Journal of Intelligent & Fuzzy Systems 41 (2021) 4051–4068
DOI:10.3233/JIFS-210349
IOS Press

RSNet: Rail semantic segmentation network for extracting aerial railroad images

R.S. Rampriya[a,*], Sabarinathan[b] and R. Suganya[c]
[a]*Department of Computer Technology, Anna University (MIT Campus), Chennai, Tamilnadu, The India*
[b]*Couger Inc, Tokyo, The Japan*
[c]*Deparment of Information Technology, Thiagarajar College of Engineering, Madurai, Tamilnadu, The India*

Abstract. In the near future, combo of UAV (Unmanned Aerial Vehicle) and computer vision will play a vital role in monitoring the condition of the railroad periodically to ensure passenger safety. The most significant module involved in railroad visual processing is obstacle detection, in which caution is obstacle fallen near track gage inside or outside. This leads to the importance of detecting and segment the railroad as three key regions, such as gage inside, rails, and background. Traditional railroad segmentation methods depend on either manual feature selection or expensive dedicated devices such as Lidar, which is typically less reliable in railroad semantic segmentation. Also, cameras mounted on moving vehicles like a drone can produce high-resolution images, so segmenting precise pixel information from those aerial images has been challenging due to the railroad surroundings chaos. RSNet is a multi-level feature fusion algorithm for segmenting railroad aerial images captured by UAV and proposes an attention-based efficient convolutional encoder for feature extraction, which is robust and computationally efficient and modified residual decoder for segmentation which considers only essential features and produces less overhead with higher performance even in real-time railroad drone imagery. The network is trained and tested on a railroad scenic view segmentation dataset (RSSD), which we have built from real-time UAV images and achieves 0.973 dice coefficient and 0.94 jaccard on test data that exhibits better results compared to the existing approaches like a residual unit and residual squeeze net.

Keywords: Railroad aerial images, efficient net, modified residual net, attention layer, semantic segmentation

1. Introduction

In railways, an accident causes significant damages to the railway concerning harshness and death rate and has been on the upswing for the past ten years. The number of accidents is rising rottenly, which is crucial for perception and avoiding casualties. Observing activities has become feasible with manual monitoring of railway tracks. In real-time, this is a tedious job that needs more labor to inspect, especially in a dangerous area, and it is not sufficient to stop the accidents. UAV with deep neural networks will become a significant research area for dynamically and frequently monitoring the railway track condition to safeguard travelers. In this development, the drone-based obstacle detection system will be acted as an alternative to the existing one due to its high mobility and low cost. The main goal of obstacle detection is accident prevention and railroad environment perception. An essential preprocessing phase of railroad observation is identifying and segmenting the regions in the aerial railroad image as it permits the drone to recognize the railway track and only perform effectual obstacle detection. In this view, implementing precise railroad segmentation is a vital role of a drone-based railroad monitoring system.

UAVs turned out to be extensively suited in several areas, mainly in transportation. The initial

[*]Corresponding author. R. S. Rampriya, Department of Computer Technology, Anna University (MIT Campus), Chennai, Tamilnadu, The India. E-mail: mail2rampriya@gmail.com.

applications comprise traffic monitoring, security surveillance, an inspection of a forest, plant detection [1], etc. UAVs are equipped with cameras to afford a proficient data acquisition process for intelligent transport systems [2]. Rail cars are used in railways [3]. The monitoring rail track is used with various sensors containing Light Detection and Ranging (LIDAR) [4] and vision-based systems. The main weakness of these methods is their limited inspection cycle and high cost. However, believe UAVs keep the promise to examine the railway tracks at more regular intervals and special conditions like bad storms. Therefore, UAV is used for rail track extraction and inspection, which is more meaningful shortly. The Indian Railways has planned to use UAV for security surveillance and passenger safety (22 August 2020, The Economic Times).

Current data collection is restricted to limited regions, so it is lavish and labor demanding to observe railroad behaviors through vast areas. Major railroad problems will go unnoticed because of irregularity in manual track monitoring and maintenance. The various limitations of manual railroad monitoring include less frequent monitoring and maintenance, less accuracy in the data observed, delay in communication, time is taken for monitoring, and limited intelligence. Also, workers are faced with unstable ground, slopes, hidden hazards, low lighting, and barriers to carry out visual monitoring. P.K. Sen et al. [5] suggested that incorporating advanced technology to increase safety and prevent accidents also added that monitoring should be conducted regularly.

In the existing methods, [11, 12] linear features of railway track are detected that provide poor performance due to the varying complex background and lighting conditions. Semantic segmentation [6] is a familiar technique for allocating labels to each pixel in the input images. It comprises an extensive diversity of applications, i.e., traffic management, object detection, urban planning, scene understanding, autonomous driving, road monitoring, etc. Most importantly, this method has set the benchmark on the rail dataset [7] as well. Over the last few years, several techniques have been developed for this intention, with few noticeable instances being the SegNet [8], Fully Convolutional Network (FCN) [9], and DeepLabv3 + [10]. In [13, 14] convolutional neural networks (CNN) based semantic segmentation is proposed for detecting railroad or track elements which exhibit the efficiency of the structure of semantic segmentation in segmenting basic track elements or rails. The latest development of deep learning for segmentation and object detection has shown great promise because of the above.

RSNet is an attempt to realize the powerful deep neural network for railroad segmentation in a more responsible way. The proposed model is an idea to implement multi-level residual feature fusion for railroad semantic segmentation using U-Net-based encoder and decoder architecture. The encoder depends on the efficientnetb4 model [15], used for feature extraction with only six convolution layers and five efficient layers. Likewise, the decoder depends on Sabari et al. [16] who upsampled only important features using a modified residual network based on six attention layers called convolution block attention module (CBAM) [46] that obtain adaptive feature refinement. In addition, developed a railroad scenic view segmentation dataset (RSSD), which consists of 2214 annotated real-time aerial railway track images with varying lighting conditions and complex backgrounds. Using this dataset, many experiments have been done for analysis. Thus, it is evaluated that an attention-based modified convolution network offers efficient and higher performance in terms of accuracy metrics (like dice coefficient, jaccard, and intersection over union) in railroad drone imagery compared to earlier network models.

Three main tasks are performed in this study, which is illustrated in Fig. 1 they are dataset preparation, training, and validation, prediction, and comparison with other existing network models. This paper contributions and novelties are: (1) The creation of new dataset consists of annotations and binary mask creation using color threshold function for the semantic segmentation of the railroad environment. (2) A novel semantic segmentation network model targeting real-time aerial railroad images considering only important features and produces higher performance in the railway environment. Testing results proved that the enhanced model accomplishes superior to the existing models. (3) The different illumination effects of input images with varying backgrounds are predicted. The testing results confirmed the varying background state and illumination impact the model performance; also, the automatic creation of class labelled aerial railroad images will be useful to robust obstacle detection and upgrade the maintenance of railroad amenities.

This paper is organized as follows: Section II discusses the related works. Section III describes the core algorithms. Section IV illustrates the proposed work of the RSNet model, which comprises various steps used for processing the RSSD dataset. Section

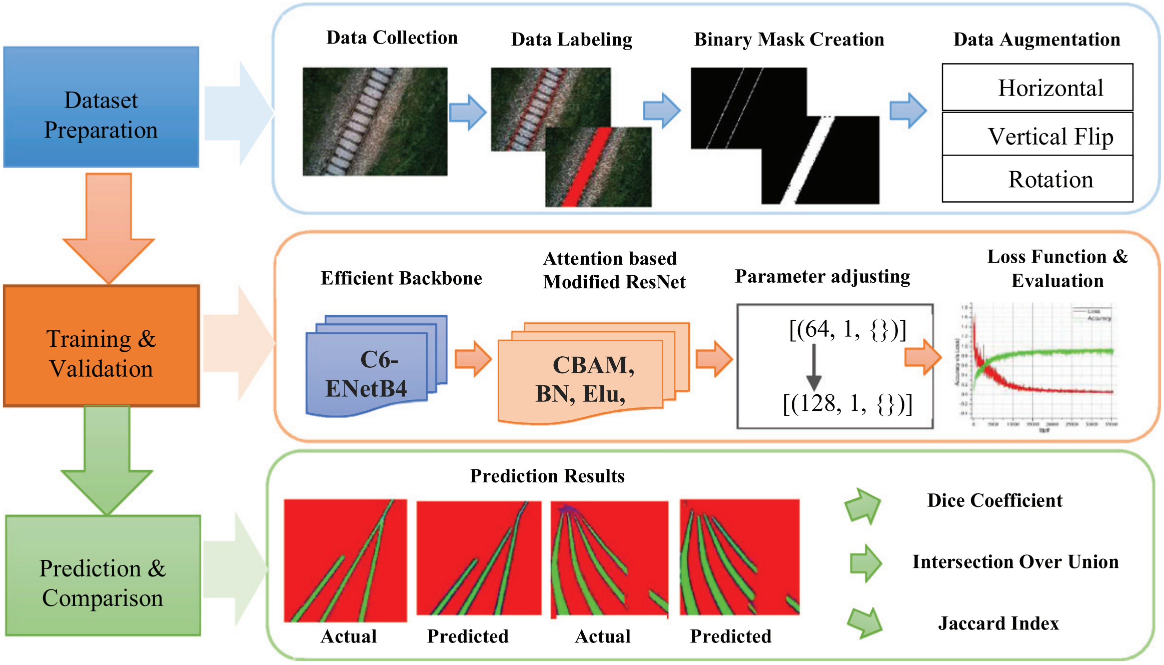


Fig. 1. Pipeline of this study.

V explains the results and discussion with the picture and table-based explanation. Section VI provides the conclusion and future enhancements of the proposed network model.

2. Related work

Traditional computer vision algorithms and few deep neural techniques are used to segment the railway environment, and some related works are associated to this paper directly. RSNet is a dedicated segmentation network intended to extract aerial railway track images consisting of rails and gauge feature extraction and semantic segmentation of the railroad environment. Table 1 describes the comparison of proposed model with other state-of-art models with respect to its contributions, merits and demerits. There are many views of existing mechanisms that are correlated to this study, which are reviewed below.

2.1. Railroad segmentation using computer vision

Railroad segmentation is a powerful module for obstacle detection in the railway track. Due to the shortage of railroad annotation datasets, most of the railroad extraction techniques are depend on geometric characteristics of the railroad and features of computer vision. The general pipeline, followed by the traditional vision technique, identifies the railroad's gradient by edge detection approaches after it detects the last railroad by using Hough transform [42].

Beyond this technique, there are many other techniques available for railroad extraction in recent years. In [17] onboard camera in front of the train videos is used in which railroad track space and train course are extracted using dynamic programming. This method initially computes the gradient of the railway environment image using sobel operator then Hough transform is processed over binary image to detect rail lines. At last, dynamic programming is used to extract the railroad track. In [18] first HOG features are used to establish integral images of the railroad, a region growing algorithm is used to extract railroad images.

Nassau et al. [19] proposed an approach for railroad extract that matches edge features of track for modelling rail pattern as a series of parabola segments. In [20], at first, Hough Transform is applied to identify rail lines and then implement a line clustering algorithm that comprises theta and rho parameters for railroad detection. In [21] superpixels based railroad detection technique is proposed in which TF-IDF (term frequency-inverse document

Table 1
Qualitative comparative analysis of proposed model with other state-of-art models

Related Work	Year	Key Contribution	Merits	Demerits
[21]	2016	To propose superpixels based railroad detection technique.	Compared several pairs of features with and without transformation, which produces detailed outcome.	Only few data's are considered for experimentation that too not emphasis on aerial data's.
[25]	2017	To attempts various transfer learning models with varying parameters that detects railroad	Multilayer filters helps to transmit more information to subsequent layers and tends to better accuracy.	Semantic segmentation is not attained here.
[20]	2017	To propose a railroad detection algorithm tailored for high altitude and large enduring drones.	Obtained reasonable detection results with all tested sequences.	Need to perform parameter adjustment with respect to the drone camera and flight information data.
[13]	2018	To exhibit the efficiency of the structure of semantic segmentation in railroad detection.	Use high-resolution images comprising geographical, user supervision, and thematic constraints.	Not suitable for all weather conditions.
[70]	2018	To propose encoder decoder model which executes pixel wise class predictions.	Encoder exhibits flexibility, which can be modified and trained for arbitrary size of images and minimize the trainable parameters.	Only road scene is explored for semantic segmentation.
[69]	2018	To precisely locate and widely define main part of the railway scene instead of distinct detection.	Extract numerous context details from images and enhance performance of the segmentation.	Role of aerial railroad segmentation is not deliberated here.
[68]	2018	To build network with residual block which has similar features of U-Net.	Residual block and skip connections helps for ease training and information propagation with fewer parameters that leads to better performance.	Concentrate only on aerial road extraction.
[67]	2019	To describe the track boundary area with very few computation burden.	Accurately and automatically restricts the scene boundaries in real time and increases the efficiency.	Aerial railroad extraction is not discussed here.
[66]	2019	To effectively segment scenes along with railway lines.	IoU and segmentations results are higher which has close link with color characteristics and geometry shape.	Accuracy and running time of the model need to be improved.
[65]	2019	To extract multiple convolution features through modified backbone network and pyramid structure.	Increases railroad detection speed without creating more number of regions.	This proposal is computationally intensive.
[64]	2020	To identify the alarm region from segmented rails through changes in vanishing point.	Its processing speed encounters the real time desires.	Very sensitive during influence of light changes conditions.
[63]	2020	To enhance the receptive field between pixels of the rail features.	Network model improves the performance of rail recognition.	Experimented only frontend view of datasets
[62]	2020	To propose framework for high-resolution railroad segmentation.	Delivers robust railway track detection within the requisite distance and inference speed	Does not focus on UAV railroad datasets.
[61]	2020	To utilize data from different frames to cogitate temporal continuity in semantic segmentation.	Class boundaries are clearer to estimate the semantic labels using class likelihoods of corresponding pixels.	Flow estimation error occurred when the number of frames increased that leads to performance degradation.
[60]	2021	To approach cutting-edge performance on lane line segmentation tasks.	Optimal fusion network attains better accuracy and fills the performance gap between the models	No examination about UAV railroad images.
[59]	2021	To capture efficient and detailed semantic features of remote sensing images.	Improves representation ability of extracted features and preservation of spatial details.	Need better optimization on whole segmentation model.
Proposed Model	2021	Targets real-time aerial railroad images considering only important features and produces higher performance in the railway environment	Reusing the convolution at the side branch, which leads to giving efficient accurate results.	Not integrated with the on-board processor. However, this is a future work direction.

frequency) is used for feature extraction and support vector machine (SVM) is employed for railroad classification besides, intracellular decision representation is used to create decisions on a superpixel utilizing feature prediction within the superpixel.

Segmentation of the railroad is identical with lane segmentation under road scene understanding [30] in that lengthy and tinny lane marks could be identified on the road. Mostly, techniques used for railroad segmentation are hired from the lane segmentation. Nevertheless, in the railroad environment, the gap between rails is designed as stricter forced, and side offset is typically continuous, which creates differences in the segmentation' post-processing as mentioned earlier. In the existing lane segmentation techniques, initially ROI is extracted [41] from the road image, then apply suitable edge detection operators such as canny or Sobel for extracting the lane region from the road environment, finally sequence of post-processing aids to produces lane line. Other trademarked algorithms [22–24] have been proposed with additional sensing equipment like Lidar, remote sensing data, or laser to segment the railroad.

2.2. *Railroad segmentation using deep learning*

In the railway environment, different kinds of backgrounds and various lightning effects are present in the input images that disturb the detection and segmentation of the railroad. In the current years, deep learning techniques growth has been gradually implemented for accurate and efficient segmentation and obstacle detection in the railroad environment. For instance, in [25] attempts Fast Recurrent CNN for ROI extraction that detects railroad. In CNN, features are extracted from each stage except the final stage termed the fully connected layer stage, opting for classification. Especially for segmentation and obstacle detection, features extracted from multiple network layers are influenced and provide superior outcomes.

To visualize higher-layer features [26] of segmentation, the sum of partial scores of every class on multiple scale features are unified in the fully convolutional network (FCN). In [27], feature pyramid network is developed to increase extraction in various performance applications. Various other models that use multiuser features are U-Net [28], Laplacian Pyramid [29], etc. Mask RCNN [37] also uses a feature pyramid model to perform human pose estimation and instance segmentation on the COCO dataset. Many researchers have been shown their interest in multiple features that can significantly increase the network model performance.

Like lane segmentation, railroad segmentation needs the segmentation of the region that comprises of railroad alone. In [38], dense up sampling convolution (DUC) module is implemented for segmentation that gives better resolution. The DUC module is also used in [39] for pixel-level prediction and proposed hybrid dilated convolution (HDC) framework for enlarging the receptive fields of the model that leads to group the global variables. In [40], spatial information between features is identified using a conditional random field, enabling greater receptive field and robust spatial connections. However, this recursive process restricts the speediness of the entire network.

In [16], FCN is used to execute image segmentation, whereas in [8], the encoder and decoder network models are used to increase the image resolution of segmentation. Without losing any resolution, dilated convolution [31] is used for the systematic aggregate multi-scale feature for contextual information. The Atrous separable convolution method [32] is proposed for robust object segmentation at multiple features. Likewise, a pyramid scene parsing network (PSPNet) [33] is proposed, which handles global context information. Fast and effective semantic segmentation for videos is proposed in [34]. Besides, some sequences of semantic segmentation solutions that improve the model's performance are proposed in [35, 36].

Though the railroad comprises stricter constraints such as shape and features, the segmentation faces more challenging scenes like ground embedding, weed coverage, track crossing, etc. Besides, the recognition of robust railroad segmentation is a greater task that needs to be achieved through multilevel residual feature fusion [71]. This paper is inspired by more efficient modules such as efficient net, convolution block attention module, and modified residual UNet, which consider only important features and at the same time, provides higher performance.

3. Algorithm overview

The RSNet is an attention-based encoder and decoder U-Net structure. As mentioned before, the encoder includes six convolution layers of the efficientnetb4 model, and the decoder uses a modified residual network based on six attention layers that consider only important features for classification. Input images are initiated by an encoder algorithm and treated over a set of convolution layers followed by efficient layers. The encoder network is based on the inverted residual block with a squeeze excitation phase that improves performance and computational efficiency. Next, the attention layer called CBAM is used for refining adaptive features. Finally, the decoder algorithm processes the feature maps generated by CBAM using a set of deconvolution and modified residual layers.

3.1. Algorithm 1

RSNet Training: Training starts with $X_I \times H_I \times W_I$. During training, the required significant additional parameters are several filters (K), Stride rate (S), and zero paddings (P). The outcome of this network is $X_O \times H_O \times W_O$

Prerequisite: (1) Conv() and Deconv() function performs convolution and deconvolution processes respectively with constant rate of filter_size = 2, stride_rate = 2 and zero_padding = 1. (2) Then the activation function ELU() and Batch Normalization (BN()) processes are carried out for the given parameters. (3) In addition, MPool() function performs the max pooling with hyperparameters stride rate = 2, filter size = 2 and padding = 0. (4) The core backbone function EffNetB4() is used for scaling all the attributes and improves the performance of the model and filters the important features. (5) CBAM() is utilized as the intermediate layer of the encoder and decoder part. (6) M_Residual() function which is used to improve the accuracy of shallow sub-network using skip connectors. (7) PSPBlock() is implemented, which comprises basic functions such as average pooling with convolution, BN(), and RELU activation function. (8) Finally, concatenate the outcome of the PSP block with the outcome of the top M_Residual layer and perform addition for the concatenated output and attention layer.

Uphold: Weight updating W^{i+1}, learning rate updating η^{i+1}, batch normalization parameters is updating by $\ominus^{i+1}$ and drop-out updating D^{i+1}.

Input: $X_I \times H_I \times W_I$ size image

Output: Rail Segmented output image of size $X_O \times H_O \times W_O$

```
Begin
InputX_I × H_I × W_I size image
for k = 0 to L − 1 do
    $W_{ek} \leftarrow ELU(W_k)$
    $W_{bk} \leftarrow BN(W_{ek})$
    $W_{c1k} \leftarrow Conv(W_{ek,} W_{bk})$
    $W_{enk} \leftarrow EffNetB4(W_{c1k})$
    $W_{mk} \leftarrow MPool(W_{enk})$
    if $k \leq L$ then
        $W_{c2} \leftarrow Conv(W_{mk})$
        $W_{r1} \leftarrow M\mathrm{Re}sidual(W_{c2})$
        $W_{cb1} \leftarrow Cbam(W_{r1})$
        $W_{r2} \leftarrow M\mathrm{Re}sidul(W_{cb1})$
    end if
end for
for $k = L - 1$ to $K > 0$ do
        $W_{dk} \leftarrow DeConv(W_{r2k})$
        $W_{cb2k} \leftarrow Cbam(W_{mk})$
        $W_{cc1k} \leftarrow Concat(W_{dk,} W_{cb2k})$
        $W_{c3k} \leftarrow Conv(W_{cc1k})$
        for x in range 2
            $W_{rkx} \leftarrow M\mathrm{Re}sidual(W_{c3kx})$
        end for
$W_{pk} \leftarrow PSPBlock(W_{rkx})$
end for
    $W_{cc2} \leftarrow Concat(W_{pk}, W_{rkx})$
    $W_{cb2} \leftarrow Cbam(W_{cc2})$
    $W_a \leftarrow W_{cb2} \oplus W_{cc2}$
return$W_a$
END
```

3.2. Algorithm 2

EffNetB4() [15]: The main module weak layer is the mobile inverted residual block, discussed below with the squeeze excitation function.

Prerequisite: Kernel size (K), Input Filters (I_f), Output Filters (O_f), Expand ratio (E_r), Stride rate (S), and Squeeze Excitation ratio (SE_r)

Uphold: Weight updating W^{i+1}, learning rate updating η^{i+1}, batch normalisation parameters is updating by $\ominus^{i+1}$ and drop-out updating D^{i+1}.

Input: $X_I \times H_I \times W_I$ size image

Output: Feature map of size $X_E \times H_E \times W_E$

```
Begin
for $k = 0$ to $L - 1$ do
 //Expansion Phase
    $W_{efk} \leftarrow ExpandedFilter(W_{I_f} \times W_{E_r})$
    $W_{ck} \leftarrow Conv(W_{efk})$
    $W_{b1k} \leftarrow BN(W_{ck})$
    $W_{s1k} \leftarrow SWISH(W_{b1k})$//Depthwise Convolution Phase
    $W_{dk} \leftarrow DWConv(W_{s1k})$
    $W_{b2k} \leftarrow BN(W_{dk})$
```

```
        $W_{s2k} \leftarrow SWISH(W_{b2k})$//
    Squeeze and Excitation Phase
        $W_{sek} \leftarrow f_{avg_{pool}}(W_{s2k})$
        $SqueezeFilter \leftarrow Max(1, (I_f \times SE_r)$
        $W_{sek} \leftarrow Conv(SqueezeFilter, W_{sek})$
        $W_{sek} \leftarrow SWISH(W_{sek})$
        $W_{sek} \leftarrow SIGMOID(W_{sek})$
        $W_{sek} \leftarrow W_{s2k} \otimes W_{sek})$//Output Phase
        $W_{sek} \leftarrow Conv(W_{sek})$
        $W_{sek} \leftarrow BN(W_{sek})$
    end for
    return $W_{sek}$
    END
```

3.3. Algorithm 3

CBAM()[46]: It computes an intermediate feature map $M(F) \in X_E \times H_E \times W_E$ and generates channel attention map $M_c(F) \in X_A \times 1 \times 1$ and spatial attention map $M_S(F) \in 1 \times H_A \times W_A$

Prerequisite: Input Feature map (I_f), Weight (W), bias(b) are initialized.

Uphold: Weight updating W^{i+1}, learning rate updating η^{i+1}, batch normalization parameters is updating by $\ominus^{i+1}$ and drop-out updating D^{i+1}.

Input: $I_f \in X_E \times H_E \times W_E$

Output: Refined outcome $F \in X_A \times H_A \times W_A$ which comprises of $M_c(F) \in X_A \times 1 \times 1$ and $M_S(F) \in 1 \times H_A \times W_A$

```
    Begin//
    Channel Attention Map
    for $k = L - 1$ to $k > 0$ do
        $W_{sl1k} \leftarrow Dense(ELU(I_{fk}), b_0)$
        $W_{sl2k} \leftarrow Dense(I_{fk}, b_1)$
        $W^c_{ak} \leftarrow f^c_{avg\_pool}(I_{fk})$
        $W^c_{ak} \leftarrow W_{sl1}(W^c_{ak})$
        $W^c_{ak} \leftarrow W_{sl2}(W^c_{ak})$
        $W^c_{mk} \leftarrow f^c_{\max\_pool}(I_{fk})$
        $W^c_{mk} \leftarrow W_{sl1}(W^c_{mk})$
        $W^c_{mk} \leftarrow W_{sl2}(W^c_{mk})$
        $M_{ck}(F) \leftarrow \sigma(W^c_{ak} + W^c_{mk})$
        $F' \leftarrow M_{ck}(F) \otimes I_{fk}$
    end for// Spatial Attention Map
    for $k = L - 1$ to $k > 0$ do
        $W^s_{ak} \leftarrow f^s_{avg_{pool}}(I_{fk})$
        $W^s_{mk} \leftarrow f^s_{\max\_pool}(I_{fk})$
        $M_{sk}(F) \leftarrow \sigma[Conv_{7\times7}(W^s_{ak}, W^s_{mk})]$
        $M_{sk}(F) \leftarrow M_{sk}(F) \otimes I_{fk}$
    end for
    $F'' \leftarrow M_{sk}(F) \otimes M_{ck}(F)$
    return $F''$
    END
```

3.4. Algorithm 4

M_Residual() [16]: Instead of learning the true distribution of output, the modified2 residual layer learns residual of the outcome.

Prerequisite: Encoder network with attention layer extracts mini-batch of feature attention maps $X_A \times H_A \times W_A$. Along with attention maps BN(), W, and Filter Count(FC) should be initialized.

Uphold: Weight updating W^{i+1}, learning rate updating η^{i+1}, batch normalization parameters is updating by $\ominus^{i+1}$ and drop-out updating D^{i+1}.

Input: Input feature map $M(F) \in X_A \times H_A \times W_A$

Output: The outcome will be $X_R \times H_R \times W_R$

```
    Begin
    Input $I_b$, $FC$
    for $k = L - 1$ to $k > 0$ do
        $W_{ek} \leftarrow ELU(I_{bk})$
        $W_{bk} \leftarrow BN(W_{ek})$
        for $W_{bk}$ in range 2
            $W_{rk} \leftarrow Conv_{3\times3}(W_{bk}, FC)$
        end for
    $W_{sb_rk} \leftarrow Conv_{3\times3}(I_b, FC)$
    $W_{rk} \leftarrow W_{rk} + W_{sb_rk}$
    end for
    return $W_{rk}$
    END
```

4. System design

Traditional techniques use edge or line features to detect the railroad maybe got good outcomes for a fixed scene if different scenes occur, and performance gets degraded. Especially in drone monitoring over some kilometers, the background is changed, and the manual feature extraction is unable to meet the desires. Thus the detection of the railroad is converted from edge detection or line detection into railroad segmentation problem. In this study, RSNet a multi-level residual feature fusion semantic segmentation model is proposed, which combines the strengths of efficientnetb4, U-Net, modified residual block, side branch layers and attention layer called CBAM that accurately identifies railroad border structure. This combination provides five major benefits 1) efficient layer makes a good design with much fewer parameters, and 2) attention layer is allowed to choose only significant features for post-processing 3) the modified residual unit provided ease training of the network; 4) the skip connections between low and high levels of the network and within a modified residual unit will enable information propagation without any performance degradation and achieve better performance on semantic segmentation 5) multi-level residual feature fusion is employed through side branches, concatenation, an attention layer and add it to achieve high performance accurate results. Here, efficientnetb4 is used as an encoder that adapts a backbone architecture that comprises compound scaling for improving computational efficiency, and the attention-based modified residual layer is used as a decoder that chooses only important features during the process and produces good accuracy for the real-time aerial railroad images compared to other existing related models.

An 11 – stage architecture of RSNet with an attention layer for railroad extraction is utilized, as shown in Fig. 2. Specifically, the RSNet model comprises of three main functions: backbone (feature extractor) an encoding function, attention layer (CBAM) serves as a bridge between encoding and decoding functions, and modified residual layer a decoding function. The first function encodes the input image into dense illustrations consisting of 3×3 convolution blocks and different settings of the efficient layer. The middle function attends like a bridge connecting the backbone and decoder functions which is worth grasping unique features. The last function improves semantic segmentation representation, i.e., pixel-wise categorization through upsampling feature maps from the lower unit and creates concatenation with feature maps from the matching encoding way. This last function comprises a 4×4 deconvolution block concatenated with the attention layer's outcome, and the result is passed to identity mapping. Each convolution block includes the ELU activation layer, Batch normalization, and convolutional layer. After that, the last decoding unit, a set of a process is performed in each PSPBlock (mentioned in Algorithm 1) followed by a concatenation of its result with the last decoding layer. Finally, it is fused with the attention layer that proposes the multiple channel feature maps into the required semantic railroad segmentation. The parameters and output size of each stage are presented in Table 2. The further explanation about the main modules is as follows.

4.1. The efficient layer

Encoder part plays a major role in extracting the features in which this efficient net layer acts as a backbone along with the convolution block. The fea-

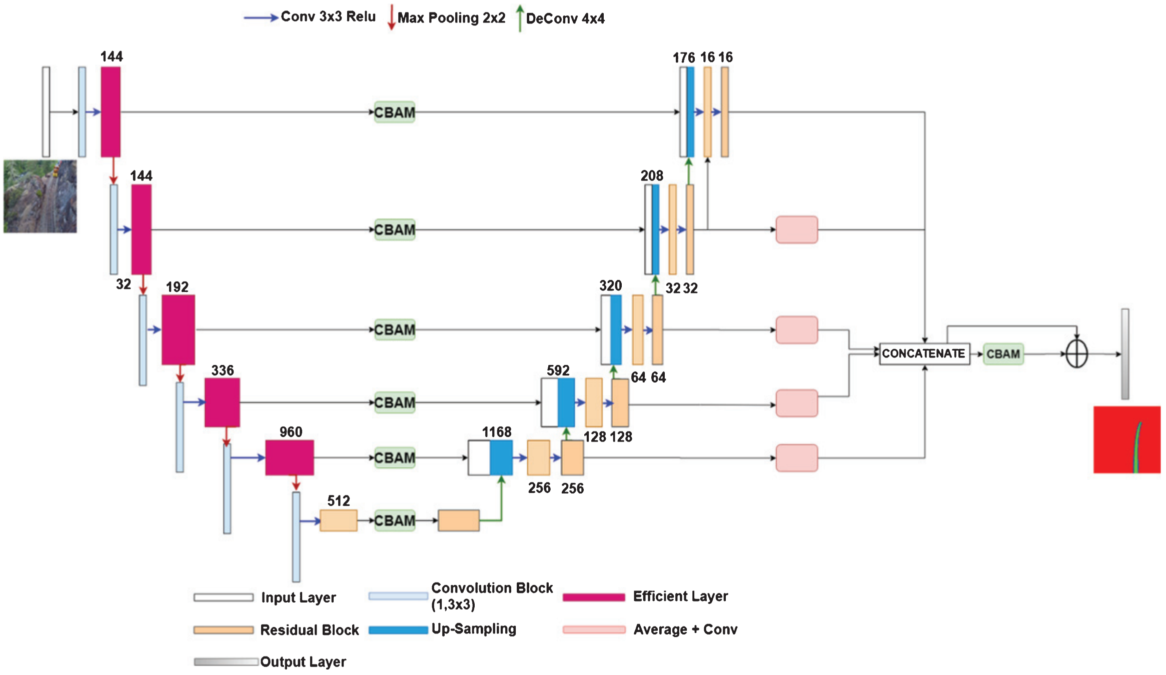


Fig. 2. Main design of the RSNet Model.

Table 2
Network structure of RSNet

	Stages	Filter size	Stride	Output size
Input				384×384×3
Encoder- Conv with EfficientNetB4	1	3×3 / 32	1	384×384×32
	2	3×3 / 144	1	192x192×144
	3	3×3 / 192	1	96x96×192
	4	3×3 / 336	1	48x48×336
	5	3×3 / 960	1	24x24×960
Bridge – CBAM	6	3×3 / 512		12x12×512
Decoder – Upsampling and Modified Residual Block	7	4×4 / 1168/ 256	2	24x24×256
	8	4×4 / 592/128	2	48x48×128
	9	4×4 / 320/64	2	96x96×64
	10	4×4 / 208/32	2	192x192×32
	11	4×4 / 176/16	2	384x384×16
Output				384x384x3

ture extraction intention is to influence the robust feature extraction potential of a CNN trained on the given dataset. Compared to widely used VGG16 [43] and ResNet50 [44], efficientnetb4 is an efficient choice of backbone feature extraction network to scaled up for better accuracy and increases top-1 accuracy up to 83% with fewer parameters (19M), and thus it is computationally efficient. Moreover, VGG16 and ResNet50 use traditional manual scaling methods such as model scaling and depth-wise scaling in which scaling works pretty good, but after a certain limit, it does not increase the performance of the model and also starts to degrade the performance adversely. The compound scaling method is used in the proposed backbone layer, which scales the attributes such as depth, width, and resolution using a compound co-efficient ϕ. The mathematical formulation of scaled attributes is given in Equation (1)

$$\begin{aligned} &\text{Depth: } d = \alpha^{\phi} \\ &\text{Width: } w = \beta^{\phi} \\ &\text{Resolution: } r = \gamma^{\phi} \end{aligned} \tag{1}$$

where, $\alpha\,\beta^2\,\gamma^2 \approx 2$ and $\alpha \geq 1, \beta \geq 1, \gamma \geq 1$. Figure 3 illustrates the main structure of the EfficientNet-B4

4.2. Attention layer

CBAM is used to build CNN to learn and emphasize the railroad information, rather than learning unwanted background information. It serves as a bridge between encoder and decoder functions. This is applied to channel and spatial dimensions sequentially, in which the process consumes a C×H×W feature map as an input and generates 1×H×W and C×1×1 as output attention maps. At last, these output maps perform element-wise multiplication with input feature map to get a more adaptive refined and highlighted output.

$$F''' = M_s(F'') \otimes M_c(F) \otimes F \tag{2}$$

$\otimes$ – The element-wise multiplication and innermost method of channel attention module Mc and spatial attention module Ms is formulated as follows.

$$M_c(F) = \sigma(MLP(F^c_{avg}) + MLP(F^c_{\max})) \tag{3}$$

$$M_s(F) = \sigma(Conv7 \times 7(F^s_{avg}; F^s_{\max})) \tag{4}$$

MLP refers to a multi-layer perceptron, a shared network with one hidden layer, F^c_{avg} $F^c_{\max}$ representing average pooling and max pooling respectively for channel attention map. Similarly F^s_{avg} and $F^s_{\max}$ signifies average pooling and max pooling concerning spatial attention map. σ denotes sigmoid function and $Conv7 \times 7$ is a convolution process with filter size 7×7.

4.3. Modified residual layer

Each layer supplies the next layer in the encoder function, whereas in the modified residual layer, each layer supplies the next layer and directly passes to the layers that are away from two to three hops. Already encoder function and attention part comprise of many layers. Besides, the decoder function encourages con-

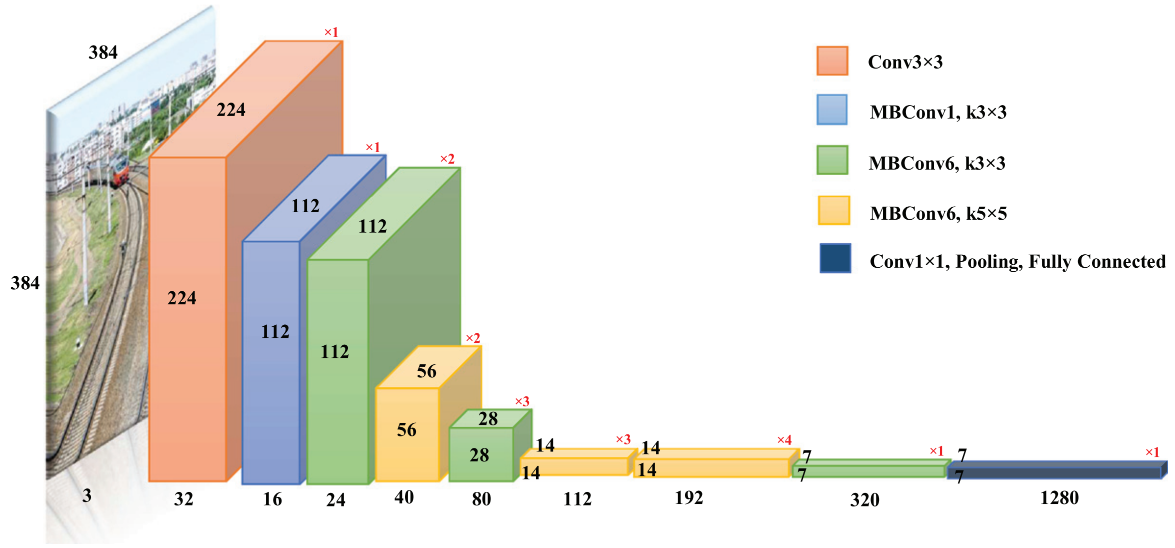


Fig. 3. Main structure of EfficientNet-B4.

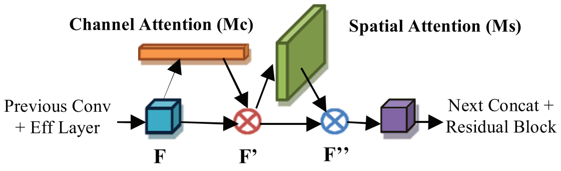


Fig. 4. Attention Layer CBAM integration with the RSNet.

catenation of upsampling and the outcome of CBAM. If the number of layers is getting increased, then the accuracy becomes degraded. To solve this problem, a skip connection is introduced that skips the training of a few layers. This is illustrated in Fig. 4, where the identity function has relied on the skip connection.

$$O(x) = R(x) + S(x) \tag{5}$$

Many state-of-art techniques have been proved the accuracy improvisation in residual units. The modified residual layer used non-identity mapping for repeatedly infusing the missing information of the last layer to the network. This reinstated information can be used to enrich the performance of semantic segmentation. Here, Exponential Linear Unit (ELU) (0.1) is used as an activation task in the modified residual block as enthused by [45].

4.4. Loss function

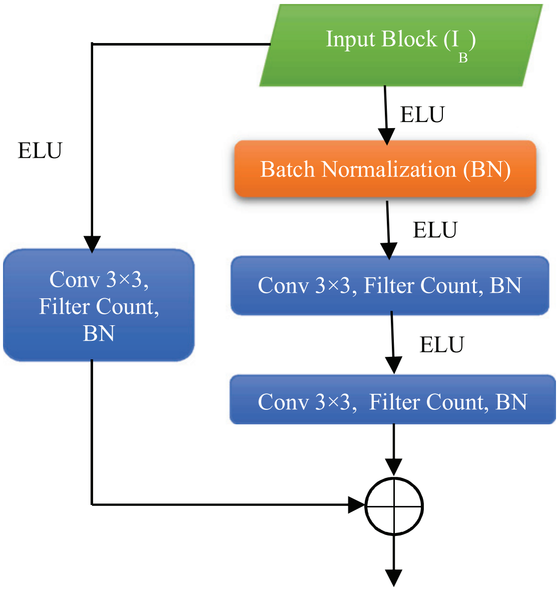

Input Block (I_B)
ELU
Batch Normalization (BN)
ELU
Conv 3×3, Filter Count, BN
ELU
Conv 3×3, Filter Count, BN
ELU
Conv 3×3, Filter Count, BN

Fig. 5. Structure of Modified Residual Block.

The total loss function of this work is defined in Equation (6) as the sum of categorical cross-entropy loss L (in Equation (8)) and channel-wise dice loss (in Equation (7)) [47] such as dice loss channel zero, dice loss channel one and dice loss channel two which represents three different classes like background, gauge (rail inside part) and rails. The ultimate goal of this loss function is to minimize the softmax dice loss during the RSNet model training.

$$\begin{aligned} Loss &= L + DiceLoss(C_0) + \\ & DiceLoss(C_1) + DiceLoss(C_2) \end{aligned} \tag{6}$$

$$DiceLoss(C_i) = 1 - \frac{2\sum_{n=k}^{i=0} y_i p_i + \in}{\sum_{n=k}^{i=0} y_i + \sum_{n=k}^{i=0} p_i + \in} \quad (7)$$

$$L = -\sum_{j=0}^{M}\sum_{i=0}^{N} y_{ij} \log p_{ij} \quad (8)$$

y_{ij} The binary mask, i.e., target images, p_{ij} is the predicted mask images and ε is the coefficient for ensuring the reliability of defined loss function.

5. Experimental results and discussion

In this section, the first railroad dataset and annotation of this dataset is described then data augmentation is discussed for increasing the number of datasets that provides greater accuracy followed by implementation details of the network training and validation is explained after that the performance of the RSNet is evaluated using various metrics for railroad semantic segmentation. Moreover, the comparison of the proposed network model with other related models in our dataset is discussed. Finally, this section ends up with a discussion of various results obtained for railroad detection.

5.1. Dataset preparation

The proposed methodology for railroad semantic segmentation is based on supervised machine learning techniques under computer vision. This technique needs an aerial railroad dataset with annotated images. This dataset is used for training the RSNet model. To the best of our knowledge, a dedicated and easily approachable dataset for aerial railroad segmentation does not become available. Hence, we began off to generate a dedicated dataset from scratch by gathering online free access aerial railroad images and images from YouTube public available aerial railroad videos [48], which are taken by different UAVs.

To train and validate the RSNet model, an aerial railroad dataset calls it a railroad scenic view segmentation dataset (RSSD) built, which includes 2214 annotated real-time aerial railway track images with various resolutions. This dataset is created considering diversity on complex backgrounds and varying light conditions, i.e., color tones and brightness.

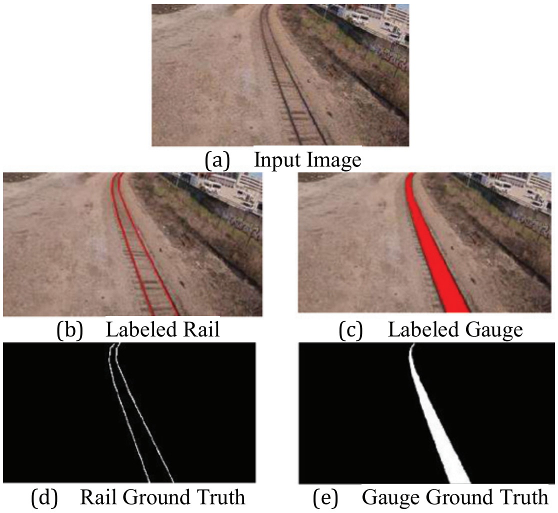

(a) Input Image
(b) Labeled Rail
(c) Labeled Gauge
(d) Rail Ground Truth
(e) Gauge Ground Truth

Fig. 6. An example of labeled and ground truth aerial railroad image.

The collected video frames are annotated using frame cropping and preprocessing. In frame cropping, the required frame is cropped using a snipping tool after that preprocessing step is carried out in that each image is labelled as two classes such as gauge and rails using Adobe Photoshop. Mainly, these two classes are generated for obstacle detection, which is occurred nearby rails. This work is primarily done for implementing obstacle detection in the future.

To make a binary mask for the labelled images, color threshold application is used, which is provided by Matlab 2016a. An example of a labelled and ground truth image is shown in Fig. 6. Based on the best practices for segmentation models and the general ratio of cross-validation principle; [49, 50] from the total dataset, the training dataset size is 80%, and the test dataset size is remaining 20%. Specifically, in the RSSD dataset, 1,660 annotated images are in the training dataset, and 554 images are in the test dataset.

Data augmentations are used to enhance the training samples during training the RSNet model, which prevents overfitting of the training images. RSSD dataset is augmented using the argumentation library that was started as a firm and adaptable execution for image augmentation [51]. The affine transformations from this library include horizontal flip, vertical flip, and rotation are utilized to obtain augmented images.

5.2. Training and validation

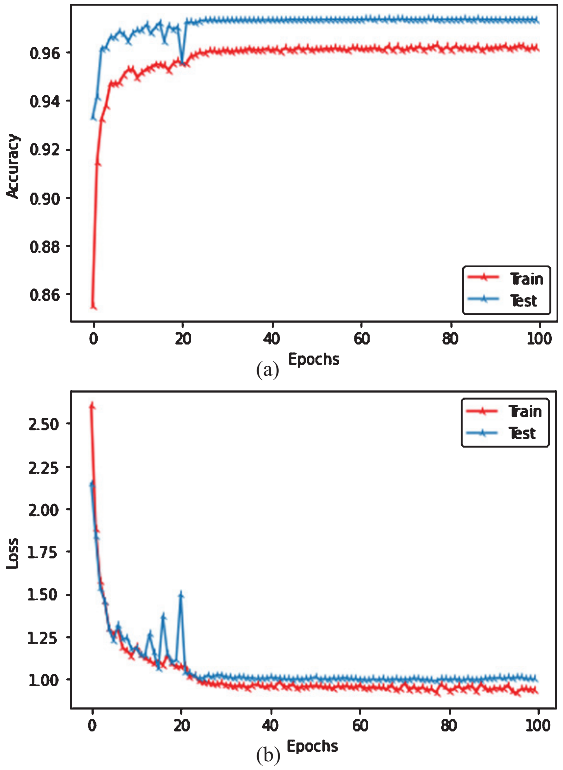
A pre-trained model is a suitable timesaving technique to train deep neural networks. Since various neural networks require to be trained and assessed, pre-trained networks aids to provide good test results better than training each network from scratches that are engaged in the evaluation. During model initialization itself, the pre-trained weights are executed for the backbone of the proposed model. The pre-trained weight file desires to be updated with the RSNet structure using its key and values because the backbone is adapted in the model.

Table 3 illustrates the training hyperparameters of the proposed model. The training process mostly reduces the overall loss by implementing the optimization over the model parameters [52] that is minimizing the overall loss leads to maximizing the performance of the model. In this study, a familiar combination of momentum and root mean square propagation heuristics called Adaptive Momentum (Adam) Optimization algorithm is employed to train the RSNet model. Table 2 illustrates training hyperparameters, the number of epochs required for both training and validation is 1k, and learning rate plays a major role in the performance of the model in which the range of learning rate varies from 1×10^{-3} to 1×10^{-5}.

The RSNet is executed using Tensorflow as the backend with Keras Library. An NVIDIA Geforce RTX 2070 GPU is utilized in a Windows P.C. with an Intel i7 processor to accomplish the training process. Then the weights of the whole network are initialized by two phases in which the first phase is to initialize the weights of the RSNet backbone network, i.e., EfficientNetB4 using the weights which are trained by the ImageNet dataset. GitHub provides the weights of mentioned backbone model. The next step is to initialize other parameters randomly. To improve the parameters of efficient layers, the RSSD training set is used for training the entire network.

Table 3
Training hyper parameters of RSNet Model

Parameters	Value
Input Size	384×384
Learning rate	One$\times10^{-3}$ to 1×10^{-5}
Epochs	1000
Batch Size	2
Number of Channels	3
Model Parameters	32,817,601
β1, β2, ε in Adam	0.9, 0.99 and 1e-10
Weight Decay	1e-4
Output Size	384×384



(a)

(b)

Fig. 7. Representative training and validation of accuracy and loss of RSNet Model.

Figure 7(a) shows the proposed model's training and validation accuracy on a randomly selected training and testing set. The plot of accuracy defines that the RSNet model provided a well-trained and better performance on the trend for accuracy on the RSSD dataset, and it is still maintaining its state till the last epochs.

Likewise, Fig. 7(b) illustrates the training and validation loss of the model on a randomly selected training and testing set in which the model shows comparable performance on both train and test datasets. Also, the loss of both training and testing values decreases until the last epochs that lead to an efficient solution. The RSNet model holds good training and validation accuracy value as 96.17 and 97.33, respectively.

5.3. Performance evaluation metrics

In this study, the performance of RSNet is evaluated using common semantic segmentation evaluation metrics such as dice coefficient, Intersec-

Table 4
Comparison of evaluation of the proposed model with other state-of-art models

Sl. no.	Network	Dice coefficient	IoU	Jaccard index
1.	Residual Squeezed U-net [72]	0.945	0.472	0.907
2.	Residual U-Net [68]	0.951	0.475	0.915
3.	PSPNet [33]	0.964	0.482	0.934
4.	SegNet [8]	0.968	0.482	0.934
5.	RSNet [Proposed]	**0.973**	**0.486**	**0.949**

tion over Union (IoU), and Jaccard index [53]. The arguments of evaluation metrics are defined as N_{TP}, N_{FN}, and N_{FP} which represents the number of true positive that is number of railroad pixels are identified correctly, number of false-negative that is number of railroad pixels are missed to identify, and the number of false-positive that is number of railroad pixels identified wrongly. This way, the following evaluation mathematical formula is defined where k is the number of test images. The IoU metric is used to quantify the overlap percentage between the ground truth and prediction result by using the above Equation (9).

$$IoU = \frac{\sum_k N_{TP}^i}{\sum_k \left(N_{TP}^i + N_{FN}^i + N_{FP}^i\right)} \tag{9}$$

$$DiceCoefficient = \frac{2\sum_k N_{TP}^i}{\sum_k \left(2N_{TP}^i + N_{FN}^i + N_{FP}^i\right)} \tag{10}$$

Dice coefficient is one of the important evaluation metrics used to quantify the proposed railroad semantic segmentation model's performance. To validate the model dice score is estimated, which is a measure of defining how similar the channels are. It divides the size of the overlap of the two correctly identified segmentations by the total size of two channels, and it is defined in Equation (10).

$$JaccardIndex = \frac{\dfrac{2\sum_k N_{TP}^i}{\sum_k \left(2N_{TP}^i + N_{FN}^i + N_{FP}^i\right)}}{2 - \dfrac{2\sum_k N_{TP}^i}{\sum_k \left(2N_{TP}^i + N_{FN}^i + N_{FP}^i\right)}} \tag{11}$$

$$= \frac{DiceCoefficient}{2 - DiceCoefficient}$$

Jaccard Index (defined in Equation (11)) and Dice both are risk minimization metrics used to reduce the learning optimization objective during training and evaluate the performance at test time. In Fig. 8, the various dice coefficient in terms of rails, gauge and the overall proposed model is represented in which the dice coefficient of the proposed model is 0.97. Similarly, the dice coefficient of rails and gauge is increased as 0.56 and 0.85 respectively, leading to an effective semantic segmentation of railroad with two different classes as rails and gauge.

Table 4 shows the results of validation data concerning the considered performance evaluation metrics. RSNet model achieves 0.97 dice coefficient and 0.94 Jaccard respectively, which leads to a good outcome. Thus, it is confirmed that the proposed model has attained the highest score compared to other state-of-art models. Figure 8 shows the training and validation of dice coefficient of each class that produces better accuracy till the last epochs.

Each year, more powerful and new network models emerged in the energetic research fields, namely computer vision and deep learning. During processing, all the features in the layers are considered at each stage, leading to computational workload. RSNet model key idea is to consider only important features during dataset processing with the help of attention layers that will not affect the accuracy and performance. Yin et al. [54] made full use of top-level, abstract features to bottom-level-specific features. It offered higher performance, however, led to a computationally intensive one. As stated at the start of this study, the proposed model's fundamental aim is to offer less overhead with a higher performance model that executes even for real-time railroad drone imagery. All features processing at each stage can slow down and increase the system's size, which is illustrated with [55, 56] model design features.

5.4. Results and discussion

Mainly this paper considers the semantic segmentation of aerial railroad images. Figure 9 visually outlines some of the railroad semantic segmentation results compared to other models such as residual block [57] and residual squeeze [58]. All the test aerial railroad images are derived from the test

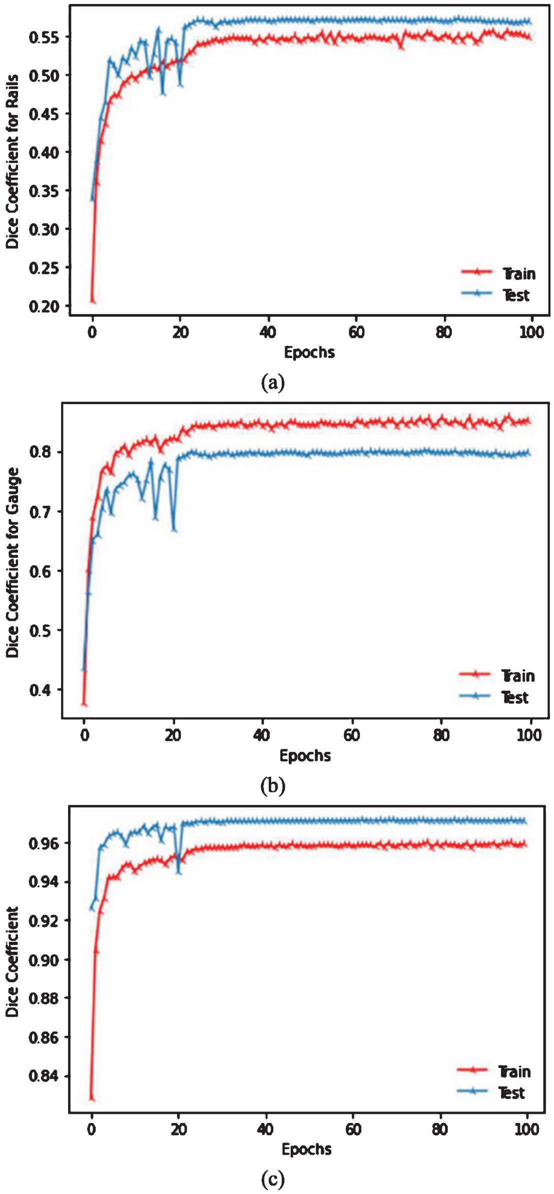


Fig. 8. Representative training and validation of dice coefficients of rails (a), gauge (b), and overall effect (c) of the proposed model.

dataset of RSSD. This figure illustrates the class-based semantic segmentation of a raw input image. The proposed model's high efficiency is reflected in the predicted image that comprises of sharp boundary line and turnouts, which can be utilized for efficient obstacle detection which is found nearby rails or gauge since it causes derailment.

RSNet model offered a perfect boundary outlining and good quality results compared to the conventional results. The figure illustrates that the predicted images were effectively segmented into diverse coloured objects as rail, gauge, and background, and it is recognized through the colour code set at the foot of the image. In the predicted images, it is cleared that the boundary illustrates the model which was not predicting any incorrect labels. The major reason for railroad semantic segmentation uncertainty can be the influence of lighting conditions. Thus the proposed RSNet offers more certainty levels in most of the lighting conditions.

These results exhibit that the RSNet model produces consistent qualitative results for the major two classes like rails and gauge, even at far distances in the scene.

The proposed model provides coarser predictions for segmentation even it is captured from a far distance. As stated before dice and Jaccard metrics used in the quantitative outcomes is a challenging metric that takes in to account the influence of lighting conditions and diverse backgrounds, but it does not affect the fact that the total dice score is over 97%, which can be well appreciated in the qualitative results. However, a failure case is illustrated in the yellow box of the second row. This method missed turnouts mask at the vanishing point. This is due to grouped turnout rails are not labelled.

Visual monitoring is very challenging one in field practice especially when the environmental conditions are complex. Hence, to assist tough environmental conditions in such field applications, segmentation model has to be dynamic and potential. Light condition is one of the typical challenge in which influence of various light conditions images are processed, segmented using the proposed model and illustrated in Fig. 10. Moreover, the prediction time of the proposed work is only 0.071 second, which is reasonable one.

6. Conclusion

This paper offered the idea of RSNet, a serious attention based encoder and decoder UNet model for semantic segmentation. The vital objective is to design an RSNet network that constructs productive attention based CNN for aerial railroad segmentation. Our proposal of a novel multi-level feature fusion network model that takes complete merit of modified residual model with skip connections that segments effective multi-modal CNN features for railroad semantic segmentation. With a single GPU, the proposed model facilitates an efficient training

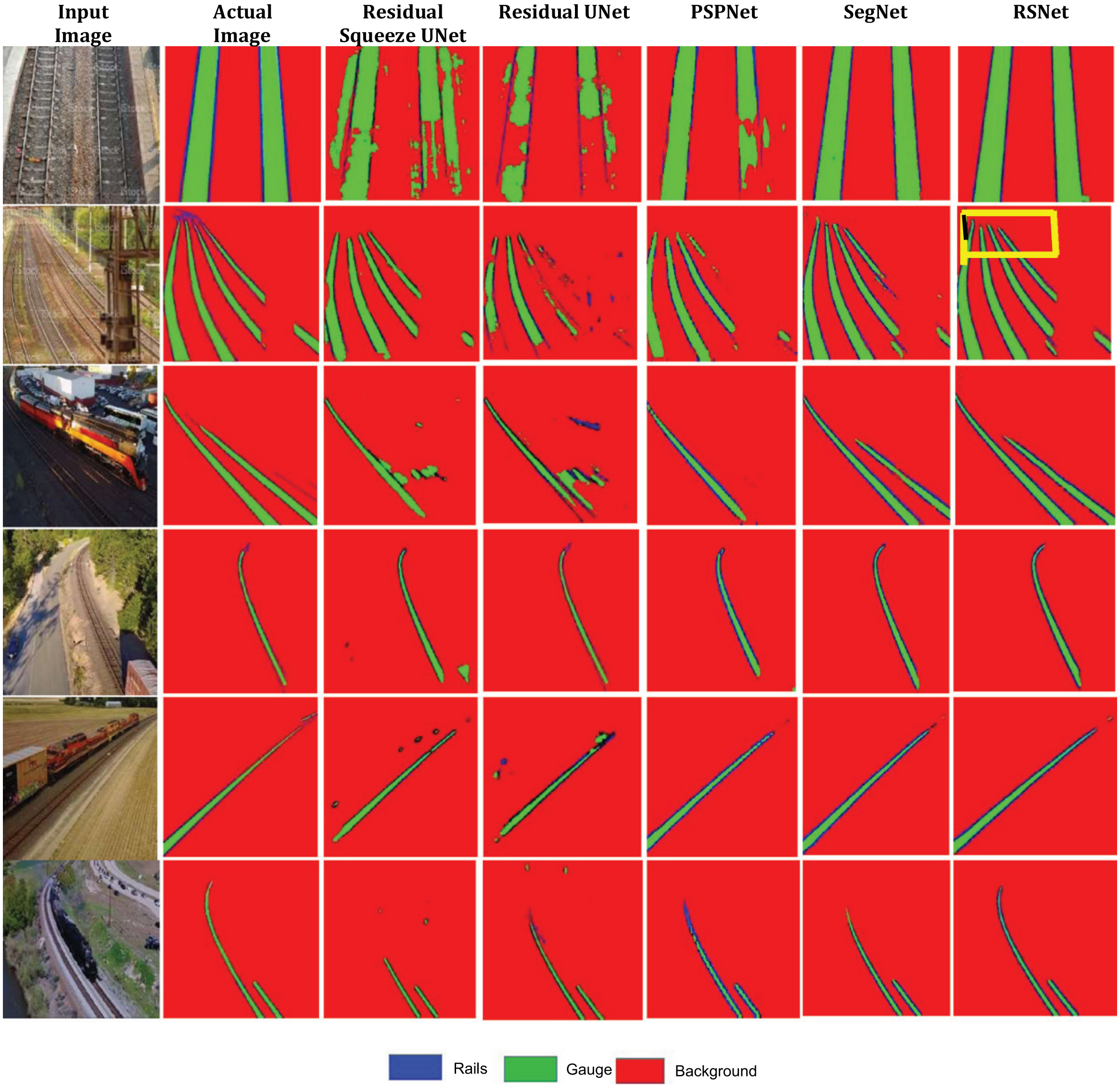


Fig. 9. Qualitative visual comparison results on the RSSD test set. From left to right: Original input images with the influence of the light condition, the actual image, results of Residual Squeeze Layer, Residual UNet Layer results, results of PSP Net, SegNet results and results of the proposed RSNet model.

of deep CNN features. The additional core idea is to propose better attention based CNN model, that considers only important features during training and validation of datasets. This results in higher performance in terms of storage requirements and accuracy. RSNet model is analyzed and evaluated its performance with former state-of-art variants concerning efficiency. It is evaluated that RSNet provides an abundant contract of improved performance above former outcomes. A special technique called modified residual layer is used for reusing the convolution at the side branch, which leads to giving even better accurate results. The proposed RSNet model is well appropriated for aerial railroad segmentation and obstacle detection. It could be employed for avoiding derailments to offer passenger safety and, more generally, railroad safety. This model is appropriate to train heterogeneous data with image-level class labels or pixel-wise segmentation annotations. The design of RSNet offers a deep learning attention based automatic aerial railroad segmentation model that can operate with diverse lighting conditions and backgrounds. Overall, the proposed result is a superior performing semantic segmentation system for futur-

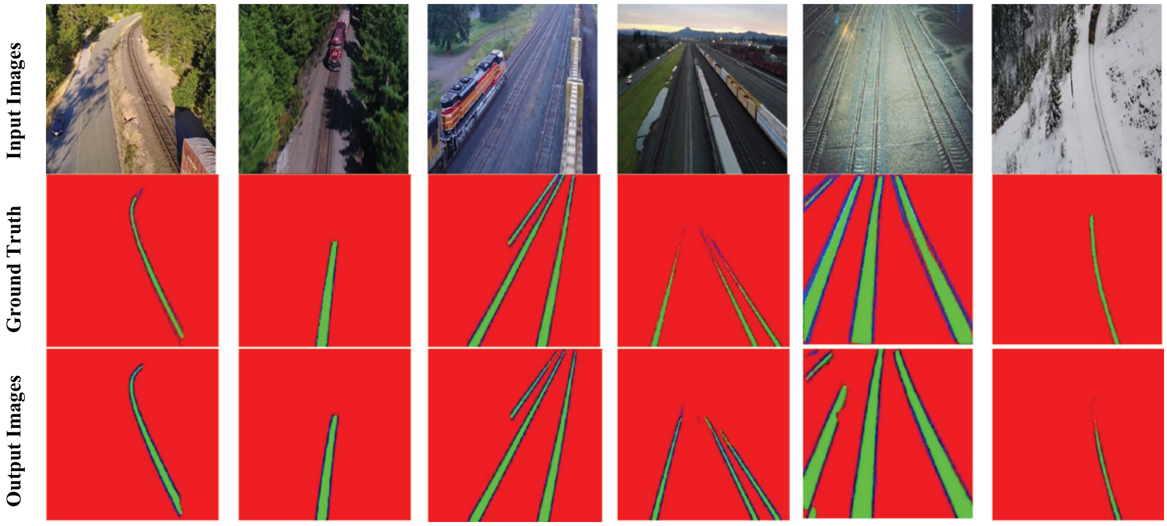


Fig. 10. Results of proposed model implemented on various light influence conditions of input images.

istic aerial railroad monitoring systems. The future direction will be focusing on integration of our proposed model with the UAV on-board processor for segmenting aerial railroad images during real time monitoring. This also helpful to find out the obstacles that are present nearby gauge or rail that causes calamities.

References

[1] B. Hosseiny, H. Rastiveis and S. Homayouni, An Automated Framework for Plant Detection Based on Deep Simulated Learning from Drone Imagery, *Remote Sensing, MDPI* **12** (2020), 1–21.

[2] A.J. Puppala, Surya Sarat Chandra Congress, A Holistic Approach for Visualization of Transportation Infrastructure Assets Using UAV-CRP Technology, *International Conference on Inforatmion technology in Geo-Engineering*, Springer Series in Geomechanics and Geoengineering, 2019, pp. 3–17.

[3] D. Liu, Z. Lu, T. Cao and T. Li, A real-time posture monitoring method for rail vehicle bodies based on machine vision, *International Journal of Vehicle Mechanics and Mobility, Vehicle System Dynamics*, Taylor and Francis, **55** (2017), 853–874.

[4] S. Sahebdivani, H. Arefi and M. Maboudi, Rail Track Detection and Projection-Based 3D Modeling from UAV Point Cloud, *Sensors* **20** (2020), 1–15.

[5] P.K. Sen, M. Bhiwapurkar and S.P. Harsha, Analysis of Causes of Rail Derailment in India and Corrective Measures, *Reliability and Risk Assessment in Engineering Proceedings of INCRS* 2018, Springer, 2018, pp 305–314.

[6] Y. Lyu, G. Vosselman, G.-S. Xia, A. Yilmaz and M.Y. Yang, UAVid: A semantic segmentation dataset for UAV imagery, *ISPRS Journal of Photogrammetry and Remote Sensing, Elsevier* **165** (2020), 108–119.

[7] Oliver Zendel Markus Murschitz Marcel Zeilinger Daniel Steininger Sara Abbasi, RailSem19: A Dataset for Semantic Rail Scene Understanding, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019, pp. 1221–1227.

[8] V. Badrinarayanan, A. Kendall and R. Cipolla, SegNet: A deep convolutional encoder-decoder architecture for image segmentation, *IEEE Trans. on Pattern Analysis and Machine Intelligence* **39** (2017), 2481–2495.

[9] J. Long, E. Shelhamer and T. Darrell, Fully convolutional networks for semantic segmentation, *Proc. 2015 IEEE Conf. on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[10] L.C. Chen, Y. Zhu, G. Papandreou, F. Schroff and H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, *Proc. of 15th European Conf. on Computer Vision*, 2018, pp. 801–818.

[11] Z. Qi, Y. Tian and Y. Shi, Efficient railway tracks detection and turnouts recognition method using HOG features, Neural Computer Applications, 2013, pp. 245–254.

[12] B.T. Nassu and M. Ukai, Rail extraction for driver support in railways, *P*roc. of IEEE Intell. Vehicles Symp. (IV), Baden-Baden, Germany, 2011, pp. 83–88.

[13] B. Le Saux, A. BeaupŁre, A. Boulch, J. Brossard, A. Manier and G. Villemin, Railway detection: From filtering to segmentation networks, Proc. IEEE Int. Geosci. Remote Sens. Symp., Valencia, Spain, 2018, pp. 4819–4822.

[14] X. Giben, V.M. Patel and R. Chellappa, Material classification and semantic segmentation of railway track images with deep convolutional neural networks, Proc. of IEEE Int. Conf. Image Process. (ICIP), 2015, pp. 621–625.

[15] M. Tan and Q.V. Le, EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks, *Proc. of the International Conference on Machine Learning*, 2019.

[16] S. Nathan and P. Kansal, Eyenet: Attention based Convolutional Encoder-Decoder Network for Eye Region Segmentation, *2019 IEEE/CVF International Conference on Computer Vision Workshop*, 2019.

[17] F. Kaleli and Y.S. Akgul, Vision-based railroad track extraction using dynamic programming, *Proc. of 12th Int. IEEE*

Conf. Intell. Transp. Syst., St. Louis, MO, USA, 2009, pp. 1–6.

[18] Z. Qi, Y. Tian and Y. Shi, Efficient railway tracks detection and turnouts recognition method using HOG features, *Neural Comput Appl* **23** (2013), 245–254.

[19] B.T. Nassu and M. Ukai, 'Rail extraction for driver support in railways, *Proc. of IEEE Intell. Vehicles Symp. (IV)*, Baden-Baden, Germany, Jun. 2011, pp. 83–88.

[20] A.I. Purica, B. Pesquet-Popescu and F. Dufaux, A railroad detection algorithm for infrastructure surveillance using enduring airborne systems, *Proc. of IEEE Int. Conf. Acoust.*, Speech Signal Process, 2017, pp. 2187–2191.

[21] Z. Teng, F. Liu and B. Zhang, Visual railway detection by superpixel based intracellular decisions, *Multimedia Tools Appl* **75** (2016), 2473–2486.

[22] M. Arastounia, Automated recognition of railroad infrastructure in rural areas from LIDAR data, *Remote Sens* **7** (2015), 14916–14938.

[23] B. Yang and L. Fang, Automated extraction of 3-D railway tracks from mobile laser scanning point clouds, *IEEE J Sel Topics Appl Earth Observ Remote Sens* **7** (2014), 4750–4761.

[24] B. Le Saux, A. BeaupŁre, A. Boulch, J. Brossard, A. Manier and G. Villemin, Railway detection: From filtering to segmentation networks, *Proc. of IEEE Int. Geosci. Remote Sens. Symp.*, Valencia, Spain, 2018, pp. 4819–4822.

[25] S. Mittal and D. Rao, Vision based railway track monitoring using deep learning, arXiv:1711.06423, 2017.

[26] D. Erhan, Y. Bengio, A. Courville and P. Vincent, Visualizing higher layer features of a deep network, University of Montreal, Canada, 2009.

[27] T. Lin, P. DollÆr, R. Girshick, K. He, B. Hariharan and S. Belongie, Feature pyramid networks for object detection, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 2117–2125.

[28] O. Ronneberger, P. Fischer and T. Brox, U-net: Convolutional networks for biomedical image segmentation, Proc. MICCAI, 2015, pp. 234–241.

[29] G. Ghiasi and C. C. Fowlkes, Laplacian pyramid reconstruction and refinement for semantic segmentation, Proc. of ECCV, 2016, pp. 519–534.

[30] D. Neven, B. De Brabandere, S. Georgoulis, M. Proesmans and L. Van Gool, Towards end-to-end lane detection: An instance segmentation approach, *P*roc. IEEE Intell. Vehicles Symp. (IV), 2018, pp. 286–291.

[31] F. Yu and V. Koltun, Multi-scale context aggregation by dilated convolutions, in Proc. ICLR, 2016, pp. 1–13.

[32] L.C. Chen, Y. Zhu, G. Papandreou, F. Schroff and H. Adam, Encoder decoder with atrous separable convolution for semantic image segmentation, Proc. of ECCV, 2018, pp. 801–818.

[33] H. Zhao, J. Shi, X. Qi, X. Wang and J. Jia, Pyramid scene parsing network, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 6230–6239.

[34] Y.-S. Xu, T.-J. Fu, H.-K. Yang and C.-Y. Lee, Dynamic video segmentation network, Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2018, pp. 6556–6565.

[35] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi and A. Agrawal, Context encoding for semantic segmentation, Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2018, pp. 7151–7160.

[36] P. Lyu, C. Yao, W. Wu, S. Yan and X. Bai, Multi-oriented scene text detection via corner localization and region segmentation, Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2018, pp. 7553–7563.

[37] K. He, G. Gkioxari, P. DollÆr and R. Girshick, Mask R-CNN, Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2017, pp. 2980–2988.

[38] H. Li, D. Zhao, Y. Chen and Q. Zhang, An efficient network for lane segmentation, Proceedings of the IEEE Conference on Cognitive Systems and Signal Processing, 2018, pp. 177–185.

[39] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou and G. Cottrell, Understanding convolution for semantic segmentation, Proceedings of the IEEE Winter Conference on Applications of Computer Vision, 2018, pp. 1451–1460.

[40] X. Pan, J. Shi, P. Luo, X. Wang and X. Tang, Spatial as deep: Spatial CNN for traffic scene understanding, Thirty-Second AAAI Conference on Artificial Intelligence, arXiv:1712.06080, 2018.

[41] V. Gaikwad and S. Lokhande, Lane Departure Identification for Advanced Driver Assistance, *IEEE Transactions on Intelligent Transportation Systems* **16** (2015), 910–918.

[42] L.A.F. Rodriguez, J.A. Uribe, J.F.V. Bonilla, Obstacle detection over rails using hough transform, *2012 XVII Symposium of Image, Signal Processing, and Artificial Vision (STSIVA)*, IEEE, 2012

[43] K. Simonyan and A. Zisserman, Very deep convolutional networks for large-scale image recognition, Proc. of ICLR, 2015, pp. 1–14.

[44] K. He, X. Zhang, S. Ren and J. Sun, Deep residual learning for image recognition, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 770–778.

[45] A. Shah, E. Kadam, H. Shah, S. Shinde and S. Shingade, Deep Residual Networks with Exponential Linear Unit, VisionNet'16: *Proceedings of the Third International Symposium on Computer Vision and the Internet, ACM Digital Library*, 2016, pp. 59–65.

[46] S. Woo, J. Park, J.-Y. Lee and S. Kweon, CBAM: Convolutional Block Attention Module, *Europea Conference on Computer Vision*, arXiv: 1807.06521v2, 2018.

[47] R.A. Naqvi, M. Arsalan, G. Batchuluun, H.S. Yoon and K.R. Park, Deep learning based gaze detection system for automobile drivers using a NIR camera sensor, *Sensors* **18** (2018), 1–34.

[48] Railfandan YouTube videos [Online]

[49] Google's best practices on splitting data. [Online]. https://tinyurl.com/y7yqfhxu

[50] S. Li, X. Zhao and G. Zhou, Automatic pixel-level multiple damage detection of concrete structure using fully convoltional network, *Computer-Aided Civil and Infrastructure Engineering*, 2019, pp. 616–634.

[51] E.K.V.I.I.A. Buslaev, A. Parinov and A.A. Kalinin, Albumentations: fast and flexible image augmentations, arXiv e-prints arXiv:1809.06839, 2018.

[52] M. Wang and J.C. Cheng, A unified convolutional neural network integrated with conditional random field for pipe defect segmentation, Computer-Aided Civil and Infrastructure Engineering, Wiley, 2020, pp. 162–177.

[53] J. Bertels, T. Eelbode, M. Berman, D. Vandermeulen, F. Maes, R. Bisschops and M.B. Blaschko, Optimizing the Dice Score and Jaccard Index for Medical Image Segmentation: Theory & Practice, 22nd International Conference on Medical Image Computing and Computer Assisted Intervention, 2019.

[54] Y. Wang, L. Wang, Y.H. Hu and J. Qiu, RailNet: A Segmentation Network for Railroad Detection, *IEEE Access* 2019, pp. 143772–143779.

[55] A. Kirillov, K. He, R. Girshick, C. Rother and P. Dollár, Panoptic Segmentation, arxiv:1801.00868, 2018.

[56] L.C. Chen, G. Papandreou, I. Kokkinos, K. Murphy and A.L. Yuille, Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, IEEE Trans. Pattern Anal. Mach. Intell, 2018, pp. 834–848,

[57] E. Romera, J.M. Álvarez, L.M. Bergasa and R. Arroyo, ERFNet: Efficient Residual Factorized ConvNet for Real-Time Semantic Segmentation, *IEEE Transactions on Intelligent Transportation Systems* **19** (2018), 263–272.

[58] Z. Zhong, Z.Q. Lin, R. Bidart, X. Hu, I.B. Daya, Z. Li, W.-S. Zheng, J. Li and A. Wong, Squeeze-and-Attention Networks for Semantic Segmentation, *Computer Vision Foundation*, 2020, pp. 13065–13074.

[59] J. Cai, C. Liu, H. Yan, X. Wu, W. Lu, X. Wang and C. Sang, Real-Time Semantic Segmentation of Remote Sensing Images Based on Bilateral Attention Refined Network, *IEEE Access* **9** (2021), 28349–28360.

[60] Z. Zou, X. Zhang, H. Liu, Z. Lia, A. Hussain and J. Li, A novel multimodal fusion network based on a joint coding model for lane line segmentation, arXiv:2103.11114v1, 2021.

[61] Y. Furitsu, D. Deguchi, Y. Kawanishi, I. Ide, H. Murase, H. Mukojima and N. Nagamine, *Asian Conference on Pattern Recognition*, 2020, pp. 639–652.

[62] S. Belyaev, I. Popov, V. Shubnikov, P. Popov, E. Boltenkova and D. Savchuk, Railroad semantic segmentation on high-resolution images, *IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, 2020.

[63] H. Li, Q. Zhang, D. Zhao and Y. Chen, RailNet: An Information Aggregation Network for Rail Track Segmentation, *IEEE International Joint Conference on Neural Networks (IJCNN)*, 2020.

[64] X. Li, L. Zhu, Z. Yu, B. Guo and Y. Wan, Vanishing Point Detection and Rail Segmentation Based on Deep Multi-Task Learning, *IEEE Access* **8** (2020), 163015–163025.

[65] Y. Wang, L. Wang, Y.H. Hu and J. Qiu, RailNet: A Segmentation Network for Railroad Detection, *IEEE Access* **7** (2019), 143772 – 143779.

[66] Y. Wang, L. Zhu, Z. Yu and B. Guo, An Adaptive Track Segmentation Algorithm for a Railway Intrusion Detection System, *Sensors, MDPI* **19** (2019), 1–21.

[67] L. Tong, L. Jia, Z. Wang, Y. Wu and Nin, Research on the Segmentation and Extraction of Scenes Along Railway Lines Based on Remote Sensing Images of UAVs, *International Conference on Electrical and Information Technologies for Rail Transportation*, Springer, **639** (2019), 481–492.

[68] Z. Zhang and Q. Liu, Road Extraction by Deep Residual U-Net, *IEEE Geoscience and Remote Sensing Letters* **15** (2018), 749 – 753.

[69] Z. He, P. Tang, W. Jin, C. Hu and W. Li, Deep Semantic Segmentation Neural Networks of Railway Scene, *37th Chinese Control Conference (CCC)*, IEEE, 2018, pp. 9096–9100.

[70] R. Yasrab, ECRU: An Encoder-Decoder Based Convolution Neural Network (CNN) for Road-Scene Understanding, *Journal of Imaging, MDPI* **4** (2018), 1–19.

[71] Y. Sun, Y. Weng, B. Luo, G. Li, B. Tao, D. Jiang and D. Chen, Gesture recognition algorithm based on multiscale feature fusion in RGB-D images, *IET Image Processing* **14** (2020), 3662–3668.

[72] Y. Lin, D. Xu, N. Wang, Z. Shi and Q. Chen, Road Extraction from Very-High-Resolution Remote Sensing Images via a Nested SE-Deeplab Model, *Remote sensing,* MDPI, 2020, pp. 1–20.