Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

RESEARCH ARTICLE

Semantic segmentation algorithm based on transformer In Mobile Edge Computing

XiBei Jia

[1]School of Computer Science and Engineer, Nanjing University of Science and Technology, Nanjing, China

Correspondence
XiBei Jia, School of Computer Science and Engineer, Nanjing University of Science and Technology, Nanjing, China.
Email: subee24@qq.com

Present Address
Nanjing University of Science and Technology, Nanjing 210094, China.

Abstract

The semantic segmentation task is a basic task in the field of Mobile Edge Computing, which requires the classification of each pixel in the image, which has higher requirements for classification accuracy than the image classification task. Fine-grained classification tasks requires more detailed information, in addition to classifying according to the semantic information and spatial information of each pixel unit and the surrounding pixels, it is also necessary to distinguish from adjacent pixels, which is one of the main difficulties of the current segmentation task. However, high-resolution input images can bring more detailed information, but they are often accompanied by expensive computing costs, so smaller resolution images will be put in practical applications to ensure computing speed. As another task of computer vision, super-resolution recovery focuses on extracting information from low-resolution pictures and reasoning into higher-resolution feature maps. Its recovered detail features contribute to the high-precision classification of semantic segmentation tasks. Considering the complementarity of the two tasks, considering the use of transformer as a feature extractor, the design algorithm realizes semantic segmentation and super-resolution recovery tasks at the same time, multi-task learning can ensure that the backbone network obtains more common high-dimensional information, and then we use the results of super-resolution recovery branches to guide the semantic segmentation task to provide more detailed information and finally obtain an effective improvement on the original baseline.

KEYWORDS:
Deep Learning, TransformerSemantic SegmentationContractive LearningSuper Resolution

1 | INTRODUCTION

Semantic segmentation is an intensive image classification task that plays a important role in multi-level scene understanding. Its pixel-level accuracy often requires more high-resolution detail information. In order to obtain more information, most complex models require large memory consumption and computational costs. How to ensure real-time segmentation without increasing the amount of computation is a hot research direction.

At present, most of the semantic segmentation algorithms based on deep learning use an end-to-end structure, input the picture into a feature extractor for encoding, gradually obtain a high-dimensional feature map, and then decode and enlarge it, and finally output the original image size. Most methods extract feature representations for density prediction by designing complex feature

extraction networks or deepening the network. The high-resolution network proposed by HRNet[1] maintains the resolution of the image throughout the process of feature extraction and prediction. While extracting low-resolution feature maps, high-resolution feature maps are also retained, and detailed information is effectively guaranteed not to be lost by combining paths and horizontal links with different resolutions. However, its network framework is relatively large. To improve the segmentation accuracy, many algorithms design a variety of more complex auxiliary modules on the decoding branch of downstream tasks, such as hole convolution, attention mechanism, context encoding, etc. However, these methods always involve more expensive computing costs, limiting their application on resource-constrained devices. At the same time, the time cost of testing is getting longer and longer. Especially for video segmentation tasks that require continuous processing, the long time cost cannot meet the needs of real-time monitoring. In some scenarios, images with a lower resolution of the model are selected for real-time semantic segmentation. Therefore, how to better use the existing image information for detailed feature analysis under the premise that the model inference speed is not affected, remains to be studied.

In the semantic segmentation task, the dimension of the feature map is greatly reduced after passing through the encoder. A low resolution of the input image causes more challenging for the back-end network. Most networks use parameter-free operations to upsample the image size when restoring it, such as bilinear interpolation. This operation makes it hard to recover the details for the final segmentation result. The SR single-image super-resolution recovery task is also a pixel-level processing task, which requires understanding and optimizing the detailed features of the original image. This task is similar to the semantic segmentation task.Therefore, we design an additional SR single-map super-resolution recovery branch, which can learn and recover the detailed characteristics of the resolved map, and the resulting high-resolution feature map can provide more detailed information for the semantic segmentation task. This branch can be deleted at the time of prediction, which can ensure that no additional time is added.

In this work, we designed a semantic segmentation based on transformer. This method restores the detailed information of the feature map by adding super-resolution auxiliary branches, which helps the semantic segmentation to obtain better results. The main flowchart is shown below, In summary, our main contributions include.

1. We design a bottom-up multi-layer feature fusion module, which can decode and fuse the multi-layer feature map extracted by the transformer encoder, and is suitable for both semantic segmentation and super-resolution recovery.

2. Consider how to combine with the semantic segmentation task to design a reasonable and effective super-resolution recovery task to effectively improve the accuracy of semantic segmentation.

3. In the Cityscapes urban landscape dataset, our algorithm is fully experimented, combined with the qualitative and quantitative results of comparative experiments, and finally it is proved that adding super-resolution recovery branch for assistance can effectively improve the accuracy of the semantic segmentation algorithm.

2 | RELATED WORK

2.1 | Semantic Segmentation

With the development of the CNN convolutional neural network, its powerful local feature extraction ability is soon applied to semantic segmentation task. Full convolutional network[2] (FCN) proposes to eliminate the full connection layer from a variety of traditional networks, and the network only extracts deep semantics due to multiple convolutional layers and activation layers, thus realizing the semantic segmentation network. Most mainstream semantic segmentation models are symmetrical U-shaped codec structures, represented by U-Net[3], SegNet[4], RefineNet[5] and other segmentation models. The mainstream backbone network has the following shortcomings for semantic segmentation tasks, such as the loss of location label information within the feature layer, the inability to process global context knowledge, and the lack of multi-scale processing, etc. Most subsequent studies have proposed various architectural technologies to solve these problems. PSPNet[6] pyramidal pool network makes use of the features of multi-layer networks. DeepLab[7] proposed by Chen et al in 2017 proposed the void convolution structure, and DeepLabv2 proposed the ASPP void convolution pyramid structure, which uses the void convolution to extract and integrate multi-scale features. For the feature fusion stage of different layers of pyramid features, ICNet[8] added an additional convolution structure to realize the alignment of features at different scales, and also used the cascade label method to guide model training. CBAM[9] proposed the channel attention mechanism and the spatial attention mechanism to emphasize the concerned information and suppress the invalid information in the feature map to obtain the appropriate feature prediction map.

2.2 | Transformer network

In computer vision, Dosovitskiy[10] first proposed a fully transformer-based image classification network, the Visual Transformer (ViT), and broke the ImageNet best performance. ViT chooses to cut the input image into a series of non-overlapping patches. ViT performs a linear projection of the block into a sequence vector. After that, the vector is put into the Transformer encoder for the calculation of self-attention, and the vector output of the module is obtained and classified. This paper realizes the first application of pure Transformer structure model in computer vision, and provides a new research direction for many fields of deep learning. SETR[11] first introduced pure Transformer encoders into the field of semantic segmentation, and made several attempts at the design of the decoder, and the final results show that progressive upsampling works best. PVT[?] introduced a pyramid structure into ViT, which makes transformers better applied to dense prediction tasks. Different from the mainstream semantic segmentation methods that perform mask prediction at the pixel level, Cheng et al. treat the semantic segmentation task as a mask prediction problem and propose maskformer[12].

2.3 | Super-Resolution

Image resolution is used to evaluate how much detailed information a picture can contain, which reflects the ability of the picture to reflect detailed information. High-resolution images often contain richer texture information and larger pixel density. Super-resolution (SR) restoration task refers to the process of using digital image processing and other related knowledge to reconstruct the details in the image and improve the Low-Resolution image to a High-Resolution image. In 2016, Dong et al. first applied convolutional neural network to super-resolution recovery task and proposed SRCNN[13]. Firstly, multiple convolution layers are used to extract the features of the image, and the extracted feature maps are converted into high-dimensional feature vectors through nonlinear mapping. The high-dimensional feature vectors are convolved again to reconstruct the high-resolution feature images as the output. The VDSR[14] model first proposed that the residual network was used in the super-resolution reconstruction network, and the input image was directly connected to the high-frequency features of the reconstruction through the residual structure. While reducing the amount of calculation, the supplement of low-frequency information in the reconstructed image was completed, and the efficiency of the model was significantly improved. In 2016, Shi et al. proposed ESPCN[15], which proposed a sub-pixel convolutional layer. Firstly, the feature channels of the image were changed by convolution, and then the feature layers of different channels were fused by pixel rearrangement to realize the enlargement of the feature map. Compared with the common upsampling methods, the proposed method has more context information, which is conducive to the recovery of image detail features.

2.4 | Multi-task learning

Multi-tasking learning means that the model learns multiple tasks with related representations at the same time to improve learning efficiency, prediction accuracy, and generalization performance. Multitasking learning is widely used in machine learning, natural language processing, computer vision and other fields[4,5,6,7]. MultiNet[4] has designed a kind of network structure, that can simultaneously image segmentation, target detection, such as semantic segmentation visual tasks. Cross-stitch Networks[5] studied the sharing method of neurons in multi-tasking networks and proposed that sharing layer can be automatically determined by end-to-end learning. Multitasking learning neural networks focus on parallel learning of computer vision tasks, requiring trade-offs between multiple task outcomes. In order to achieve optimal results for a single task, AAA puts forward the concept of task affinity. By calculating the affinity between different tasks, it sets groups for multiple tasks to optimize the final training results of tasks. The multi-task learning proposed in this paper needs to distinguish the main task from the auxiliary task in multi-task learning, and only focus on the training effect of the main task. We use the domain-specific information in the training signal of the auxiliary task to improve the generalization effect of the main task and optimize the final result.

3 | PROPOSED APPROACH

3.1 | Approach Overview

This paper designs a semantic segmentation algorithm based on transformer, which adds a super-resolution auxiliary branch to restore and supplement the detail information of the feature map, and guides the main branch to segment to obtain better results. We choose to use mix-transformer as the feature extractor to perform semantic segmentation and super-resolution recovery

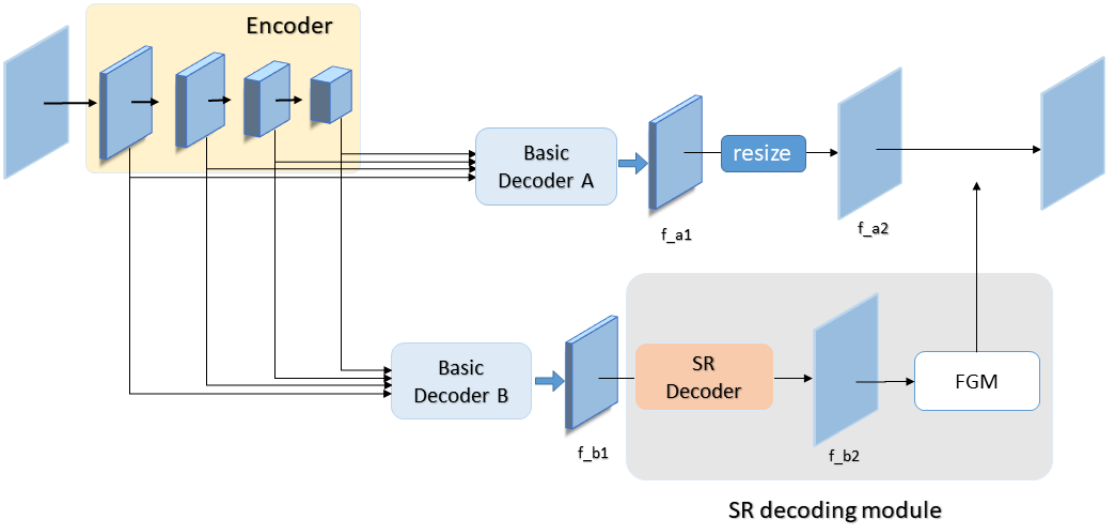

Figure 1 Network framework

tasks simultaneously, assuming that the input image size is H*W, The encoder outputs multi-scale feature maps with the size of (1/4H*1/4W,1/8H*1/8W,1/16H*1/16W,1/32H*1/32W) for the two tasks. In order to ensure that the results of the two tasks can learn from each other, we design the basic feature decoder that is suitable for both tasks. In the decoding process, feature maps of different scales are used, and a bottom-up feature alignment module is used to perform intensive reading promotion on the two tasks at the same time, and the final output feature map size is (1/4H*1/4W). In the main branch, we use bilinear interpolation to restore the feature map to its original size (H*W) as the result r1 of semantic segmentation. In addition to the basic decoder, SRB (super-resolution-branch) also includes a super resolution decoder, which performs super-resolution restoration on the low-resolution feature output to obtain high-resolution feature map (H*W) r2. The process is all supervised by the original image. Finally, this high-resolution feature map is used to supplement the semantic segmentation result f2 with fine-grained structural features, and the optimized semantic segmentation result r3 is obtained.

3.2 | Design of auxiliary branches

Semantic segmentation requires sufficient detailed features, and super-resolution recovery tasks can effectively deduce detailed features, so we consider using it as an aid. The input low-resolution feature map is recovered by SR super-resolution to improve its resolution and supplement more detailed information, which helps the semantic segmentation to obtain more accurate results in the subsequent process. We try to assist the original task with series and parallel.The experimental results show that concatenation of the two tasks after the same decoder can ensure that the features of the two tasks are in the same feature space. However, because the feature information required by the two tasks is not completely consistent, it is difficult to use one encoder to achieve the prediction of the two tasks. Our main task is semantic segmentation, and too much attention to super-resolution recovery task will lead to a decline in the final semantic segmentation results. Therefore, we use parallel to perform two tasks at the same time, which reduces the weight of the auxiliary task and ensures the original effect of segmentation.By comparing the final feature map similarity matrix of the two branch tasks, the detail part of the main task is supplemented to optimize our main task.

3.3 | Basic decoder design

The range of receptive field provided by CNN is limited. Therefore, in order to obtain more comprehensive context information, context modules such as ASPP are often needed. These modules expand the receptive field, but make the decoding

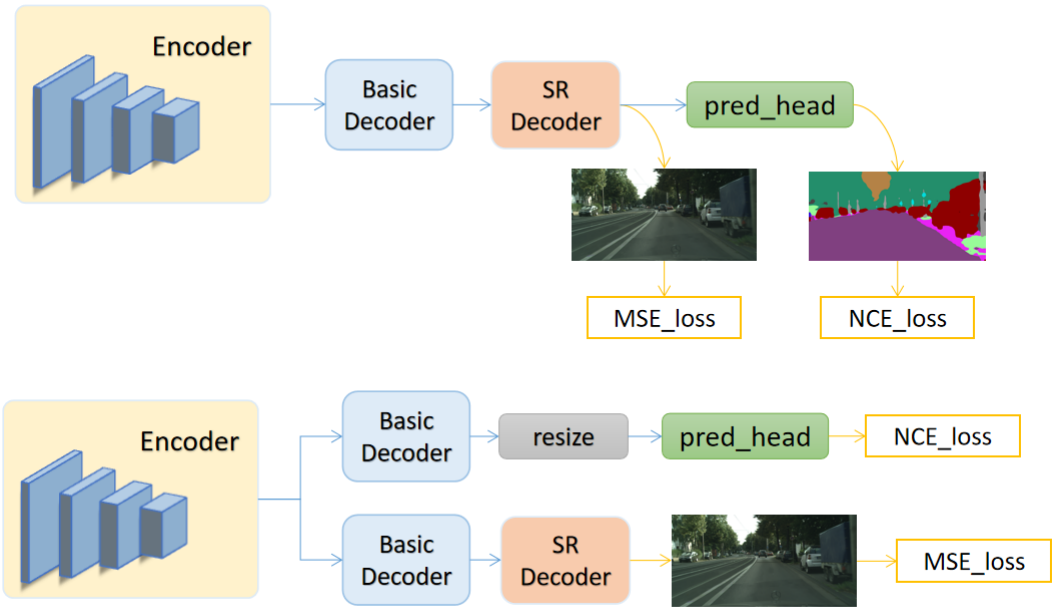

Figure 2 Two design ideas

terminal redundant. Benefit from the self-attention mechanism in Transformer to generate both high local and non-local attention. Therefore, we use MLP multilayer perceptron for the main module of feature decoding. Firstly, considering that both semantic segmentation and high-resolution tasks are fine-grained tasks, which require multiple layers of features to obtain certain context information at the same time, in order to ensure that the final segmentation graph can align the original image and segmentation graph at the same time, we choose to design a bottom-up alignment mode. A As shown in Figure 3, the original

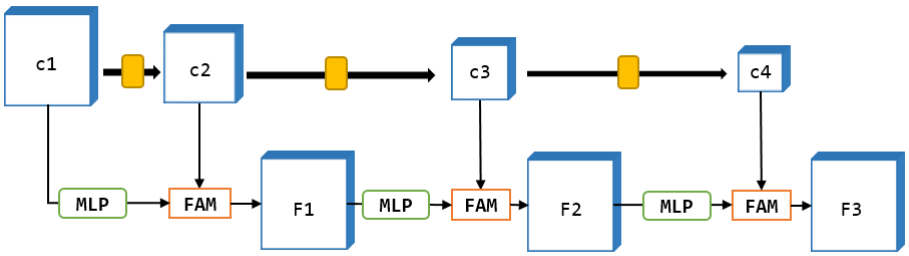

Figure 3 Basic Decoder design

image forms two branches after feature extraction, which are c1,c2,c3,c4 and f1,f2,f3 extracted by multiple Transformer blocks. The top to bottom path of the first layer in the figure is the feature layers with different resolutions extracted through the self-attention mechanism, which is the same as the Transformer backbone network part. Considering that the lower level feature map has richer spatial structure information than the higher level feature map, we choose the bottom-up method to recover the feature map in the second level path. The low-dimensional features with rich spatial structure information are used to generate bias fields to guide the alignment of adjacent layer features. As shown in the figure, there is no upsampling operation in the recovery path of the lower layer, and the feature layer always keeps the initial size, so as to ensure that the initial spatial information is not affected. With the fusion of the upper feature layers, the semantic information in the lower graph path is gradually enriched, forming the final feature map. The proposed decoding method can be effectively used in semantic segmentation and

super-resolution recovery tasks. Therefore, the decoding part of the two branches can use this decoding method to obtain the final feature fusion map for further downstream task expansion.

3.4 | SR decoder

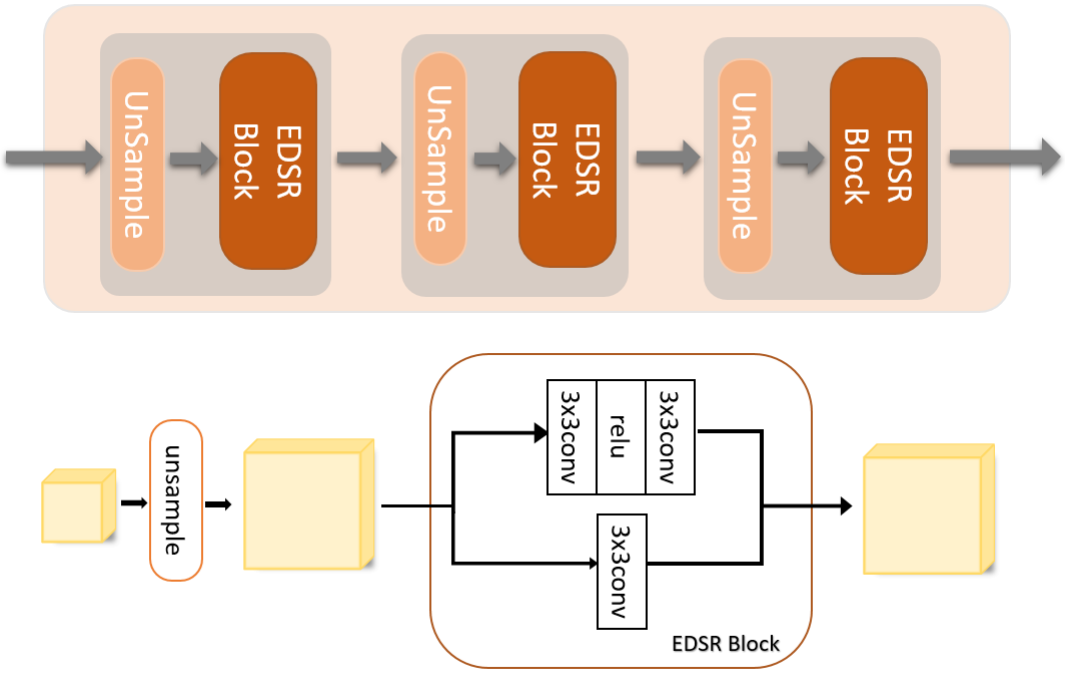

Figure 4 Design of SR Decoder

As shown above, only relying on the feature map obtained by the Decoder A decoding module is not enough to obtain the high-resolution semantic features of the original image as input, because the second half of the decoder uses bilinear upsampling or simple subnetworks to recover the features, which does not bring additional information. Here, we design the Decoder B2 part to recover the low-resolution feature output results. This branch aims to construct high-resolution information with low-resolution feature layers, which is supervised by the original image, so that it can not only capture part of the semantic information, but also effectively reconstruct the fine-grained structural information of the image. Although the focus of the task is super-resolution recovery, the part of the contour information obtained may not contain enough categories, but it can still be effectively divided into groups by pixel and pixel, and the relationship between the region and the region, which can effectively recover part of the segmentation boundary problem, so as to refine the semantic segmentation results. Therefore, we use the results recovered by the final super-resolution to guide the semantic segmentation results.

Specifically, as shown in Figure 4, it is composed of three super-resolution recovery modules. Each module includes an up-sampling layer for resolution improvement and an EDSR Block, which is mainly used to extract the mapping relationship from low-frequency information to high-frequency information. Because most of the low-frequency information of the high-resolution and low-resolution images is similar, EDSR does not repeat the learning of this part of the content, but uses the difference part to reconstruct the high-frequency content in the network path. On the premise of saving the learning cost, the required segmentation mapping relationship is effectively extracted. Compared with the residual network, the BN normalization layer is also removed. This is because when BN is used in traditional image classification tasks, the absolute differences between image pixel features will be ignored and the relative differences will be considered more, while in the super-resolution recovery task, the absolute detail information is more important.

3.5 | Feature guidance module(FGM)

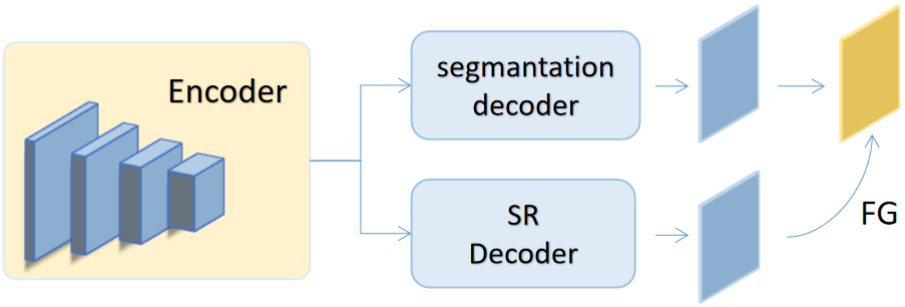

Figure 5 Design of FGM

Since the super-resolution auxiliary branch has more complete structural information than the branch obtained by the semantic branch, we add the FA feature guidance module, and use the result graph obtained by the auxiliary branch to guide the main branch to learn the high-resolution representation of the details in the image. Feature similarity learning is mainly used to reduce the similarity of the two feature layers. We first calculate the feature similarity matrix of the two branch feature maps. The similarity matrix describes the pairwise relationship between pixel values on the same feature layer, as shown in the following equation(1), where S_{ij} represents the feature similarity between pixels i and j. Then, the distance between the two-point similarity matrices is calculated, as shown in Equation (2).

$$L_{fa} = \frac{1}{W^2H^2} \sum_{i=1}^{W \cdot H} \sum_{j=1}^{W \cdot H} \parallel S_{ij}^{seg} - S_{ij}^{sr} \parallel_q \tag{1}$$

$$S_{ij} = (\frac{F_i}{\parallel F_i \parallel_p})^T \cdot (\frac{F_i}{\parallel F_i \parallel_p}) \tag{2}$$

$S_{ij}^{s}eg$ and $s_ij^{s}r$represent the similarity matrix of features obtained by semantic segmentation branch and super resolution auxiliary branch, and p and q represent norm normalization of features. $W \cdot H$ represents the size of the feature graph. The P and q representations are used to normalize the features.

4 | EXPERIMENTS

4.1 | valuation Dataset

We choose the public datasets Cityscapes and ADE20K as the main validation datasets. Cityscapes is a dataset of urban scenes with high resolution, which is mainly used in the field of autonomous driving. The dataset provides a variety of semantic segmentation annotations and instance segmentation annotations with 30 categories, including lane, car, pedestrian, building, sky bike, etc. ADE20K [16] contains annotations for a variety of tasks and was published by a team at MIT. The ADE20K dataset is collected from the network, and has large differences in size, resolution, and scene. The dataset provides pixel-level annotations of more than 3,000 categories. In addition to city streets, it also includes outdoor scenes, indoor home scenes, lakes, tables, doors, Windows, books, etc. The dataset not only annotates the semantic level of some meaningful object categories, but also subdivides the parts of the same category, such as the wheel of a bicycle, the frame of a bicycle, etc. Common evaluation metrics for semantic segmentation include mIoU, mAcc, allAcc. In this experiment, the Mean Intersection over Union (mIoU) is mainly used as the main evaluation index. The union of the real value set and the predicted value set is used as the numerator, and the union is used as the denominator. The ratio of the intersection and union obtained is called IoU value. mIoU calculates the IoU of each class separately and then averages it.

4.2 | Implementation Details

In order to ensure fairness, the same training strategy was adopted throughout the experiment. Two data sets of ADE20K and Cityscapes were used as the main data set, and all the main stems were initialized using pre-training weights. In the data enhancement part, we randomly adjusted the size of the image at a ratio of $0.5-2.0$, and expanded the data by random horizontal flip and random cropping to a specific size. AdamW optimizer was used for experiments. The initial learning rate was set to $base_lr = 1e^(-5)$, and the learning rate was dynamically adjusted according to Poly of $base_lr \bullet ((1 - iter/(total_iter))^power$. The $weight_decay$ was set to 0.01. *momentum* was set to 0.9. All the ablation experiments in this chapter were carried out on two A10 cards, using a distributed framework and a multi-threaded approach to accelerate the training. Due to hardware limitations, batchsize was set to 8 for cityscapes and 16 for ade20K. In the ablation experiment, in order to maintain the balance between accuracy and experiment time, we generally set the number of iterations as 40K. The final experimental effect will be demonstrated with 160K. We do not use any additional data sets during validation.

4.3 | Ablation Study on Network Structure

4.3.1 | Ablation experiment of key modules

In order to prove the effectiveness of our main improvement modules, we conducted ablation experiments on the overall model, as shown in Table 1. Our main improvement modules included adding super resolution auxiliary recovery branch and feature guide module. Therefore, we took the model with two key parts removed as the base, where SR represents super resolution auxiliary branch. FA indicates the feature boot module. It can be found through the experiment that adding auxiliary branches can effectively improve the accuracy, but the high-resolution feature map is more important for the guidance of the original image. After adding the guidance module, the mIoU of 1.2 can be improved.

Table 1 Ablation experiment of key modules

Model	Backbone	Iter	mIoU(%)	$mIoU_noback$(%)	allAcc (%)
Base	Transformer	16K	80.78	87.63	93.45
Base+SR	Transformer	16K	81.02	88.01	94.57
Base+SR+FA	Transformer	16K	81.98	90.45	95.45

4.3.2 | Super Resolution restoration branch ablation experiment

First of all, we designed the auxiliary branch of recovery superresolution. At first, we tried to conduct the superresolution recovery module directly through the feature map of backbone network, without the same DecoderB1 as semantic segmentation. The EDSR module is directly transplanted and the upsampling is used to restore the superresolution alternately. Meanwhile, the weight of the superresolution restoration branch loss is reduced, but it will affect the effect of the main branch. As shown in the table above, baseline represents the performance of the baseline model. We directly input the features extracted from the backbone network into the EDSR module and conduct upsampling for super resolution recovery. At this time, the loss of the auxiliary branch has been set to 0.1, which still affects the performance of the main branch of semantic segmentation, resulting in a 3.15 reduction in accuracy. It is considered that both the predicted result loss and the predicted branch format differ greatly between the two tasks. In order to ensure the stability of the model, we decided to add Decoder B1, which is exactly the same as the main branch, before the super resolution recovery module. The results are shown in the table, which can effectively maintain the stability of the model and improve the mIOU by about 0.1 compared with the original model

4.3.3 | Ablation experiment of feature-guided module

After determining the stability of the model, we used the superresolution results of the auxiliary branch to guide the detailed features of the main branch. As shown above, FA represents the feature guidance module. We added the feature fusion module

Table 2 Super Resolution restoration branch ablation experiment

Model	Iter	mIoU	mIoU_noback(%)	allAcc(%)
baseline	4K	72.59	81.87	95.27
Base+EDSR*3	4K	69.44	80.56	96.45
Base+decoderb1+EDSR*3	4K	72.68	80.45	95.12

on the basis of adding the auxiliary branch, and the results showed that the accuracy of the feature fusion module could be improved by about 1.0 more effectively.

Table 3 Ablation experiment of feature-guided module

Model	Iter	mIoU	mIoU_noback(%)	allAcc(%)
base	4K	72.59	81.87	95.27
Base+Decoder2+EDSR	4K	72.68	81.65	95.45
base+Decoder2 +EDSR+FA	4K	74.31	83.06	95.53

4.3.4 | Decoder ablation experiment

Considering that the direct use of semantic? segmentation branches may not be fully applicable to the superresolution recovery task, we also make further attempts for the decoder of this branch. In Segformer, the original text uses four MLPS to decode the four layers of features obtained, and then splice them, and then use one layer of convolution as the final feature fusion result. In the previous experiment, we directly used this decoding structure. Here, base_sr represents the model that uses MLP decoding terminal and adds super resolution assisted recovery and feature guidance, and then synchronically changes the structure of the two branches.

Table 4 Decoder ablation experiment

Model	Iter	mIoU	mIoU_noback(%)	allAcc(%)
base_sr	4K	72.59	83.9	95.72
base_sr_td	4K	73.98	83.46	95.06
base_sr_bu	4K	74.71	84.35	95.46

4.3.5 | Comparison with Other Methods

We compared our results with existing methods on ADE20K[17], Cityscapes. To better compare existing methods, we used four RTX3090s to train our model for the 16w era in order to maximize its effectiveness. Table 5 records how our model compares to the existing model on the results of the two data sets. We replicated some of the methods on four cards to compare with our experimental results. It lists scores from Cityscapes tests, in two widely used training Settings (by training or training +val). Appropriate model results are selected from the mmsegmentation[3] for comparison. The results show that we are 3 points higher than the classical method on ade20K data set, and 3 points higher than SETR[18] in a network where the transformer is the backbone network. For our baseline Segformer[19], we uniform the input size to 512*1024 for replication, which is easy to compare. As a result we improved by 1.48 points. On the ade20K dataset, we typically improved by 5 points compared to the classical approach, and by 2.38 points compared to the baseline segmented.

Table 5 Comparison with Other Methods

Method	Backbone	mIoU	
		Cityscapes	**ADE20K**
FCN	ResNet-101	76.58	41.4
PSPNet	ResNet-101	81.01	44.39
DeepLabv3	ResNet-101	80.2	45
OCRNet	HRNetv2	81.35	43.25
DeepLabv3+	ResNest101	79.62	46.47
PVT	Transformer	-	44.8
SETR	Transformer	79.21	47.34
Segformer	Transformer	80.78	48.8
Ours	Transformer	82.26	51.18

5 | CONCLUSION

Aiming at the semantic segmentation algorithm based on Transformer, this paper designs a super resolution recovery auxiliary branch to assist the semantic segmentation task. The network uses Transformer as a feature extractor to carry out semantic segmentation task and super-resolution recovery task at the same time. The feature map of auxiliary task recovery is used to supplement specific details with semantic segmentation feature map. We analyzed the similarities between the two tasks. Various attempts were made to design appropriate decoding branches, and a bottom-up feature fusion module was used to integrate multi-layer feature information, which improved the accuracy without spending extra time and cost in the prediction. The effectiveness of various components was verified on Cityscapes data set. The same type of high-density prediction tasks can play a similar auxiliary role, including panoramic segmentation, pose estimation and other scene understanding tasks. How to take into account the commonality of multitasking and optimize for specific tasks is also a meaningful direction to try.

References

1. Wang J, Sun K, Cheng T, et al. Deep High-Resolution Representation Learning for Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2019; 43: 3349-3364.

2. Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2014: 3431-3440.

3. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *ArXiv* 2015; abs/1505.04597.

4. Badrinarayanan V, Kendall A, Cipolla R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2015; 39: 2481-2495.

5. Lin G, Milan A, Shen C, Reid ID. RefineNet: Multi-path Refinement Networks for High-Resolution Semantic Segmentation. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2016: 5168-5177.

6. Zhao H, Shi J, Qi X, Wang X, Jia J. Pyramid Scene Parsing Network. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2016: 6230-6239.

7. Chen LC, Papandreou G, Kokkinos I, Murphy KP, Yuille AL. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2016; 40: 834-848.

8. Zhao H, Qi X, Shen X, Shi J, Jia J. ICNet for Real-Time Semantic Segmentation on High-Resolution Images. *ArXiv* 2017; abs/1704.08545.

9. Woo S, Park J, Lee JY, Kweon IS. CBAM: Convolutional Block Attention Module. In: ; 2018.

10. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ArXiv* 2020; abs/2010.11929.

11. Zheng S, Lu J, Zhao H, et al. Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 2020: 6877-6886.

12. Cheng B, Schwing AG, Kirillov A. Per-Pixel Classification is Not All You Need for Semantic Segmentation. In: ; 2021.

13. Dong C, Loy CC, He K, Tang X. Learning a Deep Convolutional Network for Image Super-Resolution. In: ; 2014.

14. Kim J, Lee JK, Lee KM. Accurate Image Super-Resolution Using Very Deep Convolutional Networks. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2015: 1646-1654.

15. Shi W, Caballero J, Huszár F, et al. Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2016: 1874-1883.

16. Zhou B, Zhao H, Puig X, Fidler S, Barriuso A, Torralba A. Semantic Understanding of Scenes Through the ADE20K Dataset. *International Journal of Computer Vision* 2016; 127: 302-321.

17. Saenz-Gamboa JJ, Iglesia-Vayá d. lM, Gómez JA. Automatic Semantic Segmentation of Structural Elements related to the Spinal Cord in the Lumbar Region by using Convolutional Neural Networks. *2020 25th International Conference on Pattern Recognition (ICPR)* 2021: 5214-5221.

18. Lipkin BS, Rosenfeld A. Picture Processing and Psychopictorics. *Transactions of the American Microscopical Society* 1970; 91: 244.

19. Harris CG, Stephens MJ. A Combined Corner and Edge Detector. In: ; 1988.

20. Bevilacqua M, Roumy A, Guillemot CM, Alberi-Morel ML. Low-Complexity Single-Image Super-Resolution based on Nonnegative Neighbor Embedding. In: ; 2012.

21. Lim B, Son S, Kim H, Nah S, Lee KM. Enhanced Deep Residual Networks for Single Image Super-Resolution. *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* 2017: 1132-1140.

AUTHOR BIOGRAPHY

Zhichao Lian. Zhichao Lian received the Bachelor's and Master's degrees in computer science from Jilin University, Changchun, China, in 2005 and 2008, respectively, and the Ph.D. degree from Nanyang Technological University, Singapore, in 2013. From 2012 to 2014, he was a Postdoctoral Associate with the Department of Statistics, Yale University. He is currently an Associate Professor with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China. His research interests include machine learning , explainable AI and cyberspace security.

Ling Wang. Ling Wang received the BS degree in computer science and technology from Nanjing University of Science and Technology, NanJing, China, in 2021. She is currently pursuing the MS degree in School of Computer Science and Engineering, Nanjing University of Science and Technology. Her research is deepfake detection.

How to cite this article: Williams K., B. Hoskins, R. Lee, G. Masato, and T. Woollings (2016), A regime analysis of Atlantic winter jet variability applied to evaluate HadGEM3-GC2, *Q.J.R. Meteorol. Soc.*, *2017;00:1–6*.