Efficient Fire Segmentation for Internet-of-Things-Assisted Intelligent Transportation Systems

Khan Muhammad, *Senior Member, IEEE,* Hayat Ullah, *Student Member, IEEE,* Salman Khan, *Student Member, IEEE,* Mohammad Hijji, *Member, IEEE,* Jaime Lloret, *Senior Member, IEEE*

***Abstract*—Rapid developments in deep learning (DL) and the Internet-of-Things (IoT) have enabled vision-based systems to efficiently detect fires at their early stage and avoid massive disasters. Implementing such IoT-driven fire detection systems can significantly reduce the corresponding ecological, social, and economic destruction; they can also provide smart monitoring for intelligent transportation systems (ITSs). However, deploying these systems requires lightweight and cost-effective convolutional neural networks (CNNs) for real-time processing on artificial intelligence (AI)-assisted edge devices. Therefore, in this paper, we propose an efficient and lightweight CNN architecture for early fire detection and segmentation, focusing on IoT-enabled ITS environments. We effectively utilize depth-wise separable convolution, point-wise group convolution, and a channel shuffling strategy with an optimal number of convolution kernels per layer, significantly reducing the model size and computation costs. Extensive experiments on our newly developed and other benchmark fire segmentation datasets reveal the effectiveness and robustness of our approach against state-of-the-art fire segmentation methods. Further, the proposed method maintains a balanced trade-off between the model efficiency and accuracy, making our system more suitable for IoT-driven fire disaster management in ITSs.**

***Index Terms*—Convolutional Neural Networks, Deep Learning, Edge Intelligence, Fire Segmentation, Intelligent Transportation Systems, Internet of Things (IoT), Semantic Segmentation.**

I. Introduction

Recent advancements in cutting-edge camera technologies have empowered today's surveillance cameras with next-level processing capabilities, offering real-time processing of video streams and other artificial intelligence (AI) algorithms for a variety of applications, including abnormal activities recognition [1, 2], fire detection [3], safety [4, 5], traffic management [6, 7], intelligent transportation of unmanned vehicles [8-10], and scene classification [11]. Such intelligent cameras play vital roles in Internet-of-Things (IoT)-enabled smart surveillance systems, e.g., for processing the real-time visual data of any disaster (e.g., fires, floods, and earthquakes) and instantly notifying the appropriate disaster management departments. Among disasters, fire is the most severe threat to densely populated areas, airports, and forests, owing to its high frequency and destructive nature. Therefore, edge-driven smart monitoring is urgently needed to prevent fire disasters in their early stages, i.e., before they lead to massive damage in terms of human lives and financial losses.

Aiming to save human lives, researchers have been working for the past two decades on both conventional sensors and vision-based fire and smoke detection methods. Conventional sensors often include smoke, fire, and temperature sensors [12, 13], which are economical and easy to deploy for real-time fire detection. However, these sensors are restricted to small geographical areas and cannot detect fires in large areas, such as large industrial sectors, intelligent transportation systems (ITSs), and outdoor IoT environments. Researchers have presented several vision-based approaches, including traditional handcrafted features and learning-based methods, to detect fires in outdoor and large geographical areas. Current literature reports that traditional methods [14-21] use motion, texture, and color features of flames for fire region detection. For instance, Celik et al. [22] presented an enhanced variant of a generic color model by adding fuzzy logic to their fire-specific pixel classification method [19]. The replacement of heuristic rules with fuzzy logic significantly improved the classification performance of their method, allowing them to effectively distinguish the colors between fire flames and other flames-like objects. Byoung et al. [23] presented a probabilistic color-driven method using the YUV color space and a support vector machine (SVM). Their proposed approach first detected fire-specific pixels in the moving regions of an image using high luminance information. Subsequently, they created a temporal fire model with wavelet coefficients and employed a binary-class SVM for the final fire-specific pixel classification. The main issue with this method was the high false-alarm rate in non-fire regions, which requires further reduction. In general, the performance of the above methods relies on the quality of the manually engineered handcrafted features, thereby restricting them from more challenging scenarios. Therefore,

Manuscript received April 18, 2022; Revised July 28, 2022; Accepted August 30, 2022; Published: XXXX. (*Corresponding author: Khan Muhammad*).

Khan Muhammad is with the Visual Analytics for Knowledge Laboratory (VIS2KNOW Lab), Department of Applied Artificial Intelligence, School of Convergence, College of Computing and Informatics, Sungkyunkwan University, Seoul 03063, Republic of Korea (e-mail: khan.muhammad@ieee.org).

Hayat Ullah is with the Intelligent Systems, Computer Architecture, Analytics, and Security Laboratory (ISCAAS Lab), Department of Computer Science, Kansas State University, Manhattan, Kansas State, USA (e-mail: hayatuallah@ieee.org).

Salman Khan is with the Visual Artificial Intelligence Laboratory (VAIL), School of Engineering, Computing and Mathematics, Faculty of Technology, Design and Environment, Oxford Brookes University, Oxford, United Kingdom (e-mail: salmank@ieee.org).

Mohammad Hijji is with the Industrial Innovation and Robotic Center (IIRC), University of Tabuk, Tabuk 47711, Saudi Arabia and also with the Faculty of Computers and Information Technology (FCIT), University of Tabuk, Tabuk 47711, Saudi Arabia (e-mail: m.a.hijji@gmail.com).

Jaime Lloret is with the Universitat Politecnica de Valencia, Spain (e-mail: jlloret@dcom.upv.es).

obtaining a stable trade-off between the false-alarm rate and accuracy using traditional approaches remains challenging. Furthermore, these traditional methods cannot detect fires from afar or in small volumes in a video stream. Moreover, they cannot detect high volumetric fires from near distances and often fail to detect flames with varying colors.

Several deep learning-based fire detection and segmentation methods have been presented to address the limitations of traditional color-based approaches. For example, Sharma et al. [24] proposed a learning-based approach and investigated two pre-trained deep convolutional neural network (CNN) architectures (VGG16 and ResNet50) for fire detection. Their proposed method obtained a reasonable classification accuracy, but its higher computational complexity makes it infeasible for real-time fire detection. Mao et al. [25] explored a multi-channel CNN for fire scene classification, where they analyzed each channel of the input image at multiple convolutional layers and chose the most discriminative feature maps for accurate classification of the flames. However, owing to the high time complexity, their method is limited to still images and cannot process videos for fire scene classification. Recently, a computationally efficient learning-based unified architecture was presented in [26] for fire-specific region detection and localization. Their system first performed fire scene classification and then inspected the fire image using multiple activation maps of the convolutional layer for localizing the fire-specific regions.

Considering fire detection/segmentation literature, recent CNN-based studies have significantly improved the detection rate and localization performance relative to traditional fire detection methods. However, there remains a need to improve fire segmentation performance from both qualitative and quantitative perspectives, with a focus on deployment in edge-centric IoT surveillance environments. With these motivations, we analyzed the existing CNN models for fire segmentation tasks and adopted a lightweight yet robust architecture as the backbone of our framework for an IoT-enabled surveillance environment. The key contributions of this study are as follows.

1. We comprehensively analyzed various state-of-the-art semantic segmentation architectures in terms of their computational complexity, model size, and model performance for fire segmentation to perform fire recognition over an edge-centric computing platform. As a result, we proposed a computationally efficient framework for an IoT-enabled ITS environment. For the efficient segmentation of fire flames in the IoT environment, we proposed a segmentation architecture with minimal computational complexity and overwhelming segmentation results.
2. Several benchmark datasets have been reported in the literature; however, these datasets have a limited number of images containing small volumes of fire as captured in normal environments. In this study, we created our own fire semantic segmentation dataset (pixel-wise annotation for fire regions) for benchmarking purposes. The dataset contained images of both street fire and wildfire environments, covering the scope of ITSs. Our newly created dataset is publicly available for research purposes for mature fire-segmentation systems.
3. The proposed framework used depth-wise separable convolution, point-wise group convolution, and channel shuffling to optimize the size of the model, thereby significantly reducing computational complexity while maintaining a satisfactory level of accuracy. Our training strategy reduced the model size from 187 MB to 1.49 MB, making the proposed framework a better candidate for real-time processing in an IoT-enabled surveillance environment for fire segmentation in ITSs.

The rest of this article is organized as follows. Section II presents an overview of our proposed framework and its main components. Section III discusses the experimental settings, datasets, and critiques of the results. Finally, Section IV concludes the article with possible future research directions.

II. Proposed Framework

This section describes the working procedure of the proposed framework and its major components. For better understanding, the proposed framework is divided into two sections. In the first section, we provide the architectural details of the proposed architecture for fire segmentation. The second section presents the details of our customized shuffleNetV1 architecture used as an encoder in the proposed framework. A detailed graphical outline of the proposed fire-segmentation approach is shown in Fig. 1.

A. Architectural Details of Proposed Segmentation Method

In this section, we present the details of the proposed segmentation network for fire segmentation on resource-constrained devices. Our model was inspired by the UNet architecture, which was originally introduced as an improved version of a fully convolutional network (FCN) for the semantic segmentation of medical images. The overall UNet architecture comprises two subnetworks—encoder and decoder—responsible for feature extraction and saliency prediction, respectively. The encoder part of the UNet architecture comprises four modules with two unpadded 3 × 3 2D convolutions, followed by rectified linear unit (ReLU) activation and a batch normalization layer. Next, a 2 × 2 max pooling layer downscales the receptive field of the extracted feature maps produced by the convolutional layers at different levels and encodes the image (2D representation) into a feature vector (1D representation) in the final layer. In contrast, the decoder part of the UNet architecture contains four distinct modules with 2 × 2 up-sampling (transpose convolution) and 3 × 3 standard convolution layers, followed by a ReLU activation function. Each module in the decoder part is concatenated with its corresponding module in the encoder part to transfer the deep discriminative features learned by the encoder part of the UNet architecture. The decoder takes the latent feature vector as an input and reconstructs the image with the localized saliency of an object using transpose and standard 2D convolutions. Finally, the output of the last module in the decoder part of the UNet architecture is convolved with 1 × 1 (point convolution) to transform the number of output channels of the last layer into the total number of classes for segmentation.

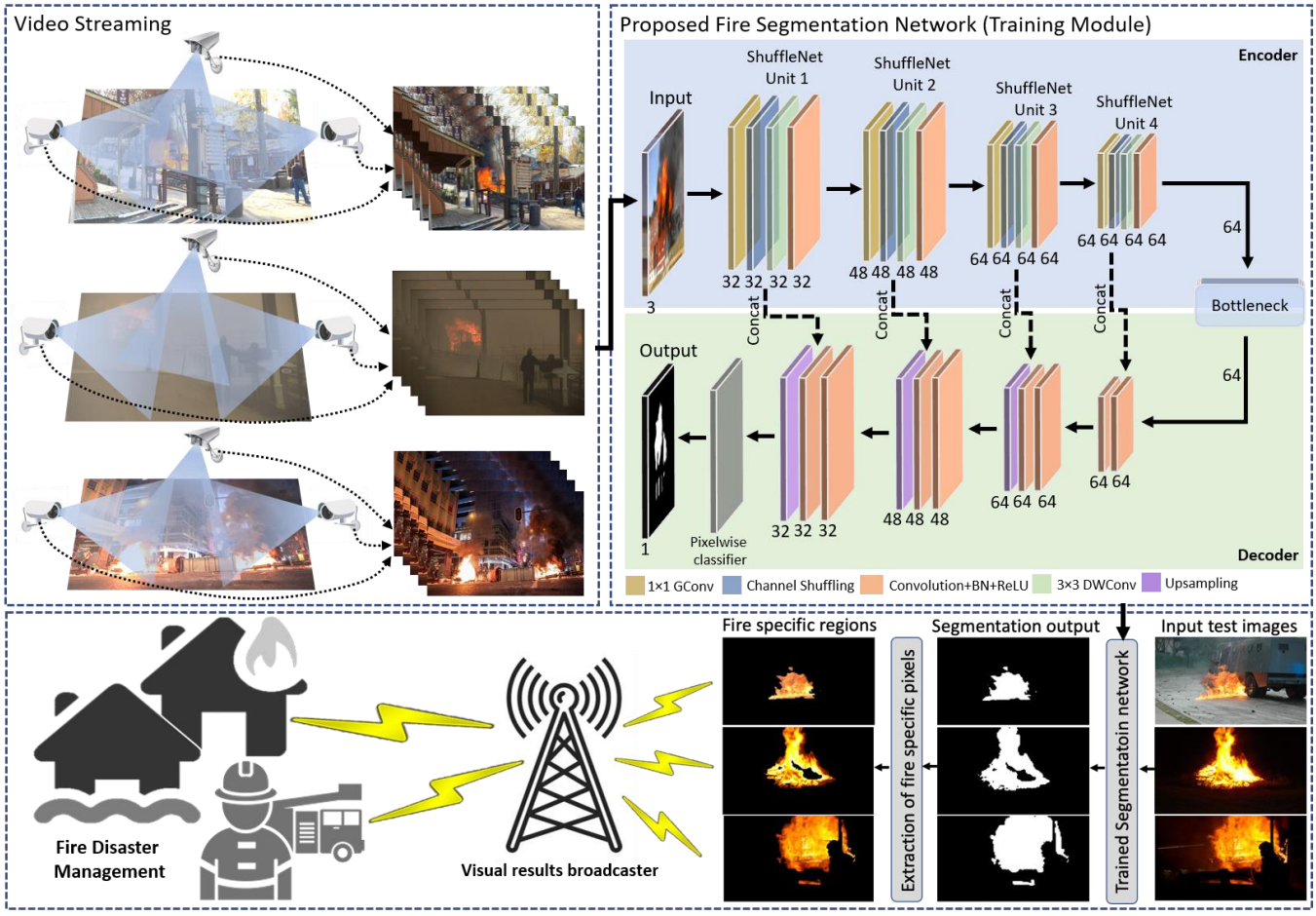

Fig. 1. Proposed deep convolutional neural network (CNN) architecture for fire segmentation in surveillance videos with two tiers: encoder and decoder. The encoder takes an RGB image captured in uncertain environment having intensive fire as an input and processes it through four ShuffleNet units, which generate a latent representation. The latent representation of the given fire image from the architecture bottleneck is then forwarded to the subnetwork (decoder), which processes it on eight convolutional layers followed by four up-sampling layers. These layers are designed in such a way that after each two convolutional layers, there is an up-sampling layer. At the end, there is a pixel-wise classifier that generates the final binary mask for the fire. The predicted segmentation masks are then used to extract fire-specific regions from the input image as a final output of our proposed method.

Considering the technical aspects of the UNet framework, it is worth noting that the encoder is the key component and plays an important role in the segmentation accuracy and computational complexity of the overall architecture. In addition, the encoder network has a set of convolutional blocks (containing several convolutional layers per block); these can be substituted with any lightweight pre-trained CNN network to boost segmentation performance and minimize the overall computational complexity of the proposed approach. Therefore, we investigated different CNN architectures in this study, including VGG16 [27], ResNet50 [28], MobileNetV1 [29], and shuffleNetV1 [30], as the encoder (backbone feature extractor) for the UNet architecture. The statistical details of each CNN are listed in Table I. We analyzed each CNN architecture from two different perspectives: computational complexity (number of training parameters and memory requirements for training) and segmentation performance (quantitative and qualitative evaluations). After extensive experimental evaluations, we found the shuffleNetV1 architecture to be the most computationally efficient yet accurate among all the investigated CNNs. Therefore, we replaced the encoder part (with standard CNN layers) of the UNet architecture with shuffleNetV1 units, except for the first convolution layer (consisting of depth-wise separable convolution, point-wise group convolution, and channel shuffling), significantly reducing the overall computational complexity while preserving the same level of accuracy. A detailed explanation of the proposed shuffleNetV1 encoder is provided in Section II (B).

B. Proposed ShuffleNetV1 Encoder

Considering the real-time task achievements required on embedded edge devices, the proposed fire segmentation framework should offer edge-based computing facilities by utilizing an efficient CNN to process the input stream in real-time and segment the fire-specific regions. To obtain a computationally efficient segmentation network with reduced number of parameters, we replaced the encoder part of the UNet architecture with four shuffleNetV1 units. Each unit comprises three distinct modules, i.e., one for depth-wise separable convolution, point-wise group convolution, and channel shuffling, respectively.

TABLE I
STATISTICAL COMPARISON OF SHUFFLENETV1 AGAINST OTHER ARCHITECTURES

Backbone Architecture	Number of parameters (millions)	Accuracy (%) Top-1	Accuracy (%) Top-5
VGG16 [27]	138	70.5	91.0
ResNet50 [28]	25	77.1	93.2
MobileNetV1 [29]	4.2	70.9	89.9
ShuffleNetV1 [30]	3.4	70.9	90.0

The depth-wise separable convolution is introduced to factorize the standard convolution into a depth-wise convolution, followed by the point-wise convolution (1 × 1 convolution). The depth-wise separable convolution layer processes each channel of the input image individually and then combines the resultant feature maps of the depth-wise convolutions using the point-wise (1 × 1) convolution. In contrast, the point-wise group convolution performs a group operation using a 1×1 kernel, such that each convolution operates on the receptive field of the corresponding channel group, thus drastically reducing the computational complexity of the architecture. The channel shuffling unit is placed immediately after the first point-wise group convolution to help the information from the previous layer flow uniformly across all feature channels. Furthermore, we modified the internal structure of the shuffleNetV1 architecture to further reduce the overall complexity of the proposed fire segmentation framework. In particular, the standard shuffleNetV1 utilizes five different groups to acquire the desired number of output channels. The formation of these groups is configurable and can be adjusted according to the problem to obtain an optimal, computationally efficient, and precise model. Therefore, we modified the shuffleNetV1 units in two different ways: 1) instead of five groups (g = 1,2,3,4,8), we used only one group (g = 2) with no repetition, and 2) we reduced the number of output channels of each layer while maintaining the same level of performance. In our approach, the shuffleNet unit initiates processing on the input with 1 × 1 group convolution (GConv), followed by batch normalization (BN) and a ReLU layer. Next, the channel shuffling module performs a shuffling operation on the output channels of the feature maps of the previous layer and forwards them to the 3 × 3 depth-wise separable convolution (DWConv) layer. The DWConv layer applies a computationally efficient 3 × 3 depth-wise convolution with stride = 2 and BN on the feature maps from the channel shuffling layer. The second 1 × 1 GConv reshuffles the channel dimensions of the feature maps of the previous layer to match the output channel dimensions of the 3 × 3 average pooling layer (AVG pool) at a shortcut path. Concatenation is used to form the final output by combining the shortcut path channel and 1×1 GConv layer output channel. Table II lists the statistical details of the backbone feature extractor employed in the UNet architecture.

III. EXPERIMENTS, RESULTS, AND DISCUSSION

This section provides the details of the implementation setup, followed by a detailed overview of our annotated and existing benchmark datasets used in the experimental evaluations. Following this, we present a detailed comparative analysis of our method with conventional and state-of-the-art fire-segmentation methods. Finally, to validate the efficiency and generalization of our method, we evaluate the performance of our proposed system based on its computational and time complexity, focusing on suitability and deployment for smart surveillance settings in IoT-assisted ITS.

TABLE II
STATISTICAL OVERVIEW OF MODIFIED SHUFFLENETV1 ARCHITECTURE

Layer	Output size	KSize	Stride	Repeat	Output channels (*g* groups)	
					Original *g* = 2	**Our *g* = 2**
Input	224 × 224	-	-	-	3	**3**
Conv1	112 × 112	3 × 3	2	1	24	**24**
MaxPool	56 × 56	3 × 3	2	-	-	**-**
Stage2	28 × 28	-	2	1	200	**32**
Stage3	14 ×14	-	2	1	400	**48**
Srage4	7 × 7	-	2	1	800	**64**

A. Implementation Details

All experiments were conducted on a computer system equipped with an NVIDIA graphic card GeForce GTX 1060 (6GB), 16 GB of onboard memory, and a 3.60-GHz processor. The proposed framework was implemented in Python (version 3) using the well-known deep learning framework Keras with TensorFlow running in the backend. We initialized the training process with a random normal weight initializer; the values of the hyperparameters were set to optimizer = Adam, learning rate = 0.0001, batch size = 32, and epochs = 50. As we were considering a pixel-wise classification problem, cross-entropy was used as the loss function.

B. Details of the Datasets

To investigate the quantitative and qualitative performance of our proposed system, we created a new fire segmentation dataset from 20 videos downloaded from YouTube containing outdoor fire scenes captured at different times of the day with varying lighting conditions. The newly created dataset contained 600 images of different fire incidents, including buildings, vehicles, and forest fires, with different levels of impairment. Each image in the dataset was manually annotated to obtain corresponding ground truth binary masks with fire-specific information. We uploaded our dataset to GitHub, which can be used by the research community to further enhance fire segmentation models. During the experiments, we used 70% of the data for training, whereas the remaining 30% was used for testing purposes. Furthermore, to validate the generalization of our proposed method on other datasets, we collected a test set from two other benchmark datasets (given in [14] and [31]). The first test set comprised 8033 images randomly selected from [14], with 1845 fire-incident images and 6188 without fires. The second test set [31] consisted of 226 images, with 119 fire images and 107 normal images with fire-like visuals, including sunlight reflecting off clouds and sunset, and street lights at night. The statistical details of the training and testing sets are listed in Table III.

TABLE III
STATISTICAL DESCRIPTION OF THE DATASETS USED FOR TRAINING AND EVALUATING OUR PROPOSED METHOD

	Dataset	**Fire**	**Non-fire**	**Total data**
Train set	Our dataset	600	500	1100
Test set	Chino et al. [31]	119	107	226

C. Effectiveness of Our Method with Different Convolutional Neural Network (CNN) Baselines for Fire Segmentation

This section presents a detailed comparative evaluation of our proposed architecture with other CNNs, under the setting of the UNet for fire segmentation. We used two different performance evaluation schemes to evaluate the performance of our proposed CNN architecture against others. The first evaluation scheme used four different performance assessment metrics—pixel accuracy ($Pixel_{accuracy}$), mean accuracy ($Mean_{accuracy}$), mean intersection-over-union ($Mean_{IoU}$), and frequency-weighted intersection-over-union (FW_{IoU})—commonly used for semantic segmentation performance evaluations based on pixel-wise accuracy and region intersection-over-union. $Pixel_{accuracy}$ computes the total number of pixels correctly classified, as given in (1). $Mean_{accuracy}$ is defined as the number of correctly classified pixels over the total number of classes, as shown in (2). $Mean_{IoU}$ first estimates the intersection-over-union (IoU) value for each class and then approximates the average IoU over the total number of classes, as formulated in (3). FW_{IoU} is an extended version of $Mean_{IoU}$, where IoU is weighted based on the frequency of each object class, as given in Equation (4).

$$Pixel_{accuracy} = \frac{\sum_i n_{ii}}{\sum_i t_i} \tag{1}$$

$$Mean_{accuracy} = \frac{1}{n_{cl}} \sum_i \frac{n_{ii}}{t_i} \tag{2}$$

$$Mean_{IoU} = \frac{1}{n_{cl}} \sum_i \frac{n_{ii}}{(t_i + \sum_j n_{ji} - n_{ii})} \tag{3}$$

$$FW_{IoU} = \sum_k \frac{1}{t_k} \sum_i \frac{t_i n_{ii}}{(t_i + \sum_j n_{ji} - n_{ii})} \tag{4}$$

In the above, n_{ii} indicates correctly classified pixels, and t_i is the total number of pixels in class i. n_{ij} indicates the number of pixels belonging to class i but predicted as class j. n_{cl} represents the total number of classes. In our case, we had two classes: 0) background and 1) fire. We compared the adopted CNN (shuffleNetV1) with other state-of-the-art CNNs using the aforementioned evaluation metrics, and the results are listed in Table IV. The results presented in Table IV compare the performance of the proposed CNN architecture with those of three classification CNNs for fire segmentation, namely, VGG16, ResNet50, and MobilenetV1. It can be seen that VGG16 attains reasonable results in terms of $Pixel_{accuracy}$ and FW_{IoU}; however, its $Mean_{accuracy}$ and $Mean_{IoU}$ scores are lower than those of the other architectures. Although ResNet50 and MobileNetV1 obtain identical scores for $Pixel_{accuracy}$, MobileNetV1 performed better than ResNet50 for $Mean_{accuracy}$, $Mean_{IoU}$, and FW_{IoU}. Compared to the other three architectures, our proposed method performs better and obtained the highest $Pixel_{accuracy}$, $Mean_{accuracy}$, $Mean_{IoU}$, and FW_{IoU} scores of 89.54%, 74.27%, 67.39%, and 82.64%, respectively, demonstrating its superiority.

D. Comparison of our Method with other State-Of-The-Art Semantic Segmentation Architectures

To analyze the effectiveness of our method for fire segmentation tasks relative to those of existing methods, we conducted a comparative analysis of different state-of-the-art segmentation networks, including SegNet [32], FCN [33], and PSPNet [34]. The results obtained from the comparative analysis are presented in Table V. Notably, SegNet obtains comparatively lower $Pixel_{accuracy}$, $Mean_{accuracy}$, and FW_{IoU} values than FCN and PSPNet. However, its $Mean_{IoU}$ is higher than those of the FCN and PSPNet. The FCN and PSPNet obtain nearly similar FW_{IoU} values; however, PSPNet is better than the FCN in terms of $Pixel_{accuracy}$, $Mean_{accuracy}$, and $Mean_{IoU}$. The proposed system obtains the highest $Pixel_{accuracy}$, $Mean_{accuracy}$, $Mean_{IoU}$, and FW_{IoU} values among all the deep learning-based segmentation methods. A set of visual segmentation results obtained by our approach and those from other fire segmentation approaches are depicted in Fig. 2.

TABLE IV
COMPARATIVE RESULTS OF OUR METHOD AND OTHER STATE-OF-THE-ART CNN ARCHITECTURES ON OUR NEWLY CREATED FIRE SEGMENTATION DATASET

Model	$Pixel_{accuracy}$	$Mean_{accuracy}$	$Mean_{IoU}$	FW_{IoU}
UNet+VGG16	85.22	61.30	56.19	76.84
UNet+ResNet50	88.43	69.17	62.47	79.92
UNet+MobileNetV1	88.29	71.05	63.56	80.15
Proposed	**89.54**	**74.27**	**67.39**	**82.64**

E. Fire Region Extraction Qualitative Analysis

In this section, we present a detailed discussion of the fire-specific region extraction results relative to those of other fire segmentation methods. For illustration purposes, we visualize the results of seven images randomly picked from the test set, along with their corresponding methods. The fire-specific regions from the original image were extracted using the corresponding pixel retrieval method; the obtained visual results are shown in Fig. 3. For each foreground pixel of the binary segmented image, we selected the pixel value per channel from the original image using the corresponding pixel location. The obtained fire-specific region extraction results were visually compared with those of state-of-the-art fire localization methods, including Chino et al. [31], Rossi et al. [35], Celik et al. [19], Rudz et al. [36], Chen et al. [17], and CNNFire [26], as shown in Fig. 4. As shown, the visual results obtained by Rossi et al. [35] and Chen et al. [17] are affected by a high misclassification rate, whereas the results of Chino et al. (BoWFire) and color classification [31], Celik et al. [19], Rudz et al. [36], and CNNFire [26] are approximately similar, with minor differences in the boundary regions. Fig. 5 illustrates a comparative analysis of the results obtained by our method and those of other fire segmentation methods based on another sample from the test set. It can be perceived from the results that the methods of Rossi, Rudz, and Chen fail to extract fire regions; however, Chen's method is better than Rossi's and Rudz's in terms of the false positive rate. Chino's method extracts fire-specific regions with no false positive predictions, but its results suffer from misclassifications of fire pixels. The results obtained using our proposed method, Celik's method, color classification, and CNNFire have similar fire regions.

However, based on the false-positive rate, our proposed approach outperforms the other three fire segmentation methods. The visual comparisons in Figs. 4 and 5 validate the effectiveness of our proposed method over the existing traditional and deep learning-based methods. In addition, we compared our results with those obtained from state-of-the-art segmentation networks, as depicted in Fig. 6. By analyzing the presented visual results depicted in Fig. 6, it can be seen that our method obtains comparatively better results than those of the other segmentation networks, including SegNet, FCN, and PSPNet.

F. Running Time Analysis and Feasibility Assessment

Experiments were conducted on a machine equipped with the specifications given in Section III (Implementation Details) to verify the effectiveness and robustness of our proposed system for real-time scenarios in IoT environments. Our proposed method obtained 27 frames per second (FPS), thereby enabling real-time processing of fire videos/streams. An average running time comparison of our proposed method and existing methods for five selected fire videos from the Foggia et al. [14] dataset is shown in Fig. 7. It can be observed that UNet+ResNet50 has the worst average running time of 88.99 ms for processing a single frame. UNet+VGG16 performs comparatively better than UNet+ResNet50 with a 73.81-ms average running time but is dominated by UNet+MobileNetV1, with the second-lowest average running time of 50.67 ms. The proposed method obtains the lowest average running time of 47.13 ms compared to the other CNN-based fire segmentation/localization approaches. Furthermore, to analyze the feasibility of our method in real-time environments, we used the model size, FPS, and mega floating-point operation per second (MFLOPS) parameters. The quantitative results are reported in Table VI, where it can be seen that UNet+ResNet50 performs a lower number of MFLOPS per image than UNet+VGG16; however, in terms of the model size and FPS, UNet+VGG16 is better than UNet+ResNet50. UNet+ MobileNetV1 is comparatively better than UNet+ResNet50, and the model size of CNNFire is better than that of EFDNet. Unlike the other comparative methods, our proposed system maintains a better tradeoff between MFLOPS/image, model size, and FPS by obtaining values of 140, 1.49, and 27, respectively. Its characteristics, including the low computational requirements, low storage requirements, and real-time processing capability, make our method sufficiently efficient to run over edge devices in IoT-enabled environments in real time.

TABLE V
COMPARATIVE RESULTS OF OUR PROPOSED METHOD AND OTHER STATE-OF-THE-ART SEGMENTATION NETWORKS ON TEST SET (Chino et al. [31])

Method	$Pixel_{accuracy}$	$Mean_{accuracy}$	$Mean_{IoU}$	FW_{IoU}
SegNet [32]	84.63	75.92	80.41	87.66
FCN [33]	85.76	75.47	72.65	89.20
PSPNet [34]	88.17	78.62	74.19	89.58
Proposed	**94.54**	**85.27**	**83.35**	**93.96**

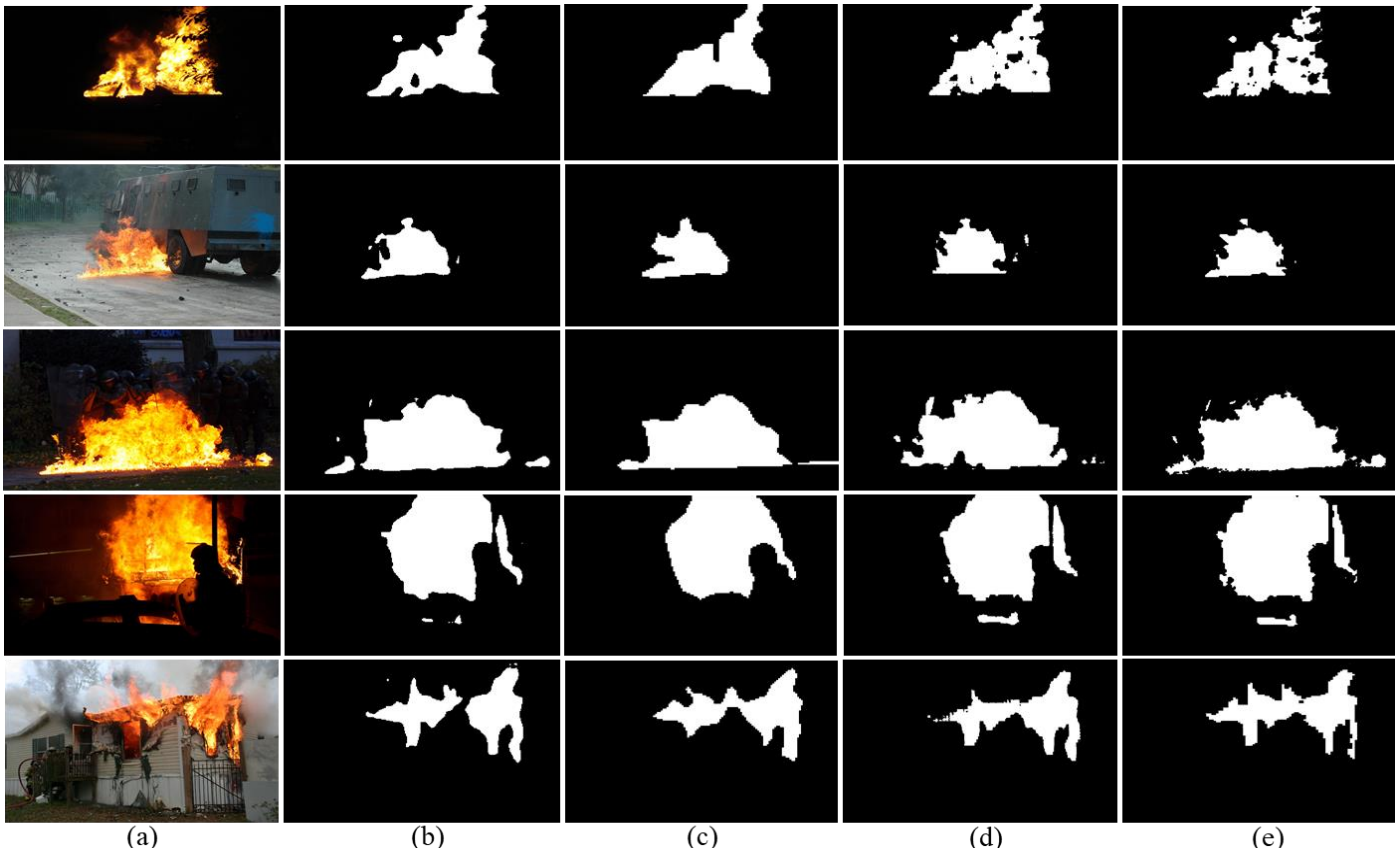

Fig. 2. Visual comparison of segmentation results obtained by our proposed approach and other comparative fire segmentation approaches (a) Input image. (b) SegNet [32], (c) fully convolutional network (FCN) [33], (d) PSPNet [34], and (e) Proposed.

Fig. 3. Visual results obtained by our method for fire-specific region extraction, where the first row represents input images and the second and third rows contain segmentation and fire-specific region extraction results, respectively.

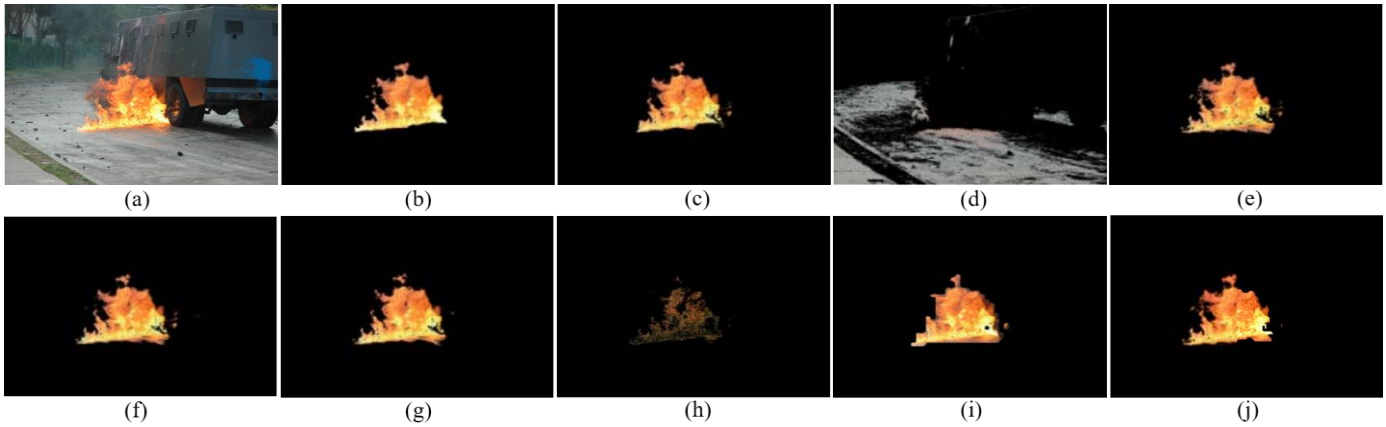

Fig. 4. Visual fire region extraction results obtained by our approach and other comparative fire segmentation approaches. (a) Input image (fire021), (b) Ground truth, (c) Chino [31], (d) Rossi [35], (e) Celik [19], (f) Color classification [31], (g) Rudz [36], (h) Chen [17], (i) CNNFire [26], and (j) Proposed.

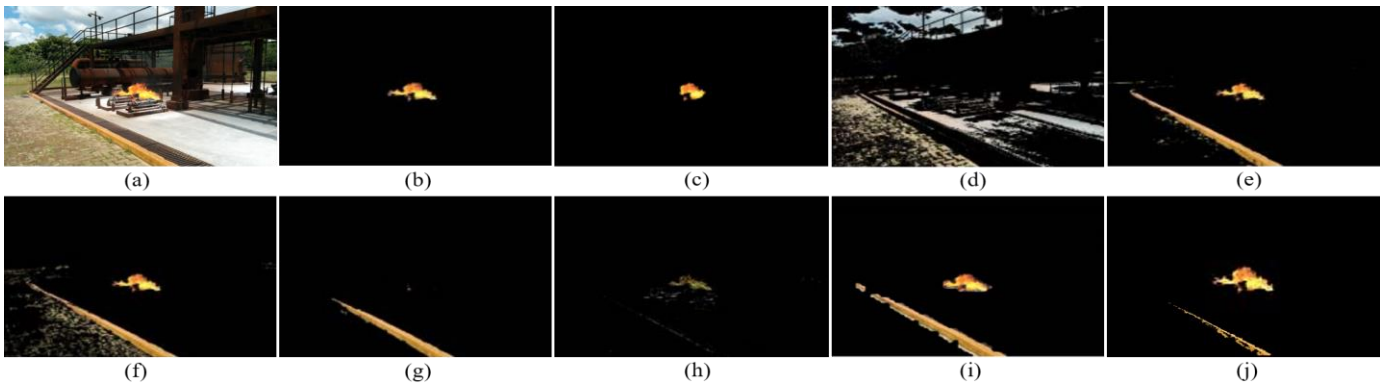

Fig. 5. Visual comparison of fire-specific region extraction results obtained by our method and other fire segmentation approaches. (a) Input image (fire092), (b) Ground truth, (c) Chino [31], (d) Rossi [35], (e) Celik [19], (f) Color classification [31], (g) Rudz [36], (h) Chen [17], (i) CNNFire [26], and (j) Proposed.

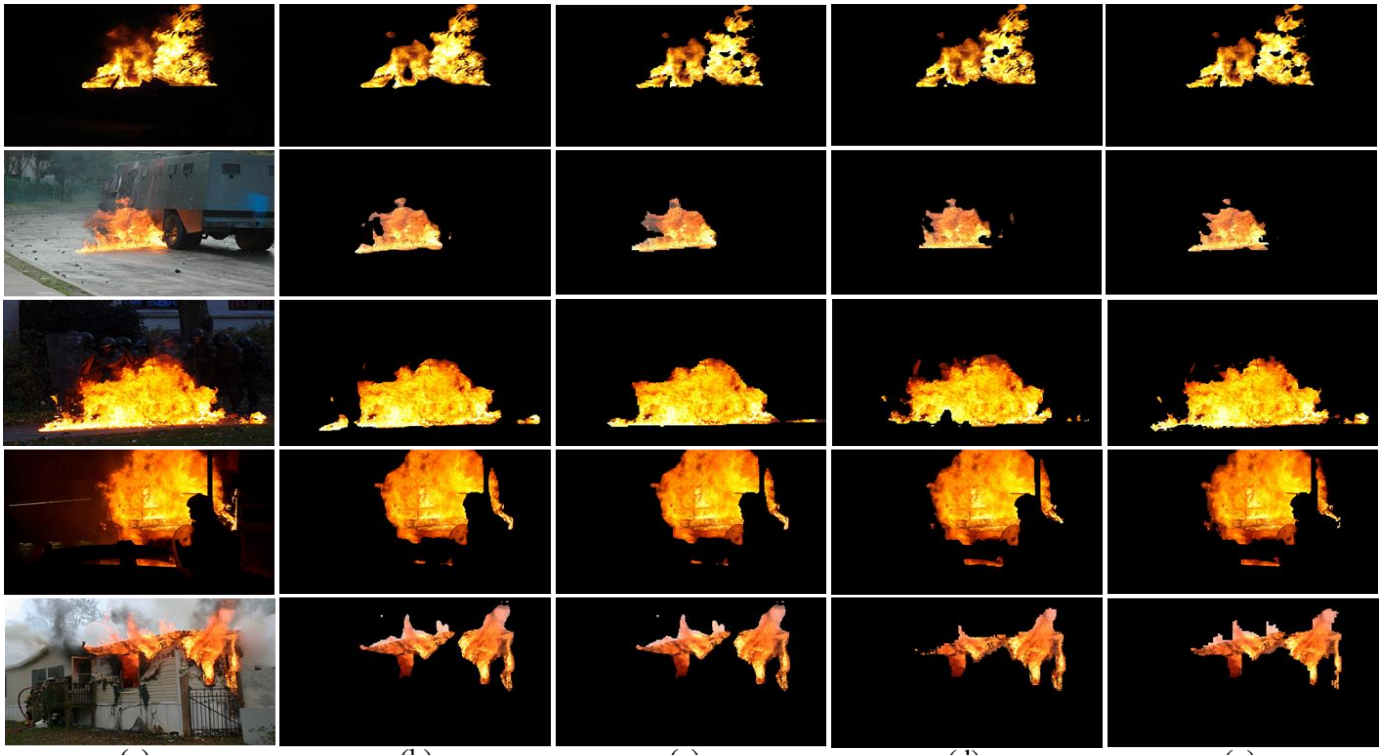

Fig. 6. Visual comparison of fire-specific region extraction results obtained by our method and other state-of-the-art segmentation networks. (a) Input image. (b) SegNet [32], (c) FCN [33], (d) PSPNet [34], and (e) Proposed.

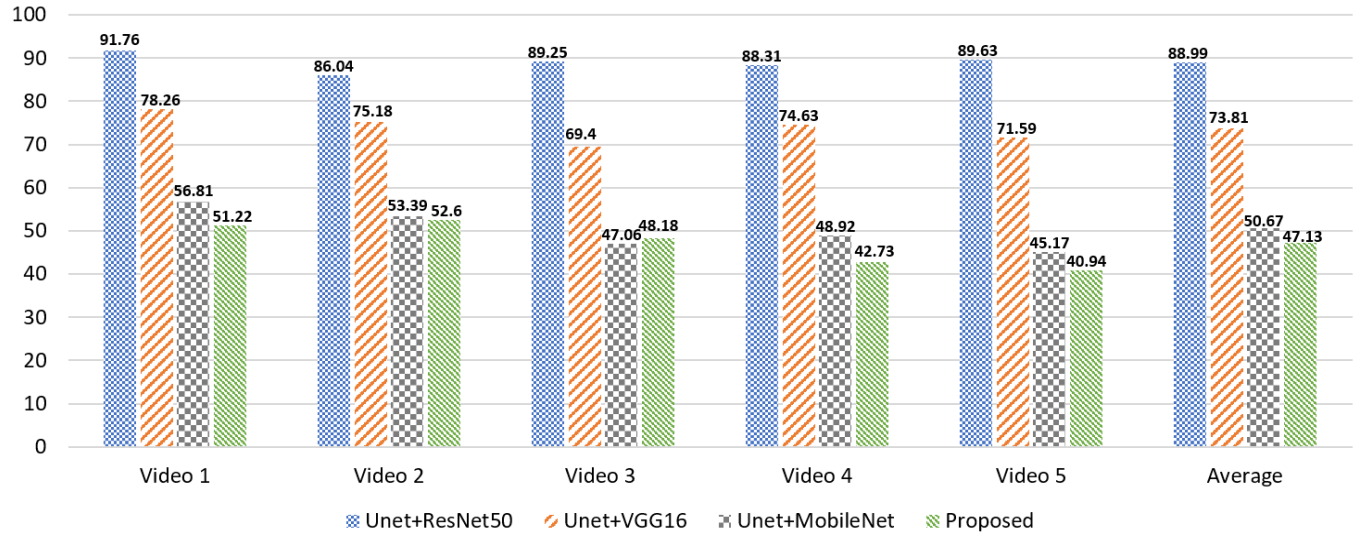

Fig. 7. Average running time per frame in milliseconds taken by our method and other fire segmentation methods using five videos from Foggia et al. [14] dataset.

TABLE VI
COMPARATIVE QUANTITATIVE RESULTS OF OUR METHOD AND OTHER FIRE SEGMENTATION METHODS BASED ON MFLOPS/IMAGE, MODEL SIZE, AND FRAMES PER SECOND (FPS)

Method	MFLOPS/image	Model Size (MB)	FPS
UNet+ResNet50	3860	74.2	21
UNet+VGG16	15300	62.9	22
UNet+MobileNetV1	569	17.2	26
EFDNet [37]	-	4.80	**63**
CNNFire [26]	833	3.06	20
Proposed	**140**	**1.49**	27

IV. CONCLUSION AND FUTURE WORK

Instant fire detection and analysis using computer vision techniques is an effective approach to saving human lives and properties, with recent CNNs exhibiting astonishing performance for vision-based fire detection and localization. However, deploying these networks on edge nodes is a challenging task for researchers focusing on edge devices functional in IoT networks of ITSs because these require real-time processing. These challenges are resolved by proposing a framework suitable for efficient fire detection and segmentation. The proposed CNN is lightweight, with an optimal number of convolutional kernels per layer to reduce the

model size and ensure real-time processing. Extensive experiments over benchmark datasets and our newly created dataset demonstrate that our proposed model can be implemented in real time with trustworthy accuracy, validating its deployment in IoT surveillance environments and ITSs.

Although our current system has the best trade-off between model performance and complexity, it is dominated by EFDNet [33] in terms of the FPS score. Thus, further improvements can be made in terms of the FPS by enhancing the inference time of the proposed system. In addition, our system struggles with fire-like visuals, which can be solved by introducing more effective feature-discrimination techniques inside the network. Our future studies will focus on intelligent decision-sharing in industrial IoT setups by employing 5G technologies for interconnectivity in public places and industries, extending to forests by employing more data using generative networks [38] and efficient and economical hardware [39].

REFERENCES

[1] J. Liu, J. Bai, H. Li, and B. Sun, "Improved LSTM-based Abnormal Stream Data Detection and Correction System for Internet of Things," *IEEE Transactions on Industrial Informatics*, 2021.

[2] F. Kong, J. Li, B. Jiang, H. Wang, and H. Song, "Integrated Generative Model for Industrial Anomaly Detection via Bi-directional LSTM and Attention Mechanism," *IEEE Transactions on Industrial Informatics*, 2021.

[3] A. K. Jain and A. Srivastava, "Privacy-Preserving Efficient Fire Detection System for Indoor Surveillance," *IEEE Transactions on Industrial Informatics*, 2021.

[4] K. Muhammad, A. Ullah, J. Lloret, J. Del Ser, and V. H. C. de Albuquerque, "Deep learning for safe autonomous driving: Current challenges and future directions," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 7, pp. 4316-4336, 2020.

[5] W. Wang, X. Li, L. Xie, H. Lv, and Z. Lv, "Unmanned aircraft system airspace structure and safety measures based on spatial digital twins," *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[6] A. Rego, L. Garcia, S. Sendra, and J. Lloret, "Software Defined Network-based control system for an efficient traffic management for emergency situations in smart cities," *Future Generation Computer Systems*, vol. 88, pp. 243-253, 2018.

[7] I. García-Magariño, R. Lacuesta, M. Rajarajan, and J. Lloret, "Security in networks of unmanned aerial vehicles for surveillance with an agent-based approach inspired by the principles of blockchain," *Ad Hoc Networks*, vol. 86, pp. 72-82, 2019.

[8] Z. Zhou, A. Gaurav, B. B. Gupta, M. D. Lytras, and I. Razzak, "A fine-grained access control and security approach for intelligent vehicular transport in 6g communication system," *IEEE transactions on intelligent transportation systems*, 2021.

[9] H. Fatemidokht, M. K. Rafsanjani, B. B. Gupta, and C.-H. Hsu, "Efficient and secure routing protocol based on artificial intelligence algorithms with UAV-assisted for vehicular ad hoc networks in intelligent transportation systems," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 7, pp. 4757-4769, 2021.

[10] A. Al-Qerem, M. Alauthman, A. Almomani, and B. B. Gupta, "IoT transaction processing through cooperative concurrency control on fog–cloud computing environment," *Soft Computing*, vol. 24, no. 8, pp. 5695-5711, 2020.

[11] X. Xu, J. Zhang, Y. Li, Y. Wang, Y. Yang, and H. T. Shen, "Adversarial Attack Against Urban Scene Segmentation for Autonomous Vehicles," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 6, pp. 4117-4126, 2020.

[12] D. Rashkovetsky, F. Mauracher, M. Langer, and M. Schmitt, "Wildfire Detection from Multisensor Satellite Imagery Using Deep Semantic Segmentation," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 7001-7016, 2021.

[13] J. Zhang, H. Zhu, P. Wang, and X. Ling, "ATT Squeeze U-Net: A Lightweight Network for Forest Fire Detection and Recognition," *IEEE Access*, vol. 9, pp. 10858-10870, 2021.

[14] P. Foggia, A. Saggese, and M. Vento, "Real-time fire detection for video-surveillance applications using a combination of experts based on color, shape, and motion," *IEEE TRANSACTIONS on circuits and systems for video technology*, vol. 25, no. 9, pp. 1545-1556, 2015.

[15] B. U. Töreyin, Y. Dedeoğlu, U. Güdükbay, and A. E. Cetin, "Computer vision based method for real-time fire and flame detection," *Pattern recognition letters*, vol. 27, no. 1, pp. 49-58, 2006.

[16] D. Han and B. Lee, "Development of early tunnel fire detection algorithm using the image processing," in *International Symposium on Visual Computing*, 2006: Springer, pp. 39-48.

[17] T.-H. Chen, P.-H. Wu, and Y.-C. Chiou, "An early fire-detection method based on image processing," in *2004 International Conference on Image Processing, 2004. ICIP'04.*, 2004, vol. 3: IEEE, pp. 1707-1710.

[18] G. Marbach, M. Loepfe, and T. Brupbacher, "An image processing technique for fire detection in video images," *Fire safety journal*, vol. 41, no. 4, pp. 285-289, 2006.

[19] T. Celik and H. Demirel, "Fire detection in video sequences using a generic color model," *Fire safety journal*, vol. 44, no. 2, pp. 147-158, 2009.

[20] A. Rafiee, R. Dianat, M. Jamshidi, R. Tavakoli, and S. Abbaspour, "Fire and smoke detection using wavelet analysis and disorder characteristics," in *2011 3rd International Conference on Computer Research and Development*, 2011, vol. 3: IEEE, pp. 262-265.

[21] Y. H. Habiboğlu, O. Günay, and A. E. Çetin, "Covariance matrix-based fire and flame detection method in video," *Machine Vision and Applications*, vol. 23, no. 6, pp. 1103-1113, 2012.

[22] T. Celik, H. Ozkaramanlt, and H. Demirel, "Fire pixel classification using fuzzy logic and statistical color model," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, 2007, vol. 1: IEEE, pp. I-1205-I-1208.

[23] B. C. Ko, K.-H. Cheong, and J.-Y. Nam, "Fire detection based on vision sensor and support vector machines," *Fire Safety Journal*, vol. 44, no. 3, pp. 322-329, 2009.

[24] J. Sharma, O.-C. Granmo, M. Goodwin, and J. T. Fidje, "Deep convolutional neural networks for fire detection in images," in *International Conference on Engineering Applications of Neural Networks*, 2017: Springer, pp. 183-193.

[25] W. Mao, W. Wang, Z. Dou, and Y. Li, "Fire recognition based on multi-channel convolutional neural network," *Fire technology*, vol. 54, no. 2, pp. 531-554, 2018.

[26] K. Muhammad, J. Ahmad, Z. Lv, P. Bellavista, P. Yang, and S. W. Baik, "Efficient deep CNN-based fire detection and localization in video surveillance applications," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 49, no. 7, pp. 1419-1434, 2018.

[27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.

[29] A. G. Howard *et al.*, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[30] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6848-6856.

[31] D. Y. Chino, L. P. Avalhais, J. F. Rodrigues, and A. J. Traina, "Bowfire: detection of fire in still images by integrating pixel color and texture analysis," in *2015 28th SIBGRAPI Conference on Graphics, Patterns and Images*, 2015: IEEE, pp. 95-102.

[32] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481-2495, 2017.

[33] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431-3440.

[34] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881-2890.
[35] L. Rossi, M. Akhloufi, and Y. Tison, "On the use of stereovision to develop a novel instrumentation system to extract geometric fire fronts characteristics," *Fire Safety Journal*, vol. 46, no. 1-2, pp. 9-20, 2011.
[36] S. Rudz, K. Chetehouna, A. Hafiane, H. Laurent, and O. Séro-Guillaume, "Investigation of a novel image segmentation method dedicated to forest fire applications," *Measurement Science and Technology*, vol. 24, no. 7, p. 075403, 2013.
[37] S. Li, Q. Yan, and P. Liu, "An Efficient Fire Detection Method Based on Multiscale Feature Extraction, Implicit Deep Supervision and Channel Attention Mechanism," *IEEE Transactions on Image Processing*, vol. 29, pp. 8467-8475, 2020.
[38] J. Wang, W. Su, C. Luo, J. Chen, H. Song, and J. Li, "CSG: Classifier-Aware Defense Strategy Based on Compressive Sensing and Generative Networks for Visual Recognition in Autonomous Vehicle Systems," *IEEE Transactions on Intelligent Transportation Systems*, 2022.
[39] M. Sajjad *et al.*, "An efficient and scalable simulation model for autonomous vehicles with economical hardware," *IEEE transactions on intelligent transportation systems*, vol. 22, no. 3, pp. 1718-1732, 2020.

BIOGRAPHIES

Khan Muhammad (S'16–M'18, SM'22) received his Ph.D. in Digital Content from Sejong University, Republic of Korea in February 2019. He was an Assistant Professor in the Department of Software, Sejong University, from March 2019 to February 2022. He is currently the director of the Visual Analytics for Knowledge Laboratory (VIS2KNOW Lab) and an Assistant Professor (Tenure-Track) with the Department of Applied AI, School of Convergence, College of Computing and Informatics, Sungkyunkwan University, Seoul, Republic of Korea. His research interests include intelligent video surveillance, information security, video summarization, and smart cities. He has registered 10 patents and contributed 220+ papers in peer-reviewed journals and conference proceedings in his research areas. His contributions have received 10,000+ citations to date, with an H-index of 56. He is an Associate Editor/Editorial Board Member for more than 14 journals. He is among the most highly cited researchers in 2021, according to the Web of Science.

Hayat Ullah (Student Member, IEEE) received his Bachelor's degree in Computer Science from Islamia College Peshawar, Peshawar, Pakistan, in 2018, and his Master's degree in Computer Science from Sejong University, Seoul, Republic of Korea, in 2021. He is currently pursuing his Ph.D. in Computer Science at Kansas State University, Manhattan, KS, USA. He is also a Research Assistant with the Intelligent Systems, Computer Architecture, Analytics, and Security (ISCAAS) Laboratory, Kansas State University, exclusively working on multi-model human actions modeling and activity recognition. His research interests include image processing, video analytics, and applied computer vision.

Salman Khan (Student Member, IEEE) received his master's degree in Computer Vision from Sejong University, Seoul, Republic of Korea in 2020 with research in vision-based fire/smoke detection. Currently, he is pursuing a Ph.D. degree in Computer Vision (Deep Learning for Modelling Complex Video Activities) from Oxford Brookes University, Oxford, United Kingdom. He has been working as a research assistant at Visual Artificial Intelligence Laboratory (VAIL) since February 2020. His research interests include complex video actions/activities recognition, fire/smoke scene analysis, and medical image analysis.

Mohammad Hijji (Member, IEEE) received his Ph.D. degree in Computing from Coventry University, UK, in July 2017. He is currently a Faculty of Computers and Information Technology (FCIT), University of Tabuk, Saudi Arabia. His research interests include artificial intelligence, cybersecurity, IoT, smart cities, energy optimization, and disaster and emergency management.

Prof. Jaime Lloret (M'07–SM'10) received his B.Sc.+ M.Sc. in Physics in 1997, his B.Sc.+ M.Sc. in electronic Engineering in 2003, and his Ph.D. in telecommunication engineering (Dr. Ing.) in 2006. He is a Cisco Certified Network Professional Instructor, and he has seven Cisco Networking Academy Certifications. He also has the Hewlett-Packard IT Architect Certification. He worked as a network designer and administrator in several enterprises. He is Full Professor at the Polytechnic University of Valencia. He has been the Chair of the Integrated Management Coastal Research Institute (IGIC) since January 2017. He was the founder of the "Communications and Networks" research group of the IGIC, and he is the head (and founder) of the "Active and collaborative techniques and use of technologic resources in the education (EITACURTE)" Innovation Group. He is the director of the University Diploma "Redes y Comunicaciones de Ordenadores," and he has been the director of the University Master "Digital Post Production" for the term of 2012–2016. He was Vice-chair for the Europe/Africa Region of Cognitive Networks Technical Committee (IEEE Communications Society) for the term of 2010–2012, and Vice-chair of the Internet Technical Committee (IEEE Communications Society and Internet

society) for the term of 2011–2013. He was Internet Technical Committee chair (IEEE Communications Society and Internet society) for the term of 2013–2015. He has authored 14 books and has more than 650 research papers published in national and international conferences and international journals (more than 375 with Clarivate Analytics JCR). He has been the co-editor of 54 conference proceedings and guest editor of several international books and journals. He is editor-in-chief of the "Ad Hoc and Sensor Wireless Networks" (with Clarivate Analytics Impact Factor), the international journal "Networks Protocols and Algorithms," and the International Journal of Multimedia Communications. Moreover, he is Associate editor of "Sensors" in the Section Sensor Networks and in Wireless Communications and Mobile Computing, he is an advisory board member of the "International Journal of Distributed Sensor Networks" (both with Clarivate Analytics Impact factor), and he is an IARIA Journals Board Chair (eight Journals). Furthermore, he is (or has been) associate editor of 46 international journals (16 of them with Clarivate Analytics Impact Factor). He has led many local, regional, national, and European projects. He was the chair of the Working Group of the Standard IEEE 1907.1 from 2013–2018. From 2016 to the present, he is the Spanish researcher with highest h-index in the TELECOMMUNICATIONS journal list, according to the Clarivate Analytics Ranking. Moreover, he has been included in the world's top 2% of scientists according to the Stanford University List since 2020. He has been involved in more than 500 program committees at international conferences, and more than 160 organizations and steering committees. He has been a general chair (or co-chair) of 75 international workshops and conferences. He is an ACM Senior, IARIA Fellow, and EAI Fellow.