Vehicle-Related Scene Understanding Using Deep Learning

Xiaoxu Liu, Minh Neuyen, and Wei Qi Yan(✉)

Auckland University of Technology, Auckland 1010, New Zealand
weiqi.yan@aut.ac.nz

Abstract. Automated driving is an inevitable trend in future transportation, it is also one of the eminent achievements in the matter of artificial intelligence. Deep learning produces a significant contribution to the progression of automatic driving. In this paper, our goal is to primarily deal with the issue of vehicle-related scene understanding using deep learning. To the best of our knowledge, this is the first time that we utilize our traffic environment as an object for scene understanding based on deep learning. Moreover, automatic scene segmentation and object detection are joined for traffic scene understanding. The techniques based on deep learning dramatically decrease human manipulations. Furthermore, the datasets in this paper consist of a large amount of our collected traffic images. Meanwhile, the performance of our algorithms is verified by the experiential results.

Keywords: Traffic scene understanding · Deep learning · Automatic driving · Image segmentation · Object detection

1 Introduction

With the development of artificial intelligence, autonomous vehicles have already been associated with the field of computer vision. Due to complex traffic environment, the capability of traffic scene understanding has become a significant indicator of autonomous vehicles.

Scene understanding is a process of cognizing and inferring the environment based on spatial perception [1]. In vehicle-related scene understanding, a scene is the environment in which the vehicle is currently located including location, person, focus, event, and relationships between them. Scene understanding mainly includes object detection and recognition, semantic segmentation, toplogical relationship exploration and discovery between objects.

Scene information in a video is extremely dense which has great discrepancy and complexity. Recently, owing to the development of deep learning, the use of scene understanding can significantly ameliorate the performance of video analysis, which is a method of using machine perception [2].

In addition, deep learning has an active merit that a myriad of pretrained networks and public datasets have provided benefits for training numerous traffic scenes [19–21]. For vehicle-related scenes, in order to understand the objects, scenes, and events in a video, deep neural networks emulate high-level abstraction from the visual data and

M. Cree et al. (Eds.): ACPR 2019 Workshops, CCIS 1180, pp. 61–73, 2020.
https://doi.org/10.1007/978-981-15-3651-9_7

encode them using a robust representation [3]. Therefore, deep learning methods have unique advantages in the field of vehicle-related scene understanding.

For an example, the end-to-end nature of deep learning is one of such advantages, which achieves faster and universal information processing than traditional methods under the premise of a particularly accurate recognition of specific scenes. For autonomous vehicle that demands to understand the information in complex traffic scene, the method of deep learning can effectively satisfy the accuracy and real-time requirements [4].

Automated driving using deep learning is not mature yet to understand visual objects in complex traffic scenes due to the diversities of global traffic rules and transportation facilities [22]. Currently, it is difficult to apply all traffic scenes using only a single algorithm [5]. The two essential branches of scene understanding give us inspirations, object recognition identifies all objects of a predefined category on the image and positions through a bounding box. Semantic segmentation operates at a fine scale, its purpose is to segment images and associate each region with class labels [6]. Albeit these are two similar tasks, few studies currently merged the two categories of work together.

In this paper, automatic image segmentation and vehicle detection for vehicle-related scene understanding are developed using deep learning so as to reduce human workload. The datasets in this paper provide a great deal of traffic scenes. Simultaneously, adjustments and ameliorations have been implemented with the proposed neural network for scene understanding.

In this paper, literature review will be provided in Sect. 2, our methodology is shown in Sect. 3, the experimental results will be demonstrated in Sect. 4, our conclusion will be drawn in Sect. 5.

2 Literature Review

In this paper, we explain the reasons why the characteristics of deep learning play an essential role in scene understanding and why high-quality scene understanding models are often achieved through deep learning. First, the layer-by-layer processing of deep learning enables the model to better express the information in the current traffic scene. By simulating the structure of human brain and its gradual cognitive process, deep learning models obtain higher-level expressions through a linear or nonlinear combination. Therefore, deep learning enables the models to analyze complex traffic scenes, it has a hierarchical structure for information proceeding similar to our human brain, progressively extracts internal features and sufficient model complexity. These characteristics of deep learning enable the proposed model to understand the high-level semantics of traffic scenes (traffic event analysis, logical relationships of objects in traffic scenes). Currently, deep learning can accurately segment lanes to understand road conditions in the scene [7, 19–23]. It is also possible to predict overtaking, lane changing, and braking events by dynamically detecting the positional relationship between the two vehicles [8].

Secondly, the end-to-end characteristic of deep learning has also made an extraordinary contribution to the development of scene understanding. As the number of vehicles globally increases, the complexity of traffic scenes continues rising. Therefore, autonomous vehicles have higher demands in terms of real-time performance. The model

of traffic scene understanding based on deep learning utilizes the end-to-end process to optimize all tasks (vehicle detection, pedestrian detection, path planning, etc.). For example, an end-to-end vehicle controller can detect obstacles in the scene and navigate them accurately following the curved lanes [9].

Thirdly, deep learning algorithms have strong versatility [10]. Faced with the tasks involved in scene understanding, deep learning models do not require redesigning new algorithms for each task like traditional algorithms. Currently, each deep learning algorithm is suitable for a variety of scene understanding. For example, Faster R-CNN model achieves excellent results such as vehicle detection, pedestrian detection, and lane detection.

Finally, deep learning models have active mobility. A mature scheme of autonomous vehicle must contain a large number of functionalities related to scene understanding which utilizes human experience to train a scene so as to understand the scene from scratch. Deep learning models learn neural network parameters from one task and can transfer them very well to another. For example, the deep learning parameters and knowledge learned based on ImageNet dataset can achieve superior results in scene understanding by using other datasets [11].

In this paper, the understanding of vehicle-related scenes is implemented in conjunction with a deep learning-based vehicle detector and a semantic segmenter. Compared with other machine learning methods, deep neural networks can improve the accuracy by increasing the amount of training data and introducing sophisticated methods to betterment efficacy and accuracy [12]. Additionally, deep learning naturally is an end-to-end model, because the visual data is imported directly to the input layer, the well-trained network can export excellent outcomes. Finally, the underlying concepts and techniques using deep learning are universally transferable. Hence, deep learning can be much adaptive to various datasets for vehicle-related scene understanding.

3 Methodology

In order to deeply discover the merits of deep learning in vehicle-related scene understanding, we detail a computable method in this section based on deep learning for semantic segmentation and vehicle detection.

3.1 Vehicle-Related Scene Understanding

For the understanding of vehicle-related scene, a high-performance model is not only to detect and identify single isolated object, but also to understand advanced semantics in the vehicle-related scenes. Therefore, we make full use of deep learning to simulate the characteristics of our human brain. The high-level semantics in complex traffic scenes are employed through layer-by-layer processing and a stepwise abstraction of the feature map.

If we use a hierarchical system to emulate human cognition of the scenes, the entire human cognitive process of information needs to be carried out through several levels. At a lower level, our humans extract visual and auditory information as basic features, which is similar to the idea of extracting concepts from the scene and forming the basic

layer of the ontology; at a higher level, our human brain unifies these features and makes judgement from these features. At the highest level, our human can obtain the implicit semantics of the information through reasoning, which confirms to semantic expression and semantic understanding [13].

Therefore, in vehicle-related scene understanding, the positional relationship between objects in a scene is very useful for understanding the high-level semantics. It is also one of the necessary steps to use deep learning to analyze traffic scenes. The end-to-end nature of deep learning allows the model to handle multiple tasks. Deep learning models can complete multiple progressive tasks, the results of the previous task are used as an aid to the later. In order to explore the vehicle-related scene more deeply, we explore positional relationship of the objects in the scene for semantic segmentation.

Figure 1 utilizes topological relationships to correlate the classes in scene understanding. According to prior knowledge, visual objects such as trees and buildings should normally appear on both sides of the road. Bus lanes are usually drawn on the road. Most vehicles only travel on the road and do not appear on trees or in the sky. The sky is always above all objects, this is the fact that it never changes.

The relationship in the topological map plays a decisive role in scene understanding based on deep learning. According to the topological relationship between objects in the scene, the model can be used to clarify the object positions in the traffic scene. Deep neural network, as a typical ANN model, can achieve more human-like cognition by learning the logical relationship between objects. Moreover, the topological map constrains the range of the output, reduces the output of unrealistic scene, thereby improves the accuracy of scene understanding.

Fig. 1. The topological relationships in the scene

3.2 Methodology of Vehicle-Related Scene Understanding Model

Our model learns parameters by minimizing the value of the loss function. In order to optimize the extremely complex nonconvex function of convolutional neural networks, deep learning models typically exploit feedforward operations to abstract scene information, utilize stochastic gradient descent (SGD), error backpropagation, and chain rules to update the parameters of the model [14].

In the process of feedforward propagation, assume that input x, output y, and the cost $J(\theta)$ are given, the gradient during backward propagation is $\nabla_\theta J(\theta)$ [15]

$$J(\theta) = -E\left[\|y - f(x,\theta)\|^2\right] \tag{1}$$

According to the chain rules, if $h = g(k), z = f(g(k)) = f(h)$ then,

$$\frac{dz}{dk} = \frac{dz}{dh}\frac{dy}{dx} = \dot{z}\dot{h} \tag{2}$$

The loss function of SGD is $J(\theta) = L(f_\theta(k_i),\ h_i), (k_i, h_i)$ are samples $i = 1, \ldots, m$ with regard to θ in

$$\frac{\partial J(\theta)}{\partial \theta} = 0 \tag{3}$$

If α is a learn rate, we can construct the weight decay function as

$$\theta_{t+1} = \theta_t - \alpha \cdot \nabla_\theta J_i(\theta) \tag{4}$$

In summary, we can describe the SGD algorithm in Fig. 2.

Algorithm 1 Stochastic Gradient Descent Algorithm.

Input:
Learn rate, α;
Parameter, θ;
Training set (k_i, h_i) with m training samples, i=1,...,m;

Output:
while $i \neq$m **do**
Calculating gradient estimate: $\nabla_\theta \mathrm{J}(\theta) = +\frac{1}{m}\nabla_\theta \sum_i L(f_\theta(\mathrm{k}_i), h_i)$;
Weight update: $\theta \leftarrow \theta - \alpha \nabla_\theta J_i(\theta)$;
end while

Fig. 2. The algorithm of stochastic gradient descent (SGD)

3.3 Methodology of Semantic Segmentation

We apply a combination of VGG19 and SegNet as a semantic segmentation model and employ our datasets to train and test the model, our dataset consists of 91 images as shown in Fig. 3. This paper reflects each image at a 50% probability level through data

Fig. 3. The dataset of semantic segmentation

augmentation. The translation is performed in the unit of pixels from the horizontal and vertical directions, the translation is randomly selected from the continuous uniform distribution within the interval [−10, 10].

The structure of our neural network proposed in this paper utilizes a combination of encoders and decoders to produce feature maps of images. The encoder of SegNet includes convolutional layer, batch normalization, ReLU activation, and max pooling. Max pooling is used for reducing the size of feature maps. Even though the object boundary in the image may be blurred during the operation of max pooling, the pooling is indeed the best way to reduce the size of the feature maps. For reducing the feature map size while retaining the complete boundary information, SegNet extracts boundary information from the feature maps before performing the downsampling. During the decoding process, the upsampling operation of the decoder preserves the size of the original input. The max pooling memory index stored in each encoder map is used for upsampling the feature map. The last decoder is connected to the softmax classifier so as to assign the label of each class for the image [16].

Assume a is an array received from the upper layer, a_k is the k-th element in the array, and i is the total number of the array [17]

$$soft\max(k) = \frac{\exp(a_k)}{\sum_i \exp(a_i)} . \tag{5}$$

This paper created a SegNet network whose weight was initialized from the VGG-19 network. The additional layer required for semantic segmentation replaces the last pooling layer.

In summary, our semantic segmentation model takes advantage of the encoder-decoder structure, which combines visual information with high-level features through pooling to understand the surrounding scene in detail.

3.4 Methodology of Vehicle Detection

Our vehicle detector is a Faster RCNN-based model as shown in Fig. 4, which uses 337 images from our traffic scenes as datasets as shown in Fig. 5. The size of the input layer is 32 × 32 for better convolution operations. All convolutional layers of this model use a 3 × 3 convolution kernel, we set the step size as 1. Moreover, in the convolutional layer of the Faster R-CNN, all the convolutions are subject to one padding expansion, resulting in an increase of 2 in length and width. This setup does not change the size of the inputs and outputs.

Similarly, the convolution kernel size and step size of the pooling layer in the model are both set to 2. Thus, the size of each matrix passing through the pooling layer becomes one-half of the original. In other words, the convolutional layer and the ReLU layer maintain the size of the feature maps, the pooling layer reduces the size of the feature maps to 0.25.

Moreover, this model converts the collected proposals in the RoI Pooling layer into 7 × 7 proposal feature maps and sends them to the classification layer. Feature maps are classified using the cross-entropy as a loss function in the classification layer.

By adjusting a series of network parameters, vehicle detector can better learn semantics from shallow to deep and learn feature maps from abstract to specific.

Moreover, Faster R-CNN has two fully connected layers for classification and regression, respectively. Similarly, the most usage of two loss functions is fine-tuning. If i is assumed to be the index of an anchor in a mini-batch, p_i is the probability that the algorithm output, p_i^* is the ground truthing label of the anchor I, t_i is a four-element vector, which is bounded by the algorithm. The parameterized coordinates of the box t_i^* are the ground truthing box associated with a positive anchor [18],

$$L(\{p_i\},\{t_i\}) = \frac{1}{N_{cls}}\sum\nolimits_i L_{cls}\left(p_i, p_i^*\right) + \lambda\frac{1}{N_{reg}}\sum\nolimits_i p_i^* L_{reg}\left(t_i, t_i^*\right) \tag{6}$$

where L_{cls} (p_i, p_i*) is the logarithmic loss of the classification

$$L_{cls}\left(p_i, p_i^*\right) = -\log\left[p_i, p_i^* + \left(1 - p_i^*\right)\left(1 - p_i\right)\right]. \tag{7}$$

The classification calculates logarithm loss for each anchor, which is summed and divided by the total number of anchors N_{cls}.

In the regression loss,

$$L_{reg}\left(t, t_i^*\right) = R\left(t_i - t_i^*\right) \tag{8}$$

where R is defined as smooth L_1 and σ is 3, $x = t_i - t_i^*$,

$$R = smoothL_1(\mathrm{x}) = \begin{cases} 0.5\mathrm{x}^2 \times 1/\sigma^2 & \text{if } |\mathrm{x}| < 1/\sigma^2 \\ |\mathrm{x}| - 0.5 & \text{otherwise.} \end{cases} \tag{9}$$

For each anchor, we calculate $L_{reg}\left(t_i, t_i^*\right)$ and multiply it by p^*, then sum and multiply it by using a factor λ/N_{reg}. p^* has an object (+1) and no object (−1), which means that only the foreground is used to calculate the cost, the background does not calculate the loss. N_{reg} is the size of feature maps, λ controls the weight for classification and regression at a stable level.

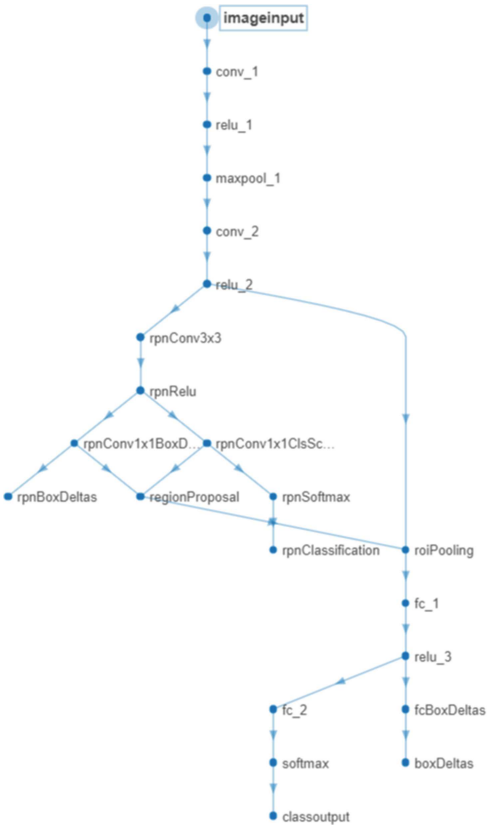

Fig. 4. The network structure of vehicle detector

4 Results

We use deep learning to achieve scene understanding. In order to verify the advantages of deep learning, we explore semantic segmentation and vehicle detection through experiments. As shown in Fig. 6, visually, the model in the semantic segmentation is satisfactory for the segmentation of buildings, sky, bus lane, and vehicles.

However, visual objects such as pedestrians, lanes, and trees are not accurate. This paper uses IoU to measure the amount of overlap for each class. Taken the first segmentation result in Fig. 6, the IoU of each category shows that the building IoU is up to 89% with the highest accuracy among the 11 classes. At the same time, the

proposed model scored higher accuracy on roads, sky, bus lane, and vehicles with 66%, 81%, 86%, and 66% respectively. However, the scores for segmentations of lane, tree, and turning sign are lower. The reason why we get these results is that the objects in the dataset lane and traffic sign are relatively smaller, the possibility of occurrence is rare. In future, we will focus on collecting the visual data including lane, tree, and turning sign images.

Fig. 5. The dataset of vehicle detection

For each class, the accuracy is the ratio of correctly categorized pixels to the total number of pixels in that class. The accuracy of sky detection is the highest among all classes at 91%. Secondly, bus lane, road, lane, and building classes have achieved satisfactory levels of accuracy 90%, 86%, 70%, and 81% respectively. Meanwhile, in this paper we exploited IoU to measure the overlapping between the ground truth and the detected regions, generated the results from the proposed model. The overlapping rates of buildings, sky, and bus lane are 71%, 74%, and 76%, respectively. In contrast, the accuracy and IoU of lane detection are worse than those of other classes, because the labelled area is small, its occurrence is lower than other classes.

From the experimental results of vehicle detection, the detector can successfully detect vehicles in multiple directions, sizes, and types under normal conditions, even detect vehicles with only a half of the vehicle appeared in the imaging range.

However, if there is an interference in the detection environment, our detection accuracy will be decreased. When a road tax sticker appears on the car windscreen, the detector incorrectly detects the road tax sticker as a vehicle. On the other hand, the detector can roughly detect the presence or absence of a vehicle in a particular area, but the position of the label boxes is offset.

This paper takes advantage of average precision to measure the accuracy of the detector. Both precision and recall are based on an understanding and measure of relevance. The best training result of average precision for this paper is 81%, which is based on a 22-layer Faster R-CNN network.

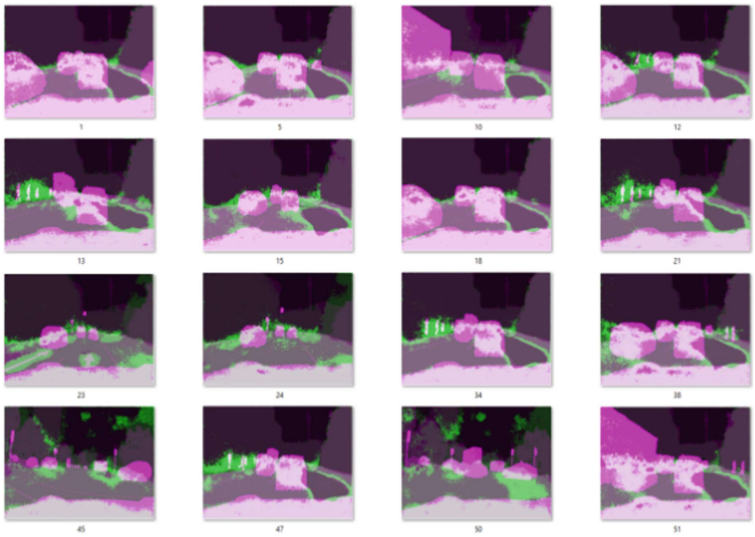

Fig. 6. The segmentation results

The log-average miss rate of the detection results is compared to the ground truth, which is adopted to measure the performance of the object detector in this paper, miss rate decreases as the false positive per image (FPPI) grows, and the log-average miss rate is 0.4.

In the vehicle identification, the test results were up to 81%, the average error rate was as low as 0.4. As shown in Fig. 7, compared with several models and models of different deep networks of the same class, this model is better in the control of the position and size of the bounding box.

Table 1. Comparison of VGG16-SegNet and VGG19-SegNet in IoU

Class	VGG16-SegNet IoU	VGG19-SegNet IoU
Sky	66%	74%
Building	70%	71%
Buslane	59%	76%
Road	59%	61%
Lane	50%	50%
Vehicle	67%	60%

Fig. 7. Comparison of four models with the same image

In the semantic segmentation, multiple evaluation indicators are taken as the metrics. We compare the network performance using VGG16 and VGG19 as the basic models shown in Table 1, respectively. The results of IoU show that segmentation results of VGG19 on Sky and Buslane are significantly higher than VGG16-SegNet. VGG16 is higher than VGG19 only in the segmentation result of the vehicle. After measuring all the evaluation indicators in general, we choose VGG19 as the basic model for our segmentation.

Through the exploration of vehicle detection and semantic segmentation, we find that deep learning has a great positive effect on scene understanding which relies on a layered processing mechanism and powerful transfer capabilities to improve the performance of the model in scene understanding.

5 Conclusion

The goal of this paper is to achieve traffic scene understanding using deep learning including semantic segmentation and vehicle detection. We fulfil each phase of the paper such as dataset preprocessing, neural network design and training, model evaluation, resultant comparisons. This paper provides a large number of original images and annotated images of our traffic environment for neural network computations, including annotations for image segmentation and vehicle identification. Furthermore,

SegNet shows high accuracy with small datasets. Faster R-CNN adopts two sets of convolution units to achieve high-precision segmentation and recognition. In future, we will use ensemble learning to integrate our experimental results together so as to get better results [24–27].

References

1. Li, Y., Dong, G., Yang, J., Zhang, L., Gao, S.: 3D point cloud scene data acquisition and its key technologies for scene understanding. Laser Optoelectron. Progr. **56**(4), 040002 (2019)
2. Chen, H., et al.: The rise of deep learning in drug discovery. Drug Discov. Today **23**(6), 1241–1250 (2019)
3. Husain, F., Dellen, B., Torras, C.: Scene understanding using deep learning, pp. 373–382. Academic Press (2017)
4. Yang, S., Wang, W., Liu, C., Deng, W.: Scene understanding in deep learning-based end-to-end controllers for autonomous vehicles. IEEE Trans. Syst. Man Cybern.: Syst. **49**(1), 53–63 (2019)
5. Jin, Y., Li, J., Ma, D., Guo, X., Yu, H.: A semi-automatic annotation technology for traffic scene image labelling based on deep learning pre-processing. In: IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC), pp. 315–320 (2017)
6. Nikita, D., Konstantin, S., Julien, M., Cordelia, S.: BlitzNet: A real-time deep network for scene understanding. In: IEEE International Conference on Computer Vision (ICCV), pp. 4154–4162 (2017)
7. Yao, W., Zeng, Q., Lin, Y., Guillemard, F., Geronimi, S., Aioun, F.: On-road vehicle trajectory collection and scene-based lane change analysis. IEEE Trans. Intell. Transp. Syst. **18**(1), 206–220 (2017)
8. Wei, Y., Tian, Q., Guo, J., Huang, W., Cao, J.: Multi-vehicle detection algorithm through combining Harr and HOG features. Math. Comput. Simul. **155**, 130–145 (2017)
9. Lecun, Y., Muller, U., Ben, J., Cosatto, E., Flepp, B.: Off-road obstacle avoidance through end-to-end learning. In: International Conference on Neural Information Processing Systems, pp. 739–746 (2005)
10. Ohsugi, H., Tabuchi, H., Enno, H., Ishitobi, N.: Accuracy of deep learning, a machine-learning technology using ultra-wide-field fundus ophthalmoscopy for detecting hematogenous retinal detachment. Sci. Rep. **7**(1), 9425 (2017)
11. Li, F., Deng, J., Li, K.: ImageNet: constructing a largescale image database. J. Vis. **9**(8), 1037–1038 (2009)
12. Samui, P., Roy, S.S., Balas, V.E.: Handbook of Neural Computation, pp. 12–34. Academic Press, Cambridge (2017)
13. Yu, Y., Cao, K.: A method for semantic representation of dynamic events in traffic scenes. Inf. Control **44**(1), 83–90 (2015)
14. Newton, A., Pasupathy, R., Yousefian, F.: Recent trends in stochastics gradient descent for machine learning and big data. In: Winter Simulation Conference (WSC), pp. 366–380 (2018)
15. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning, pp. 44–56. The MIT Press, Cambridge (2016)
16. Tran, S., Kwon, O., Kwon, K., Lee, S., Kang, K.: Blood cell images segmentation using deep learning semantic segmentation. In: IEEE International Conference on Electronics and Communication Engineering (ICECE), pp. 13–16 (2018)

17. Badrinarayanan, V., Handa, A., Cipolla, R.: SegNet: a deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling. IEEE Trans. Pattern Anal. Mach. Intell. **39**(12), 2481–2495 (2017)
18. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. Pattern Anal. Mach. Intell **39**(6), 1137–1149 (2017)
19. Ji, H., Liu, Z., Yan, W., Klette, R.: Early diagnosis of Alzheimer's disease based on selective kernel network with spatial attention. In: ACPR (2019)
20. Al-Sarayreh, M., Reis, M.M., Yan, W.Q., Klette, R.: A sequential CNN approach for foreign object detection in hyperspectral ımages. In: Vento, M., Percannella, G. (eds.) CAIP 2019. LNCS, vol. 11678, pp. 271–283. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-29888-3_22
21. Ji, H., Liu, Z., Yan, W., Klette, R.: Early diagnosis of Alzheimer's disease using deep learning. In: ICCCV 2019, pp. 87–91 (2019)
22. Shen, Y., Yan, W.: Blindspot monitoring using deep learning. In: IVCNZ (2018)
23. Pan, C., Li, X., Yan, W.: A learning-based positive feedback approach in salient object detection. In: IVCNZ (2018)
24. Wang, X., Yan, W.Q.: Cross-view gait recognition through ensemble learning. Neural Comput. Appl. (2019). https://doi.org/10.1007/s00521-019-04256-z
25. Liu, X.: Vehicle-related Scene Understanding. Masters thesis, Auckland University of Technology, New Zealand (2019)
26. Wang, X., Yan, W.: Human gait recognition based on frame-by-frame gait energy images and convolutional long short-term memory. Int. J. Neural Syst. **30**(1), 1950027:1–1950027:12 (2020)
27. Wang, X., Zhang, J., Yan, W.Q.: Gait recognition using multichannel convolution neural networks. Neural Comput. Appl. (2019). https://doi.org/10.1007/s00521-019-04524-y