



DFA-UNet: Efficient Railroad Image Segmentation

Yan Zhang ¹, Kefeng Li ¹, Guangyuan Zhang ^{1,*}, Zhenfang Zhu ¹  and Peng Wang ^{1,2} ¹ School of Information Science and Electrical Engineering, Shandong Jiaotong University, Jinan 250357, China² Institute of Automation, Shandong Academy of Sciences, Jinan 250013, China

* Correspondence: zhanggy@sdjtu.edu.cn; Tel.: +86-137-0893-2917

Abstract: In computer vision technology, image segmentation is a significant technological advancement for the current problems of high-speed railroad image scene changes, low segmentation accuracy, and serious information loss. We propose a segmentation algorithm, DFA-UNet, based on an improved U-Net network architecture. The model uses the same encoder–decoder structure as U-Net. To be able to extract image features efficiently and further integrate the weights of each channel feature, we propose to embed the DFA attention module in the encoder part of the model for the adaptive adjustment of feature map weights. We evaluated the performance of the model on the RailSem19 dataset. The results showed that our model showed improvements of 2.48%, 0.22%, 3.31%, 0.97%, and 2.2% in mIoU, F1-score, Accuracy, Precision, and Recall, respectively, compared with U-Net. The model can effectively achieve the segmentation of railroad images.

Keywords: deep learning; image segmentation; U-Net; depthwise separable convolution

1. Introduction

With the continuous development of intelligent technology, driverless technology has been widely used in the field of transportation. In the future, high-speed rails will be a very critical mode of transportation. Cameras on the front of the vehicle help visualize obstacles on the track for driverlessness. To achieve driverlessness, the first step is to process the relevant images. Image segmentation is a complex and critical step in the field of image processing and analysis, the purpose of which is to segment the parts of an image that have some special meaning and extract the relevant features. Splitting out image targets is a difficult task. Some common methods of traditional image segmentation are threshold-based segmentation, region-based segmentation, model-based deformation, etc. However, due to the complex background and changing scenes of railroad images, they often contain a large amount of noise, which is challenging to be accurately segmented by traditional image segmentation algorithms. Therefore, it is urgently needed to develop a fast and accurate segmentation method to improve the accuracy of segmentation in complex scenes and achieve real-time railroad segmentation.

CNN-based encoder-decoder architecture is a commonly used image processing method. CNN has excellent feature extraction capability that overcomes the limitations of manual feature extraction. However, many feature extractions can only extract high-level semantic information while ignoring the underlying semantic information, thus not having a good segmentation effect.

In this paper, we propose an efficient segmentation network, DFA-UNet. The module can focus on the image's main features at the encoder stage and further process them by downsampling. We present the GAMP module, which improves the network feature extraction capability by global average pooling as well as maximum global pooling for the channels.

The network model proposed in this paper consists of the following two main works:

1. A new convolution module, GAMP



Citation: Zhang, Y.; Li, K.; Zhang, G.; Zhu, Z.; Wang, P. DFA-UNet: Efficient Railroad Image Segmentation. *Appl. Sci.* **2023**, *13*, 662. <https://doi.org/10.3390/app13010662>

Academic Editor: Byung-Gyu Kim

Received: 11 December 2022

Revised: 21 December 2022

Accepted: 28 December 2022

Published: 3 January 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

In this paper, we propose a new convolutional module, GAMP, which is able to increase the perceptual field of the network without increasing the number of network parameters compared to other convolutional blocks. It enables the network to extract richer semantic information from the images effectively. Compared with other modules, the GAMP module can better capture the features and detailed areas of the input image in the results of image segmentation, and the segmentation is perfect.

2. Depthwise fusion-attention block

A fused attention mechanism is added to our model. By fusing the globally averaged pooling and the globally maximal pooling feature maps, it is possible to output feature maps that include both the spatial information of the input channels and capture the salient semantic information in the inputs.

2. Materials and Methods

2.1. Related Work

2.1.1. Traditional Railroad Segmentation

In recent years, there has been a proliferation of methods for railroad segmentation. Kaleli et al. [1] used the Sobel operator to calculate the gradient of the input image and applied it to the binary image by Hough transform processing to segment the railroad track. Qi et al. [2] used the feature extraction method of HOG to extract image features and used a local area growth-based method to segment the tracks. Teng et al. [3] proposed a visual rail detection using super pixels with dynamic planning to extract the left and right tracks without any calibration process. A feature extraction network with a pyramid structure that was proposed by Wang et al. [4] can increase the detection of railroads without creating a lot of regions, but the network's processing speed prevents it from being used for real-time railroad detection.

2.1.2. Deep Learning Image Segmentation

Image segmentation is a breakthrough in computer vision, and CNN-based image segmentation algorithms are increasingly being studied. A U-shaped network structure to obtain contextual information and location information for U-Net was proposed by Ronneberger et al. [5]. The encoder performs feature extraction through convolution and pooling operations, and the decoder recovers the features of the image through upsampling operations. To prevent information loss, the feature map of the decoder is spliced with the feature map of the encoder, which can effectively realize image segmentation. Zhou et al. [6] argued that direct concatenation in U-Net is too coarse and will cause the two connected convolutional layer inputs to have large semantic differences. To address these issues, they proposed UNet++ to reduce the semantic differences and the learning difficulty of the network. The Attention U-Net proposed by Oktay et al. [7] adds the mechanism of attention to the basic U-Net. By automatically learning parameters to adjust the activation value, it can suppress irrelevant regions in the image, highlight the features of the region of interest, and segment effectively. The deeper the network, the more likely it is to suffer from network capacity degradation. In order to solve this problem, He et al. [8] proposed Resnet for solving the gradient disappearance problem. Jha et al. [9] proposed ResUNet++ with a ResNet structure, capable of achieving good results from indistinguishable images. The transformer for sequence-to-sequence prediction has emerged as an alternative architecture with an innate global self-attentive mechanism by Vaswani et al. [10]. However, the lack of low-level detail may result in limited localization capability. Because convolution cannot learn global and long-range semantic information very well, Cao et al. [11] proposed the SwinUNet, with local and global semantic feature learning by feeding tokenized image blocks into an encoder module with transformer results. To make the model more lightweight, Li et al. [12] proposed Rail-Net. although their network has very good detection speed, their accuracy is relatively low

2.1.3. Attention Mechanism

Attentional mechanisms are now widely used in computer vision. Hu et al. [13] focused on the relationship between channels and proposed a new structural unit, SE-Net, which can model the interdependence between feature map channels. Oktat et al. [7] offered an Attention Gate in the Attention U-Net network, able to improve local regions of interest and suppress certain non-interest regions. Due to the popularity of Transformer, Chen et al. [14] proposed TransUNet based on the transformer architecture. To take advantage of Transformer and CNNs, the strategy of the TransUNet encoder is a mixture of CNN and Transformer to build the encoder. Transformer focuses more on global information but tends to ignore image details at low resolution. This hurts the decoder more to recover the pixel size, which will result in a coarse segmentation result. However, CNNs can make up for this shortcoming of Transformer, so mixed coding is of great benefit in the author's opinion. For the decoder, it is relatively simple, and it is the conventional transpose convolutional upsampling to recover the image pixels; simultaneous downsampling from the CNN of the encoder corresponds to the cascade over the same layer's resolution. These are all inherent operations of the original U-Net.

Self-attention is a unique attention mechanism that plays an increasingly important role in computer vision due to its long-range dependence and adaptability. However, self-attention was originally designed for NLP. When dealing with computer vision tasks, it treats images as one-dimensional sequences, thus ignoring the two-dimensional structure of images. For high-resolution images, only spatial adaptation is achieved, ignoring the adaptation of channel dimensions. Different channels often represent other objects for vision tasks, and channel adaptation has also been shown to be necessary. To resolve this problem, Guo et al. [15] proposed Large Kernel Attention (LKA) to exploit the advantages of self-attention and large kernel convolution. The decomposition of the LKA operation is proposed to capture long-term relationships. The LKA achieves adaptation not only in the spatial dimension but also in the channel dimension.

2.1.4. Depthwise Separable Convolution

Chollet et al. [16] proposed Depthwise Separable Convolution as an efficient convolutional neural network structure consisting of depthwise (DW) convolution and pointwise (PW) convolution. The calculation of DW convolution is straightforward: it uses one convolution kernel for each channel of the input feature map and then splices the output of all the convolution kernels to obtain its final output. To combine the information between the channels and the feature map, a layer of 1×1 PW convolution is added at the end of the feature map, which connects the maps from the previous step in the depth direction with weighting to generate a new feature map. Compared with a traditional CNN, DW separable convolution not only reduces the number of parameters of the model but also reduces the size of the model.

3. Methods

Most image segmentation methods are processed by simple convolution in the feature extraction stage. U-Net goes through a simple two-layer convolution process at the encoder stage to extract the low-level semantic information of the image. In the ResUNet++ downsampling stage, the perceptual image field is enhanced by adding a residual link to prevent the image gradient from disappearing. However, the stacking of a large number of 3×3 convolution blocks increases the number of parameters. In DCSAU-Net [10,17], the authors proposed the PFC block, by using a 7×7 DW convolution and a 1×1 PW convolution to perform downsampling. DW convolution has strong low-level feature extraction capabilities. Still, these cannot be extracted from deeper layers of the image. To take into account local contextual information, large receptive domains, and dynamic processes, The LKA module combines the advantages of convolution and self-attention to achieve the adaptation of the channel dimension as well as the spatial dimension. To increase the perceptual image field and reduce the number of parameters, we propose a

more efficient architecture of a depthwise fusion-attention block. We achieve access to low-level semantic information by adding the GAMP module. Our modules mainly consist of 7×7 DW convolution and 1×1 PW convolution, which can reduce the number of parameters and costs. In addition, 3×3 convolution blocks are added to the initial block for downsampling the input image. To avoid gradient disappearance, we also add residual connections in the GAMP module, which can improve the network performance without increasing the parameters. Finally, we demonstrate the reason for our choice of 7×7 DW convolution by ablation experiments. The different module pairs are shown in Figure 1.

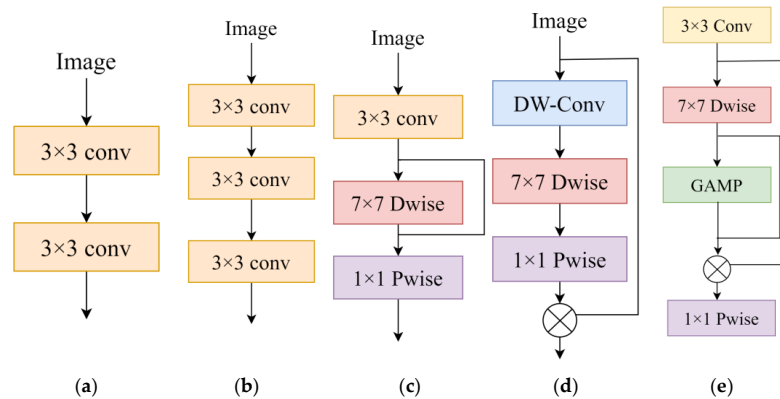


Figure 1. Comparing our DFA (e) with U-Net (a), Stem Block (b), LKA (c), and PFC (d) designs used to extract the low-level semantic information from the input images.

3.1. GAMP Module

SE-Net is an attention mechanism that focuses on the channel domain and has been widely used in many networks to reinforce the feature information of different channels in the feature map. The standard SE-Net first performs the global averaging pooling of the input features, which is usually used to aggregate spatial information in different channels of the feature map so that the segmentation result can retain more details of the original map. However, this tends to ignore the local information within each channel and thus dilute some significant or unique features of the input. In order to focus channel attention on the meaningful regions of the input image, we add a global maximum pooling operation to the original SE-Net to capture local robust feature information within each channel and combine it with the global average pooling results, thus compensating for the initial shortcomings of SE-Net. In addition, SE-Net uses full connectivity to find the correlation between different channels, and the number of parameters in the entire connectivity layer is relatively large, which also increases the overall computational effort of the network. Therefore, we use a 1×1 convolutional network instead of a fully connected layer. This will not only achieve the same effect but also reduce the number of parameters of the original SE-Net. The specific GAMP module is shown in Figure 2.

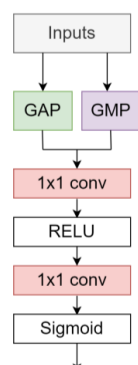


Figure 2. The framework of the GAMP block.

3.2. Deep Fusion Feature Module

The standard U-Net network uses a stacked 3×3 convolution to extract features. However, such a U-Net can only capture a limited number of receptive fields. The common practice of expanding the receptive field of a neural network is to use convolutional layers with large kernels or to stack more convolutional layers. However, using a larger convolutional kernel size can make the number of parameters in the network rise sharply, and blindly stacking convolutional layers can make the model suffer from gradient dispersion during deep gradient transfer. To enable the model to obtain richer semantic information from the input and to balance the complexity of the whole network, we propose a deep fusion feature module. In this module, we use a set of large-size depth-separable convolutions to extract new features from the input image. The input is first subjected to a 7×7 channel-by-channel convolution for feature map computation. In this process, we add the GMAP module and enable it to capture not only the spatial information in the feature map but also the significant semantic information in the channels, so that the whole network can focus on some important regional parts of the input image and increase the regression rate of the segmented image. The feature map extracted by this convolution will be input into a 1×1 point-by-point convolution to integrate the feature information on each channel and finally output the new features. Since the depth-separable convolution does not perform well with low-dimensional feature maps, we add a 7×7 convolution before this convolution to extract the initial features and boost the features' dimensionality. The specific DFA module is shown in Figure 1e. The GAP and GMP formulas are as follows:

$$GAP_c = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W u_c(i, j) \tag{1}$$

$$GMP_c = \text{Max}(u_c) \tag{2}$$

3.3. DFA-Unet Architecture

The complete DFA-UNet architecture is shown in the figure below, with the encoder in the left half of the figure and the decoder in the right half. The DFA attention module is used to extract complex features from the input image, and 2×2 maximum pooling in steps of 2 is performed after each block for downsampling. After completing downsampling 4 times, DFA-UNet will start decoding and each block is up-sampled with bilinear interpolation to gradually recover the size of the original image. A skip connection is used to stitch these feature maps with the feature maps from the corresponding encoders, which mixes low-level and high-level semantic information to generate accurate masks. Finally, a 1×1 convolution followed by a sigmoid function is used to output the image segmentation mask. The complete DFA-UNet architecture is shown in Figure 3.

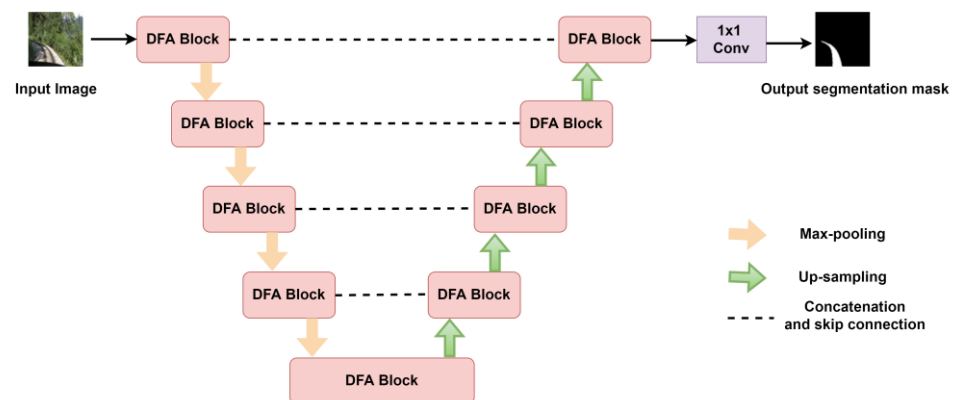


Figure 3. DFA-UNet network architecture designs.

4. Experimental Data and Results

4.1. Datasets

To test the validity of the model, we used the public Railsem19 railroad dataset. This dataset consists of 8500 images. In our experiments, we divided the dataset into 7650 training images as well as 850 test images.

4.2. Experimental Details

All experiments were implemented using the PyTorch 1.11.0 framework on a single RTX A5000 Tensor Core GPU, AMD EPYC 7543 32-Core Processor, and 24 GB RAM. We used a typical segmentation loss function, dice loss, and an SGD optimizer with a learning rate of 1×10^{-3} to train all models. The number of batch sizes and epochs was set to 16 and 150, respectively. During training, we set the image size of all input models to 256×256 , except for SwinUNet, which has an image size of 224×224 .

4.3. Model Evaluation

The evaluation metrics commonly used for image segmentation tasks are mIoU, Accuracy, Precision, Recall, and F1-score. One of the more dominant indicators is mIoU. To evaluate the performance of the models, we compared other more popular models and tested the metrics of each model for this dataset. Compared with other models, our model shows significant improvements in terms of mIoU, Precision, Recall, and F1-score. The specific experimental results are shown in Table 1. For the image segmentation task, the performance of the network on mIoU and F1-score metrics usually receives more attention. From Table 1, DFA-UNet achieves an F1-score of 0.8997 and a mIoU of 0.8662, which outperforms ResUNet++ by 1.84% in terms of F1-score and 2.12% in mIoU. Particularly, our proposed model provides a significant improvement over the two recent transformer-based architectures, where the mIoU of DFA-UNet is 11.28% and 4.27% higher than SwinUNet and LeViTUNet, and the F1-score of DFA-UNet is 7.83% and 3.13% higher than these two models, respectively.

We visualized the mIoU during the training of different models, as shown in Figure 4. We use a violin drawing, which shows the overall distribution of the data in addition to the above statistics.

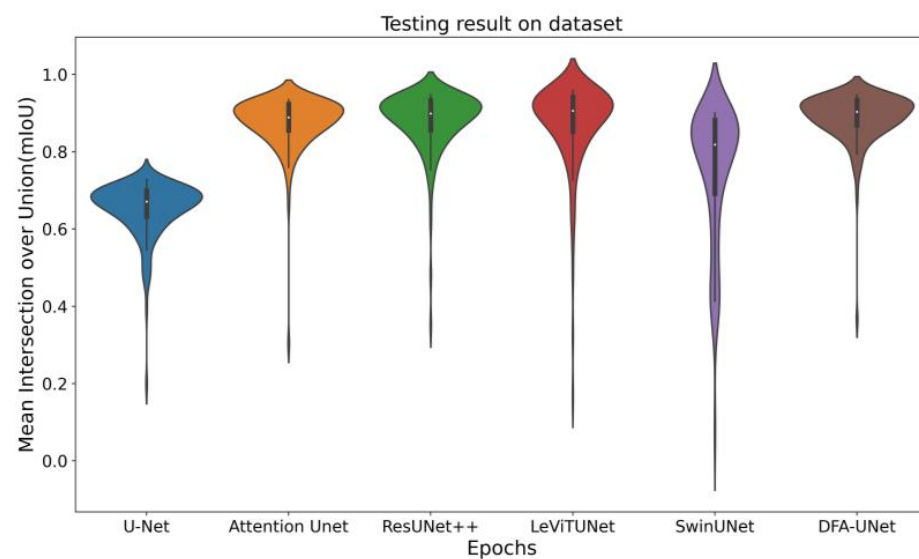


Figure 4. MIoU performance on different models.

Table 1. The results of the evaluation on the dataset.

Model	mIoU	Accuracy	Precision	Recall	F1-Score
U-Net	0.8414	0.9902	0.8706	0.8960	0.8777
Attention UNet	0.8635	0.9924	0.8919	0.9020	0.8928
LeViTUNet	0.8235	0.9905	0.8812	0.8711	0.8684
ResUNet++	0.8450	0.9921	0.8940	0.8814	0.8813
SwinUNet	0.7534	0.98338	0.8480	0.8196	0.8214
UNet+SEnet	0.8466	0.9912	0.8878	0.88677	0.8795
DFA-UNet	0.8662	0.9924	0.9037	0.9057	0.8997

The following Figure 5 shows the loss plots of different models during training.

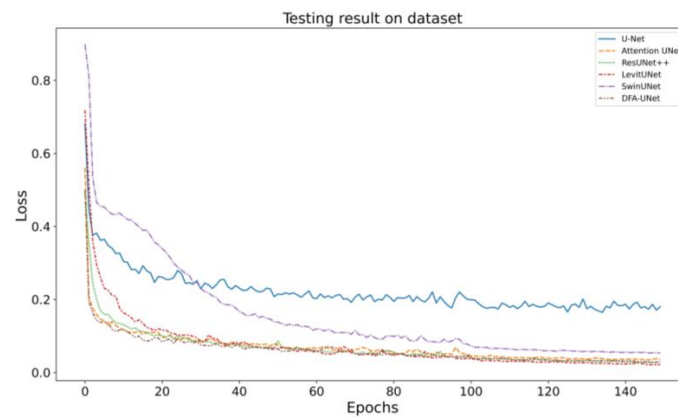


Figure 5. Training loss on different models.

4.4. Model Results

The following Figure 6 shows the results of our model compared with other models. As you can see from Figure 6, our model is also able to identify well the railroad environment for dark conditions. For the second row of data in Figure 6, the original image has two tracks, but the label only labels one track, and our model is able to predict all tracks clearly.

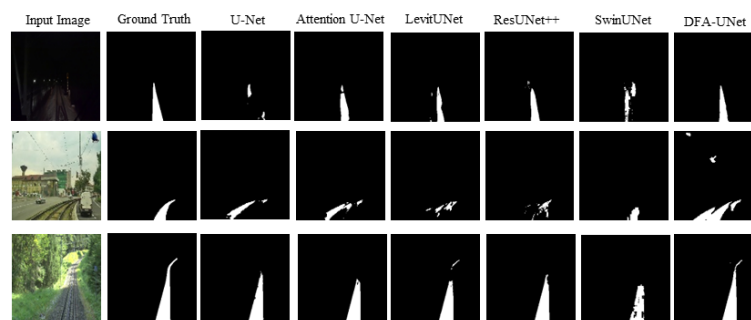


Figure 6. Qualitative comparison results between DFA-Net and other models on rail segmentation datasets.

4.5. Ablation Study

In this section, we performed the DFA-Unet ablation experiments. We analyzed the number of parameters per model (Params), the number of floating point operations per second (Flops), and the number of frames per second (FPS) transmitted by the model. Table 2 shows that our model has fewer parameters. To justify our choice of 7×7 convolutional kernels, we compared different convolutional kernel sizes, and the final result shows the effectiveness of our choice of 7×7 -sized convolutional kernels. The results are shown in Table 3.

Table 2. Comparison of other parameters of different models.

Model	FLOPs	Params	FPS
U-Net	31.1191G	13.3953M	64.67
Attention UNet	66.6318G	34.8785M	34.47
LevitUNet	33.2142G	52.1438M	43.33
Resunet++	70.9938G	14.4825M	36.61
SwinUNet	10.7228G	27.1458M	51.14
UNet+SE	31.1229G	13.4711M	48.69
DFA-UNet	24.9564G	8.7517M	49.61

Table 3. Study of different kernel sizes in DFA modules of the DFA-UNet architecture on railway datasets.

Kernel Size	mIoU	Accuracy	Precision	Recall	F1-Score
3 × 3	0.8463	0.9909	0.8766	0.8952	0.8805
5 × 5	0.8562	0.9912	0.8872	0.9089	0.8935
7 × 7	0.8662	0.9924	0.9037	0.9057	0.8997
9 × 9	0.8455	0.9907	0.8780	0.8950	0.8813

5. Discussion

Semantic segmentation has been widely used in the field of image analysis, and most segmentation models are composed in the form of encoder-decoder results and fuse low-level to high-level semantic information by jumping connections. This simple encoding processing may lose a lot of valid information. We can retain this valid feature information by introducing the DFA attention module. Our proposed GAMP module, which introduces global average pooling, can greatly reduce the number of parameters and optimization efforts of the model. It can suppress noise, reduce information redundancy, improve the network's ability to judge categories, and find all the target distinguishable regions for prediction. GMP is introduced to maximize the global and ignore other regions with low scores. After that, we combine the extracted different features by the attention channel fusion strategy for deeper feature extraction. At the downsampling stage, the feature maps at different scales can be better extracted by our module. In the upsampling stage, the same module can recover features at different scales. In the middle of downsampling and upsampling, we use a simple skip connection to concatenate the features of different layers to complete the image segmentation. There are also many complex forms of jump connections, which we will explore further in the future. Table 2 shows that our proposed model has fewer Flops and Params and can be used as a more lightweight model result. Our FPS is not very high with the same device, but it can identify some devices in real time. Future research will focus further on the model for quick real-time detection. As shown in Figure 4, it can be demonstrated that our proposed model shows higher performance in most scenarios and is more robust than other SOTA methods.

6. Conclusions

In this paper, we propose a new encoder-decoder architecture for image segmentation: DFA-UNet. The proposed model consists of the GAMP module and the DFA module. The GAMP module can effectively extract the low-level features of the image and reduce the number of parameters of the model, and the DFA module can realize the fusion of different scale feature maps. We evaluated our model on the Remrail19 dataset, and the results showed that our model has higher mIoU evaluation metrics compared to other models. Our model can perform better for complex image segmentation tasks. In the future, we will focus on optimizing its performance so that it can segment images more accurately in real time.

Author Contributions: Methodology, Y.Z.; software, K.L.; validation, G.Z., Z.Z. and P.W.; formal analysis, Y.Z. and G.Z.; investigation, Y.Z.; resources, Y.Z. and G.Z.; data curation, Y.Z.; writing—original draft preparation, Y.Z.; writing—review and editing, Y.Z., K.L. and G.Z.; visualization, K.L. and G.Z.; supervision, Z.Z. and P.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Science Foundation of Shandong Province (ZR2021MF064) and the China Postdoctoral Science Foundation (2021M702030).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Dataset source: <https://wilddash.cc/railsem19>.

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

References

1. Kaleli, F.; Akgul, Y.S. Vision-based railroad track extraction using dynamic programming. In Proceedings of the 2009 12th International IEEE Conference on Intelligent Transportation Systems, St. Louis, MO, USA, 4–7 October 2009; pp. 1–6.
2. Qi, Z.; Tian, Y.; Shi, Y. Applications: Efficient railway tracks detection and turnouts recognition method using HOG features. *Neural Comput. Appl.* **2013**, *23*, 245–254. [[CrossRef](#)]
3. Teng, Z.; Liu, F.; Zhang, B. Applications: Visual railway detection by superpixel based intracellular decisions. *Multimed. Tools Appl.* **2016**, *75*, 2473–2486. [[CrossRef](#)]
4. Wang, Y.; Wang, L.; Hu, Y.H.; Qiu, J. RailNet: A segmentation network for railroad detection. *IEEE Access* **2019**, *7*, 143772–143779. [[CrossRef](#)]
5. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
6. Zhou, Z.; Rahman Siddiquee, M.M.; Tajbakhsh, N.; Liang, J. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 3–11.
7. Oktay, O.; Schlemper, J.; Folgoc, L.L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N.Y.; Kainz, B. Attention u-net: Learning where to look for the pancreas. *arXiv* **2018**, arXiv:1804.03999.
8. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
9. Jha, D.; Smedsrud, P.H.; Riegler, M.A.; Johansen, D.; De Lange, T.; Halvorsen, P.; Johansen, H.D. Resunet++: An advanced architecture for medical image segmentation. In Proceedings of the 2019 IEEE International Symposium on Multimedia (ISM), San Diego, CA, USA, 9–11 December 2019; pp. 225–2255.
10. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; Volume 30.
11. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Japa: Swin-unet: Unet-like pure transformer for medical image segmentation. *arXiv* **2021**, arXiv:2105.05537.
12. Li, X.; Peng, X. Rail Detection: An Efficient Row-based Network and a New Benchmark. In Proceedings of the 30th ACM International Conference on Multimedia, Lisbon, Portugal, 10–14 October 2022; pp. 6455–6463.
13. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
14. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. Japa: Transunet: Transformers make strong encoders for medical image segmentation. *arXiv* **2021**, arXiv:2102.04306.
15. Guo, M.-H.; Lu, C.-Z.; Liu, Z.-N.; Cheng, M.-M.; Hu, S.-M. Japa: Visual attention network. *arXiv* **2022**, arXiv:2202.09741.
16. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.
17. Xu, Q.; Duan, W.; He, N. Japa: DCSAU-Net: A Deeper and More Compact Split-Attention U-Net for Medical Image Segmentation. *arXiv* **2022**, arXiv:2202.00972.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.