testing practice.

Flame regions were manually annotated on selected frames using pixel-level masks to create the ground-truth dataset. To improve model robustness and account for inter-laboratory variations in lighting and exposure, data augmentation techniques were applied during training. These included random rotations ($\pm 5^\circ$), horizontal flipping, scaling, brightness and contrast adjustments, and the addition of mild Gaussian noise. No temporal smoothing or tracking was applied during training, and each frame was treated independently. Annotated datasets were created manually using LabelMe, where flame regions were outlined pixel by pixel. Pixel counts are used as an intermediate representation of flame size because they form the direct output of segmentation models and provide a transparent link between model performance metrics and subsequent geometric reconstruction. Conversion to physical dimensions is performed through camera calibration, but reporting pixel-level agreement enables comparison across models independently of geometric assumptions.

Five segmentation models were trained and evaluated: U-Net [22], Enet [23], DeepLab-B0 [24,25], DeepLab-B3 [24,25], and Fast-SCNN [26] using PyTorch [27] and trained using identical datasets and evaluation metrics to ensure a fair comparison. The training and validation datasets were fixed throughout training, following a commonly used 80/20 training-validation split in machine learning model development [27]; no independent test set was used due to dataset size where the validation set was used independently for model evaluation. Training was performed with a loss function combining Dice loss and Binary cross-entropy. Model performance was evaluated using data from confusion matrices, Dice coefficients and Intersection-over-Union (IoU) scores. For flame geometry estimation, pixel-based flame masks were converted to height and area using camera calibration, while volume was estimated either through Shape from Silhouette or geometric approximations treating the flame as cones or paraboloids. Although, in this particular study only the Shape from Silhouette method is used. Segmentation performance, accuracy of volume estimation, and computational costs were all systematically recorded.

Due to limited space, details of the segmentation model and results will be shown for U-Net however brief descriptions of models and results are provided for all models.

2.1. Description of models

U-Net is a convolutional neural network architecture originally developed for biomedical image segmentation. It features a symmetric

encoder–decoder structure in which each down-sampling step in the encoder is mirrored by an up-sampling step in the decoder. The encoder employed two successive 3×3 convolutional layers with ReLU activation at each level, followed by max-pooling. The number of feature channels started at 64 and doubled at each down-sampling step. The decoder mirrored the encoder using transposed convolutions and skip connections to recover spatial detail. Skip connections link corresponding layers, allowing recovery of fine spatial details lost in deep networks. This design enables U-Net to generate high-resolution segmentation maps from limited annotated data and makes it highly effective when detailed object boundaries are essential, and data is scarce or expensive to collect, see Fig. 1.

ENet (Efficient Neural Network) was designed for fast and resource-efficient semantic segmentation, prioritizing real-time inference with minimal computational cost. The network achieves this by aggressively down-sampling early in the architecture, significantly reducing complexity compared to heavier models such as U-Net or DeepLab. The network consisted of an initial block followed by bottleneck modules with asymmetric and dilated convolutions, resulting in significantly fewer trainable parameters than the other evaluated models. ENet requires only a fraction of the parameters, making it suitable for embedded systems and mobile applications, see Table 1. Its main trade-off is lower accuracy in fine detail segmentation due to reduced resolution, which can be problematic in tasks such as flame edge detection.

The DeepLab-B0 model is a semantic-segmentation network which is less complex compared to the DeepLab-B3 however it has significantly more parameters than the other tested models, See Table 1. A distinctive feature of DeepLab-type models is its use of the Atrous Spatial Pyramid Pooling (ASPP) module, which enables the model to integrate multi-scale contextual information [24]. The decoder structure followed a U-Net–inspired design to refine object boundaries. The B3 variant has a higher parameter count and representational capacity than the B0 model, at the cost of increased computational demand. This approach is particularly effective for segmenting objects of varying size, an advantage when analyzing fire imagery where both small and large flame regions may occur within the same frame. Another major strength of DeepLab v3+ is its ability to handle complex and heterogeneous backgrounds, making it well suited to scenes where object boundaries are indistinct, or the background is highly dynamic. The ASPP module provides a broader contextual understanding of the image, which contributes to greater robustness compared with traditional segmentation architectures, especially in environments with high visual variability [24]. The encoder employs depthwise-separable convolutions and atrous spatial pyramid pooling to capture multi-scale contextual information, while the decoder refines object boundaries for precise segmentation. This design provides a favorable balance between accuracy and computational efficiency, making the model fairly well suited for applications that require real-time inference or deployment on

Table 1

The total amount of available trainable parameters in the models.

	U-Net	ENet	DeepLab-B3	DeepLab-B0	Fast-SCNN
Parameters	1,862,849	46,689	13,769,577	7,026,621	63,841

Note that the number of parameters do not indicate the “intelligence” nor the abilities of the models but rather give a general indicator of their sizes and computational complexities. The reported parameter counts include all trainable weights in the segmentation networks, excluding preprocessing and post-processing steps.

resource-constrained platforms.

DeepLab-B3 follows the same hybrid design strategy as DeepLab-B0 but employs a larger EfficientNet-B3 backbone. This increases the model capacity and accuracy at the expense of higher computational cost and parameter count (under 14M). It retains the U-Net–inspired decoder and skip connections, striking a middle ground between performance and efficiency. While heavier than B0, it remains significantly leaner than the full DeepLabv3+ architecture. Note that DeepLabv3+ was only used in combination with U-Net to construct the hybrid models. All DeepLab models were implemented using standard configurations without modification of backbone depth or ASPP dilation rates.

Fast-SCNN (Fast Semantic Segmentation Convolutional Neural Network) is designed specifically for high-speed segmentation on resource-limited hardware. It employs a two-branch structure: a “learning to down-sample” module for efficient feature extraction, and a lightweight global feature extractor for capturing contextual information. The design enables very fast inference while keeping accuracy competitive. Fast-SCNN is particularly suitable for real-time tasks such as flame segmentation in video streams, though its accuracy may be lower than heavier architectures. The architecture is described in detail in Ref. [26].

It is important to note the significant difference in relative size and complexity of the evaluated models. For this purpose, the number of trainable parameters constitutes a particularly relevant metric, as presented in Table 1. While the parameter count provides an indication of computational complexity and deployment feasibility, it does not directly correspond to segmentation accuracy. ENet and Fast-SCNN can be classified as lightweight, computationally efficient models with relatively few parameters, whereas the two DeepLab variants represent substantially larger and more complex architectures with considerably higher parameter counts. U-Net occupies an intermediate position, being moderately large and resource-demanding, yet still significantly less complex than even the lighter DeepLab model (DeepLab-B0).

2.2. Loss function and assessment criteria

The loss function, accuracy and performance of the models have

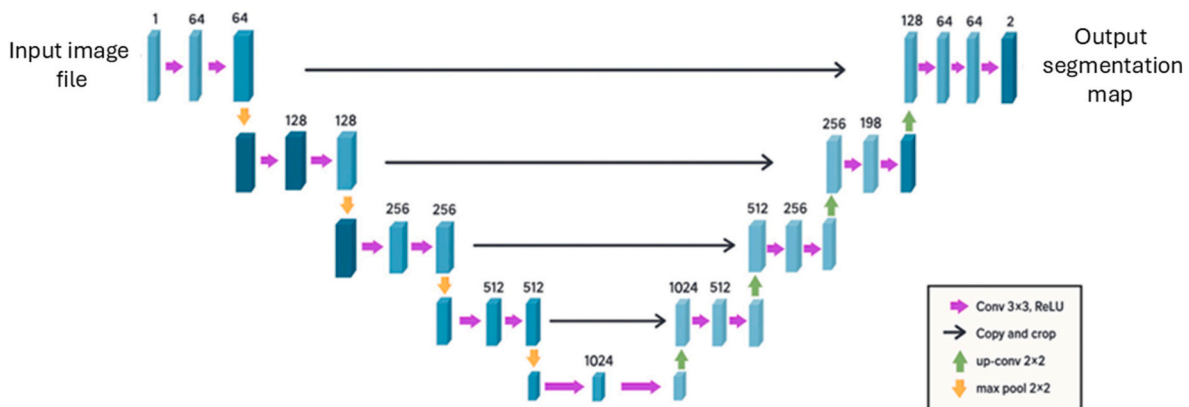


Fig. 1. The image shows a classic U-Net architecture. Inspiration taken from the original work by Ronneberger et al. [22].

been explored by a few metrics, namely, the Binary Cross-Entropy (BCE), DICE loss and Intersection over Union (IoU). BCE is a commonly employed loss function for binary classification tasks, quantifying the discrepancy between ground-truth labels and the model's probabilistic predictions. The loss is defined as the average over all data points, where incorrect predictions are penalized via logarithmic terms. Specifically, when the true label is $y_i = 1$ and the model assigns a high probability ($\hat{y}_i \approx 1$), the loss remains low. Here, y_i represents the true value and the \hat{y}_i the predicted value. Conversely, if the model assigns a low probability ($\hat{y}_i = 0$), the loss increases substantially, thereby penalizing the misclassification. Owing to its effectiveness, BCE is widely utilized in binary classification and image segmentation tasks, particularly in conjunction with sigmoid activation functions in neural networks. The BCE can be expressed as,

$$BCE = -\frac{1}{N} \sum_{i=1}^N [y_i \log \hat{y}_i + (1 - y_i) \log (1 - \hat{y}_i)]. \quad (1)$$

The evaluation metric is based on the Dice coefficient, which penalizes poor overlap by comparing the intersection and the union of predicted and ground-truth pixels in the segmentation. The term ϵ (epsilon) represents an arbitrarily small constant that prevents division by zero. As before, y_i denotes the true class label for pixel i (0 or 1), while \hat{y}_i represents the predicted probability for pixel i of the DICE coefficient,

$$DICE = \left(2 \sum (y_i * \hat{y}_i) + \epsilon \right) / \left(\sum y_i + \sum \hat{y}_i + \epsilon \right) \quad (2)$$

The IoU metric is the common area between prediction and ground truth, the value is doubled and then divided by the sum of the predicted and ground-truth areas. The resulting value ranges between 0 and 1, where 1 indicates a perfect match and 0 indicates no overlap. In this study, a model achieving an average Dice coefficient above 0.9 is considered to demonstrate satisfactory performance. IoU,

$$IoU = \left(2 \sum (y_i * \hat{y}_i) + \epsilon \right) / \left(\sum (y_i + \hat{y}_i - y_i * \hat{y}_i) + \epsilon \right) \quad (3)$$

The Jaccard index, also referred to as the IoU, is a closely related metric that quantifies the overlap between prediction and ground truth. Compared to the Dice coefficient, IoU is slightly stricter, since Dice doubles the numerator and sums both areas in the denominator. Due to this increased sensitivity, a model in this study is regarded as maintaining good performance when the IoU exceeds 0.8. Traditionally, segmentation validation employs either Dice or IoU alone; however, to capture their complementary properties where IoU being more sensitive to small errors and Dice being more robust for very small structures and therefore both metrics are included in the analysis. Another interesting parameter to observe is Precision, which is a commonly used metric for model validation. Its components, True Positives (TP) and False Positives (FP) are obtained from the confusion matrix presented later in the results. Precision quantifies how often the model is correct when it predicts a positive outcome, i.e., the proportion of predicted positives that are actually correct:

$$\text{Precision} = TP / (TP + FP). \quad (4)$$

This metric is particularly relevant in applications where false alarms are costly, such as in medical diagnosis. Sensitivity, also referred to as recall, measures the model's ability to correctly identify all positive cases. It reflects the proportion of actual positive pixels that are successfully captured by the model:

$$\text{Sensitivity} = TP / (TP + FN). \quad (5)$$

Here, FN is the number of false negatives. This metric is especially important in contexts where missed detections are critical, for example in fire detection or cancer screening. The F1-score is the final validation metric applied in this work. It represents the harmonic mean of precision and sensitivity, thereby balancing the two by penalizing large

discrepancies between them:

$$F1 = 2 \times (\text{Precision} \times \text{Sensitivity}) / (\text{Precision} + \text{Sensitivity}). \quad (6)$$

This metric is particularly useful in scenarios where both false alarms and missed detections carry significant consequences.

3. Results and discussion

The segmentation results demonstrated that DeepLab-B0 consistently achieved the highest accuracy, with Dice values approaching 0.92 and IoU values around 0.85 on the validation dataset. Fast-SCNN, while slightly less accurate, provided significantly higher inference speed, making it suitable for real-time applications. U-Net and ENet achieved reasonable performance but struggled to generalize across varying lighting conditions. When comparing volume estimation methods, Shape from Silhouette proved the most precise, producing reconstructions closely matching reference data, but at the cost of long computation times. Paraboloid approximations, although less accurate with errors of about 10–15%, offered much faster performance, thereby providing a practical alternative where speed is critical. Cone approximations systematically underestimated flame volume due to oversimplification.

A comparative assessment of the segmentation models is presented in Table 2, where three validation metrics derived from each model's confusion matrix, as introduced in chapter 2.2 (Equations (1)–(6)). In brief, precision quantifies the proportion of predicted positives that are correct, sensitivity (recall) measures the proportion of true positive pixels correctly identified by the model, and the F1-score represents a harmonic balance between precision and sensitivity. The highest precision is achieved by DeepLab-B0 at approximately 0.96, while the lowest is observed for Fast-SCNN at approximately 0.89. This indicates that, among the pixels predicted as flames, DeepLab-B0 achieved a higher proportion of correct classifications. Notably, DeepLab-B0 even outperforms DeepLab-B3, despite the latter being a larger and more complex model. Equally surprising is that ENet outperforms Fast-SCNN, despite its comparatively smaller architecture.

When examining sensitivity, the ranking shifts to a more expected outcome, with DeepLab-B3 achieving the highest result at approximately 0.99, followed closely by DeepLab-B0, and thereafter by Fast-SCNN, U-Net, and ENet. The F1-score provides a comprehensive summary of the models' predictive accuracy, as it combines the contributions of both precision and sensitivity. Based on this metric, the ranking is DeepLab-B0, DeepLab-B3, U-Net, Fast-SCNN, and ENet, an order that closely reflects the authors' observations throughout the study.

Furthermore, when the Dice and IoU scores are presented in descending order, they mirror the results indicated by the F1-score, with DeepLab-B0 ranked first and ENet ranked fifth (see Table 3).

As an example, U-Net achieved a mean Dice coefficient of 0.94 and IoU of 0.89. A Dice value above 0.9 indicates very close agreement with the reference segmentation, and an IoU above 0.8 shows that the vast majority of the segmented area is correct (1.0 represents perfect overlap). Performance was lower in the second half of the dataset, likely because the last 60 validation images were captured with the camera rotated 90°, introducing a more complex background with lights and reflective metal surfaces. The generalizability of the models is tested for each on new data, it is found that U-Net performs reasonably well in new

Table 2
Performance of the different models, obtained by equations (4)–(6).

Model	Precision	Sensitivity	F1 Measure
U-Net	0.9181	0.9846	0.9502
ENet	0.9023	0.9566	0.9287
DeepLab-B3	0.9292	0.9922	0.9597
DeepLab-B0	0.9589	0.9784	0.9686
F-SCNN	0.8862	0.9860	0.9334

Table 3

The averaged results from the DICE coefficient and the IoU values.

Average	U-Net	ENet	DeepLab-B3	DeepLab-B0	Fast SCNN
DICE	0.94	0.92	0.96	0.97	0.93
IoU	0.89	0.85	0.92	0.93	0.86

settings. It should be noted that while DeepLab-B0 achieved highest accuracy on known environments it struggled heavily with generalization as unknown environments gave far worse results. This would be expected however, as a more complex model like this requires a larger and more diverse training dataset to be able to utilize its larger set of trainable parameters. Otherwise, it risks overfitting like we observe here leading to bad results in new environments. However, it should be acknowledged that the training data are from specific façade fire tests and small scale calibration tests; different camera types, lighting, or fire chemistries could lower performance, so further data diversity is needed for full real-world generalization.

3.1. Results exemplified for U-Net

The results are exemplified by U-Net in detail mainly due to its F1 ranking in the middle of the tested segmentation models, it is not the best nor the worst. In Fig. 1, the U-Net predictions are shown at three different time points. The middle example depicts a single continuous flame, whereas the upper and lower examples contain smaller, disconnected segments. The upper case is noteworthy because the gap between flame regions is very narrow; some other models interpret the flame here as one large body. This leads to substantial variation in estimated flame height, since height is measured from the highest point of the largest connected region. Consequently, two models can produce nearly identical area estimates while reporting markedly different heights from the same image, see Fig. 2.

The scatter plot in Fig. 3 compares the total number of predicted flame pixels to the ground-truth masks, without assessing pixel-wise overlap. It reveals a slight systematic overestimation of flame size relative to the annotations.

A confusion matrix categorizes pixels from the 120 validation images as true background, false background, false flame, or true flame. As expected, true background dominates because most pixels belong to the background, consistent with the observed distribution, see Fig. 4.

Flame volume has previously been linked to heat release rate and thermal radiation through empirical and semi-empirical relationships

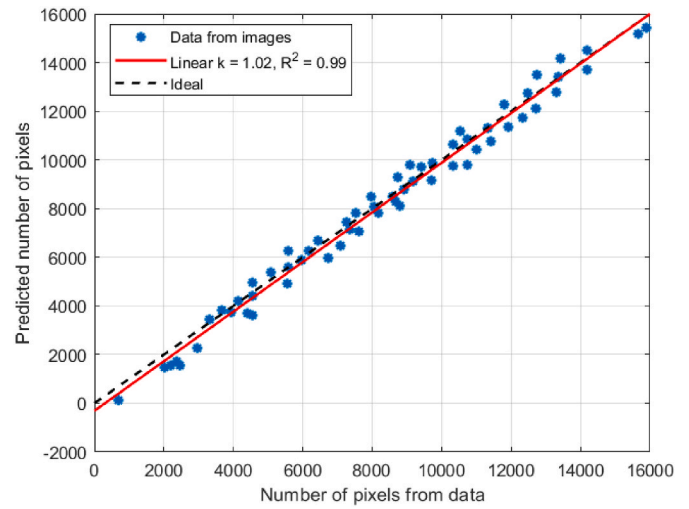


Fig. 3. Comparison between the segmentation and the model prediction for U-Net.

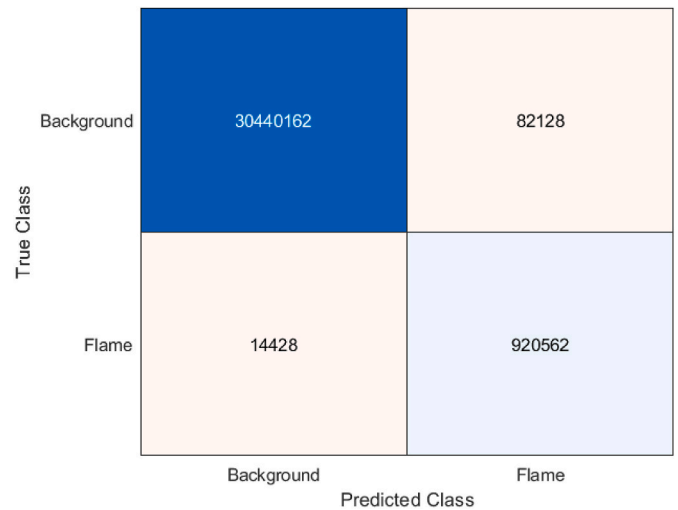


Fig. 4. Confusion chart for U-Net.



Fig. 2. The figure depicts three distinct time points during the fire test, where a prediction is generated by U-Net. The predicted segmentation mask is presented in binary form in the center and is subsequently overlaid on the original image to the right.

(e.g., Refs. [28–31]). However, such approaches typically rely on simplified flame geometries or indirect measurements. The present work does not seek to redefine these relationships, but rather to explore whether modern image-based techniques can provide a consistent and automated means of estimating flame volume from experimental video data, thereby supporting or complementing existing fire characterization methods. In the validation example presented here, the flame volume was converted to heat release rate (HRR) by assuming a representative volumetric heat release rate of approximately 1000 kW/m^3 , consistent but in the lower range compared with values reported in previous studies on flame power density [27–29]. This conversion is used only for approximate comparison of growth behaviour rather than precise HRR prediction. Next, to be able to determine volume several visual sources are needed. In this work a calibration test of a small burner is used where cameras are placed orthogonal to each other. The volume is then based on the segmentation results; two segmented images were combined to create binary masks for 3D volume approximation (Fig. 5). Four different methods to determine the volume are tested, analytical approximations of a paraboloid, cone, the Shape from Silhouette, and space integration. The volume for the paraboloid and the cone is determined by formulas where the base and height are obtained from the video. The Shape from Silhouette is a 3D reconstruction method that by combining 2D video data from at least two sources can numerically compute the volume through extrusion of 2D features and a global intersection. The space integral is found by using the solid of revolution integral. Among the tested methods, the paraboloid approach provided the shortest inference time, making it suitable for real-time applications, and yielded volume estimates between those of the silhouette-based and cone approximations. The silhouette method produced the most visually accurate flame shape, while the space-integral method likely overestimated the volume. Because flame geometry is difficult to validate directly, quantitative evaluation of the volume estimates was not possible, although height and area measurements support their plausibility.

4. Model test and validation

A larger experimental test was conducted where a free-standing tray ($1000 \text{ mm} \times 500 \text{ mm} \times 100 \text{ mm}$) was filled with 30 L of heptane, regularly used in the SP Fire 105 test for external wall claddings. It is to be expected that the heat release rate (HRR) is lower than half of that generated in the SP Fire 105 fire test (approximately 2.5 MW) due to the missing back radiation from the combustion chamber enhancing the gasification of the fuel [32].

The pan containing the heptane was ignited and video recorded from two vertically separated viewpoints, as previously described for the small flame tests, see Fig. 5. The fire development was captured with standard video cameras from these angles and, additionally, with an infrared (IR) camera from one angle to enable future extensions of the

work on quantifying flame volume. The models were not trained on the data from these viewpoints; they were used solely to validate the methodology.

The results showed that the models had a reasonably good ability to identify the flame, but they did not produce consistently high-quality segmentations. The flame sizes were overestimated by lights mounted on the hall wall and also by reflection from some other surfaces, which confounded the models and led them to misclassify some areas as fire. The test is evaluated using the DeepLab-B0 model, which is the best performing model, to compute the flame volume, where one instantaneously explored area can be seen in Fig. 6.

The recorded HRR is displayed in Fig. 7 with the estimated HRR computed from the volume approximation. Due to the mischaracterization of some flames, the flame volume is normalized to have a small start value however some additional areas are unfortunately included in the calculation due to reflections.

The calculated growth rate of the flame volume approximately matches the growth of the fire however quantitative prediction of HRR is impeded by misclassification and thus often overestimated.

5. Conclusions

This study demonstrates the feasibility of estimating flame volume from fire test videos using ML-based semantic segmentation and geometric modeling. DeepLab-B0 provided the most accurate segmentation results, while Fast-SCNN proved most suitable for real-time applications due to its substantially lower computational cost. Shape from Silhouette produced the most accurate volume estimates, whereas paraboloid approximations offered a faster but less precise alternative. The findings suggest that AI-driven methods can enrich both experimental fire research and real-time fire safety monitoring.

Overall, the results suggest two distinct pathways for application. For scientific fire testing where precision is paramount, DeepLab-B0 in combination with Shape from Silhouette provides the most reliable solution. For real-time monitoring in operational settings, Fast-SCNN combined with paraboloid approximation strikes the best balance between accuracy and computational efficiency. Future considerations include the need for careful validation of models on diverse datasets before deployment, as well as transparency about limitations to avoid over-reliance in safety-critical contexts.

The models depend on high-quality, diverse training data to avoid overfitting to specific conditions such as lighting, background, or fuel type, and camera calibration is critical for accurate geometric measurements. DeepLab-B0 delivered the best precision and F1 score on familiar data, while U-Net showed the strongest generalization to new environments, outperforming more complex DeepLab variants. Data augmentation improved all models' F1 scores but increased their tendency to over-segment. Model complexity also affected practical considerations: lightweight networks (Fast-SCNN, ENet) offered faster

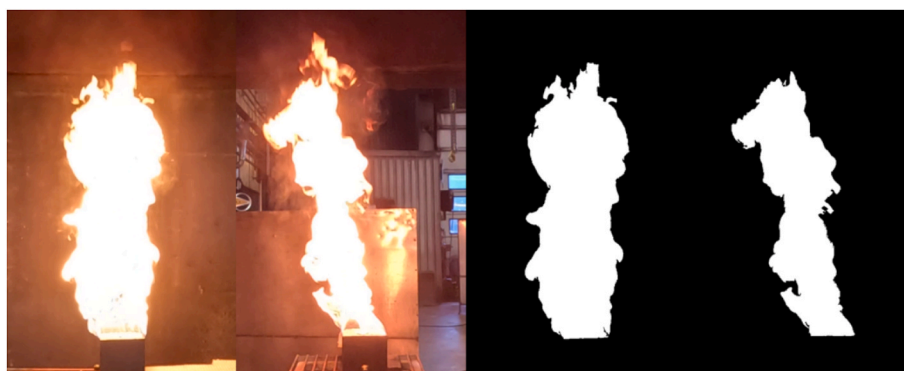


Fig. 5. Data from small-scale fire tests with two orthogonal video streams (left) and masking (right).

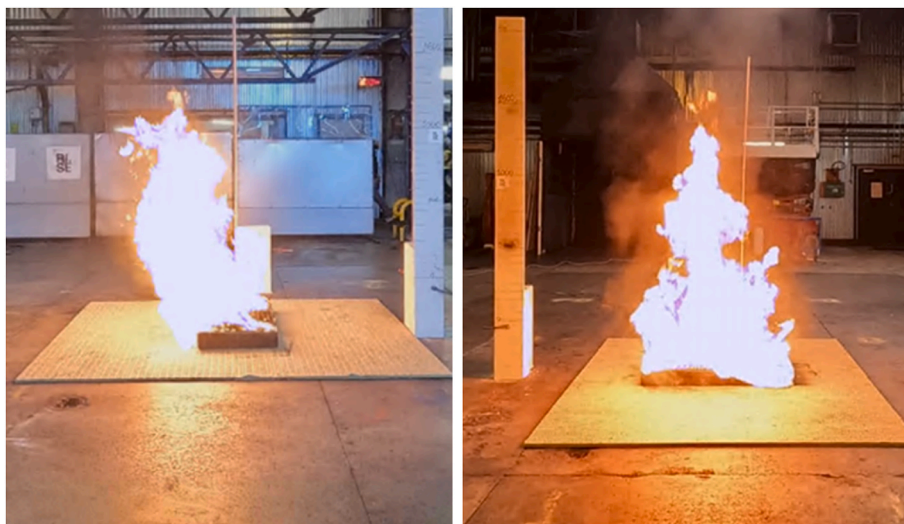


Fig. 6. Validation test, open free burning test using heptane with overlaid masking.

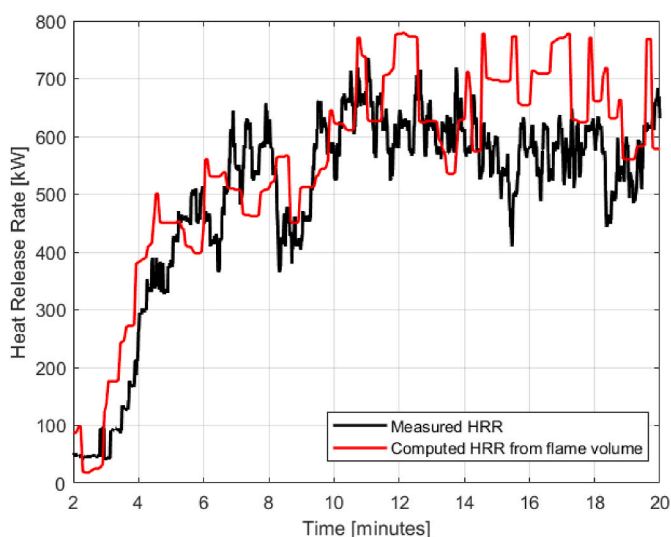


Fig. 7. A comparison of computed HRR from flame volume and the measured HRR in the validation experiment.

training and inference, whereas heavier models (DeepLab, U-Net) provided higher accuracy when speed is less critical.

Volume estimation tests indicated that the paraboloid volume approximation balanced accuracy and real-time performance, and optical distortions could be corrected effectively. As part of the preprocessing of all data a radial/lens distortion error produced by the GoPro fish-eye lens was corrected through the program GyroFlow with the help of GoPros official data logs or through a visual examination to preserve all parallel lines within the captured environment. Overall, the study demonstrates that careful calibration, consistent annotation, and robust preprocessing yield segmentation methods and volume estimates can be transferred to other fire-related or industrial monitoring applications.

Annotation methods also introduce uncertainty: early data were manually labeled with LabelMe, while later data used automated HSV (Hue-Saturation-Value) masking. Mixing these methods may have caused inconsistent flame boundaries and a tendency toward over-segmentation, since the LabelMe masks were generally larger.

Despite these challenges, the work demonstrates how general machine-learning techniques can be adapted to fire research and

emphasizes the importance of careful preprocessing, calibration, and consistent annotation for robust flame-segmentation models. Future work should focus on expanding datasets, incorporating infrared imagery, and exploring hybrid models that combine data-driven and physics-informed approaches. To advance this research, larger and more diverse datasets are needed from fires with different fuels, environments, camera angles, and lighting. Such diversity would improve model generalizability and reduce overfitting. The current training set of only 750 images is small, and models especially complex ones like U-Net and DeepLab would benefit from thousands of varied training images. Volume estimates could also be refined by incorporating temperature measurements or multi-camera 3D reconstruction to increase accuracy. Finally, extending the work to real-time analysis with fast inference is important for practical applications such as industrial fire-alarm systems or for analyzing drone data obtained from outdoor fires. Combining it with, for example, drone-based observations could provide emergency services and other stakeholders with better information to make faster, more informed decisions, ultimately helping to save lives and reduce damage during wildfires. This type of data could be used to predict how a fire will develop under different weather conditions.

CRediT authorship contribution statement

Samuel Tistelgren Holappa: Writing – review & editing, Investigation, Formal analysis, Data curation. **Gustav Schelin:** Writing – review & editing, Methodology, Investigation, Formal analysis. **Johan Anderson:** Writing – original draft, Supervision, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Johan Anderson reports was provided by RISE. Johan Anderson reports a relationship with RISE that includes: employment. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The Authors are grateful for a well-supported collaboration between RISE and University of Borås.

References

- [1] N. Ketkar, *Deep Learning with Python: a hands-on Introduction*, Apress, Bangalore, Karnataka, India, 2017.
- [2] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, *Nature* 521 (2015) 436–444, <https://doi.org/10.1038/nature14539>.
- [3] J.L. Hodges, B.Y. Lattimer, K.D. Luxbacher, Compartment fire predictions using transpose convolutional neural networks, *Fire Saf. J.* 108 (2019) 102854, <https://doi.org/10.1016/j.firesaf.2019.102854>.
- [4] K. McGrattan and S. Hostikka, "FDS Fire Dynamics Simulator (Version 6.7.9) User's Guide," NIST Special Publication 1019 sixth ed.
- [5] L. Pulga, G. Bianchi, S. Falfari, C. Forte, A machine learning methodology for improving the accuracy of laminar flame simulations with reduced chemical kinetics mechanisms, *Combust. Flame* 216 (2020) 72–81, <https://doi.org/10.1016/j.combustflame.2020.02.021>.
- [6] H. Li, X. Fei, C. He, Study on Most important factor and Most vulnerable location for A forest fire case using various machine learning techniques, in: 2018 Sixth International Conference on Advanced Cloud and Big Data, 2018, pp. 298–303.
- [7] H. Pourghasemia, A. Gayenb, R. Lasaponarac, J.P. Tiefenbacherd, Application of learning vector quantization and different machine learning, *Environ. Res.* (2020) 109321, <https://doi.org/10.1016/j.envres.2020.109321>.
- [8] J. Hodges, Predicting Large Domain Multi-Physics Fire Behavior Using Artificial Neural Networks, Virginia Polytechnical Institute and State University, 2018. PhD thesis.
- [9] B. Lattimer, J. Hodges, A. Lattimer, Using machine learning in physics-based simulation of fire, *Fire Saf. J.* 114 (2020) 102991, <https://doi.org/10.1016/j.firesaf.2020.102991>.
- [10] J. Anderson, A. Mossberg, E. Gard, et al., Investigating Machine Learning for Fire Sciences - Literature Review and Examples, RISE Report, vol. 59, 2021.
- [11] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, P. Martinez-Gonzalez, J. Garcia-Rodriguez, A survey on deep learning techniques for image and video semantic segmentation, *Appl. Soft Comput.* 70 (2018) 41–65.
- [12] I. Goodfellow, Y. Bengio, A. Courville, *Deep learning*. Cambridge, MA: MIT press. Gragnaniello, D., greco, A., sansone, C. & vento, B. (2025). 'FLAME: fire detection in videos combining a deep neural network with a model-based motion analysis', *Neural Comput. Appl.* (2016). <https://link.springer.com/article/10.1007/s00521-024-10963-z>.
- [13] H. Guo, Z. Sun, T. Zhang, S. Zhang, Application of deep learning in remote sensing fire detection: a review, *ISPRS J. Photogrammetry Remote Sens.* 175 (2021) 276297.
- [14] A.E. Çetin, K. Dimitropoulos, B. Gouverneur, N. Grammalidis, O. Günay, Y. H. Habiboğlu, B.U. Töreyn, S. Verstockt, Video fire detection - review, *Digit. Signal Process.* 23 (2013) 1827–1843, <https://doi.org/10.1016/j.dsp.2013.07.003>.
- [15] Z. Wu, R. Xue, H. Li, Real-time video fire detection via modified YOLOv5 network model, *Fire Technol.* 58 (2022) 2377–2403, <https://doi.org/10.1007/s10694-022-01260-z>.
- [16] D. Gragnaniello, A. Greco, C. Sansone, B. Vento, Fire and smoke detection from videos: a literature review under a novel taxonomy, *Expert Syst. Appl.* 255 (2024) 124783, <https://doi.org/10.1016/j.eswa.2024.124783>.
- [17] R. Hartley, A. Zisserman, *Multiple View Geometry in Computer Vision*, 2 Ed., Cambridge University Press, Cambridge, 2003.
- [18] J. Anderson, J. Sjöström, R. Chiva, F. Dumont, A. Hofmann-Böllinghaus, P. Tóth, O. Lulu, L. Boström, Finalization of the European Approach to Assess the Fire Performance of Facades, 2024, <https://doi.org/10.2873/7300386>, 978-92-68-20808-3. Final report. October 2024.
- [19] Deutsches Institut für Normung, DIN 4102-20: Fire Behaviour of Building Materials and Elements – Part 20: Supplementary Verification for the Assessment of External Wall Claddings, 2018.
- [20] British Standards Institution, BS 8414-1: Fire Performance of External Cladding Systems – Test Method for Non-loadbearing External Cladding Systems Applied to the Masonry Face of a Building, 2020.
- [21] Gyroflow Developers, Gyroflow (version 1.6). <https://gyroflow.xyz/>, 2024.
- [22] O. Ronneberger, P. Fischer, T. Brox, U-Net: convolutional networks for biomedical image segmentation, *MICCAI* (2015) arXiv:1505.04597.
- [23] A. Paszke, A. Chaurasia, S. Kim, E. Culurciello, Enet: a Deep Neural Network Architecture for Real-Time Semantic Segmentation, 2016 arXiv:1606.02147.
- [24] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation (DeepLabv3+), *ECCV* 2018 (2018) [arXiv:1802.02611].
- [25] A. Howard, M. Sandler, et al., Searching for MobileNetV3. *ICCV* 2019. This Introduces MobileNetV3, 2019.
- [26] R.P.K. Poudel, S. Liwicki, R. Cipolla, Fast-Semantic Segmentation Network (Fast-SCNN), 2019 arXiv:1902.04502.
- [27] PyTorch, PyTorch. <https://pytorch.org/>, 2025.
- [28] [a] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, O'Reilly Media, Canada 2019, 2019; [b] L. Hu, X. Zhang, Q. Wang, A. Palacios, Flame size and volumetric heat release rate of turbulent buoyant jet diffusion flames in normal- and a sub-atmospheric pressure, *Fuel* 150 (2015) 278–287, <https://doi.org/10.1016/j.fuel.2015.01.081>.
- [29] Y. Xin, Estimation of chemical heat release rate in rack storage fires based on flame volume, *Fire Saf. J.* 63 (2014) 29–36, <https://doi.org/10.1016/j.firesaf.2013.11.004>.
- [30] J. De Ris, L. Orloff, Flame heat transfer between parallel panels, *Fire Saf. Sci.* 8 (2005) 999–1010, <https://doi.org/10.3801/IAFSS.FSS.8-999>.
- [31] G. Shen, K. Zhou, F. Wu, J. Jiang, Z. Dou, A model considering the flame volume for prediction of thermal radiation from pool fire, *Fire Technol.* 55 (2019) 129–148, <https://doi.org/10.1007/s10694-018-0779-y>.
- [32] SP FIRE 105, Method for Fire Testing of Facade Materials, Swedish National Testing and Research Institute (SP), Borås, Sweden, 1994.