remote sensing

MDPI

Article

Wildfire Segmentation Using Deep Vision Transformers

Rafik Ghali [1,2], Moulay A. Akhloufi [1,*], Marwa Jmal [3], Wided Souidene Mseddi [2] and Rabah Attia [2]

[1] Perception, Robotics and Intelligent Machines Research Group (PRIME), Department of Computer Science, Université de Moncton, 18 Antonine-Maillet Ave, Moncton, NB E1A 3E9, Canada; rafik.ghali@ept.rnu.tn
[2] SERCOM Laboratory, Ecole Polytechnique de Tunisie, Université de Carthage, La Marsa 77-1054, Tunisia; wided.souidene@ept.rnu.tn (W.S.M.); Rabah.attia@enit.rnu.tn (R.A.)
[3] Telnet Innovation Labs, Telnet Holding, Parc Elghazela des Technologies de la Communication, Ariana 2088, Tunisia; jmal.marwa@gmail.com
* Correspondence: moulay.akhloufi@umoncton.ca

Abstract: In this paper, we address the problem of forest fires' early detection and segmentation in order to predict their spread and help with fire fighting. Techniques based on Convolutional Networks are the most used and have proven to be efficient at solving such a problem. However, they remain limited in modeling the long-range relationship between objects in the image, due to the intrinsic locality of convolution operators. In order to overcome this drawback, Transformers, designed for sequence-to-sequence prediction, have emerged as alternative architectures. They have recently been used to determine the global dependencies between input and output sequences using the self-attention mechanism. In this context, we present in this work the very first study, which explores the potential of vision Transformers in the context of forest fire segmentation. Two vision-based Transformers are used, TransUNet and MedT. Thus, we design two frameworks based on the former image Transformers adapted to our complex, non-structured environment, which we evaluate using varying backbones and we optimize for forest fires' segmentation. Extensive evaluations of both frameworks revealed a performance superior to current methods. The proposed approaches achieved a state-of-the-art performance with an F1-score of 97.7% for TransUNet architecture and 96.0% for MedT architecture. The analysis of the results showed that these models reduce fire pixels mis-classifications thanks to the extraction of both global and local features, which provide finer detection of the fire's shape.

Keywords: forest fire detection; fire segmentation; vision Transformer; TransUNet; MedT; wildfires

Citation: Ghali, R.; Akhloufi, M.A.; Jmal, M.; Souidene Mseddi, W.; Attia, R. Wildfire Segmentation Using Deep Vision Transformers. *Remote Sens.* **2021**, *13*, 3527. https://doi.org/10.3390/rs13173527

Academic Editors: Bardia Yousefi and Rubén Usamentiaga

Received: 11 July 2021
Accepted: 31 August 2021
Published: 5 September 2021

1. Introduction

Statistically, forest fire accidents and arsons result in frightful damage. They are the cause of human and financial losses, the death of animals, and the destruction of woods and houses. Fires also affect 350 million to 450 million hectares every year [1]. Thus, several researchers focused on reducing this negative impact number by developing systems for fire detection at an early stage.

The first existing fire detection systems employed numerous fire sensing technologies such as gas, flame, heat, and smoke detectors. While these systems have managed to detect fire, they have faced some limitations related to coverage areas, false alarms, and slow time response [2]. Fortunately, the aforementioned problems were partially solved by using vision sensors that detect visual features of fire such as shape, color, and dynamic texture of flame.

In recent years, Deep Learning (DL) approaches have been proposed to replace hand-crafted techniques in computer vision applications. They showed impressive results in various tasks such as autonomous vehicles [3], pedestrian detection [4], and video surveillance [5,6]. DL approaches are used in forest fire segmentation tasks to extract the geometrical characteristics of the fire, such as height, width, angle, and so forth. These

models, especially Convolutional Neural Networks (ConvNets), were also successfully employed to predict and detect the boundaries of fire as well as to identify and segment each fire pixel [7,8]. Their impressive results help to develop metrology tools, which can be used in modeling fire preparation as well as providing the necessary inputs to the mathematical propagation models. Nonetheless, due to the intrinsic locality of convolution operators, ConvNets remain limited when modelling the long-range relationship between elements in the image. In order to overcome this problem, vision Transformers, designed for sequence-to-sequence prediction were explored. Based on the self-attention mechanism, they determine the global dependencies between input and output sequences. Transformers witnessed a considerable success in the field of machine translation [9] and Natural Language Processing (NLP) [10–12]. This explains the multiple attempts of researchers to adapt them to several image recognition tasks such as object detection [13], text-image synthesis [14], image super-resolution [15], and medical imaging [12,16,17]. Indeed, these models proved their robustness and accuracy and outperformed the ConvNets in several applications [18].

In this paper, we present the very first study exploring the potential of Transformers in the context of forest fire segmentation using visible spectrum images. Two vision-based Transformers are considered—TransUNet [17] and MedT [16]. In order to exploit their strengths, these models were adapted to our problem. We used the Dice loss [19] as a loss function and the CorsicanFire dataset [20]. We show the feasibility of Transformers for fire segmentation under different settings, varying input size and backbone models that are either pure or hybrid Transformers, which combine a Convolutional Neural Network (CNN) and a Transformer. Then, we evaluate the two Transformers based approaches with state-of-the-art models U-Net [21], a fusion color space method [22], U^2-Net [23], and EfficientSeg [24], which provided excellent results for object segmentation.

The remainder of this work is as follows: Related works are reviewed in Section 2. Then, employed models are presented in Section 3. In Section 4, experimental results are presented. In Section 5, results are discussed. Finally, Section 6 draws the conclusions of this study.

2. Related Work

In this section, we present state-of-the-art fire detection and segmentation methods that we divided into four categories: feature-based, deep learning-based, vision Transformers-based methods, and post-fire mapping based satellite remote sensing imagery. Table 1 summarizes the reviewed fire detection and segmentation methods.

2.1. Feature-Based Fire Detection and Segmentation Methods

Initially, color features were the most used techniques to identify fire from the background and to detect its pixels. Different color spaces were employed to represent fire pixels such as HSI (Hue, Saturation, Intensity) [25], YCbCr [26], and RGB (Red, Green, Blue) [27,28]. However, these methods fail to efficiently detect fire due to their sensitivity to illumination changes and the difficulty of setting the exact range of fire pixels' color. Other methods were proposed to combine several color spaces to detect the fire [22]. The obtained results confirmed that color feature alone is insufficient for this task.

Other studies explored the merge of spatial and temporal features. for instance, Chino et al. [29] proposed a method called "BoWFire" (Best of Both Worlds Fire detection) to identify fire. They combined color and texture features to reduce the false positive rate [30]. Jamali et al. [31] combined the same features to detect fire. They used the HSV (Hue Saturation Value) color space as a color descriptor and LBP-TOP (Local Binary Pattern on Three Orthogonal Planes) to model the texture of fire. Ko et al. [32] also merged multiple features. They used color and motion features to detect fire pixels. Their method proved to be robust to noise [32]. The number of features to combine varied among researchers with the aim of finding the best representation of fire pixels. Foggia et al. [33] combined three features: color, motion, and shape variation, to detect fire in a video surveillance

network. Using large data of real fire videos, the experiment's results proved the reliability of this method in terms of accuracy and false alarms [33]. Recently, Khondaker et al. [34] used the same features to solve this problem. While chromatic segmentation and shape analysis extracted visual features of fire, optical flow analysis determined its motion. This combination of features succeeded in reducing the false alarm rate compared to past state-of-the-art models. Moreover, Emmy et al. [35] developed a method that integrated color, motion, and both static and dynamic textures. The results showed an accurate performance to detect fire with the presence of fire-like colored moving objects. Teng et al. [36] used flame color dispersion and the similarity of consecutive frames in a region to detect fire in a video. In [37], spatial, motion, and temporal features are extracted from different regions to identify smoke and fire in IR (Infra-Red) videos.

Table 1. Fire detection and segmentation methods (Seg. represents the segmentation techniques and Det. represents the detection techniques).

Ref.	Methodology	Object Segmented	Dataset	Results
[25]	HSI color	Flame	Private: 7 videos	Det. accuracy = 96.97 %
[26]	YCbCr color	Flame	Private: 751 images	Flame Det. rate = 99%
[29]	YCbCr color, LBP, KNN, Naive-Bayes	Flame	BowFire: 226 images	Det. precision = 80%, Det. recall = 65%
[38]	AlexNet, simple CNN	Flame	Private: 560 images	Seg. accuracy = 94.76%, Seg. F1-score = 90.31%
[35]	YCbCr color, 2D DWT, 3D DWT, ELM classifier	Flame	VisiFire, VisionFire: 85 videos	Det. accuracy = 95.65%
[21]	U-Net	Flame	CorsicanFire: 419 images	Seg. accuracy = 97.09%, Seg. F1-score = 91%
[22]	RGB, HSV, HSL, HWB color	Flame	CorsicanFire: 500 images	Det. accuracy = 93%, Det. F1-score = 79%
[39]	RGB, YCbCr, optical flow	Flame	Private: 30 videos	Det. accuracy = 96.09%
[37]	SIFT Flow, optical flow, GMM, K-means, MRF, GMRF	Flame, smoke	Private: 10 IR videos	Seg. accuracy = 95.39%
[40]	Lightweight CNN based on the SqueezeNet	Flame	BowFire, MIVIA: 68457 images	Det. accuracy = 94.50%, Det. F1-score = 91%
[31]	HSV color, LBP-TOP	Flame	Firesense: 27 videos	Det. accuracy = 98.5%
[34]	YUV color, LKT optical flow	Flame	Zenodo, Mivia, web	Det. accuracy = 97.2%
[41]	DeepLab V3	Flame	FlickFire, Firesense, Visifire, private: 7561 images	Seg. accuracy = 98.78%, Seg. mean IoU = 70.51%
[42]	DeepLab V3+	Flame	CorsicanFire: 1775 images	Seg. accuracy = 97.67%, Seg. mean IoU = 89.64%
[43]	wUUNet	Flame	Private: 6250 images	Seg. accuracy = 95.34%
[44]	Yolov3, R-FCN, Faster R-CNN, SSD	Flame, smoke	Public smoke/fire: 29180 images	Det. accuracy = 83.7%
[45]	Yolov5, efficient-Det, efficientNet	Flame	BowFire, VisiFire, FD, ForestryImages: 10581 images	Det. accuracy = 99.4%

Dang-Ngoc et al. [39] presented a two stages method to detect and segment fires in aerial videos. At first, fire color information is extracted using multi-color spaces, RGB, HSI, and YCbCr. Then, optical flow is used to detect forest fire motion features from a

dynamic background. Using a large aerial video dataset, which includes 9070 frames, this method showed high performances and low false alarm rates in detecting forest fires.

Khan et al. [46] also presented a real time fire detection method based on video processing. This method integrated color information and spatiotemporal features to distinguish between fire areas and moving backgrounds. Experimental results showed a high accuracy of 97.7% and outperformed some previous state-of-the-art models.

While the aforementioned methods showed their ability to detect fires, they still suffer from false alarms, mainly in the case of small fire portions and fire-like objects that get confused with a real fire.

2.2. Deep Learning Based Fire Detection and Segmentation Methods

In the last decade, Deep Learning methods witnessed exploding success in many computer vision tasks such as object detection [47], satellite image analysis [48,49], road monitoring [50], medical diagnosis [51], and object segmentation [52]. This is due mainly to the rich feature map extracted using convolutional layers. As for the task of object segmentation, DL methods were able to efficiently classify each pixel in the whole image and determine the exact shape of the objects. They exceeded the performance of classical machine learning models [52].

Through the years, many studies addressed the problem of fire detection based on deep learning techniques. For instance, Gonzalez et al. [38] proposed the SFEwAN-SD (Simple Feature Extraction with FCN AlexNet, Single Deconvolution) for the UAV (Unmanned Aerial Vehicle) fire monitoring system. Two convolutional neural networks, AlexNet [53] and a simple CNN, were employed to find fire features. While AlexNet extracts the shape and texture of input images to identify the fire regions, the second CNN, composed from numerous convolutional layers of size 3*3 each, and using the ReLU (Rectified-Linear Unit) activation function, handles the color and texture features of fire. This method outperformed state-of-the-art methods, at that time, in terms of processing time and accuracy. Muhammad et al. [40] presented an energy-friendly CNN method based on the SqueezeNet [54] model for fire localization and detection in closed-circuit television (CCTV) networks. This model was selected thanks to its portability to be deployed on FPGAs (Field-Programmable Gate Arrays). Using two benchmarks containing videos or images of fire and fire-like objects, a reasonable performance was obtained.

Recently, Encoder–Decoder architectures become one of the most used models for semantic segmentation. The encoder first extracts a high dimensional feature map of input images. It consists of convolutional and pooling layers. Then, the decoder decodes generated features and determines the mask of objects. It includes Upsampling blocks that contain unpooling and/or deconvolution (transpose convolution) layers [55]. Thanks to their performance in various tasks such as medical image segmentation, plenty of works were based on the Encoder–Decoder structure to segment objects. As an example, U-Net [56], the Encoder–Decoder applied for segmenting biomedical images, won the ISBI cell tracking challenge in 2015. While, the encoder consists of a 3*3 convolutional layer, ReLU activation function and a max-pooling layer, the decoder contains Upsampling layer, 3*3 convolutional layer, ReLU activation function, and up-convolution layer. Finally, a 1*1 convolutional layer produced the mask of the input. In 2018, Akhloufi et al. [21] integrated this Encoder–Decoder into a Deep-Fire model due to its great performance when applied to the task of forest fire detection and segmentation. Using a small data (419 images of the CorsicanFire dataset) as learning data and Dice loss as a loss function, this model reached an F1-score value of 97.09% and 91% for training and testing, respectively. It also showed its excellent efficiency in segmenting wildfire in a non-structured environment, which is affected by uncontrolled forest fires [21].

Bochkov et al. [43] proposed a novel model, wUUNet (wide-UUNet concatenative), to detect fire regions and flame areas. This method includes a UUNet model as a modernization of U-Net, with maximum skip connections number between encoder and decoder. The UUNet model consists of two UNet networks. The first detects fire areas as binary

segmentation. The second identifies fire colors (orange, red, yellow) as a multiclass segmentation, using input images and the output of first U-Net. Using VGG16 [57] as a backbone collected learning dataset, cross-entropy and soft-Jaccard as a loss function, this model outperformed the UNet by 3% and 2% in the case of multiclass segmentation and binary segmentation, respectively.

In order to reduce the excessive downsampling as a result of pooling operations, Chen et al. [58] proposed a DeepLab model, which employs the atrous convolution (dilated convolution), fully connected Conditional Random Field (CRF), and atrous spatial pyramid pooling (ASPP). The atrous convolution controls the resolution of feature maps and extracts objects at multiple scales. The ASPP computes multiple scales feature maps. The CRF overcomes the invariance of the combination of downsampling and pooling operation, and improves localization performance [58]. DeepLabV3 [59] outperformed DeepLab model. It applies four parallel dilated convolutions with various atrous rate to extract multi-scale features [59]. Recently, DeepLabV3+ [60] applies an Encoder–Decoder structure. It employs the atrous separable convolution, which consists of depth-wise convolution followed point-wise convolution (1*1 convolution). This model reached a better performance than DeepLabV3 [60]. Two models of the DeepLab series, DeepLabV3 [41] and DeepLabV3+ [42], were explored to detect and segment fire pixels. According to the experiments, the two detectors showed efficient results thanks to the multi-scale and rich feature map. These networks proved to be suitable for fire segmentation. The first model was trained using a learning dataset, which contains a 1775 fire images and non-fire images from the SUN397 dataset [61], data augmentation techniques such as rotation, Hue, saturation, brightness, and flip, and three pre-trained backbones, ResNet152, ResNet101, and ResNet50. This model reached an accuracy of 98.78% and a mean IoU (Intersection Over Union) of 70.51% [41]. The second was tested using three loss functions, which are Tversky loss, Dice loss, and Cross-entropy loss, and the CorsicanFire dataset, which contains RGB and IR images. Experiment results showed that DeepLabV3+ with Dice or Tversky loss function and ResNet50 presents the best mean boundary F1 (BF) contour matching score value of 92.23% [42].

Li et al. [44] presented novel image fire detection methods based on the advanced object detection CNN models. Four CNNs, R-FCN [62], SSD [63], Yolov3 [64], and Faster R-CNN [65] are used to detect and localize fire and smoke. Using a large dataset, which contains 13,400 fire images (7742 smoke and 9695 fire objects) and 15 780 non-fire images, Yolov3 achieved the highest accuracy with 83.7% and the best detection time with 28 FPS compared to other models. It proved its ability to detect and localize fire and smoke with lower false alarms (fire/smoke-like objects) [44].

Xu et al. [45] proposed CNN techniques to detect and localize forest fire. Their approach integrated two object detectors, Yolov5 [66] and EfficientDet [67], and a classifier based on EfficientNet [68]. Using a dataset containing 2976 forest images and 7605 non-fire images, the proposed method achieved a higher detection accuracy with 99.4%. It outperformed some state-of-the-art models such as Yolov3, EffcientDet, Yolov5 and SSD. It also showed its ability to detect forest fires in different scenarios and reduce the false-positive rate [45].

2.3. Methods Based on Transformers

During the last decade, researchers have developed a plethora of ConvNets models for image segmentation. These models rely on the convolution operator, which extracts the local features of the image. However, they are still limited at modeling global context and have the limitation of an expensive computational cost.

Recently, Transformers were proposed to avoid this limitation and to model the long-range interactions between input patches using self-attention mechanism, which is at the core of Transformers. This mechanism models the relevance of each input element to other elements. It determines the global contextual information of each item by capturing its interaction amongst all items. At first, Transformers showed great performances in

NLP tasks. Then, Transformers were applied in computer vision tasks such as video processing [69], image super-resolution [15], object detection [13] and segmentation [70], and image classification [71] thanks to their excellent performance.

Various Transformers were developed. They are divided into pure Transformers and hybrid Transformers, which combine a ConvNets and a Transformer. The size of learning data is the main element, which affects the performance of Transformers. Vision Transformer (ViT) [72] showed an excellent performance compared to state-of-the-art ConvNets. However, it still depends on pre-trained models in very large-scale dataset like JFT-300M [73], which contains 300 million labelled images. DeiT (Data-efficient Image Transformers) [74] showed the ability of Transformers to perform well using only a mid-sized dataset. More recently, Medical Transformer (MedT) [16] also reached an excellent performance to segment objects from scratch with no need for pre-training models.

2.4. Post-Fire Mapping Based Satellite Remote Sensing Imagery

Thanks to its ability to monitor large areas, satellite remote sensing imagery is used to determine and report post-fires, which defines the perimeter of fire areas and the severity of the damage [8]. The damage level is defined by five levels, which are unburned areas with no damages, burned areas with negligible damage, burned areas with moderate damage, burned areas with high damage, and burned areas destroyed. Numerous solutions were developed to map and date the burned areas such as the FireCCI51 [75], the Global Fire Emission Database [76], and the MCD64A1 Collection 6 [77].

Various techniques are used to compute the degree of change in soil and vegetation caused by wildfire, for example, the Composite Burned Index (CBI) [78], the Normalized Burn Ratio (NBR) [79], the delta Normalized Burn Ratio (dNBR) [80], and Normalized Difference Vegetation Index (NDVI) [81]. These techniques compute the features from pre and post-fire acquisitions to determine the severity of burned areas. Machine learning methods are also adopted to predict the damage severity using pre and post-wildfire satellite imagery such as Support Vector Regressor (SVR) [82] and Random Forest [83].

Zanetti et al. [84] proposed a multitemporal automatic and unsupervised method to detect burned areas using image time series acquired by Sentinel-2 and Landsat-8. At first, the normalized fire index method detected the candidate burned areas in each input image. Then, the temporal harmonization method was used to confirm the burned areas and reduce the false alarms. Experimental results proved the ability of this method to correctly detect the burned areas and reduce false alarms [84].

Recently, Deep Learning models were used in various remote sensing applications such as land cover classification [85,86], reconstruction of missing data [87], segmentation of small objects [88], and identification of clouds [89]. Motivated by their high performance, numerous contributions adopted deep learning to compute the severity of the damage. Pinto et al. [90] proposed a deep learning method, which integrates CNNs and the Long Short-Term Memory (LSTM) [91] method with an architecture based on U-Net. This method showed excellent results in dating and mapping burned areas. It also proved its ability to overcome the limitations of traditional methodologies such as the filtering of lower quality and no shadows and clouds mask [90]. Farasin et al. [92] also developed a novel deep learning model, Double-Step U-Net, to estimate the wildfire damage severity from the sentinel-2 satellite. Two subtasks, Binary Classification U-Net and Regression U-Net, are proposed. The first task identifies affected areas by fire. The second provides the severity level estimation. This model is validated across five European regions. It also showed a comparable result to the threshold of the delta Normalized Burn Ratio method [92].

Rahmatov et al. [93] proposed a fire prevention method over large forest areas. At first, a CNN detected the fire and its geographic location. Then, an intelligent agent determined the optimal path to reach the forest fire's location at the optimal time. The analysis of the results showed the efficiency of this method to help in controlling disaster situations and reduce the damage caused by the fire [93].

Khennou et al. [94] also proposed a forest fire spread prediction model based on U-Net called FU-Netcast. This method predicted the next burned forest areas over a time frame of 24 h using Landsat images. Using Digital Elevation Model maps, weather data, satellite images, and 120 consecutive wildfire perimeters, the proposed approach achieved an accuracy of 92.73% and proved its efficiency in forecasting the spread of fires [94].

3. Methods

In this section, we describe the models employed for our wildfire segmentation task and we also present the training dataset, and the evaluation metrics used in this work.

Vision transformers, designed for sequence-to-sequence prediction, have emerged as alternative architectures to ConvNet models, which still show some limitations in explicitly modeling long-range dependency. In the previous section, we reviewed some of the most used visual transformers. Two of them, TransUNet [17] and MedT [16], used for medical imaging, have shown interesting performances in medical image segmentation. In this section, we will analyse these two vision Transformers since we will explore their performances when applied to our problem. We propose, here, the very first study using visual Transformers in forest fires segmentation task.

Furthermore, we present U^2-Net [23] and EfficientSeg [24], two models used to carry a comparative study with the two aforementioned Transformers. It is worth mentioning that, nowadays, both U^2-Net and EfficientSeg achieved excellent performances and outperformed the state-of-the-art object segmentation methods based on Convolutional Neural Networks [23,24]. These models are fed with RGB images, which are used to segment fire pixels and detect the exact fire areas' shape. The result is a binary mask, which defines the segmented forest fire area in the input image.

3.1. TransUNet

TransUNet [17] is a hybrid CNN-Transformer model, which integrates both the Transformer and U-Net network. It adopted a high resolution of local features extracted by a CNN and the global information encoded by Transformers.

This model employs a Hybrid CNN-Transformer as an encoder. The CNN model is the first to extract the features. Then, patch embedding is employed to encode the positional information. The Transformer encoder contains twelve Transformer layers, which include a normalization layer, a Multi-Layer Perceptron (MLP), and a Multihead Self-Attention (MSA). The skip-connections from the encoder and the output of the Transformer are feeding the decoder, which consists of multiple 3*3 convolutional layers, upsamling operator, and ReLU activations. All feature extractors were pretrained on a very large dataset, namely, ImageNet [95], which provides numerous images [17]. Figure 1 illustrates the TransUNet architecture.

3.2. Medical Transformer (MedT)

The majority of vision Transformers require large-scale learning data for a better performance. MedT [16] is proposed to avoid this problem. This transformer adopts two concepts, Local-Global (LoGo) learning strategy and gated axial transformer layer. LoGo methodology is made up of two branches, global branch and local branch. The first branch uses the original resolution of the input image. The second employs the patches of the input image.

Figure 2 shows the MedT architecture. At first, the feature map is extracted from the input images using two convolutional blocks. Each block contains three convolutional layers, batch normalization layer, and ReLU activation. Then, the extracted feature map feeds local and global branches. The two branches include, respectively, five Encoder–Decoder blocks and two Encoder–Decoder blocks, connected by skip connection. The encoder consists of 1*1 convolutional layer, normalization layer, and two layers of multi-head attention, which operate on both width and height axis. The decoder also contains convolutional layer, batch normalization layer, and ReLU activation.

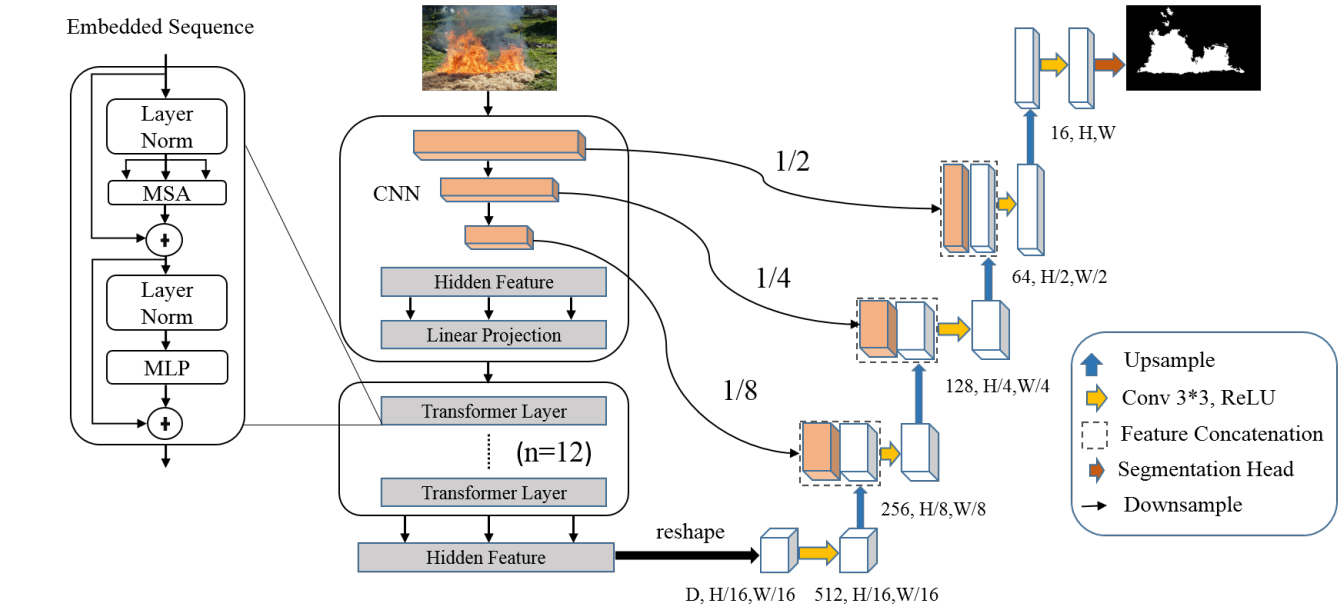

Figure 1. The proposed TransUNet architecture.

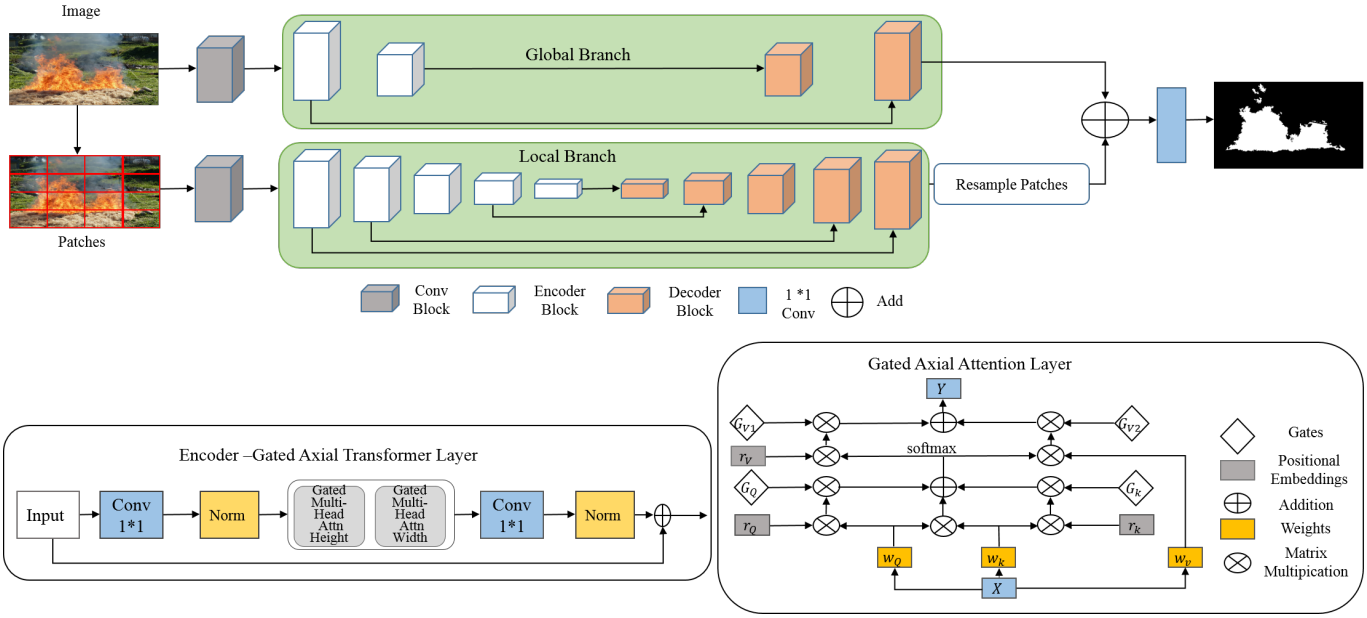

Figure 2. The proposed MedT architecture.

3.3. U²Net Architecture

U²-Net [23] is a deep network architecture. It is a two level nested U-structure, which employs RSU (Residual U-blocks) to extract multi scale information. The RSU block consists of a convolution layer, residual connection, U-Net structure, which contains a convolutional layer, a batch normalisation layer, and a ReLU activation function to extract the multi-scale features. U²-Net consists of six encoders, five decoders, and a saliency map fusion block. Each encoder and decoder are filled by the RSU. The saliency map block contains the Sigmoid function and 3*3 convolution layer [23]. Figure 3 presents the U²-Net architecture.

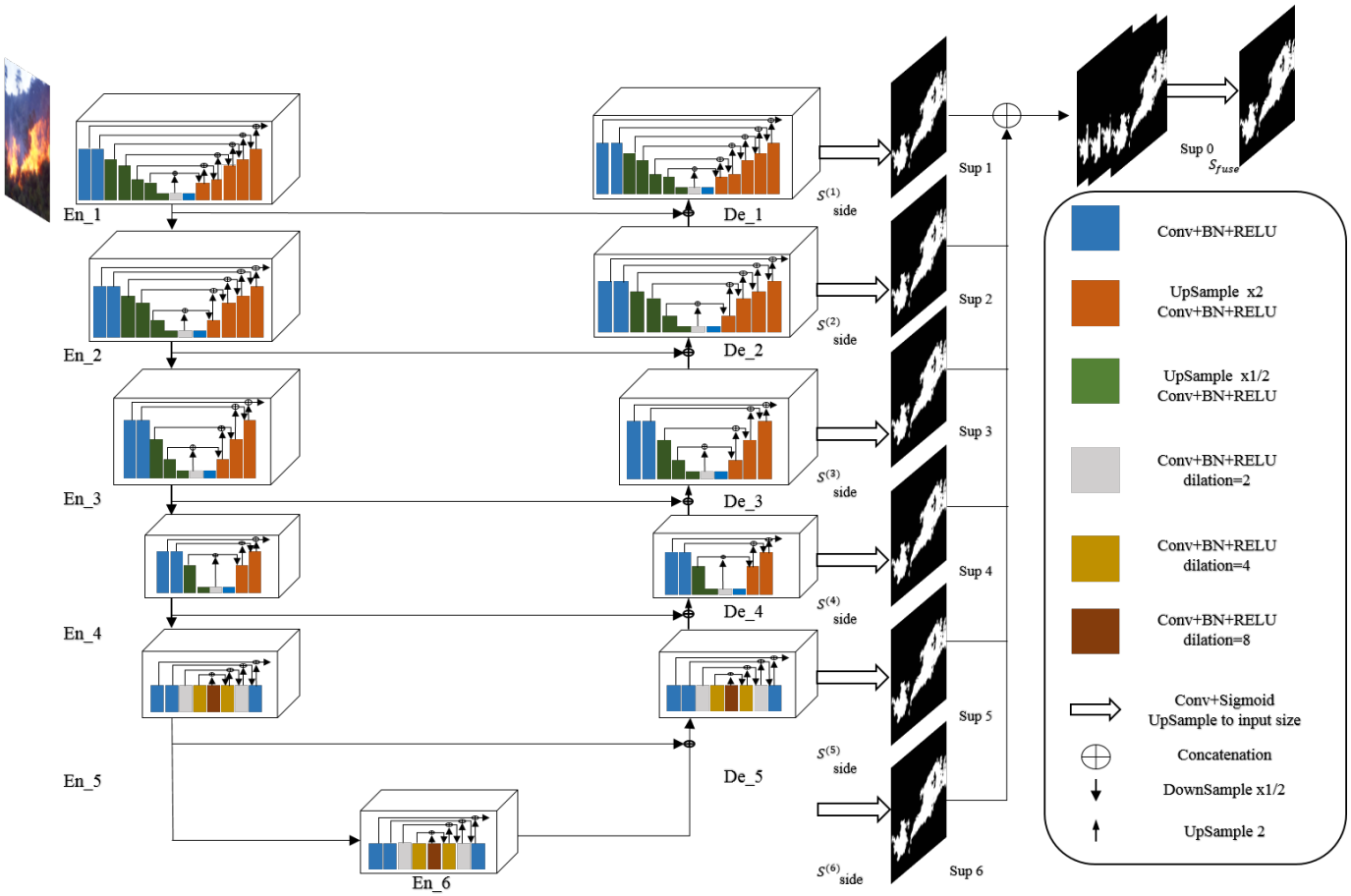

Figure 3. The proposed U^2-Net architecture.

3.4. EfficientSeg

EfficientSeg [24] is a modified and scalable model based on U-Net structure and MobileNetV3 [96] blocks. It has an Encoder–Decoder structure. It proved its great performance and outperformed U-Net model [24].

As presented in Figure 4, the EfficientSeg architecture consists of different blocks, 3*3 and 1*1 convolution layers followed by a batch normalization layer and a ReLU activation function, Inverted Residual Blocks, four shortcut connections between encoder and decoder, and Downsampling and Upsampling operations.

3.5. Dataset

For the fire segmentation problem, there is still a limited number of datasets available. We use the CorsicanFire dataset [20] to train and evaluate the proposed models. The CorsicanFire dataset contains RGB and NIR (near-infrared) images. In this dataset, NIR images are captured with a longer integration/exposure time, which increases the brightness of the fire area and makes the segmentation easier with simple image processing techniques. Still, the obtained shape is less precise than with visible spectrum images, since it integrates the fire areas during the full exposure time (the fire shape covers a larger area in the image). In this work, we are interested in RGB images which provide a large number of images in the dataset and are widely used in the context of capturing fires. Our dataset consists of 1135 images and their corresponding masks. It describes the visual features of fire pixels such as color (red, orange, and white-yellow), the different weather conditions, the brightness, the distance to the fire, and the presence of smoke. Figure 5 depicts samples of the CorsicanFire dataset and their corresponding masks.

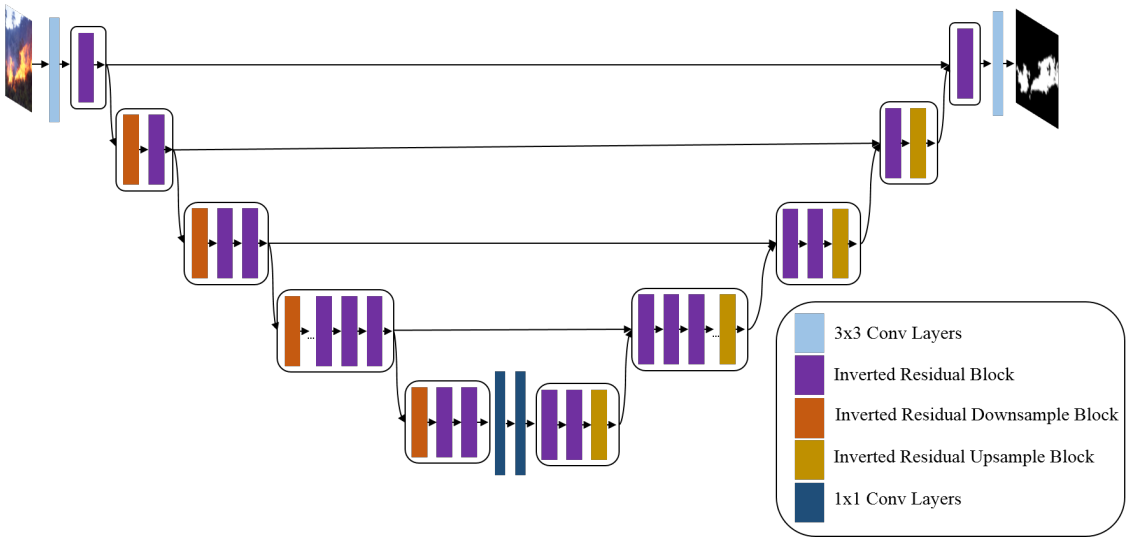

Figure 4. The proposed EfficientSeg architecture.

Figure 5. Examples from the CorsicanFire dataset. From (**top**) to (**bottom**): RGB images and their corresponding masks.

3.6. Evaluation Metrics

We evaluate the proposed approaches using *F*1-*score* and inference time.

- *F*1-*score* combines recall and precision metrics to calculate the performance of the model (as given by Equation (1)).

$$F1 - Score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN}, \quad (3)$$

where FN is the false negative rate, FP is the false positive rate, and TP is the true positive rate.

- Inference time is defined as the average segmentation's time using our test data.

4. Experiments and Results

In this section we present the experimental settings, illustrate the implementation details and present the experimental results.

4.1. Experimental Results

The proposed Transformers were developed using Pytorch [97]. Training and testing was performed on a machine with NVIDIA Geforce RTX 2080Ti GPU. The training data were split as follows: 815 images for training, 111 images for validation, and 209 images for testing. For all experiments, we adopted Dice loss [19], which maximizes the overlap between the predicted images set and the ground truth image set as a loss function, patch size p value of 16 and simple data augmentation techniques that are a rotation of 20 degrees and a horizontal flip.

Dice loss computes the similarity between predicted image and ground truth image (given by Equation (4)).

$$Dice - loss = 1 - \frac{2|Y \bigcap X|}{|Y| + |X|}, \tag{4}$$

where X is the predicted image, Y is the ground truth input, and $\bigcap$ is the intersection of the ground truth Y and the predicted mask X.

The TransUNet Transformer was tested using a learning rate value of 10^{-3}, input resolution of 224*224 and 512*512, and two backbones, which were pure transformer ViT and Hybrid CNN-Transformer ResNet50-ViT, pretrained on the large-scale dataset ImageNet [95].

The MedT Transformer was tested from scratch (no pretraining) using a learning rate value of 10^{-2}, Hybrid ConvNet-Transformer as a backbone, and input resolution images of 224*224 and 256*256.

To evaluate the performance of the proposed Transformers, we first compared the F1-score values between the two Transformers, TransUNet and MedT, by varying backbones, input resolution, and the size of the dataset. Then, we compared their results with various models: U^2-Net, EfficientSeg, U-Net [21], and a color space fusion method [22].

4.1.1. Quantitative Results

The performances of both Transformers (TransUNet and MedT) are reported in Table 2. We can see that the Transformers, TransUNet and MedT, reach F1-score values of 97.7% and 96.0% respectively, and prove their accurate and robust performance to detect and segment fire pixels. These models segment wildfire pixels well, thanks to the use of both global and local features. They provide finer details of the fire.

Table 2. Quantitative results of TransUNet and MedT on CorsicanFire dataset

Model	Backbone	Input Resolution	F1-Score (%)
TransUNet	Res50-ViT	224*224	97.5
TransUNet	Res50-ViT	512*512	97.7
TransUNet	ViT	224*224	94.1
TransUNet	ViT	512*512	94.8
MedT	simple CNN-Transformer	224*224	95.5
MedT	simple CNN-Transformer	256*256	96.0

Transformers, TransUNet-Res50-ViT and MedT, with a hybrid backbone, extract more details from the input image due to spatial and local information extracted by ConvNets. These models show a better performance than TransUNet-ViT with a pure Transformer as backbone.

Changing the input resolution from 224*224 to 256*256 or 512*512 shows some improvements in the F1-score value (between 0.2% and 0.7%), due to the larger num-

ber of input patches. However, larger computational capacity and time are required during training.

Using a LoGo methodology for learning, MedT extracts high level and finer features. This Transformer presents a great performance and proves its efficiency in segmenting fire pixels without pre-training.

TransUNet with a hybrid backbone obtains a better performance than MedT and proves its excellent ability to localize and segment forest fire pixels. However, it still depends on a pre-trained backbone on a very large dataset.

Table 3 presents a comparative analysis of TransUNet, MedT, U-Net, Color fusion method, U^2-Net, and EfficientSeg, in terms of F1-score, using the CorsicanFire dataset. We can see that the results of both Transformers, TransUNet and MedT, are better than the performance of deep CNN models (U-Net, U^2-Net, and EfficientSeg) and classical machine learning models (color fusion method).

Table 3. Comparative analysis of TransUNet, MedT, U-Net, Color fusion method, U^2-Net and EfficientSeg using CorsicanFire dataset with image size of 224*224.

Model	Backbone	Learning Data	F1-Score (%)	Inference Time (s)
TransUNet	Res50-ViT	1135 images	97.5	1.20
TransUNet	ViT	1135 images	94.1	0.13
MedT	simple CNN-Transformer	1135 images	95.5	2.72
U-Net [21]	——	419 images	91.0	—
Color fusion method [22]	——	500 images	79.0	—
U-Net	——	1135 images	92.0	0.02
U^2-Net	——	1135 images	82.93	1.41
EfficientSeg	——	1135 images	94.27	2.00

TransUNet-Res50-ViT and MedT obtain the best F1-score values of 97.5% and 95.5%, respectively. They outperform the color fusion method and convolutional networks, U-Net, EfficientSeg, and U^2-Net. EfficientSeg shows a great result compared to the color fusion method, U-Net, and U^2-Net. However, it is still more limited to modeling global information than TransUNet and MedT. It also obtains an inference time of 2 s, which is higher than the inference time of TransUNet. U-Net with 1135 images, which outperforms recently developed models, U-Net with 419 images, the color fusion method and U^2-Net, thanks to its diverse feature maps. It also shows the best value of inference time, which is 0.02 s. U^2-Net obtains an F1-score value of 82.93%, which is better than the color fusion method. However, it has a lower performance compared to ConvNets models (U-Net and EfficientSeg) and Transformers (TransUNet and MedT).

4.1.2. Qualitative Results

Similar to the quantitative results presented in Table 3, we can see in Figure 6, that TransUNet, with Res50-ViT as a backbone, segments fire pixels even better than manual annotation. This Transformer can correctly distinguish between fire and background under different conditions such as the presence of smoke, different weather conditions, and various brightnesses of the environment. It also proves its better ability to identify small fire areas and detect the precise shape of a fire.

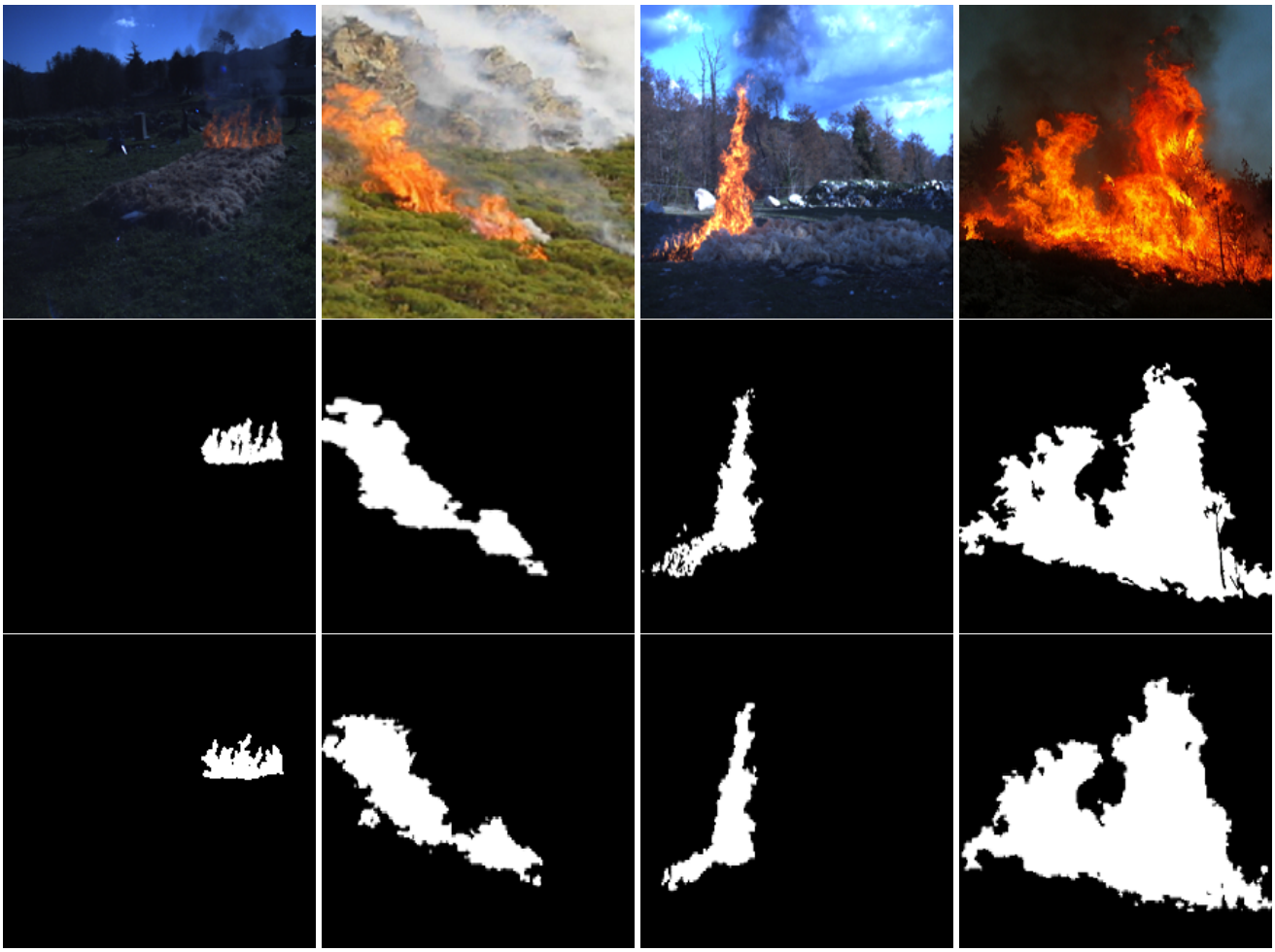

Figure 6. Results of TransUNet-Res50-ViT. From (**top**) to (**bottom**): RGB images, their corresponding mask, and the predicted images of TransUNet-Res50-ViT.

Figure 7 depicts the results of TransUNet with pure Transformer ViT as a backbone. We can see that this Transformer correctly segments fire pixels under various conditions. However, it misclassifies the fire border pixels (highlighted in red box) and misdetects the exact shape of the fire area.

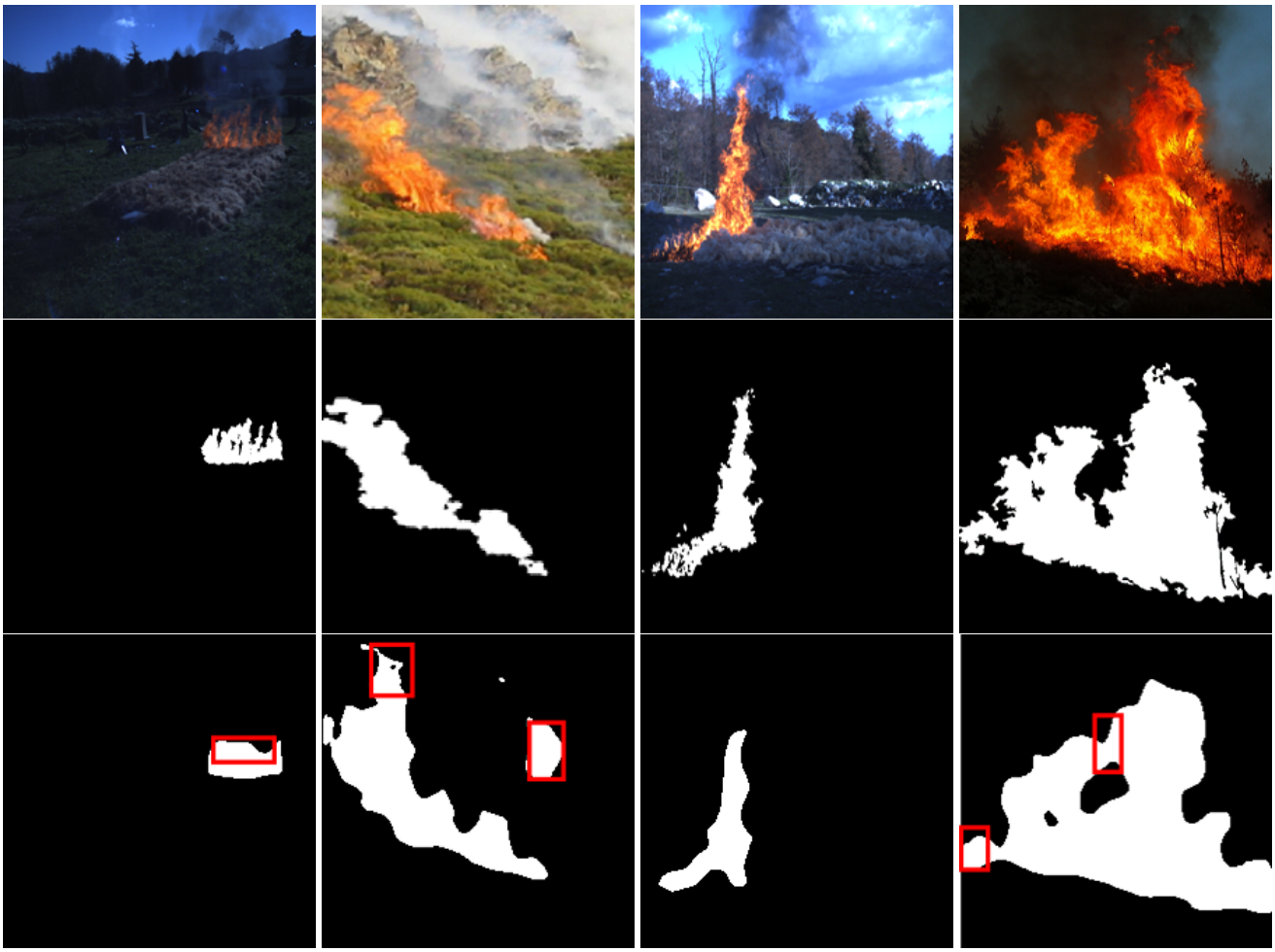

Figure 7. Results of TransUNet-ViT. From (**top**) to (**bottom**): RGB images, their corresponding mask, and the predicted images of TransUNet-ViT.

Figure 8 shows examples of the segmentation of MedT. Similar to TransUNet-Res50-ViT, we can see that MedT proves its efficiency in segmenting fire pixels and detects the precise shape of the fire. However, it still misclassifies some small areas of fire (highlighted in red box).

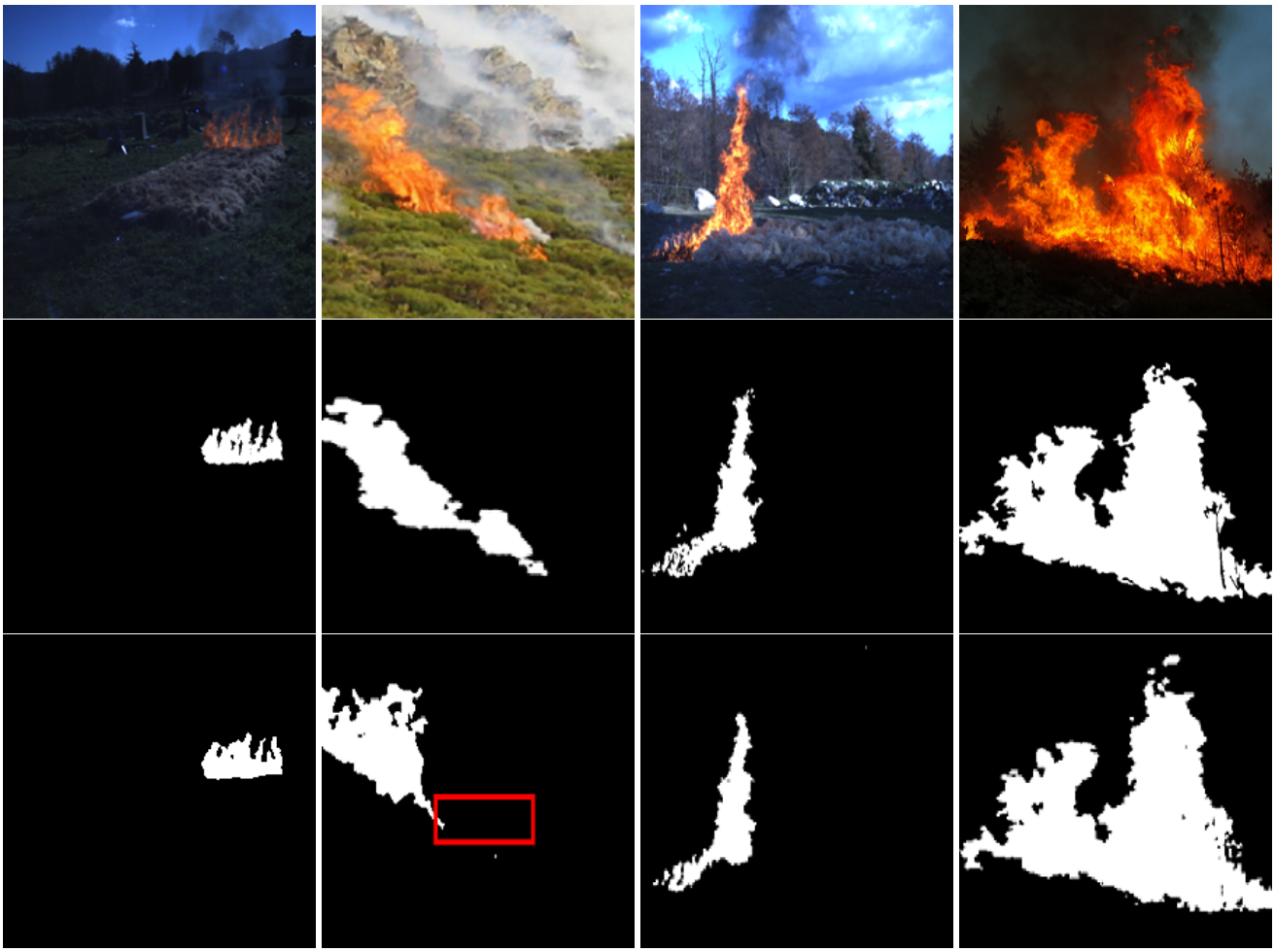

Figure 8. Results of MedT. From (**top**) to (**bottom**): RGB images, their corresponding mask, and the predicted images of MedT.

The results of the U-Net model are depicted in Figure 9. We can see that the model correctly segments fire pixels and identifies the shape of a flame. However, it still misclassifies some small areas.

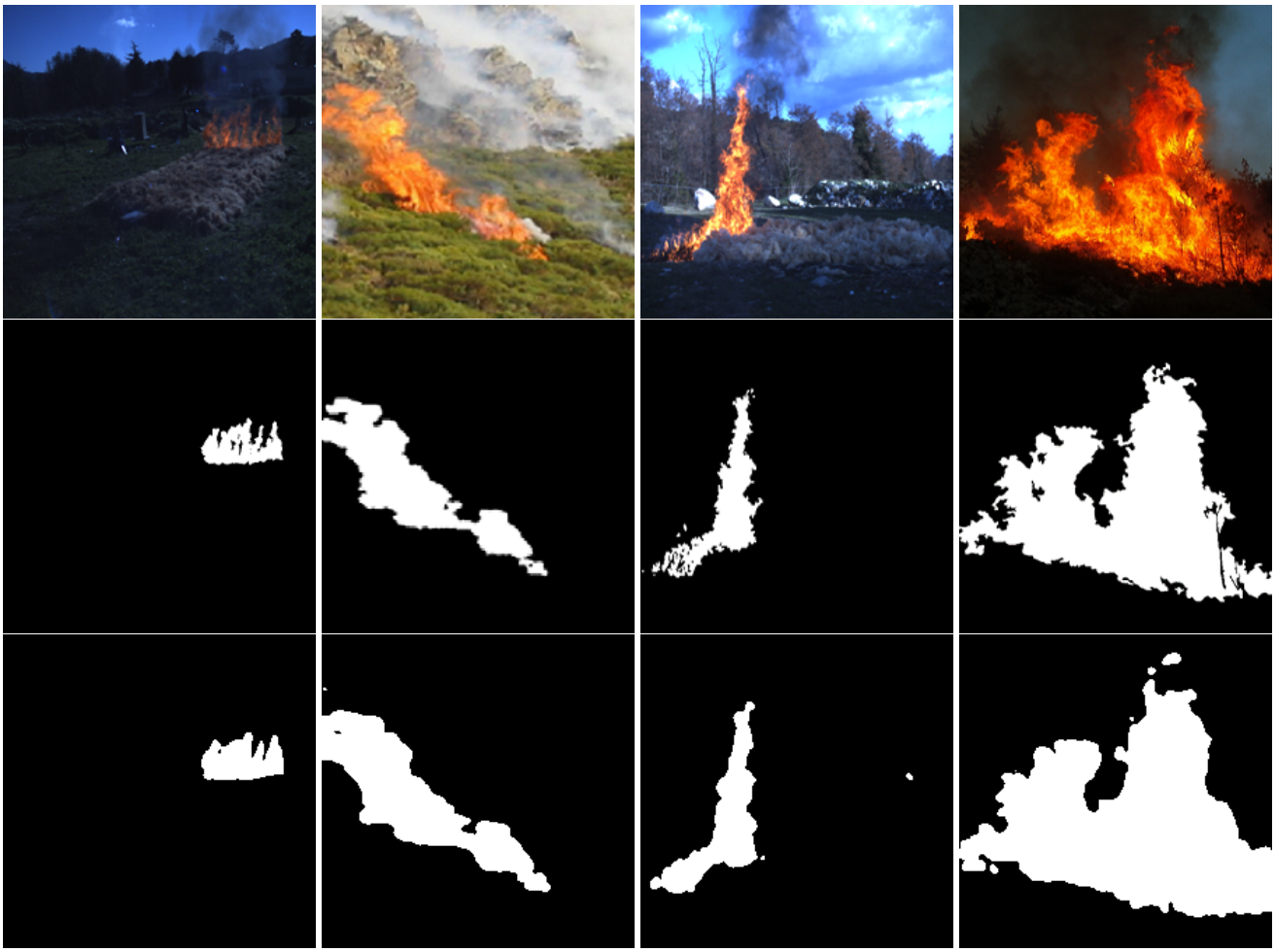

Figure 9. Results of U-Net. From (**top**) to (**bottom**): RGB images, their corresponding mask, and the predicted images of U-Net.

Figure 10 presents the results of the U^2-Net model. We can see that this model does not identify the precise shape of the fire and misdetects some small fire areas.

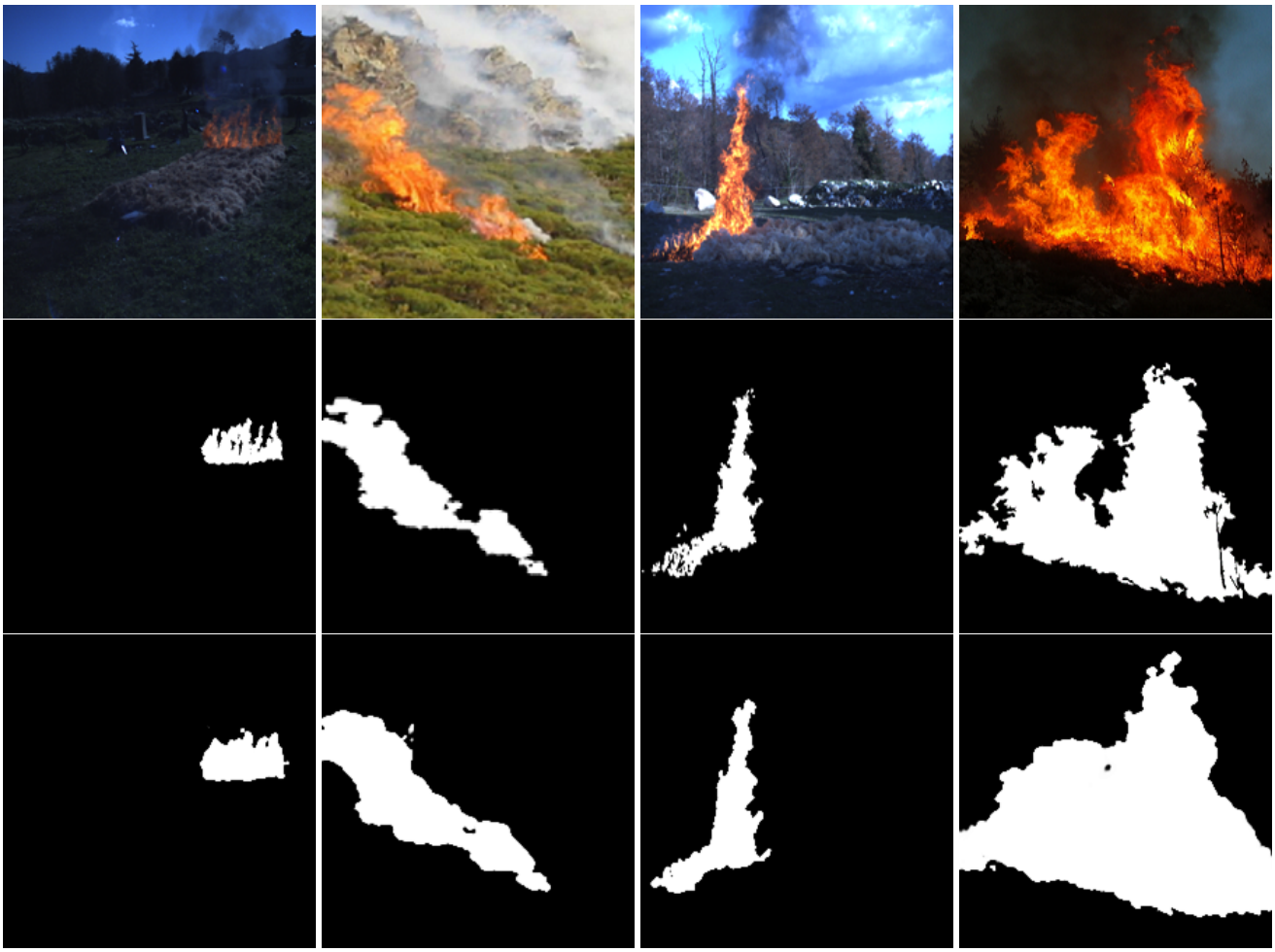

Figure 10. Results of U^2-Net. From (**top**) to (**bottom**): RGB images, their corresponding mask, and the predicted images of U^2-Net.

Figure 11 illustrates some EfficientSeg results. We can see that EfficientSeg shows an excellent performance in segmenting fire pixels under different conditions, similar to Transformers TransUNet-Res50-ViT and MedT. It also correctly identifies the small areas of fire and their precise shapes.

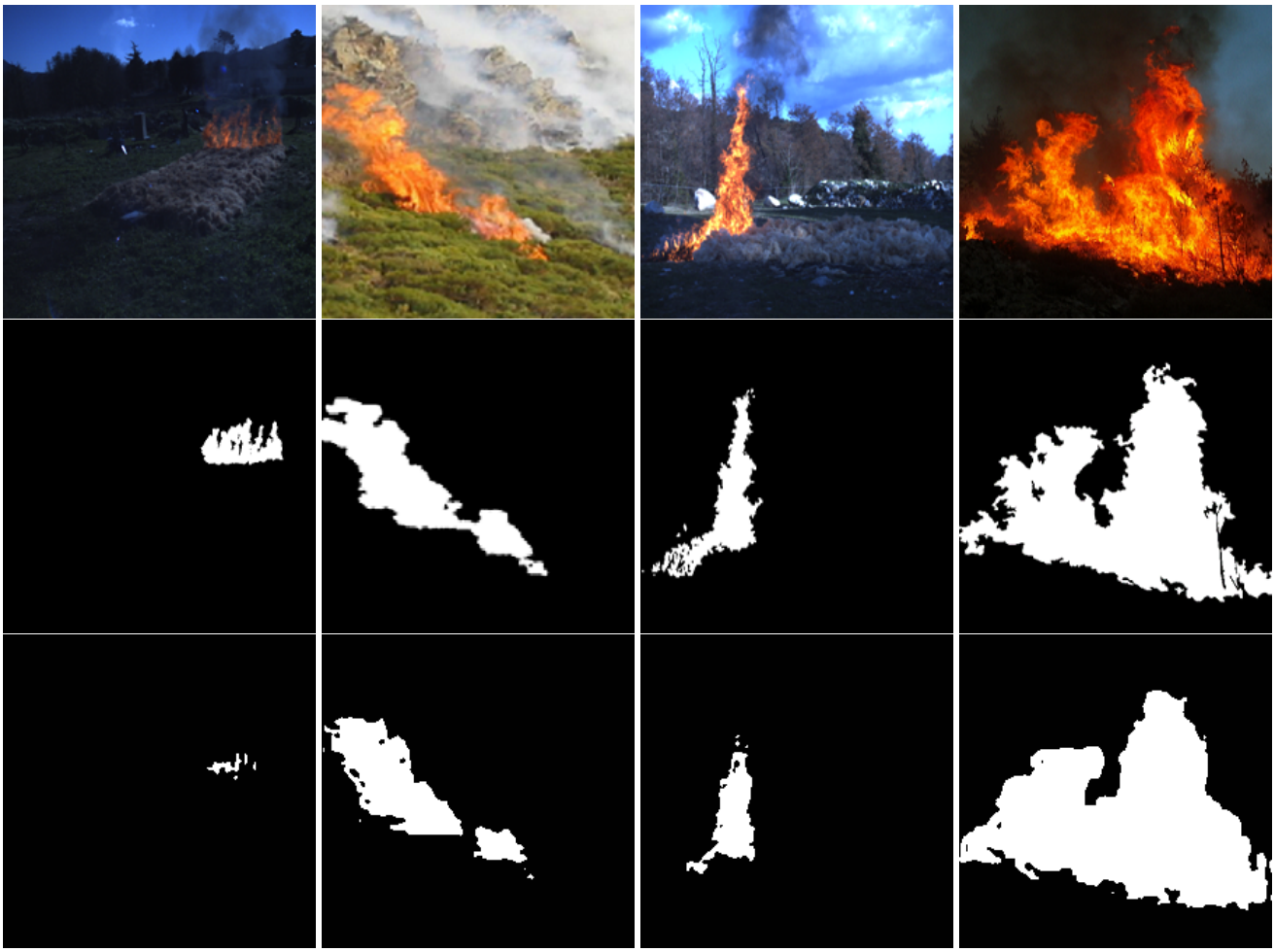

Figure 11. Results of EfficientSeg. From (**top**) to (**bottom**): RGB images, their corresponding mask, and the predicted images of EfficientSeg.

In addition, we evaluated TransUNet and MedT using images downloaded from the web. We can see in Figure 12 that TransUNet-Res50-ViT accurately segments fire pixels and detects the precise shape of the fire under various conditions such as in the presence of smoke. It shows a better visual result than TransUNet–ViT, which fails to detect small fire areas. MedT also shows a great performance in identifying fire pixels. However, it misclassifies some small fire areas.

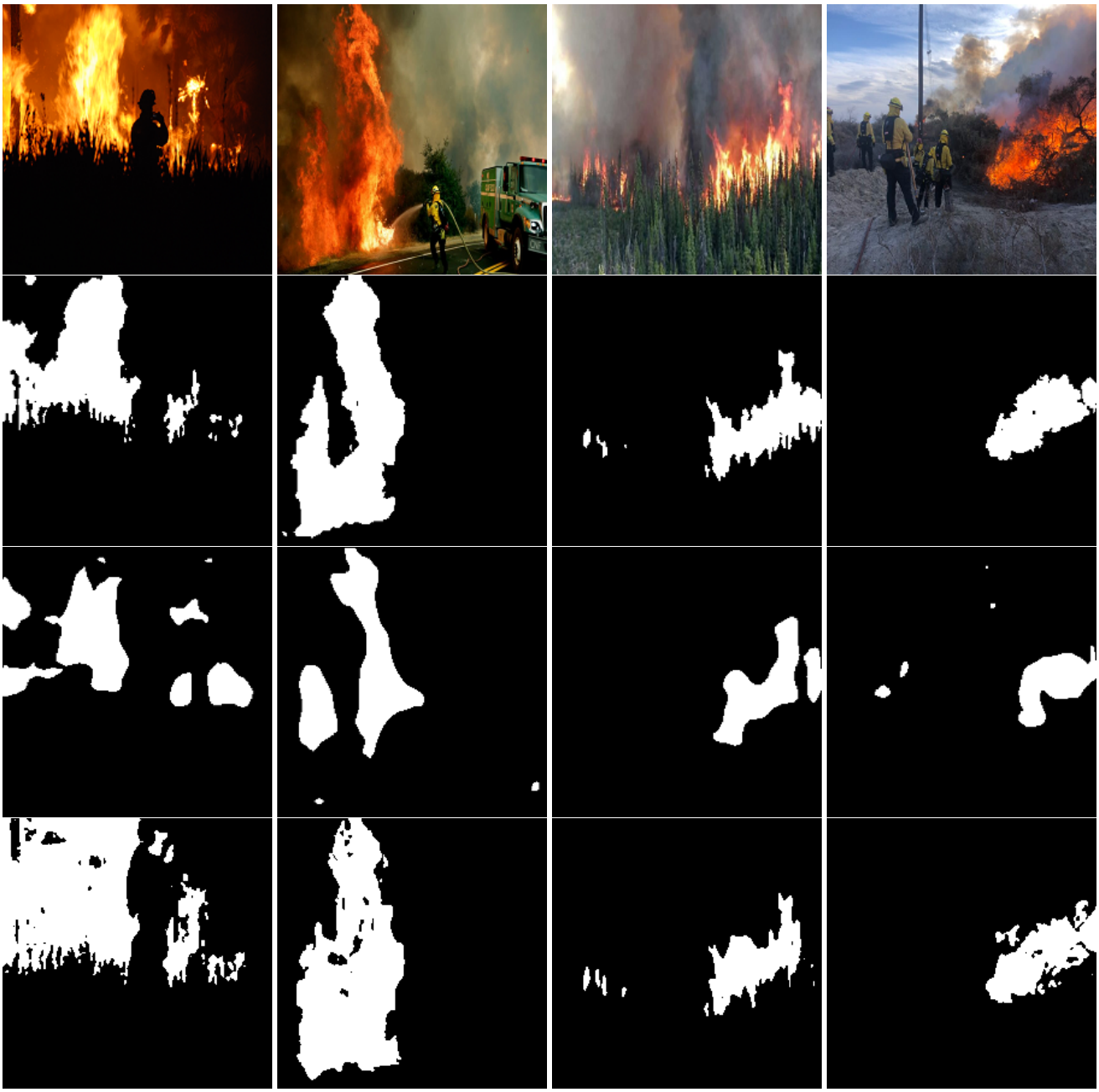

Figure 12. Results of TransUNet and MedT using web images. From (**top**) to (**bottom**): real RGB images, TransUNet-Res50-ViT results, TransUNet-ViT results, and MedT results.

5. Discussion

Transformers techniques, TransUNet-Res50-ViT, TransUNet-ViT and MedT, achieve an excellent performance compared to deep CNNs U^2-Net, U-Net, and EfficientSeg and compared to classical machine learning methods (the color fusion method) thanks to their rich extracted feature maps. However, they need higher inference times. For example, TransUNet R50-ViT and MedT obtain 1.41 and 2.72 s, respectively. TransUNet-Res50-ViT achieves the best F1-score with 97.7% thanks to the use of both global and local features. This Transformer proves an excellent ability to localize and segment forest fire pixels as well as small fire areas. It correctly distinguishes between forest fires and background under different conditions such as in the presence of smoke and in different weather conditions. However, it still depends on a pre-trained backbone, which requires a large computational capacity and more time during training.

MedT achieves an F1-score of 96%, which is better than TransUNet-ViT, thanks to spatial and local information extracted by ConvNets. This model proves its efficiency in segmenting fire pixels and detecting the precise fire's shape without pre-training. However, it misdetects some small fire areas. It also requires larger computational capacity and more time during training using high input resolution images. TransUNet-ViT extracts fire features using pure Transformer ViT. This model shows an excellent performance compared to deep CNNs, U-Net, U^2-Net, and EfficientSeg. It obtains a lower inference time, which is 0.13 s. It also proves its ability to segment forest fire pixels under various conditions. However, it misdetects the border pixels of the fire and the exact fire area's shape. EfficientSeg, U-Net, and U^2-Net show their ability to segment forest fire pixels and detect the precise fire's shape. However, they still misclassify some small fire areas.

To conclude, vision Transformers, TransUNet and MedT, show their excellent ability to segment forest fires and detect the shape of the fire front. These models outperform current state-of-the-art architectures. TransUNet-Res50-ViT and MedT, which adopted a ConvNets and Transformer as a backbone, identify the fine details of forest fires. They prove their promising use for segmenting fire pixels under various conditions such as in the presence of smoke and with changes of brightness in the environment.

6. Conclusions

In this paper, we propose a new approach based on the use of vision Transformers for forest fire segmentation. We explore two Transformers, TransUNet and MedT, adapted to segment and identify fire pixels using the CorsicanFire dataset. We evaluate the performance of our proposed Transformers by varying backbones and input size. Then, we present a comparative analysis of the two Transformers with current state-of-the-art models: U-Net, a fusion color space method, U^2-Net, and EfficientSeg. TransUNet and MedT with a hybrid CNN-Transformer as a backbone outperformed state-of-the-art methods and showed an excellent ability to segment fire pixels and identify the precise shape of a fire under different conditions such as different brightnesses and in the presence of smoke. For future work, we first aim to adapt our Transformers-based algorithms to detect and track wildfire in videos, which consists of a sequence of frames using the spatiotemporal features of fires. We will also work on evaluating our models for the detection and segmentation of both smoke and fire pixels in urban environments.

Author Contributions: Conceptualization, M.A.A. and R.G.; methodology, R.G. and M.A.A.; software, R.G.; validation, R.G. and M.A.A.; formal analysis, R.G., M.A.A., M.J. and W.S.M.; writing—original draft preparation, R.G.; writing—review and editing, M.A.A., M.J., W.S.M. and R.A.; funding acquisition, M.A.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was enabled in part by support provided by the Natural Sciences and Engineering Research Council of Canada (NSERC), funding reference number RGPIN-2018-06233.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: This work uses a publicly available dataset CorsicanFire, see reference [20] for data availability. More details about the data are available under Section 3.5.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Dimitropoulos, S. Fighting fire with science. *Nature* **2019**, *576*, 328–329. [CrossRef] [PubMed]
2. Gaur, A.; Singh, A.; Kumar, A.; Kulkarni, K.S.; Lala, S.; Kapoor, K.; Srivastava, V.; Kumar, A.; Mukhopadhyay, S.C. Fire Sensing Technologies: A Review. *IEEE Sens. J.* **2019**, *19*, 3191–3202. [CrossRef]
3. Kuutti, S.; Bowden, R.; Jin, Y.; Barber, P.; Fallah, S. A Survey of Deep Learning Applications to Autonomous Vehicle Control. *IEEE Trans. Intell. Transp. Syst.* **2021**, *22*, 712–733. [CrossRef]
4. Tian, Y.; Luo, P.; Wang, X.; Tang, X. Deep Learning Strong Parts for Pedestrian Detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1904–1912.

5. Pérez-Hernández, F.; Tabik, S.; Lamas, A.; Olmos, R.; Fujita, H.; Herrera, F. Object Detection Binary Classifiers methodology based on deep learning to identify small objects handled similarly: Application in video surveillance. *Knowl.-Based Syst.* **2020**, *194*, 105590. [CrossRef]
6. Nawaratne, R.; Alahakoon, D.; De Silva, D.; Yu, X. Spatiotemporal Anomaly Detection Using Deep Learning for Real-Time Video Surveillance. *IEEE Trans. Ind. Inform.* **2020**, *16*, 393–402. [CrossRef]
7. Gaur, A.; Singh, A.; Kumar, A.; Kumar, A.; Kapoor, K. Video flame and smoke based fire detection algorithms: A literature review. *Fire Technol.* **2020**, *56*, 1943–1980. [CrossRef]
8. Ghali, R.; Jmal, M.; Souidene Mseddi, W.; Attia, R. Recent Advances in Fire Detection and Monitoring Systems: A Review. In Proceedings of the 18th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT'18), Genoa, Italy, 20–22 December 2018; Springer International Publishing: Berlin/Heidelberg, Germany, 2018; Volume 1, pp. 332–340.
9. Ott, M.; Edunov, S.; Grangier, D.; Auli, M. Scaling Neural Machine Translation. *arXiv* **2018**, arXiv:1806.00187.
10. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762.
11. Devlin, J.; Chang, M.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2018**, arXiv:1810.04805.
12. Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; et al. A Survey on Visual Transformer. *arXiv* **2020**, arXiv:2012.12556.
13. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. In *Computer Vision—ECCV*; Springer International Publishing: Cham, Switzerland, 2020; pp. 213–229.
14. Ding, M.; Yang, Z.; Hong, W.; Zheng, W.; Zhou, C.; Yin, D.; Lin, J.; Zou, X.; Shao, Z.; Yang, H.; et al. CogView: Mastering Text-to-Image Generation via Transformers. *arXiv* **2021**, arXiv:2105.13290.
15. Yang, F.; Yang, H.; Fu, J.; Lu, H.; Guo, B. Learning Texture Transformer Network for Image Super-Resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 5791–5800.
16. Valanarasu, J.M.J.; Oza, P.; Hacihaliloglu, I.; Patel, V.M. Medical Transformer: Gated Axial-Attention for Medical Image Segmentation. *arXiv* **2021**, arXiv:2102.10662.
17. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. *arXiv* **2021**, arXiv:2102.04306.
18. Khan, S.; Naseer, M.; Hayat, M.; Zamir, S.W.; Khan, F.S.; Shah, M. Transformers in Vision: A Survey. *arXiv* **2021**, arXiv:2101.01169.
19. Sudre, C.H.; Li, W.; Vercauteren, T.; Ourselin, S.; Jorge Cardoso, M. Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*; Springer International Publishing: Cham, Switzerland, 2017; pp. 240–248.
20. Toulouse, T.; Rossi, L.; Campana, A.; Celik, T.; Akhloufi, M.A. Computer vision for wildfire research: An evolving image dataset for processing and analysis. *Fire Saf. J.* **2017**, *92*, 188–194. [CrossRef]
21. Akhloufi, M.A.; Tokime, R.B.; Elassady, H. Wildland fires detection and segmentation using deep learning. Pattern recognition and tracking xxix. *Int. Soc. Opt. Photonics Proc. SPIE* **2018**, *10649*, 106490B. [CrossRef]
22. Dzigal, D.; Akagic, A.; Buza, E.; Brdjanin, A.; Dardagan, N. Forest Fire Detection based on Color Spaces Combination. In Proceedings of the 11th International Conference on Electrical and Electronics Engineering (ELECO), Bursa, Turkey, 28–30 November 2019; pp. 595–599. [CrossRef]
23. Qin, X.; Zhang, Z.; Huang, C.; Dehghan, M.; Zaiane, O.R.; Jagersand, M. U2-Net: Going deeper with nested U-structure for salient object detection. *Pattern Recognit.* **2020**, *106*, 107404. [CrossRef]
24. Yesilkaynak, V.B.; Sahin, Y.H.; Unal, G.B. EfficientSeg: An Efficient Semantic Segmentation Network. *arXiv* **2020**, arXiv:2009.06469.
25. Horng, W.B.; Peng, J.W.; Chen, C.Y. A new image-based real-time flame detection method using color analysis. In Proceedings of the IEEE Networking, Sensing and Control, Tucson, AZ, USA, 19–22 March 2005; pp. 100–105. [CrossRef]
26. Çelik, T.; Demirel, H. Fire detection in video sequences using a generic color model. *Fire Saf. J.* **2009**, *44*, 147–158. [CrossRef]
27. Chen, T.H.; Wu, P.H.; Chiou, Y.C. An early fire-detection method based on image processing. In Proceedings of the International Conference on Image Processing, Singapore, 24–27 October 2004; Volume 3, pp. 1707–1710. [CrossRef]
28. Collumeau, J.F.; Laurent, H.; Hafiane, A.; Chetehouna, K. Fire scene segmentations for forest fire characterization: A comparative study. In Proceedings of the 18th IEEE International Conference on Image Processing, Brussels, Belgium, 11–14 September 2011; pp. 2973–2976. [CrossRef]
29. Chino, D.Y.T.; Avalhais, L.P.S.; Rodrigues, J.F.; Traina, A.J.M. BoWFire: Detection of Fire in Still Images by Integrating Pixel Color and Texture Analysis. In Proceedings of the 28th SIBGRAPI Conference on Graphics, Patterns and Images, Salvador, Brazil, 6–29 August 2015; pp. 95–102. [CrossRef]
30. Chen, J.; He, Y.; Wang, J. Multi-feature fusion based fast video flame detection. *Build. Environ.* **2010**, *45*, 1113–1122. [CrossRef]
31. Jamali, M.; Karimi, N.; Samavi, S. Saliency Based Fire Detection Using Texture and Color Features. In Proceedings of the 28th Iranian Conference on Electrical Engineering (ICEE), Tabriz, Iran, 4–6 August 2020; pp. 1–5. [CrossRef]
32. Ko, B.C.; Cheong, K.H.; Nam, J.Y. Fire detection based on vision sensor and support vector machines. *Fire Saf. J.* **2009**, *44*, 322–329. [CrossRef]

33. Foggia, P.; Saggese, A.; Vento, M. Real-Time Fire Detection for Video-Surveillance Applications Using a Combination of Experts Based on Color, Shape, and Motion. *IEEE Trans. Circuits Syst. Video Technol.* **2015**, *25*, 1545–1556. [CrossRef]
34. Khondaker, A.; Khandaker, A.; Uddin, J. Computer Vision-based Early Fire Detection Using Enhanced Chromatic Segmentation and Optical Flow Analysis Technique. *Int. Arab. J. Inf. Technol. (IAJIT)* **2020**, *17*, 947–953.
35. Emmy Prema, C.; Vinsley, S.S.; Suresh, S. Efficient Flame Detection Based on Static and Dynamic Texture Analysis in Forest Fire Detection. *Fire Technol.* **2018**, *54*, 255–288. [CrossRef]
36. Wang, T.; Shi, L.; Yuan, P.; Bu, L.; Hou, X. A new fire detection method based on flame color dispersion and similarity in consecutive frames. In Proceedings of the Chinese Automation Congress (CAC), Jinan, China, 20–22 October 2017; pp. 151–156. [CrossRef]
37. Ajith, M.; Martínez-Ramón, M. Unsupervised Segmentation of Fire and Smoke From Infra-Red Videos. *IEEE Access* **2019**, *7*, 182381–182394. [CrossRef]
38. Gonzalez, A.; Zuniga, M.D.; Nikulin, C.; Carvajal, G.; Cardenas, D.G.; Pedraza, M.A.; Fernandez, C.A.; Munoz, R.I.; Castro, N.A.; Rosales, B.F.; et al. Accurate fire detection through fully convolutional network. In Proceedings of the 7th Latin American Conference on Networked and Electronic Media (LACNEM), Valparaiso, Chile, 6–7 November 2017; pp. 1–6.
39. Dang-Ngoc, H.; Nguyen-Trung, H. Evaluation of Forest Fire Detection Model using Video captured by UAVs. In Proceedings of the 19th International Symposium on Communications and Information Technologies (ISCIT), Ho Chi Minh City, Vietnam, 25–27 September 2019; pp. 513–518.
40. Muhammad, K.; Ahmad, J.; Lv, Z.; Bellavista, P.; Yang, P.; Baik, S.W. Efficient Deep CNN-Based Fire Detection and Localization in Video Surveillance Applications. *IEEE Trans. Syst. Man Cybern. Syst.* **2019**, *49*, 1419–1434. [CrossRef]
41. Mlích, J.; Koplík, K.; Hradiš, M.; Zemčík, P. Fire segmentation in Still images. In *International Conference on Advanced Concepts for Intelligent Vision Systems*; Springer International Publishing: Berlin/Heidelberg, Germany, 2020; pp. 27–37.
42. Harkat, H.; Nascimento, J.; Bernardino, A. Fire segmentation using a DeepLabv3+ architecture. Image and Signal Processing for Remote Sensing XXVI. *Int. Soc. Opt. Photonics Proc. SPIE* **2020**, *11533*, 134–145.
43. Bochkov, V.S.; Kataeva, L.Y. wUUNet: Advanced Fully Convolutional Neural Network for Multiclass Fire Segmentation. *Symmetry* **2021**, *13*, 98. [CrossRef]
44. Li, P.; Zhao, W. Image fire detection algorithms based on convolutional neural networks. *Case Stud. Therm. Eng.* **2020**, *19*, 100625. [CrossRef]
45. Xu, R.; Lin, H.; Lu, K.; Cao, L.; Liu, Y. A Forest Fire Detection System Based on Ensemble Learning. *Forests* **2021**, *12*, 217. [CrossRef]
46. Khan, R.A.; Uddin, J.; Corraya, S. Real-time fire detection using enhanced color segmentation and novel foreground extraction. In Proceedings of the 4th International Conference on Advances in Electrical Engineering (ICAEE), Dhaka, Bangladesh, 28–30 September 2017; pp. 488–493.
47. Zhao, Z.Q.; Zheng, P.; Xu, S.T.; Wu, X. Object Detection With Deep Learning: A Review. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *30*, 3212–3232. [CrossRef]
48. Pan, Z.; Xu, J.; Guo, Y.; Hu, Y.; Wang, G. Deep Learning Segmentation and Classification for Urban Village Using a Worldview Satellite Image Based on U-Net. *Remote Sens.* **2020**, *12*, 1574. [CrossRef]
49. Bragilevsky, L.; Bajić, I.V. Deep learning for Amazon satellite image analysis. In Proceedings of the IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM), Victoria, BC, Canada, 21–23 August 2017; pp. 1–5. [CrossRef]
50. Abdollahi, A.; Pradhan, B.; Shukla, N.; Chakraborty, S.; Alamri, A. Deep Learning Approaches Applied to Remote Sensing Datasets for Road Extraction: A State-Of-The-Art Review. *Remote Sens.* **2020**, *12*, 1444. [CrossRef]
51. Bakator, M.; Radosav, D. Deep Learning and Medical Diagnosis: A Review of Literature. *Multimodal Technol. Interact.* **2018**, *2*, 47. [CrossRef]
52. Minaee, S.; Boykov, Y.Y.; Porikli, F.; Plaza, A.J.; Kehtarnavaz, N.; Terzopoulos, D. Image Segmentation Using Deep Learning: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, [CrossRef] [PubMed]
53. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [CrossRef]
54. Iandola, F.N.; Moskewicz, M.W.; Ashraf, K.; Han, S.; Dally, W.J.; Keutzer, K. SqueezeNet: AlexNet-level accuracy with 50× fewer parameters and <1 MB model size. *arXiv* **2016**, arXiv:1602.07360.
55. Xing, Y.; Zhong, L.; Zhong, X. An Encoder-Decoder Network Based FCN Architecture for Semantic Segmentation. *Wirel. Commun. Mob. Comput.* **2020**, *2020*, 8861886. [CrossRef]
56. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI*; Springer International Publishing: Berlin/Heidelberg, Germany, 2015; pp. 234–241. [CrossRef]
57. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2015**, arXiv:1409.1556.
58. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 834–848. [CrossRef]
59. Chen, L.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv* **2017**, arXiv:1706.05587.

60. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
61. Xiao, J.; Hays, J.; Ehinger, K.A.; Oliva, A.; Torralba, A. SUN database: Large-scale scene recognition from abbey to zoo. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 3485–3492. [CrossRef]
62. Dai, J.; Li, Y.; He, K.; Sun, J. R-FCN: Object Detection via Region-based Fully Convolutional Networks. *arXiv* **2016**, arXiv:1605.06409.
63. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the European conference on computer vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.
64. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
65. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
66. Jocher, G.; Stoken, A.; Chaurasia, A.; Borovec, J.; Chanvichet, V.; Kwon, Y.; TaoXie, S.; Changyu, L.; Abhiram, V.; Skalski, P.; et al. Yolov5. 2021. Available online: https://github.com/ultralytics/yolov5 (accessed on 20 August 2021).
67. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and Efficient Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10781–10790.
68. Tan, M.; Le, Q. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 6105–6114.
69. Girdhar, R.; Carreira, J.; Doersch, C.; Zisserman, A. Video Action Transformer Network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 9–15 June 2019; pp. 244–253.
70. Ye, L.; Rochan, M.; Liu, Z.; Wang, Y. Cross-Modal Self-Attention Network for Referring Image Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 9–15 June 2019; pp. 10502–10511.
71. He, X.; Chen, Y.; Lin, Z. Spatial-Spectral Transformer for Hyperspectral Image Classification. *Remote Sens.* **2021**, *13*, 498. [CrossRef]
72. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.
73. Sun, C.; Shrivastava, A.; Singh, S.; Gupta, A. Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 843–852.
74. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jégou, H. Training data-efficient image transformers & distillation through attention. *arXiv* **2020**, arXiv:2012.12877.
75. Chuvieco, E.; Mouillot, F.; van der Werf, G.R.; San Miguel, J.; Tanase, M.; Koutsias, N.; García, M.; Yebra, M.; Padilla, M.; Gitas, I.; et al. Historical background and current developments for mapping burned area from satellite Earth observation. *Remote Sens. Environ.* **2019**, *225*, 45–64. [CrossRef]
76. van der Werf, G.R.; Randerson, J.T.; Giglio, L.; van Leeuwen, T.T.; Chen, Y.; Rogers, B.M.; Mu, M.; van Marle, M.J.E.; Morton, D.C.; Collatz, G.J.; et al. Global fire emissions estimates during 1997–2016. *Earth Syst. Sci. Data* **2017**, *9*, 697–720. [CrossRef]
77. Giglio, L.; Boschetti, L.; Roy, D.P.; Humber, M.L.; Justice, C.O. The Collection 6 MODIS burned area mapping algorithm and product. *Remote Sens. Environ.* **2018**, *217*, 72–85. [CrossRef]
78. Key, C.H.; Benson, N.C. Landscape assessment (LA). In *FIREMON: Fire Effects Monitoring and Inventory System. Gen. Tech. Rep. RMRS-GTR-164-CD*; Lutes, D.C., Keane, R.E., Caratti, J.F., Key, C.H., Benson, N.C., Sutherland, S., Gangi, L.J., Eds.; U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station: Fort Collins, CO, USA, 2006; Volume 164, p. LA-1-55.
79. Roy, D.; Boschetti, L.; Trigg, S. Remote sensing of fire severity: Assessing the performance of the normalized burn ratio. *IEEE Geosci. Remote Sens. Lett.* **2006**, *3*, 112–116. [CrossRef]
80. Miller, J.D.; Thode, A.E. Quantifying burn severity in a heterogeneous landscape with a relative version of the delta Normalized Burn Ratio (dNBR). *Remote Sens. Environ.* **2007**, *109*, 66–80. [CrossRef]
81. Frampton, W.J.; Dash, J.; Watmough, G.; Milton, E.J. Evaluating the capabilities of Sentinel-2 for quantitative estimation of biophysical variables in vegetation. *ISPRS J. Photogramm. Remote Sens.* **2013**, *82*, 83–92. [CrossRef]
82. Zheng, Z.; Wang, J.; Shan, B.; He, Y.; Liao, C.; Gao, Y.; Yang, S. A New Model for Transfer Learning-Based Mapping of Burn Severity. *Remote Sens.* **2020**, *12*, 708. [CrossRef]
83. Rebecca, G.; Tim, D.; Warwick, H.; Luke, C. A remote sensing approach to mapping fire severity in south-eastern Australia using sentinel 2 and random forest. *Remote Sens. Environ.* **2020**, *240*, 111702. [CrossRef]
84. Zanetti, M.; Marinelli, D.; Bertoluzza, M.; Saha, S.; Bovolo, F.; Bruzzone, L.; Magliozzi, M.L.; Zavagli, M.; Costantini, M. A high resolution burned area detector for Sentinel-2 and Landsat-8. In Proceedings of the 10th International Workshop on the Analysis of Multitemporal Remote Sensing Images (MultiTemp), Shanghai, China, 5–7 August 2019; pp. 1–4.
85. Kussul, N.; Lavreniuk, M.; Skakun, S.; Shelestov, A. Deep Learning Classification of Land Cover and Crop Types Using Remote Sensing Data. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 778–782. [CrossRef]

86. Marcos, D.; Volpi, M.; Kellenberger, B.; Tuia, D. Land cover mapping at very high resolution with rotation equivariant CNNs: Towards small yet accurate models. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 96–107. [CrossRef]
87. Zhang, Q.; Yuan, Q.; Zeng, C.; Li, X.; Wei, Y. Missing Data Reconstruction in Remote Sensing Image With a Unified Spatial–Temporal–Spectral Deep Convolutional Neural Network. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 4274–4288. [CrossRef]
88. Kampffmeyer, M.; Salberg, A.B.; Jenssen, R. Semantic Segmentation of Small Objects and Modeling of Uncertainty in Urban Remote Sensing Images Using Deep Convolutional Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 680–688.
89. Jeppesen, J.H.; Jacobsen, R.H.; Inceoglu, F.; Toftegaard, T.S. A cloud detection algorithm for satellite imagery based on deep learning. *Remote Sens. Environ.* **2019**, *229*, 247–259. [CrossRef]
90. Pinto, M.M.; Libonati, R.; Trigo, R.M.; Trigo, I.F.; DaCamara, C.C. A deep learning approach for mapping and dating burned areas using temporal sequences of satellite images. *ISPRS J. Photogramm. Remote Sens.* **2020**, *160*, 260–274. [CrossRef]
91. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
92. Farasin, A.; Colomba, L.; Garza, P. Double-Step U-Net: A Deep Learning-Based Approach for the Estimation of Wildfire Damage Severity through Sentinel-2 Satellite Data. *Appl. Sci.* **2020**, *10*, 4332. [CrossRef]
93. Rahmatov, N.; Paul, A.; Saeed, F.; Seo, H. Realtime fire detection using CNN and search space navigation. *J. Real-Time Image Process.* **2021**, *18*, 1331–1340. [CrossRef]
94. Khennou, F.; Ghaoui, J.; Akhloufi, M.A. Forest fire spread prediction using deep learning. Geospatial Informatics XI. *Int. Soc. Opt. Photonics Proc. SPIE* **2021**, *11733*, 106–117.
95. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255. [CrossRef]
96. Howard, A.; Sandler, M.; Chu, G.; Chen, L.C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. Searching for MobileNetV3. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 1314–1324.
97. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *arXiv* **2019**, arXiv:1912.01703.