electronics

MDPI

Article

Electric Vehicle Fire Trace Recognition Based on Multi-Task Semantic Segmentation

Jiankun Pu and Wei Zhang *

School of Microelectronics, Tianjin University, Tianjin 300072, China; 3015204014@tju.edu.cn
* Correspondence: tjuzhangwei@tju.edu.cn

Abstract: Conflagration is the major safety issue of electric vehicles (EVs). Due to their well-kept appearance and structure, which demonstrate salient visual changes after combustion, EV bodies are recognized as an important basis for on-spot inspection of burnt EVs and make application using semantic segmentation possible. The combination of deep learning-based semantic segmentation and recognition of visual traces of burnt EVs would provide preliminary analytical results of fire spread trends and output status descriptions of burnt EVs for further investigation. In this paper, a dataset of image traces of burnt EVs was built, and a two-branch network structure that splits the whole task into two sub-tasks separately concentrated on foreground extraction and severity segmentation is proposed. The proposed network is trained on the dataset via the transfer learning method and is tested using 5-fold cross validation. The foreground extraction branch achieved a mean intersection over union (mIoU) of 95.16% in the burnt EV foreground extraction task, and the burnt severity branch achieved a mIoU of 66.96% for the severity segmentation task. By jointly training two branches and applying a foreground mask to 3-class severity output, the mIoU was improved to 68.92%.

Keywords: deep learning; semantic segmentation; electric vehicle fire

Citation: Pu, J.; Zhang, W. Electric Vehicle Fire Trace Recognition Based on Multi-Task Semantic Segmentation. *Electronics* **2022**, *11*, 1738. https://doi.org/10.3390/electronics11111738

Academic Editors: Luis Hernández-Callejo, Sergio Nesmachnow and Sara Gallardo Saavedra

Received: 5 May 2022
Accepted: 27 May 2022
Published: 30 May 2022

1. Introduction

Vehicles are necessities in human life and are extensively utilized in logistics, transportation and travel. The termination of the production of traditional internal combustion engine vehicles (ICEVs) is being gradually implemented worldwide under the pressure of the global energy shortage and environment pollution issues, and electric vehicles (EVs) are recognized ideal alternatives in this situation. Partially or fully driven by Li-ion batteries, EVs have presented the potential hazard of fire, which heavily affects the safety of passengers under various scenarios, e.g., parking, charging and driving. Fire incidents in EVs and plug-in hybrid electric vehicles (PHEVs) mostly begin in the battery power system. Compared with gasoline-caused vehicle fires, battery-caused vehicle fires contain more energy, extremely high temperatures, and the release of combustible and toxic gas, thus leading to higher risks and difficulty in extinguishing the fire [1,2].

In order to eliminate potential fire hazards and improve the manufacturing safety of EVs, correlative research should not only focus on prevention of combustion, but also on analysis and research of existing cases of burnt EVs. Recently, the on-spot investigation of burnt EVs has become an important method for analysis and research. Fire or damage traces remaining on the body panels and vehicle frames are frequently used to locate the origin of fire [3]. When the vehicle is not burnt extensively, traces with salient appearances, e.g., burnt-off paint and rusted metal, can provide reliable clues for the determination of fire origin [4]. Due to the similarity of material and paint utilized in EVs and conventional vehicles, fire traces of bodies of burnt EVs are also applicable and credible for investigation. Moreover, fire traces can be conveniently captured as digital images, which also provides possibilities for using a computer vision method for recognition.

Semantic segmentation is one of the major computer vision tasks that applies end-to-end classification of every pixel of the image input and outputs a corresponding segmentation map, in which a cluster of pixels classified as the same class is called semantic. With fully convolutional network (FCN) [5] first introduce convolutional neural network (CNN) into semantic segmentation, multiple advanced network structures with various optimization methods were proposed, e.g., contextual information-reinforced PSPNet [6] and DeepLab [7,8] and attention mechanism-based DANet [9] and PSANet [10]. Multiple backbones are also implemented in semantic segmentation tasks for different purposes, e.g., ResNet [11,12] with deep architecture, MobileNet [13] as a lightweight framework, and HRNet [14] for high-resolution feature extraction.

With the improvement of computer performance and the emergence of in-depth research on deep learning, semantic segmentation has been utilized in various practical tasks and has achieved par excellence performance. In the medical field, Ronneberger et al. [15] proposed U-Net with an encoder-decoder architecture for biomedical segmentation tasks. Milletari et al. [16] proposed a variant called V-Net that utilized residual blocks. Zhou et al. [17] proposed a much more complex UNet++ with sub-networks connected through a series of nested, dense skip pathways. Apart from the structures, the targets for medical segmentation also varies, e.g., lungs, lesions, lobes, tumours, and vessels. In the scene parsing and automatic driving field, Zhao et al. [6] proposed PSPNet with a classic pyramid pooling module. Charles et al. [18] expanded the input of the network to 3d point sets and proposed a related structure named PointNet. Kirillov et al. [19] combined sematic segmentation and instance segmentation tasks and proposed a new task called panoptic segmentation. Semantic segmentation is also in large-scale use for fire and smoke detection and recognition. Wang et al. [20] proposed a model concentrated on small fire and smoke regions in video data. Zhang et al. [21] proposed a lightweight U-Net-based network for forest fire detection and recognition. Mseddi et al. [22] proposed a method combining YOLOV5 and U-Net for fire detection and segmentation. Moreover, in the remote sensing field, Chen et al. [23] proposed symmetrical dense-shortcut frameworks for very-high-resolution images, and Zhang et al. [24] proposed a dual lightweight attention network for high-resolution remote sensing images.

Currently, no semantic segmentation-based research on the recognition of EV fire traces has been implemented, and no corresponding dataset has been built for the task. However, according to the forementioned analogous tasks, semantic segmentation would be compatible with the EV fire trace recognition task of this paper. The combination of semantic segmentation would not only output a preliminary analytical result of burnt EVs by collecting images conveniently, but also make its output a status description of burnt EVs for further archive and research. In summary, the main contributions of this paper can be summarized as follows:

1. A deep learning-based semantic segmentation technique was novelly applied to the recognition of fire image traces on EVs, and a dataset was labeled according to the different visual appearances of burnt EVs for corresponding tasks;
2. A multi-task learning-based two-branch network architecture was proposed. The first branch of the network was used for the foreground extraction task, and the other was built for distinguishing different severities of the burnt vehicle body. The best configuration of training and output of this architecture was found;
3. A connectivity-based weighted cross entropy loss function was proposed in the foreground branch for eliminating false true regions and keeping the main vehicle body for further processing;
4. A densely connected module with the expectation maximum attention (EMA) mechanism was proposed for better extracting multi-scale features in the severity segmentation branch.

The proposed model and an executable demo are available in Supplementary Materials at: https://github.com/Jkreat/EVFTR (accessed on 27 May 2022).

2. Materials and Methods

2.1. Dataset of Burnt EVs

Original images of burnt EVs were collected from various accident cases of EV combustion in China and burning tests conducted by Tianjin Fire Research Institute of M.E.M. The dataset contains 314 raw images with pixel-level annotations of burnt EVs. Vehicle bodies of the dataset are labeled into 3 different levels of severity and background into pixel-level according to their visual appearance after combustion. Blue stands for intact (IN), brown stands for mild and moderate burnt (MB) regions, red stands for severely burnt (SB) regions, and black stands for background (BG). The proportion of the numbers of pixels in different classes is shown in Table 1. Detailed regions of different labels are shown in Figure 1. The distinction between MB and SB is mainly based on the visual appearance of the painting. In short, regions with painting burnt into yellow or black were labeled as MB, and regions with painting entirely burnt out and bottom metal exposed were labeled as SB. As for tires and glasses, MB and SB were labeled according to whether their basic structure were kept after burning. All images with labeled masks were resized to 560 × 420 to fit the input of the proposed network. Moreover, the whole dataset was divided into five folds uniformly for five-fold cross validation. While training the foreground extraction branch, the labeled images were transferred into foreground masks. More images with corresponding labeled masks for different tasks are shown in Figure 2.

Table 1. Proportion of numbers of pixels in different classes (%).

BG	IN	MB	SB
57.09	25.70	7.55	9.67

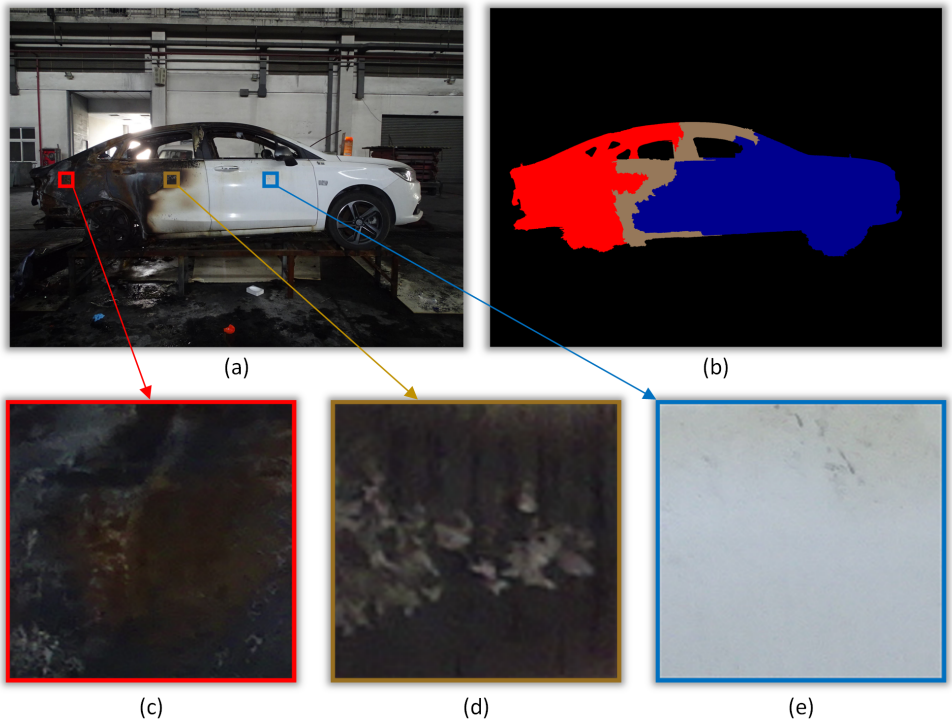

Figure 1. Original image, labeled image and details. (**a**) Original image of burnt EV. (**b**) Labeled image of different severity. (**c**) Detail of region labeled as SB. (**d**) Detail of region labeled as MB. (**e**) Detail of region labeled as IN.

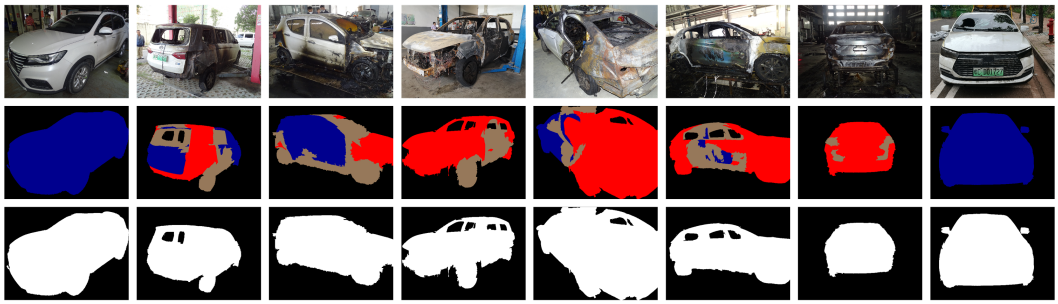

Figure 2. Images from the dataset with corresponding labels. First row: original images, second row: labeled masks for fire trace segmentation, third row: labeled masks for foreground extraction.

2.2. Backbone and Transfer Learning

Many public datasets for semantic segmentation task contain classes annotated as vehicles or cars. Due to the similarity of burnt vehicles in the tasks of this paper and intact vehicles annotated in public datasets, initializing pretrained weights from these public datasets for training the proposed network of this paper via fine-tuning method will not only lead to quick convergence, but significantly improve the overall accuracy by transferring knowledge learned from abundant corresponding data. Therefore, rather than training from scratch, transfer learning was used for training. To obtain benefits from pretrained weights and extract features better, a mainstream backbone network with deep architecture was needed. Therefore, ResNet101 with dilated convolution was selected as the backbone of the proposed architecture. Weights of the backbone were initialized using pretrained weights from COCO dataset.

Compared with the original ResNet101, the dilated version has the same number of layers and number of parameters but replaces the normal convolution operation with the dilated convolution operation in the last two groups of convolution blocks. Such a replacement increased the resolution of the output feature map without reducing the reception field. As for the semantic segmentation task, the feature map with higher spatial resolution contains more context representation; thus, the dilated ResNet101 better fits the task of this paper. The detailed configuration of the selected backbone is listed in Table 2.

Table 2. Configuration of backbone.

Layer Name	Block Configuration	Number of Blocks	Output Size
Layer0	$\begin{bmatrix} Conv, (7\times 7), 64, stride = 2 \\ Maxpool, (3\times 3), 64, stride = 2 \end{bmatrix}$	1	280×210
Layer1	$\begin{bmatrix} Conv, (3\times 3), 64, stride = 1 \\ Conv, (3\times 3), 64, stride = 1 \\ Conv, (3\times 3), 256, stride = 1 \end{bmatrix}$	3	140×105
Layer2	$\begin{bmatrix} Conv, (3\times 3), 128, stride = 1 \\ Conv, (3\times 3), 128, stride = 1 \\ Conv, (3\times 3), 512, stride = 1 \end{bmatrix}$	4	70×53
Layer3	$\begin{bmatrix} Conv, (3\times 3), 256, stride = 1 \\ Conv, (3\times 3), 256, dilation = 2 \\ Conv, (3\times 3), 1024, stride = 1 \end{bmatrix}$	23	70×53
Layer4	$\begin{bmatrix} Conv, (3\times 3), 512, stride = 1 \\ Conv, (3\times 3), 512, dilation = 4 \\ Conv, (3\times 3), 2048, stride = 1 \end{bmatrix}$	3	70×53

2.3. Foreground Extraction Branch

2.3.1. Network Structure

A modified atrous spatial pyramid pooling (ASPP) module from DeeplabV3 was connected after the backbone in this branch for capturing the multi-scale context. To fit the size of the feature map from the backbone, the original ASPP module with a dilation rate of

(6, 12, 18) was modified to a larger module with a dilation rate of (4, 11, 18, 25). Moreover, the number of output channels of each layer was promoted from 256 to 512. The overall structure of the foreground extraction branch is shown in Figure 3.

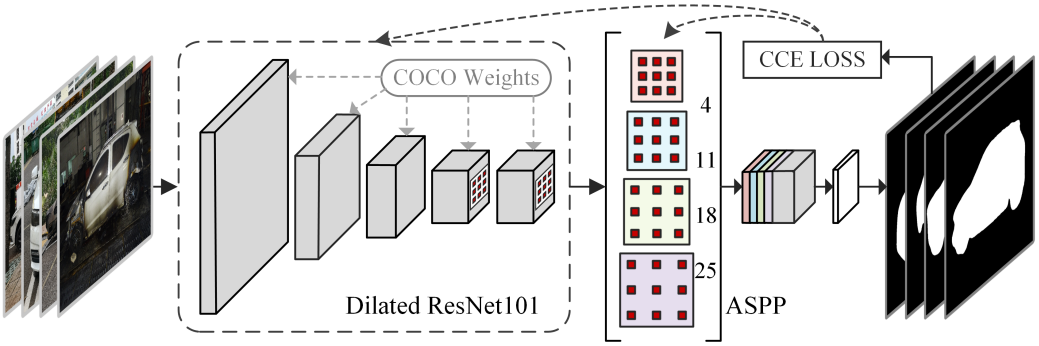

Figure 3. Structure of the foreground extraction branch.

2.3.2. CCE Loss Function

The characteristics of the foreground extraction task in this paper are summarized as follows:

- Every input image has only one main EV body as the target for processing. Other partially or fully captured vehicle bodies in the image should all be regarded as background and be minimized;
- The body of the target EV in each image is always at the center of image, i.e., the farther a predicted foreground pixel cluster is from the center of image, the less possible it would be for it to be considered the main vehicle target;
- Compared to false negative (FN) areas, false positive (FP) areas are a major issue that influence overall accuracy and should be eliminated.

To restrain the FP areas of the results from the foreground branch, a cross entropy loss function with connectivity-based weights was proposed to increase the penalization of FP domains according to their area and distance from center of the image.

The proposed loss function works when the model is "nearly converged", i.e., $N < threshold$ connected domains exist in the output image. In this condition, a connectivity analysis algorithm is applied to split output foreground into N sorted domains according to their area, and the domain with the largest area is regarded as the main body of the vehicle.

$$\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_N\} \tag{1}$$

The weighted binary cross entropy loss function for 2-class segmentation task could be described as below:

$$L = -\frac{1}{N}\sum_{i=1}^{N}(y_i \log p_i + w(1-y_i)\log(1-p_i)) \tag{2}$$

In the equation above, w is the weight value. When $w < 1$, the function concentrates more on FNs; on the contrary, the function pays more attention on FPs when $w > 1$. Moreover, the function degenerates into normal cross entropy loss if w tends to 1. When one pixel belongs to the domain $\mathcal{D}_k$, w is calculated as follows:

$$w = 1 + \log\left(1 + \frac{d_k}{\gamma}\sqrt{\frac{A_k}{A_1}}\right) \tag{3}$$

In the equation above, d_k stands for the distance between the centroid of the minimum bounding rectangle of $\mathcal{D}_k$ and the center of image, A_k is the area of $\mathcal{D}_k$, and A_1 is the

domain that possesses the largest area, i.e., the main body of the EV. γ is a hyperparameter for controlling the value of the weight.

2.4. Severity Segmentation Branch

Considering that the features of burnt EV bodies are close to the features of intact vehicles from the source dataset used for pretraining, the transfer learning method is effective in the foreground extraction task, and a simple ASPP module would result in good accuracy. However, in the severity segmentation task, the features of burnt regions are amorphous and abstract, and the number of classes for classification also increase from 2 to 4. Therefore, a network architecture with a better feature representation capability is in need.

Contextual information reinforcement and attention mechanism utilization are two major research priorities in semantic segmentation research. Inspired by DenseASPP, a densely connected multi-scale structure with an attention module named DA-EMA was proposed in this paper. The overall structure of the severity segmentation branch, including the DA-EMA module, is shown in Figure 4.

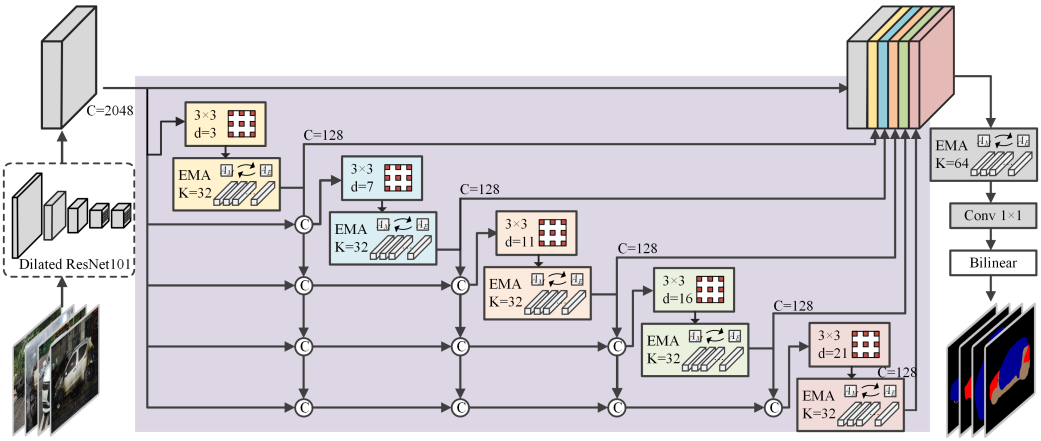

Figure 4. Structure of the severity segmentation branch. C is the output channels, K is number of bases contained in the EMA unit, and d is the dilation rate.

Simply improving the dilation rate of the ASPP module to improve the receptive field may cause a drop in overall model performance caused by the loss of modelling capability. To solve the problem and enlarge the receptive field further, Yang et al. [25] proposed a DenseNet [26]-like densely connected ASPP (DenseASPP) module.

Attention mechanisms have been proven effective in many semantic segmentation scenarios by performing feature recalibration and feature enhancement [27]. In this paper, an attention module is added to every level of a densely connected structure for enhancing multi-scale feature representation. However, traditional attention-based modules need to generate a large attention map that has high computation complexity and high GPU memory cost. A lightweight expectation maximization attention (EMA) module [28] is a good alternative in this case. Instead of treating all pixels as the reconstruction bases of the attention map, the EMA module uses the expectation maximization algorithm to find a set of compact basis in an iterative manner and then largely reduces computational complexity. A typical EMA unit consists of three operations, including responsibility estimation (A_E), likelihood maximization (A_M) and data re-estimation (A_R). Given the input $\mathbf{X} \in \mathbb{R}^{N \times C}$ and the initial bases $\boldsymbol{\mu} \in \mathbb{R}^{K \times C}$, A_E estimates the latent variables $\mathbf{Z} \in \mathbb{R}^{N \times K}$ as 'responsibility', the step functions as the E step in the expectation maximization (EM) algorithm. A_M uses the estimation to update the bases $\boldsymbol{\mu}$, which works as the M step in the EM algorithm. The A_E and A_M steps execute alternately for a pre-specified number of iterations. Then, with the converged $\boldsymbol{\mu}$ and $\mathbf{Z}$, A_R reconstructs the original $\mathbf{X}$ as $\mathbf{Y}$ and outputs it. The detailed structure of one EMA unit is shown in Figure 5.

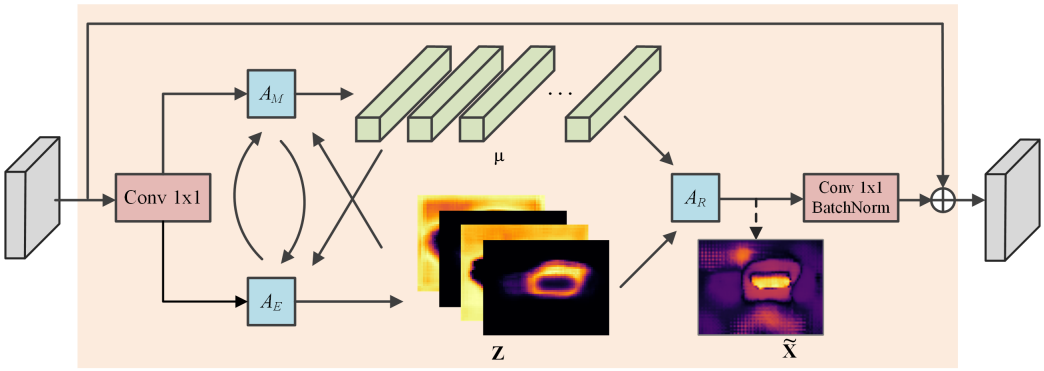

Figure 5. Detailed structure of an expectation maximization attention unit.

To improve the contextual representation, dilated convolution is frequently utilized in the proposed network. Wang et al. [29] found a "gridding" issue in the dilated convolution framework: as zeros are padded in the dilated convolution layer, the receptive field of the kernel only covers locations with a non-zero value and makes other neighboring information become lost. In this paper, dilation rates in the proposed DA-EMA module were modified from $(3, 6, 12, 18, 24)$ to $(3, 7, 11, 16, 21)$, which had no common divisor larger than 1 to improve the information used in the densely connected convolution layers with alleviation of the gridding effect. According to Figure 4, the overall DA-EMA module contains 5 EMA units with dilated convolution, and the sixth EMA unit is utilized to process the concatenated feature map. The detailed configuration of the dilated convolution layers and EMA units is shown in Table 3.

Table 3. Detailed configuration of DA-EMA units.

Block Name	Convolution Kernel Size	Dilation	Number of EMA Bases	Input Channels	Output Channels
DA-EMA1	3×3	3	32	2048	128
DA-EMA2	3×3	7	32	$2048 + 128 \times 1$	128
DA-EMA3	3×3	11	32	$2048 + 128 \times 2$	128
DA-EMA4	3×3	16	32	$2048 + 128 \times 3$	128
DA-EMA5	3×3	21	32	$2048 + 128 \times 4$	128
Output EMA	1×1	1	64	$2048 + 128 \times 5$	4

2.5. Multi-Task Learning-Based Two-Branch Architecture

Multi-task learning is a learning mechanism that enables multiple learning tasks to improve their generalization performance by sharing common knowledge learned from other tasks and maintaining their own features. The proposed model combines branches introduced above together with a shared backbone feature. In the foreground extraction branch, the result is accurate enough by training with the transfer learning method; thus, the output of this branch is used as a mask for further processes. In the severity branch, the background class is set as ignored, i.e., the parameters of the background class are not reckoned in back propagation; only parameters of three different severity levels are learned. Finally, to get the final results, the mask from the foreground extraction branch is applied to the output image of the severity segmentation branch.

Two different training methods were adopted for comparison to get better results. The overall architecture and training methods are listed in Figure 6.

Two-stage training: Train the backbone and foreground extraction branch using transfer learning first, then fully freeze parameters of the backbone and train the severity segmentation branch.

Joint training: Train the two branches and background together, then calculate the weighted sum of loss from the two branches for back propagation. Assuming L_1 is the loss

from the foreground extraction branch, and L_2 is the loss from the severity segmentation branch, the overall loss is calculated as:

$$L = \lambda L_1 + (1 - \lambda) L_2 \tag{4}$$

Moreover, two output methods were also implemented and taken into comparison. The first output method did not set the background label as ignored; thus, the severity branch also output the prediction of the background, and the number of classes of this branch output is 4. On the contrary, the second method set the background label as ignored, i.e., background was not included for back propagation; thus the severity branch barely output the prediction result containing the background class. Two different methods are shown in Figure 7.

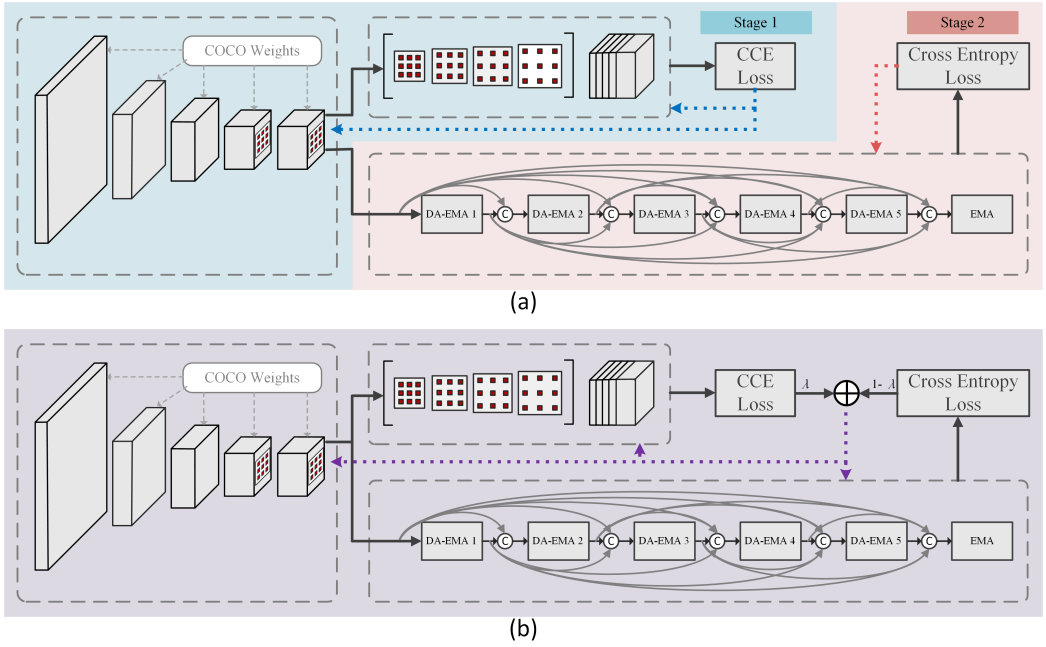

Figure 6. Two training methods implemented in this paper; dotted lines stand for back propagation. (**a**) Two-stage training; (**b**) joint training.

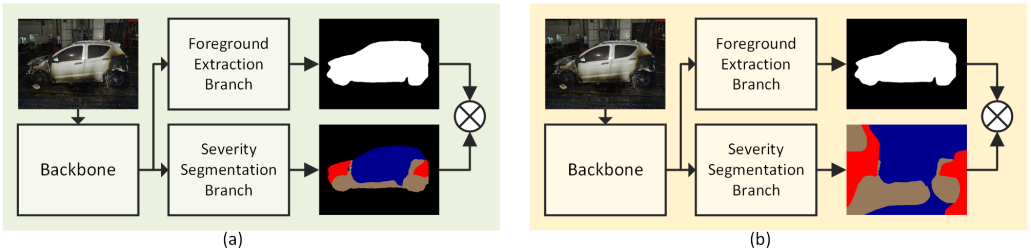

Figure 7. Two output methods (for the severity segmentation branch) in this paper: (**a**) 4-class output; (**b**) 3-class output.

2.6. K-fold Cross Validation

Generally, to evaluate the performance of a model, the dataset is randomly split into two subsets for training and testing according to a certain ratio. Test set obtained through this method may be unreliable to estimate the real performance of the model, especially when the size of the dataset is relatively small. K-fold cross validation utilizes all data to test the model, and thus could better estimate the generalization ability of the model. The fold number K is usually set to 5 or 10 [30,31]. In this paper, K was set to 5 as a trade-off between the bias of the result and time consumption for training. The leave-one-out method, a

special case of K-fold cross validation, was utilized. In this case, the number of folds equals the number of instances.

3. Results

3.1. Experimental Configuration and Evaluation Metrics

All experiments were conducted on a server running the Ubuntu 16.04 operation system. The server was equipped with two Tesla p40 GPUs and a Xeon Gold 5118 CPU. The resolutions of images from the dataset were resized to 560 × 420. Due to the utilization of transfer learning, the model converged rapidly, and the number of training epochs was set to 10 while each branch was separately trained. When two branches were trained jointly, the number was increased to 20. For all experiments, the initial learning rate was set to 0.0001 and the Adam optimizer was used. Additionally, 5-fold cross validation was implemented. The training group with fold K set for testing was named training group K.

Intersection over union (IoU) was utilized as the metric form of segmentation tasks of this paper to evaluate the accuracy of the outputs. IoU is calculated as follows:

$$IoU = \frac{TP}{TP + FP + FN} \tag{5}$$

In experiments of the foreground extraction task, only the IoU of foreground that represented bodies of target vehicles were counted.

In the EV fire trace recognition task, the number of classes was set to 4, so the mean IoU (mIoU) of 4 classes was calculated to evaluate the performance. As discussed in 3.1, the 4 classes were IN, MB, SB, BG, and the mIoU could be calculated as follows:

$$mIoU = \frac{1}{4}(IoU_{BG} + IoU_{IN} + IoU_{MB} + IoU_{SB}) \tag{6}$$

Additionally, to evaluate the accuracy of vehicle body segmentation, the union regions of IN, MB and SB were regarded as "Vehicle Body" (VB) regions; to evaluate the segmentation accuracy of burnt regions as a whole, the union of MB regions and SB regions were regarded as "Fire Trace" (FT) regions. Their IoU was thus calculated as follows:

$$IoU_{VB} = \frac{I_{IN \cup MB \cup SB}}{U_{IN \cup MB \cup SB}} \tag{7}$$

$$IoU_{FT} = \frac{I_{MB \cup SB}}{U_{MB \cup SB}} \tag{8}$$

3.2. Experiments of the Foreground Extraction Branch

In this group of experiments, to evaluate the performance of the foreground extraction branch, the backbone was connected to the modified ASPP module only, and the proposed CCE loss function was utilized.

3.2.1. Parameter Experiments of the CCE Loss Function

γ is an important component of the proposed CCE loss function in the foreground extraction branch. The value of γ was adjusted in a reasonable range, and the results obtained from different values are shown in Table 4.

3.2.2. Ablation Study

To conduct an ablation study for the foreground extraction branch, we compared the impact of the modified ASPP module and the proposed CCE loss function. The value of γ in this experiment was set to 20 according to the results above. The comparison results are listed in Table 5.

Table 4. Detailed configuration of DA-EMA units.

γ	Training Group 1	Training Group 2	Training Group 3	Training Group 4	Training Group 5	Average	Standard Deviation
1	95.69	95.33	94.31	94.78	93.88	94.80	0.74
2	95.14	95.15	94.89	95.09	94.24	94.90	0.39
3	95.51	95.35	95.31	94.75	94.20	95.02	0.54
5	95.85	95.27	95.04	94.61	94.04	94.96	0.68
10	95.80	95.53	94.96	94.63	94.22	95.03	0.65
15	95.84	95.47	95.19	94.78	94.20	95.10	0.63
20	96.03	95.62	95.24	94.80	94.11	95.16	0.74
30	95.90	95.42	95.20	94.70	94.06	95.06	0.70
50	95.66	95.33	94.63	95.02	93.91	94.91	0.68

Table 5. Detailed configuration of DA-EMA units.

Modified ASPP	CCE Loss	Training Group 1	Training Group 2	Training Group 3	Training Group 4	Training Group 5	Average	Standard Deviation
		95.34	95.27	94.58	94.65	93.81	94.73	0.62
✓		95.81	95.37	94.69	94.78	94.02	94.93	0.69
	✓	95.48	95.59	94.88	94.73	94.09	94.95	0.61
✓	✓	96.03	95.62	95.24	94.80	94.11	95.16	0.74

3.3. Experiments of the Severity Segmentation Branch

In this group of experiments, to evaluate the performance of the severity segmentation branch, the backbone was connected to the proposed DA-EMA module only, and the number of classes for training and output was set to 4, i.e., no class was ignored in the back propagation process.

3.3.1. Performance Comparison

The proposed DA-EMA module and multiple mainstream semantic segmentation network structures were trained in the same configuration including the same backbone network. The results are shown in Table 6 and Figure 8.

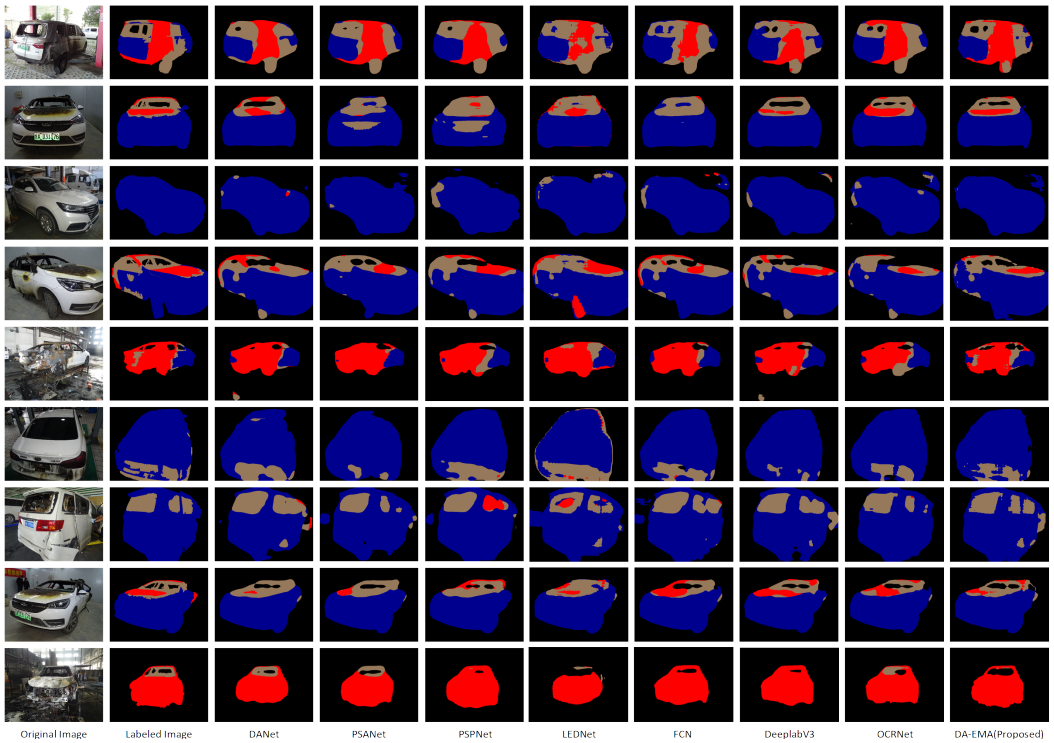

Figure 8. Results of the proposed DA-EMA and other semantic segmentation models.

Table 6. Comparison of the proposed DA-EMA and other semantic segmentation models.

Method	BG	IN	MB	SB	mIoU	VB	FT
FCN [5]	93.97	76.07	38.89	48.56	64.37	93.61	67.95
PSPNet [6]	94.15	75.07	37.63	46.90	63.44	92.82	67.85
DeepLabV3 [7]	93.84	76.23	39.11	49.06	64.56	93.44	68.47
DANet [9]	94.85	76.52	42.74	49.58	65.92	93.42	70.71
PSANet [10]	94.61	76.00	36.92	49.00	64.13	93.29	67.88
LEDNet [32]	93.18	76.91	37.82	48.82	64.18	91.25	65.75
OCRNet [33]	94.76	76.67	39.65	44.89	63.99	93.33	67.98
DA-EMA	95.30	77.12	42.68	52.73	66.96	94.03	71.00

3.3.2. Ablation Study

To examine the contribution of different modules in the proposed DA-EMA module, an ablation study was conducted. The first experiment used the structure of DenseASPP with a modified dilation rate without the EMA module (DA), the second experiment only utilized one EMA module to process the feature map from the backbone (EMA), and the third experiment was conducted using the proposed DA-EMA module. As per the results shown in Table 7, both the EMA module and the densely connected structure helped to improve the overall performance.

Table 7. Detailed configuration of DA-EMA units.

DA	EMA	BG	IN	MB	SB	mIoU	VB	FT
✓		95.16	76.92	40.03	51.22	65.83	93.86	69.15
	✓	94.52	75.78	40.79	48.51	64.90	92.87	69.78
✓	✓	95.30	77.12	42.68	52.73	66.96	94.03	71.00

3.3.3. Responsibility Map Visualization

In the EMA module, each basis corresponds to an abstract concept of the image. To examine whether the EMA mechanism functioned in the proposed DA-EMA module, multiple responsibility maps, i.e., latent variables **Z** generated from different EMA bases, were extracted. These were concluded from responsibility maps from different levels of the EMA module, as shown in Figure 9.

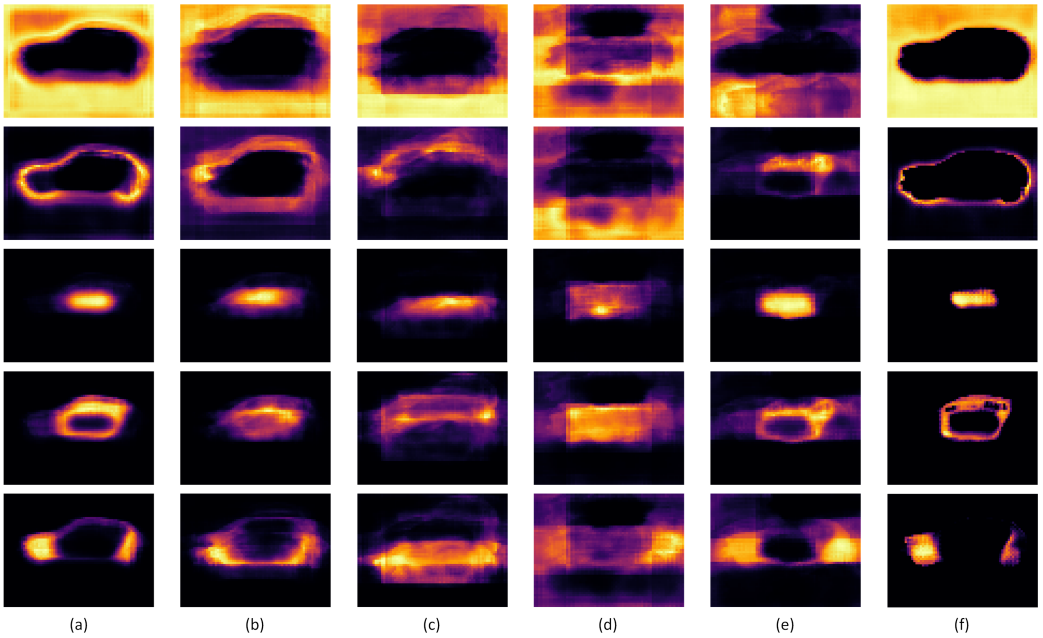

Figure 9. Converged responsibility maps collected from EMA units from different levels. (**a**) Maps from DA-EMA block 1. (**b**) Maps from DA-EMA block 2. (**c**) Maps from DA-EMA block 3. (**d**) Maps from DA-EMA block 4. (**e**) Maps from DA-EMA block 5. (**f**) Maps from output EMA block.

3.4. Experiment of the Entire Network

Benefiting from the multi-task learning mechanism, the entire network for EV fire trace recognition combined two branches and achieved better performance than using the single severity segmentation branch only. To demonstrate this improvement, different configurations of training and output were implemented using the proposed network, and the results are shown in Table 8 and Figure 10. In the joint training method, the λ value for loss calculation was set to 0.25 based on the ratio of the loss value while each branc converged.

Table 8. Results of different training methods and output methods. "Branch#2" stands for training the severity segmentation branch only.

Training Method	Output Classes	BG	IN	MB	SB	mIoU	VB	FT
Branch#2	4	95.30	77.12	42.68	52.73	66.96	94.03	71.00
2-Stage	4	95.84	78.52	45.54	52.59	68.12	94.72	72.44
Joint	4	95.63	77.86	43.81	53.57	67.72	94.86	72.79
2-Stage	3	96.15	79.17	45.11	53.80	68.56	95.10	71.79
Joint	3	95.70	78.92	45.96	55.11	68.92	94.50	73.17

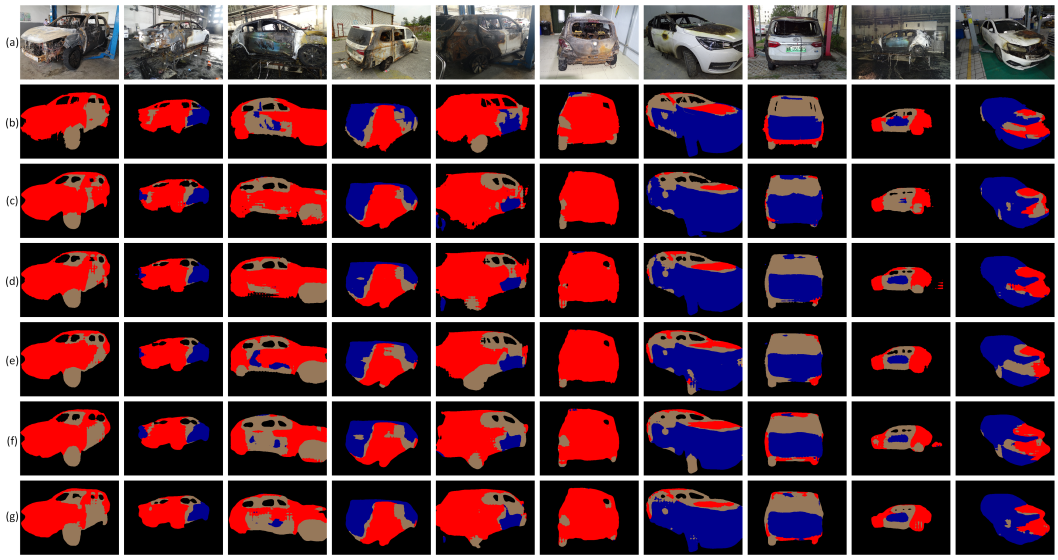

Figure 10. Results of the proposed DA-EMA in different training and output configurations. (**a**) Original images. (**b**) Labeled images. (**c**) Results of the single severity segmentation branch. (**d**) Results of the two-stage training and 4-class output. (**e**) Results of the joint training and 4-class output. (**f**) Results of the two-stage training and 3-class output. (**g**) Results of the joint training and 3-class output.

4. Discussion

To evaluate the foreground extraction branch, two experiments were conducted: a loss function parameter experiment and an ablation experiment. By tuning the value of hyperparameters in the proposed CCE loss function, we concluded that by using the loss function with an appropriate value of hyperparameters, the performance of the foreground extraction branch was improved. While the value is relatively small, it, on the contrary, hindered the convergence of the network. Once the value was extremely big, the function degenerated into normal cross entropy loss and lost its ability. Moreover, by conducting the ablation experiment, we found both the modified ASPP module and the CCE loss function had a positive effect on the branch.

The transfer learning method is essential in this paper, especially in the foreground extraction branch. By using weights pretrained on an enormous public dataset including labeled intact vehicles, the branch converged rapidly, and obtained great IoU results of

over 95%. Due to the fine results obtained from the selected backbone and modified ASPP using pretrained weights, it is enough to use the simple ASPP module and CCE loss function for the foreground extraction task. More complicated models could not cause considerable excessive improvement. However, the severity segmentation task would be less benefited from the transfer learning method, which was also the reason for splitting the whole fire trace recognition task into two sub-tasks and focusing on a new module for the enhancement of feature extraction and expression. Therefore, the DA-EMA module with densely connected dilated convolution layers and a lightweight expectation maximization attention mechanism was proposed in the severity segmentation branch for the EV fire trace recognition task.

Regarding the experiments on the severity segmentation branch, we first compared the performance of the proposed DA-EMA module and other mainstream semantic segmentation models. The results in Table 6 showed that the proposed DA-EMA module achieved better accuracy in comparison to many mainstream networks. Moreover, according to Figure 8, due to the combination of the contextual mechanism and attention mechanism, outputs of the proposed DA-EMA module were more detailed than models with attention models, e.g., DANet and PSANet, and emphasized burnt regions more than models with contextual information, e.g., PSPNet and DeeplabV3. In addition, for EVs with slightly burnt bodies, the proposed DA-EMA module generated less error when classifying intact regions into burnt regions. For EVs with windows broken and internal structures or background exposed behind the glass, the proposed DA-EMA could better recognize regions behind the broken windows. Moreover, some models might wrongly recognize components, e.g., air inlets and intact tires, as burnt regions, but these issues were barely present with the proposed DA-EMA module. The other experiment evaluating the performance of the proposed DA-EMA module was an ablation experiment conducted by separately utilizing the DenseASPP-like structure with multiple dilated convolution layers and only one EMA module without a multi-scale structure. As a result, both the dense structure and EMA module had a positive impact on the overall performance. Moreover, the visualization of responsibility maps showed that bases of EMA units were converged to a certain concept of the input image, e.g., regions of different severities, contours of EV, and backgrounds. Though responsibility maps became more abstract and diffused as dilation rate increased, representations of different concepts were not reduced.

To prove that the performance improvement benefited from the multi-task learning mechanism by combining two branches, different training methods and number of classes of the severity branch were tested. According to the results shown in Table 8, by setting the background as an ignored label and predicting only three classes of severity levels, the severity segmentation branch output fewer errors than when taking the background class into consideration. When the two-stage training method was applied, backbone parameters were frozen after the foreground branch was trained, and the parameters did not change while training the severity branch. Therefore, the output of the foreground mask was much more close to the best performance achieved by training the foreground only. However, by training the two branches jointly and making the severity segmentation branch output only three classes, the whole model achieved the best performance.

Although the proposed DA-EMA module achieved better accuracy than other mainstream semantic segmentation models and the two-branched model also improved the overall performance further, the model still has some room for improvement. Firstly, the number of parameters of the network, especially the number of parameters of the backbone and the modified ASPP with more output channels in the foreground extraction branch is large, thus raising the time consumption of model training and inference. Though the task of this paper does not have a real-time requirement, there is still room for simplifying the model by reducing redundant components. Secondly, the size of dataset is relatively small, and white is the major color of EV bodies. Therefore, a lack of EV samples of different colors may lead to error when inferring EVs with rare colors or complicated paintings. Thirdly, restricted by the computing capacity, the resolution of images was relatively insufficient

for expressing many detailed features. To solve this problem, a modified model with the capacity of processing larger images should be implemented.

5. Conclusions

In this paper, we used semantic segmentation techniques for recognizing traces of different severity levels from burnt EV images. A corresponding model with two branches separately concentrating on the foreground extraction task and the severity segmentation task was proposed, the backbone of which was ResNet101 with dilated convolution. Benefiting from the feature similarity between intact vehicles from a public dataset for pretraining and burnt vehicles from a dataset built in this paper, transfer learning considerably improved the overall accuracy of the foreground extraction task. Along with the modified ASPP module and proposed CCE loss function, the foreground extraction branch achieved an IoU of 95.16%. In the severity segmentation branch, to better enhance the feature representation capacity, a module combining the DenseASPP-like dense architecture and attention module named EMA was proposed. Achieving a mIoU of 66.96%, the proposed severity segmentation branch was tested and found to fit the task of the paper better than the other mainstream networks. Finally, by combining the two branches together, the whole multi-task based model was evaluated under different configurations of training and output, and the mIoU was finally improved to 68.92% while jointly training two branches and setting the background as ignored in the severity segmentation branch.

However, the proposed model has some limitations in certain scenarios. First, it is limited by the scale of dataset, as the majority of EV bodies are white. The lack of images of EVs with rare colors in dataset may cause errors when recognizing fire traces on EVs with these colors. To solve this problem, continuing to expand the dataset is the most efficient method. Second, although the gridding effect of the DA-EMA module was alleviated by modifying the dilation rates, the dilated convolution layers of the backbone were not optimized, and thus, the gridding effect still existed, especially in the foreground mask output from the foreground extraction branch. Third, the proposed CCE loss function in the foreground extraction branch did assist in eliminating FP areas, but when jointly training two branches, the λ was set to 0.25, which may weaken the function of CCE loss. As many FP areas were caused by other vehicle bodies, the best solution would be to apply the instance segmentation method to the foreground segmentation branch. Instance segmentation would classify pixel clusters of vehicle and distinguish which cluster belongs to which vehicle. By using this, the FP areas of other vehicle bodies can be conveniently removed. The problems above are shown in Figure 11.

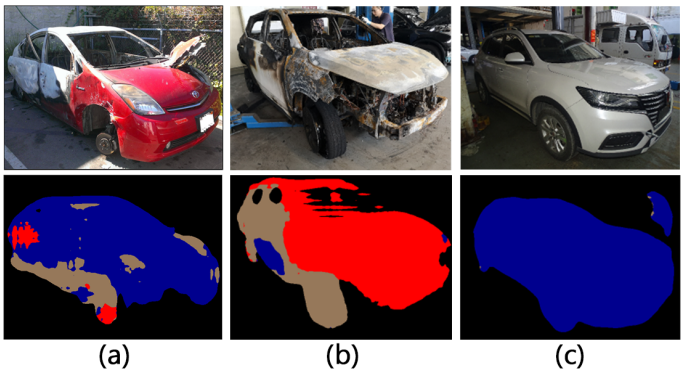

Figure 11. Limitations of the model. (**a**) Error of a red vehicle. (**b**) Gridding. (**c**) FP area from other vehicles.

Supplementary Materials: The proposed model and an executable demo are available at: https://github.com/Jkreat/EVFTR (accessed on 27 May 2022).

Author Contributions: Conceptualization, W.Z. and J.P.; methodology, J.P.; formal analysis, J.P.; investigation, J.P.; resources, W.Z.; software, J.P.; validation, J.P.; data curation, J.P.; writing—original draft preparation, J.P.; writing—review and editing, W.Z.; visualization, J.P.; supervision, W.Z.; project administration, W.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Restrictions apply to the availability of these data. Data was obtained from Tianjin Fire Research Institute of M.E.M. and are available from the authors with the permission of Tianjin Fire Research Institute of M.E.M.

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

References

1. Sun, P.; Bisschop, R.; Niu, H.; Huang, X. A Review of Battery Fires in Electric Vehicles. *Fire Technol.* **2020**, *56*, 1361–1410. [CrossRef]
2. Li, H.; Peng, W.; Yang, X.; Chen, H.; Sun, J.; Wang, Q. Full-scale experimental study on the combustion behavior of lithium ion battery pack used for electric vehicle. *Fire Technol.* **2020**, *56*, 2545–2564 [CrossRef]
3. Nicholas, J.S.; William, H.; Gregory, E.G.; Ronald, L.H.; Patrick, M.K. Vehicle Fire Burn Pattern Study. In Proceedings of the International Symposium on Fire Investigation Science and Technology, Adelphi, MD, USA, 27–29 September 2010.
4. Shields, L.E.; Scheibe, R.R. Computer-Based Training in Vehicle Fire Investigation Part 2: Fuel Sources and Burn Patterns. *SAE Tech. Paper* **2006**. [CrossRef]
5. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015.
6. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
7. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
8. 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, Munich, Germany, 8–14 September 2018; pp. 801–818.
9. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3146–3154.
10. Zhao, H.; Zhang, Y.; Liu, S.; Shi, J.; Loy, C.C.; Lin, D.; Jia, J. Psanet: Point-wise spatial attention network for scene parsing. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 267–283.
11. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
12. Yu, F.; Koltun, V.; Funkhouser, T. Dilated residual networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–27 July 2017; pp. 472–480.
13. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
14. Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. Deep high-resolution representation learning for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 3349–3364. [CrossRef] [PubMed]
15. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 234–241.
16. Milletari, F.; Navab, N.; Ahmadi, S.A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In Proceedings of the International Conference on 3D Vision, Stanford, CA, USA, 25–28 October 2016; pp. 565–571.
17. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. Unet++: A nested U-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*; Springer: Granada, Spain, 2018; pp. 3–11.
18. Charles, R.Q.; Su, H.; Kaichun, M.; Guibas, L.J. PointNet: Deep learning on point sets for 3D classification and segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Los Alamitos, CA, USA, 21–26 July 2017; pp. 77–85.
19. Kirillov, A.; He, K.; Girshick, R.; Rother, C.; Dollár, P. Panoptic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9396–9405.
20. Wang, G.; Zhang, Y.; Qu, Y.; Chen, Y.; Maqsood, H. Early forest fire region segmentation based on deep learning. In Proceedings of the Chinese Control And Decision Conference, Nanchang, China, 3–5 June 2019; pp. 6237–6241.

21. Zhang, J.; Zhu, H.; Wang, P.; Ling, X. Att squeeze u-net: A lightweight network for forest fire detection and recognition. *IEEE Access* **2021**, *9*, 10858–10870. [CrossRef]
22. Mseddi, W.S.; Ghali, R.; Jmal, M.; Attia, R. Fire detection and segmentation using YOLOv5 and U-net. In Proceedings of the European Signal Processing Conference, Dublin, Ireland, 23–27 August 2021; pp. 741–745.
23. Chen, G.; Zhang, X.; Wang, Q.; Dai, F.; Gong, Y.; Zhu, K. Symmetrical dense-shortcut deep fully convolutional networks for semantic segmentation of very-high-resolution remote sensing images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 1633–1644. [CrossRef]
24. Zhang, Y.; Chen, Y.; Ma, Q.; He, C.; Cheng, J. Dual lightweight network with attention and feature fusion for semantic segmentation of high-Resolution remote sensing images. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, Brussels, Belgium, 11–16 July 2021; pp. 2755–2758.
25. Yang, M.; Yu, K.; Zhang, C.; Li, Z.; Yang, K. Denseaspp for semantic segmentation in street scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3684–3692.
26. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
27. Ji, Y.; Zhang, H.; Wu, Q.J. Salient object detection via multi-scale attention CNN. *Neurocomputing* **2018**, *322*, 130–140. [CrossRef]
28. Li, X.; Zhong, Z.; Wu, J.; Yang, Y.; Lin, Z.; Liu, H. Expectation-maximization attention networks for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 9167–9176.
29. Wang, P.; Chen, P.; Yuan, Y.; Liu, D.; Huang, Z.; Hou, X.; Cottrell, G. Understanding convolution for semantic segmentation. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision, Lake Tahoe, NV, USA 12–15 March 2018; pp. 1451–1460.
30. Kuhn, M.; Johnson, K. *Applied Predictive Modeling*; Springer: New York, NY, USA, 2013.
31. Rodriguez, J.D.; Perez, A.; Lozano, J.A. Sensitivity analysis of k-fold cross validation in prediction error estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 569–575. [CrossRef] [PubMed]
32. Wang, Y.; Zhou, Q.; Liu, J.; Xiong, J.; Gao, G.; Wu, X.; Latecki, L.J. Lednet: A lightweight encoder-decoder network for real-time semantic segmentation. In Proceedings of the IEEE International Conference on Image Processing, Taipei, Taiwan, 22–25 September 2019; pp. 1860–1864.
33. Yuan, Y.; Chen, X.; Wang, J. Object-contextual representations for semantic segmentation. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 173–190.