Study on Optimized Lane Detection Algorithm based on U-Net

Yuanzhou Yao[1, a], Yihang Zhao[1], Ao Feng[1], Xinyue Su[3], Yuhang Yang[1], Jie Sun[4] and Haibo Pu[1,2, b, *]

[1] College of Information Engineering, Sichuan Agricultural University, Ya'an, Sichuan 625000, China

[2] Sichuan Key Laboratory of Agricultural Information Engineering, Ya'an, Sichuan 625000, China

[3]College of Humanities, Sichuan Agricultural University, Ya'an, Sichuan 625000, China

[4] College of Science, Sichuan Agricultural University, Ya'an, Sichuan 625000, China

[a]yaoyuanzhou@stu.sicau.edu.cn, [b, *] puhb@sicau.edu.cn,

Abstract. Lane detection has always been one of the important researches in semantic segmentation, but there are many problems in traditional lane detection algorithms, such as the much larger image pixels, the poor detection effect and so on. Based on the U-Net semantics segmentation network model, this paper redesigns two U-Net optimization network models based on RESNET residual module, and puts forward a series of image preprocessing methods aiming at the dataset's much larger pixels and some other problems. In the training process, the training data are adjusted Besides, date cleaning, data enhancement, data exposure and other operations are added. The final training model performs well on Apollos capes dataset.

Keywords: Lane detection; lane detection algorithms; U-Net semantics segmentation network model; RESNET; Apollos capes.

1. Introduction

With the rapid development of science and technology, our live is gradually becoming intelligent. Of course, there will be many breakthroughs in people's life, especially in the most important aspects like clothing, food, housing and transportation. With the explosive development in automatic driving technology, in order to improve the situation that every year, many people lose their lives in traffic accidents due to negligence, the intelligent driving assistance system, which takes lane detection system as its core, emerges as the times require. The system obtains road pictures including lanes by network camera, then obtains lane coordinates by lane detection algorithm, calculates the relative position of vehicle and lanes, and timely reminds drivers to correct their heading when deviates from its setting course, so as to avoid traffic accidents caused by lane deviation as far as possible.

In the lane detection mentioned in this paper, besides pre-processing the dataset to a certain extent, two u-net [1] networks based on RESNET residual module [2] are redesigned with previous experience based on the simplified deep lab v3p model provided by paddle. Finally, the average results of three networks are adopted to improve the training accuracy.

2. Introduction of Lane Detection Algorithms

Traditional Lane Detection Algorithms

It includes two network models, LaneNet [3]and H-Net [4]. LaneNet is a multi-task model that combines semantics segmentation and the vector representation of pixels, which is used to segment lanes in pictures. H-Net is a network model composed of convolution layer and full-connected layer, which is used to predict transformation matrix H, using transformation matrix H to regress the pixels of the same lane.

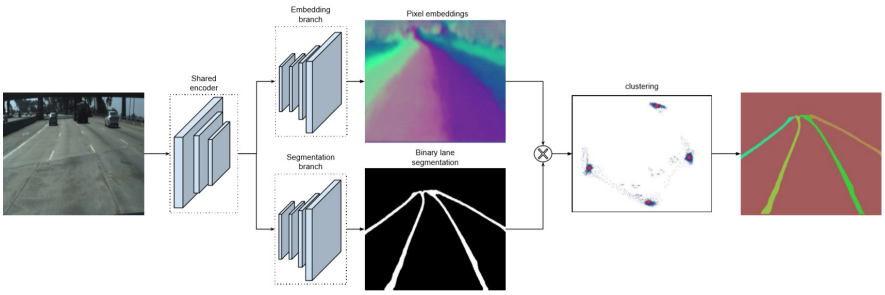

Figure 1. LaneNet architecture

As shown in Figure 1, in LaneNet, decoder is divided into two branches. Embedding branch conducts the embedded representation of the pixels. The trained embedding vectors are used for clustering. Segmentation branch is used to conduct semantics segmentation to input images (classify the pixels into two categories to determine whether the pixels belong to lane or background). Finally, the results of the two branches are combined to get the segmentation [5] results.

The output of aneNet is a set of pixels for each lane line, and a lane line needs to be regressed according to these pixels. Traditionally, the image is projected into an aerial view and then fitted by second or third order polynomials. In this method [6], the transformation matrix H is calculated only once, and all images use the same transformation matrix, which will lead to errors in the change of the ground surface (mountain, hill). The neural network H-Net, which can predict the transpose matrix H, emerges. The input of the network is a picture, and the output is the transpose matrix H:

$$H = \begin{bmatrix} \mathrm{a} & b & c \\ 0 & d & e \\ 0 & f & 1 \end{bmatrix}$$

The transpose matrix is constrained by zeroing, that is, the horizontal line remains horizontal under transformation (the transformation of coordinate y is not affected by coordinate X). As can be seen from Figure 2, the transpose matrix H has only six parameters, so the output of H-Net is a six-dimensional vector. H-Net consists of six layers of ordinary convolution network and one layer of fully connected network. Its network structure is shown in Table 1.

Table 1. H-Net network structure

Type	Filters	Size/Stride	Output
Conv+BN+ReLU	16	3×3	128×64
Conv+BN+ReLU	16	3×3	128×64
Maxpool		2×2/2	64×32
Conv+BN+ReLU	32	3×3	64×32
Conv+BN+ReLU	32	3×3	64×32
Maxpool		2×2/2	32×16
Conv+BN+ReLU	64	3×3	32×16
Conv+BN+ReLU	64	3×3	32×16
Maxpool		2×2/2	16×8
Linear+BN+ReLU		1×1	1024
Linear		1×1	6

3. What Improvement have been Made based on U-Net

3.1 U-Net Network based on Residual Block

The image segmentation problem can be simply summarized as using the corresponding algorithm to segment the image into an accurate object contour, so as to classify at pixel-level. FCN is the earliest network of deep learning in semantics segmentation. The solution of FCN network [7] is mainly to integrate pool 4, pool 3 and feature map, which can retain image information to a large extent [8]. The advantage of FCN is to segment from end-to-end. U-Net network is also an improvement based on FCNs.[9] U-Net network structure chart are as follows:

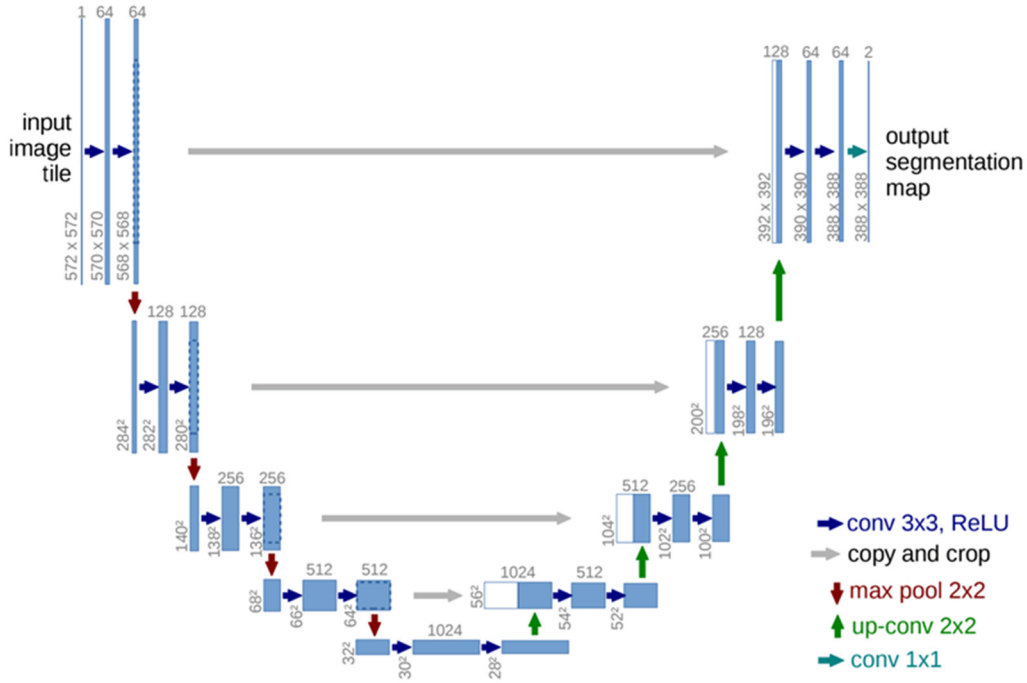

Figure 2. U-net architecture (example for 32x32 pixels in the lowest resolution). Each blue box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box. The x-y-size is provided at the lower left edge of the box. White boxes represent copied feature maps. The arrows denote the different operations

Because the network structure is similar to U-shaped, it is called U-Net. This network consists of two parts. The first part is feature extraction, which is similar to VGG. Some lightweight networks [10] can also be used for feature extraction. The second parties up-sampling. In feature extraction, each pooling layer is a scale, and there are total five scales, including that of the original map. After each up-sampling, the channel number of the corresponding part extracted from the feature is fused at the same scale. Before fusion, crop it [11]. Fusion can also be seen as mosaic. U-Net has the same back propagation process as traditional neural networks. Up-sampling and deconvolution can be used for the second part. Deconvolution, the transposition convolution, is a process from small size to large size [12].

Due to the clear outline of the lane data, U-Net can be used to better segment lane, so that the network can learn features from pixel level and improve the computing speed [13].

Residual block:

Based on theoretical reasoning, the deeper a network is, the more abstract the results of each layer after it will be. When the data is more and more abstract, the machine will be used to learn. But in reality, the deep network will be over-fitted, and RESNET [14] can better solve this problem, so that the deep network can be used to train in reality. Residual network is to jump input over one layer and directly enter the next layer. For example, the original input is Z [l], after the activation function is a [l+1] =g (z [l]). But if we add the residual block, a [l+1]=g (z [l]+a [l]), that is, the output of the upper layer jumps into the activation function of the next layer, and is activated after adding the output of

the layer, so that each feature of the upper layer can be transmitted directly to the next layer. There will be no over-fitting due to the fast dropping weight. The residual block diagram is as follows:

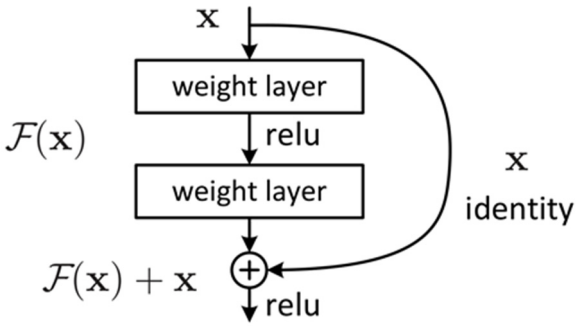

Figure 3. Residual block

Now the residual block is introduced into the feature extraction part of U-Net, so that the feature extraction part adopts the deep network to extract features as abstractly as possible, which is conducive to the subsequent segmentation task [15].

3.2 Introduction to the Paddlepaddle Framework

PaddlePaddle, which was developed by Baidu and belongs to a deep-learning open source platform under Baidu, has been developed and used only by engineers in Baidu before. In recent years, it has been opened to others. Globally, many technology companies, such as Google, have their own deep-learning open source platforms, which is highly connected with its own research field, having different technical characteristics. For Baidu, because of its various business and technical direction in search, image recognition, recognition and understanding of phonology and semantic, speech emotion analysis, machine translation, user portrait recommendation and some other fields. Therefore, PaddlePaddle is more versatile, more widely used, more comprehensive, covering the above content, which is a relatively full-featured framework for deep-learning. The advantage of PaddlePaddle lies in its lower requirements for data. It not only supports the parallel training of large scale deep-learning in dense and sparse parameter settings, but also supports the efficient parallel training of hundreds of billions of scale parameters and hundreds of points. It also provides a deep-learning framework for deep-learning parallel technology. So, PaddlePaddle has a strong deployment capability and can be carried out on multiple terminals. It also has great advantages in predicting performance because it supports high-speed inference of hardware devices with different structures such as the server and mobile end. At present, the stability and backward compatibility of API have been implemented in PaddlePaddle, and there are rather mature documents for bilingual use in Chinese and English.

3.3 Innovations

In the dee plab v3p model provided by paddle, the complexity and accuracy of the model need improving. This paper designs two U-Net networks based on RESNET residual module. As the depth of convolution network increases, the problem of gradient dispersion is particularly serious. But Residual network can effectively avoid it. As for the strategy of setting learning rate, this paper sets learning rate manually. In the first three epochs, default parameter training is adopted. In the next three epochs training, each epoch equally distributes six places to change learning rate. The overall change trend is still in accordance with one cycle strategy. The learning rate increases first and then decreases. By comparison, it is found that there is a certain visual difference between the test set and the training set. However, if the learning rate is reduced by several orders of magnitude according to the 1cycle strategy, it will result in over-fitting. So, the manual control of the learning rate will make the training model more robust. In the selection of loss function, it is determined in batches, and different loss function can be used in different training stages of the network to better determine the learning direction of the model. The difference of training sets makes the loss of the model fluctuate greatly. Selecting the classifying training of datasets properly will improve the predictive capability of the model, so the dataset road03 is abandoned. Using bilinear difference to zoom the image and

align the geometric center of the source image with that of the target image can also improve the accuracy. In traditional Lane detection, the offset in predicting image is widespread. The overall offset pixels are calculated to correct and improve the accuracy of the model.

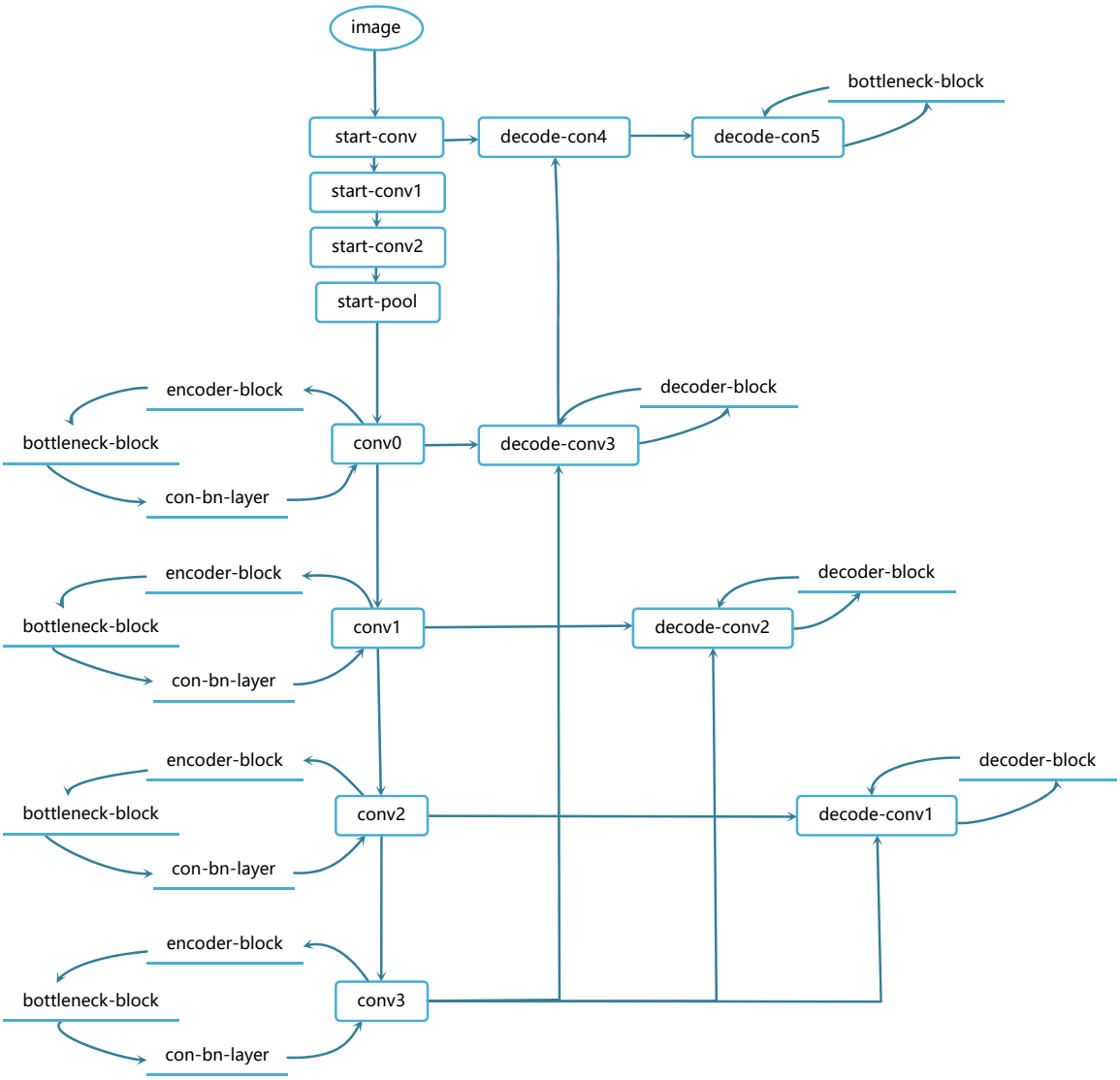

Figure 4. Improved network framework diagram

4. Algorithm Experiments

4.1 Experimental Platform

8GB memory, NVIDIA GeForce RTX 2080 GPU, inter (R) Core (TM) i7-9700K CPU are used as hardware platform; The operating system is Linux Ubuntu 16.04; Deep Neural Network library version is CUDA8.0. GPU-accelerated library is CUDNN V6.0. Paddlepaddle- gpu==1.3.0. post 97, which is based on Baidu Deep Learning Open Source Framework, mainly relies on open cv-python 3.4.3.18, imgaug 0.2.7 and other expansion kits. Now this platform trains and detects the improved U-Net lane detection algorithm.

4.2 Data Set Introduction

This paper selected the data set given by Baidu as the experimental dataset. The dataset file contains the surface line patterns in different time and space, in which includes the root folder of the apolloscapes dataset, the root directory. Type in the type/form of data, such as color images and "labels". The road ID specifies the road identifier, such as road02, the folder name of the record ID image subset. Defined by data acquisition system, camera ID images are grouped by the cameras that

capture them. In Apolloscape, there are two cameras: "Camera 5" and "Camera 6". The timestamp captures the time each image. Extend the file extensions: Jpg'for RGB images and. png'for basic facts. Take the advantage of all the picture information provided by Baidu, including training pictures and test pictures.

4.3 Image Preprocessing

Because the image resolution of the data set is 3384 * 1710 and the available display memory of the experimental equipment is limited, this paper mainly makes use of clipping to preprocess the reduced image. After analyzing the experimental data, because the upper part of the image is the sky and trees, there is no positive sample. So, the top part of the image, 3384 * 694, is cut. After clipping, in order to keep the image aspect as far as possible, three kinds of resolution (768 * 256, 1024 * 384, 1536 * 512) are used to train separately. The large resolution model is based on the pre-training model of the previous small resolution model. It is difficult to train them because of the receptive field problem in large resolution, so starting with small resolution model can not only quickly test the effectiveness of the model, but also provide better features and distributions. The experimental results show that the pre-training based on small resolution models can indeed help the convergence of the large resolution models.

4.4 Training Parameter Setting Instructions

In this paper, we take the modified Cycle LR strategy. The optimizer uses the adaptive gradient descent algorithm. The first three epochs take default parameters training, the learning rate is 0.001. In the following three epochs training, each epoch equally distributes six places to change the learning rate, the way of changing is: 0.001-0.0006-0.0003-0.0001-0.0004-0.0008-0.001. The last two epochs take the learning-rate training strategy which is between 0.0004 and 0.0001. Because of the large difference between test set and training set in image quality and visual perception, low learning rate can easily lead to over-fitting. So, the minimum learning rate is 0.0001. After 8 to 10 epochs, the training is basically over.

4.5 Results and Analysis

The detected Miou is 0.61234. Because the result of local CV does not correspond to that of Miou, the label is superimposed on the original map, but there are corresponding problems. Because the training resolution is 1536 x 512 and the original image is 3384 x 1020, the nearest interpolation is used to zoom label first. But the bilinear are used to get the result. The resolution itself is not divisible by the original resolution. This causes all predictions to shift 4-5 pixels to the right of the image. So, in results_correction.py, label is corrected by four pixels, and the current 0.63547 is obtained.

Table 2**.** Experimental data

Models	Loss Function	Base LR	Batch Size	Resolution	Miou
Unet-base	bce+dice	0.001	8	768×256	0.52231
Unet-base	bce+dice	0.001	4	1024×384	0.55136
Unet-base	bce+dice	0.001	2	1536×521	0.60577
Unet-Simple	bce+dice	0.001	2	1536×521	0.60223
Deeplabv3p	bce+dice	0.001	2	1536×521	0.59909
Ensemble	-	-	-	1536×521	0.61234
Correction	-	-	-	1536×521	0.63547

5. Conclusion

Based on the U-Net semantics segmentation network model, two U-Net optimization network models are redesigned based on Resnet residual module, and a series of image preprocessing methods are proposed aiming at the problem of the large dataset pixels. In the training process, the training data are adjusted, and data cleaning, data enhancement, data exposure and other operations are added. The final training model test Miou on Apolloscapes dataset and a good result, 0.6354, is achieved, proving that this optimization improves some performances on Apolloscapes dataset, but this model is still difficult to converge on large-scale pixels. This disadvantage is also one of the research directions in the future.

References

[1]. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. (2016).

[2]. Heinrich, M.P., Blendowski, M., Oktay, O.: TernaryNet: Faster deep model inference without GPUs for medical 3D segmentation using sparse and binary convolutions. arXiv preprint arXiv:1801.09449 (2018).

[3]. K. Zhao, M. Meuter, C. Nunn, D. Mller, S. Mller-Schneiders, and J. Pauli. A novel multi-lane detection and tracking system. In Intelligent Vehicles Symposium, pages 1084–1089, 2012.

[4]. Hani Altwaijry, Eduard Trulls, James Hays, Pascal Fua, and Serge Belongie.Learning to match aerial images with deep attentive architectures. In CVPR, pages 3539–3547, 2016.

[5]. Bertasius, G., J. Shi, L. Torresani. Semantic segmentation with boundary neural fields. In the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016.

[6]. W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network.IEEE Conference on Computer Vision and Pattern Recognition, pages 1874–1883, 2016.

[7]. B. Dai, Y. Zhang, and D. Lin. Detecting visual relationships with deep relational networks. In CVPR, 2017. 1, 2.

[8]. R. Hu, M. Rohrbach, J. Andreas, T. Darrell, and K. Saenko. Modeling relationships in referential expressions with compositional modular networks. In CVPR, 2017. 1, 2.

[9]. T. Darrell, and K. Saenko. Large scale visual recognition through adaptation using joint representation and multiple [13] J. Hoffman, D. Pathak, E. Tzeng, J. Long, S. Guadarrama.

[10]. F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In Proc. CVPR, 2015.

[11]. Esser, S. K., Merolla, P. A., Arthur, J. V., Cassidy, A. S., Appuswamy, R., Andreopoulos, A., et al. (2016). Convolutional networks for fast, energy-efficient neuromorphic computing. Proceedings of the National Academy of Sciences, 201604850.

[12]. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and vqa. arXiv preprint arXiv:1707.07998 (2017).

[13]. Nicolas Ballas, Li Yao, Chris Pal, and Aaron Courville. Delving deeper into convolutional networks for learning video representations. In Proc. ICLR, 2016.

[14]. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. Int. J. Comput. Vision, 115(3):211–252.

[15]. Heinrich, M.P., Blendowski, M., Oktay, O.: TernaryNet: Faster deep model inference without GPUs for medical 3D segmentation using sparse and binary convolutions. arXiv preprint arXiv:1801.09449 (2018).